Universidade de Brasília

Mixed Finite Element Methods for Steady Granular Flows: Numerical Analysis and Applications

Saulo Rodrigo Medrado

Advisor: Prof. Dr. Yuri Dumaresq Sobral

Co-Advisor: Prof. Gabriel N. Gatica Pérez

Department of Mathematics
University of Brasília

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematics

Brasília, April 23, 2025

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Medrado, Saulo Rodrigo

MM492mm Mixed Finite Element Methods for Steady Granular Flows: Numerical Analysis and Applications / Saulo Rodrigo Medrado; orientador Yuri Dumaresq Sobral; co-orientador Gabriel Nibaldo Gatica. Brasília, 2025.
160 p.

Tese(Doutorado em Matemática) Universidade de Brasília, 2025.

1. Finite element, Rayleigh-Ritz, and Galerkin methods for boundary value problems (PDEs). 2. Navier-Stokes equations (incompressible viscous fluids). 3. Stability and convergence of numerical methods for PDEs. 4. Multiphase flows (gas-liquid, particle-laden, etc.). 5. PDEs in fluid mechanics (non-Newtonian fluids, turbulence, etc.). I. Sobral, Yuri Dumaresq, orient. II. Gatica, Gabriel Nibaldo, co-orient. III. Título.

Mixed Finite Element Methods for Steady Granular Flows: Numerical Analysis and Applications

by

Saulo Rodrigo Medrado*

Thesis presented to the Department of Mathematics at the University of Brasília, as part of the requirements for obtaining the degree of

Ph.D. in Mathematics

Brasília, April 23, 2025.

Examining Committee:

Yuri Dumaresq Sobral

Prof. Dr. Yuri Dumaresq Sobral

(Advisor)

Prof. Dr. Gabriel Nibaldo Gatica – DIM/UdeC-CL

(Co-Advisor)

Prof. Dr. Antonio Cesar Pinho Brasil Junior – FT/UNB-BR

(Member)

Prof. Dr. Maicon Ribeiro Correa – DMA/IMECC/UNICAMP-BR

(Member)

Rafael Alves Bonfim de Queiroz

Prof. Dr. Rafael Alves Bonfim de Queiroz– DECOM/UFOP-BR

(Member)

* The author was a CAPES fellowship recipient during the preparation of this thesis.

Acknowledgements

To Professor Yuri Dumaresq Sobral, for the conception, support, and coordination of this work, as well as for the trust placed in the author. His contribution was essential for the beginning of this research.

To Professor Gabriel Nibaldo Gatica Pérez, likewise for the conception, support, and coordination of this work, and for sharing the knowledge that enabled its development. His contribution was fundamental to the realization of this research.

To Professor Sergio Andrés Caucao Paillán, for his significant collaboration, without which the completion of this work would not have been possible.

And finally, to the esteemed reader, for your interest in this work.

Abstract

We propose and analyze new mixed finite element methods for a regularized $\mu(I)$-rheology model of granular flows with an equivalent viscosity depending nonlinearly on the pressure and the norm of the strain rate tensor. To this end, and besides the velocity, the pressure and the strain rate, we introduce a modified stress tensor, and the skew-symmetric vorticity, as auxiliary tensor unknowns, thus yielding a mixed variational formulation within a Banach spaces framework. The pressure is obtained through an postprocess suggested by the incompressibility condition of the fluid. A fixed-point strategy combined with a solvability result for a class of nonlinear twofold saddle point operator equations in Banach spaces, are employed to show, along with the classical Banach fixed-point theorem, the well-posedness of the continuous and discrete formulations. In particular, PEERS and AFW elements of order $\ell \geq 0$ for the stress, the velocity, and the skew-symmetric vorticity, and piecewise polynomials of degree $\leq \ell + n$ (resp. $\leq \ell + 1$) for the strain rate with PEERS (resp. with AFW), yield stable Galerkin schemes. Optimal a priori error estimates are derived and associated rates of convergence are established. Numerical results confirming the latter and illustrating the good performance of the methods, are reported. Additionally, we develop the first reliable and efficient residual-based *a posteriori* error estimator for its associated mixed finite element scheme in both 2D and 3D. For the reliability analysis, we employ the first-order Gâteaux derivative of the global operator involved in the problem, a stable Helmholtz decomposition in Banach spaces, and local approximation properties of the Raviart–Thomas and Clément interpolants. In turn, the localization technique based on bubble functions in local L^p-spaces, and results from previous works are the main tools yielding the efficiency estimate. Numerical examples illustrating the performance of the associated adaptive algorithms are reported.

Keywords: granular flows, nonlinear viscosity, mixed finite elements, twofold saddle point, fixed-point theory, a priori error analysis, a posteriori error analysis, reliability, efficiency.

Mathematics Subject Classifications (2020): 65N30, 65N12, 65N15, 47H10, 47J26, 76D05, 76T25, 76R05, 35Q79.

Resumo

Propomos e analisamos um novo método de elementos finitos mistos para um modelo regularizado de reologia $\mu(I)$ de escoamentos granulares, com viscosidade equivalente dependendo não linearmente da pressão e do tensor taxa de deformação. Para isso, além da velocidade, da pressão e da taxa de deformação, introduzimos um tensor de tensão modificado e a vorticidade como incógnitas auxiliares, obtendo uma formulação variacional mista no contexto de espaços de Banach. A pressão é calculada através de um pós-processamento sugerido pela condição de incompressibilidade do fluido. Uma estratégia de ponto fixo, combinada com um resultado de solubilidade para uma classe de equações de operadores não lineares de ponto de sela duplo em espaços de Banach, é empregada para demonstrar a boa colocação das formulações contínua e discreta. Em particular, os elementos PEERS e AFW de ordem $\ell \geq 0$ para tensão, velocidade e vorticidade antissimétrica, e polinômios por partes de grau $\leq \ell + n$ (resp. $\leq \ell + 1$) para a taxa de deformação com PEERS (resp. AFW), produzem esquemas de Galerkin estáveis. Estimativas de erro *a priori* ótimas e taxas de convergência associadas são estabelecidas, com resultados numéricos confirmando sua validade e ilustrando o bom desempenho dos métodos. Adicionalmente, desenvolvemos o primeiro estimador de erro residual *a posteriori* confiável e eficiente para o esquema de elementos finitos mistos associado, em 2D e 3D. Para a análise de confiabilidade, utilizamos a derivada de Gâteaux de primeira ordem do operador global do problema, uma decomposição de Helmholtz estável em espaços de Banach e propriedades de aproximação local dos interpolantes de Raviart-Thomas e Clément. Por sua vez, a técnica de localização baseada em funções *bubble* em espaços L^p locais e resultados de trabalhos anteriores são as principais ferramentas para a estimativa de eficiência. Exemplos numéricos ilustram o desempenho dos algoritmos adaptativos associados.

Palavras-chave: fluxos granulares, viscosidade não linear, elementos finitos mistos, ponto de sela duplo, teoria do ponto fixo, análise de erro *a priori*, análise de erro *a posteriori*, confiabilidade, eficiência.

Classificações de Assunto (2020): 65N30, 65N12, 65N15, 47H10, 47J26, 76D05, 76T25, 76R05, 35Q79.

Título em português: Métodos de Elementos Finitos Mistos para Escoamentos Granulares Estacionários: Análise Numérica e Aplicações

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Granular Materials

Granular materials consist of macroscopic particles, visible to the naked eye, ranging in size from a few micrometers to several millimeters or more.

As a first example of granular materials, we highlight the production of grains and cereals. The origins of cereal cultivation date back approximately 10,000 years, a process involving transportation, storage, and currently, industrialization. According to reports by the Food and Agriculture Organization (FAO) [3], cereals — corn, rice, wheat, barley, and sorghum — lead global agricultural production with 3.1 billion tons. In comparison, other crops show smaller volumes: 2.3 billion for sugar, 1.2 billion for vegetables, 1.2 billion for oilseeds, 1 billion for fruits, and 900 million for roots and tubers. It is further clarified that improving cereal production techniques will be essential to meet the increasing demand by 2033. The report emphasizes that around 70% of the projected increase will come from grains productivity, not territorial expansion. This highlights the need to refine agricultural techniques to reduce losses and enhance efficiency (Food and Agriculture Organization of the United Nations [3]).

Another significant example is sand mining. Sand, considered the second most extracted resource in the world, is widely used from construction to the manufacture of silicon chips. Global sand consumption reaches between 40 and 50 billion tons per year, with demand tripling in recent decades due to urbanization and infrastructure growth. Most of this resource is extracted from rivers, beaches, and seabeds (Gallagher and Peduzzi [4]).

Numerous other examples of granular materials are found in industrial processes, such as in pharmaceutical, agricultural inputs industries, in construction, in coastal sediment management through the addition or removal of sand on beaches, in geological phenomena like landslides, desert dynamics, and hail (Fig. 1.1).

Considering the importance and abundance of these materials, it is essential to study the mechanics involved in their flows, allowing for the optimization of the treatment, management, and processing of granular materials, contributing to greater efficiency, cost reduction, and sustainability in their industrial applications.

However, the physics of granular materials is challenging due to their hybrid behavior between solid, liquid, and gas, to their disordered nature, and to the nonlinear interactions between particles. Factors such as friction, cohesion, and energy dissipation complicate their modeling. Moreover, the absence of a unified theory and the variation in behavior at different scales require specific and interdisciplinary approaches for their study (Andreotti et al. [2]).

Figure 1.1 Examples of granular materials. (a) Image of grains from agricultural production: corn, barley, rice, wheat, millet, beans, lentils; (b) a silo is a closed storage structure for granular material such as grains, cement, etc.; (c) loading of a bulk carrier ship; (d) sand production by the Brazilian company Vale S.A.; (e) production of medication capsules; (f) landslide in 2022 in the state of Santa Catarina, Brazil.

1.2 Discrete models

There are two types of mathematical models widely used to study granular materials: discrete models and continuous models (studied in this work). Discrete Element Method

(DEM) are essential tools for studying the behavior of granular media. In this context, the work of Cundall and Strack [1] is considered one of the first revolutionary studies on the mathematical and computational modeling of granular flows. They proposed a method to model the movement of individual granular particles, where each particle is deformable and interacts with others primarily by contact, including friction, collisions, and cohesive effects, for example. In this model, interactions between particles are explicitly represented, considering each contact individually. Particles only interact when there is an overlap, treated as a small deformation at the contact point (Figure 1.2 (a)).

Contact forces are divided into normal forces, F_N, and tangential forces, F_T. The normal force is proportional to the normal overlap, δ_N, and is modeled by a spring and a viscous damper, while the tangential force is proportional to the tangential displacement, δ_T, and is represented by a spring coupled with a sliding block to capture Coulomb friction, which depends on the friction coefficient and cohesion between particles (Figure 1.2 (a) and (b)). The movement of the particles is calculated using Newton's Second Law, where acceleration is determined by the sum of contact forces and gravitational force. Position and velocity are updated numerically Cundall and Strack [1]. Furthermore, the model considers interactions with rigid boundaries, calculating contact forces similarly to particle-particle interactions.

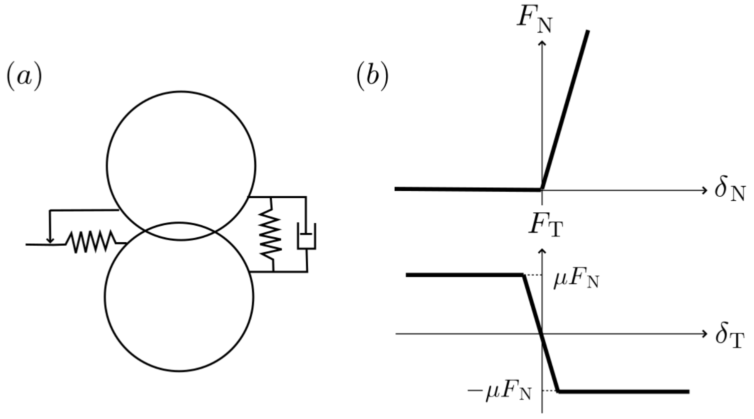


Figure 1.2 (a) Interactions between particles implemented in the method of Cundall and Strack [1]. (b) Normal and tangential forces as functions of the relative normal and tangential displacements (Andreotti et al. [2]).

With this model, various flows of granular materials have been studied, yielding remarkable results, as highlighted by Andreotti et al. [2]. These advancements have allowed a deeper understanding of the dynamics of granular materials, including phenomena such as the formation of force chains, the transition between flow regimes, and behavior under different loading conditions. However, the simulation of large-scale granular flows still represents a significant computational challenge. The processing power required to model a sufficient number of particles to capture all relevant features of these flows is not yet widely available. This limitation restricts the application of DEM in large-scale problems, such as flows in silos, landslides, or large-scale industrial processes. Thus, while DEM has revolutionized the study of granular materials, the search for more efficient and scalable methods remains an active area of research.

1.3 Continuum models

The idea of proposing continuous equations, similar to the Navier-Stokes equations for Newtonian fluids, has always attracted researchers of granular materials. One of the first significant contributions was made by Savage and Hutter [5], who proposed conservation equations for mass and momentum. In this model, it is considered that the pressure within the granular layer depends only on depth, due to the small aspect ratio (height much smaller than horizontal extent). The model assumes that the dissipative nature is determined by basal friction, which follows Coulomb's law, associating the shear force with the normal weight on the inclined plane, based on the friction coefficient and inclination. The layer of granular material is considered thin relative to the horizontal extent, which allows simplifying the equations using the shallow water hypothesis and ignoring less relevant terms. Finally, it is assumed that the velocity along the depth is uniformly distributed, except in regions close to the base.

But it was the study conducted by GDR-MiDi-Group [6] that established consistent rheological measurements of dense granular flow properties, unifying experimental and numerical data obtained in six distinct geometric configurations: plane shear, annular shear, vertical channel flow, inclined plane flow, heap flow, and rotating drum. The goal was to identify general patterns, even amidst differences in experimental or simulation conditions.

Granular flows were characterized into three main regimes, based on the relationship between inertial and confinement effects, using the inertial number I, defined as

$$I = \frac{\sqrt{2}\, d\, |\mathbf{D}|}{\sqrt{p/\rho}} \,,$$

where d is the particle diameter, $\mathbf{D}$ is the shear rate, p is the confining pressure, and ρ the material density. This dimensionless number allows distinguishing the following regimes: quasi-static regime, dominated by stable contact networks between particles, where inertia is negligible; dense inertial regime, where there is balance between inertial and contact forces; gaseous regime, dominated by binary collisions at high agitation rates.

It was proposed by Jop et al. [7] that the dissipative nature of granular flows is intrinsically associated with the frictional behavior between particles. In granular flows, mechanical energy is continuously dissipated mainly through friction and inelastic collisions, which makes these systems highly dissipative, especially in dense flow regimes. To describe this behavior, a constitutive relation for the friction coefficient based on the inertial number I is defined as:

$$\mu(I) \;:=\; \mu_s + \left(\frac{\mu_d - \mu_s}{I + I_0}\right) I\,,$$

where μ_s is a critical parameter that defines the static friction coefficient, setting the critical threshold for dense granular flow to initiate. The parameter μ_d is the upper limit value of the friction coefficient when flow occurs at high shear rates, reflecting the dynamic equilibrium in the high shear regime. The parameter I_0 is an adjustment parameter that controls the transition between low and high shear regimes. Analogous to viscoplastic fluids, the shear stress τ is a generalization of Coulomb's law, written as:

$$\tau = \mu(I)p\frac{\mathbf{D}}{|\mathbf{D}|}\,,$$

thus, the effective viscosity η of the granular continuum is defined by:

$$\eta(p, |\mathbf{D}|) \;:=\; \frac{\sqrt{2}\mu(I)p}{|\mathbf{D}|}\,.$$

In the complete formulation of the continuum mechanics, the Cauchy momentum conservation equation is:

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} \;+\; (\nabla\mathbf{u})\mathbf{u}\right) \;=\; \mathbf{div}(\boldsymbol{\sigma}) \;+\; \rho\,\mathbf{g}\,,$$

where ρ is the material density, $\mathbf{g}$ is gravity, and with the stress tensor $\boldsymbol{\sigma}$ decomposed as:

$$\boldsymbol{\sigma} \;=\; \underbrace{\eta(p, |\mathbf{D}|)\,\mathbf{D}}_{\text{deviatoric term}} \;-\; \underbrace{p\,\mathbb{I}}_{\text{isotropic term}}\quad,$$

where the deviatoric term captures viscous effects due to friction. In this model, friction plays a crucial role in the formation of static zones, where the material exhibits solid behavior, and in defining the effective viscosity, which increases with applied pressure. Although other mechanisms, such as inelastic collisions and drag caused by interstitial fluids, also contribute to energy dissipation, friction is dominant in the dense flow regime. This frictional model, besides describing energy dissipation, governs transitions between solid and liquid states, being essential for modeling granular flows in geophysical and industrial contexts.

The approach proposed by Jop et al. [7] provides a robust quantitative basis and is validated by experiments in complex three-dimensional configurations, demonstrating its applicability in a wide range of scenarios, as we will see next.

1.4 Numerical methods for continuum models

In the study presented by Jop et al. [7], the Finite Difference Method was employed, validated through experimental data obtained from flows on inclined planes and in an inclined channel with rough walls. The experiments measured surface inclination, velocity profiles, and flow layer thickness. The simulations demonstrated high accuracy in reproducing experimental results, confirming the effectiveness of the method and the constitutive law used to describe dense granular flows in complex configurations.

Since the publication of Jop et al. [7], numerous studies have explored numerical solutions for the $\mu(I)$ rheology equations in different physical scenarios, for example Lagrée et al. [8], Staron et al. [9], Chauchat and Médale [10], Franci and Cremonesi [11], Yang et al. [12, 13], whose main characteristics we highlight below.

In the work of Lagrée et al. [8], the problem of granular column collapse was investigated. For this, the Volume of Fluid (VOF) method was used, an effective technique for tracking and modeling interfaces between different phases, as in this case, the interaction between grains and air. The VOF method is based on the concept of volume fraction c, which indicates the proportion of each phase within the computational mesh element, that is, $c = 1$ for the granular fluid, and $c = 0$ for air, and when $0 < c < 1$, the element comprises the grain-air interface. The volume fraction is transported by the velocity field $\mathbf{u}$ through the advection equation,

$$\frac{\partial c}{\partial t} + \nabla \cdot (c\mathbf{u}) = 0 \,.$$

Additionally, the density ρ_{VOF} and viscosity η_{VOF} of the grain-air mixture are calculated as weighted and harmonic averages, respectively, of the properties of each phase:

$$\rho_{VOF} = c\rho_{\text{grains}} + (1-c)\rho_{\text{air}},$$

$$\eta_{VOF} = \frac{1}{c/\eta_{\text{grains}} + (1-c)/\eta_{\text{air}}}.$$

Note that if $c = 1$, then $\rho_{VOF} = \rho_{\text{grains}}$ and $\eta_{VOF} = \eta_{\text{grains}}$, and if $c = 0$ then $\rho_{VOF} = \rho_{\text{air}}$ and $\eta_{VOF} = \eta_{\text{air}}$ allowing for an accurate representation of the fluid behavior in each region. Note that $\eta_{\text{grains}} \gg \eta_{\text{air}}$, this means that at the interface between the two fluids, the mixture flow is governed by air, which has less resistance to movement. The harmonic average captures this behavior, as it gives more weight to the fluid with lower viscosity. This is important to ensure that the average viscosity at the interface is consistent with the physics of the problem. Two configurations were used to validate simulations: a stationary and incompressible granular layer on an inclined plane, with analytical solutions for velocity $\mathbf{u}$ and pressure p; and a granular layer under a Newtonian fluid on an inclined plane, simulating a free surface and tracking the interface via the VOF method. Comparisons were made with analytical and semi-analytical solutions, the latter obtained by solving the momentum equations for each layer, considering the boundary conditions at the interface, validating the approach and preparing the model for more complex problems, such as granular column collapse.

The method was applied to the collapse of granular columns in two dimensions, with different aspect ratios, and compared with discrete simulations. The results showed good agreement in the temporal evolution of the column shape and internal deformations, although the model systematically underestimated the final flow runout, especially for taller columns.

In the work of Staron et al. [9], models of confined flows in silos were analyzed. Instead of regularizing the effective viscosity of the granular material, to avoid physically unrealistic and numerically problematic situations, a limitation of η_{grains} by a parameter $\eta_{\max}$ was considered,

$$\eta_{\text{grain}} = \min\left\{\frac{\mu(I)p}{|\mathbf{D}|}, \eta_{\max}\right\}.$$

The VOF method's density and viscosity equations, in this case, are given by,

$$\frac{\partial c}{\partial t} + \nabla \cdot (c\mathbf{u}) = 0\,,$$

$$\rho_{VOF} = c\rho_{\text{grains}} + (1-c)\rho_{\text{air}}\,,$$

$$\eta_{VOF} = c\eta_{\text{grains}} + (1-c)\eta_{\text{air}} \,.$$

The validation compares the $\mu(I)$ model with discrete models for granular flow in silos. The continuum model qualitatively captures the discharge rate, velocity, and pressure. However, discrepancies may occur in areas of slow deformation.

In turn, Chauchat and Médale [10] proposed a three-dimensional model based on $\mu(I)$ rheology, using the Finite Element Method (FEM) with primal formulation. They used d as the length scale, $\sqrt{d/|\mathbf{g}|}$ as the time scale, and $\rho|\mathbf{g}|d$ as the stress scale. Four different regularization methods were studied:

Simple Regularization

$$\eta_p^s = \frac{\mu_s p}{|\mathbf{D}| + \epsilon} + \frac{(\mu_d - \mu_s)p}{I_0\sqrt{p} + |\mathbf{D}| + \epsilon},$$

where ϵ is a small regularization parameter.

Mixed Bercovier-Engelman Regularization

$$\eta_p^{be} = \frac{\mu_s p}{\sqrt{|\mathbf{D}|^2 + \epsilon^2}} + \frac{(\mu_d - \mu_s)p}{I_0\sqrt{p} + |\mathbf{D}| + \epsilon}.$$

Mixed Papanastasiou Regularization

$$\eta_p^{papa} = \mu_s p \frac{1 - e^{-|\mathbf{D}|/\epsilon}}{|\mathbf{D}|} + \frac{(\mu_d - \mu_s)p}{I_0\sqrt{p} + |\mathbf{D}| + \epsilon}.$$

Chauchat-Médale Regularization (based on Bercovier-Engelman)

$$\eta_p^{mc} = \left[\mu_s + \frac{(\mu_d - \mu_s)|\mathbf{D}|}{I_0\sqrt{p} + |\mathbf{D}|}\right] \frac{p}{\sqrt{|\mathbf{D}|^2 + \epsilon^2}}.$$

The numerical model was validated against analytical solutions for vertical chute and inclined plane flows, showing excellent agreement with theoretical velocity profiles. Applied to granular heap flows and cylinder interactions, it accurately predicted flow behavior and drag forces while demonstrating 30-50% faster convergence than non-regularized approaches. Though limited to moderate deformations by its fixed mesh, the method's stability and efficiency make it particularly suitable for industrial granular transport and geophysical flow simulations.

In the work of Franci and Cremonesi [11] two regularizations are considered and applied only to the first term of η, since this is responsible for the divergent behavior:

Exponential Regularization

$$\eta^E = \frac{p\mu_s(1 - e^{-|\mathbf{D}|/\epsilon})}{|\mathbf{D}|} + \frac{pd(\mu_d - \mu_s)}{I_0\sqrt{p/\rho} + |\mathbf{D}|}.$$

Penalty Regularization

$$\eta^p = \frac{p\mu_s}{\sqrt{|\mathbf{D}|^2 + \epsilon^2}} + \frac{pd(\mu_d - \mu_s)}{I_0\sqrt{p/\rho} + |\mathbf{D}|}.$$

The problem was implemented in the *Particle Finite Element Method* (PFEM), suitable for handling large deformations and free surfaces, which operates in iterative cycles: first, the domain is discretized into particles carrying material information; then, a finite element mesh is generated from these particles using algorithms. The governing equations are solved on the mesh, and the particles are moved and updated based on the results. The mesh is reconstructed at each time step, allowing the method to handle large deformations, enabling the simulation of complex granular flows with free surfaces. For validation, the authors simulated the collapse of granular columns in 2D and 3D, comparing the results with experimental data and methods such as DEM. The results showed good agreement with experiments, confirming the accuracy of the models and the effectiveness of the regularizations, which improved the conditioning of the linear system and allowed the use of iterative solvers even on refined meshes.

In the study by Yang et al. [12], the authors developed the LBGrain model, combining the Lattice Boltzmann Method (LBM) with $\mu(I)$ rheology to efficiently simulate granular flows with free surfaces. The LBM, which models fluid dynamics through the evolution of distribution functions on a structured mesh, proved significantly faster (up to 23×) than traditional Navier-Stokes-based methods. The treatment of the grain-air interface was simplified through dynamic cell classification (fluid, empty, or interface), avoiding the need to explicitly resolve the gas phase. Granular collapse simulations showed excellent agreement with DEM results, outperforming models with Bingham rheology. The approach demonstrated potential for large-scale applications, such as geophysical landslide simulations, with prospects for extension to 3D and inclusion of non-local effects.

In the work of Yang et al. [13], the problem of granular column collapse was investigated using the LBM with a new friction boundary condition. To model dense granular flow, the $\mu(I)$ rheology was implemented, which describes the material behavior as a viscoplastic fluid, where shear stress is limited by the Coulomb criterion. The proposed friction boundary condition calculates the wall slip velocity based on the Coulomb criterion, which states that slip occurs when the wall shear stress exceeds the wall friction coefficient. For validation, a planar Couette flow was simulated, where a fluid is sheared between two parallel plates, with the upper plate moving and the lower plate stationary and frictional. In this study, the lower plate was modeled with the new friction condition, reproducing analytical velocity profiles and capturing the transition between no-slip and partial slip regimes. The model was then applied to 2D granular column collapse, comparing with DEM simulations. The extended LBM model (LBGrain) accurately predicted the temporal evolution of the column shape and internal flow structures for different initial aspect ratios and inclination angles. The approach proved computationally efficient and generalizable to complex problems, such as avalanches.

All the analyzed studies were fundamental for the development and consolidation of the $\mu(I)$ rheology. In particular, the use of classical numerical methods played a crucial role, as it allowed consolidating the proposed mathematical model, validating its ability to predict complex behaviors in granular flows in the examples considered in each work. These methods provided precise tools to simulate and analyze phenomena such as phase transitions, flow regimes, and particle interactions, demonstrating the robustness and versatility of the $\mu(I)$ model in the studied scenarios. Thus, numerical methods not only reinforced the theoretical foundation of the model but also expanded its applicability in practical and complex contexts.

1.5 Mixed finite elements

The $\mu(I)$-rheology model presents significant numerical challenges due to its pressure-dependent dissipative terms, which complicate the application of classical pressure-correction schemes Hinch [14] and the classical primal finite element methods designed for linear problems. Recent advances in Banach spaces-based mixed formulations have proven particularly effective for analyzing nonlinear continuum mechanics problems, as demonstrated by applications to Brinkman-Forchheimer, Darcy-Forchheimer, Navier-Stokes, Boussinesq, and coupled flow-transport, and fluidized beds are some of the respective models addressed, and a non-exhaustive list of the corresponding references includes Benavides et al. [15], Camaño et al. [16], Caucao et al. [17], Caucao and Yotov [18], Colmenares et al. [19, 20], Colmenares and Neilan [21], Gatica et al. [22]. The most distinctive feature of a mixed formulation is the incorporation of additional unknowns, usually depending on the original ones of the model, for either physical or analytical reasons, obtaining a saddle-point problem where the associated operator have the form

$$\begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix}.$$

These mixed approaches offer several advantages: they eliminate the need for artificial augmentation techniques required in classical formulations, provide more physically consistent frameworks through natural function spaces, and enable momentum-conservative schemes with direct approximation of physically relevant variables. For the $\mu(I)$-rheology model specifically, this mixed Banach framework could allow direct computation of key quantities like strain rate tensor, shear rate, inertia number, and vorticity without the accuracy loss associated with numerical differentiation.

It is well known that adaptive algorithms based on *a posteriori* error estimates are particularly effective in recovering the loss of convergence orders often observed in standard Galerkin procedures, such as finite element and mixed finite element methods. This is especially true when these methods are applied to nonlinear problems, where singularities or high gradients in the exact solutions are present. In this context, the study of *a posteriori* error estimators for saddle-point problems has been widely developed in the literature by various authors (see, e.g., Ainsworth and Oden [23], Alonso [24], Carstensen [25], Carstensen and Dolzmann [26], Lonsing and Verfürth [27], Repin et al. [28], and references therein). In particular, this powerful approach has been successfully applied to the Navier–Stokes equations, both with constant and nonlinear viscosity, as well as to related models. We refer to pioneering works such as Oden et al. [29], Verfürth [30], and Verfürth [31], as well as to [32, Section 9.3], where the first contributions to derive an *a posteriori* error analysis for the incompressible Navier–Stokes problem in its classical velocity-pressure formulation were introduced. Later, the *a priori* and *a posteriori* error analysis for the dual mixed finite element method of the Navier–Stokes problem were proposed and developed in Farhloul et al. [33]. Additionally, we mention Allendes et al. [34], where the authors extend these contributions to the case of Dirac measures, and Kanschat and Schötzau [35], which provides an *a posteriori* error analysis for a Discontinuous Galerkin scheme that offers exactly divergence-free approximations of the velocity. Meanwhile, adaptive methods for augmented-mixed formulations for the Navier–Stokes problem with constant and variable viscosity were developed in Gatica et al. [36] and Camaño et al. [37], respectively. We also refer to Caucao et al. [38], where the authors developed an *a posteriori* error analysis for a fully-mixed formulation of the Navier–Stokes/Darcy coupled problem with nonlinear viscosity. In this work, a suitable first-order Gâteaux derivative of the global operator involved is employed to derive the corresponding reliability of the estimator. Furthermore, Camaño et al. [39] is particularly notable for its *a posteriori* error analysis of a momentum-conservative Banach space-based mixed finite element method for the Navier–Stokes problem. In this work, standard duality-based arguments, a suitable Helmholtz decomposition within Banach frameworks, and classical approximation properties are combined with small data assumptions to establish the reliability of the estimators. Similar techniques have been employed in Caucao et al. [40] and Gatica et al. [41] to develop reliable and efficient residual-based *a posteriori* error estimators in both 2D and 3D for Banach space-based mixed finite element methods applied to the stationary Boussinesq and Oberbeck-Boussinesq systems. Lastly, we refer to Caucao et al. [42] for a recent *a posteriori* error analysis of a Banach space-based mixed formulation for the coupled Brinkman–Forchheimer and double-diffusion equations.

1.6 Thesis objectives

The general objective of this thesis is the development and analysis of mixed finite element methods for the numerical resolution of $\mu(I)$ rheology equations applied to stationary granular flows. The main focus of this work is creating stable methods with optimal convergence and proven robustness, capable of handling the inherent complexities of the $\mu(I)$ model.

1.7 Specific objectives

1. **Presentation of the Physical Problem and Mathematical Model Formulation**

    - Discuss the most important models for granular material flow.
    - Present the physical problem of stationary granular flows and its regularized mathematical formulation, highlighting its main characteristics such as nonlinearities, singularities, and the dependence of the effective friction coefficient on the inertial number I.

2. **Mixed Variational Formulation and Choice of Functional Spaces**

    - Develop a mixed variational formulation for the problem, where the considered derivatives are in the weak sense.

- Select appropriate functional spaces for the involved variables, such as Lebesgue and Sobolev spaces, ensuring that the mathematical properties of the continuous problem are preserved.

3. **Solvability Analysis of the Associated Variational Problem**

   - Use classical theorems of functional analysis, such as the Lax-Milgram Theorem, Babuška-Brezzi Theorem, Hölder, Schwarz, and Poincaré inequalities, to demonstrate the existence and uniqueness of the dual solution associated with the mixed variational problem.

4. **Discretization of the Variational Problem**

   - Perform the discretization of the variational problem using Lagrange and Raviart-Thomas interpolants, ensuring compatibility between the discrete spaces.
   - Ensure that the discretization preserves the stability and convergence properties of the method.

5. **Solvability Analysis of the Discretized Problem, Stability and A Priori Error**

   - Demonstrate the existence and uniqueness of the solution to the discretized problem using appropriate functional analysis theorems.
   - Perform the stability analysis of the numerical method, ensuring its robustness against variations in the problem parameters.
   - Estimate the a priori error, establishing optimal convergence rates for the proposed method, independent of the problem parameters.

6. **Numerical Implementation and Validation**

   - Implement computationally the developed mixed finite element methods.
   - Validate the methods using manufactured analytical solutions that reproduce the main characteristics of the $\mu(I)$ model, such as nonlinearities, singularities, and high pressure gradients.

7. **A Posteriori Error Analysis**

   - Define local and global estimators based on the problem residuals.

- Demonstrate the efficiency and reliability of the residual estimators, ensuring they provide accurate indicators for mesh refinement.
- Implement an adaptive mesh refinement method, based on residual estimators, to improve solution accuracy in critical regions.
- Compare the adaptive method with the uniform method, showing the recovery of lost precision in regions determined by the estimators.

8. **Numerical Implementation in Practical Applications**

   - Apply the developed methods to practical problems of granular flows, such as flow regime transitions and shear zone formation.
   - Evaluate the robustness of the methods in complex situations, verifying their ability to reproduce physically observed phenomena.

1.8 Thesis organization

In Chapter 2 we begin with a brief introduction to the classical theory of mixed finite elements, considering a model variational problem originating from a problem of dissipative nature, and applying the Babuška-Brezzi Theory to show the well-posedness of the problem, both in its continuous version and in its discretized version. We propose discretized spaces based on Lagrange and Raviart-Thomas interpolations, then we perform the a priori error analysis. Finally, we implement the numerical method to an example showing the results predicted by the theory, such as stability and optimal convergence.

In Chapter 3, we propose and analyze new mixed finite element methods for a regularized $\mu(I)$ rheology model of granular flows, with an equivalent viscosity depending nonlinearly on the pressure and the Euclidean norm of the symmetric part of the velocity gradient. For this, in addition to the velocity, pressure, and aforementioned deformation rate, we introduce a modified stress tensor that includes the convective term and the antisymmetric vorticity as auxiliary tensor unknowns, resulting in a mixed variational formulation in the context of Banach spaces. Then, the pressure is obtained through an iterative post-processing suggested by the fluid incompressibility condition, which allows us to express this unknown in terms of the aforementioned stress tensor and velocity. A fixed-point strategy, combined with a solvability result for a class of nonlinear double saddle-point operator equations in Banach spaces, is employed to demonstrate, along with the classical Banach fixed-point theorem, the well-posedness of the continuous and discrete formulations. In particular, PEERS and AFW elements of order ℓ greater than or equal to 0 for the stress tensor, velocity, and antisymmetric vorticity, and piecewise polynomials of degree less than or equal to $\ell + n$ (resp. $\ell + 1$) for the deformation rate with PEERS (resp. with AFW), provide stable Galerkin schemes. Optimal *a priori* error estimates are derived, and associated convergence rates are established. Finally, numerical results confirming these estimates and illustrating the good performance of the methods are reported.

The contents of this chapter resulted in the following published article:

- [43] S. Caucao, G.N. Gatica, S.R. Medrado, and Y.D. Sobral, *Nonlinear twofold saddle point-based mixed finite element methods for a regularized $\mu(I)$-rheology model of granular materials.* Journal of Computational Physics 520 (2025) 113462.

In Chapter 4, we develop the first reliable and efficient residual *a posteriori* error estimator for the 2D and 3D versions of the mixed finite element scheme applied to $\mu(I)$ rheology. This estimator, denoted by Θ, was determined for the 2D and 3D versions of the mixed finite element methods introduced in Chapter 3. Specifically, we derive the global estimator Θ formulated in terms of computable local indicators Θ_K, each associated with an element K of a triangulation $\mathcal{T}_h$. This allows the identification of error sources and the design of an adaptive mesh algorithm to improve computational efficiency. In this context, the estimator Θ is considered efficient and reliable if there exist positive constants $C_{\mathtt{eff}}$ and $C_{\mathtt{rel}}$, independent of the mesh sizes, such that

$$C_{\mathtt{eff}}\,\Theta \,+\, \mathtt{h.o.t.} \;\leq\; \|\text{error}\| \;\leq\; C_{\mathtt{rel}}\,\Theta \,+\, \mathtt{h.o.t.}\,,$$

where $\mathtt{h.o.t.}$ represents one or more higher-order terms. For the reliability analysis, and due to the nonlinear nature of the problem, we employ the first-order Gâteaux derivative of the involved global operator, combined with small data assumptions, a stable Helmholtz decomposition in non-standard Banach spaces, and local approximation properties of Raviart-Thomas and Clément interpolants. In turn, inverse inequalities, the localization technique based on "bubble" functions in local L^p spaces, and known results from previous works are the main tools to obtain the efficiency estimate. Finally, several numerical examples confirm the theoretical properties of the estimator and illustrate the performance of the associated adaptive algorithms.

To the best of our knowledge, this work presents the first *a posteriori* error analysis of Banach space-based mixed finite element methods for the stationary $\mu(I)$ rheology equations governing granular materials.

Chapter 2

Elements of Classical Finite Element Theory

2.1 Chapter Introduction

Before addressing the main problem of this work, we will introduce the classical Finite Element Theory for an abstract problem associated with a Partial Differential Equation (PDE). The main objective is to ensure that the equivalent abstract problem can be solved approximately, guaranteeing that this solution is sufficiently accurate and that, under certain conditions, the approximate solution converges to the exact solution. The mathematical tools employed in this context are commonly explored in master's level PDE courses in Mathematics, with an emphasis on Analysis.

For readers unfamiliar with PDE-related methods, it is relevant to emphasize that in this work, the presented mathematical objects do not necessarily have a direct physical interpretation. The associated abstract problem may lack natural or intuitive justification, as sometimes occurs, for example, with concepts of vector spaces and linear transformations, frequently addressed in undergraduate Linear Algebra courses.

We emphasize that the mathematical formalism used will be explored without restriction, following the traditional approach found in the works that form the basis of this area. The idea is to direct the reader to these sources for more detailed information, as complete proofs of the theorems employed here will not be presented, except in some specific cases, such as in convergence error theorems. Among the classical references for this chapter, we highlight the works in: Raviart and Thomas [44], Ern and Guermond [45], Gatica [46], Ciarlet [47].

We will introduce some notations and definitions used in the classical theory of mixed finite elements in Section (2.1). Subsequently, in Section (2.2), we will present a variational problem associated with a PDE, highlighting its well-posedness (existence and uniqueness of the solution, as well as continuous dependence on the data). The discretization of the problem by the Galerkin method will also be performed. In Section (2.3), we will construct discretized spaces using the Lagrange and Raviart-Thomas interpolation theory, emphasizing the approximation errors of the interpolation operators. Finally, in Section (2.4), we will apply the results to a Poisson problem, implementing the method and demonstrating its effectiveness.

Preliminary notations

In what follows, Ω is a bounded domain of R^n, $n \in \Big\{2,3\Big\}$, with Lipschitz-continuous boundary Γ, and corresponding outward normal denoted $\boldsymbol{\nu}$. Then, we adopt the usual notation for Lebesgue spaces $\mathrm{L}^t(\Omega)$ and Sobolev spaces $\mathrm{W}^{l,t}(\Omega)$ and $\mathrm{W}^{l,t}_0(\Omega)$, with $l \geq 0$ and $t \in [1,+\infty)$, whose corresponding norms, either for the scalar and vectorial case, are denoted by $\|\cdot\|_{0,t;\Omega}$ and $\|\cdot\|_{l,t;\Omega}$, respectively. In particular, $\mathrm{W}^{0,t}(\Omega) = \mathrm{L}^t(\Omega)$, and when $t = 2$ we write $\mathrm{H}^l(\Omega)$ instead of $\mathrm{W}^{l,2}(\Omega)$, with the corresponding norm and seminorm denoted by $\|\cdot\|_{l,\Omega}$ and $|\cdot|_{l,\Omega}$, respectively. In addition, given any generic scalar function space M, we let $\mathbf{M}$ and $\mathbb{M}$ be its vectorial and tensorial counterparts, respectively, whereas M$'$ represents its dual space, whose norm is defined by $\|f\|_{\mathrm{M}'} := \sup\limits_{0\neq v\in \mathrm{M}} \dfrac{|f(v)|}{\|v\|_{\mathrm{M}}}$. Also, $\mathbb{I}$ stands for the identity tensor in $\mathrm{R}^{n\times n}$, and, besides denoting the absolute value in R, $|\cdot|$ stands for the norms in R^n to $\mathrm{R}^{n\times n}$. In turn, for any vector fields $\mathbf{v} = (v_i)_{i=1,n}$ and $\mathbf{w} = (w_i)_{i=1,n}$, we set the gradient, divergence, and tensor product operators, respectively, as

$$\nabla\mathbf{v} := \left(\frac{\partial v_i}{\partial x_j}\right)_{i,j=1,n}, \quad \mathrm{div}(\mathbf{v}) := \sum_{j=1}^{n} \frac{\partial v_j}{\partial x_j}, \quad \text{and} \quad \mathbf{v}\otimes\mathbf{w} := (v_i w_j)_{i,j=1,n}.$$

On the other hand, for any tensor fields $\boldsymbol{\tau} = (\tau_{ij})_{i,j=1,n}$ and $\boldsymbol{\zeta} = (\zeta_{ij})_{i,j=1,n}$, we let $\mathbf{div}(\boldsymbol{\tau})$ be the divergence operator div acting along the rows of $\boldsymbol{\tau}$, and define the transpose, the trace, the tensor inner product operators, and the deviatoric tensor, respectively, as

$$\boldsymbol{\tau}^{\mathtt{t}} = (\tau_{ji})_{i,j=1,n}, \quad \mathrm{tr}(\boldsymbol{\tau}) = \sum_{i=1}^{n} \boldsymbol{\tau}_{ii}, \quad \boldsymbol{\tau}:\boldsymbol{\zeta} := \sum_{i,j=1}^{n} \tau_{ij}\zeta_{ij}, \quad \text{and} \quad \boldsymbol{\tau}^{\mathtt{d}} := \boldsymbol{\tau} - \frac{1}{n}\mathrm{tr}(\boldsymbol{\tau})\,\mathbb{I}. \tag{2.1}$$

Furthermore, given $t \in (1, +\infty)$, we introduce the Banach space

$$\mathbb{H}(\mathbf{div}_t;\Omega) \;:=\; \Big\{\boldsymbol{\tau} \in \mathbb{L}^2(\Omega): \quad \mathbf{div}(\boldsymbol{\tau}) \in \mathbf{L}^t(\Omega)\Big\}, \tag{2.2}$$

which is endowed with the natural norm defined by

$$\|\boldsymbol{\tau}\|_{\mathbf{div}_t;\Omega} \;:=\; \|\boldsymbol{\tau}\|_{0,\Omega} \,+\, \|\mathbf{div}(\boldsymbol{\tau})\|_{0,t;\Omega} \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega)\,.$$

Then, proceeding as in [46, (1.43), Section 1.3.4], one can easily verify that the following holds for each $t \in \begin{cases} (1,+\infty) & \text{if} \quad n \,=\, 2 \\ [6/5,+\infty\,) & \text{if} \quad n \,=\, 3 \end{cases}$,

$$\langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \mathbf{v}\rangle \;=\; \int_\Omega \Big\{\boldsymbol{\tau} : \nabla\mathbf{v} \,+\, \mathbf{v}\cdot\mathbf{div}(\boldsymbol{\tau})\Big\} \qquad \forall\,(\boldsymbol{\tau},\mathbf{v}) \in \mathbb{H}(\mathbf{div}_t;\Omega)\times\mathbf{H}^1(\Omega)\,, \tag{2.3}$$

where $\langle\cdot,\cdot\rangle$ stands for the duality pairing between $\mathbf{H}^{-1/2}(\Gamma)$ and $\mathbf{H}^{1/2}(\Gamma)$.

Fundamental Concepts in Functional Analysis

Let us list some fundamental notations and definitions for this section. Consider a vector space X with a norm $\|.\|_X$.

Basic Spaces and Linear Operators

In functional analysis, a normed vector space $(X, \|\cdot\|_X)$ is called a **Banach space** when it is complete - meaning every Cauchy sequence converges in X. This completeness can be characterized through the convergence of iterative processes: a sequence $\{x_n\}$ in X converges to some limit $x \in X$ if and only if the distance between consecutive elements tends to zero. Formally, this is expressed as:

$$\lim_{n\to\infty} \|x_n - x_{n+1}\|_X = 0 \quad \Leftrightarrow \quad \lim_{n\to\infty} x_n = x.$$

An important subclass of Banach spaces are **Hilbert spaces**, where the norm is induced by an inner product. Specifically, a Banach space $(H, \langle\cdot,\cdot\rangle_H)$ is Hilbert if its norm satisfies:

$$\|x\|_H = \sqrt{\langle x, x\rangle_H}.$$

When considering linear mappings between these spaces, we say a linear form $T : X \to \mathrm{R}$ is **bounded** if there exists a constant $C > 0$ such that:

$$|T(x)| \le C\|x\|_X \quad \text{for all } x \in X.$$

The collection of all such linear and bounded forms constitutes the **dual space** of X, denoted by X':

$$X' := \{f : X \to \mathrm{R} \mid f \text{ is linear and bounded}\}.$$

This construction can be iterated to obtain the **bidual space** $X'' := (X')'$, which consists of all linear and bounded functionals on X':

$$X'' := \{g : X' \to \mathrm{R} \mid g \text{ is linear and bounded}\}.$$

The duality pairing between $x \in X$ and $f \in X'$ is often denoted by $\langle f, x\rangle_{X',X}$. There is a canonical injection $J : X \to X''$ that satisfies: given $x \in X$,

$$J(x) = \langle J(x), f\rangle_{X'',X} = \langle f, x\rangle_{X',X}, \quad \forall f \in X', \quad \text{and} \quad \|J(x)\|_{X''} = \|x\|_X\,.$$

When $J : X \to X''$ is a bijection, we say that the space X is **reflexive**, then we can indeed considerer $X = X''$. For linear operators between Banach spaces, the **transpose operator** of $T : X \to Y$ is defined as the mapping:$T^{\mathsf{t}} : Y' \to X'$ such that

$$\langle T^{\mathsf{t}} g, x\rangle = \langle g, Tx\rangle\,, \quad \forall g \in Y', \quad \forall x \in Y.$$

Regarding order structure, for any subset $Y \subseteq X$, the **infimum** $\inf(Y)$ represents the greatest lower bound of Y in X, while the **supremum** $\sup(Y)$ is the least upper bound. When these bounds belong to Y itself, they coincide with the minimum and maximum of Y respectively.

2.2 Operator Equation

Operators in infinite-dimensional spaces generalize fundamental concepts of linear algebra, such as matrices and eigenvalues, to broader contexts, such as Hilbert spaces and Banach spaces. We will define a variational problem, which involves bounded bilinear forms, and obtain the associated operator problem. Then, we will establish the main theorems that guarantee the solvability of the operator problem. Finally, we will apply the Galerkin Method, which is based on the idea of approximating the solution of a variational problem using finite-dimensional subspaces of the original problem spaces.

2.2.1 Abstract Problem

Let H and Q be two Banach spaces. Consider the bounded bilinear forms $a : H \times H \to \mathrm{R}$ and $b : H \times Q \to \mathrm{R}$ and the linear and bounded forms $f : H \to \mathrm{R}$ and $g : Q \to \mathrm{R}$. We want to solve the following variational problem:

$$\begin{cases} \text{Find } \sigma \in H \text{ and } u \in Q \text{ such that} \\ \begin{array}{lcl} a(\sigma,\tau) + b(\tau,u) & = & f(\tau), \quad \forall \tau \in H, \\ b(\sigma,v) & = & g(v), \quad \forall v \in Q. \end{array} \end{cases} \tag{2.4}$$

We say the variational problem (2.4) is well-posed when its solution exists, is unique, and depends continuously on the data f and g. If this problem can be solved, then it makes sense to continue the procedure. To verify if the problem (2.4) is well-posed, we need to consider an abstract problem associated with (2.4), which is obtained by defining the operators $A : H \to H'$ and $B : H \to Q'$, such that

$$\begin{array}{rcl} A : H & \to & H' \\ \sigma & \mapsto & A(\sigma) : \quad H \to \mathrm{R} \\ & & \qquad\qquad \tau \mapsto \langle A\sigma, \tau \rangle_{H',H} = a(\sigma,\tau)\,, \end{array}$$

and

$$\begin{array}{rcl} B : H & \to & Q' \\ \sigma & \mapsto & B(\sigma) : \quad Q \to \mathrm{R} \\ & & \qquad\qquad v \mapsto \langle B\sigma, v \rangle_{Q',Q} = b(\sigma,v). \end{array}$$

If Q is reflexive, i.e., $Q'' = Q$, we can define the transpose operator of B, $B^{\mathsf{t}} : Q'' = Q \to H'$ where:

$$\begin{array}{rcl} B^{\mathsf{t}} : Q & \to & H' \\ u & \mapsto & B^{\mathsf{t}}(u) : \quad Q \to \mathrm{R} \\ & & \qquad\qquad u \mapsto \langle \tau, B^{\mathsf{t}} u \rangle_{H,H'} = b(\tau,u). \end{array}$$

Then the problem (2.4) can be written as:

$$\begin{cases} \text{Find } \sigma \in H \text{ and } u \in Q \text{ such that} \\ \langle A\sigma, \tau \rangle_{H',H} + \langle B^{\mathsf{t}} u, \tau \rangle_{Q',H} \;=\; f(\tau), \quad \forall \tau \in H \\ \langle B\sigma, v \rangle_{H',Q} \;=\; g(v) \quad \forall v \in Q. \end{cases}$$

and the abstract problem associated with (2.4) becomes:

$$\begin{cases} \text{Find } \sigma \in H \text{ and } u \in Q \text{ such that} \\ A\sigma + B^{\mathsf{t}} u \;=\; f, \quad \text{in } H' \\ B\sigma, \;=\; g, \quad \text{in } Q'. \end{cases} \tag{2.5}$$

The equivalence between (2.4) and (2.5) is in the sense that $\sigma \in H$ and $u \in Q$ are a solution of the first if and only they are is a solution of the second. Therefore, to verify if the problem (2.4) is well-posed, we will study the associated abstract problem (2.5). For this, consider the kernel of the operator B, denoted by $\operatorname{Ker}(B)$ and defined by

$$\operatorname{Ker}(B) := \{\tau \in H;\, B\tau = 0\} \;=\; \{\tau \in H;\, b(\tau, v) = 0,\, \forall\, \tau \;\in Q\}.$$

Also consider the projection operator $\Pi : H \to \operatorname{Ker}(B)$ such that the composition $\Pi A : \operatorname{Ker}(B) \subset H \to \operatorname{Ker}(B)'$ is given by $\langle \Pi A\sigma, \tau \rangle_{H',H} := a(\Pi\sigma, \tau)$ with $\sigma \in \operatorname{Ker}(B)$, $\tau \in \operatorname{Ker}(B)$. Note that we could have defined $\Pi A : H \to H'$, but we prefer to be more objective. Thus:

$$\langle \Pi A\sigma, \tau \rangle_{H',H} = \langle A\sigma, \tau \rangle_{H',H}, \quad \forall\, \sigma, \tau \;\in \operatorname{Ker}(B).$$

We are now in a position to state the theorem of existence and uniqueness of solutions to the problem (2.4):

Theorem 2.2.1. The problem (2.4) is well-posed if and only if:

1. $\Pi A : \operatorname{Ker}(B) \to \operatorname{Ker}(B)'$ is injective and norm-preserving (Isomorphism);

2. $B : H \to Q$ is surjective.

The proof that (2.2.1) derives from the Gauss technique for systems of linear equations, where the upper triangular matrix is solved by back substitution followed by the application of the Closed Range Theorem for Banach spaces that relates the image of the operator B^{t} with the kernel of B, $\operatorname{Ker}(B)$ (cf. [45, Theorem A.34]). See details in [45, Theorem A.56].

Note that for the particular case where a is a continuous and coercive (or H-elliptic) bilinear form, i.e., there exists $\alpha > 0$ such that

$$a(\sigma,\sigma) \geq \alpha\|\sigma\|_H^2, \quad \forall\, \sigma \ \in \ H,$$

it is easy to see that ΠA is injective. Indeed, given $\sigma, \tilde{\sigma} \in \mathrm{Ker}(B)$ that have the same image in $\mathrm{Ker}(B)'$ by A, choosing $\tau = \sigma - \tilde{\sigma}$ we obtain

$$0 = a(\sigma - \tilde{\sigma}, \tau) = a(\sigma - \tilde{\sigma}, \sigma - \tilde{\sigma}) \geq \alpha\|\sigma - \tilde{\sigma}\|_H^2$$

which means that $\sigma = \tilde{\sigma}$ and A is injective.

Inf-sup conditions

To verify the surjectivity hypotheses of Theorem (2.2.1), we use equivalent hypotheses known as inf-sup conditions. The equivalence is a consequence of two classical theorems of Functional Analysis: the Closed Range Theorem (cf. [45, Theorem A.34], [48, Theorem 2.19]) and the Open Mapping Theorem (cf. [45, Theorem A.35], [48, Theorem 2.6]), which allow characterizing surjective operators. With these new hypotheses, Theorem (2.2.1) is known as the Babuška-Brezzi Theorem (cf. [45, Theorem 2.34],[49, Theorem 4.1]). These inf-sup conditions are given by:

1.
$$\begin{cases} \exists \alpha > 0, & \displaystyle\inf_{\sigma\in\mathrm{Ker}(B)} \sup_{\tau\in\mathrm{Ker}(B)} \frac{a(\sigma,\tau)}{\|\sigma\|_H\|\tau\|_H} \geq \alpha, \\ \forall \tau \in \mathrm{Ker}(B), & (\forall \sigma \in \mathrm{Ker}(B), a(\sigma,\tau) = 0) \Rightarrow (\tau = 0), \end{cases}$$

2.
$$\exists \beta > 0, \quad \inf_{q\in Q} \sup_{\tau\in H} \frac{b(\tau,v)}{\|\tau\|_H\|v\|_Q} \geq \beta.$$

Note that the second statement in 1. only provides an injectivity condition for the abstract operator ΠA.

Now, with the above inf-sup conditions, we can obtain the continuous dependence on the data according to Theorem (2.2.1), showing that there exists a constant $C > 0$, $C := C(\|A\|, \alpha, \beta)$ such that

$$\begin{cases} \|\sigma\|_H \leq c_1\|f\|_{H'} + c_2\|g\|_{Q'}, \\ \|u\|_Q \leq c_3\|f\|_{H'} + c_4\|g\|_{Q'}, \end{cases}$$

with $c_1 = \dfrac{1}{\alpha}$, $c_2 = \dfrac{1}{\beta}\left(1 + \dfrac{\|A\|}{\alpha}\right)$, $c_3 = \dfrac{1}{\beta}\left(1 + \dfrac{\|A\|}{\alpha}\right)$, and $c_4 = \dfrac{\|A\|}{\beta^2}\left(1 + \dfrac{\|A\|}{\alpha}\right)$.

Remark 2.1. (**Nonlinear case**). For the nonlinear operator A, results on existence, uniqueness, and approximation for dual-dual mixed variational formulations can be found in [50]. In these formulations, the nonlinear operator A is characterized by being strongly monotone and Lipschitz-continuous in the appropriate spaces.

2.2.2 Approximate Abstract Problem - Galerkin Method

Now consider two families of finite-dimensional subspaces $\{H_h\}_{h>0} \subset H$ and $\{Q_h\}_{h>0} \subset Q$, say N_1 and N_2, respectively. Then, for each $h > 0$ we can write the variational problem:

$$\begin{cases} \text{Find } \sigma_h \in H_h \text{ and } u_h \in Q_h \text{ such that} \\ a(\sigma_h, \tau_h) + b(\tau_h, u_h) \quad = \quad f(\tau_h), \quad \forall \tau_h \in H_h, \\ b(\sigma_h, v_h) \quad = \quad g(v_h), \quad \forall v_h \in Q_h. \end{cases} \tag{2.6}$$

If for each $h > 0$ the problem (2.6) is well-posed, and if the families of spaces $\{H_h\}_{h>0} \subset H$ and $\{Q_h\}_{h>0} \subset Q$ satisfy the following approximability condition

$$\forall \tau \in H, \quad \lim_{h\to 0}\left(\inf_{\tau_h \in H_h} \|\tau - \tau_h\|_H\right) = 0, \quad \lim_{h\to 0}\left(\inf_{w_h \in Q_h} \|w - w_h\|_Q\right) = 0\,, \tag{2.7}$$

then the solution (σ_h, u_h) of (2.6) is an approximate solution to the problem (2.4).

Remark 2.2. First, recall that given a subset $Y \subset X$ where X is a normed space, the distance between x and the subset Y is defined by

$$\text{dist}(x, Y) := \inf_{y \in Y} \|x - y\|_X.$$

Therefore, the approximation conditions (2.7) mean that for $h > 0$ increasingly smaller, the subspaces H_h and Q_h become increasingly larger. We also emphasize that $h > 0$ only characterizes a family of indices related to the dimensions N_1 and N_2 of the considered finite-dimensional spaces, H_h and Q_h, such that $h \to 0$ results in $N_1, N_2 \to \infty$. We will provide a geometric interpretation for h later in our discussion.

The verification of the well-posedness of the approximate problem (2.6) is analogous to that of the problem (2.4), that is, by writing the equivalent abstract problem and then using Theorem (2.2.1) (cf. [46, Theorem 2.4]). Thus, for each $h > 0$, we can find constants α_h and β_h depending on $h > 0$.

The linear system

The approximate problem (2.6) is simply a linear system. To illustrate, let us consider the particular case where $b = 0$ and a is coercive (or H-elliptic). Let $\{\phi_1, ...\phi_N\}$ be a basis of the finite-dimensional space H_h. Then there exist unique real numbers $\{U_1, ...U_N\}$ such that the solution u_h can be written as:

$$u_h = \sum_{i=1}^{N} U_i \phi_i$$

Let $\mathcal{A} \in \mathrm{R}^{n \times n}$ be the stiffness matrix such that

$$\mathcal{A}_{ij} = a(\phi_i, \phi_j), \quad 1 \leq i, j \leq N$$

and $F \in \mathrm{R}^n$ the vector with components

$$F_i = f(\phi_i), \quad 1 \leq i \leq N.$$

It is easy to see that

$$u_h \text{ is a solution of } (2.6) \text{ if and only if } \mathcal{A}U = \mathcal{F}.$$

Due to the approximability condition (2.7), we must have $N \to \infty$, and the existence and uniqueness of a solution to this linear system may not be guaranteed for every $h > 0$ (equivalently for every N). This is why PDE Theory is important - it ensures that by appropriately selecting the finite-dimensional subspaces, the existence of a solution to the linear system is guaranteed for any N.

A priori error

In this section, we will obtain *a priori* estimates for the approximation error $\|(\sigma, u) - (\sigma_h, u_h)\|_{H \times Q}$, where (σ, u) solves the exact problem (2.4) and (σ_h, u_h) solves the approximate problem (2.6). Using $\tau = \tau_h \in H$ and $v = v_h \in Q$ in (2.4) and (2.6), we have the following equality:

$$\begin{array}{llll} a(\sigma_h, \tau_h) + b(\tau_h, u_h) & = f(\tau_h) = & a(\sigma, \tau_h) + b(\tau_h, v) & \forall \tau_h \in H_h, \\ b(\sigma_h, \nu_h) & = g(\tau_h) = & b(\sigma, \nu_h) & \forall \nu_h \in Q_h. \end{array}$$

This induces the definition of a Galerkin operator $G_h : H \times Q \to H_h \times Q_h$ such that for each $(\zeta, w) \in H \times Q$, $G_h(\zeta, w)$ is the solution of the approximate variational problem

$$\begin{array}{llll} a(\zeta_h, \tau_h) + b(\tau_h, w_h) & = f_{\zeta,w}(\tau_h) := & a(\zeta, \tau_h) + b(\tau_h, w) & \forall \tau_h \in H_h, \\ b(\zeta_h, \nu_h) & = g_{\zeta,w}(\tau_h) := & b(\zeta, \nu_h) & \forall \nu_h \in Q_h. \end{array} \tag{2.8}$$

Remark 2.3. Rewriting the expression (2.8) as

$$\begin{array}{lll} a(\zeta - \zeta_h, \tau_h) + b(\tau_h, w - w_h) & = 0 & \forall \tau_h \in H_h, \\ b(\zeta - \zeta_h, \nu_h) & = 0 & \forall \nu_h \in Q_h, \end{array} \tag{2.9}$$

the left-hand side is a duality in $H \times Q$, that is:

$$\Big\langle (\zeta, w) - (\zeta_h, w_h), (\tau_h, v_h) \Big\rangle = (0, 0) \quad \forall \tau_h \in H_h, \, \forall \nu_h \in Q_h,$$

where $(\zeta_h, w_h) = G_h(\zeta, w)$. Thus, the operator G_h defines a projection.

Due to Theorem (2.2.1) for approximate problems, G_h is well-defined and bounded with $\|G_h\|$ depending on $\|A_h\|$, $\|(\Pi A_h)^{-1}\|$, β_h, $\|A\|$, and $\|B\|$. Taking $(\zeta, w) = (\sigma, u)$, the solution of (2.4), and $(\zeta_h, w_h) = (\sigma_h, u_h)$, the solution of (2.6), we have $G_h(\sigma, u) = (\sigma_h, u_h)$. It is also easy to see that $G_h(\zeta_h, w_h) = (\zeta_h, w_h)$. Consequently, we have the equality:

$$(\sigma, u) - (\sigma_h, u_h) = (I - G_h)\Big((\sigma, u) - (\zeta_h, w_h)\Big) \quad \forall (\zeta_h, w_h) \in H_h \times Q_h$$

Finally, using that $\|I - G_h\| = \|G_h\|$ (cf. [46, Theorem 2.5]), we obtain Cea's Estimate (see [46]):

$$\|(\sigma, u) - (\sigma_h, u_h)\|_{H \times Q} \leq \|G_h\| \, dist\Big((\sigma, u), X_h \times Q_h\Big). \tag{2.10}$$

Certainly, to confirm the convergence of the Galerkin scheme, i.e.,

$$\lim_{h \to 0} \|(\sigma, u) - (\sigma_h, u_h)\|_{H \times Q} = 0, \tag{2.11}$$

$\|G_h\|$ must be independent of h, which means requiring that all the involved constants, including the norms of the operators $\|A_h\|$, $\|(\Pi A_h)^{-1}\|$, β_h, $\|A\|$, and $\|B\|$, and the discrete inf-sup conditions, α_h and β_h, be independent of the subspace $H_h \times Q_h$. In fact, the need for h-independence is better perceived when, instead of deriving Cea's estimate through the Galerkin projector G_h, it is obtained by individually analyzing each of the errors $\|\sigma - \sigma_h\|_H$ and $\|u - u_h\|_Q$. More precisely, with the conditions and notations of Theorems (2.2.1) and its version for the approximate problem, we obtain

$$\|\sigma - \sigma_h\|_H \leq \left(1 + \frac{\|A\|}{\alpha_h}\right)\left(1 + \frac{\|B\|}{\beta_h}\right) \inf_{\zeta_h \in H_h} \|\sigma - \zeta_h\|_H + \frac{\|B\|}{\alpha_h} \inf_{w_h \in Q_h} \|u - w_h\|_Q \tag{2.12}$$

and

$$\begin{aligned} \|u - u_h\|_Q \leq {} & \frac{\|A\|}{\beta_h}\left(1 + \frac{\|A\|}{\alpha_h}\right)\left(1 + \frac{\|B\|}{\beta_h}\right) \inf_{\zeta_h \in H_h} \|\sigma - \zeta_h\|_H \\ & + \left(1 + \frac{\|B\|}{\beta_h} + \frac{\|A\|\|B\|}{\alpha_h \beta_h}\right) \inf_{w_h \in Q_h} \|u - w_h\|_Q. \end{aligned} \tag{2.13}$$

Remark 2.4. It is important to emphasize that the subspaces H_h and Q_h, which define the Galerkin scheme (2.4), cannot be chosen arbitrarily, as they must satisfy the hypotheses of the approximate version of Theorem (2.2.1) in addition to satisfying the approximation condition (2.7). In the next section, we will define some spaces with such properties for the problems we will study in this work.

2.3 Examples of approximation spaces

The spaces H, Q considered in problem (2.4) are function spaces defined on a subset Ω of R^n, $n \in \{2, 3\}$. The finite-dimensional subspaces H_h, Q_h considered in problem (2.6) are obtained by decomposing Ω into small parts and using some approximation (for example, polynomial interpolation) on each part of the decomposition of Ω. The finite-dimensional spaces obtained in this way are the Finite Element Spaces. There is a good variety of finite element spaces in the literature, both in terms of the geometric shapes that decompose Ω and in terms of the approximation method, for example polynomial interpolation (cf. [45–47, 51]). However, we will focus on some classical examples of finite element spaces: the Raviart-Thomas space that approximates vector fields with normal continuity, and the Lagrange space that approximates continuous functions. We will consider Ω decomposed into triangles or tetrahedra. After studying this chapter, readers will be able to analyze similar finite element subspaces from the literature.

2.3.1 Local Polynomials

In what follows, Ω is a bounded and connected domain of R^n, $n \in \{2, 3\}$, with polyhedral boundary Γ, and for each $h > 0$, $\mathcal{T}_h$ is a triangulation of $\overline{\Omega}$. More precisely, $\mathcal{T}_h$ is a finite family of triangles (in R^2) or tetrahedra (in R^3), such that

(i) $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$;

(ii) for every $K \in \mathcal{T}_h$, the interior of K, denoted by $\mathring{K}$, is non-empty ($\mathring{K} \neq \varnothing$);

(iii) $\mathring{K}_i \cap \mathring{K}_j = \emptyset$ for every $K_i, K_j \in \mathcal{T}_h$, $K_i \neq K_j$;

(iv) If $F = K_i \cap K_j$, $K_i, K_j \in \mathcal{T}_h$, $K_i \neq K_j$, then F is a common face, a common edge, or a common vertex of K_i and K_j;

(v) $\mathrm{diam}(K) =: h_K \leq h$ for every $K \in \mathcal{T}_h$.

Additionally, to each $\mathcal{T}_h$ we associate a fixed reference polyhedron $\hat{K}$, which may or may not belong to $\mathcal{T}_h$, and a family of affine mappings $\{T_K\}_{K \in \mathcal{T}_h}$ such that

(a) $T_K : \mathrm{R}^n \to \mathrm{R}^n$, $T_K(\hat{x}) = B_K \hat{x} + b_K$ for every $\hat{x} \in \mathrm{R}^n$, with $B_K \in \mathrm{R}^{n \times n}$ invertible, and $b_K \in \mathrm{R}^n$;

(b) $K = T_K(\hat{K})$ for every $K \in \mathcal{T}_h$.

Given a triangle K in R^n, where $n \in \{2, 3\}$, and a non-negative integer k, we define the spaces

$$\tilde{P}_k(K) := \{p : K \to \mathrm{R} : p \text{ is a polynomial of degree } = k\},$$

and

$$P_k(K) := \{p : K \to \mathrm{R} : p \text{ is a polynomial of degree } \leq k\}.$$

Equivalently, denoting $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and using multi-index notation, we have that $p \in P_k(K)$ if and only if there exist scalars $a_\alpha \in \mathrm{R}$ for every $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{N}_0^n$ with $|\alpha| \leq k$ such that

$$p(x) = \sum_{|\alpha| \leq k} a_\alpha x^\alpha \quad \forall x \in K.$$

Similarly, $p \in \tilde{P}_k(K)$ if and only if there exist scalars $a_\alpha \in \mathrm{R}$ for every $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{N}_0^n$ with $|\alpha| = k$ such that

$$p(x) = \sum_{|\alpha| = k} a_\alpha x^\alpha \quad \forall x \in K.$$

It is easy to see that the spaces $P_k(K)$ and $\tilde{P}_k(K)$ have finite dimension, with:

$$\dim P_k(K) = \binom{n+k}{k}, \qquad \text{and} \qquad \dim \tilde{P}_k(K) = \binom{n+k-1}{k}.$$

The Raviart-Thomas polynomials are defined by:

$$RT_k(K) = \mathbf{P}_k(K) \oplus \tilde{P}_k(K)x$$

and are also finite-dimensional spaces with (see Gatica [46], Lemma 3.5)

$$\dim RT_k(K) = \frac{(n+k+1)(n+k-1)!}{(n-1)!\,k!}.$$

2.3.2 Interpolation

Interpolation theory is fundamental for understanding the order of approximation error in numerical methods, especially in finite element methods. We will detail the concepts and logic behind the interpolation error estimate.

An interpolation operator Π is a linear operator that acts on functions in an infinite-dimensional space (such as $C(K)$ or $L^2(K)$) and preserves polynomials of degree up to k. This means that, for any polynomial $p \in P_k(K)$, we have:

$$\Pi(p) = p\,.$$

Here, $P_k(K)$ is the space of polynomials of degree up to k defined on the element K. In order to construct the interpolation operator, we choose a basis for $P_k(K)$ that is determined by the geometry of the element K. For example, in finite elements, this basis can be associated with the vertices, edges, or faces of K.

The canonical dual basis of $P_k(K)$ consists of linear functionals that evaluate the coefficients of the polynomials in the chosen basis. These functionals are the coordinate functions that allow representing any function v in the space X in terms of the basis of $P_k(K)$.

Interpolation theory provides estimates for the interpolation error $\|v - \Pi(v)\|_X$. These estimates depend on the regularity of the function v and the geometry of the element K. We will introduce the Lagrange interpolants and the Raviart-Thomas interpolant next, and calculate the interpolation errors of these interpolants.

Lagrange interpolant

Given $K \in \mathcal{T}_h$ with vertices a_1, a_2 ... a_{n+1}. The barycentric coordinates $\phi_j = \lambda_j(x)$, $1 \leq j \leq n+1$, of any point $x \in \mathrm{R}^n$, relative to the $(n+1)$ points a_{ij}, are the (unique) solutions of the linear system

$$\begin{cases} \displaystyle\sum_{j=1}^{n+1} a_{ij}\lambda_j = x_i, \quad 1 \le i \le n, \\ \displaystyle\sum_{j=1}^{n+1} \lambda_j = 1\,, \end{cases}$$

where $A = (a_{ij})$ with a_{ij} for $1 \le i \le n$ are the coordinates of the vertex a_j for $1 \le j \le n+1$ and $a_{ij} = 1$ if $i = n+1$ is an invertible matrix. Then we can define $\lambda_j : \mathrm{R}^n \to \mathrm{R}$ as:

$$\lambda_j(x) \; = \sum_{j=1}^{n} b_{ij}x_j + b_{i,n+1}, \quad 1 \le i \le n+1\,,$$

where $B = (b_{ij}) = A^{-1}$ is the inverse of the matrix A.

Remark 2.5. For $n = 2$, geometrically we have

$$\lambda_1(x) = \frac{|triangle(x, a_2, a_3)|}{|triangle(a_1, a_2, a_3)|},$$

$$\lambda_2(x) = \frac{|triangle(a_1, x, a_3)|}{|triangle(a_1, a_2, a_3)|},$$

$$\lambda_3(x) = \frac{|triangle(a_1, a_2, x)|}{|triangle(a_1, a_2, a_3)|},$$

where $|triangle(.,.,.)|$ denotes the area of the triangle. For $n = 3$, we simply use the volume of the tetrahedron. Note that $\lambda_i(a_j) = \delta_{ij}$ is the Kronecker delta.

Based on [47], Section 2.2, we can state the following unisolvence theorem:

Theorem 2.3.2 (Unisolvence)**.** Let $K \in \mathcal{T}_h$ and $p \in P_1(K)$. Then p is uniquely determined by its values at the $(n+1)$ vertices a_i of K.

Proof. We need to show that for all real μ_j, $1 \le j \le n+1$, the linear system

$$p(a_j) \; = \sum_{|\alpha| \le 1} \gamma_\alpha (a_j)^\alpha = \mu_j,$$

has a unique solution γ_α, $|\alpha| \le 1$. The dimension of the space P_1 is $n+1$, which coincides with the number of vertices a_j. Since the matrix of this linear system is square, it suffices to prove the existence of the solution. The barycentric coordinates λ_i satisfy $\lambda_i(a_j) = \delta_{ij}$ for $1 \le i, j \le n+1$. Consider the polynomial defined as:

$$p(x) = \sum_{i=1}^{n+1} \mu_i \lambda_i(x).$$

This polynomial satisfies $p(a_j) = \mu_j$ for all $1 \le j \le n+1$, as required. Therefore, the polynomial p is uniquely determined by

$$\forall p \in P_1(K), \quad p(x) = \sum_{i=1}^{n+1} p(a_i)\lambda_i(x)\,, \tag{2.14}$$

this completes the proof. □

Note that $\{\lambda_i\}_{1\le i\le n+1}$ is a basis of $P_1(K)$. We denote the Lagrange interpolation operator $\Pi_K^{L,k}(v) : C(K) \to P_1(K)$ by

$$\Pi_K^{L,k}(v) := \sum_{i=1}^{n+1} v(a_i)\lambda_i(x), \quad \forall\, v \in C(K).$$

The Lagrange interpolation in P_2 is obtained by also using the midpoints of the edges of the triangles, $a_{ij} = \dfrac{1}{2}(a_i + a_j)$, $1 \le i < j \le n+1$, (cf. [47, Section 2.2]), and the Lagrange interpolation becomes:

$$\forall p \in P_2(K), \quad p(x) = \sum_{i=1}^{n+1} \lambda_i(x)(2\lambda_i(x) - 1)p(a_i) + \sum_{i<j} 4\lambda_i(x)\lambda_j(x)p(a_{ij}).$$

There exists Lagrange interpolation P_k, with $k > 2$, but it is not commonly used in applications. For other interpolations, see [47, Section 2.2].

The idea is to combine the locally defined polynomials to obtain a finite-dimensional function space X_h^k such that $X_h^k \subset C(\overline{\Omega})$. This property will allow us to calculate the approximation error estimate using the norm of the function space $C(\overline{\Omega})$. Then, given a triangulation $\mathcal{T}_h$ of $\overline{\Omega}$ and an integer $k \ge 0$, we define the global Lagrange space as

$$X_h^k := \left\{ C(\overline{\Omega}); \quad v_K \in P_k(K) \quad \forall K \in \mathcal{T}_h \right\}.$$

Then, if X is a sufficiently regular function space, the global Lagrange interpolation $\Pi_{L,h}^k : X \to X_h^k$ is naturally defined by combining the local interpolations such that

$$\forall K \in \mathcal{T}_h, \quad \Pi_{L,h}^k(v)|_K := \Pi_K^{L,k}(v|_K)\,, \tag{2.15}$$

or, equivalently,

$$\Pi_{L,h}^k v = \sum_{K\in\mathcal{T}_h} \sum_{i=1}^{n+1} v(a_{i,K})\lambda_{i,K}\,.$$

where for each $K \in \mathcal{T}_h$, $a_{i,K}$ are the $n+1$ vertices of K and $\lambda_{i,K}$ are the $n+1$ basis functions of $P_1(K)$.

Theorem 2.3.3. Suppose that Ω is a bounded open subset of R^n with piecewise C^1 boundary Γ. Then, if $m \geq \frac{n}{2}$, the space $H^m(\Omega)$ is a subspace of $C(\overline{\Omega})$ and the canonical injection of $H^m(\Omega)$ into $C(\overline{\Omega})$ is continuous.

Proof. See [44, Theorem 1.6-4]. □

Raviart-Thomas interpolant

The Raviart-Thomas space is also defined locally so that its assembly produces a function space where the normal component across each piece is continuous. The continuity of the normal component is crucial for the consistency of the method, as it allows the principle of flux conservation to be approximately satisfied in the discretization. For this, we need the domain of the RT interpolation operator to be appropriate. In this sense, we have the following theorem:

Theorem 2.3.4. Consider the function space:

$$Z := \left\{ \tau \in [L^2(\Omega)]^n : \tau|_K \in [H^1(K)]^n \ \forall K \in \mathcal{T}_h \right\}.$$

Then

$$H(\mathrm{div};\Omega) \cap Z = \Big\{ \tau \in Z : \tau \cdot \mathbf{n}_{K_i} + \tau \cdot \mathbf{n}_{K_j} = 0 \text{ in } L^2(F)$$

$$\forall K_i, K_j \in \mathcal{T}_h \text{ that are adjacent with common face/edge } F \Big\}.$$

Proof. See [46, Theorem 3.2]. □

Remark 2.6. The expression $\tau \cdot \mathbf{n}_{K_i} + \tau \cdot \mathbf{n}_{K_j} = 0$ in $L^2(F)$ implies that

$$\int_F (\tau \cdot \mathbf{n}_{K_i} + \tau \cdot \mathbf{n}_{K_j})\psi = 0 \ \forall \psi \in L^2(\Omega)$$

since $\mathbf{n}_{K_i} = -\mathbf{n}_{K_j}$, we can write $\mathbf{n}_{K_F} = \mathbf{n}_{K_i} = -\mathbf{n}_{K_j}$ and obtain

$$\int_F \tau|_{K_i} \cdot \mathbf{n}_{K_F} \psi = \int_F \tau|_{K_j} \cdot \mathbf{n}_{K_F} \psi \quad \forall \psi \in L^2(\Omega) \tag{2.16}$$

Before defining the *RT* space, let us first state a theorem that will show the unisolvence of the polynomials $RT_k(K)$, i.e., that $\tau \in RT_k(K)$ is uniquely determined by the vertices of the triangle K. Recall that given a vector space X of finite dimension N, a set $\{f_1, ..., f_N\} \in X'$ is linearly independent if and only if $\cap_{i=1}^N \mathrm{Ker}(f_i) = \{0\}$. Indeed, consider the linear transformation $\Phi : X \to \mathrm{R}^n$ defined by $\Phi(x) := (f_1(x), ..., f_N(x))$. Note that Φ is injective, since $\mathrm{Ker}(\Phi) = \cap_{i=1}^N \mathrm{Ker}(f_i) = \{0\}$. Consequently, the matrix

A such that $\Phi(x) = Ax$ for every x in X in some basis of X is invertible. Let $c_1, ..., c_N$ be real numbers such that

$$0 = \sum_{i=1}^{N} c_i f_i(x) = c^{\mathrm{t}} Ax, \quad \forall x \in X\,,$$

where c is the column vector of c_i. It is easy to see that $c = 0$, therefore $\{f_1, ..., f_N\}$ is linearly independent.

Theorem 2.3.5 (Unisolvence)**.** Let $K \in \mathcal{T}_h$ and $\tau \in RT_k(K)$ and $\{\psi_{1,F}, \psi_{2,F}, ..., \psi_{d_k,F}\}$ be a basis of $P_k(F)$ and $\{\psi_{1,K}, \psi_{2,K}, ..., \psi_{r_k,K}\}$ be a basis of $P_{k-1}(F)$ Assume that

(i) $\displaystyle\int_F \tau \cdot \mathbf{n}_K\, \psi_{i,F} = 0, \quad 1 \leq i \leq d_k \quad \forall F$ face/edge of K, when $k \geq 0$;

(ii) $\displaystyle\int_K \tau \cdot \mathbf{n}_K\, \psi_{i,K} = 0, \quad 1 \leq i \leq r_k,$ when $k \geq 1$.

Then $\tau \equiv 0$ in K.

Proof. See [46, Theorem 3.3]. □

Remark 2.7. Theorem (2.3.5) defines $\tilde{N}$ linear and bounded functionals of $RT_k(K)'$, which for now we will denote by f_i, $1 \leq i \leq \tilde{N}$ such that $\cap_{i=1}^{\tilde{N}} Ker(f_i) = \{0\}$. As a consequence, we have that $\{f_1, ..., f_{\tilde{N}}\}$ is linearly independent. Note also that

$$\tilde{N} = (n+1)\, d_k + n\, \dim(P_{k-1}(K)) \;=\; \dim(RT_K)\,,$$

therefore the set of defined functionals is a basis for the dual space of $RT_k(K)$. From the definition of the dual basis, there must exist unique $\{p_1, ..., p_{\tilde{N}}\} \subset RT_k(K)$ such that $f_i(p_j) = \delta_{ij}$, and therefore, given $\tau \in RT_k(K)$, we have

$$\tau(x) = \sum_{i=1}^{\tilde{N}} f_i(\tau) p_i(x).$$

That is, $\tau \in RT_k(K)$ is uniquely determined by the functionals f_i, defined in Theorem (2.3.5), which in turn are uniquely determined by the vertices of K.

We define the global Raviart-Thomas space as

$$H_k^h := \left\{ \tau \in H(\mathrm{div}; \Omega) : \tau|_K \in RT_k(K) \quad \forall K \in \mathcal{T}_h^b \right\}.$$

Given $\tau \in H(\mathrm{div}; \Omega) \cap Z$, the linear forms defined in Theorem (2.3.5) i and ii are called F-moments and K-moments, respectively. All F-moments of τ are $m_i(\tau)$, $i \in \{1, 2, 3, .., N_1\}$ where $N_1 = d_k \times$ number of faces of $\mathcal{T}_h$ and all K-moments of τ

are $m_i(\tau)$, $i \in \{N_1+1, N_1+2, .., N\}$ where $N - N_1 = r_k \times$ number of triangles of $\mathcal{T}_h$. Therefore, the total number of moments of τ is N.

The interpolation operator $\Pi_h^{RT,k} : H(\mathrm{div};\Omega) \cap Z \to H_h^k$ is defined as

$$\Pi_h^{RT,k}(\tau) := \sum_{j=1}^{N} m_j(\tau)\phi_j,$$

where $\phi_1, \phi_2, ..., \phi_N$ are the unique functions in H_h^k such that $m_i(\phi_j) = \delta_{ij}$. Equivalently, $\Pi_h^{RT,k}(\tau)$ is the unique function in H_h^k such that

$$m_i(\Pi_h^{RT,k}(\tau)) = m_i(\tau), \quad \forall i \in \{1, 2, ..., N\}.$$

Then, for each $K \in \mathcal{T}_h$ we define $m_{i,K}(\tau)$, $i \in \{1, 2, \cdots, N_K\}$, as the corresponding local moments, i.e., the F-moments of the faces/edges F of K and the K-moments of K. Since the number of faces/edges of K is $n+1$, we have that $N_K = (n+1)d_k + r_k$. Then we define the local interpolation operator $\Pi_K^{RT,k} : [H^1(K)]^n \to RT_k(K)$ as

$$\Pi_K^{RT,k}(\tau) := \sum_{j=1}^{N_K} m_{j,K}(\tau)\varphi_{j,K} \quad \forall \tau \in [H^1(K)]^n,$$

where, given $j \in \{1, \cdots, N_K\}$, $\varphi_{j,K}$ is the unique function in $RT_k(K)$ such that

$$m_{i,K}(\varphi_{j,K}) = \delta_{ij} \quad \forall i \in \{1, 2, \cdots, N_K\}.$$

Note that $\Pi_h^{RT,k}(\tau)|_K = \Pi_K^{RT,k}(\tau) \quad \forall \tau \in H(\mathrm{div};\Omega) \cap Z$ holds. The following lemma relates the divergences of the local and global interpolation operators in terms of the orthogonal projectors and will be used to calculate the local interpolation error:

$$\mathcal{P}_K^k : L^2(K) \to \mathcal{P}_k(K) \quad \text{and} \quad \mathcal{P}_h^k : L^2(\Omega) \to \mathcal{P}_h^k\,,$$

where

$$Y_h^k := \left\{ v \in L^2(\Omega) : v|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h \right\}.$$

Lemma 2.3.1. The following holds:

$$\mathrm{div}(\Pi_h^{RT,k}(\tau)) = \mathcal{P}_K^k(\mathrm{div}\tau) \quad \forall \tau \in [H^1(K)]^n \tag{2.17}$$

and

$$\mathrm{div}(\Pi_h^{RT,k}(\tau)) = \mathcal{P}_h^k(\mathrm{div}\tau) \quad \forall \tau \in H(\mathrm{div};\Omega) \cap Z. \tag{2.18}$$

Proof. See [46, Lemma 3.7]. □

2.3.3 Local Interpolation Error

We will now obtain the local interpolation error for Lagrange and Raviart-Thomas interpolants. This will allow us to calculate the convergence order of finite element methods that use these spaces. But first, we need some preliminary results from the general theory of interpolation.

Preliminary Results

An important result in interpolation theory is the Bramble-Hilbert Lemma, which provides an estimate for the error of linear and bounded operators defined on function spaces with two characteristics: they preserve polynomials and do not increase the regularity of the function.

Theorem 2.3.6 (Bramble-Hilbert Lemma)**.** Let m and k be non-negative integers such that $0 \leq m \leq k+1$, and let $\Pi : H^{k+1}(K) \to H^m(K)$ be a linear and bounded operator such that $\Pi(p) = p \quad \forall p \in P_k(K)$. Then there exists $C := C(\Pi, K) > 0$ such that

$$\|v - \Pi(v)\|_{m,K} \leq C|v|_{k+1,K} \quad \forall v \in H^{k+1}(K). \tag{2.19}$$

Proof. See [46, Theorem 3.5], [44, Theorem 4.4-2]. □

Piola Transformation

Let $K \in \mathcal{T}_h$, $\tau \in [H^1(K)]^n$ and the affine transformation $T_K : \mathrm{R}^n \to \mathrm{R}^n$ defined by

$$T_K(\hat{x}) := B_K\hat{x} + b_K \; \forall\hat{x} \in \mathrm{R}^n \, ,$$

with $B_K \in \mathrm{R}^{n\times n}$ invertible and $b_K \in \mathrm{R}^n$, such that $K = T_K(\hat{K})$, where $\hat{K}$ is the reference polyhedron. We introduce the Piola transformation:

$$\hat{\tau} := |\det B_K| B_K^{-1} \tau \circ T_K .$$

The affine transformation is important because it is the basis of most convergence theorems. Also, in computational practice, the calculation of coefficients is performed on a reference finite element [47, Section 4.1]. In this sense, we present two lemmas that provide estimates relating an element K with the reference element $\hat{K}$ via the Piola transformation.

Lemma 2.3.2. Let $K, \hat{K} \in \mathcal{T}_h$, and let $F : \mathrm{R}^n \longrightarrow \mathrm{R}^n$ be the affine transformation given by $F(\hat{x}) = B\hat{x} + b \ \forall \hat{x} \in \mathrm{R}^n$, with $B \in \mathrm{R}^{n\times n}$ invertible and $b \in \mathrm{R}^n$, such that $K = F(\hat{K})$. Let m be a non-negative integer, and let $v \in H^m(K)$. Then $\hat{v} := v \circ F \in H^m(\hat{K})$, and there exists $C := C(m,n) > 0$ such that

$$|\hat{v}|_{m,\hat{K}} \leq C\|B\|^m |\det B|^{-1/2} |v|_{m,K}. \tag{2.20}$$

Conversely, if $\hat{v} \in H^m(\hat{K})$ and we define $v = \hat{v} \circ F^{-1}$, then $v \in H^m(K)$, and there exists $\hat{C} := \hat{C}(m,n) > 0$ such that

$$|v|_{m,K} \leq \hat{C}\|B^{-1}\|^m |\det B|^{1/2} |\hat{v}|_{m,\hat{K}}. \tag{2.21}$$

Proof. See [46, Lemma 3.12] (see also [47, Theorem 3.1.2]). □

Remark 2.8. If $\tau \in [H^m(K)]^n$. Then $\hat{\tau} := |\det B| B^{-1} \tau \circ F \in [H^m(\hat{K})]^n$, and there exists $C := C(m,n) > 0$ such that

$$|\hat{\tau}|_{m,\hat{K}} \leq C\|B^{-1}\| \|B\|^m |\det B|^{1/2} |\tau|_{m,K}. \tag{2.22}$$

Conversely, if $\hat{\tau} \in [H^m(\hat{K})]^n$ and we define $\tau := |\det B|^{-1} B\hat{\tau} \circ F^{-1}$, then $\tau \in [H^m(K)]^n$, and there exists $\hat{C} := \hat{C}(m,n) > 0$ such that

$$|\tau|_{m,K} \leq \hat{C}\|B\| \|B^{-1}\|^m |\det B|^{-1/2} |\hat{\tau}|_{m,\hat{K}}. \tag{2.23}$$

(cf. [46, Lemma 3.13]).

The following lemma establishes geometric properties of the Piola Transformation under the elements K and $\widehat{K}$ of $\mathcal{T}_h$.

Lemma 2.3.3. Let $K, \hat{K} \in \mathcal{T}_h$, and let $F : \mathrm{R}^n \to \mathrm{R}^n$ be the affine transformation given by $F(\hat{x}) = B\hat{x} + b \ \forall \hat{x} \in \mathrm{R}^n$, with $B \in \mathrm{R}^{n\times n}$ invertible and $b \in \mathrm{R}^n$, such that $K = F(\hat{K})$. Let

$$\begin{aligned} h_K &:= \text{diameter of } K = \max_{x,y\in K} \|x - y\|, \\ \rho_K &:= \text{diameter of the largest sphere contained in } K, \\ \hat{h} &:= \text{diameter of } \hat{K}, \text{ and} \\ \hat{\rho} &:= \text{diameter of the largest sphere contained in } \hat{K}. \end{aligned}$$

Then

$$|\det B| = \frac{|K|}{|\hat{K}|}, \quad \|B\| \leq \frac{h_K}{\hat{\rho}} \quad \text{and} \quad \|B^{-1}\| \leq \frac{\hat{h}}{\rho_K}. \tag{2.24}$$

Proof. See [46, Lemma 3.14] (see [47, Theorem 3.1.3] and see [44, Lemma 4.4.1]). □

Remark 2.9. The following relation between the local interpolation on $K \in \mathcal{T}_h$ and the reference element $\widehat{K}$, for both Lagrange and RT interpolants, is easily verified using variable substitution (cf. [46, Lemma 3.11]):

$$\Pi^k_{RT,\widehat{K}}(\widehat{\tau}) = \widehat{\Pi_K^{RT,k}(\tau)} := |\det B_K|\, B_K^{-1}\, \Pi_K^{RT,k}(\tau) \circ T_K\,, \tag{2.25}$$

for all τ in domain of $\Pi_K^{RT,k}$.

Error Estimates for Polynomial-Preserving Operators

We present the Error Estimate Theorem for more general interpolation operators than Lagrange and Raviart-Thomas, which are linear, bounded, and polynomial-preserving, applied to a triangular (or tetrahedral) mesh. Its proof follows from the application of the Bramble-Hilbert Theorem (cf. Theorem (2.3.6)) and Lemma (2.3.2). Given its importance, we will present its proof in detail (cf. [44], Theorem 4.4-2 or [47], Theorem 3.1.4).

Theorem 2.3.7. Given $\widehat{K} \in \mathcal{T}_h$, and let $\widehat{\Pi}$ be a linear continuous operator from $H^{k+1}(\widehat{K})$ to $H^m(\widehat{K})$, $0 \leq m \leq k+1$, such that

$$\forall \widehat{p} \in P_k(\widehat{K}), \quad \widehat{\Pi}\widehat{p} = \widehat{p}. \tag{2.26}$$

If $K \in \mathcal{T}_h$ such that $F(K) = \widehat{K}$, and if the operator Π is defined by

$$\forall v \in H^{k+1}(K), \quad \widehat{\Pi v} = \widehat{\Pi}(\widehat{v}), \tag{2.27}$$

then there exists a constant $C := C(\widehat{K}, \widehat{\Pi})$, independent of F (and therefore of the geometric characteristics of K), such that

$$\forall v \in H^{k+1}(K), \quad |v - \Pi v|_{m,K} \leq C \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,K}. \tag{2.28}$$

Proof. Let $\tau \in [H^{k+1}(K)]^n$. Using the estimate (2.21) and (2.27), we obtain

$$|\tau - \Pi(\tau)|_{m,K} \leq C\|B^{-1}\|^m|\det B|^{1/2}|\hat{\tau} - \widehat{\Pi}(\widehat{\tau})|_{m,\widehat{K}}. \tag{2.29}$$

Therefore, Theorem (2.3.6) implies that

$$|\widehat{\tau} - \widehat{\Pi}(\widehat{\tau})|_{m,\widehat{K}} \leq C|\widehat{\tau}|_{k+1,\widehat{K}}. \tag{2.30}$$

Then, applying the estimate (2.20), we obtain

$$|\widehat{\tau}|_{k+1,\widehat{K}} \leq C\|B_K\|^{k+1}|\det B_K|^{-1/2}|\tau|_{k+1,K}. \tag{2.31}$$

Thus, inserting (2.31) into (2.30) and then the resulting bound into (2.28), we deduce that

$$|\tau - \Pi(\tau)|_{m,K} \leq C\|B_K\|^{k+1}\|B_K^{-1}\|^m|\tau|_{k+1,K}, \tag{2.32}$$

from which, using $\|B_K^{-1}\| \leq \dfrac{\widehat{h}}{\rho_K}$ and $\|B_K\| \leq \dfrac{h_K}{\widehat{\rho}}$ (cf. Eqs. (2.24)), we arrive at (2.28). □

Remark 2.10. If $\widehat{\Pi}$ is a linear continuous operator from vectorial spaces, $[H^{k+1}(\widehat{K})]^n$ to $[H^m(\widehat{K})]^n$, using Remark (2.8) instead of Lemma (2.3.2), we obtain

$$\forall \tau \in [H^{k+1}(K)]^n, \quad |\tau - \Pi\tau|_{m,K} \leq C\frac{h_K^{k+2}}{\rho_K^{m+1}}|\tau|_{k+1,K}. \tag{2.33}$$

Local Error Estimates for Lagrange

The local error of Lagrange interpolation that we will present next will follow as a particular case of Theorem (2.3.7) (cf. [44], Theorem 4.4-3 or [47], Theorem 3.1.5).

Theorem 2.3.8 (Lagrange Interpolation)**.** Let m, n, and k be non-negative integers such that $n \leq 3$, $k > 1$, and $0 \leq m \leq k+1$. If $\Pi_K^{L,k}$ is the Lagrange interpolant, then there exists $C := C(\widehat{K}, \Pi_{\widehat{K}}^{L,k}, k, m, n) > 0$ such that

$$\forall v \in H^{k+1}(K), \quad |v - \Pi_K^{L,k}v|_{m,K} \leq C\,\frac{h_K^{k+1}}{\rho_K^m}|v|_{k+1,K}. \tag{2.34}$$

Proof. Clearly $\widehat{\Pi_K^{L,k}}(\widehat{p}) = \widehat{p}$ for all $\widehat{p} \in P_k$. Now we will show that $\Pi^k_{RT,\widehat{K}} : H^{k+1}(\widehat{K}) \to H^m(\widehat{K})$ is bounded. Given $\widehat{v} \in H^{k+1}(\widehat{K})$, we have

$$\|\widehat{\Pi_K^{L,k}}(\widehat{v})\|_{m,\widetilde{K}} \leq \sum_{i=1}^{N} |\widehat{v}(a_i)| \|\phi_i(x)\|_{m,\widehat{K}} \, .$$

Since $k \geq 1$, then from Theorem (2.3.3), $H^{k+1}(\widehat{K}) \subset C(\widehat{K})$, we can use $|\widehat{v}(a_i)| \leq \|\widehat{v}\|_{\infty,\widehat{K}} \leq C_1 \|\widehat{v}\|_{k+1,\widehat{K}}$

$$\|\widehat{\Pi_K^{L,k}}(\widehat{v})\|_{m,\widetilde{K}} \leq C(\Pi_K^{L,k}, \widehat{K}) \|\widehat{v}\|_{k+1,\widehat{K}},$$

which proves that $\Pi_K^{L,k}$ is bounded. Finally, it is clear that $\widehat{\Pi_K^{L,k}(v)} = \Pi^k_{L,\widehat{K}}(\widehat{v})$ for all v in domain of $\Pi_K^{L,k}$) (cf. (2.25)). Therefore, it suffices to apply Theorem (2.3.7) to obtain (2.34). □

Local Error Estimates for RT

The local error of Raviart-Thomas interpolation, as in the case of the Lagrange interpolant, can be obtained based on Theorem (2.3.7). However, the analysis requires additional care due to the norm of the space $H(\text{div}, \Omega)$, which involves not only the interpolated function but also its divergence. Based on [46], Lemma 3.16, we can state the following theorem.

Lemma 2.3.4 (Local Interpolation Error)**.** Let m and k be non-negative integers such that $0 \leq m \leq k+1$. Then there exists $C := C(\widehat{K}, \Pi^k_{RT,\widehat{K}}, k, m, n) > 0$ such that

$$|\tau - \Pi_K^{RT,k}(\tau)|_{m,K} \leq C \frac{h_K^{k+2}}{\rho_K^{m+1}} |\tau|_{k+1,K} \quad \forall \tau \in [H^{k+1}(K)]^n. \tag{2.35}$$

Moreover, for each $\tau \in [H^1(K)]^n$, with $\operatorname{div} \tau \in H^{k+1}(K)$, we have

$$|\operatorname{div} \tau - \operatorname{div} \Pi_K^{RT,k}(\tau)|_{m,K} \leq C \frac{h_K^{k+1}}{\rho_K^{m}} |\operatorname{div} \tau|_{k+1,K}. \tag{2.36}$$

Proof. Since $\Pi_K^{RT,k} \in \mathcal{L}([H^{k+1}(\widehat{K})]^n, [H^m(\widehat{K})]^n)$ (cf. Lemma 3.15), $\Pi^k_{RT,\widehat{K}}(\widehat{p}) = \widehat{p}$ $\forall \, \widehat{p} \in RT_k(\widehat{K})$, and $[\mathbb{P}_k(\widehat{K})]^n \subseteq RT_k(\widehat{K})$. It suffices to use Remark (2.10) of Theorem (2.3.7) to obtain (2.35).

On the other hand, let $\tau \in [H^1(K)]^n$, with $\operatorname{div}\tau \in H^{k+1}(K)$. From the chain rule, we have

$$\nabla\widehat{\tau}(\widehat{x}) = |detB_K|\, B_K^{-1}\, \nabla\tau(T_K(\widehat{x}))\, B_K\,.$$

Then, using that $tr(B^{-1}TB) = tr(T)$ and $div(\widehat{\tau}) = tr(\nabla\widehat{\tau})$, we can deduce that

$$\operatorname{div}\widehat{\tau} = |\det B_K|\operatorname{div}\tau\circ T_K \quad \forall\tau\in[H^1(K)]^n, \tag{2.37}$$

and then, also using Lemma (2.25), we find that

$$\operatorname{div}\tau - \widehat{\operatorname{div}\Pi}_K^{RT,k}(\tau) = |\det B_K|^{-1}\left\{\operatorname{div}\widehat{\tau} - \operatorname{div}\Pi^k_{RT,\widehat{K}}(\widehat{\tau})\right\}.$$

Moreover, we know from Lemma (2.3.1) (applied to $\widehat{K}$) that $\operatorname{div}\Pi^k_{RT,\widehat{K}}(\tau) = \mathcal{P}_k^k(\operatorname{div}\widehat{\tau})$, where $\mathcal{P}_k^k : L^2(\widehat{K}) \to \mathcal{P}_k(\widehat{K})$ is the orthogonal projector. Then, employing the estimate (2.23) (cf. Remark (2.8)) and the preceding identity, we obtain

$$\begin{aligned}|\operatorname{div}\tau - \operatorname{div}\Pi_K^{RT,k}(\tau)|_{m,K} &\le \widehat{C}|\widehat{B}_K^{-1}|^m|\det B_K|^{1/2}|\operatorname{div}\widehat{\tau} - \operatorname{div}\Pi^k_{RT,\widehat{K}}(\widehat{\tau})|_{m,\widehat{K}}\\ &= \widehat{C}|\widehat{B}_K^{-1}|^m|\det B_K|^{1/2}|\operatorname{div}\widehat{\tau} - \mathcal{P}_k^k(\operatorname{div}\widehat{\tau})|_{m,\widehat{K}}.\end{aligned} \tag{2.38}$$

Now it is easy to see that $\mathcal{P}_K^k \in \mathcal{L}(H^{k+1}(\widehat{K}), H^m(\widehat{K}))$, for example, by writing

$$\mathcal{P}_K^k(\widehat{v}) := \sum_{i=1}^{m_k}\langle\widehat{v},\varphi_{i,k}\rangle_{0,\widehat{K}}\varphi_{i,k} \quad \forall\widehat{v}\in L^2(\widehat{K}),$$

where $\langle\cdot,\cdot\rangle_{0,\widehat{K}}$ is the inner product of $L^2(\widehat{K})$ and $\{\varphi_{1,k},\varphi_{2,k},\cdots,\varphi_{m_k,k}\}$ is an orthonormal basis of $\mathcal{P}_k(\widehat{K})$. Moreover, it is clear that $\mathcal{P}_K^k(\widehat{p}) = \widehat{p} \quad \forall\widehat{p}\in\mathcal{P}_k(\widehat{K})$. Thus, applying the Bramble-Hilbert lemma, the identity (2.37), and the estimate (2.22) (cf. Remark (2.8)), we conclude that

$$|\text{div}\widehat{\tau} - \mathcal{P}_K^k(\text{div}\widehat{\tau})|_{m,\widehat{K}} \le C|\text{div}\widehat{\tau}|_{k+1,\widehat{K}}$$

$$= C|\det B_K||\widehat{\text{div}\tau}|_{k+1,\widehat{K}} \le C|\det B_K|^{1/2}\|B_K\|^{k+1}|\text{div}\tau|_{k+1,K},$$

which, substituted into (2.38), implies

$$|\text{div}\tau - \text{div}\Pi_K^{RT,k}(\tau)|_{m,K} \le C\|B_K^{-1}\|^m\|B_K\|^{k+1}|\text{div}\tau|_{k+1,K}. \tag{2.39}$$

Finally, using again the geometric constraints given by Lemma (2.3.3), we obtain (2.36) directly from (2.39). □

Remark 2.11. The following result extends Theorem (2.3.8) to all intermediate seminorms (cf. [46, Lemma 3.17]). For non-negative integers m, k, and l with $0 \leq l \leq k$ and $0 \leq m \leq l+1$, there exists a constant $C := C(\widehat{K}, \Pi_K^{RT,k}, k, m, n) > 0$ such that

$$\left| \tau - \Pi_K^{RT,k}(\tau) \right|_{m,K} \leq C \frac{\eta_K^{l+2}}{\rho_K^{m+1}} |\tau|_{l+1,K} \quad \forall \tau \in [H^{l+1}(K)]^n.$$

Moreover, for each $\tau \in [H^l(K)]^n$ with div $\tau \in H^{l+1}(K)$, the following estimate holds:

$$\left| \text{div } \tau - \text{div } \Pi_K^{RT,k}(\tau) \right|_{m,K} \leq C \frac{\eta_K^{l+1}}{\rho_K^{m}} |\text{div } \tau|_{l+1,K} \, .$$

2.3.4 Global Interpolation Error

Having estimated the local interpolation error, we are now in a position to estimate the global interpolation error for the considered examples. For this, we recall that a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of Ω is said to be regular if there exists $c > 0$ such that

$$\frac{h_K}{\rho_K} \leq c \quad \forall K \in \mathcal{T}_h, \quad \forall h > 0.$$

With this, we will state the two main Theorems that establish the convergence order of the interpolation errors for Lagrange and Raviart-Thomas interpolations applied to Sobolev spaces $H^{k+1}(\Omega)$, with $k > 1$. (cf. [47], Theorem 3.2.1).

Global error estimates for Lagrange interpolation

Theorem 2.3.9. Let $\{\mathcal{T}_h\}_{h>0}$ be a regular family of triangulations of Ω, assume that there exist integers $k \geq 1$ and $m \geq 0$ with $m \leq k$. Then, there exists a constant C independent of h such that, for any function $v \in H^{k+1}(\Omega)$,

$$\|v - \Pi_h^{RT,k} v\|_{m,\Omega} \leq C h^{k+1-m} |v|_{k+1,\Omega}, \quad 0 \leq m \leq 1, \tag{2.40}$$

$$\left(\sum_{K \in \mathcal{T}_h} \|v - \Pi_h^{RT,k} v\|_{m,K}^2 \right)^{1/2} \leq C h^{k+1-m} |v|_{k+1,\Omega}, \quad 2 \leq m \leq k+1, \tag{2.41}$$

where $\Pi_h^{RT,k} v \in V_h$ is the X_h interpolant of the function v.

Remark 2.12. We can extend Theorem (2.3.9) to intermediate seminorms. Just consider an integer $l > 0$ such that $l \leq k$ and the same estimates remain but with $0 \leq m \leq \min 1, l$ in equation (2.40) and $2 \leq m \leq \min l, k+1$ in equation (2.41).

Proof. (*Proof of Theorem* (2.3.9)) Applying Theorem (2.3.8), we obtain

$$\|v - \Pi_k v\|_{m,K} \leq C h_K^{k+1-m} |v|_{k+1,K}, \quad 0 \leq m \leq k+1.$$

Using the relations $(\Pi_h^{RT,k} v)|_K = \Pi_K^{RT,k} v$, $K \in \mathcal{T}_h$ (cf. (2.15)) and the inequalities $h_K \leq h$, $K \in \mathcal{T}_h$ (cf. (3.2.2)), we obtain

$$\begin{aligned} \left(\sum_{K \in \mathcal{T}_h} \|v - \Pi_h^{RT,k} v\|_{m,K}^2 \right)^{1/2} &\leq C h^{k+1-m} \left(\sum_{K \in \mathcal{T}_h} |v|_{k+1,K}^2 \right)^{1/2} \\ &= C h^{k+1-m} |v|_{k+1,\Omega}, \quad 0 \leq m \leq k+1. \end{aligned}$$

Thus, inequality (2.41) is proven.

From $X_h \subset H^1(\Omega)$ for $m = 0$ and for $m = 1$, we have

$$\left(\sum_{K \in \mathcal{T}_h} \|v - \Pi_h^{RT,k} v\|_{m,K}^2 \right)^{1/2} = \|v - \Pi_h^{RT,k} v\|_{m,\Omega},$$

and therefore we obtain (2.40). □

Global error estimates for RT interpolation

Based on [46], Theorem 3.6, we can state the following theorem.

Theorem 2.3.10 (Global RT Interpolation Error). Let $\{\mathcal{T}_h\}_{h>0}$ be a regular family of triangulations of Ω, and let k be a non-negative integer. Then there exists $C > 0$, independent of h, such that

$$\|\tau - \Pi_h^{RT,k}(\tau)|_{\mathrm{div},\Omega} \leq C h^{m+1} \left\{ |\tau|_{m+1,\Omega} + |\operatorname{div} \tau|_{m+1,\Omega} \right\} \tag{2.42}$$

for each $\tau \in [H^{m+1}(\Omega)]^n$, with $\operatorname{div} \tau \in H^{m+1}(\Omega)$, $0 \leq m \leq k$.

Remark 2.13. For intermediate norms, we consider l such that $0 \leq l \leq k$ and $0 \leq m \leq l+1$, and we obtain

$$\|\tau - \Pi_h^{RT,k}(\tau)|_{\mathrm{div},\Omega} \leq C h^{l+1} \left\{ |\tau|_{l+1,\Omega} + |\operatorname{div} \tau|_{l+1,\Omega} \right\} \tag{2.43}$$

for each $\tau \in [H^{m+1}(\Omega)]^n$, with $\operatorname{div} \tau \in H^{m+1}(\Omega)$, $0 \leq m \leq k$.

Proof. (*Proof of Theorem* (2.3.10)) Let $0 \leq m \leq k$ and $\tau \in [H^{m+1}(\Omega)]^n$ such that $\mathrm{div}\tau \in H^{m+1}(\Omega)$ (consequently $\tau \in H_{div}(\Omega) \cap Z$). Then, applying (2.35) and (2.36) (cf.

Lemma (2.3.4)), with $m = 0$, we obtain

$$\|\tau - \Pi_K^{RT,k}(\tau)\|_{0,K} \leq C \frac{h_K^{m+2}}{\rho_K} |\tau|_{m+1,K} \quad \forall K \in \mathcal{T}_h$$

and

$$\|\mathrm{div}\tau - \mathrm{div}\Pi_K^{RT,k}(\tau)\|_{0,K} \leq C h_K^{m+1} |\mathrm{div}\tau|_{m+1,K} \quad \forall K \in \mathcal{T}_h,$$

from which, using the regularity of the family $\{\mathcal{T}_h\}_{h>0}$, we deduce that

$$\|\tau - \Pi_K^{RT,k}(\tau)\|_{\mathrm{div},K}^2 = \|\tau - \Pi_K^{RT,k}(\tau)\|_{0,K}^2 + \|\mathrm{div}\tau - \mathrm{div}\Pi_K^{RT,k}(\tau)\|_{0,K}^2$$
$$\leq \tilde{C}^2 h_K^{2(m+1)} \left\{ |\tau|_{m+1,K}^2 + |\mathrm{div}\tau|_{m+1,K}^2 \right\},$$

Then, recalling that $\Pi_h^{RT,k}(\tau)|_K = \Pi_K^{RT,k}(\tau|_K)$ and $h_K \leq h \ \forall K \in \mathcal{T}_h$, we find that

$$\|\tau - \Pi_h^{RT,k}(\tau)\|_{\mathrm{div},\Omega}^2 = \sum_{K \in \mathcal{T}_h} \|\tau - \Pi_K^{RT,k}(\tau)\|_{\mathrm{div},K}^2$$
$$\leq \sum_{K \in \mathcal{T}_h} \tilde{C}^2 h_K^{2(m+1)} \left\{ |\tau|_{m+1,K}^2 + |\mathrm{div}\tau|_{m+1,K}^2 \right\}$$
$$\leq \tilde{C}^2 h^{2(m+1)} \left\{ |\tau|_{m+1,\Omega}^2 + |\mathrm{div}\tau|_{m+1,\Omega}^2 \right\},$$

which gives (2.42) and completes the proof. □

2.3.5 Approximability and Order of Convergence

With the interpolation properties of Lagrange and Raviart-Thomas interpolants from Section (2.3.4), we can now show the approximability conditions (cf. (2.7)), consequently finding the order of convergence of the approximation error for the spaces of interest. The central idea is that given a normed vector space X and a finite-dimensional subspace X_h, we consider the orthogonal projection $\mathcal{P} : X \to X_h$, and the interpolation operator $\Pi : X \to X_h$. For $\tau \in X$ we have

$$\|\tau - \mathcal{P}(\tau)\|_X := \inf_{\tau_h \in X_h^k} \|\tau - \tau_h\|_X \leq \|\tau - \Pi(\tau)\|_X.$$

where $\|\tau - \mathcal{P}(\tau)\|_X := \mathrm{dist}(\tau, X_h)$. Then, we use the interpolation error estimates to obtain the order of convergence.

Given a non-negative integer k, we are interested in the following orthogonal projectors (in each case with respect to the inner products of the projected spaces):

$$\begin{aligned}\mathcal{P}^k_{\mathrm{div},h} : H(\mathrm{div};\Omega) \to H^k_h &:= \Big\{\tau \in H(\mathrm{div};\Omega) : \tau|_K \in RT_k(K)\,\forall K \in \mathcal{T}^b_h\; k \geq 0\Big\}, \\ P^k_{1,h} : H^1(\Omega) \to X^k_h &:= \Big\{v \in C(\overline{\Omega}) : v_h|_K \in P_k(K)\,\forall K \in \mathcal{T}_h\; k \geq 1\Big\}, \\ P^k_h : L^2(\Omega) \to X^k_h &:= \Big\{v \in C(\overline{\Omega}) : v_h|_K \in P_k(K)\,\forall K \in \mathcal{T}_h\; k \geq 1\Big\}, \\ \mathcal{P}^k_h : L^2(\Omega) \to Y^k_h &:= \Big\{v \in L^2(\Omega) : v_h|_K \in P_k(K)\,\forall K \in \mathcal{T}_h\; k \geq 0\Big\}\end{aligned} \tag{2.44}$$

Remark 2.14. The finite-dimensional spaces defined above are named as follows:

The X^k_h is the Continuous Finite Element Space of Degree k (See [46, Section 4.1]).

The H^k_h is the Raviart-Thomas Finite Element Space of Degree k (See [46, Section 4.1]).

The Y^k_h is the Discontinuous Finite Element Space of Degree k (See [46, Section 4.1]).

Approximability of the Raviart-Thomas Finite Element Space of Degree k (H^k_h) in $H(\mathrm{div};\Omega)$

Lemma 2.3.5. Let $\Pi^{RT,k}_h : H(\mathrm{div};\Omega) \cap Z \to H^k_h$ be the global Raviart-Thomas interpolation operator, where $Z := \Big\{\tau \in \Big[L^2(\Omega)\Big]^n : \tau|_K \in \Big[H^1(K)\Big]^n \;\forall K \in \mathcal{T}_h\Big\}$ (cf. Theorem 3.2). Then for all $\tau \in [H^{l+1}(\Omega)]^n$ with $\mathrm{div}\tau \in H^{l+1}(\Omega)$, $0 \leq l \leq k$, we have

$$\|\tau - \mathcal{P}^k_{\mathrm{div},h}(\tau)\|_{\mathrm{div},\Omega} \leq Ch^{l+1}\Big\{|\tau|_{l+1,\Omega} + |\mathrm{div}\,\tau|_{l+1,\Omega}\Big\}. \tag{2.45}$$

Proof. For $\tau \in H(\mathrm{div};\Omega) \cap Z$

$$\|\tau - \mathcal{P}^k_{\mathrm{div},h}(\tau)\|_{\mathrm{div},\Omega} := \inf_{\tau_h \in H^k_h} \|\tau - \tau_h\|_{\mathrm{div},\Omega} \leq \|\tau - \Pi^{RT,k}_h(\tau)\|_{\mathrm{div},\Omega}.$$

Since $\Big\{\tau \in \Big[H^{l+1}(\Omega)\Big]^n ; \mathrm{div}\tau \in H^{l+1}(\Omega)\Big\} \subset H(\mathrm{div};\Omega) \cap Z$, then, according to Theorem (2.3.10), it implies (2.42). □

Approximability of the Continuous Finite Element Space of Degree k (X_h^k) in $H^1(\Omega)$

Lemma 2.3.6. Let $\Pi_{L,h}^k : C(\overline{\Omega}) \to X_h^k$ denote the global Lagrange interpolation operator. Then for each $v \in H^{l+1}(\Omega), 0 \leq l \leq k$, we have

$$\|v - P_{1,h}^k(v)\|_{1,\Omega} \leq Ch^l|v|_{l+1,\Omega}. \tag{2.46}$$

Proof. For $l > 1$ we can use a global version of (2.40). If $l = 0$ we apply the Bramble-Hilbert lemma (cf. Theorem (2.3.6)) to $S = \Omega$ and $\Pi := P_{1,h}^k$ (with $m = 1$ and $k = 0$), and observing that $\Pi(p) = p \; \forall p \in P_0(\Omega) \subset X_h$, we deduce that

$$\|v - P_{1,h}^k(v)\|_{1,\Omega} \leq C|v|_{1,\Omega},$$

which proves that (2.46) can be extended to $l = 0$. □

Approximability of the Continuous Finite Element Space of Degree k (X_h^k) in $L^2(\Omega)$

For this case, we will need a technical lemma that establishes the error estimate for the projection $I - P_{1,h}^k$ in the norm $\|.\|_{0,\Omega}$.

Lemma 2.3.7. Let Ω be a convex domain, and let $k \geq 1$. Then there exists $C > 0$, independent of h, such that for each $v \in H^{k+1}(\Omega)$, $0 \leq l \leq k$, we have

$$\|v - P_{1,h}^k(v)\|_{0,\Omega} \leq Ch^{l+1}|v|_{l+1,\Omega}. \tag{2.47}$$

Proof. See [46, Lemma 4.1] □

Therefore, we can establish the approximability of the Continuous Finite Element Space of Degree k (X_h^k) in $L^2(\Omega)$.

Lemma 2.3.8. Then, if $\Pi_{L,h}^k : C(\overline{\Omega}) \to X_h^k$ denotes the global Lagrange interpolation operator, for each $v \in H^{l+1}(\Omega), 0 \leq l \leq k$, we have

$$\|v - P_h^k(v)\|_{0,\Omega} \leq Ch^{l+1}|v|_{l+1,\Omega}. \tag{2.48}$$

Proof. Recall that $H^{l+1}(\Omega)$ is continuously embedded in $C(\overline{\Omega})$ for $l > 1$ (cf. Theorem 2.3.3). Thus, we can use a global version of (2.40) to obtain

$$\|v - P_h^k(v)\|_{0,\Omega} \leq Ch^{l+1}|v|_{l+1,\Omega}, \quad \forall \tau \in H^{l+1}.$$

Now, by definition it is clear that for $v \in H^1$ we have

$$\|v - P_h^k(v)\|_{0,\Omega} \leq \|v - P_{1,h}^k(v)\|_{0,\Omega},$$

using (2.47) with $l = 0$, it follows that

$$\|v - P_{1,h}^k(v)\|_{0,\Omega} \leq Ch|v|_{1,\Omega}, \quad \forall v \in H^1(\Omega),$$

and therefore we can obtain (2.48). □

Approximability of the Discontinuous Finite Element Space of Degree k (Y_h^k) in $L^2(\Omega)$

Lemma 2.3.9. Finally, consider the projector $\mathcal{P}_h^k : L^2(\Omega) \to Y_h^k$ for $k \geq 0$. Then, for each $v \in H^{l+1}(\Omega), \quad 0 \leq l \leq k$, we have

$$\|v - \mathcal{P}_h^k(v)\|_{0,\Omega} \leq Ch^{l+1}|v|_{l+1,\Omega}. \tag{2.49}$$

Proof. It is easy to see that

$$\mathcal{P}_h^k(v)|_K = \mathcal{P}_K^k(v|_K) \quad \forall v \in L^2(\Omega), \quad \forall K \in \mathcal{T}_h,$$

Now, applying Lemma (2.3.8) to $\Omega = K \in \mathcal{T}_h$, which is obviously convex, we find that

$$\|v - \mathcal{P}_K^k(v)\|_{0,K} \leq Ch_K^{l+1}|v|_{l+1,K} \quad \forall v \in H^{l+1}(K).$$

Thus, for each $v \in H^{l+1}(\Omega)$, $0 \leq l \leq k$, we have

$$\|v - \mathcal{P}_h^k(v)\|_{0,\Omega}^2 = \sum_{K \in \mathcal{T}_h} \|v - \mathcal{P}_K^k(v)\|_{0,K}^2$$

$$\leq \sum_{K \in \mathcal{T}_h} C^2 h_K^{2(l+1)}|v|_{l+1,K}^2 \leq Ch^{2(l+1)}|v|_{l+1,\Omega}^2,$$

resulting in (2.49) □

Summary

When τ and v possess sufficient regularity, $\tau \in [H^{l+1}(\Omega)]^n$ and $v \in H^{l+1}(\Omega)$, $0 \leq l \leq k$, the finite element spaces H_h^k, X_h^k, and Y_h^k (cf. (2.44)) deliver the following optimal approximation estimates:

$$
\begin{aligned}
\operatorname{dist}(\tau, H_h^k) &\leq Ch^{l+1}\left(|\tau|_{l+1,\Omega} + |\mathrm{div}\tau|_{l+1,\Omega}\right)\,, \\
\operatorname{dist}(v, X_h^k) &\leq Ch^{l}|v|_{l+1,\Omega}\,, \\
\operatorname{dist}(v, X_h^k) &\leq Ch^{l+1}|v|_{l+1,\Omega}\,, \\
\operatorname{dist}(v, Y_h^k) &\leq Ch^{l+1}|v|_{l+1,\Omega}\,.
\end{aligned}
$$

2.4 Examples

Example 2.1: Poisson Problem in a 2D domain

In this example, we will apply the results we have seen so far. Our starting point is the mixed formulation to obtain an abstract variational problem in the form of (2.4) with the respective spaces H and Q. Then, we will apply Theorem (2.2.1) to show the solvability of (2.4). Next, we choose the finite-dimensional subspaces $H_h \subset H$ and $Q_h \subset Q$ to obtain the discretized problem. It is important to emphasize that the subspaces H_h and Q_h defining the Galerkin scheme cannot be chosen arbitrarily, as they obviously need to satisfy the hypotheses of Theorem (2.2.1) for solvability and the approximation condition (2.7). Regarding solvability, the most demanding of all is the discrete inf-sup condition for b. In particular, since it is equivalent to the surjectivity of $B_h : H_h \to Q_h$, we deduce that a necessary condition for its occurrence is that $\dim H_h \geq \dim Q_h$. Thus, in this example, we will use a technical lemma known as Fortin's trick [46, Lemma 2.3], which will provide a sufficient condition for the surjectivity of the operator B_h. The approximation will follow from the results of Subsection (2.3.5), and consequently, we will obtain the convergence rates.

Mathematical Model

Let Ω be a bounded domain in R^n, $n \geq 2$, with a Lipschitz-continuous boundary Γ. Then, given $f \in L^2(\Omega)$ and $g \in H^{1/2}(\Gamma)$, we consider the Poisson problem

$$
-\Delta u = f \quad \text{in} \quad \Omega, \quad u = g \quad \text{on} \quad \Gamma. \tag{2.50}
$$

We will use the mixed formulation by adding the unknown $\sigma = \nabla u$, thus obtaining the equivalent problem,

$$
\sigma = \nabla u \quad \text{in} \quad \Omega, \quad \mathrm{div}\,\sigma = -f \quad \text{in} \quad \Omega, \quad u = g \quad \text{on} \quad \Gamma.
$$

Then, multiplying the first equation by $\tau \in H(\mathrm{div}, \Omega)$, integrating by parts, and using the Dirichlet boundary conditions for u, and the second equation by $v \in L^2(\Omega)$, we obtain

$$\int_\Omega \sigma \cdot \tau + \int_\Omega u \,\mathrm{div}\, \tau = \langle \gamma_{\mathbf{n}}(\tau), g \rangle \quad \forall\, \tau \in H(\mathrm{div}; \Omega).$$

$$\int_\Omega \nu \,\mathrm{div}\, \sigma = -\int_\Omega f\, \nu \quad \forall\, \nu \in L^2(\Omega).$$

Continuous Formulation

The mixed variational formulation of (2.50) reduces to the following: find $(\sigma, u) \in H \times Q$ such that

$$\begin{aligned} a(\sigma,\tau) + b(\tau,u) &= F(\tau) \quad \forall \tau \in H, \\ b(\sigma, v) &= G(v) \quad \forall v \in Q. \end{aligned} \tag{2.51}$$

where

$$H := H(\mathrm{div}; \Omega), \quad Q := L^2(\Omega)\,.$$

Here, a and b are the bilinear forms defined by

$$a(\sigma,\tau) := \int_\Omega \sigma \cdot \tau \quad \forall (\sigma,\tau) \in H \times H,$$

$$b(\tau,v) := \int_\Omega v \,\mathrm{div}\, \tau \quad \forall (\tau, v) \in H \times Q,$$

and the functionals $F \in H'$ and $G \in Q'$ are given by

$$F(\tau) := \langle \gamma_n(\tau), g \rangle \quad \forall \tau \in H, \quad G(v) := -\int_\Omega f v \quad \forall v \in Q.$$

Continuous Solvability Analysis

Is the particular case of the Babuška-Brezzi theory (cf. Theorem (2.2.1)) (cf. [46, Section 4.2]) Therefore, Theorem (2.2.1) implies that there exists a unique pair $(\sigma, u) \in H \times \mathcal{Q}$ solution of the mixed variational formulation (2.51) satisfying

$$||(\sigma,u)||_{H \times \mathcal{Q}} \leq C \left\{ ||g||_{1/2,\Gamma} + ||f||_{0,\Omega} \right\}.$$

Galerkin Scheme

If $\{\mathcal{T}_h\}_{h>0}$ is a regular family of triangulations of Ω and k is an integer ≥ 0, we introduce the following finite element spaces:

$$\begin{cases} H_h := H_h^k := \{\tau_h \in H(\mathrm{div};\Omega): \quad \tau_h|_K \in RT_k(K) \quad \forall K \in \mathcal{T}_h\}, \\ Q_h := Y_h^k := \Big\{v_h \in L^2(\Omega): \quad v_h|_K \in \mathbb{P}_k(K) \quad \forall K \in \mathcal{T}_h\Big\}, \end{cases}$$

so that the associated Galerkin scheme is the following: find $(\sigma_h, u_h) \in H_h \times Q_h$ such that

$$\begin{aligned} a(\sigma_h, \tau_h) + b(\tau_h, u_h) &= F(\tau_h) \quad \forall \tau_h \in H_h, \\ b(\sigma_h, v_h) &= G(v_h) \quad \forall v_h \in Q_h. \end{aligned} \tag{2.52}$$

Consequently, a direct application of the discrete version of Theorem (2.2.1) implies that there exists a unique solution $(\sigma_h, u_h) \in H_h \times Q_h$ of (2.6) and a constant $C > 0$, independent of h, such that

$$\|(\sigma_h, u_h)\|_{H\times Q} \leq C\Big\{\|f\|_{0,\Omega} + \|g\|_{1/2,\Gamma}\Big\}.$$

A priori error analysis

Using Cea's estimate (cf. (2.10)), we obtain:

$$\|\sigma - \sigma_h\|_H + \|u - u_h\|_Q \leq C\left\{\mathrm{dist}(\sigma, H_h) + \mathrm{dist}(u, Q_h)\right\},$$

where C depends on $\|A\|$, $\|B\| \leq 1$, $\tilde{\alpha}$, and $\tilde{\beta}$. According to the upper bounds for projection errors given by (2.45) and (2.49), we have respectively,

$$\mathrm{dist}(\sigma, H_h) := \|\sigma - \mathcal{P}_{\mathrm{div},h}^k(\sigma)\|_{\mathrm{div},\Omega} \leq Ch^{l+1}\left\{|\sigma|_{l+1,\Omega} + |\operatorname{div}\sigma|_{l+1,\Omega}\right\} \tag{2.53}$$

if $\sigma \in [H^{l+1}(\Omega)]^n$, with $\operatorname{div}\sigma \in H^{l+1}(\Omega)$, $0 \leq l \leq k$, and

$$\mathrm{dist}(u, Q_h) := \|u - \mathcal{P}_h^k(u)\|_{0,\Omega} \leq Ch^{l+1}|u|_{l+1,\Omega} \tag{2.54}$$

if $u \in H^{l+1}(\Omega)$, $0 \leq l \leq k$. Therefore, under these regularity assumptions on the exact solution $(\sigma, u) \in H \times Q$, we deduce that the convergence rate of the Galerkin method (2.6) is given by the estimate following from (2.53)-(2.54), namely,

$$\|\sigma - \sigma_h\|_{\mathrm{div},\Omega} + \|u - u_h\|_{0,\Omega} \leq Ch^{l+1}\Big\{|\sigma|_{l+1,\Omega} + |\operatorname{div}\sigma|_{l+1,\Omega} + |u|_{l+1,\Omega}\Big\}. \tag{2.55}$$

On the other hand, if (σ, u) is not sufficiently regular, the convergence of the Galerkin scheme (2.6), but without any convergence rate, can still be proved using appropriate density arguments. More precisely, we have the following result.

Lemma 2.4.10. Let $(\sigma, u) \in H \times Q$ and $(\sigma_h, u_h) \in H_h \times Q_h$ be the solutions of the continuous `and` discrete formulations, respectively. Then

$$\lim_{h \to 0} \{\|\sigma - \sigma_h\|_{\mathrm{div},\Omega} + \|u - u_h\|_{0,\Omega}\} = 0 \,.$$

Proof. See [46, Lemma 4.5]. □

Numerical Results

To illustrate the performance of the mixed finite element method on a set of uniform domain triangulations, we consider a function u such that Δu exists, and define the source term f so that equation (2.50) is satisfied. Then u is called a manufactured solution. We implement the numerical method for the discretized problem (2.52) with f defined as above, obtaining u_h and σ_h, and compare the errors $u - u_h$ and $\sigma - \sigma_h$. We use the open-source finite element library `FEniCS` [52]. We use a laptop with an Intel Core i5 10th generation processor and 16 GB of memory. The execution time for $l = 1$ and $131,072$ elements $(1,049,600$ degrees of freedom) was 34 seconds. The code is provided at the end of this section.

The individual errors are denoted by

$$\mathsf{e}(\sigma) := \|\sigma - \sigma_h\|_{\mathrm{div},\Omega}, \quad \mathsf{e}(u) := \|u - u_h\|_{0,\Omega},$$

and, for each $\star \in \{\sigma, u\}$ we define $r(\star)$ as the experimental convergence rate given by

$$r(\star) := \frac{\log(\mathsf{e}(\star)/\widehat{\mathsf{e}}(\star))}{\log(h/\widehat{h})},$$

where h and $\widehat{h}$ denote two consecutive mesh sizes with errors e and $\widehat{\mathsf{e}}$, respectively. In this test, we confirm the convergence rates on a two-dimensional domain defined by the square $\Omega = (0,1)^2$. We adjust the data f so that

$$u(x,y) = \left(\frac{10}{6}\right) \sin\Big(2\pi(x+0.5)\Big) \sin(2\pi y) \cos\Big(\pi(x+0.5)\Big) \sin(\pi y)$$

Figure (2.2) shows the potential function u, and the new unknown $\sigma = \nabla u$, which represents the field associated with the potential function. Meanwhile, Figure (2.1) confirms that the optimal convergence rates $\mathcal{O}(h^{\ell+1})$, predicted by Equation (2.55), are achieved for $\ell = \{0, 1\}$.

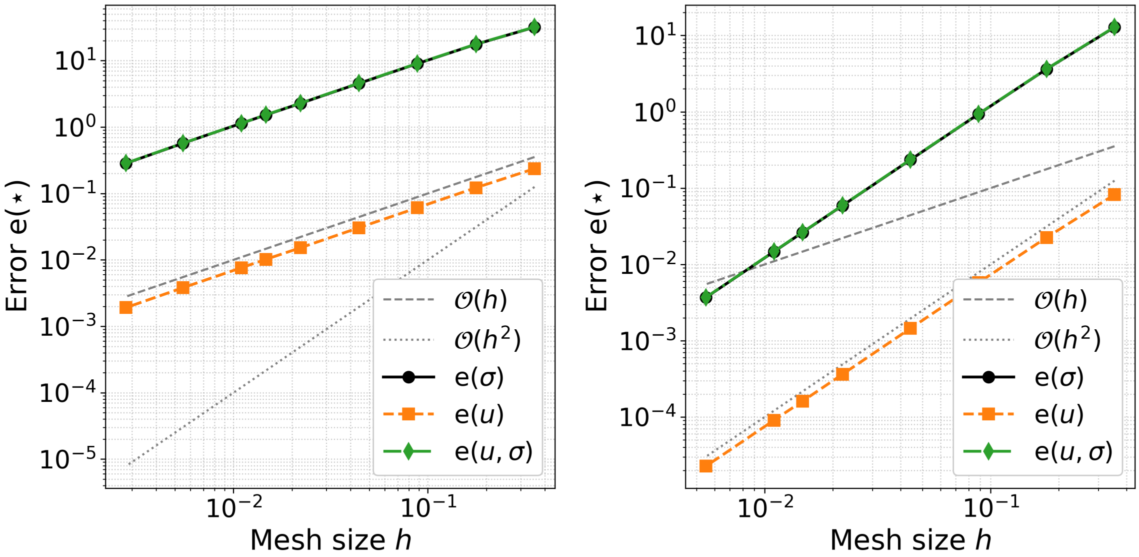


Figure 2.1 [Example 2.1] Convergence rates of the errors for each unknown u and σ and the total error, for $l = 0$ and $l = 1$.

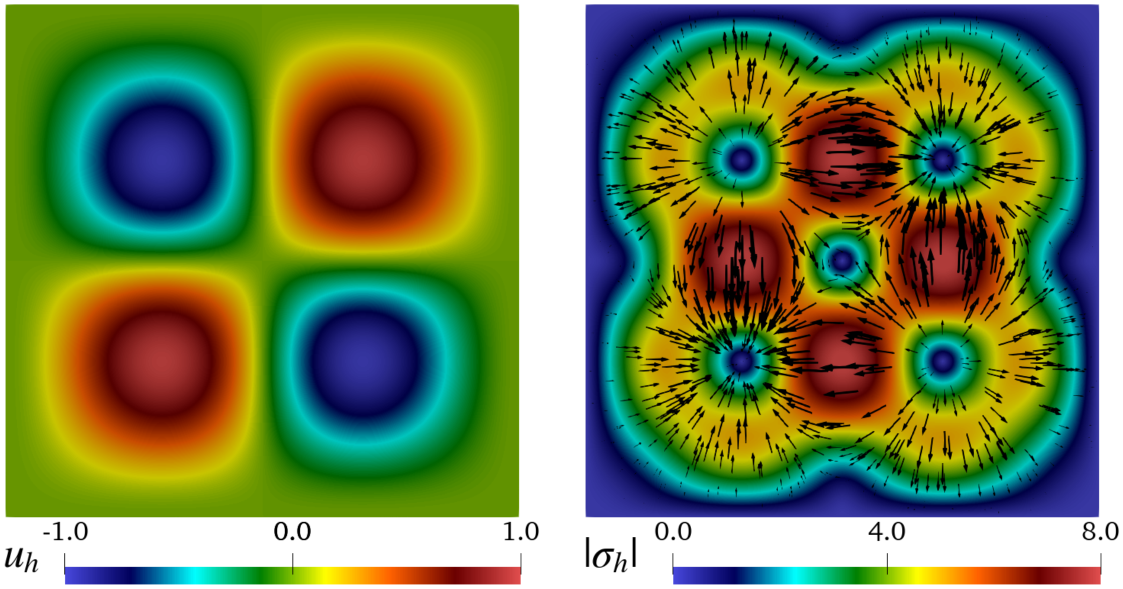


Figure 2.2 [Example 2.1] Potential u and field $\sigma = \nabla u$.

Code

```
from fenics import *
import matplotlib.pyplot as plt
import sympy as sp
import numpy as np
from math import log
import time

# <----- Model parameters and auxiliary symbolic expressions ----->
Id  = Identity(2)  # 2x2 identity matrix
x, y = sp.symbols(’x[0] x[1]’)  # Symbolic variables for x and y coordinates
pi = sp.pi  # Define pi in SymPy

# <----- Manufactured solution (known exact solution) ----->
ue = (10/6)*sp.sin(2*pi*(x+0.5))*sp.sin(2*pi*y)*sp.cos(1*pi*(x+0.5))*sp.sin(1*pi*y)

# Gradient of exact solution (x and y components)
grad_ue_1 = ue.diff(x, 1)  # Partial derivative of ue with respect to x
grad_ue_2 = ue.diff(y, 1)  # Partial derivative of ue with respect to y

# Exact flux (sigma = grad(ue))
sigmae_1 = grad_ue_1  # x-component of flux
sigmae_2 = grad_ue_2  # y-component of flux

# Divergence of exact flux (div(sigmae))
div_sigmae = sigmae_1.diff(x, 1) + sigmae_2.diff(y, 1)

# <----- Manufactured source term (fe = -div(sigmae)) ----->
fe = -div_sigmae

# <----- Converting symbolic expressions to FEniCS mathematical functions ----->
f = Expression(sp.printing.ccode(fe), degree=5)  # Source term
ue = Expression(sp.printing.ccode(ue), degree=5)  # Exact solution
sigmae = Expression((sp.printing.ccode(sigmae_1), sp.printing.ccode(sigmae_2)), degree=5)  # Exact flux
div_sigmae = Expression(sp.printing.ccode(div_sigmae), degree=5)  # Divergence of exact flux

# <----- Variational Poisson solver (mixed formulation) ----->
def PoissonSolver(W, f):
    # Define test and trial functions
    (sigma, u) = TrialFunctions(W)  # Trial functions (unknowns)
    (tau, v) = TestFunctions(W)  # Test functions

    # Variational form
    a = (dot(sigma, tau) + div(tau)*u + div(sigma)*v) * dx  # Bilinear form
    L = -f * v * dx  # Linear form

    # Solve the system
    w = Function(W)  # Function to store the solution
    solve(a == L, w)  # Solve the linear system

    # Extract sigma and u from solution
    sigma, u = w.split()
    return sigma, u

# <----- Initialization of vectors to store results ----->
vec_nelem = []  # Number of elements
vec_hh = []  # Mesh size (h)
vec_dofs = []  # Degrees of freedom
vec_time = []  # Processing time
vec_err_sig = []  # Flux error
vec_err_u = []  # Solution error
vec_err_tot = []  # Total error
```

```python

# <----- Mesh refinement ----->
NN = [4, 8, 16, 32, 64, 96, 128, 256]  # Number of divisions per mesh side

for i in NN:
    # <----- Create regular mesh ----->
    mesh = UnitSquareMesh(i, i)  # Unit square mesh
    h = mesh.hmax()  # Maximum mesh size
    nelem = mesh.num_cells()  # Number of elements in mesh

    # <----- Define function spaces ----->
    order = 1  # Finite element order
    RT = FiniteElement("RT", mesh.ufl_cell(), order + 1)  # Space for sigma (flux)
    DG = FiniteElement("DG", mesh.ufl_cell(), order)  # Space for u (potential)
    W = FunctionSpace(mesh, RT * DG)  # Mixed space

    # <----- Mesh information ----->
    h = mesh.hmax()
    nelem = mesh.num_cells()
    print('Number of elements: ', nelem)
    dim = W.dim()  # Degrees of freedom

    # <----- Measure processing time ----->
    start_time = time.perf_counter()  # Start time counting
    sigma, u = PoissonSolver(W, f)  # Solve the problem
    end_time = time.perf_counter()  # End time counting
    elapsed_time = end_time - start_time  # Elapsed time

    # <---- Calculate errors ---->
    err_u = pow(assemble((u - ue)**2 * dx), 1./2.)  # Solution error (L2 norm)
    err_sig_L2 = pow(assemble(inner(sigma - sigmae, sigma - sigmae) * dx), 1./2.)  # Flux error (L2 norm)
    err_sig_div = pow(assemble(dot(div_sigmae - div(sigma), div_sigmae - div(sigma)) * dx), 1./2.)  # Divergence error
    err_sig = err_sig_L2 + err_sig_div  # Total flux error
    err_tot = err_u + err_sig  # Total error

    # <------ Store results ------->
    vec_hh.append(h)
    vec_nelem.append(nelem)
    vec_dofs.append(dim)
    vec_time.append(elapsed_time)
    vec_err_sig.append(err_sig)
    vec_err_u.append(err_u)
    vec_err_tot.append(err_tot)

# <----- Display/export data ----->
mytable = [["#elements", "h", "dofs", "time", "e_sig", "r_sig", "e_u", "r_u", "e_tot", "r_tot" ]]

i = 0
while i < len(vec_err_u):
    if i == 0:
        # First row (no convergence rate)
        mytable.append([
            "%6.0f" % vec_nelem[i],  # Number of elements
            "%2.4f" % vec_hh[i],  # Mesh size
            "%6.0f" % vec_dofs[i],  # Degrees of freedom
            "%2.4f" % vec_time[i],  # Processing time
            "%2.2e" % vec_err_sig[i], 0,  # Flux error and convergence rate
            "%2.2e" % vec_err_u[i], 0,  # Solution error and convergence rate
            "%2.2e" % vec_err_tot[i], 0  # Total error and convergence rate
        ])
    else:
        # Calculate convergence rates
        rate_sig = log(vec_err_sig[i] / vec_err_sig[i-1]) / log(vec_hh[i] / vec_hh[i-1])
        rate_u = log(vec_err_u[i] / vec_err_u[i-1]) / log(vec_hh[i] / vec_hh[i-1])
        rate_tot = log(vec_err_tot[i] / vec_err_tot[i-1]) / log(vec_hh[i] / vec_hh[i-1])

        # Add row to table
        mytable.append([
            "%6.0f" % vec_nelem[i],  # Number of elements
            "%2.4f" % vec_hh[i],  # Mesh size
            "%6.0f" % vec_dofs[i],  # Degrees of freedom
            "%2.4f" % vec_time[i],  # Processing time
            "%2.2e" % vec_err_sig[i], "%2.3f" % rate_sig,  # Flux error and convergence rate
            "%2.2e" % vec_err_u[i], "%2.3f" % rate_u,  # Solution error and convergence rate
            "%2.2e" % vec_err_tot[i], "%2.3f" % rate_tot  # Total error and convergence rate
        ])
    i = i + 1

# Display table
for row in mytable:
    print("{:<10} {:<8} {:<8} {:<8} {:<10} {:<10} {:<10} {:<10} {:<10} {:<10}".format(*row))

# <----- Export graphics for visualization in Paraview ----->
sig_file = File("Data_Paraview_2D/approx_sig.pvd") << sigma  # Approximate flux
u_file = File("Data_Paraview_2D/approx_u.pvd") << u  # Approximate solution

# <----- Interpolate exact solutions for visualization in Paraview ----->
V1 = FunctionSpace(mesh, "CG", 2)  # Continuous space for u
V2 = VectorFunctionSpace(mesh, "CG", 2)  # Continuous vector space for sigma
SIGMA = interpolate(sigmae, V2)  # Interpolate exact flux
U = interpolate(ue, V1)  # Interpolate exact solution

# <----- Export exact solutions ----->
SIG_file = File("Data_Paraview_2D/exact_sig.pvd") << SIGMA  # Exact flux
U_file = File("Data_Paraview_2D/exact_u.pvd") << U  # Exact solution
```

Chapter 3

A priori error analysis for $\mu(I)$-rheology

3.1 Chapter Introduction

The major difficulty imposed by the $\mu(I)$-rheology model is the dependence of the dissipative terms on the pressure of the flow. This will be presented in more detail in the following section. However, it is clear that this poses an extra complication to the numerical algorithms that are normally based on pressure-correction projection schemes [14]. In other words, the strong non-linearity of the $\mu(I)$-rheology model prevents us from guaranteeing in advance successful applications of classical numerical methods, such as primal finite elements and related techniques, which are known to be usually more suitable for linear problems, particularly if they are posed within a Hilbertian framework. In this regard, we find it important to stress that the suitability of Banach spaces-based approaches to analyze the continuous and discrete solvabilities of diverse nonlinear problems in continuum mechanics, including several coupled models, and employing mainly mixed formulations, has been confirmed by a significant amount of contributions in recent years. Brinkman-Forchheimer, Darcy-Forchheimer, Navier-Stokes, Boussinesq, coupled flow-transport, and fluidized beds are some of the respective models addressed, and a non-exhaustive list of the corresponding references includes [15–22]. Needless to say, the most distinctive feature of a mixed formulation is the incorporation of additional unknowns, usually depending on the original ones of the model, for either analytical or physical reasons.

Furthermore, one of the main advantages of employing a Banach framework is the fact that no augmentation is required, a common “trick" of Hilbert spaces-based formulations to force them to become, for instance, elliptic or strongly monotone, and hence the spaces to which the unknowns belong are the natural ones arising simply from the testing of the equations of the model along with the use of the Cauchy-Schwarz and Hölder inequalities. In this way, simpler and closer to the original physical model formulations are derived. In turn, the main benefits of employing a mixed approach include the derivation of momentum-conservative numerical schemes, and the possibility of obtaining direct approximations of further variables of physical interest, either by incorporating them into the formulation, or by employing a postprocessing formula in terms of the remaining unknowns. In the particular case of our model of interest, to be described below in Section 3.2, the above might certainly mean to be able to obtain direct calculations of strain rate tensor, shear rate, inertia number, and vorticity, among other variables of interest, thus avoiding numerical differentiation and its consequent loss of accuracy, to approximate them.

According to the previous discussion in the Introduction of this thesis, the goal of the present Chapter is to introduce and analyze mixed finite element methods for numerically solving the steady-state $\mu(I)$-rheology equations for granular flows. The Chapter is organized as follows. In the rest of this section we collect some notations to be employed throughout the chapter. In Section 3.2 we describe the mathematical model, which includes the setting of a regularized sity, and introduce, besides the velocity and the pressure, the further unknowns to be considered. Next, in Section 3.3 we develop the mixed variational formulation, which is shown to have a twofold saddle point-type structure. The corresponding solvability analysis is carried out in Section 3.4 by adopting a fixed-point strategy in terms of the velocity and the pressure, and by employing an abstract result on the well-posedness of Banach spaces-based twofold saddle point operator equations, along with the classical Banach theorem. Lipschitz-continuity and motononicity properties of the viscosity function are also required for the analysis. In turn, in Section 3.5 we define the associated Galerkin scheme, and assume suitable hypotheses on the finite element subspaces in order to prove the corresponding well-posedness by means of a discrete fixed-point approach. A priori error estimates are also obtained here. Then, specific finite element subspaces satisfying the aforementioned assumptions, are derived in Section 3.6 by applying a useful connection with the discrete stability of the usual Hilbertian mixed formulation for linear elasticity, and optimal rates of convergence are established as well. Finally, numerical experiments illustrating the theoretical findings are reported in Section 3.7, whereas the fulfillment of the hypotheses on the viscosity is discussed in Appendix A.1.

3.2 The mathematical model

We recall the $\mu(I)$-rheology equations introduced in Introduction. We are interested in the flows of granular materials based on the $\mu(I)$-rheology approach introduced in [7]. This rheological model arose from the fundamental hypothesis that the corresponding stresses can be described by a viscoplastic constitutive equation in which the internal friction μ of the material, which governs the yield stress, is not constant and depends on a flow parameter called the inertial number I. In order to introduce the corresponding mathematical model, we consider the flow of particles of constant density ρ_p and diameter d in Ω, denote by $\mathbf{u}$ the velocity of the flow, and assume that the latter is incompressible, that is, the volume fraction ϕ of particles is constant throughout the flow, so that the overall density is $\rho = \phi\rho_p$. The governing equations are then given by:

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\nabla\mathbf{u})\mathbf{u}\right) = \mathbf{div}(\boldsymbol{\sigma}) + \rho\,\mathbf{g} \quad \text{in} \quad \Omega\,, \tag{3.1}$$

and

$$\operatorname{div}(\mathbf{u}) = 0 \quad \text{in} \quad \Omega\,. \tag{3.2}$$

In turn, the stress tensor $\boldsymbol{\sigma}$ is composed of two terms, a deviatoric one associated to dissipation due to the internal friction of the medium, which is inspired by a Coulomb friction-like law, and an isotropic one related to the pressure p on the medium. More precisely, there holds

$$\boldsymbol{\sigma} = \sqrt{2}\,\mu\,p\,\frac{\mathbf{D}}{|\mathbf{D}|} - p\,\mathbb{I} \quad \text{in} \quad \Omega\,, \tag{3.3}$$

where μ is the internal friction coefficient of the granular continuum, $\mathbf{D}$ is the symmetric part of the velocity gradient, namely

$$\mathbf{D} := \frac{1}{2}\Big(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathtt{t}}\Big), \tag{3.4}$$

which is also known as the rate of strain tensor, and

$$|\mathbf{D}| = \sqrt{\mathbf{D} : \mathbf{D}}\,. \tag{3.5}$$

Note, thanks to the incompressibility condition (3.2), that there holds

$$\operatorname{tr}(\mathbf{D}) = \operatorname{div}(\mathbf{u}) = 0\,. \tag{3.6}$$

Now, if the friction coefficient is constant, we have the traditional Coulomb model for granular materials [53]. However, there is strong evidence [6] that μ actually depends on the local properties of the flow through the inertial number I, in the form

$$\mu(I) \,:=\, \mu_s + \left(\frac{\mu_d - \mu_s}{I + I_0}\right) I \qquad \text{with} \qquad I = \frac{\sqrt{2}\, d\, |\mathbf{D}|}{\sqrt{p/\rho}}\,, \tag{3.7}$$

where the coefficients μ_s and μ_d correspond, respectively, to the static and dynamic friction limits, and I_0 is a reference (experimental) constant. It is easy to see from the above expression for $\mu(I)$ that

$$\min\Big\{\mu_s, \mu_d\Big\} \,\le\, \mu(I) \,\le\, \max\Big\{\mu_s, \mu_d\Big\}\,,$$

so that, assuming from now on, for simplicity, that $\mu_s \le \mu_d$, there holds

$$\mu_s \,\le\, \mu(I) \,\le\, \mu_d\,.$$

Then, substituting (3.7) in the constitutive relation (3.3), we arrive at

$$\boldsymbol{\sigma} \,=\, \eta(p, |\mathbf{D}|)\,\mathbf{D} \,-\, p\,\mathbb{I} \quad \text{in} \quad \Omega\,, \tag{3.8}$$

where $\eta : \mathrm{R}^+ \times \mathrm{R}^+ \longrightarrow \mathrm{R}^+$ is defined as

$$\eta(\varrho, \omega) \,:=\, \frac{a_1\,\varrho}{\omega} \,+\, \frac{a_2\,\varrho}{a_3\,\sqrt{\varrho} + a_4\,\omega} \qquad \forall\,(\varrho, \omega) \in \mathrm{R}^+ \times \mathrm{R}^+\,, \tag{3.9}$$

with positive coefficients a_i, $i \in \Big\{1, 2, 3, 4\Big\}$, given by

$$a_1 \,:=\, \sqrt{2}\,\mu_s\,, \quad a_2 \,:=\, 2\,d(\mu_d - \mu_s)\,, \quad a_3 \,:=\, \rho^{-1/2}\,I_0\,, \quad \text{and} \quad a_4 \,:=\, \sqrt{2}\,d\,. \tag{3.10}$$

It is important to stress here that, due to the fact that μ is defined in terms of I, which, in turn, depends on p and $|\mathbf{D}|$ (cf. (3.7)), the function η, and thus its evaluation $\eta(p, |\mathbf{D}|)$, have been introduced to emphasize that the expression $\dfrac{\sqrt{2}\,\mu\,p}{|\mathbf{D}|}$ (which multiplies $\mathbf{D}$ in (3.3)) depends only on those unknowns, and that this dependence can be explicitly stated, as (3.9) shows. Hence, being (3.8) just a rewriting of (3.3), working with one or the other is basically the same, but the former is much more suitable for identifying later on the assumptions needed for the analysis.

We now notice that the term $\eta(p, |\mathbf{D}|)$ in (3.8), which can be understood as an equivalent viscosity, is singular when $|\mathbf{D}| = 0$. Indeed, it is expected that some regions of the granular flows are static, as granular materials can exhibit a solid-like behavior [2], just as in a sand pile. In this particular case, the flow of grains only happens near the surface of the dunes, while in the inner core of flow, the material remains static (and resist stresses). In these static regions, the $\mu(I)$-rheology model, which is valid for fluid-like flows of granular materials [7], breaks drown. Similar problems are also observed in flows of different visco-plastic materials [54]. In addition to the theoretical constitutive problem, the singularity of $\eta(p, |\mathbf{D}|)$ also poses technical computational difficulties, as the very large values it can assume in the domain of the flow can lead to ill-posed linear systems that undermine the performance of standard solvers [10]. Therefore, a regularization technique has to be used in order to avoid the presence of the afore-mentioned singularity. This can be done in different ways [10, 11, 54], although the underlying assumption in all cases is that the unyielded regions should be treated as practically unyielded, i.e. creeping, regions [54] with a limited maximum value of $\eta(p, |\mathbf{D}|)$. For instance, one way is to add a small parameter $0 < \varepsilon \ll 1$ to the denominators in (3.9), thus yielding

$$\eta(\varrho, \omega) \,:=\, \frac{a_1\,\varrho}{\omega + \varepsilon} + \frac{a_2\,\varrho}{a_3\,\sqrt{\varrho} + a_4\,\omega + \varepsilon} \qquad \forall\,(\varrho, \omega) \in \mathrm{R}^+ \times \mathrm{R}^+ \,. \tag{3.11}$$

Finally, regarding boundary conditions, and knowing that recent evidence [55] suggests that there can be some slip between the grains and the boundaries, we proceed accordingly and assume this condition for the steady-state regime that we consider below.

In virtue of the above discussion, the governing equations of the stationary model arising from (3.1), (3.2), and (3.8), are given by

$$\begin{gathered} \rho\,(\nabla\mathbf{u})\mathbf{u} \,=\, \mathbf{div}\Big(\eta(p, |\mathbf{D}|)\,\mathbf{D}\Big) \,-\, \nabla p \,+\, \rho\,\mathbf{g} \quad \text{in} \quad \Omega\,, \\ \operatorname{div}(\mathbf{u}) \,=\, 0 \quad \text{in} \quad \Omega\,, \quad \mathbf{u} \,=\, \mathbf{u}_D \quad \text{on} \quad \Gamma\,, \end{gathered} \tag{3.12}$$

where $\mathbf{u}_D \in \mathbf{H}^{1/2}(\Gamma)$ constitutes a non-necessarily null Dirichlet boundary condition for $\mathbf{u}$. In addition, since our main interest is to develop a fully-mixed finite element method for (3.12), we now introduce a modified stress tensor, still denoted $\boldsymbol{\sigma}$, as the further unknown defined by

$$\boldsymbol{\sigma} \,:=\, \eta(p, |\mathbf{D}|)\,\mathbf{D} \,-\, p\,\mathbb{I} \,-\, \rho\,(\mathbf{u} \otimes \mathbf{u})\,. \tag{3.13}$$

In this way, recalling that the overall density is constant, and noting that the incompressibility condition allows us to show that $\mathbf{div}\big(\mathbf{u}\otimes\mathbf{u}\big) = (\nabla\mathbf{u})\mathbf{u}$, we deduce that the momentum equation can be rewritten as

$$\mathbf{div}(\boldsymbol{\sigma}) + \rho\,\mathbf{g} = 0 \quad \text{in} \quad \Omega\,. \tag{3.14}$$

Moreover, applying deviatoric operator (cf. (2.1)) to (3.13), and using (3.6), which obviously yields $\mathbf{D}^{\mathrm{d}} = \mathbf{D}$, we find that

$$\boldsymbol{\sigma}^{\mathrm{d}} := \eta(p,|\mathbf{D}|)\,\mathbf{D} - \rho\,(\mathbf{u}\otimes\mathbf{u})^{\mathrm{d}} \quad \text{in} \quad \Omega\,. \tag{3.15}$$

In turn, applying now matrix trace to (3.13), we obtain an explicit formula for the pressure p in terms of $\boldsymbol{\sigma}$ and $\mathbf{u}$, namely

$$p = -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma} + \rho\,(\mathbf{u}\otimes\mathbf{u})\Big)\,. \tag{3.16}$$

We remark here that (3.13) and the incompressibility condition (3.2) are jointly equivalent to (3.15) - (3.16). On the other hand, in order to perform the usual integration by parts procedure required by a mixed formulation, which reduces to be able to test $\nabla\mathbf{u}$, we now decompose $\mathbf{D}$ as

$$\mathbf{D} = \nabla\mathbf{u} - \boldsymbol{\gamma}\,, \tag{3.17}$$

where $\boldsymbol{\gamma}$ is the auxiliary known given by

$$\boldsymbol{\gamma} := \frac{1}{2}\Big(\nabla\mathbf{u} - (\nabla\mathbf{u})^{\mathrm{t}}\Big)\,. \tag{3.18}$$

Note that the diagonal entries of $\boldsymbol{\gamma}$ are all null, and that the off diagonal ones include the components of the vorticity $\nabla\times\mathbf{u}$. Summarizing, (3.12) can be equivalently reformulated as: Find $\mathbf{D}$, $\boldsymbol{\sigma}$, $\mathbf{u}$, p, and $\boldsymbol{\gamma}$ in suitable spaces, to be defined later on, such that

$$\begin{aligned}
\mathbf{D} - \nabla\mathbf{u} + \boldsymbol{\gamma} &= 0 \qquad \text{in} \quad \Omega\,,\\
\eta(p,|\mathbf{D}|)\,\mathbf{D} - \boldsymbol{\sigma}^{\mathrm{d}} - \rho\,(\mathbf{u}\otimes\mathbf{u})^{\mathrm{d}} &= 0 \qquad \text{in} \quad \Omega\,,\\
\mathbf{div}(\boldsymbol{\sigma}) + \mathbf{f} &= 0 \qquad \text{in} \quad \Omega\,,\\
p = -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma} + \rho\,(\mathbf{u}\otimes\mathbf{u})\Big) \quad \text{in} \quad \Omega\,, \quad \mathbf{u} &= \mathbf{u}_D \qquad \text{on} \quad \Gamma\,,
\end{aligned} \tag{3.19}$$

where, for sake of generality as well as for convenience of the numerical experiments to be reported later on, we have replaced $\rho\,\mathbf{g}$ by a source term $\mathbf{f}$, which belongs to a space to be precised in due course. We end this section by remarking that, because of (3.2), the datum $\mathbf{u}_D$ must satisfy the compatibility condition

$$\int_\Gamma \mathbf{u}_D \cdot \boldsymbol{\nu} \,=\, 0\,. \tag{3.20}$$

3.3 The continuous formulation

In this section we derive a variational formulation for the system (3.19). To this end, we first proceed analogously to [56, Section 3] and look originally for $\mathbf{u}$ in $\mathbf{H}^1(\Omega)$. In this way, multiplying the first equation of (3.19) by $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega)$, where $t \in \begin{cases} (1,+\infty) & \text{if } \; n \,=\, 2 \\ [6/5,+\infty) & \text{if } \; n \,=\, 3 \end{cases}$, and then applying the integration by parts formula (2.3) along with the Dirichlet boundary condition for $\mathbf{u}$, we obtain

$$\int_\Omega \boldsymbol{\tau} : \mathbf{D} \,+\, \int_\Omega \mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau}) \,+\, \int_\Omega \boldsymbol{\tau} : \boldsymbol{\gamma} \,=\, \langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \mathbf{u}_D\rangle \qquad \forall\,\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega)\,. \tag{3.21}$$

We notice that the first and third terms make sense for $\mathbf{D}$, $\boldsymbol{\gamma} \in \mathbb{L}^2(\Omega)$, which, due to the free trace property of $\mathbf{D}$ (cf. (3.6)) and the skew symmetry of $\boldsymbol{\gamma}$ (cf. (3.18)), leads to look for $\mathbf{D} \in \mathbb{L}^2_{\mathtt{tr}}(\Omega)$ and $\boldsymbol{\gamma} \in \mathbb{L}^2_{\mathtt{sk}}(\Omega)$, where

$$\mathbb{L}_{\mathtt{tr}}(\Omega) \,:=\, \Big\{ \mathbf{E} \in \mathbb{L}^2(\Omega) : \quad \mathrm{tr}(\mathbf{E}) \,=\, 0 \Big\}\,, \tag{3.22}$$

and

$$\mathbb{L}_{\mathtt{sk}}(\Omega) \,:=\, \Big\{ \boldsymbol{\xi} \in \mathbb{L}^2(\Omega) : \quad \boldsymbol{\xi}^{\mathtt{t}} \,=\, -\boldsymbol{\xi} \Big\}\,. \tag{3.23}$$

In turn, since $\mathbf{div}(\boldsymbol{\tau}) \in \mathbf{L}^t(\Omega)$, we realize by Hölder's inequality that the second term from (3.21) is actually well defined for $\mathbf{u} \in \mathbf{L}^{t'}(\Omega)$, where $t' \in (1,+\infty)$ is the conjugate of t. On the other hand, in order to continue the present derivation, we need to introduce the following hypothesis:

(**H.1**) there exist constants η_1, η_2 such that

$$0 \,<\, \eta_1 \,\leq\, \eta(\varrho,\omega) \,\leq\, \eta_2 \qquad \forall\,(\varrho,\omega) \in \mathrm{R}^+ \times \mathrm{R}^+\,. \tag{3.24}$$

Certainly, the above assumption might imply the need to suitably redefine η in (3.11). Next, testing the second equation of (3.19) against $\mathbf{E} \,\in\, \mathbb{L}^2_{\mathtt{tr}}(\Omega)$, and using that

$\boldsymbol{\zeta}^{\mathrm{d}} : \mathbf{E} \,=\, \boldsymbol{\zeta} : \mathbf{E}$ for all $\boldsymbol{\zeta} \in \mathbb{L}^2(\Omega)$, we formally obtain

$$\int_\Omega \eta(p, |\mathbf{D}|)\, \mathbf{D} : \mathbf{E} \,-\, \int_\Omega \boldsymbol{\sigma} : \mathbf{E} \,-\, \rho \int_\Omega (\mathbf{u} \otimes \mathbf{u}) : \mathbf{E} \,=\, 0\,, \tag{3.25}$$

which says, thanks to (3.24), that the first term is well defined, whereas the second one makes sense if $\boldsymbol{\sigma}$ is sought in $\mathbb{L}^2(\Omega)$. Regarding the last term, we first notice, thanks to Cauchy-Schwarz's inequality in $\mathrm{L}^2(\Omega)$ and R^n, that there holds

$$\|\mathbf{w} \otimes \mathbf{v}\|_{0,\Omega} \,\leq\, n^{1/2}\, \|\mathbf{w}\|_{0,4;\Omega}\, \|\mathbf{v}\|_{0,4;\Omega} \qquad \forall\, \mathbf{w},\, \mathbf{v} \in \mathbb{L}^4(\Omega)\,. \tag{3.26}$$

It follows that

$$\left| \int_\Omega (\mathbf{u} \otimes \mathbf{u}) : \mathbf{E} \right| \,\leq\, \|(\mathbf{u} \otimes \mathbf{u})\|_{0,\Omega}\, \|\mathbf{E}\|_{0,\Omega} \,\leq\, n^{1/2}\, \|\mathbf{u}\|^2_{0,4;\Omega}\, \|\mathbf{E}\|_{0,\Omega}\,, \tag{3.27}$$

from which we deduce that it suffices to consider $t' \,=\, 4$, thus looking for $\mathbf{u}$ in $\mathbf{L}^4(\Omega)$ (equivalently $(\mathbf{u} \otimes \mathbf{u}) \in \mathbb{L}^2(\Omega)$), and then $t \,=\, 4/3$, whence the test space of (3.21) becomes $\mathbb{H}(\mathbf{div}_{4/3};\Omega)$. The above suggests to seek $\boldsymbol{\sigma}$ in this same space, which requires $\mathbf{f}$ to belong to $\mathbf{L}^{4/3}(\Omega)$, so that the third equation of (3.19) is tested as

$$\int_\Omega \mathbf{v} \cdot \mathbf{div}(\boldsymbol{\sigma}) \,=\, -\int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad \forall\, \mathbf{v} \in \mathbf{L}^4(\Omega)\,. \tag{3.28}$$

Now, having identified the spaces to which $\boldsymbol{\sigma}$ and $\mathbf{u}$ belong, we realize from the first equation in the last row of (3.19) that the pressure p must be sought in $\mathrm{L}^2(\Omega)$. Furthermore, the symmetry of $\boldsymbol{\sigma}$ (cf. (3.13)) is weakly imposed by

$$\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\xi} \,=\, 0 \qquad \forall\, \boldsymbol{\xi} \in \mathbb{L}^2_{\mathtt{sk}}(\Omega)\,. \tag{3.29}$$

Finally, we resort to the decomposition

$$\mathbb{H}(\mathbf{div}_{4/3};\Omega) \,=\, \mathbb{H}_0(\mathbf{div}_{4/3};\Omega) \,\oplus\, \mathrm{R}\,\mathbb{I}\,, \tag{3.30}$$

where

$$\mathbb{H}_0(\mathbf{div}_{4/3};\Omega) \,:=\, \left\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_{4/3};\Omega) : \quad \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) \,=\, 0 \right\}. \tag{3.31}$$

In this way, the unknown $\boldsymbol{\sigma}$ can be decomposed as $\boldsymbol{\sigma} \,=\, \boldsymbol{\sigma}_0 \,+\, c_0\,\mathbb{I}$, where $\boldsymbol{\sigma}_0 \in \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$ and, according to the expression for p in (3.19), there holds

$$c_0 \,:=\, \frac{1}{n\,|\Omega|} \int_\Omega \mathrm{tr}(\boldsymbol{\sigma}) \,=\, -\frac{1}{|\Omega|} \int_\Omega p \,-\, \frac{\rho}{n\,|\Omega|} \int_\Omega \mathrm{tr}(\mathbf{u} \otimes \mathbf{u})\,, \tag{3.32}$$

which means that, given p, the constant c_0 can be computed once the velocity is known. Thus, it only remains to find $\boldsymbol{\sigma}_0$, which can be placed instead of $\boldsymbol{\sigma}$ in (3.25), (3.28), and (3.29) without altering the validity of these equations. Moreover, it is easy to see that for each $\boldsymbol{\tau} \in \mathrm{R}\,\mathbb{I}$ both sides of (3.21) vanish, in particular the right one because of the compatibility condition (3.20), and hence testing (3.21) against $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_{4/3};\Omega)$ is equivalent to doing it against $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$. Consequently, redenoting from now on $\boldsymbol{\sigma}_0$ as simply $\boldsymbol{\sigma} \in \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$, and suitably gathering (3.21), (3.25), (3.28), and (3.29), we deduce the following mixed variational formulation of (3.19): Given $p \in \mathrm{L}^2(\Omega)$, find $(\mathbf{D},\boldsymbol{\sigma},\mathbf{u},\boldsymbol{\gamma}) \in \mathbb{L}^2_{\mathtt{tr}}(\Omega) \times \mathbb{H}_0(\mathbf{div}_{4/3};\Omega) \times \mathbf{L}^4(\Omega) \times \mathbb{L}^2_{\mathtt{sk}}(\Omega)$ such that

$$\begin{array}{llcl}
\displaystyle\int_\Omega \eta(p,|\mathbf{D}|)\,\mathbf{D}:\mathbf{E} & \displaystyle -\int_\Omega \boldsymbol{\sigma}:\mathbf{E} - \rho\int_\Omega (\mathbf{u}\otimes\mathbf{u}):\mathbf{E} & = & 0\,,\\
\displaystyle -\int_\Omega \boldsymbol{\tau}:\mathbf{D} & \displaystyle -\int_\Omega \mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau}) - \int_\Omega \boldsymbol{\tau}:\boldsymbol{\gamma} & = & -\langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \mathbf{u}_D\rangle\,,\\
 & \displaystyle -\int_\Omega \mathbf{v}\cdot\mathbf{div}(\boldsymbol{\sigma}) - \int_\Omega \boldsymbol{\sigma}:\boldsymbol{\xi} & = & \displaystyle\int_\Omega \mathbf{f}\cdot\mathbf{v}\,,
\end{array} \tag{3.33}$$

for all $(\mathbf{E},\boldsymbol{\tau},\mathbf{v},\boldsymbol{\xi}) \in \mathbf{L}^2_{\mathtt{tr}}(\Omega) \times \mathbb{H}_0(\mathbf{div}_{4/3};\Omega) \times \mathbf{L}^4(\Omega) \times \mathbb{L}^2_{\mathtt{sk}}(\Omega)$. Next, in order to emphasize the particular structure of (3.33), we set the spaces

$$\mathcal{H}_1 := \mathbb{L}^2_{\mathtt{tr}}(\Omega)\,, \quad \mathcal{H}_2 := \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)\,, \quad \text{and} \quad \mathcal{Q} := \mathbf{L}^4(\Omega) \times \mathbb{L}^2_{\mathtt{sk}}(\Omega)\,, \tag{3.34}$$

which are endowed with the norms

$$\|\mathbf{E}\|_{\mathcal{H}_1} := \|\mathbf{E}\|_{0,\Omega}\,, \quad \|\boldsymbol{\tau}\|_{\mathcal{H}_2} := \|\boldsymbol{\tau}\|_{\mathbf{div}_{4/3};\Omega}\,, \quad \text{and} \quad \|(\mathbf{v},\boldsymbol{\xi})\|_{\mathcal{Q}} := \|\mathbf{v}\|_{0,4;\Omega} + \|\boldsymbol{\xi}\|_{0,\Omega}\,,$$

respectively, and introduce the notations

$$\vec{\mathbf{u}} := (\mathbf{u},\boldsymbol{\gamma})\,, \quad \vec{\mathbf{v}} := (\mathbf{v},\boldsymbol{\xi}) \in \mathcal{Q}\,.$$

Then, denoting from now on by $[\,\cdot,\cdot\,]$ the duality pairing between X' and X for any Banach space X, the system (3.33) can be rewritten as: Given $p \in \mathrm{L}^2(\Omega)$, find $(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}}) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ such that

$$\begin{array}{lllcll}
[\mathcal{A}_p(\mathbf{D}),\mathbf{E}] & +\mathcal{B}_1(\mathbf{E},\boldsymbol{\sigma}) & & = & \mathcal{F}_{\mathbf{u}}(\mathbf{E}) & \forall\,\mathbf{E} \in \mathcal{H}_1\,,\\
\mathcal{B}_1(\mathbf{D},\boldsymbol{\tau}) & & +\mathcal{B}(\boldsymbol{\tau},\vec{\mathbf{u}}) & = & \mathcal{G}(\boldsymbol{\tau}) & \forall\,\boldsymbol{\tau} \in \mathcal{H}_2\,,\\
 & \mathcal{B}(\boldsymbol{\sigma},\vec{\mathbf{v}}) & & = & \mathcal{F}(\vec{\mathbf{v}}) & \forall\,\vec{\mathbf{v}} \in \mathcal{Q}\,,
\end{array} \tag{3.35}$$

where the nonlinear operator $\mathcal{A}_p : \mathcal{H}_1 \to \mathcal{H}_1'$, the bilinear forms $\mathcal{B}_1 : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathrm{R}$ and $\mathcal{B} : \mathcal{H}_2 \times \mathcal{Q} \to \mathrm{R}$, and the functionals $\mathcal{F}_{\mathbf{z}} : \mathcal{H}_1 \to \mathrm{R}$, for each $\mathbf{z} \in \mathbf{L}^4(\Omega)$, $\mathcal{G} : \mathcal{H}_2 \to \mathrm{R}$, and $\mathcal{F} : \mathcal{Q} \to \mathrm{R}$, are defined by

$$[\mathcal{A}_p(\mathbf{D}), \mathbf{E}] := \int_\Omega \eta(p, |\mathbf{D}|)\, \mathbf{D} : \mathbf{E} \qquad \forall\, \mathbf{D}, \mathbf{E} \in \mathcal{H}_1\,, \tag{3.36}$$

$$\mathcal{B}_1(\mathbf{E}, \boldsymbol{\tau}) := -\int_\Omega \boldsymbol{\tau} : \mathbf{E} \qquad \forall\,(\mathbf{E}, \boldsymbol{\tau}) \in \mathcal{H}_1 \times \mathcal{H}_2\,, \tag{3.37}$$

$$\mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{v}}) := -\int_\Omega \mathbf{v} \cdot \mathbf{div}(\boldsymbol{\tau}) - \int_\Omega \boldsymbol{\tau} : \boldsymbol{\xi} \qquad \forall\,(\boldsymbol{\tau}, \vec{\mathbf{v}}) \in \mathcal{H}_2 \times \mathcal{Q}\,, \tag{3.38}$$

$$\mathcal{F}_{\mathbf{z}}(\mathbf{E}) := \rho \int_\Omega (\mathbf{z} \otimes \mathbf{z}) : \mathbf{E} \qquad \forall\, \mathbf{E} \in \mathcal{H}_1\,, \tag{3.39}$$

$$\mathcal{G}(\boldsymbol{\tau}) := -\langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \mathbf{u}_D \rangle \qquad \forall\, \boldsymbol{\tau} \in \mathcal{H}_2\,, \tag{3.40}$$

and

$$\mathcal{F}(\vec{\mathbf{v}}) := \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad \forall\, \vec{\mathbf{v}} \in \mathcal{Q}\,. \tag{3.41}$$

Note that the upper bound of η (cf. (3.24)) guarantees that $\mathcal{A}_p$ is well-defined in the sense that $\mathcal{A}_p(\mathbf{D}) \in \mathcal{H}_1'$ for all $\mathbf{D} \in \mathcal{H}_1$. In turn, regarding the boundedness properties of the above bilinear forms and linear functionals, we employ the Cauchy-Schwarz and Hölder inequalities, along with (3.27), and the continuity of both the normal trace operator in $\mathbb{H}(\mathbf{div}_{4/3}; \Omega)$ and the injection $\mathbf{i}_4 : \mathbf{H}^1(\Omega) \to \mathbf{L}^4(\Omega)$, to deduce the existence of positive constants, denoted and given as

$$\begin{aligned} \|\mathcal{B}_1\| &:= 1\,, \qquad \|\mathcal{B}\| := 1\,, \qquad \|\mathcal{F}_{\mathbf{z}}\| := \rho\, n^{1/2}\, \|\mathbf{z}\|^2_{0,4;\Omega}\,, \\ \|\mathcal{G}\| &:= \max\Big\{1, \|\mathbf{i}_4\|\Big\}\, \|\mathbf{u}_D\|_{1/2,\Gamma}\,, \quad \text{and} \quad \|\mathcal{F}\| := \|\mathbf{f}\|_{0,4/3;\Omega}\,, \end{aligned} \tag{3.42}$$

such that

$$\begin{aligned} |\mathcal{B}_1(\mathbf{E}, \boldsymbol{\tau})| &\le \|\mathcal{B}_1\|\, \|\mathbf{E}\|_{\mathcal{H}_1}\, \|\boldsymbol{\tau}\|_{\mathcal{H}_2} && \forall\,(\mathbf{E}, \boldsymbol{\tau}) \in \mathcal{H}_1 \times \mathcal{H}_2\,, \\ |\mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{v}})| &\le \|\mathcal{B}\|\, \|\boldsymbol{\tau}\|_{\mathcal{H}_2}\, \|\vec{\mathbf{v}}\|_{\mathcal{Q}} && \forall\,(\boldsymbol{\tau}, \vec{\mathbf{v}}) \in \mathcal{H}_2 \times \mathcal{Q}\,, \\ |\mathcal{F}_{\mathbf{z}}(\mathbf{E})| &\le \|\mathcal{F}_{\mathbf{z}}\|\, \|\mathbf{E}\|_{\mathcal{H}_1} && \forall\, \mathbf{E} \in \mathcal{H}_1\,, \\ |\mathcal{G}(\boldsymbol{\tau})| &\le \|\mathcal{G}\|\, \|\boldsymbol{\tau}\|_{\mathcal{H}_2} && \forall\, \boldsymbol{\tau} \in \mathcal{H}_2\,, \quad \text{and} \\ |\mathcal{F}(\vec{\mathbf{v}})| &\le \|\mathcal{F}\|\, \|\vec{\mathbf{v}}\|_{\mathcal{Q}} && \forall\, \vec{\mathbf{v}} \in \mathcal{Q}\,. \end{aligned} \tag{3.43}$$

We stress here that (3.35) can be seen as a twofold saddle point-type formulation with a nonlinear operator $\mathcal{A}_p$. Furthermore, once this system is solved, and because of its dependence on the given p, we propose to update the pressure unknown according to the expression provided in the last row of (3.19). More precisely, bearing in mind that the stress tensor appearing there is actually $\boldsymbol{\sigma} + c_0\,\mathbb{I}$, with $\boldsymbol{\sigma} \in \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$ being part of the solution of (3.35), and c_0 given by (3.32), we find that the new pressure, say p_N, becomes

$$p_N \,=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma} + \rho\,(\mathbf{u}\otimes\mathbf{u})\Big) \,+\, \frac{1}{|\Omega|}\int_\Omega p \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\mathbf{u}\otimes\mathbf{u})\,.$$

Note from the foregoing equation that p_N, and hence all the subsequent updates of it, keep the same mean value of p, that is $\displaystyle\int_\Omega p_N \,=\, \int_\Omega p$, so that from now on we assume a given positive value, say κ, and define

$$\mathrm{L}^2_\kappa(\Omega) \,:=\, \left\{ q \in \mathrm{L}^2(\Omega) : \quad \int_\Omega q \,=\, \kappa \right\}.$$

In this way, after solving (3.35) with a given $p \in \mathrm{L}^2_\kappa(\Omega)$, we simply define

$$p_N \,=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma} + \rho\,(\mathbf{u}\otimes\mathbf{u})\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\mathbf{u}\otimes\mathbf{u})\,. \tag{3.44}$$

We will go back to the above when introducing below in Section 3.4 a suitable fixed-point approach to analyze the solvability of (3.35).

We end this section by remarking that the variational formulations resulting from other boundary conditions, say, for instance, mixed ones, instead of the no-slip condition for the velocity, are just minor modifications of (3.33) (or (3.35)). Mixed boundary conditions, often called frictional boundary conditions, are fairly frequent in flows of granular materials in frictional walls, where some slip velocity and shear limited by Coulomb friction can occur simultaneously [13, 57]. In fact, letting Γ_D and Γ_N be disjoint parts of Γ, both with non-null measures, such that $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, we consider first:

$$\mathbf{u} \,=\, \mathbf{u}_D \quad \text{on} \quad \Gamma_D\,, \qquad \boldsymbol{\sigma}\,\boldsymbol{\nu} \,=\, \mathbf{0} \quad \text{on} \quad \Gamma_N\,, \tag{3.45}$$

with datum $\mathbf{u}_D \in \mathbf{H}^{1/2}(\Gamma_D)$. Then, given $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega)$ such that $\boldsymbol{\tau}\,\boldsymbol{\nu}$ is null on Γ_N, it follows that $\boldsymbol{\tau}\,\boldsymbol{\nu}|_{\Gamma_D} \in \mathbf{H}^{-1/2}(\Gamma_D)$ (cf. [58, Lemma 2.4, Remark 2.5] or [59, Section 3.1]), and hence, instead of (3.21), the integration by parts formula (2.3), for which the Dirichlet boundary condition on Γ_D is still natural, yields

$$\int_\Omega \boldsymbol{\tau} : \mathbf{D} \,+\, \int_\Omega \mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau}) \,+\, \int_\Omega \boldsymbol{\tau} : \boldsymbol{\gamma} \,=\, \langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \mathbf{u}_D\rangle_D \qquad \forall\,\boldsymbol{\tau} \in \mathbb{H}_N(\mathbf{div}_t;\Omega)\,, \tag{3.46}$$

where $\langle \cdot, \cdot \rangle_D$ stands for the duality pairing between $\mathbf{H}^{-1/2}(\Gamma_D)$ and $\mathbf{H}^{1/2}(\Gamma_D)$, and

$$\mathbb{H}_N(\mathbf{div}_t;\Omega) \,:=\, \Big\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega) : \quad \boldsymbol{\tau}\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \Gamma_N \Big\}.$$

In this way, the changes of (3.33) are basically the right hand side of its second row, and the space where the unknown $\boldsymbol{\sigma}$ is sought, which, because of the Neumann boundary condition on Γ_N, becomes the same of its associated test function $\boldsymbol{\tau}$, that is $\mathbb{H}_N(\mathbf{div}_t;\Omega)$. Secondly, and exchanging the null condition in (3.45), we can also look at:

$$\mathbf{u} \,=\, \mathbf{0} \quad \text{on} \quad \Gamma_D\,, \qquad \boldsymbol{\sigma}\,\boldsymbol{\nu} \,=\, \mathbf{g}_N \quad \text{on} \quad \Gamma_N\,, \tag{3.47}$$

with datum $\mathbf{g}_N \,\in\, \mathbf{H}_{00}^{-1/2}(\Gamma_N)$. Proceeding similarly as above, but introducing the auxiliary unknown $\boldsymbol{\varphi} \,:=\, -\mathbf{u}|_{\Gamma_N} \in \mathbf{H}_{00}^{1/2}(\Gamma_N)$, we now arrive at

$$\int_\Omega \boldsymbol{\tau} : \mathbf{D} \,+\, \int_\Omega \mathbf{u}\cdot \mathbf{div}(\boldsymbol{\tau}) \,+\, \int_\Omega \boldsymbol{\tau} : \boldsymbol{\gamma} \,+\, \langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \boldsymbol{\varphi}\rangle_N \,=\, 0 \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_t;\Omega)\,, \tag{3.48}$$

where $\langle \cdot, \cdot \rangle_N$ stands for the duality pairing between $\mathbf{H}_{00}^{-1/2}(\Gamma_N)$ and $\mathbf{H}_{00}^{1/2}(\Gamma_N)$. In addition, being the Neumann boundary condition on Γ_N essential, we impose it weakly as

$$\langle \boldsymbol{\sigma}\,\boldsymbol{\nu}, \boldsymbol{\psi}\rangle_N \,=\, \langle \mathbf{g}_N, \boldsymbol{\psi}\rangle_N \qquad \forall\, \boldsymbol{\psi} \in \mathbf{H}_{00}^{1/2}(\Gamma_N)\,. \tag{3.49}$$

Consequently, the extra terms given by $\langle \boldsymbol{\tau}\,\boldsymbol{\nu}, \boldsymbol{\varphi}\rangle_N$ (cf. (3.48)) and those from (3.49), are incorporated into the second and third rows, respectively, of (3.33) (equivalently, (3.35)), thus yielding the space $\mathcal{Q}$, the bilinear form $\mathcal{B}$, and the functional $\mathcal{F}$ to be slightly modified. Finally, we could also deal with the more general case of mixed boundary conditions, namely:

$$\mathbf{u} \,=\, \mathbf{u}_D \quad \text{on} \quad \Gamma_D\,, \qquad \boldsymbol{\sigma}\,\boldsymbol{\nu} \,=\, \mathbf{g}_N \quad \text{on} \quad \Gamma_N\,, \tag{3.50}$$

with data $\mathbf{u}_D \in \mathbf{H}^{1/2}(\Gamma_D)$ and $\mathbf{g}_N \in \mathbf{H}_{00}^{-1/2}(\Gamma_N)$, for which the direct sum decompositions of $\mathbf{H}^{1/2}(\Gamma)$ and its dual $\mathbf{H}^{-1/2}(\Gamma)$ provided in [58, Lemma 2.2] (see, also [59, Section 3.1]), should be employed when applying the integration by parts formula. Alternatively, one could also resort to suitable trace liftings to reduce (3.50) to either (3.45) or (3.47). We omit further details and just stress that, for any of the above described situations, the corresponding continuous and discrete analyses will follow very closely the ones to be developed in what follows.

3.4 The continuous solvability analysis

In this section we employ a fixed-point approach along with an abstract result on the well-posedness of the aforementioned type of nonlinear operator equations in Banach spaces, to analyze the solvability of the mixed variational formulation (3.35).

3.4.1 The fixed point strategy

We begin by introducing the operator $\mathbf{T} : \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega) \longrightarrow \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega)$ defined as

$$\mathbf{T}(\mathbf{z}, r) \,:=\, (\mathbf{u}, p) \qquad \forall\, (\mathbf{z}, r) \,\in\, \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega)\,, \tag{3.51}$$

where $(\mathbf{D}, \boldsymbol{\sigma}, \vec{\mathbf{u}}) \;:=\; \big(\mathbf{D}, \boldsymbol{\sigma}, (\mathbf{u}, \gamma)\big) \,\in\, \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ is the unique solution (to be confirmed later on) of the problem arising from (3.35) when $\mathcal{A}_p$ and the functional $\mathcal{F}_\mathbf{u}$ are replaced by $\mathcal{A}_r$ and $\mathcal{F}_\mathbf{z}$, respectively, that is

$$\begin{array}{llcll}
[\mathcal{A}_r(\mathbf{D}), \mathbf{E}] & + \mathcal{B}_1(\mathbf{E}, \boldsymbol{\sigma}) & = & \mathcal{F}_\mathbf{z}(\mathbf{E}) & \forall\, \mathbf{E} \in \mathbb{H}_1\,, \\
\mathcal{B}_1(\mathbf{D}, \boldsymbol{\tau}) & \qquad\qquad + \mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{u}}) & = & \mathcal{G}(\boldsymbol{\tau}) & \forall\, \boldsymbol{\tau} \in \mathbb{H}_2\,, \\
 & \mathcal{B}(\boldsymbol{\sigma}, \vec{\mathbf{v}}) & = & \mathcal{F}(\vec{\mathbf{v}}) & \forall\, \vec{\mathbf{v}} \in \mathbb{Q}\,,
\end{array} \tag{3.52}$$

and p is computed according to (3.44), that is

$$p \,:=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma} + \rho\,(\mathbf{u} \otimes \mathbf{u})\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|} \int_\Omega \mathrm{tr}(\mathbf{u} \otimes \mathbf{u})\,. \tag{3.53}$$

Then, it is readily seen that solving (3.35) is equivalent to finding a fixed point of $\mathbf{T}$, that is $(\mathbf{u}, p) \in \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega)$ such that

$$\mathbf{T}(\mathbf{u}, p) \;=\; (\mathbf{u}, p)\,. \tag{3.54}$$

3.4.2 Well-definedness of the fixed point operator

In this section we prove that the operator $\mathbf{T}$ (cf. (3.51) - (3.52)) is well-defined, for which we make use of the following abstract result establishing sufficient conditions for the well-posedness of a class of twofold saddle point operator equations.

Theorem 3.4.1. Let $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{Y}$ be reflexive and separable Banach spaces, and let $\mathbf{A} : \mathbf{X}_1 \to \mathbf{X}_1'$ be a nonlinear operator, and $\mathbf{B}_1 : \mathbf{X}_1 \times \mathbf{X}_2 \to \mathrm{R}$ and $\mathbf{B} : \mathbf{X}_2 \times \mathbf{Y} \to \mathrm{R}$ be bounded bilinear forms. In addition, let $\mathbf{V}$ be the null space of the operator induced by $\mathbf{B}$, and assume that

i) $\mathbf{A}$ is Lipschitz-continuous, that is there exists a positive constant $L_{\mathbf{A}}$ such that

$$\|\mathbf{A}(\mathbf{r}) - \mathbf{A}(\mathbf{s})\|_{\mathbf{X}_1'} \,\le\, L_{\mathbf{A}}\,\|\mathbf{r}-\mathbf{s}\|_{\mathbf{X}_1} \qquad \forall\,\mathbf{r},\,\mathbf{s}\in\mathbf{X}_1\,,$$

ii) the family of operators $\big\{\mathbf{A}(\mathbf{t}+\cdot)\big\}_{\mathbf{t}\in\mathbf{X}_1}$ is uniformly strongly monotone, that is there exists a positive constant $\alpha_{\mathbf{A}}$ such that

$$[\mathbf{A}(\mathbf{t}+\mathbf{r}) - \mathbf{A}(\mathbf{t}+\mathbf{s}), \mathbf{r}-\mathbf{s}] \,\ge\, \alpha_{\mathbf{A}}\,\|\mathbf{r}-\mathbf{s}\|^2_{\mathbf{X}_1} \qquad \forall\,\mathbf{t},\,\mathbf{r},\,\mathbf{s}\in\mathbf{X}_1\,,$$

iii there exists a positive constant β such that

$$\sup_{\substack{\tau\in\mathbf{X}_2\\ \tau\neq 0}} \frac{\mathbf{B}(\tau,v)}{\|\tau\|_{\mathbf{X}_2}} \,\ge\, \beta\,\|v\|_{\mathbf{Y}} \qquad \forall\,v\in\mathbf{Y}\,,$$

iv) and there exists a positive constant β_1 such that

$$\sup_{\substack{\mathbf{r}\in\mathbf{X}_1\\ \mathbf{r}\neq 0}} \frac{\mathbf{B}_1(\mathbf{r},\tau)}{\|\mathbf{r}\|_{\mathbf{X}_1}} \,\ge\, \beta_1\,\|\tau\|_{\mathbf{X}_2} \qquad \forall\,\tau\in\mathbf{V}\,.$$

Then, for each $(\mathbf{F}_1,\mathbf{F}_2,\mathbf{G})\in\mathbf{X}_1'\times\mathbf{X}_2'\times\mathbf{Y}'$ there exists a unique $(\mathbf{t},\sigma,u)\in\mathbf{X}_1\times\mathbf{X}_2\times\mathbf{Y}$ such that

$$\begin{array}{lllll} [\mathbf{A}(\mathbf{t}),\mathbf{s}] & +\,\mathbf{B}_1(\mathbf{s},\sigma) & & = & \mathbf{F}_1(\mathbf{s}) & \forall\,\mathbf{s}\in\mathbf{X}_1\,, \\ \mathbf{B}_1(\mathbf{t},\tau) & & +\,\mathbf{B}(\tau,u) & = & \mathbf{F}_2(\tau) & \forall\,\tau\in\mathbf{X}_2\,, \\ & \mathbf{B}(\sigma,v) & & = & \mathbf{G}(v) & \forall\,v\in\mathbf{Y}\,. \end{array} \tag{3.55}$$

Moreover, there exists a positive constant C, depending only on $L_{\mathbf{A}}$, $\alpha_{\mathbf{A}}$, β, β_1, and the boundedness constant of $\mathbf{B}_1$, say $\|\mathbf{B}_1\|$, such that

$$\|(\mathbf{t},\boldsymbol{\sigma},\mathbf{u})\|_{\mathbf{X}_1\times\mathbf{X}_2\times\mathbf{Y}} \,\le\, C\Big\{\|\mathbf{F}_1\|_{\mathbf{X}_1'} + \|\mathbf{F}_2\|_{\mathbf{X}_2'} + \|\mathbf{G}\|_{\mathbf{Y}'} + \|\mathbf{A}(0)\|_{\mathbf{X}_1'}\Big\}\,. \tag{3.56}$$

Proof. It is a particular case of [60, Theorem 3.4]. □

As already announced, we plan to apply Theorem 3.4.1 to conclude the well-posedness of (3.52), for which we proceed next to show that the respective hypotheses are satisfied. In particular, for those involving $\mathcal{A}_r$, we need to incorporate additional assumptions on the function η, namely

(**H.2**) with the same positive constants η_1 and η_2 from (**H.1**), there holds

$$0 < \eta_1 \le \eta(\varrho,\omega) + \omega \frac{\partial}{\partial\omega}\eta(\varrho,\omega) \le \eta_2 \qquad \forall\,(\varrho,\omega)\in \mathrm{R}^+\times\mathrm{R}^+\,, \quad \text{and} \tag{3.57}$$

(**H.3**) there exists a positive constant L_η such that

$$\Big|\eta(\varrho,\omega) - \eta(\chi,\omega)\Big|\,\omega \le L_\eta\,|\varrho-\chi| \qquad \forall\,\varrho,\,\chi,\,\omega\in\mathrm{R}^+. \tag{3.58}$$

In the Appendix A.1 we prove that η, as defined by (3.11), satisfies (**H.3**) and that, under a suitable modification of its domain, it accomplishes (**H.1**) and (**H.2**) as well.

Then, we can prove the following lemma establishing continuity and strong-monotonicity properties of the nonlinear operator $\mathcal{A}_r$.

Lemma 3.4.1. Let $L_{\mathcal{A}} := 2\eta_2 - \eta_1$ and $\alpha_{\mathcal{A}} := \eta_1$. Then, there holds

$$\|\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_r(\mathbf{E})\|_{\mathcal{H}_1'} \le L_{\mathcal{A}}\,\|\mathbf{D}-\mathbf{E}\|_{\mathcal{H}_1} \qquad \forall\, r\in\mathrm{L}^2(\Omega)\,, \quad \forall\,\mathbf{D},\,\mathbf{E}\in\mathcal{H}_1\,, \tag{3.59}$$

$$[\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_r(\mathbf{E}), \mathbf{D}-\mathbf{E}] \ge \alpha_{\mathcal{A}}\,\|\mathbf{D}-\mathbf{E}\|^2_{\mathcal{H}_1} \qquad \forall\, r\in\mathrm{L}^2(\Omega)\,, \quad \forall\,\mathbf{D},\,\mathbf{E}\in\mathcal{H}_1\,, \tag{3.60}$$

and

$$\Big|[\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_q(\mathbf{D}), \mathbf{E}]\Big| \le L_\eta\,\|r-q\|_{0,\Omega}\,\|\mathbf{E}\|_{\mathcal{H}_1} \qquad \forall\, r,\,q\in\mathrm{L}^2(\Omega)\,, \quad \forall\,\mathbf{D},\,\mathbf{E}\in\mathcal{H}_1\,. \tag{3.61}$$

Proof. For the proofs of (3.59) and (3.60) we refer to [61, Theorem 3.8]. In turn, given $r,\,q\in\mathrm{L}^2(\Omega)$, and $\mathbf{D},\,\mathbf{E}\in\mathcal{H}_1$, bearing in mind the definition of $\mathcal{A}_r$ (cf. (3.36)), and using (3.58) with $\varrho = r$, $\chi = q$, and $\omega = |\mathbf{D}|$, we deduce that

$$\begin{aligned}\Big|[\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_q(\mathbf{D}), \mathbf{E}]\Big| &= \left|\int_\Omega \Big\{\eta(r,|\mathbf{D}|) - \eta(q,|\mathbf{D}|)\Big\}\mathbf{D}:\mathbf{E}\right| \\ &\le \int_\Omega \Big|\eta(r,|\mathbf{D}|) - \eta(q,|\mathbf{D}|)\Big|\,|\mathbf{D}|\,|\mathbf{E}| \le L_\eta \int_\Omega |r-q|\,|\mathbf{E}|\,,\end{aligned}$$

from which, applying Cauchy-Schwarz's inequality, we obtain (3.61) and end the proof. □

We now observe from (3.59) and (3.60) that, for each $r\in\mathrm{L}^2(\Omega)$, $\mathcal{A}_r$ verifies the hypotheses i) and ii) of Theorem 3.4.1 with the constants $L_{\mathcal{A}}$ and $\alpha_{\mathcal{A}}$ (cf. Lemma

3.4.1), respectively. In particular, for ii) we simply notice that there holds

$$\begin{aligned} [\mathcal{A}_r(\mathbf{J}+\mathbf{D}) - \mathcal{A}_r(\mathbf{J}+\mathbf{E}), \mathbf{D}-\mathbf{E}] &= [\mathcal{A}_r(\mathbf{J}+\mathbf{D}) - \mathcal{A}_r(\mathbf{J}+\mathbf{E}), (\mathbf{D}+\mathbf{J}) - (\mathbf{E}+\mathbf{J})] \\ \geq \alpha_{\mathcal{A}} \|(\mathbf{D}+\mathbf{J}) - (\mathbf{E}+\mathbf{J})\|^2_{\mathcal{H}_1} &= \alpha_{\mathcal{A}} \|\mathbf{D}-\mathbf{E}\|^2_{\mathcal{H}_1} \qquad \forall\, \mathbf{J},\, \mathbf{D},\, \mathbf{E} \in \mathcal{H}_1 \,. \end{aligned}$$

Next, we recall from [22] the following lemma establishing the continuous inf-sup condition for $\mathcal{B}$.

Lemma 3.4.2. There exists a positive constant $\widetilde{\beta}$ such that

$$\sup_{\substack{\boldsymbol{\tau}\in\mathcal{H}_2 \\ \boldsymbol{\tau}\neq 0}} \frac{\mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{v}})}{\|\boldsymbol{\tau}\|_{\mathcal{H}_2}} \geq \widetilde{\beta}\, \|\vec{\mathbf{v}}\|_{\mathcal{Q}} = \widetilde{\beta} \Big\{ \|\mathbf{v}\|_{0,4;\Omega} + \|\boldsymbol{\xi}\|_{0,\Omega} \Big\} \qquad \forall\, \vec{\mathbf{v}} := (\mathbf{v}, \boldsymbol{\xi}) \in \mathcal{Q}\,. \tag{3.62}$$

Proof. See [22, Lemma 3.5] for details. □

Regarding the continuous inf-sup condition for $\mathcal{B}_1$, we first observe from the definition of $\mathcal{B}$ (cf. (3.38)) that the null space of its induced operator is given by

$$\mathcal{V} := \Big\{ \boldsymbol{\tau} \in \mathcal{H}_2 : \quad \mathbf{div}(\boldsymbol{\tau}) = 0 \quad \text{and} \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathrm{t}} \quad \text{in} \quad \Omega \Big\}.$$

Then, we recall from [62] the following result.

Lemma 3.4.3. There exists a positive constant $\widetilde{\beta}_1$ such that

$$\sup_{\substack{\mathbf{E}\in\mathcal{H}_1 \\ \mathbf{E}\neq 0}} \frac{\mathcal{B}_1(\mathbf{E}, \boldsymbol{\tau})}{\|\mathbf{E}\|_{\mathcal{H}_1}} \geq \widetilde{\beta}_1\, \|\boldsymbol{\tau}\|_{\mathcal{H}_2} \qquad \forall\, \boldsymbol{\tau} \in \mathcal{V}\,. \tag{3.63}$$

Proof. See [62, Lemma 3.3] for details. □

We remark here that the proof of Lemma 3.4.3 makes use of the inequality establishing the existence of a positive constant c_1 such that

$$c_1\, \|\boldsymbol{\tau}\|_{0,\Omega} \leq \|\boldsymbol{\tau}^{\mathrm{d}}\|_{0,\Omega} + \|\mathbf{div}(\boldsymbol{\tau})\|_{0,4/3;\Omega} \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)\,. \tag{3.64}$$

The well-posedness of (3.52), equivalently the well-definedness of $\mathbf{T}$, is stated now as follows.

Theorem 3.4.2. For each $(\mathbf{z}, r) \in \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega)$ there exists a unique tuple $(\mathbf{D}, \boldsymbol{\sigma}, \vec{\mathbf{u}}) := \big(\mathbf{D}, \boldsymbol{\sigma}, (\mathbf{u}, \boldsymbol{\gamma})\big) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ solution to (3.52), and hence one can define $\mathbf{T}(\mathbf{z}, r) := (\mathbf{u}, p) \in \mathbf{L}^4(\Omega) \times \mathrm{L}^2_\kappa(\Omega)$, where p is computed according to (3.53).

Moreover, there exists a positive constant $C_{\mathbf{T}}$, depending only on $L_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, $\widetilde{\beta}$, $\widetilde{\beta}_1$, n, and $\|\mathbf{i}_4\|$, such that

$$\|\mathbf{u}\|_{0,4;\Omega} \,\le\, \|(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}})\|_{\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}} \,\le\, C_{\mathbf{T}}\left\{\rho\,\|\mathbf{z}\|^2_{0,4;\Omega} \,+\, \|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega}\right\}. \tag{3.65}$$

Proof. Having already checked that (3.52) verifies the assumptions i) and ii) of Theorem 3.4.1, and noting that Lemmas 3.4.2 and 3.4.3 confirm that ii) and iv) also hold, the proof is a straightforward application of that abstract result. In particular, the a priori estimate (3.65) follows from (3.56), the boundedness properties of the functionals involved (cf. (3.42), (3.43)), and the fact that $\mathcal{A}_p(\mathbf{0}) \,= \mathbf{0} \in \mathcal{H}_1'$. Regarding $C_{\mathbf{T}}$, note that we omit its dependence on $\|\mathcal{B}_1\|$ since this latter value equals 1 (cf. (3.42)). □

3.4.3 Solvability analysis of the fixed point equation

Knowing that $\mathbf{T}$ is well-defined, we now address the solvability of the fixed-point equation (3.54). We begin the analysis deriving sufficient conditions on $\mathbf{T}$ to map a complete metric subspace of $\mathbf{L}^4(\Omega\times \mathrm{L}^2_\kappa(\Omega)$ into itself. Indeed, given $\delta>0$, we set

$$\mathrm{W}(\delta) \,:=\, \left\{\mathbf{z}\in\mathbf{L}^4(\Omega):\quad \|\mathbf{z}\|_{0,4;\Omega}\,\le\,\delta\right\} \quad\text{and}\quad \mathrm{S}(\delta)\,:=\,\mathrm{W}(\delta)\times\mathrm{L}^2_\kappa(\Omega)\,. \tag{3.66}$$

Then, proceeding as in [56, Lemma 4.7], we are able to prove the following result.

Lemma 3.4.4. Assume that

$$\rho\,\delta\,\le\,\frac{1}{2\,C_{\mathbf{T}}}\quad\text{and}\quad C_{\mathbf{T}}\left\{\|\mathbf{u}_D\|_{1/2,\Gamma}\,+\,\|\mathbf{f}\|_{0,4/3;\Omega}\right\}\,\le\,\frac{\delta}{2}\,. \tag{3.67}$$

Then, $\mathbf{T}\big(\mathrm{S}(\delta)\big)\,\subseteq\,\mathrm{S}(\delta)$.

Proof. Given $(\mathbf{z},r)\in\mathrm{S}(\delta)$, we know from Theorem 3.4.2 that $\mathbf{T}(\mathbf{z},r):=(\mathbf{u},p)$ is well-defined and that, in virtue of (3.65) and the assumptions from (3.67), there holds

$$\|\mathbf{u}\|_{0,4;\Omega}\,\le\,C_{\mathbf{T}}\left\{\rho\,\|\mathbf{z}\|^2_{0,4;\Omega}\,+\,\|\mathbf{u}_D\|_{1/2,\Gamma}\,+\,\|\mathbf{f}\|_{0,4/3;\Omega}\right\}\,\le\,C_{\mathbf{T}}\,\rho\,\delta^2+\frac{\delta}{2}\,\le\,\delta\,,$$

whereas (3.53) guarantees that $p\in\mathrm{L}^2_\kappa(\Omega)$, and hence $(\mathbf{u},p)\in\mathrm{S}(\delta)$. □

The continuity property of $\mathbf{T}$ is established next.

Lemma 3.4.5. Under the same assumption of Lemma 3.4.4, that is (3.67), there exist positive constants $L_j(\mathbf{T})$, $j\in\big\{1,2\big\}$, depending only on $L_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, $\widetilde{\beta}$, $\widetilde{\beta}_1$, n, and $\|\mathbf{i}_4\|$, such that

$$\|\mathbf{T}(\mathbf{z},r)-\mathbf{T}(\utilde{\mathbf{z}},\utilde{r})\| \,\le\, L_1(\mathbf{T})\,\rho\,\delta\,\|\mathbf{z}-\utilde{\mathbf{z}}\|_{0,4;\Omega} \,+\, L_2(\mathbf{T})\,L_\eta\,\|r-\utilde{r}\|_{0,\Omega} \tag{3.68}$$

for all $(\mathbf{z},r),\ (\utilde{\mathbf{z}},\utilde{r})\in \mathrm{S}(\delta)$.

Proof. Given $(\mathbf{z},r)$, $(\utilde{\mathbf{z}},\utilde{r})\in\mathrm{S}(\delta)$, we let

$$\mathbf{T}(\mathbf{z},r) \,:=\, (\mathbf{u},p) \quad\text{and}\quad \mathbf{T}(\utilde{\mathbf{z}},\utilde{r}) \,:=\, (\utilde{\mathbf{u}},\utilde{p})\,, \tag{3.69}$$

where $(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}}) = \big(\mathbf{D},\boldsymbol{\sigma},(\mathbf{u},\boldsymbol{\gamma})\big)\in\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}$ is the unique solution of (3.52) and p is defined by (3.53), and, analogously, $(\utilde{\mathbf{D}},\utilde{\boldsymbol{\sigma}},\utilde{\vec{\mathbf{u}}}) = \big(\utilde{\mathbf{D}},\utilde{\boldsymbol{\sigma}},(\utilde{\mathbf{u}},\utilde{\boldsymbol{\gamma}})\big)\in\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}$ is the unique solution of (3.52) with $\mathcal{A}_{\utilde{r}}$ and $\mathcal{F}_{\utilde{\mathbf{z}}}$ instead of $\mathcal{A}_r$ and $\mathcal{F}_{\mathbf{z}}$, respectively, and, following (3.53),

$$\utilde{p} \,:=\, -\frac{1}{n}\,\mathrm{tr}\Big(\utilde{\boldsymbol{\sigma}}+\rho\,(\utilde{\mathbf{u}}\otimes\utilde{\mathbf{u}})\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\utilde{\mathbf{u}}\otimes\utilde{\mathbf{u}})\,. \tag{3.70}$$

Then, subtracting from each other the aforementioned systems (3.52) whose solutions are $(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}})$ and $(\utilde{\mathbf{D}},\utilde{\boldsymbol{\sigma}},\utilde{\vec{\mathbf{u}}})$, we obtain

$$\begin{array}{llcll}
[\mathcal{A}_r(\mathbf{D})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{E}] & +\mathcal{B}_1(\mathbf{E},\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}}) & = & \Big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\Big)(\mathbf{E}) & \forall\,\mathbf{E}\in\mathbb{H}_1\,,\\
\mathcal{B}_1(\mathbf{D}-\utilde{\mathbf{D}},\boldsymbol{\tau}) & +\mathcal{B}(\boldsymbol{\tau},\vec{\mathbf{u}}-\utilde{\vec{\mathbf{u}}}) & = & 0 & \forall\,\boldsymbol{\tau}\in\mathbb{H}_2\,,\\
 & \mathcal{B}(\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}},\vec{\mathbf{v}}) & = & 0 & \forall\,\vec{\mathbf{v}}\in\mathbb{Q}\,.
\end{array} \tag{3.71}$$

Next, taking $\boldsymbol{\tau}=\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}}$, we get from the second and third rows of the foregoing equation that

$$\mathcal{B}_1(\mathbf{D}-\utilde{\mathbf{D}},\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}}) \,=\, -\mathcal{B}(\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}},\vec{\mathbf{u}}-\utilde{\vec{\mathbf{u}}}) \,=\, 0\,,$$

which, along with the first row applied to $\mathbf{E}=\mathbf{D}-\utilde{\mathbf{D}}$, yields

$$[\mathcal{A}_r(\mathbf{D})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}] \,=\, \Big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\Big)(\mathbf{D}-\utilde{\mathbf{D}})\,.$$

Thus, subtracting and adding $\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}})$, we see that

$$\begin{aligned}
[\mathcal{A}_r(\mathbf{D})-\mathcal{A}_r(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}] &= [\mathcal{A}_r(\mathbf{D})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}]-[\mathcal{A}_r(\utilde{\mathbf{D}})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}]\\
&= \Big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\Big)(\mathbf{D}-\utilde{\mathbf{D}}) \,-\, [\mathcal{A}_r(\utilde{\mathbf{D}})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}]\,,
\end{aligned}$$

so that, using (3.60) and (3.61), we find that

$$\begin{aligned}\alpha_{\mathcal{A}}\,\|\mathbf{D}-\utilde{\mathbf{D}}\|^2_{0,\Omega} \,&\le\, [\mathcal{A}_r(\mathbf{D})-\mathcal{A}_r(\utilde{\mathbf{D}}),\mathbf{D}-\utilde{\mathbf{D}}]\\ &\le\, |\big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\big)(\mathbf{D}-\utilde{\mathbf{D}})| \,+\, L_\eta\,\|r-\utilde{r}\|_{0,\Omega}\,\|\mathbf{D}-\utilde{\mathbf{D}}\|_{0,\Omega}\,.\end{aligned}\tag{3.72}$$

In turn, it is clear from (3.39) that

$$\big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\big)(\mathbf{D}-\utilde{\mathbf{D}}) \,=\, \rho\int_\Omega\Big((\mathbf{z}\otimes\mathbf{z})-(\utilde{\mathbf{z}}\otimes\utilde{\mathbf{z}})\Big):(\mathbf{D}-\utilde{\mathbf{D}})\,,\tag{3.73}$$

from which, subtracting and adding $\utilde{\mathbf{z}}$ to one of the factors of $(\mathbf{z}\otimes\mathbf{z})$, and using Cauchy-Schwarz's inequality, (3.26), and the fact that $\mathbf{z}$, $\utilde{\mathbf{z}}\in \mathrm{W}(\delta)$, we readily deduce that

$$\begin{aligned}|\big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\big)(\mathbf{D}-\utilde{\mathbf{D}})| \,&\le\, n^{1/2}\,\rho\,\Big(\|\mathbf{z}\|_{0,4;\Omega}+\|\utilde{\mathbf{z}}\|_{0,4;\Omega}\Big)\,\|\mathbf{z}-\utilde{\mathbf{z}}\|_{0,4;\Omega}\,\|\mathbf{D}-\utilde{\mathbf{D}}\|_{0,\Omega}\\ &\le\, 2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z}-\utilde{\mathbf{z}}\|_{0,4;\Omega}\,\|\mathbf{D}-\utilde{\mathbf{D}}\|_{0,\Omega}\,.\end{aligned}\tag{3.74}$$

In this way, employing (3.74) in (3.72), we arrive at

$$\|\mathbf{D}-\utilde{\mathbf{D}}\|_{0,\Omega} \,\le\, \alpha_{\mathcal{A}}^{-1}\Big\{2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z}-\utilde{\mathbf{z}}\|_{0,4;\Omega} \,+\, L_\eta\,\|r-\utilde{r}\|_{0,\Omega}\Big\}\,.\tag{3.75}$$

On the other hand, using the continuous inf-sup condition for $\mathcal{B}$ (cf. (3.62)) and the second row of (3.71), we get

$$\widetilde{\beta}\,\|\vec{\mathbf{u}}-\utilde{\vec{\mathbf{u}}}\|_{\mathcal{Q}} \,\le\, \sup_{\substack{\boldsymbol{\tau}\in\mathcal{H}_2\\ \boldsymbol{\tau}\neq 0}}\frac{\mathcal{B}(\boldsymbol{\tau},\vec{\mathbf{u}}-\utilde{\vec{\mathbf{u}}})}{\|\boldsymbol{\tau}\|_{\mathcal{H}_2}} \,=\, \sup_{\substack{\boldsymbol{\tau}\in\mathcal{H}_2\\ \boldsymbol{\tau}\neq 0}}\frac{-\mathcal{B}_1(\mathbf{D}-\utilde{\mathbf{D}},\boldsymbol{\tau})}{\|\boldsymbol{\tau}\|_{\mathcal{H}_2}} \,\le\, \|\mathbf{D}-\utilde{\mathbf{D}}\|_{0,\Omega}\,,$$

which, along with (3.75), implies

$$\|\vec{\mathbf{u}}-\utilde{\vec{\mathbf{u}}}\|_{\mathcal{Q}} \,\le\, \alpha_{\mathcal{A}}^{-1}\,\widetilde{\beta}^{-1}\Big\{2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z}-\utilde{\mathbf{z}}\|_{0,4;\Omega} \,+\, L_\eta\,\|r-\utilde{r}\|_{0,\Omega}\Big\}\,.\tag{3.76}$$

Next, noting from the third row of (3.71) that $\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}}$ belongs to $\mathcal{V}:=N(\mathcal{B})$, we have from the continuous inf-sup condition for $\mathcal{B}_1$ (cf. (3.63)) and the first row of (3.71), that

$$\widetilde{\beta}_1\,\|\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}}\|_{\mathcal{H}_2} \,\le\, \sup_{\substack{\mathbf{E}\in\mathcal{H}_1\\ \mathbf{E}\neq 0}}\frac{\mathcal{B}_1(\mathbf{E},\boldsymbol{\sigma}-\utilde{\boldsymbol{\sigma}})}{\|\mathbf{E}\|_{\mathcal{H}_1}} \,=\, \sup_{\substack{\mathbf{E}\in\mathcal{H}_1\\ \mathbf{E}\neq 0}}\frac{\big(\mathcal{F}_{\mathbf{z}}-\mathcal{F}_{\utilde{\mathbf{z}}}\big)(\mathbf{E})-[\mathcal{A}_r(\mathbf{D})-\mathcal{A}_{\utilde{r}}(\utilde{\mathbf{D}}),\mathbf{E}]}{\|\mathbf{E}\|_{\mathcal{H}_1}}\,.\tag{3.77}$$

Then, exactly as for the derivation of (3.74), we deduce that

$$\big|\Big(\mathcal{F}_{\mathbf{z}} - \mathcal{F}_{\underset{\sim}{\mathbf{z}}}\Big)(\mathbf{E})\big| \,\le\, 2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z} - \underset{\sim}{\mathbf{z}}\|_{0,4;\Omega}\,\|\mathbf{E}\|_{0,\Omega}\,. \tag{3.78}$$

In turn, similarly as previously done in the present proof, it is easily seen that

$$[\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_{\underset{\sim}{r}}(\underset{\sim}{\mathbf{D}}), \mathbf{E}] \,=\, [\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_r(\underset{\sim}{\mathbf{D}}), \mathbf{E}] \,+\, [\mathcal{A}_r(\underset{\sim}{\mathbf{D}}) - \mathcal{A}_{\underset{\sim}{r}}(\underset{\sim}{\mathbf{D}}), \mathbf{E}]\,,$$

from which, employing (3.59) and (3.61), it follows that

$$\Big|[\mathcal{A}_r(\mathbf{D}) - \mathcal{A}_{\underset{\sim}{r}}(\underset{\sim}{\mathbf{D}}), \mathbf{E}]\Big| \,\le\, \Big\{L_{\mathcal{A}}\,\|\mathbf{D} - \underset{\sim}{\mathbf{D}}\|_{0,\Omega} \,+\, L_\eta\,\|r - \underset{\sim}{r}\|_{0,\Omega}\Big\}\,\|\mathbf{E}\|_{0,\Omega}\,. \tag{3.79}$$

In this way, replacing the estimates (3.78) and (3.79) back into (3.77), we conclude that

$$\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{\mathcal{H}_2} \,\le\, \widetilde{\beta}_1^{-1}\Big\{2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z} - \underset{\sim}{\mathbf{z}}\|_{0,4;\Omega} \,+\, L_{\mathcal{A}}\,\|\mathbf{D} - \underset{\sim}{\mathbf{D}}\|_{0,\Omega} \,+\, L_\eta\,\|r - \underset{\sim}{r}\|_{0,\Omega}\Big\}\,, \tag{3.80}$$

which, combined with the estimate for $\|\mathbf{D} - \underset{\sim}{\mathbf{D}}\|_{0,\Omega}$ (cf. (3.75)), leads to

$$\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{\mathcal{H}_2} \,\le\, \Big(1 + L_{\mathcal{A}}\,\alpha_{\mathcal{A}}^{-1}\Big)\,\widetilde{\beta}_1^{-1}\Big\{2\,n^{1/2}\,\rho\,\delta\,\|\mathbf{z} - \underset{\sim}{\mathbf{z}}\|_{0,4;\Omega} \,+\, L_\eta\,\|r - \underset{\sim}{r}\|_{0,\Omega}\Big\}\,. \tag{3.81}$$

Furthermore, invoking (3.69), (3.53), and (3.70), and performing some simple algebraic computations, which include the use of Cauchy-Schwarz's inequality and the fact that $\|\mathrm{tr}(\boldsymbol{\tau})\|_{0,\Omega} \,\le\, n^{1/2}\,\|\boldsymbol{\tau}\|_{0,\Omega}$, we easily deduce that

$$\|\mathbf{T}(\mathbf{z},r) - \mathbf{T}(\underset{\sim}{\mathbf{z}},\underset{\sim}{r})\| \,\le\, \|\mathbf{u} - \underset{\sim}{\mathbf{u}}\|_{0,4;\Omega} + n^{-1/2}\,\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{0,\Omega} + 2\,n^{-1/2}\,\rho\,\|(\mathbf{u}\otimes\mathbf{u}) - (\underset{\sim}{\mathbf{u}}\otimes\underset{\sim}{\mathbf{u}})\|_{0,\Omega}\,, \tag{3.82}$$

from which, subtracting and adding $\mathbf{u}$ to one of the factors of $\mathbf{u}\otimes\mathbf{u}$, employing (3.26), recalling that $\mathbf{u},\ \underset{\sim}{\mathbf{u}} \in \mathrm{W}(\delta)$, and using from (3.67) that $\rho\,\delta \,\le\, \dfrac{1}{2\,C_{\mathbf{T}}}$, we arrive at

$$\begin{aligned}\|\mathbf{T}(\mathbf{z},r) - \mathbf{T}(\underset{\sim}{\mathbf{z}},\underset{\sim}{r})\| \,&\le\, \Big(1 \,+\, 4\,\rho\,\delta\Big)\,\|\mathbf{u} - \underset{\sim}{\mathbf{u}}\|_{0,4;\Omega} \,+\, n^{-1/2}\,\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{0,\Omega}\\ &\le\, \Big(1 \,+\, 2\,C_{\mathbf{T}}^{-1}\Big)\,\|\mathbf{u} - \underset{\sim}{\mathbf{u}}\|_{0,4;\Omega} \,+\, n^{-1/2}\,\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{0,\Omega}\,.\end{aligned}$$

Finally, replacing the estimates for $\|\mathbf{u} - \underset{\sim}{\mathbf{u}}\|_{0,4;\Omega}$ (cf. (3.76)) and $\|\boldsymbol{\sigma} - \underset{\sim}{\boldsymbol{\sigma}}\|_{0,\Omega}$ (cf. (3.81)) into the foregoing inequality, and recalling from Theorem 3.4.2 that $C_{\mathbf{T}}$ depends on $L_{\mathcal{A}}$,

$\alpha_{\mathcal{A}}$, $\widetilde{\beta}$, $\widetilde{\beta}_1$, n, and $\|\mathbf{i}_4\|$, we conclude the required inequality (3.68) with the constants

$$L_1(\mathbf{T}) \,:=\, 2\left(1 + 2\,C_{\mathbf{T}}^{-1}\right)\alpha_{\mathcal{A}}^{-1}\,\widetilde{\beta}^{-1}\,n^{1/2} \,+\, 2\left(1 + L_{\mathcal{A}}\,\alpha_{\mathcal{A}}^{-1}\right)\widetilde{\beta}_1^{-1}$$

and

$$L_2(\mathbf{T}) \,:=\, \left(1 + 2\,C_{\mathbf{T}}^{-1}\right)\alpha_{\mathcal{A}}^{-1}\,\widetilde{\beta}^{-1} \,+\, \left(1 + L_{\mathcal{A}}\,\alpha_{\mathcal{A}}^{-1}\right)\widetilde{\beta}_1^{-1}\,n^{-1/2}\,.$$

□

We are now in a position to state the first main result of this section.

Theorem 3.4.3. Assume that $\rho\,\delta$, L_η, and the data are sufficiently small so that

$$\begin{gathered}\rho\,\delta \,<\, \min\left\{\frac{1}{2\,C_{\mathbf{T}}}, \frac{1}{L_1(\mathbf{T})}\right\}, \qquad L_\eta \,<\, \frac{1}{L_2(\mathbf{T})}\,, \quad \text{and} \\ C_{\mathbf{T}}\left\{\|\mathbf{u}_D\|_{1/2,\Gamma} + \|\mathbf{f}\|_{0,4/3;\Omega}\right\} \,\leq\, \frac{\delta}{2}\,.\end{gathered} \tag{3.83}$$

Then, the operator $\mathbf{T}$ has a unique fixed point $(\mathbf{u},p) \in \mathrm{S}(\delta)$. Equivalently, given this $p \in \mathrm{L}^2_\kappa(\Omega)$, the system (3.35) has a unique solution $(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}}) \,:=\, \big(\mathbf{D},\boldsymbol{\sigma},(\mathbf{u},\gamma)\big) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ with $\mathbf{u} \in \mathrm{W}(\delta)$ and p satisfying (3.53). Moreover, there holds

$$\|(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}})\|_{\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}} \,\leq\, 2\,C_{\mathbf{T}}\left\{\|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega}\right\}. \tag{3.84}$$

Proof. According to the assumptions stipulated in (3.83), we deduce from Lemmas 3.4.4 and 3.4.5 that $\mathbf{T}$ is a contraction mapping $\mathrm{S}(\delta)$ into itself. Hence, a straightforward application of the classical Banach theorem implies the existence of a unique fixed point $(\mathbf{u},p) \in \mathrm{S}(\delta)$ of this operator, thus yielding the indicated consequences regarding the system (3.35). In turn, thanks to (3.65) (cf. Theorem 3.4.2) we have

$$\|(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}})\|_{\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}} \,\leq\, C_{\mathbf{T}}\left\{\rho\,\|\mathbf{u}\|^2_{0,4;\Omega} \,+\, \|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega}\right\},$$

whereas the fact that $\mathbf{u} \in \mathrm{W}(\delta)$ and the first assumption in (3.83) lead to

$$\rho\,\|\mathbf{u}\|^2_{0,4;\Omega} \,\leq\, \rho\,\delta\,\|\mathbf{u}\|_{0,4;\Omega} \,\leq\, \frac{1}{2\,C_{\mathbf{T}}}\,\|(\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}})\|_{\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}}\,,$$

so that from these two inequalities we readily obtain (3.84) and conclude the proof. □

Regarding the smallness data assumptions specified in (3.83), we notice that the fact that $C_{\mathbf{T}}$, $L_1(\mathbf{T})$, and $L_2(\mathbf{T})$ depend on other constants and parameters, some of which might not be known explicitly, makes hard, for not saying impossible, to actually verify those constraints in practice. Certainly, there do exist radius δ, constant L_η, and data $\|\mathbf{u}_D\|_{1/2,\Gamma}$ and $\|\mathbf{f}\|_{0,4/3;\Omega}$ satisfying them, but, unless all the aforementioned constants are explicitly known, we ignore how small they need to be in order to accomplish (3.83). The same comments apply to similar constraints along the paper, in particular those ensuring later on the unique solvability of the Galerkin scheme (cf. Theorem 3.5.5) and the corresponding *a priori* error estimates (cf. Theorem 3.5.6).

3.5 The Galerkin scheme

In this section we introduce the Galerkin scheme of the fully-mixed variational formulation (3.35), analyze its solvability by means of a discrete version of the fixed-point approach employed in Section 3.4, and derive the corresponding a priori error estimate.

3.5.1 Preliminaries

We begin by letting $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ be arbitrary finite dimensional subspaces of $\mathbb{L}^2_{\mathtt{tr}}(\Omega)$, $\mathbb{H}(\mathbf{div}_{4/3};\Omega)$, $\mathbf{L}^4(\Omega)$, and $\mathbb{L}^2_{\mathtt{sk}}(\Omega)$, respectively, and let $\mathcal{P}_h := \widetilde{\mathcal{P}}_h \oplus \left\{\dfrac{\kappa}{|\Omega|}\right\}$, where $\widetilde{\mathcal{P}}_h$ is a finite dimensional subspace of $\mathrm{L}^2_0(\Omega) := \left\{q \in \mathrm{L}^2(\Omega) : \quad \int_\Omega q = 0\right\}$. Hereafter, h stands for both the sub-index of each subspace and the size of each member of a regular family $\left\{\mathcal{T}_h\right\}_{h>0}$ of triangulations of $\overline{\Omega}$ made up of triangles K (when $n = 2$) or tetrahedra K (when $n = 3$) of diameters h_K, so that $h := \max\left\{h_K : \quad K \in \mathcal{T}_h\right\}$. Now, defining

$$\mathcal{H}_{2,h} := \mathbb{H}_0(\mathbf{div}_{4/3};\Omega) \cap \widetilde{\mathcal{H}}_{2,h} \quad \text{and} \quad \mathcal{Q}_h := \mathcal{Q}_{1,h} \times \mathcal{Q}_{2,h}\,,$$

and letting $p_h \in \mathcal{P}_h$ be a given discrete approximation of the pressure p, the Galerkin scheme associated with (3.35) reads: Find $(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h) := \Big(\mathbf{D}_h, \boldsymbol{\sigma}_h, (\mathbf{u}_h, \gamma_h)\Big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ such that

$$\begin{array}{llcll} [\mathcal{A}_{p_h}(\mathbf{D}_h), \mathbf{E}_h] & + \mathcal{B}_1(\mathbf{E}_h, \boldsymbol{\sigma}_h) & = & \mathcal{F}_{\mathbf{u}_h}(\mathbf{E}_h) & \forall\, \mathbf{E}_h \in \mathcal{H}_{1,h}\,, \\ \mathcal{B}_1(\mathbf{D}_h, \boldsymbol{\tau}_h) & \qquad\qquad + \mathcal{B}(\boldsymbol{\tau}_h, \vec{\mathbf{u}}_h) & = & \mathcal{G}(\boldsymbol{\tau}_h) & \forall\, \boldsymbol{\tau}_h \in \mathcal{H}_{2,h}\,, \\ & \mathcal{B}(\boldsymbol{\sigma}_h, \vec{\mathbf{v}}_h) & = & \mathcal{F}(\vec{\mathbf{v}}_h) & \forall\, \vec{\mathbf{v}}_h \in \mathcal{Q}_h\,. \end{array} \tag{3.85}$$

Next, we consider the discrete analogue of the fixed-point strategy employed in Section 3.4. Indeed, we introduce the discrete operator $\mathbf{T}_h : \mathcal{Q}_{1,h} \times \mathcal{P}_h \to \mathcal{Q}_{1,h} \times \mathcal{P}_h$ defined by

$$\mathbf{T}_h(\mathbf{z}_h, r_h) \,:=\, (\mathbf{u}_h, p_h) \qquad \forall\,(\mathbf{z}_h, r_h) \,\in\, \mathcal{Q}_{1,h} \times \mathcal{P}_h\,, \tag{3.86}$$

where $(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h) \,:=\, \Big(\mathbf{D}_h, \boldsymbol{\sigma}_h, (\mathbf{u}_h, \boldsymbol{\gamma}_h)\Big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ is the unique solution (to be confirmed later on) of the problem arising from (3.85) when $\mathcal{A}_{p_h}$ and the functional $\mathcal{F}_{\mathbf{u}_h}$ are replaced by $\mathcal{A}_{r_h}$ and $\mathcal{F}_{\mathbf{z}_h}$, respectively, that is

$$\begin{aligned}
[\mathcal{A}_{r_h}(\mathbf{D}_h), \mathbf{E}_h] \;\; + \mathcal{B}_1(\mathbf{E}_h, \boldsymbol{\sigma}_h) \qquad\qquad\qquad\qquad &= \; \mathcal{F}_{\mathbf{z}_h}(\mathbf{E}_h) \qquad \forall\, \mathbf{E}_h \in \mathcal{H}_{1,h}\,,\\
\mathcal{B}_1(\mathbf{D}_h, \boldsymbol{\tau}_h) \qquad\qquad\qquad\qquad + \mathcal{B}(\boldsymbol{\tau}_h, \vec{\mathbf{u}}_h) \;\; &= \;\; \mathcal{G}(\boldsymbol{\tau}_h) \qquad\;\; \forall\, \boldsymbol{\tau}_h \in \mathcal{H}_{2,h}\,,\\
\mathcal{B}(\boldsymbol{\sigma}_h, \vec{\mathbf{v}}_h) \qquad\qquad\qquad\quad\;\; &= \;\; \mathcal{F}(\vec{\mathbf{v}}_h) \qquad\;\; \forall\, \vec{\mathbf{v}}_h \in \mathcal{Q}_h\,,
\end{aligned} \tag{3.87}$$

whereas p_h is computed as suggested by the discrete version of (3.44), that is

$$p_h \,:=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma}_h + \rho\,(\mathbf{u}_h \otimes \mathbf{u}_h)\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\mathbf{u}_h \otimes \mathbf{u}_h)\,. \tag{3.88}$$

Note from (3.88) that the specific subspaces to which $\boldsymbol{\sigma}_h$ and $\mathbf{u}_h$ belong determine the choice of $\widetilde{\mathcal{P}}_h$. Then, it is readily seen that solving (3.85) is equivalent to finding a fixed point of $\mathbf{T}_h$, that is $(\mathbf{u}_h, p_h) \in \mathcal{Q}_{1,h} \times \mathcal{P}_h$ such that

$$\mathbf{T}_h(\mathbf{u}_h, p_h) \;=\; (\mathbf{u}_h, p_h)\,. \tag{3.89}$$

3.5.2 Discrete solvability analysis

In what follows we proceed analogously to Sections 3.4.2 and 3.4.3, and establish the well-posedness of the Galerkin scheme (3.85) by means of the solvability study of the equivalent fixed-point equation (3.89). In this regard, we announce in advance that, being the respective discussion similar to the one developed for the continuous formulation, here we simply collect the main results and provide selected details of their proofs. To this end, suitable hypotheses regarding the arbitrary subspaces $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, and $\mathcal{Q}_h$, need to be introduced throughout the analysis. Explicit finite element subspaces satisfying them will be specified later on in Section 3.6.

We begin by letting $\mathcal{V}_h$ be the discrete kernel of the bilinear form $\mathcal{B}$, that is

$$\mathcal{V}_h \,:=\, \Big\{\boldsymbol{\tau}_h \in \mathcal{H}_{2,h} : \quad \mathcal{B}(\boldsymbol{\tau}_h, \vec{\mathbf{v}}_h) \,=\, 0 \quad \forall\, \vec{\mathbf{v}}_h \in \mathcal{Q}_h\Big\}\,, \tag{3.90}$$

and by assuming that

(**H.4**) $\widetilde{\mathcal{H}}_{2,h}$ contains multiples of the identity tensor $\mathbb{I}$,

(**H.5**) $\mathbf{div}(\widetilde{\mathcal{H}}_{2,h}) \subseteq \mathcal{Q}_{1,h}$,

(**H.6**) $\mathcal{V}_h^{\mathtt{d}} := \left\{ \boldsymbol{\tau}_h^{\mathtt{d}} : \quad \boldsymbol{\tau}_h \in \mathcal{V}_h \right\} \subseteq \mathcal{H}_{1,h}$, and

(**H.7**) there exists a positive constant $\widetilde{\beta}_{\mathtt{d}}$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h \in \mathcal{H}_{2,h} \\ \boldsymbol{\tau}_h \neq 0}} \frac{\mathcal{B}(\boldsymbol{\tau}_h, \vec{\mathbf{v}}_h)}{\|\boldsymbol{\tau}_h\|_{\mathcal{H}_2}} \geq \widetilde{\beta}_{\mathtt{d}} \, \|\vec{\mathbf{v}}_h\|_{\mathcal{Q}} \,=\, \widetilde{\beta}_{\mathtt{d}} \left\{ \|\mathbf{v}_h\|_{0,4;\Omega} + \|\boldsymbol{\xi}_h\|_{0,\Omega} \right\} \qquad \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h, \boldsymbol{\xi}_h) \in \mathcal{Q}_h \,. \tag{3.91}$$

Then, as a consequence of (**H.4**), there holds the discrete version of the decomposition (3.30), namely $\widetilde{\mathcal{H}}_{2,h} = \mathcal{H}_{2,h} \oplus \mathrm{R}\mathbb{I}$, which confirms the validity of using $\mathcal{H}_{2,h}$ as the subspace where $\boldsymbol{\sigma}_h$ is sought. Now, according to the definition of $\mathcal{B}$ (cf. (3.38)), and noting that (**H.5**) can be equivalently rephrased as $\mathbf{div}(\mathcal{H}_{2,h}) \subseteq \mathcal{Q}_{1,h}$, it readily follows from (3.90) that

$$\mathcal{V}_h \,:=\, \left\{ \boldsymbol{\tau}_h \in \mathcal{H}_{2,h} : \quad \mathbf{div}(\boldsymbol{\tau}_h) = 0 \quad \text{and} \quad \int_\Omega \boldsymbol{\tau}_h : \boldsymbol{\xi}_h \,=\, 0 \quad \forall\, \boldsymbol{\xi}_h \in \mathcal{Q}_{2,h} \right\}, \tag{3.92}$$

which yields the discrete analogue of (3.63). Indeed, given $\boldsymbol{\tau}_h \in \mathcal{V}_h$ such that $\boldsymbol{\tau}_h^{\mathtt{d}} \neq \mathbf{0}$, we have thanks to (**H.6**) that $-\boldsymbol{\tau}_h^{\mathtt{d}} \in \mathcal{H}_{1,h}$, and thus

$$\sup_{\substack{\mathbf{E}_h \in \mathcal{H}_{1,h} \\ \mathbf{E}_h \neq 0}} \frac{\mathcal{B}_1(\mathbf{E}_h, \boldsymbol{\tau}_h)}{\|\mathbf{E}_h\|_{\mathcal{H}_1}} \geq \frac{\mathcal{B}_1(-\boldsymbol{\tau}_h^{\mathtt{d}}, \boldsymbol{\tau}_h)}{\|\boldsymbol{\tau}_h^{\mathtt{d}}\|_{\mathcal{H}_1}} \,=\, \|\boldsymbol{\tau}_h^{\mathtt{d}}\|_{0,\Omega} \,,$$

from which, employing the inequality (3.64), we arrive at

$$\sup_{\substack{\mathbf{E}_h \in \mathcal{H}_{1,h} \\ \mathbf{E}_h \neq 0}} \frac{\mathcal{B}_1(\mathbf{E}_h, \boldsymbol{\tau}_h)}{\|\mathbf{E}_h\|_{\mathcal{H}_1}} \geq \widetilde{\beta}_{1,\mathtt{d}} \, \|\boldsymbol{\tau}_h\|_{\mathcal{H}_1} \,, \tag{3.93}$$

with $\widetilde{\beta}_{1,\mathtt{d}} \,=\, c_1$. Now, if $\boldsymbol{\tau}_h \in \mathcal{V}_h$ is such that $\boldsymbol{\tau}_h^{\mathtt{d}} = \mathbf{0}$, then it follows from (3.64) that $\boldsymbol{\tau}_h = \mathbf{0}$, whence (3.93) holds trivially in this case.

Furthermore, it is not difficult to see that the Lipschitz-continuity and monotoniticity properties of $\mathcal{A}_r$ provided in Lemma 3.4.1 (cf. (3.59), (3.60), and (3.61)), are also valid in the present discrete case, and with the same constants $L_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, and L_η, that is

$$\begin{aligned} & \|\mathcal{A}_{r_h}(\mathbf{D}_h) - \mathcal{A}_{r_h}(\mathbf{E}_h)\|_{\mathcal{H}_{1,h}'}[2ex] \\ \leq \;& L_{\mathcal{A}} \, \|\mathbf{D}_h - \mathbf{E}_h\|_{\mathcal{H}_1} \qquad \forall\, r_h \in \mathcal{P}_h \,, \quad \forall\, \mathbf{D}_h,\, \mathbf{E}_h \in \mathcal{H}_{1,h} \,, \end{aligned} \tag{3.94}$$

$$\begin{aligned}&[\mathcal{A}_{r_h}(\mathbf{D}_h) - \mathcal{A}_{r_h}(\mathbf{E}_h), \mathbf{D}_h - \mathbf{E}_h][2ex]\\ &\geq \alpha_{\mathcal{A}} \|\mathbf{D}_h - \mathbf{E}_h\|^2_{\mathcal{H}_1} \qquad \forall\, r_h \in \mathcal{P}_h\,, \quad \forall\, \mathbf{D}_h,\, \mathbf{E}_h \in \mathcal{H}_{1,h}\,,\end{aligned} \tag{3.95}$$

and

$$\begin{aligned}&\Big|[\mathcal{A}_{r_h}(\mathbf{D}_h) - \mathcal{A}_{q_h}(\mathbf{D}_h), \mathbf{E}_h]\Big|[2ex]\\ &\leq L_\eta \, \|r_h - q_h\|_{0,\Omega} \, \|\mathbf{E}_h\|_{\mathcal{H}_1} \qquad \forall\, r_h,\, q_h \in \mathcal{P}_h\,, \quad \forall\, \mathbf{D}_h,\, \mathbf{E}_h \in \mathcal{H}_{1,h}\,.\end{aligned} \tag{3.96}$$

Consequently, we are now in a position to establish the discrete analogue of Theorem 3.4.2.

Theorem 3.5.4. For each $(\mathbf{z}_h, r_h) \in \mathcal{Q}_{1,h} \times \mathcal{P}_h$ there exists a unique tuple $(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h) := \big(\mathbf{D}_h, \boldsymbol{\sigma}_h, (\mathbf{u}_h, \gamma_h)\big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ which is the solution to (3.87), and hence one can define $\mathbf{T}(\mathbf{z}_h, r_h) := (\mathbf{u}_h, p_h) \in \mathcal{Q}_{1,h} \times \mathcal{P}_h$, where p_h is computed according to (3.88). Moreover, there exists a positive constant $C_{\mathbf{T},\mathtt{d}}$, depending only on $L_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, $\widetilde{\beta}_{\mathtt{d}}$, $\widetilde{\beta}_{1,\mathtt{d}}$, n, and $\|\mathbf{i}_4\|$, such that

$$\|\mathbf{u}_h\|_{0,4;\Omega} \,\leq\, \|(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h)\|_{\mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}} \,\leq\, C_{\mathbf{T},\mathtt{d}} \Big\{ \rho \, \|\mathbf{z}_h\|^2_{0,4;\Omega} \,+\, \|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega} \Big\}\,. \tag{3.97}$$

Proof. Thanks to the discrete inf-sup conditions for $\mathcal{B}$ (cf. (**H.7**)) and $\mathcal{B}_1$ (cf. (3.93)), and the properties (3.94), (3.95) and (3.96), the proof follows from a direct application of Theorem 3.4.1. We omit further details. □

Knowing that the discrete operator $\mathbf{T}_h$ is well defined, we now address the solvability of the fixed point equation (3.89). In fact, letting $\delta_{\mathtt{d}}$ be an arbitrary radius, we now set

$$\mathrm{W}(\delta_{\mathtt{d}}) \,:=\, \Big\{ \mathbf{z}_h \in \mathcal{Q}_{1,h} : \quad \|\mathbf{z}_h\|_{0,4;\Omega} \,\leq\, \delta_{\mathtt{d}} \Big\} \quad \text{and} \quad \mathrm{S}(\delta_{\mathtt{d}}) \,:=\, \mathrm{W}(\delta_{\mathtt{d}}) \times \mathcal{P}_h\,. \tag{3.98}$$

In this way, proceeding analogously to the deduction of Lemma 3.4.4, we find that, under the discrete analogue of (3.67), that is

$$\rho\, \delta_{\mathtt{d}} \,\leq\, \frac{1}{2\, C_{\mathbf{T},d}} \quad \text{and} \quad C_{\mathbf{T},\mathtt{d}} \Big\{ \|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega} \Big\} \,\leq\, \frac{\delta_{\mathtt{d}}}{2}\,, \tag{3.99}$$

$\mathbf{T}_h$ maps $\mathrm{S}(\delta_{\mathtt{d}})$ into itself. Note that the above is the same as for the continuous case (cf. (3.67)), except that the constant $C_{\mathbf{T}}$ and the radius δ are replaced by $C_{\mathbf{T},\mathtt{d}}$ and $\delta_{\mathtt{d}}$, respectively.

In addition, employing similar arguments to those from the proof of Lemma 3.4.5, we can prove the discrete version of (3.68) with the constants

$$L_{1,\mathtt{d}}(\mathbf{T}) \,:=\, 2\left(1 + 2\,C_{\mathbf{T},\mathtt{d}}^{-1}\right)\alpha_{\mathcal{A}}^{-1}\,\widetilde{\beta}_{\mathtt{d}}^{-1}\,n^{1/2} \,+\, 2\left(1 + L_{\mathcal{A}}\,\alpha_{\mathcal{A}}^{-1}\right)\widetilde{\beta}_{1,\mathtt{d}}^{-1}$$

and

$$L_{2,\mathtt{d}}(\mathbf{T}) \,:=\, \left(1 + 2\,C_{\mathbf{T},\mathtt{d}}^{-1}\right)\alpha_{\mathcal{A}}^{-1}\,\widetilde{\beta}_{\mathtt{d}}^{-1} \,+\, \left(1 + L_{\mathcal{A}}\,\alpha_{\mathcal{A}}^{-1}\right)\widetilde{\beta}_{1,\mathtt{d}}^{-1}\,n^{-1/2}\,,$$

that is

$$\|\mathbf{T}_h(\mathbf{z}_h, r_h) - \mathbf{T}_h(\utilde{\mathbf{z}}_h, \utilde{r}_h)\| \,\leq\, L_{1,\mathtt{d}}(\mathbf{T})\,\rho\,\delta_{\mathtt{d}}\,\|\mathbf{z}_h - \utilde{\mathbf{z}}_h\|_{0,4;\Omega} \,+\, L_{2,\mathtt{d}}(\mathbf{T})\,L_\eta\,\|r_h - \utilde{r}_h\|_{0,\Omega} \tag{3.100}$$

for all $(\mathbf{z}_h, r_h),\ (\utilde{\mathbf{z}}_h, \utilde{r}_h) \in \mathrm{S}(\delta_{\mathtt{d}})$.

The main result of this section, which constitutes the discrete analogue of Theorem 3.4.3, is then established as follows.

Theorem 3.5.5. Assume that $\rho\,\delta_{\mathtt{d}}$, L_η, and the data are sufficiently small so that

$$\begin{gathered}\rho\,\delta_{\mathtt{d}} \,<\, \min\left\{\frac{1}{2\,C_{\mathbf{T},\mathtt{d}}}, \frac{1}{L_{1,\mathtt{d}}(\mathbf{T})}\right\}, \qquad L_\eta \,<\, \frac{1}{L_{2,\mathtt{d}}(\mathbf{T})}\,, \quad \text{and} \\ C_{\mathbf{T},\mathtt{d}}\left\{\|\mathbf{u}_D\|_{1/2,\Gamma} + \|\mathbf{f}\|_{0,4/3;\Omega}\right\} \,\leq\, \frac{\delta_{\mathtt{d}}}{2}\,.\end{gathered} \tag{3.101}$$

Then, the operator $\mathbf{T}_h$ has a unique fixed point $(\mathbf{u}_h, p_h) \in \mathrm{S}(\delta_{\mathtt{d}})$. Equivalently, given this $p_h \in \mathcal{P}_h$, the system (3.85) has a unique solution $(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h) \,:=\, \big(\mathbf{D}_h, \boldsymbol{\sigma}_h, (\mathbf{u}_h, \boldsymbol{\gamma}_h)\big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ with $\mathbf{u}_h \in \mathrm{W}(\delta_{\mathtt{d}})$ and p_h satisfying (3.88). Moreover, there holds

$$\|(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h)\|_{\mathcal{H}_1\times\mathcal{H}_2\times\mathcal{Q}} \,\leq\, 2\,C_{\mathbf{T},\mathtt{d}}\left\{\|\mathbf{u}_D\|_{1/2,\Gamma} \,+\, \|\mathbf{f}\|_{0,4/3;\Omega}\right\}. \tag{3.102}$$

Proof. It is clear from the previous discussion and the assumptions in (3.101), that $\mathbf{T}_h$ is a contraction mapping $\mathrm{S}(\delta_{\mathtt{d}})$ into itself. Thus, a straightforward application of the classical Banach Theorem implies the existence of a unique solution to (3.89), and hence, equivalently, to the system (3.85). In turn, thanks to (3.97) (cf. Theorem 3.5.4), and performing similar algebraic manipulations to those utilized in the proof of Theorem 3.4.3, we deduce the a priori estimate (3.102). □

3.5.3 A priori error analysis

In this section we consider arbitrary finite element subspaces satisfying the assumptions specified in Section 3.5.2, and derive the Céa estimate for the Galerkin error given by

$$\|\vec{\mathbf{D}}-\vec{\mathbf{D}}_h\|_{\mathcal{H}}+\|p-p_h\|_{0,\Omega} \,:=\, \|\mathbf{D}-\mathbf{D}_h\|_{0,\Omega} + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div}_{4/3};\Omega} + \|\vec{\mathbf{u}}-\vec{\mathbf{u}}_h\|_{\mathcal{Q}} + \|p-p_h\|_{0,\Omega}\,,$$

where $\vec{\mathbf{D}} := (\mathbf{D},\boldsymbol{\sigma},\vec{\mathbf{u}}) = \big(\mathbf{D},\boldsymbol{\sigma},(\mathbf{u},\gamma)\big) \in \mathcal{H} := \mathcal{H}_1 \times \mathcal{H}_2 \times \mathbf{Q}$ is the unique solution of (3.35), with $\mathbf{u} \in \mathrm{S}(\delta)$, and $\vec{\mathbf{D}}_h := (\mathbf{D}_h,\boldsymbol{\sigma}_h,\vec{\mathbf{u}}_h) = \big(\mathbf{D}_h,\boldsymbol{\sigma}_h,(\mathbf{u}_h,\gamma_h)\big) \in \mathcal{H}_h := \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathbf{Q}_h$ is the unique solution of (3.85), with $\mathbf{u}_h \in \mathrm{S}(\delta_{\mathbf{d}})$, whereas p and p_h are computed according to (3.53) and (3.88), respectively.

We begin by defining for each $r \in \mathrm{L}^2_\kappa(\Omega)$ the operator $\Xi_r : \mathcal{H} \to \mathcal{H}'$ that arises from the left-hand side of the variational formulation (3.35) after adding all its rows, that is

$$[\Xi_r(\vec{\mathbf{C}}),\vec{\mathbf{E}}] \,:=\, [\mathcal{A}_r(\mathbf{C}),\mathbf{E}] \,+\, \mathcal{B}_1(\mathbf{E},\boldsymbol{\zeta}) \,+\, \mathcal{B}_1(\mathbf{C},\boldsymbol{\tau}) \,+\, \mathcal{B}(\boldsymbol{\tau},\vec{\mathbf{w}}) \,+\, \mathcal{B}(\boldsymbol{\zeta},\vec{\mathbf{v}})\,, \tag{3.103}$$

for all $\vec{\mathbf{C}} := (\mathbf{C},\boldsymbol{\zeta},\vec{\mathbf{w}})$, $\vec{\mathbf{E}} := (\mathbf{E},\boldsymbol{\tau},\vec{\mathbf{v}}) \in \mathcal{H}$, so that (3.35) and (3.85) can be rewritten, respectively, as

$$[\Xi_p(\vec{\mathbf{D}}),\vec{\mathbf{E}}] \,=\, \mathcal{F}_{\mathbf{u}}(\mathbf{E}) + \mathcal{G}(\boldsymbol{\tau}) + \mathcal{F}(\vec{\mathbf{v}}) \qquad \forall\, \vec{\mathbf{E}} := (\mathbf{E},\boldsymbol{\tau},\vec{\mathbf{v}}) \in \mathcal{H}\,, \tag{3.104}$$

and

$$[\Xi_{p_h}(\vec{\mathbf{D}}_h),\vec{\mathbf{E}}_h] \,=\, \mathcal{F}_{\mathbf{u}_h}(\mathbf{E}_h) + \mathcal{G}(\boldsymbol{\tau}_h) + \mathcal{F}(\vec{\mathbf{v}}_h) \qquad \forall\, \vec{\mathbf{E}}_h := (\mathbf{E}_h,\boldsymbol{\tau}_h,\vec{\mathbf{v}}_h) \in \mathcal{H}_h\,. \tag{3.105}$$

It readily follows from (3.104) and (3.105) that

$$[\Xi_p(\vec{\mathbf{D}}),\vec{\mathbf{E}}_h] - [\Xi_{p_h}(\vec{\mathbf{D}}_h),\vec{\mathbf{E}}_h] \,=\, \Big(\mathcal{F}_{\mathbf{u}} - \mathcal{F}_{\mathbf{u}_h}\Big)(\mathbf{E}_h) \qquad \forall\, \vec{\mathbf{E}}_h := (\mathbf{E}_h,\boldsymbol{\tau}_h,\vec{\mathbf{v}}_h) \in \mathcal{H}_h\,. \tag{3.106}$$

Now, the smoothness of the regularized η (cf. (3.11)) allows to show that for each $r \in \mathrm{L}^2_\kappa(\Omega)$, the operator $\mathcal{A}_r$, and hence Ξ_r as well, have first order Gâteaux derivatives $\mathcal{D}(\mathcal{A}_r) \in \mathcal{L}\Big(\mathcal{H}_1,\mathcal{L}(\mathcal{H}_1,\mathcal{H}_1')\Big)$ and $\mathcal{D}(\Xi_r) \in \mathcal{L}\Big(\mathcal{H},\mathcal{L}(\mathcal{H},\mathcal{H}')\Big)$, respectively, as well as their corresponding discrete versions denoted by $\mathcal{D}_h(\mathcal{A}_r) \in \mathcal{L}\Big(\mathcal{H}_{1,h},\mathcal{L}(\mathcal{H}_{1,h},\mathcal{H}_{1,h}')\Big)$ and $\mathcal{D}_h(\Xi_r) \in \mathcal{L}\Big(\mathcal{H}_h,\mathcal{L}(\mathcal{H}_h,\mathcal{H}_h')\Big)$. Moreover, using (3.59) and (3.60) (cf. Lemma 3.4.1), one is able to prove (see, e.g. [63, Lemma 3.1]) that for each $\mathbf{C}_h \in \mathcal{H}_{1,h}$, the operator $\mathcal{D}_h(\mathcal{A}_r)(\mathbf{C}_h) \in \mathcal{L}(\mathcal{H}_{1,h},\mathcal{H}_{1,h}')$ can be identified as a bounded and $\mathcal{H}_1$-elliptic bilinear form with constants $L_{\mathcal{A}}$ and $\alpha_{\mathcal{A}}$, respectively. It follows that for each $r \in \mathrm{L}^2_\kappa(\Omega)$, and for each $\vec{\mathbf{C}}_h \in \mathcal{H}_h$, the operator $\mathcal{D}_h(\Xi_r)(\vec{\mathbf{C}}_h) \in \mathcal{L}(\mathcal{H}_h,\mathcal{H}_h')$ satisfies the hypotheses of the discrete linear version of Theorem 3.4.1, and hence the corresponding global inf-sup condition as well with a positive constant $\alpha_{\boldsymbol{\Xi},\mathtt{d}}$, depending only on $\mathrm{L}_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, $\widetilde{\beta}_{\mathtt{d}}$, and $\widetilde{\beta}_{1,\mathtt{d}}$. In this way, proceeding analogously to the proof of [63, Theorem 3.3], which includes, in particular, applying the mean value theorem to $\boldsymbol{\Xi}_r$, we deduce that for each $r \in \mathrm{L}^2_\kappa(\Omega)$ there holds

$$\alpha_{\boldsymbol{\Xi},\mathtt{d}} \, \|\underset{\sim}{\vec{\mathbf{C}}}_h - \vec{\mathbf{C}}_h\|_{\mathcal{H}} \, \leq \sup_{\substack{\vec{\mathbf{E}}_h \in \mathcal{H}_h \\ \vec{\mathbf{E}}_h \neq \mathbf{0}}} \frac{[\boldsymbol{\Xi}_r(\underset{\sim}{\vec{\mathbf{C}}}_h) - \boldsymbol{\Xi}_r(\vec{\mathbf{C}}_h), \vec{\mathbf{E}}_h]}{\|\vec{\mathbf{E}}_h\|_{\mathcal{H}}} \qquad \forall \, \underset{\sim}{\vec{\mathbf{C}}}_h, \, \vec{\mathbf{C}}_h \in \mathcal{H}_h \, . \tag{3.107}$$

Then, we begin our derivation by employing the triangle inequality, which gives

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \, \leq \, \|\vec{\mathbf{D}} - \vec{\mathbf{C}}_h\|_{\mathcal{H}} \, + \, \|\vec{\mathbf{D}}_h - \vec{\mathbf{C}}_h\|_{\mathcal{H}} \qquad \forall \, \vec{\mathbf{C}}_h \in \mathcal{H}_h \, , \tag{3.108}$$

whereas, applying (3.107) with $r = p$, we obtain

$$\alpha_{\boldsymbol{\Xi},\mathtt{d}} \, \|\vec{\mathbf{D}}_h - \vec{\mathbf{C}}_h\|_{\mathcal{H}} \, \leq \sup_{\substack{\vec{\mathbf{E}}_h \in \mathcal{H}_h \\ \vec{\mathbf{E}}_h \neq \mathbf{0}}} \frac{[\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_p(\vec{\mathbf{C}}_h), \vec{\mathbf{E}}_h]}{\|\vec{\mathbf{E}}_h\|_{\mathcal{H}}} \, . \tag{3.109}$$

Next, subtracting and adding $[\boldsymbol{\Xi}_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}_h]$, we find that

$$[\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_p(\vec{\mathbf{C}}_h), \vec{\mathbf{E}}_h] \, = \, [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}_h] \, + \, [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}) - \boldsymbol{\Xi}_p(\vec{\mathbf{C}}_h), \vec{\mathbf{E}}_h] \, , \tag{3.110}$$

so that, employing the Lipschitz-continuity of $\mathcal{A}_p$ (cf. (3.59), Lemma 3.4.1), we deduce from (3.103) the existence of a positive constant $L_{\boldsymbol{\Xi}}$, depending on $L_{\mathcal{A}}$, $\|\mathcal{B}\|$, and $\|\mathcal{B}_1\|$, such that the second term on the right-hand side of (3.110) is bounded as

$$\Big| [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}) - \boldsymbol{\Xi}_p(\vec{\mathbf{C}}_h), \vec{\mathbf{E}}_h] \Big| \, \leq \, L_{\boldsymbol{\Xi}} \, \|\vec{\mathbf{D}} - \vec{\mathbf{C}}_h\|_{\mathcal{H}} \, \|\vec{\mathbf{E}}_h\|_{\mathcal{H}} \, . \tag{3.111}$$

In turn, subtracting and adding $[\boldsymbol{\Xi}_{p_h}(\vec{\mathbf{D}}_h), \vec{\mathbf{E}}_h]$, applying the Lipschitz-continuity of $\mathcal{A}$ with respect to the pressure (cf. (3.61), Lemma 3.4.1), and employing (3.106), we find that

$$\begin{aligned} \Big| [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}_h] \Big| \, &= \, \Big| [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_{p_h}(\vec{\mathbf{D}}_h), \vec{\mathbf{E}}_h] + [\boldsymbol{\Xi}_{p_h}(\vec{\mathbf{D}}_h) - \boldsymbol{\Xi}_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}_h] \Big| \\ &\leq \, \Big\{ L_\eta \, \|p - p_h\|_{0,\Omega} \, + \, \|\mathcal{F}_{\mathbf{u}} - \mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}'_{1,h}} \Big\} \, \|\vec{\mathbf{E}}_h\|_{\mathcal{H}} \, . \end{aligned} \tag{3.112}$$

In this way, using (3.111) and (3.112) to bound the expression in (3.110), and then replacing the resulting estimate in (3.109), we arrive at

$$\alpha_{\boldsymbol{\Xi},\mathtt{d}}\,\|\vec{\mathbf{D}}_h-\vec{\mathbf{C}}_h\|_{\mathcal{H}}\;\le\;L_{\boldsymbol{\Xi}}\,\|\vec{\mathbf{D}}-\vec{\mathbf{C}}_h\|_{\mathcal{H}}+L_\eta\,\|p-p_h\|_{0,\Omega}\;+\;\|\mathcal{F}_{\mathbf{u}}-\mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}'_{1,h}}\,,\qquad(3.113)$$

which, along with (3.108), implies

$$\|\vec{\mathbf{D}}-\vec{\mathbf{D}}_h\|_{\mathcal{H}}\;\le\;\Big(1+\alpha_{\boldsymbol{\Xi},\mathtt{d}}^{-1}\,L_{\boldsymbol{\Xi}}\Big)\,\mathrm{dist}\big(\vec{\mathbf{D}},\mathcal{H}_h\big)+\alpha_{\boldsymbol{\Xi},\mathtt{d}}^{-1}\Big\{L_\eta\,\|p-p_h\|_{0,\Omega}\;+\;\|\mathcal{F}_{\mathbf{u}}-\mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}'_{1,h}}\Big\}\,,\qquad(3.114)$$

where, as usual,

$$\mathrm{dist}\big(\vec{\mathbf{D}},\mathcal{H}_h\big)\;:=\;\inf_{\vec{\mathbf{C}}_h\in\mathcal{H}_h}\|\vec{\mathbf{D}}-\vec{\mathbf{C}}_h\|_{\mathcal{H}}\,.$$

Furthermore, according to the expressions provided by (3.53) and (3.88), and proceeding similarly to the derivation of the last two terms in (3.82), we get

$$\|p-p_h\|_{0,\Omega}\;\le\;n^{-1/2}\Big\{\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\;+\;2\,\rho\,\|(\mathbf{u}\otimes\mathbf{u})-(\mathbf{u}_h\otimes\mathbf{u}_h)\|_{0,\Omega}\Big\}\,.\qquad(3.115)$$

In addition, invoking now the definition of $\mathcal{F}_{\mathbf{z}}$ (cf. (3.39)) as in (3.73), we obtain

$$\Big(\mathcal{F}_{\mathbf{u}}-\mathcal{F}_{\mathbf{u}_h}\Big)(\mathbf{E}_h)\;=\;\rho\int_\Omega\Big((\mathbf{u}\otimes\mathbf{u})-(\mathbf{u}_h\otimes\mathbf{u}_h)\Big):\mathbf{E}_h\qquad\forall\,\mathbf{E}_h\in\mathcal{H}_{1,h}\,,$$

which gives

$$\|\mathcal{F}_{\mathbf{u}}-\mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}'_{1,h}}\;\le\;\rho\,\|(\mathbf{u}\otimes\mathbf{u})-(\mathbf{u}_h\otimes\mathbf{u}_h)\|_{0,\Omega}\,.\qquad(3.116)$$

Then, replacing the bounds from (3.115) and (3.116) back into (3.114), and denoting the constants

$$C_{1,\boldsymbol{\Xi}}\;:=\;1+\alpha_{\boldsymbol{\Xi},\mathtt{d}}^{-1}\,L_{\boldsymbol{\Xi}}\,,\quad C_{2,\boldsymbol{\Xi}}\;:=\;\alpha_{\boldsymbol{\Xi},\mathtt{d}}^{-1}\,n^{-1/2}\,,\quad\text{and}\quad C_{3,\boldsymbol{\Xi}}\;:=\;\alpha_{\boldsymbol{\Xi},\mathtt{d}}^{-1}\,\rho\,\Big(2\,n^{-1/2}\,L_\eta+1\Big)\,,$$

we conclude that

$$\begin{aligned}\|\vec{\mathbf{D}}-\vec{\mathbf{D}}_h\|_{\mathcal{H}}\;\;&\le\;\;C_{1,\boldsymbol{\Xi}}\,\mathrm{dist}\big(\vec{\mathbf{D}},\mathcal{H}_h\big)+C_{2,\boldsymbol{\Xi}}\,L_\eta\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\\&\qquad+C_{3,\boldsymbol{\Xi}}\,\|(\mathbf{u}\otimes\mathbf{u})-(\mathbf{u}_h\otimes\mathbf{u}_h)\|_{0,\Omega}\,.\end{aligned}\qquad(3.117)$$

Finally, similarly to the derivation of (3.74), there holds

$$\begin{aligned}\|(\mathbf{u}\otimes\mathbf{u})-(\mathbf{u}_h\otimes\mathbf{u}_h)\|_{0,\Omega}\;&\le\;n^{1/2}\Big(\|\mathbf{u}\|_{0,4;\Omega}+\|\mathbf{u}_h\|_{0,4;\Omega}\Big)\,\|\mathbf{u}-\mathbf{u}_h\|_{0,4;\Omega}\\&\le\;n^{1/2}\Big(\delta+\delta_{\mathtt{d}}\Big)\,\|\mathbf{u}-\mathbf{u}_h\|_{0,4;\Omega}\,,\end{aligned}\qquad(3.118)$$

and hence the inequalities (3.115) and (3.117) become, respectively,

$$\|p - p_h\|_{0,\Omega} \,\leq\, n^{-1/2}\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} \,+\, 2\,\rho\,\Big(\delta + \delta_{\mathtt{d}}\Big)\,\|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega}\,, \tag{3.119}$$

and

$$\begin{aligned} \|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \;\leq\; & C_{1,\Xi}\,\mathrm{dist}\big(\vec{\mathbf{D}}, \mathcal{H}_h\big) + C_{2,\Xi}\,L_\eta\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} \\ & + C_{3,\Xi}\,n^{1/2}\,\Big(\delta + \delta_{\mathtt{d}}\Big)\,\|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega}\,. \end{aligned} \tag{3.120}$$

We are now in position to establish the main result of this section.

Theorem 3.5.6. In addition to the notations and hypotheses of Theorems 3.4.3 and 3.5.5, assume that L_η, and the radii δ and δ_d are sufficiently small so that

$$C_{2,\Xi}\,L_\eta \,\leq\, \frac{1}{2} \quad \text{and} \quad C_{3,\Xi}\,n^{1/2}\,\Big(\delta + \delta_{\mathtt{d}}\Big) \,\leq\, \frac{1}{2}\,. \tag{3.121}$$

Then, there exists a positive constant C, independent of h, such that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} + \|p - p_h\|_{0,\Omega} \,\leq\, C\,\mathrm{dist}\big(\vec{\mathbf{D}}, \mathcal{H}_h\big)\,. \tag{3.122}$$

Proof. By employing (3.121) in (3.120), we readily deduce that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,\leq\, 2\,C_{1,\Xi}\,\mathrm{dist}\big(\vec{\mathbf{D}}, \mathcal{H}_h\big)\,,$$

whereas the corresponding estimate for $\|p - p_h\|_{0,\Omega}$ follows from (3.119) and the foregoing inequality. □

As announced at the end of Section 3.4, and similarly as for those required by Theorems 3.4.3 and 3.5.5, the assumptions in (3.121) are, in general, not verifiable in practice, but certainly there do exist constants L_η, δ, and $\delta_{\mathtt{d}}$ satisfying them. When the latter occurs, the present Theorem 3.5.6 ensures that there holds the *a priori* error estimate (3.122), and hence also the rates of convergence that are provided later on in Theorem 3.6.7. We do not know for certain whether the lack of satisfaction of any of the constraints (3.83), (3.101), or (3.121), would affect the accurateness of the method. In other words, the small data conditions are sufficient for the well-posedness of the continuous and discrete formulations, and for the expected convergence behaviour of the numerical scheme, but they might not be necessary. Indeed, this seems to be the case for the numerical results that are presented in Section 3.7, for which the eventual accomplishment of the aforementioned assumptions is not available either, and yet optimal rates of convergence are obtained.

3.6 Specific finite element subspaces

In this section we proceed as in [62, Section 4.4], where, in turn, the analysis from [22, Section 4.4] is employed, to describe two examples of finite element subspaces $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$, satisfying the hypotheses (**H.4**), (**H.5**), (**H.6**), and (**H.7**) that were introduced in Section 3.5.2. The associated rates of convergence are also provided.

3.6.1 Polynomial spaces

We first collect some definitions regarding local and global polynomial spaces, for which we make use of the regular family of triangulations $\big\{\mathcal{T}_h\big\}_{h>0}$ of $\overline{\Omega}$ introduced in Section 3.5.1. Indeed, given an integer $\ell \geq 0$ and $K \in \mathcal{T}_h$, we let $\mathrm{P}_\ell(K)$ be the space of polynomials of degree $\leq \ell$ defined on K, and denote its vector and tensor versions by $\mathbf{P}_\ell(K) := [\mathrm{P}_\ell(K)]^n$ and $\mathbb{P}_\ell(K) = [\mathrm{P}_\ell(K)]^{n\times n}$, respectively. In addition, we let $\mathbb{RT}_\ell(K) := \mathbf{P}_\ell(K) \oplus \mathrm{P}_\ell(K)\,\mathbf{x}$ be the local Raviart–Thomas space of order ℓ defined on K, where $\mathbf{x}$ stands for a generic vector in $\mathbf{R} := \mathrm{R}^n$. Also, we let b_K be the bubble function on K, which is defined as the product of its $n+1$ barycentric coordinates. Then, we define the local bubble spaces of order ℓ as

$$\begin{aligned} \mathbf{B}_\ell(K) &:= \mathrm{curl}\Big(b_K\,\mathrm{P}_\ell(K)\Big) \quad \text{if} \quad n=2\,,\\ \text{and} \quad \mathbf{B}_\ell(K) &:= \mathrm{curl}\Big(b_K\,\mathbf{P}_\ell(K)\Big) \quad \text{if} \quad n=3\,, \end{aligned} \tag{3.123}$$

where $\mathrm{curl}\,(v) := \Big(\dfrac{\partial v}{\partial x_2}, -\dfrac{\partial v}{\partial x_1}\Big)$ if $n=2$ and $v: K \to \mathrm{R}$, and $\mathrm{curl}\,(\mathbf{v}) := \nabla\times\mathbf{v}$ if $n=3$ and $\mathbf{v}: K \to \mathrm{R}^3$. The following global spaces are also needed

$$\mathbf{P}_\ell(\Omega) := \Big\{\mathbf{v}_h \in \mathbf{L}^2(\Omega): \quad \mathbf{v}_h|_K \in \mathbf{P}_\ell(K) \quad \forall\,K \in \mathcal{T}_h\Big\}\,,$$

$$\mathbb{P}_\ell(\Omega) := \Big\{\boldsymbol{\xi}_h \in \mathbb{L}^2(\Omega): \quad \boldsymbol{\xi}_h|_K \in \mathbb{P}_\ell(K) \quad \forall\,K \in \mathcal{T}_h\Big\}\,,$$

$$\mathbb{RT}_\ell(\Omega) := \Big\{\boldsymbol{\tau}_h \in \mathbb{H}(\mathbf{div};\Omega): \quad \boldsymbol{\tau}_{h,i}|_K \in \mathbb{RT}_\ell(K) \quad \forall\,i \in \big\{1,...,n\big\}, \quad \forall\,K \in \mathcal{T}_h\Big\}\,,$$

and

$$\mathbb{B}_\ell(\Omega) := \Big\{\boldsymbol{\tau}_h \in \mathbb{H}(\mathbf{div};\Omega): \quad \boldsymbol{\tau}_{h,i}|_K \in \mathbf{B}_\ell(K) \quad \forall\,i \in \big\{1,...,n\big\}, \quad \forall\,K \in \mathcal{T}_h\Big\}\,,$$

where $\boldsymbol{\tau}_{h,i}$ stands for the ith-row of $\boldsymbol{\tau}_h$. While $\mathbf{P}_\ell(\Omega)$ and $\mathbb{P}_\ell(\Omega)$ are defined here as subspaces of $\mathbf{L}^2(\Omega)$ and $\mathbb{L}^2(\Omega)$, we stress that they are also subspaces of $\mathbf{L}^4(\Omega)$

and $\mathbb{L}^4(\Omega)$, respectively. Similarly, it is easy to see that $\mathbb{RT}_\ell(\Omega)$ and $\mathbb{B}_\ell(\Omega)$ are both subspaces of $\mathbb{H}(\mathbf{div}_{4/3};\Omega)$ as well. Actually, recalling that $\mathbb{H}(\mathbf{div};\Omega)$ stands for the Hilbertian version of (2.2), that is with $t=2$, it is clear that $\mathbb{H}(\mathbf{div};\Omega)$ is contained in $\mathbb{H}(\mathbf{div}_{4/3};\Omega)$, and hence any subspace of the former is also subspace of the latter. Certainly, the same observation is valid for $\mathbb{H}_0(\mathbf{div};\Omega)$ and $\mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$, where the former is defined analogously to (3.31).

3.6.2 Connection with linear elasticity

Here we describe a useful connection between (**H.7**) and the stability of a usual mixed finite element method for the linear elasticity model. We begin by recalling that a triplet of subspaces $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ of $\mathbb{H}(\mathbf{div};\Omega)$, $\mathbf{L}^2(\Omega)$, and $\mathbb{L}^2_{\mathtt{sk}}(\Omega)$, respectively, is said to yield a stable Galerkin scheme for the Hilbertian mixed formulation of linear elasticity if it satisfies the corresponding hypotheses of the discrete Babuška-Brezzi theory (see, e.g., [46, Theorem 2.4]). In particular, the above includes the discrete inf-sup condition for the bilinear form $\mathcal{B}$ (cf. (3.38)), which, setting $\mathcal{H}_{2,h} := \widetilde{\mathcal{H}}_{2,h} \cap \mathbb{H}_0(\mathbf{div};\Omega)$, reduces to the existence of a positive constant $\widetilde{\beta}_{\mathtt{e}}$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h\in\mathcal{H}_{2,h}\\ \boldsymbol{\tau}_h\neq 0}} \frac{\mathcal{B}(\boldsymbol{\tau}_h,\vec{\mathbf{v}}_h)}{\|\boldsymbol{\tau}_h\|_{\mathbf{div};\Omega}} \geq \widetilde{\beta}_{\mathtt{e}} \Big\{ \|\mathbf{v}_h\|_{0,\Omega} + \|\boldsymbol{\xi}_h\|_{0,\Omega} \Big\} \qquad \forall\, \vec{\mathbf{v}}_h := (\mathbf{v}_h,\boldsymbol{\xi}_h) \in \mathcal{Q}_h\,. \tag{3.124}$$

Note that, though similar, (3.124) and (3.91) differ because of the different norms in which $\boldsymbol{\tau}_h$ and $\mathbf{v}_h$ are measured. However, the following result (cf. [22, Lemma 4.8]) establishes that (3.124), along with suitable further assumptions on the subspaces, constitute a sufficient condition for (3.91).

Lemma 3.6.6. Let $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ be subspaces of $\mathbb{H}(\mathbf{div};\Omega)$, $\mathbf{L}^2(\Omega)$, and $\mathbb{L}^2_{\mathtt{sk}}(\Omega)$, respectively, such that they accomplish (3.124). In addition, assume that there exists an integer $\ell \geq 0$ such that $\mathbb{RT}_\ell(\Omega) \subseteq \widetilde{\mathcal{H}}_{2,h}$ and $\mathcal{Q}_{1,h} \subseteq \mathbf{P}_\ell(\Omega)$. Then $\mathcal{H}_{2,h} := \widetilde{\mathcal{H}}_{2,h} \cap \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ satisfy (3.91) with a positive constant $\widetilde{\beta}_{\mathtt{d}}$, independent of h.

3.6.3 Examples of stable finite element subspaces

We now apply Lemma 3.6.6 to each one of the stable triplets for linear elasticity proposed in [22, Section 4.4], thus deriving two examples of finite element subspaces $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ satisfying (**H.4**), (**H.5**), (**H.6**), and (**H.7**).

Our first example is based on the plane elasticity element with reduced symmetry (PEERS) of order $\ell \geq 0$, which, denoting $\mathbb{C}(\bar{\Omega}) := [C(\bar{\Omega})]^{n\times n}$, is given by

$$\widetilde{\mathcal{H}}_{2,h} := \mathbb{RT}_\ell(\Omega) \oplus \mathbb{B}_\ell(\Omega)\,, \quad \mathcal{Q}_{1,h} := \mathbf{P}_\ell(\Omega)\,, \quad \text{and} \quad \mathcal{Q}_{2,h} := \mathbb{C}(\bar{\Omega}) \cap \mathbb{P}_{\ell+1}(\Omega) \cap \mathbb{L}^2_{\mathtt{sk}}(\Omega)\,. \tag{3.125}$$

The discrete stability of these subspaces was originally proved in [64] for $\ell = 0$ and $n = 2$, and later on for $\ell \geq 0$ and $n \in \big\{2, 3\big\}$ in [65]. It is easily seen from (3.125), in particular using due to (3.123) that $\mathbf{div}(\widetilde{\mathcal{H}}_{2,h}) = \mathbf{div}\Big(\mathbb{RT}_\ell(\Omega)\Big) \subseteq \mathbf{P}_\ell(\Omega)$, that $\widetilde{\mathcal{H}}_{2,h}$ and $\mathcal{Q}_{1,h}$ satisfy (**H.4**) and (**H.5**), and that the assumptions on them required by Lemma 3.6.6 are accomplished as well, whence (**H.7**) holds true. It remains to check (**H.6**), for which we first recall that the divergence free tensors of $\mathbb{RT}_\ell(\Omega)$ are contained in $\mathbb{P}_\ell(\Omega)$ (cf. [46, proof of Theorem 3.3]). Thus, noting again that the tensors of $\mathbb{B}_\ell(\Omega)$ are divergence free, and that this space is contained in $\mathbb{P}_{\ell+n}(\Omega)$, we deduce from (3.92) that

$$\mathcal{V}_h \subseteq \mathbb{P}_\ell(\Omega) \oplus \mathbb{B}_\ell(\Omega) \subseteq \mathbb{P}_{\ell+n}(\Omega)\,,$$

so that, in order to guarantee (**H.6**), it suffices to take

$$\mathcal{H}_{1,h} := \mathbb{P}_{\ell+n}(\Omega) \cap \mathbb{L}^2_{\mathtt{tr}}(\Omega)\,.$$

Finally, it follows from (3.88) and the above definitions of $\mathcal{H}_{2,h}$ and $\mathcal{Q}_{1,h}$, that $\mathcal{P}_h := \widetilde{\mathcal{P}}_h \oplus \left\{\dfrac{\kappa}{|\Omega|}\right\}$, where $\widetilde{\mathcal{P}}_h := \mathrm{P}_{\bar{\ell}}(\Omega) \cap \mathrm{L}^2_0(\Omega)$, with $\bar{\ell} := \max\Big\{\ell + n, 2\,\ell\Big\}$. Our second example is the Arnold-Falk-Winther (AFW) element of order $\ell \geq 0$, whose stability for the Hilbertian mixed formulation of linear elasticity is proved in [66], and which is defined as

$$\widetilde{\mathcal{H}}_{2,h} := \mathbb{P}_{\ell+1}(\Omega) \cap \mathbb{H}(\mathbf{div};\Omega)\,, \quad \mathcal{Q}_{1,h} := \mathbf{P}_\ell(\Omega)\,, \quad \text{and} \quad \mathcal{Q}_{2,h} := \mathbb{P}_\ell(\Omega) \cap \mathbb{L}^2_{\mathtt{sk}}(\Omega)\,. \tag{3.126}$$

According to the above, it is also simple to realize that (**H.4**) and (**H.5**) are satisfied, and that, thanks to the inclusion $\mathbb{RT}_\ell(\Omega) \subseteq \mathbb{P}_{\ell+1}(\Omega)$, the corresponding hypotheses of Lemma 3.6.6 are fulfilled, thus establishing (**H.7**). In turn, being in this case $\mathcal{V}_h$ (cf. (3.92)) not further simplifiable, we deduce that (**H.6**) is accomplished if we simply choose

$$\mathcal{H}_{1,h} := \mathbb{P}_{\ell+1}(\Omega) \cap \mathbb{L}^2_{\mathtt{tr}}(\Omega)\,.$$

Furthermore, it is readily seen in this case that $\mathcal{P}_h := \widetilde{\mathcal{P}}_h \oplus \left\{\dfrac{\kappa}{|\Omega|}\right\}$, where $\widetilde{\mathcal{P}}_h := \mathrm{P}_{2\ell}(\Omega) \cap \mathrm{L}^2_0(\Omega)$.

3.6.4 The rates of convergence

We now provide the rates of convergence of (3.85) for both specific examples of finite element subspaces introduced in Section 3.6.3. To this end, we first collect next the corresponding approximation properties of $\mathcal{H}_{1,h}$, $\mathcal{H}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$, which, taken mainly from [67], [68], [19, eqs. (5.37) and (5.40)], and [45, Proposition 1.135], are derived by employing the error estimates of suitable interpolation and projection operators, along with associated commuting diagram properties and interpolation estimates of Sobolev spaces.

Denoting $\ell^* := \begin{cases} \ell + n & \text{for PEERS-based} \\ \ell + 1 & \text{for AFW-based} \end{cases}$, the respective statements are as follows:

$\mathbf{AP}\Big(\mathcal{H}_{1,h}\Big)$ there exists a positive constant C, independent of h, such that for each $r \in [0, \ell^* + 1]$, and for each $\mathbf{E} \in \mathbb{H}^r(\Omega) \cap \mathbb{L}^2_{\mathrm{tr}}(\Omega)$, there holds

$$\operatorname{dist}\Big(\mathbf{E}, \mathcal{H}_{1,h}\Big) \,:=\, \inf_{\mathbf{E}_h \in \mathcal{H}_{1,h}} \|\mathbf{E} - \mathbf{E}_h\|_{0,\Omega} \,\leq\, C\, h^r \|\mathbf{E}\|_{r,\Omega}\,,$$

$\mathbf{AP}\Big(\mathcal{H}_{2,h}\Big)$ there exists a positive constant C, independent of h, such that for each $r \in (0, \ell + 1]$, and for each $\boldsymbol{\tau} \in \mathbb{H}^r(\Omega) \cap \mathbb{H}_0(\mathbf{div}_{4/3}; \Omega)$ with $\mathbf{div}(\boldsymbol{\tau}) \in \mathbf{W}^{r,4/3}(\Omega)$, there holds

$$\operatorname{dist}\Big(\boldsymbol{\tau}, \mathcal{H}_{2,h}\Big) \,:=\, \inf_{\boldsymbol{\tau}_h \in \mathcal{H}_{2,h}} \|\boldsymbol{\tau} - \boldsymbol{\tau}_h\|_{\mathbf{div}_{4/3};\Omega} \,\leq\, C\, h^r \left\{ \|\boldsymbol{\tau}\|_{r,\Omega} \,+\, \|\mathbf{div}(\boldsymbol{\tau})\|_{r,4/3;\Omega} \right\},$$

$\mathbf{AP}\Big(\mathcal{Q}_{1,h}\Big)$ there exists a positive constant C, independent of h, such that for each $r \in [0, \ell + 1]$, and for each $\mathbf{v} \in \mathbf{W}^{r,4}(\Omega)$, there holds

$$\operatorname{dist}\Big(\mathbf{v}, \mathcal{Q}_{1,h}\Big) \,:=\, \inf_{\mathbf{v}_h \in \mathcal{Q}_{1,h}} \|\mathbf{v} - \mathbf{v}_h\|_{0,4;\Omega} \,\leq\, C\, h^r \|\mathbf{v}\|_{r,4;\Omega}\,,$$

$\mathbf{AP}\Big(\mathcal{Q}_{2,h}\Big)$ there exists a positive constant C, independent of h, such that for each $r \in [0, \ell + 1]$, and for each $\boldsymbol{\xi} \in \mathbb{H}^r(\Omega) \cap \mathbb{L}^2_{\mathbf{sk}}(\Omega)$, there holds

$$\operatorname{dist}\Big(\boldsymbol{\xi}, \mathcal{Q}_{2,h}\Big) \,:=\, \inf_{\boldsymbol{\xi}_h \in \mathcal{Q}_{2,h}} \|\boldsymbol{\xi} - \boldsymbol{\xi}_h\|_{0,\Omega} \,\leq\, C\, h^r \|\boldsymbol{\xi}\|_{r,\Omega}\,.$$

As a consequence of the Céa estimate (3.122) (cf. Theorem 3.5.6), along with $\mathbf{AP}\Big(\mathcal{H}_{1,h}\Big)$, $\mathbf{AP}\Big(\mathcal{H}_{2,h}\Big)$, $\mathbf{AP}\Big(\mathcal{Q}_{1,h}\Big)$, and $\mathbf{AP}\Big(\mathcal{Q}_{2,h}\Big)$, we are now able to provide the main result of this section.

Theorem 3.6.7. In addition to the notations and hypotheses of Theorem 3.5.6, assume that there exists $r \in (0, \ell+1]$, such that $\mathbf{D} \in \mathbb{H}^r(\Omega) \cap \mathbb{L}^2_{\mathrm{tr}}(\Omega)$, $\boldsymbol{\sigma} \in \mathbb{H}^r(\Omega) \cap \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)$, $\mathbf{div}(\boldsymbol{\sigma}) \in \mathbf{W}^{r,4/3}(\Omega)$, $\mathbf{u} \in \mathbf{W}^{r,4}(\Omega)$, and $\boldsymbol{\gamma} \in \mathbb{H}^r(\Omega) \cap \mathbb{L}^2_{\mathtt{sk}}(\Omega)$. Then, there exists a positive constant C, independent of h, such that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} + \|p - p_h\|_{0,\Omega} \,\leq\, C\, h^r \left\{ \|\mathbf{D}\|_{r,\Omega} + \|\boldsymbol{\sigma}\|_{r,\Omega} + \|\mathbf{div}(\boldsymbol{\sigma})\|_{r,4/3;\Omega} + \|\mathbf{u}\|_{r,4;\Omega} + \|\boldsymbol{\gamma}\|_{r,\Omega} \right\}.$$

3.7 Numerical results

In this section we consider the two pairs of finite element subspaces detailed in Section 3.6 to present three examples illustrating the performance of the mixed finite element method (3.85) on a set of quasi-uniform triangulations of the respective domains. In what follows, we refer to the corresponding sets of finite element subspaces generated by $\ell = \{0, 1\}$ as simply PEERS$_\ell$ and AFW$_\ell$ based discretizations. The numerical methods have been implemented using the open source finite element library `FEniCS` [52]. We solve approximately the nonlinear problem (3.85) by means of a strategy combining a Picard iteration with the Newton method. More precisely, the corresponding computations are described as follows: [a),leftmargin=*]

(1) Start solving the Stokes problem arising from (3.85) by choosing $\eta = 1$ and the overall density $\rho = 0$ to obtain the initial solution $(\mathbf{D}^0_h, \boldsymbol{\sigma}^0_h, \vec{\mathbf{u}}^0_h) := \big(\mathbf{D}^0_h, \boldsymbol{\sigma}^0_h, (\mathbf{u}^0_h, \gamma^0_h)\big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$, compute p^0_h as in (3.88), that is

$$p^0_h \,:=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma}^0_h + \rho\,(\mathbf{u}^0_h \otimes \mathbf{u}^0_h)\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\mathbf{u}^0_h \otimes \mathbf{u}^0_h)\,,$$

and let $m = 1$.

(2) Set $(\mathbf{z}_h, r_h) := (\mathbf{u}^{m-1}_h, p^{m-1}_h)$ and let $(\mathbf{D}^m_h, \boldsymbol{\sigma}^m_h, \vec{\mathbf{u}}^m_h) := \big(\mathbf{D}^m_h, \boldsymbol{\sigma}^m_h, (\mathbf{u}^m_h, \gamma^m_h)\big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ be the output of a single Newton iteration applied to (3.87).

(3) Update the pressure p^m_h by employing the formula (3.88), namely

$$p^m_h \,:=\, -\frac{1}{n}\,\mathrm{tr}\Big(\boldsymbol{\sigma}^m_h + \rho\,(\mathbf{u}^m_h \otimes \mathbf{u}^m_h)\Big) \,+\, \frac{\kappa}{|\Omega|} \,+\, \frac{\rho}{n\,|\Omega|}\int_\Omega \mathrm{tr}(\mathbf{u}^m_h \otimes \mathbf{u}^m_h)\,,$$

let $m = m + 1$, and go to step (2).

The iterative procedure given by (2) and (3) is finished when the relative error between two consecutive iterations of the complete coefficient vector, namely $\mathbf{coeff}^m$

and $\mathbf{coeff}^{m+1}$, is sufficiently small, that is,

$$\frac{\|\mathbf{coeff}^{m+1} - \mathbf{coeff}^{m}\|_{\mathtt{DOF}}}{\|\mathbf{coeff}^{m+1}\|_{\mathtt{DOF}}} \leq \mathsf{tol}\,,$$

where $\|\cdot\|_{\mathtt{DOF}}$ stands for the usual Euclidean norm in $\mathrm{R}^{\mathtt{DOF}}$ with $\mathtt{DOF}$ denoting the total number of degrees of freedom defining the finite element subspaces $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ (cf. (3.125)–(3.126)), and tol is a fixed tolerance chosen as $\mathsf{tol} = 1E-06$.

We now introduce some additional notation. The individual errors are denoted by

$$\mathsf{e}(\mathbf{D}) := \|\mathbf{D} - \mathbf{D}_h\|_{0,\Omega}\,, \quad \mathsf{e}(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div}_{4/3};\Omega}\,, \quad \mathsf{e}(\mathbf{u}) := \|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega}\,,$$
$$\mathsf{e}(\boldsymbol{\gamma}) := \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,\Omega}\,, \quad \mathsf{e}(p) := \|p - p_h\|_{0,\Omega}\,,$$

and, as usual, for each $\star \in \Big\{\mathbf{D}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\gamma}, p\Big\}$ we let $\mathsf{r}(\star)$ be the experimental rate of convergence given by $\mathsf{r}(\star) := \log(\mathsf{e}(\star)/\widehat{\mathsf{e}}(\star))/\log(h/\widehat{h})$, where h and $\widehat{h}$ denote two consecutive meshsizes with errors e and $\widehat{\mathsf{e}}$, respectively.

The examples to be considered in this section are described next, for which we consider the regularized viscosity $\eta(\varrho,\omega)$ defined by (3.11), but without needing to make use of the modification described by Figure A.1. In the first two examples, for the sake of simplicity, we take $\mu_s = 0.1$, $\mu_d = 1$, $I_0 = 1$, $d = 1$ and $\rho = 1$. In addition, the null mean value of $\mathrm{tr}(\boldsymbol{\sigma}_h)$ over Ω is fixed via a real Lagrange multiplier strategy.

Example 3.1: Convergence against smooth exact solutions in a 2D domain

In this test we corroborate the rates of convergence in a two-dimensional domain set by the square $\Omega = (0,1)^2$. We choose the regularization factor $\varepsilon = 1E-08$, and adjust the datum $\mathbf{f}$ in (3.19) such that the exact solution is given by

$$\mathbf{u}(x_1,x_2) = \begin{pmatrix} \sin(x_1)\cos(x_2) \\ -\cos(x_1)\sin(x_2) \end{pmatrix} \quad \text{and} \quad p(x_1,x_2) = \exp(x_1+x_2), \tag{3.127}$$

where $p \in \mathrm{L}^2_\kappa(\Omega)$, with $\kappa = (\exp(1)-1)^2$. The model problem is then complemented with the appropriate Dirichlet boundary condition. Tables 3.1 and 3.2 show the convergence history for a sequence of quasi-uniform mesh refinements, including the number of Newton iterations. As already announced, we stress that we are able not only to approximate the original unknowns but also the pressure field through the formula (3.88). The results confirm that the optimal rates of convergence $\mathcal{O}(h^{\ell+1})$

predicted by Theorem 3.6.7 are attained for $\ell = \{0, 1\}$ for both PEERS_ℓ and AFW_ℓ based schemes. The Newton method exhibits a behavior dependent on the mesh size, converging faster for finer meshes in both discrete schemes. The latter is justified by the fact that for finer mesh a better initial data $(\mathbf{D}_h^0, \mathbf{u}_h^0)$ and p_h^0 are provided for the iterative method. In Figure 3.1 we display the discrete internal friction coefficient $\mu(I_h)$ recovered from (3.7), with $I_h = \sqrt{2}\, d\, |\mathbf{D}_h| / \sqrt{p_h/\rho}$, and some solutions obtained with the mixed PEERS_1 approximation with meshsize $h = 0.014$ and $20,000$ triangle elements (actually representing $1,081,202$ DOF).

PEERS_ℓ–based discretization with $\ell = 0$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(γ)	r(γ)	e(p)	r(p)
842	0.354	16	3.15e-01	–	1.14e+00	–	7.84e-02	–	1.08e-01	–	4.27e-01	–
3314	0.177	14	1.87e-01	0.750	5.53e-01	1.044	3.70e-02	1.085	4.58e-02	1.236	1.95e-01	1.131
13154	0.088	13	1.00e-01	0.905	2.67e-01	1.051	1.78e-02	1.057	1.74e-02	1.393	8.91e-02	1.130
46082	0.047	11	5.44e-02	0.969	1.40e-01	1.026	9.35e-03	1.021	6.83e-03	1.491	4.55e-02	1.069
183962	0.024	9	2.74e-02	0.991	6.95e-02	1.011	4.65e-03	1.006	2.38e-03	1.521	2.23e-02	1.029
510602	0.014	8	1.65e-02	0.997	4.16e-02	1.004	2.79e-03	1.002	1.09e-03	1.526	1.33e-02	1.012

AFW_ℓ–based discretization with $\ell = 0$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(γ)	r(γ)	e(p)	r(p)
609	0.354	15	5.62e-02	–	5.63e-01	–	6.94e-02	–	6.76e-02	–	3.27e-01	–
2369	0.177	14	2.65e-02	1.086	2.80e-01	1.005	3.48e-02	0.995	3.34e-02	1.018	1.63e-01	1.000
9345	0.088	12	1.30e-02	1.027	1.40e-01	1.002	1.74e-02	0.999	1.66e-02	1.006	8.17e-02	1.000
32641	0.047	10	6.89e-03	1.008	7.46e-02	1.001	9.29e-03	1.000	8.85e-03	1.002	4.36e-02	1.000
130081	0.024	7	3.44e-03	1.002	3.73e-02	1.001	4.65e-03	1.000	4.42e-03	1.001	2.18e-02	1.000
360801	0.014	6	2.06e-03	1.001	2.24e-02	1.001	2.79e-03	1.000	2.65e-03	1.000	1.31e-02	1.000

Table 3.1 [Example 3.1, $\ell = 0$] Number of degrees of freedom, meshsizes, Newton iteration count, errors, and rates of convergence for the mixed approximations.

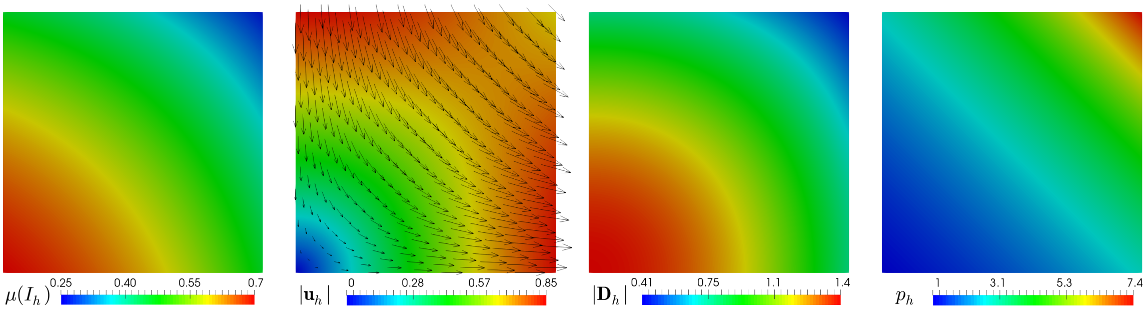


Figure 3.1 [Example 3.1] Computed internal friction coefficient, magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

PEERS$_\ell$–based discretization with $\ell = 1$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e($\boldsymbol{\gamma}$)	r($\boldsymbol{\gamma}$)	e(p)	r(p)
1778	0.354	12	1.80e-02	–	4.59e-02	–	4.59e-03	–	7.45e-03	–	1.84e-02	–
7010	0.177	10	5.36e-03	1.750	1.17e-02	1.970	1.15e-03	1.999	3.12e-03	1.257	4.51e-03	2.031
27842	0.088	8	1.48e-03	1.858	2.98e-03	1.977	2.87e-04	2.001	9.81e-04	1.668	1.12e-03	2.006
97562	0.047	6	4.42e-04	1.922	8.56e-04	1.983	8.15e-05	2.000	3.09e-04	1.840	3.19e-04	1.998
389522	0.024	4	1.14e-04	1.958	2.16e-04	1.990	2.04e-05	2.000	8.16e-05	1.919	8.00e-05	1.997
1081202	0.014	4	4.14e-05	1.977	7.78e-05	1.993	7.34e-06	2.000	3.00e-05	1.957	2.88e-05	1.998

AFW$_\ell$–based discretization with $\ell = 1$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e($\boldsymbol{\gamma}$)	r($\boldsymbol{\gamma}$)	e(p)	r(p)
1393	0.354	10	2.21e-03	–	2.49e-02	–	4.57e-03	–	2.84e-03	–	1.73e-02	–
5473	0.177	7	5.35e-04	2.046	6.12e-03	2.027	1.15e-03	1.997	7.29e-04	1.963	4.33e-03	1.996
21697	0.088	5	1.32e-04	2.020	1.52e-03	2.013	2.87e-04	1.999	1.84e-04	1.983	1.08e-03	1.999
75961	0.047	4	3.73e-05	2.009	4.29e-04	2.008	8.15e-05	2.000	5.27e-05	1.992	3.08e-04	2.000
303121	0.024	3	9.29e-06	2.007	1.07e-04	2.008	2.04e-05	2.000	1.32e-05	1.997	7.70e-05	2.000
841201	0.014	3	3.34e-06	2.002	3.84e-05	2.002	7.34e-06	2.000	4.76e-06	1.998	2.77e-05	2.000

Table 3.2 [Example 3.1, $\ell = 1$] Number of degrees of freedom, meshsizes, Newton iteration count, errors, and rates of convergence for the mixed approximations.

Example 3.2: Convergence against smooth exact solutions in a 3D domain

In the second example we consider the cube domain $\Omega = (0,1)^3$, and the regularization factor $\varepsilon = 1E - 06$. The manufactured solution is given by

$$\mathbf{u}(x_1,x_2,x_3) = \begin{pmatrix} \sin(x_1)\cos(x_2)\cos(x_3) \\ -2\cos(x_1)\sin(x_2)\cos(x_3) \\ \cos(x_1)\cos(x_2)\sin(x_3) \end{pmatrix} \quad \text{and} \quad p(x_1,x_2,x_3) = 10\,\exp(x_1+x_2+x_3)\,,$$

where $p \in \mathrm{L}^2_\kappa(\Omega)$, with $\kappa = 10\,(\exp(1) - 1)^3$. Similarly to the first example, the data $\mathbf{f}$ and $\mathbf{u}_D$ is computed from (3.19) using the above solution. The convergence history for a set of quasi-uniform mesh refinements using $\ell = 0$ is shown in Table 3.3. Again, the mixed finite element method converges optimally with order $\mathcal{O}(h)$, as it was proved by Theorem 3.6.7. We observe a considerable increasing of degrees of freedom in the PEERS$_0$-based scheme compared to the AFW$_0$ one. This is justified mainly by the fact that the symmetric part of the velocity gradient is approximated with $\mathbb{P}_3(\Omega)$ and $\mathbb{P}_1(\Omega)$, respectively. In addition, the discrete internal friction coefficient and some components of the numerical solution are displayed in Figure 3.2, which were built using the mixed AFW$_0$ approximation with meshsize $h = 0.108$ and $24,576$ tetrahedral elements (actually representing $1,390,081$ DOF).

PEERS$_\ell$-based discretization with $\ell = 0$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(γ)	r(γ)	e(p)	r(p)
8698	0.866	20	9.95e-01	–	5.05e+01	–	2.41e-01	–	6.04e-01	–	1.66e+01	–
69016	0.433	20	5.85e-01	0.766	2.73e+01	0.885	1.10e-01	1.135	2.28e-01	1.406	9.01e+00	0.884
550156	0.217	19	3.24e-01	0.852	1.36e+01	1.008	4.96e-02	1.143	7.95e-02	1.520	4.32e+00	1.061
1854688	0.144	18	2.23e-01	0.926	8.89e+00	1.046	3.19e-02	1.091	4.17e-02	1.590	2.74e+00	1.122
4393876	0.108	18	1.69e-01	0.957	6.59e+00	1.045	2.35e-02	1.056	2.62e-02	1.625	1.99e+00	1.113

AFW$_\ell$–based discretization with $\ell = 0$												
DOF	h	it	e($\mathbf{D}$)	r($\mathbf{D}$)	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(γ)	r(γ)	e(p)	r(p)
2905	0.866	12	2.09e-01	–	2.59e+01	–	1.78e-01	–	2.01e-01	–	1.43e+01	–
22369	0.433	11	8.24e-02	1.344	1.21e+01	1.104	9.12e-02	0.969	9.34e-02	1.103	7.15e+00	1.005
175489	0.217	9	3.56e-02	1.212	5.82e+00	1.050	4.59e-02	0.992	4.55e-02	1.037	3.57e+00	1.002
588385	0.144	8	2.28e-02	1.097	3.85e+00	1.023	3.06e-02	0.997	3.02e-02	1.014	2.38e+00	1.001
1390081	0.108	7	1.68e-02	1.054	2.87e+00	1.013	2.30e-02	0.999	2.26e-02	1.007	1.78e+00	1.000

Table 3.3 [Example 3.2, $\ell = 0$] Number of degrees of freedom, meshsizes, Newton iteration count, errors, and rates of convergence for the mixed approximations.

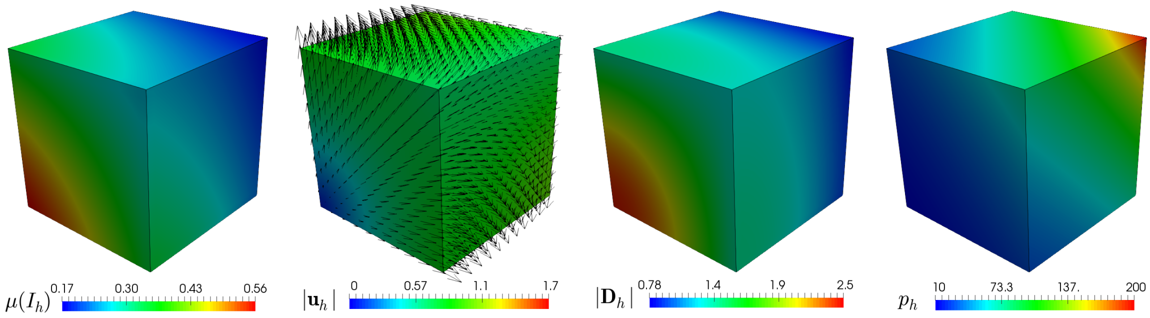


Figure 3.2 [Example 3.2] Computed internal friction coefficient, magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

Example 3.3: Fluid flow through a cavity 2D with two circular obstacles

In the last example, motivated by [69, Section 2.1], we study the behavior of the regularized $\mu(I)$-rheology model of granular materials for fluid flow through a cavity 2D with two circular obstacles without manufactured solution. More precisely, we consider the domain $\Omega = (0,1)^2 \setminus (\Omega_1 \cup \Omega_2)$, where

$$\Omega_1 = \left\{(x_1, x_2) : \ (x_1 - 1/2)^2 + (x_2 - 1/3)^2 < 0.1^2\right\}, \text{ and}$$

$$\Omega_2 = \left\{(x_1, x_2) : \ (x_1 - 1/2)^2 + (x_2 - 2/3)^2 < 0.1^2\right\},$$

with boundary Γ, whose part around the circles is given by $\Gamma_c = \partial\Omega_c$. The model parameters are chosen as $\mu_s = 0.36, \mu_d = 0.91, I_0 = 0.73, d = 0.05, \rho = 2500$, and the regularization factor is $\varepsilon = 1E - 03$. Notice that the relation between the diameter of the particles d and the width of the cavity is $1 : 20$, whereas the radius of both circular obstacles is double that of d. The mean value of p is fixed as $\kappa = 100$, no presence of gravity is assumed, that is, $\mathbf{f} = \mathbf{0}$, and the boundaries conditions are

$$\mathbf{u} = (0.2\, x_2 - 0.1, 0)^{\mathrm{t}} \quad \text{on} \quad \Gamma \setminus \Gamma_c \quad \text{and} \quad \mathbf{u} = \mathbf{0} \quad \text{on} \quad \Gamma_c\,.$$

In particular, we impose that flows cannot go in nor out through Γ_c, whereas at the top and bottom of the domain flows are faster in opposite direction. In Figure 3.3, we display the computed internal friction coefficient, magnitude of the velocity and symmetric part of the velocity gradient, and pressure field, which were built using the mixed AFW$_0$-based scheme on a mesh with meshsize $h = 0.016$ and $18,423$ triangle elements (actually representing $332,573$ DOF). We observe higher velocities at the top and bottom of the boundary going to the right and left of the domain, respectively, as we expected, but also a circulation phenomenom on the left and right boundaries since the flows cannot in nor out through the circle obstacles. In addition, most of the variations in both the magnitude of the symmetric part of the velocity gradient tensor and pressure field occur around the circular obstacles. This observation aligns with the results obtained for the discrete internal friction coefficient. Notice also that between the circle obstacles and in some parts of the middle of the domain the magnitude of the symmetric part of the velocity gradient is zero or close to it describing a region where the original viscosity η (3.9) is singular and hence the granular flows are static. The latter is in agreement with the velocity of the fluid and it is overcome by the mixed approximation considering the regularized viscosity (3.11) as it was described in Section 3.2.

Example 3.4: Fluid flow in a cubic lid-driven cavity

Finally, we conduct a simulation of the 3D lid-driven cavity flow within a unit cube $\Omega = (0,1)^3$. On the top lid $x_3 = 1$, the tangential velocity is set as $\mathbf{u} = (1,0,0)^{\mathrm{t}}$, while the rest of the boundary has no-slip conditions. The model parameters are chosen as $\mu_s = 0.1, \mu_d = 1, I_0 = 1, d = 1, \rho = 1$, and the regularization factor is $\varepsilon = 1E - 03$. The mean value of p is fixed at $\kappa = 1000$, and the right-hand side is set as $\mathbf{f} = \mathbf{0}$. The numerical results, displayed in Figure 3.4, show the computed internal friction coefficient, the magnitude of the velocity, the symmetric part of the velocity gradient,

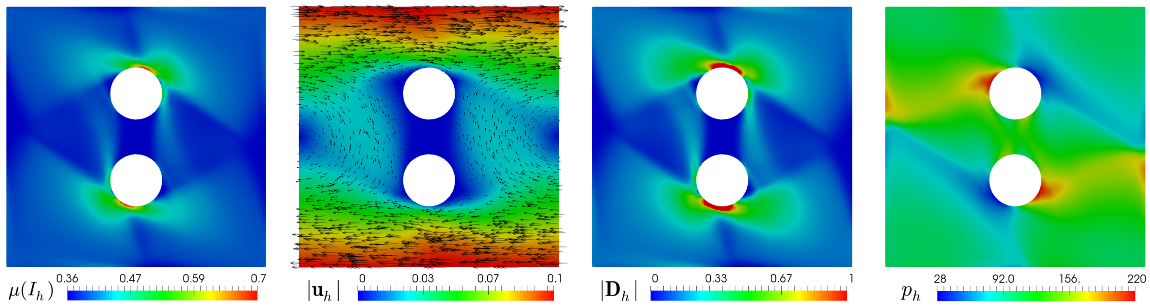


Figure 3.3 [Example 3.3] Computed internal friction coefficient, magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

and the pressure field. These were obtained using the mixed AFW$_0$-based scheme on a mesh with a mesh size of $h = 0.108$ and $24,576$ tetrahedral elements (representing $1,390,081,$ DOF). A large-scale recirculation, influenced by the momentum transfer from the top surface to the rest of the fluid, is observed. Most variations in both the magnitude of the symmetric part of the velocity gradient tensor and the pressure field occur near the top of the cube, which aligns with the results obtained for the discrete internal friction coefficient. Additionally, below the top of the domain, the magnitude of the symmetric part of the velocity gradient is zero or close to it, indicating a large region where the original viscosity η (3.9) becomes singular, rendering the granular flows static. Similarly to Example 3, this issue is addressed by the mixed approximation through the incorporation of the regularized viscosity (3.11).

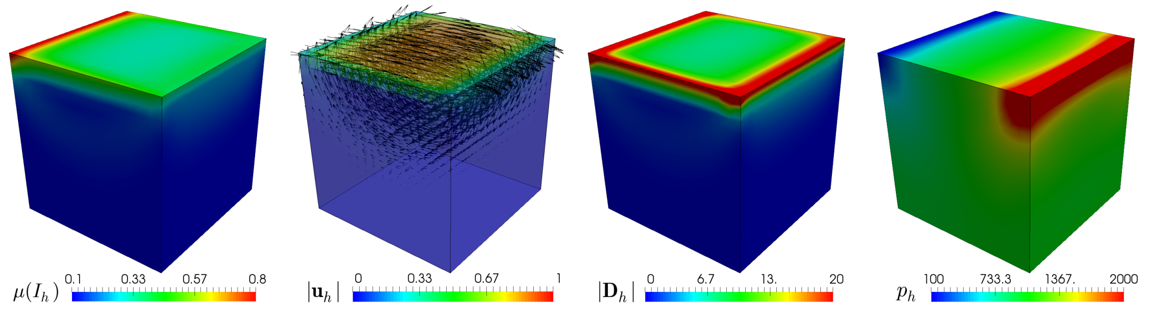


Figure 3.4 [Example 3.4] Computed internal friction coefficient, magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

Chapter 4

A posteriori error analysis for $\mu(I)$-rheology

4.1 Introduction

Building upon the previous discussion in Introduction and extending the study initiated in [43] on a regularized $\mu(I)$-rheology model for granular materials described by a Navier–Stokes-like equation, this chapter employs and adapts the *a posteriori* error analysis techniques developed in [38], [39], [41], and [42] for mixed formulations in Hilbert and Banach spaces to the current $\mu(I)$-rheology model. We construct a reliable and efficient residual-based *a posteriori* error estimator for the 2D and 3D versions of the mixed finite element methods introduced in [43]. Specifically, we derive a global quantity Θ that is formulated in terms of computable local indicators Θ_K, each associated with an element K of a given triangulation $\mathcal{T}$. This allows for the identification of error sources and the design of an adaptive meshing algorithm to enhance computational efficiency. In this setting, the estimator Θ is considered efficient (resp. reliable) if there exist positive constants $C_{\mathtt{eff}}$ (resp. $C_{\mathtt{rel}}$), independent of the mesh sizes, such that

$$C_{\mathtt{eff}}\,\Theta + \mathtt{h.o.t.} \;\leq\; \|\text{error}\| \;\leq\; C_{\mathtt{rel}}\,\Theta + \mathtt{h.o.t.}\,,$$

where $\mathtt{h.o.t.}$ represents one or more higher-order terms. To the best of the authors' knowledge, this work presents the first *a posteriori* error analysis of Banach space-based mixed finite element methods for the stationary $\mu(I)$-rheology equations governing granular materials.

This chapter is organized as follows. In Section 4.2, we provide a detailed derivation of a reliable and efficient residual-based *a posteriori* error estimator for the 2D version of the problem from [43]. In particular, the reliability analysis considers a suitable Helmholtz decomposition in a Banach space setting, with its discrete version employing PEERS and AFW-based elements. Several numerical results illustrating the reliability and efficiency of the estimator, the effectiveness of the associated adaptive algorithm, and the recovery of optimal convergence rates are reported in Section 4.3. Finally, additional properties required for the derivation of the reliability and efficiency estimates are provided in Appendices A.2 and A.3, respectively. In turn, the 3D version of the *a posteriori* error estimator, building upon the results in Section 4.2, is established in Appendix A.4.

4.2 A residual-based a posteriori error estimator

In this section, we derive a reliable and efficient residual-based *a posteriori* error estimator for the two-dimensional version of the Galerkin scheme (3.85). The corresponding *a posteriori* error analysis for the three-dimensional case, which follows from minor modifications of the analysis presented here, will be addressed in Appendix A.4. Throughout this section, we employ the notations and results from Appendix A.2.

Recalling that $\big(\mathbf{D}_h, \boldsymbol{\sigma}_h, (\mathbf{u}_h, \gamma_h)\big) \in \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ is the unique solution of the discrete problem (3.85), and that p_h is computed from (3.88), we define the global *a posteriori* error estimator Θ as

$$\Theta \,=\, \left\{ \sum_{K\in\mathcal{T}_h} \Theta_{1,K}^{4/3} \right\}^{3/4} + \left\{ \sum_{K\in\mathcal{T}_h} \Theta_{2,K}^{2} \right\}^{1/2} + \left\{ \sum_{K\in\mathcal{T}_h} \Theta_{3,K}^{4} \right\}^{1/4}, \tag{4.1}$$

where, for each $K \in \mathcal{T}_h$, the local error indicators $\Theta_{1,K}^{4/3}$, $\Theta_{2,K}^{2}$ and $\Theta_{3,K}^{4}$ are defined as

$$\Theta_{1,K}^{4/3} \,:=\, \Big\| \mathbf{f} + \mathbf{div}(\boldsymbol{\sigma}_h) \Big\|_{0,4/3;K}^{4/3}, \tag{4.2}$$

$$\begin{aligned} \Theta_{2,K}^{2} \,:=\, & \Big\| \eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho(\mathbf{u}_h \otimes \mathbf{u}_h)^{\mathtt{d}} \Big\|_{0,K}^{2} + \Big\| \boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathtt{t}} \Big\|_{0,K}^{2} \\ & + h_K^2 \Big\| \mathbf{rot}\,(\mathbf{D}_h + \gamma_h) \Big\|_{0,K}^{2} + \sum_{e\in\mathcal{E}_h(K)\cap\mathcal{E}(\Omega)} h_e \Big\| [\![\big(\mathbf{D}_h + \gamma_h\big)\mathbf{s}]\!] \Big\|_{0,e}^{2} \\ & + \sum_{e\in\mathcal{E}_h(K)\cap\mathcal{E}(\Gamma)} h_e \Big\| \nabla \mathbf{u}_D\, \mathbf{s} - \big(\mathbf{D}_h + \gamma_h\big)\mathbf{s} \Big\|_{0,e}^{2}, \end{aligned} \tag{4.3}$$

and

$$\Theta_{3,K}^4 := h_K^4 \left\| \nabla \mathbf{u}_h - \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h \Big) \right\|_{0,4;K}^4 + \sum_{e \in \mathcal{E}_h(K) \cap \mathcal{E}(\Gamma)} h_e \, \|\mathbf{u}_D - \mathbf{u}_h\|_{0,4;e}^4 . \tag{4.4}$$

Notice that the last term defining $\Theta_{2,K}^2$ (cf. (4.3)) requires that $(\nabla \mathbf{u}_D \, \mathbf{s})|_e \in \mathbf{L}^2(e)$ for all $e \in \mathcal{E}_h(\Gamma)$, which is guaranteed by simply assuming that $\mathbf{u}_D \in \mathbf{H}^1(\Gamma)$. Nevertheless, to be more precise, it suffices to assume that $\nabla \mathbf{u}_D|_\Gamma \in \mathbb{L}^2(\Gamma)$, which holds if $\nabla \mathbf{u}_D|_\Gamma$ coincides with the trace of the gradient of a function in $\mathbf{H}^t(\Omega)$ for some $t > 4/3$. In any case, the Dirichlet data used in the numerical results reported below in Section 4.3 satisfy the first-mentioned assumptions on $\mathbf{u}_D$.

From now on, we define

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} := \|\mathbf{D} - \mathbf{D}_h\|_{\mathcal{H}_1} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathcal{H}_2} + \|\vec{\mathbf{u}} - \vec{\mathbf{u}}_h\|_{\mathcal{Q}} ,$$

where $\vec{\mathbf{D}} := \Big(\mathbf{D}, \boldsymbol{\sigma}, \vec{\mathbf{u}}\Big) \in \mathcal{H} := \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{Q}$ and $\vec{\mathbf{D}}_h := \Big(\mathbf{D}_h, \boldsymbol{\sigma}_h, \vec{\mathbf{u}}_h\Big) \in \mathcal{H}_h := \mathcal{H}_{1,h} \times \mathcal{H}_{2,h} \times \mathcal{Q}_h$ denote the unique solutions of (3.35) and (3.85), respectively. The main goal of this section is to establish, under suitable assumptions, the existence of positive constants $C_{\texttt{eff}}$ and $C_{\texttt{rel}}$, independent of the mesh sizes and the continuous and discrete solutions, such that

$$C_{\texttt{eff}} \, \Theta + \texttt{h.o.t} \leq \|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} + \|p - p_h\|_{0,\Omega} \leq C_{\texttt{rel}} \, \Theta , \tag{4.5}$$

where $\texttt{h.o.t}$ is a generic expression denoting one or several terms of higher order, whereas p and p_h are computed according to (3.44) and (3.88), respectively. The upper and lower bounds in (4.5), which are known as the reliability and efficiency of Θ, are derived below in Sections 4.2.1 and 4.2.2, respectively.

4.2.1 Reliability

The main result of this section is stated in the following theorem. To this end, and as done in [43, eq. (5.19)], given $r \in \mathrm{L}^2_\kappa(\Omega)$, we first note that we can define the operator $\Xi_r : \mathcal{H} \to \mathcal{H}'$, which arises from the left-hand side of the variational formulation (3.35) after summing all its rows, that is,

$$[\Xi_r(\vec{\mathbf{C}}), \vec{\mathbf{E}}] := [\mathcal{A}_r(\mathbf{C}), \mathbf{E}] + \mathcal{B}_1(\mathbf{E}, \boldsymbol{\zeta}) + \mathcal{B}_1(\mathbf{C}, \boldsymbol{\tau}) + \mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{w}}) + \mathcal{B}(\boldsymbol{\zeta}, \vec{\mathbf{v}}) , \tag{4.6}$$

for all $\vec{\mathbf{C}} := (\mathbf{C}, \boldsymbol{\zeta}, \vec{\mathbf{w}})$, $\vec{\mathbf{E}} := (\mathbf{E}, \boldsymbol{\tau}, \vec{\mathbf{v}}) \in \mathcal{H}$, so that (3.35) can be rewritten as

$$[\Xi_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}] \,=\, \mathcal{F}_{\mathbf{u}}(\mathbf{E}) + \mathcal{G}(\boldsymbol{\tau}) + \mathcal{F}(\vec{\mathbf{v}}) \quad \forall\, \vec{\mathbf{E}} \in \mathcal{H}\,. \tag{4.7}$$

Thus, the smoothness of the regularized function η (cf. (3.11)) allows to show that for each $r \in \mathrm{L}^2_\kappa(\Omega)$, the operator $\mathcal{A}_r$ (cf. (3.36)), and hence Ξ_r as well, have first order Gâteaux derivatives $\mathcal{D}(\mathcal{A}_r) \in \mathcal{L}\Big(\mathcal{H}_1, \mathcal{L}(\mathcal{H}_1, \mathcal{H}_1')\Big)$ and $\mathcal{D}(\Xi_r) \in \mathcal{L}\Big(\mathcal{H}, \mathcal{L}(\mathcal{H}, \mathcal{H}')\Big)$, respectively. Moreover, using [43, eqs. (4.9) and (4.10) in Lemma 4.2], one is able to prove (see, e.g. [63, Lemma 3.1]) that for each $\mathbf{C} \in \mathcal{H}_1$, the operator $\mathcal{D}(\mathcal{A}_r)(\mathbf{C}) \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_1')$ can be identified as a bounded and $\mathcal{H}_1$-elliptic bilinear form with constants $L_{\mathcal{A}}$ and $\alpha_{\mathcal{A}}$, respectively. It follows that for each $r \in \mathrm{L}^2_\kappa(\Omega)$, and for each $\vec{\mathbf{C}} \in \mathcal{H}$, the operator $\mathcal{D}(\Xi_r)(\vec{\mathbf{C}}) \in \mathcal{L}(\mathcal{H}, \mathcal{H}')$ satisfies the hypotheses of the linear version of [43, Theorem 4.1], and hence, there exists a positive constant $\alpha_{\boldsymbol{\Xi}}$, depending only on $L_{\mathcal{A}}$, $\alpha_{\mathcal{A}}$, and the inf-sup constants of $\mathcal{B}$ and $\mathcal{B}_1$, namely $\widetilde{\beta}$ and $\widetilde{\beta}_1$ (cf. [43, eqs. (4.12), (4.13)]), such that the following global inf-sup condition holds:

$$\alpha_{\boldsymbol{\Xi}}\, \|\vec{\mathbf{F}}\|_{\mathcal{H}} \,\leq\, \sup_{\mathbf{0} \neq \vec{\mathbf{E}} \in \mathcal{H}} \frac{\mathcal{D}(\Xi_r)(\vec{\mathbf{C}})(\vec{\mathbf{F}}, \vec{\mathbf{E}})}{\|\vec{\mathbf{E}}\|_{\mathcal{H}}} \qquad \forall\, \vec{\mathbf{F}} \in \mathcal{H}\,. \tag{4.8}$$

In addition, we let

$$C_{1,\boldsymbol{\Xi}} \,:=\, \alpha_{\boldsymbol{\Xi}}^{-1}\, n^{-1/2} \quad \text{and} \quad C_{2,\boldsymbol{\Xi}} \,:=\, \alpha_{\boldsymbol{\Xi}}^{-1}\, \rho \left(2\, n^{-1/2}\, L_\eta + 1\right), \tag{4.9}$$

where $\alpha_{\boldsymbol{\Xi}}$ satisfies (4.8), and L_η denotes the Lipschitz continuity constant of η (cf. [43, eq. (4.8)]).

The aforementioned result is stated now.

Theorem 4.2.1. Assume that L_η and the radii δ and $\delta_{\mathtt{d}}$ are sufficiently small so that

$$C_{1,\boldsymbol{\Xi}}\, L_\eta \,\leq\, \frac{1}{2} \quad \text{and} \quad C_{2,\boldsymbol{\Xi}}\, n^{1/2} \left(\delta + \delta_{\mathtt{d}}\right) \,\leq\, \frac{1}{2}\,. \tag{4.10}$$

Then, there exists a constant $C_{\mathtt{rel}} > 0$, such that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,+\, \|p - p_h\|_{0,\Omega} \,\leq\, C_{\mathtt{rel}}\, \Theta\,. \tag{4.11}$$

We begin the proof of Theorem 4.2.1 with a preliminary lemma. Specifically, proceeding analogously to [42, Section 3.1] (see also [70, Section 1]), we first introduce the residual functional $\mathcal{R} : \mathcal{H} \to \mathrm{R}$, given by

$$\mathcal{R}(\vec{\mathbf{E}}) := \mathcal{R}_1(\mathbf{E}) + \mathcal{R}_2(\boldsymbol{\tau}) + \mathcal{R}_3(\vec{\mathbf{v}}) \quad \forall\, \vec{\mathbf{E}} = (\mathbf{E}, \boldsymbol{\tau}, \vec{\mathbf{v}}) \in \mathcal{H}\,, \tag{4.12}$$

where $\mathcal{R}_1 : \mathcal{H}_1 \to \mathrm{R}$, $\mathcal{R}_2 : \mathcal{H}_2 \to \mathrm{R}$, and $\mathcal{R}_3 : \mathcal{Q} \to \mathrm{R}$ are given by

$$\mathcal{R}_1(\mathbf{E}) \;:=\; \mathcal{F}_{\mathbf{u}_h}(\mathbf{E}) - [\mathcal{A}_{p_h}(\mathbf{D}_h), \mathbf{E}] \;-\; \mathcal{B}_1(\mathbf{E}, \boldsymbol{\sigma}_h) \quad \forall\, \mathbf{E} \in \mathcal{H}_1\,, \tag{4.13}$$

$$\mathcal{R}_2(\boldsymbol{\tau}) \;:=\; \mathcal{G}(\boldsymbol{\tau}) \;-\; \mathcal{B}_1(\mathbf{D}_h, \boldsymbol{\tau}) \;-\; \mathcal{B}(\boldsymbol{\tau}, \vec{\mathbf{u}}_h) \quad \forall\, \boldsymbol{\tau} \in \mathcal{H}_2\,, \tag{4.14}$$

and

$$\mathcal{R}_3(\vec{\mathbf{v}}) \;:=\; \mathcal{F}(\vec{\mathbf{v}}) - \mathcal{B}(\boldsymbol{\sigma}_h, \vec{\mathbf{v}}) \quad \forall\, \vec{\mathbf{v}} \in \mathcal{Q}\,, \tag{4.15}$$

respectively, which according to the discrete problem (3.85) satisfy

$$\begin{gathered} \mathcal{R}_1(\mathbf{E}_h) = 0 \quad \forall\, \mathbf{E}_h \in \mathcal{H}_{1,h}\,, \quad \mathcal{R}_2(\boldsymbol{\tau}_h) = 0 \quad \forall\, \boldsymbol{\tau}_h \in \mathcal{H}_{2,h}\,, \\ \mathcal{R}_3(\vec{\mathbf{v}}_h) = 0 \quad \forall\, \vec{\mathbf{v}}_h \in \mathcal{Q}_h\,. \end{gathered} \tag{4.16}$$

We are now in a position to establish the following aforementioned preliminary *a posteriori* error estimate.

Lemma 4.2.1. Assume that L_η and the radii δ and $\delta_{\mathbf{d}}$ satisfy (4.10). Then, there exists a positive constant C, independent of h, such that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \;+\; \|p - p_h\|_{0,\Omega} \;\leq\; C \left\{ \|\mathcal{R}_1\|_{\mathcal{H}_1'} \;+\; \|\mathcal{R}_2\|_{\mathcal{H}_2'} \;+\; \|\mathcal{R}_3\|_{\mathcal{Q}'} \right\}. \tag{4.17}$$

Proof. We begin by proceeding analogously to the proof of [63, Theorem 3.3]. In fact, given $p \in \mathrm{L}^2_\kappa(\Omega)$ satisfying (3.44) and since $\vec{\mathbf{D}}$ and $\vec{\mathbf{D}}_h$ belong to $\mathcal{H}$, a straightforward application of the mean value theorem yields the existence of a convex combination of $\vec{\mathbf{D}}$ and $\vec{\mathbf{D}}_h$, say $\vec{\mathbf{C}}_h \in \mathcal{H}$, such that

$$\mathcal{D}(\boldsymbol{\Xi}_p)(\vec{\mathbf{C}}_h)(\vec{\mathbf{D}} - \vec{\mathbf{D}}_h, \vec{\mathbf{E}}) \;=\; [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}), \vec{\mathbf{E}}] - [\boldsymbol{\Xi}_p(\vec{\mathbf{D}}_h), \vec{\mathbf{E}}] \quad \forall\, \vec{\mathbf{E}} \in \mathcal{H}\,. \tag{4.18}$$

Then, by adding and subtracting $[\boldsymbol{\Xi}_{p_h}(\vec{\mathbf{D}}_h), \vec{\mathbf{E}}]$ and $\mathcal{F}_{\mathbf{u}_h}(\mathbf{E})$ on the right-hand side of (4.18), using (4.7), and the definitions of $\boldsymbol{\Xi}_p$ and $\mathcal{R}$ (cf. (4.6), (4.12)), along with straightforward algebraic manipulations, we deduce that

$$\mathcal{D}(\boldsymbol{\Xi}_p)(\vec{\mathbf{C}}_h)(\vec{\mathbf{D}} - \vec{\mathbf{D}}_h, \vec{\mathbf{E}}) = \mathcal{R}(\vec{\mathbf{E}}) + \Big(\mathcal{F}_{\mathbf{u}} - \mathcal{F}_{\mathbf{u}_h}\Big)(\mathbf{E}) - [\mathcal{A}_p(\mathbf{D}_h) - \mathcal{A}_{p_h}(\mathbf{D}_h), \mathbf{E}] \quad \forall\, \vec{\mathbf{E}} \in \mathcal{H}\,. \tag{4.19}$$

In turn, applying (4.8) with $r = p$, $\vec{\mathbf{C}} = \vec{\mathbf{C}}_h$, and $\vec{\mathbf{F}} = \vec{\mathbf{D}} - \vec{\mathbf{D}}_h$, using (4.19) and the continuity of the operator $\mathcal{A}_p$ (cf. [43, eq. (4.11) in Lemma 4.2]), with the positive continuity constant L_η, we get

$$\alpha_{\boldsymbol{\Xi}} \, \|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,\leq\, \|\mathcal{R}\|_{\mathcal{H}'} \,+\, \|\mathcal{F}_{\mathbf{u}} - \mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}_1'} \,+\, L_\eta \|p - p_h\|_{0,\Omega} \,. \tag{4.20}$$

Next, we focus on bounding the last two terms on the right-hand side of (4.20). First, using the definition of $\mathcal{F}_{\mathbf{z}}$ (cf. (3.39)) and applying the Cauchy–Schwarz inequality, we obtain

$$\|\mathcal{F}_{\mathbf{u}} - \mathcal{F}_{\mathbf{u}_h}\|_{\mathcal{H}_1'} \,\leq\, \rho \, \|\mathbf{u} \otimes \mathbf{u} - \mathbf{u}_h \otimes \mathbf{u}_h\|_{0,\Omega} \,, \tag{4.21}$$

whereas, according to the expressions provided by (3.44) and (3.88), and proceeding similarly to [43, eq. (5.31)], the last term in (4.20) can be bounded by

$$\|p - p_h\|_{0,\Omega} \,\leq\, n^{-1/2} \Big\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} \,+\, 2\,\rho \, \|\mathbf{u} \otimes \mathbf{u} - \mathbf{u}_h \otimes \mathbf{u}_h\|_{0,\Omega} \Big\} \,. \tag{4.22}$$

Furthermore, subtracting and adding the term $(\mathbf{u} \otimes \mathbf{u}_h)$, using Cauchy–Schwarz's inequality and the fact that $\mathbf{u} \in \mathrm{W}(\delta)$ and $\mathbf{u}_h \in \mathrm{W}(\delta_{\mathtt{d}})$, there holds

$$\begin{aligned} \|\mathbf{u} \otimes \mathbf{u} - \mathbf{u}_h \otimes \mathbf{u}_h\|_{0,\Omega} \;&\leq\; n^{1/2} \Big(\|\mathbf{u}\|_{0,4;\Omega} + \|\mathbf{u}_h\|_{0,4;\Omega} \Big) \, \|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega} \\ &\leq\; n^{1/2} \Big(\delta + \delta_{\mathtt{d}} \Big) \, \|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega} \,, \end{aligned} \tag{4.23}$$

whence, combining (4.20) with (4.21), (4.22), and (4.23), and using the definition of the constants $C_{1,\boldsymbol{\Xi}}, C_{2,\boldsymbol{\Xi}}$ (cf. (4.9)), we obtain

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,\leq\, \frac{1}{\alpha_{\boldsymbol{\Xi}}} \, \|\mathcal{R}\|_{\mathcal{H}'} + C_{1,\boldsymbol{\Xi}} \, L_\eta \, \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega} + C_{2,\boldsymbol{\Xi}} \, n^{1/2} (\delta + \delta_{\mathtt{d}}) \, \|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega} \,. \tag{4.24}$$

Thus, by employing (4.10) in (4.24) and the definition of the residual $\mathcal{R}$ (cf. (4.12)) in terms of $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ (cf. (4.13), (4.14), (4.15)), we find that

$$\|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,\leq\, \frac{2}{\alpha_{\boldsymbol{\Xi}}} \Big\{ \|\mathcal{R}_1\|_{\mathcal{H}_1'} \,+\, \|\mathcal{R}_2\|_{\mathcal{H}_2'} \,+\, \|\mathcal{R}_3\|_{\mathcal{Q}'} \Big\} \,, \tag{4.25}$$

so that the corresponding estimate for $\|p - p_h\|_{0,\Omega}$ follows from (4.22), (4.23), and (4.25), thus yielding (4.17), which concludes the proof. □

Throughout the rest of this section, we provide suitable upper bounds for each one of the terms on the right-hand side of (4.17). We begin by establishing the corresponding estimates for $\|\mathcal{R}_1\|_{\mathcal{H}_1'}$ and $\|\mathcal{R}_3\|_{\mathcal{Q}'}$ (cf. (4.13) and (4.15)).

Lemma 4.2.2. There hold

$$\|\mathcal{R}_1\|_{\mathcal{H}_1'} \,\leq\, \Big\|\eta(p_h,|\mathbf{D}_h|)\,\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho\,(\mathbf{u}_h\otimes\mathbf{u}_h)^{\mathtt{d}}\Big\|_{0;\Omega} \tag{4.26}$$

and

$$\|\mathcal{R}_3\|_{\mathcal{Q}'} \,\leq\, \Big\|\mathbf{f} + \mathbf{div}(\boldsymbol{\sigma}_h)\Big\|_{0,4/3;\Omega} \,+\, \frac{1}{2}\,\Big\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathtt{t}}\Big\|_{0,\Omega}\,. \tag{4.27}$$

Proof. First, using the definition of the functionals and operators $\mathcal{R}_1$, $\mathcal{F}_{\mathbf{u}_h}$, $\mathcal{A}_{p_h}$, and $\mathcal{B}_1$ (cf. (4.13), (3.39), (3.36), (3.37)), along with the fact that $\boldsymbol{\tau}^{\mathtt{d}} : \mathbf{E} = \boldsymbol{\tau} : \mathbf{E}$, for all $\mathbf{E}\in\mathcal{H}_1$ (cf. (3.34)), and Cauchy–Schwarz's inequality, we deduce that

$$\begin{aligned}|\mathcal{R}_1(\mathbf{E})| \,&=\, \left|-\int_\Omega\Big(\eta(p_h,|\mathbf{D}_h|)\,\mathbf{D}_h \,-\, \boldsymbol{\sigma}_h^{\mathtt{d}} \,-\, \rho\,(\mathbf{u}_h\otimes\mathbf{u}_h)^{\mathtt{d}}\Big) : \mathbf{E}\right| \\ &\leq\, \Big\|\eta(p_h,|\mathbf{D}_h|)\,\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho\,(\mathbf{u}_h\otimes\mathbf{u}_h)^{\mathtt{d}}\Big\|_{0,\Omega}\,\|\mathbf{E}\|_{0,\Omega}\,,\end{aligned}$$

which yields (4.26). On the other hand, employing the definition of the functionals and bilinear form $\mathcal{R}_3$, $\mathcal{F}$, and $\mathcal{B}$ (cf. (4.15), (3.41), (3.38)), in conjunction with the decomposition of the tensor $\boldsymbol{\sigma}_h$ into

$$\boldsymbol{\sigma}_h = \frac{1}{2}\left(\boldsymbol{\sigma}_h + \boldsymbol{\sigma^{\mathtt{t}}}_h\right) + \frac{1}{2}\left(\boldsymbol{\sigma}_h - \boldsymbol{\sigma^{\mathtt{t}}}_h\right),$$

the fact that $\left(\boldsymbol{\sigma}_h + \boldsymbol{\sigma^{\mathtt{t}}}_h\right) : \boldsymbol{\xi} = \mathbf{0}$, for all $\boldsymbol{\xi}\in\mathbb{L}^2_{\mathsf{sk}}(\Omega)$, and the Cauchy–Schwarz and Hölder inequalities, we obtain

$$\begin{aligned}|\mathcal{R}_3(\vec{\mathbf{v}})| \,&=\, \left|\int_\Omega\Big(\mathbf{f} + \mathbf{div}(\boldsymbol{\sigma}_h)\Big)\cdot\mathbf{v} \,+\, \frac{1}{2}\int_\Omega\left(\boldsymbol{\sigma}_h - \boldsymbol{\sigma^{\mathtt{t}}}_h\right) : \boldsymbol{\xi}\right| \\ &\leq\, \Big\|\mathbf{f} + \mathbf{div}(\boldsymbol{\sigma}_h)\Big\|_{0,4/3;\Omega}\,\|\mathbf{v}\|_{0,4;\Omega} \,+\, \frac{1}{2}\,\Big\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathtt{t}}\Big\|_{0,\Omega}\,\|\boldsymbol{\xi}\|_{0,\Omega}\,,\end{aligned}$$

which implies (4.27) and ends the proof. □

We now turn to the derivation of the corresponding estimate for $\|\mathcal{R}_2\|_{\mathcal{H}_2'}$. To that end, we first recall from (4.16) that $\mathcal{R}_2(\boldsymbol{\tau}_h) = 0$ for all $\boldsymbol{\tau}_h\in\mathcal{H}_{2,h}$, whence in the computation of

$$\|\mathcal{R}_2\|_{\mathcal{H}_2'} \,:=\, \sup_{\mathbf{0}\neq\boldsymbol{\tau}\in\mathcal{H}_2}\frac{\mathcal{R}_2(\boldsymbol{\tau})}{\|\boldsymbol{\tau}\|_{\mathcal{H}_2}}\,, \tag{4.28}$$

we can replace each term $\mathcal{R}_2(\boldsymbol{\tau})$ by $\mathcal{R}_2(\boldsymbol{\tau} - \boldsymbol{\tau}_h)$, with a suitable $\boldsymbol{\tau}_h \in \mathcal{H}_{2,h}$ (cf. (3.125), (3.126)) depending on the given $\boldsymbol{\tau} \in \mathcal{H}_2$. Indeed, we first consider the Helmholtz decomposition (A.19) provided by Lemma A.2.2, with $p = 4/3$, which says that for each $\boldsymbol{\tau} \in \mathcal{H}_2$ there exist $\boldsymbol{\zeta} \in \mathbb{W}^{1,4/3}(\Omega)$ and $\boldsymbol{\xi} \in \mathbf{H}^1(\Omega)$, such that

$$\boldsymbol{\tau} \,=\, \boldsymbol{\zeta} + \underline{\mathbf{curl}}\,(\boldsymbol{\xi}) \quad \text{in} \quad \Omega \quad \text{and} \quad \|\boldsymbol{\zeta}\|_{1,4/3;\Omega} + \|\boldsymbol{\xi}\|_{1,\Omega} \,\leq\, C_{4/3}\|\boldsymbol{\tau}\|_{\mathbf{div}_{4/3};\Omega}\,, \tag{4.29}$$

with a positive constant $C_{4/3}$ independent of $\boldsymbol{\tau}$. Next, for simplicity of presentation, we focus on the discrete approach (3.125), which relies on PEERS-based elements of order $\ell \geq 0$. The AFW-based discretization (3.126) can be handled analogously, using the BDM interpolation operator instead of the Raviart–Thomas one. In fact, setting

$$\boldsymbol{\tau}_h \,:=\, \boldsymbol{\Pi}_h^k(\boldsymbol{\zeta}) + \underline{\mathbf{curl}}\,(\mathcal{I}_h(\boldsymbol{\xi})) \,+\, c\,\mathbb{I}\,, \tag{4.30}$$

where $\boldsymbol{\Pi}_h^k$ and $\mathcal{I}_h$ denote the tensor and vector versions of the Raviart–Thomas (or BDM, in the case of the AFW-based approach) and Clément interpolation operators, respectively (cf. Appendix A.2). The constant c is chosen so that $\mathrm{tr}(\boldsymbol{\tau}_h)$ has zero mean value, and hence $\boldsymbol{\tau}_h$ belongs to $\mathcal{H}_{2,h}$. Note that $\boldsymbol{\Pi}_h^k(\boldsymbol{\zeta})$ lies in $\mathbb{RT}_\ell(\Omega) \subseteq \widetilde{\mathcal{H}}_{2,h}$ (cf. (3.125)). Also observe that $\boldsymbol{\tau}_h$ can be interpreted as a discrete Helmholtz decomposition of $\boldsymbol{\tau}$. In this way, using the second equation of the Galerkin scheme (3.85), together with the compatibility condition (3.20), we deduce that $\mathcal{R}_2(c\,\mathbb{I}) = 0$, so that denoting

$$\widehat{\boldsymbol{\zeta}} \,:=\, \boldsymbol{\zeta} \,-\boldsymbol{\Pi}_h^k(\boldsymbol{\zeta}) \quad \text{and} \quad \widehat{\boldsymbol{\xi}} \,:=\, \boldsymbol{\xi} \,- \mathcal{I}_h(\boldsymbol{\xi})\,,$$

it follows from (4.29) and (4.30), that

$$\mathcal{R}_2(\boldsymbol{\tau}) \,=\, \mathcal{R}_2(\boldsymbol{\tau} - \boldsymbol{\tau}_h) \,=\, \mathcal{R}_2(\widehat{\boldsymbol{\zeta}}) + \mathcal{R}_2(\underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}}))\,, \tag{4.31}$$

where, bearing in mind the definition of $\mathcal{R}_2$ (cf. (4.14), (3.40)), we find that

$$\mathcal{R}_2(\widehat{\boldsymbol{\zeta}}) \,:=\, \int_\Omega \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big) : \widehat{\boldsymbol{\zeta}} \,+\, \int_\Omega \mathbf{u}_h \cdot \mathbf{div}(\widehat{\boldsymbol{\zeta}}) \,-\, \langle \widehat{\boldsymbol{\zeta}}\,\boldsymbol{\nu}, \mathbf{u}_D\rangle \tag{4.32}$$

and

$$\mathcal{R}_2(\underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}})) \,:=\, \int_\Omega \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big) : \underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}}) \,-\, \langle \underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}})\,\boldsymbol{\nu}, \mathbf{u}_D\rangle\,. \tag{4.33}$$

The following lemma establishes the residual upper bound for $\|\mathcal{R}_2\|_{\mathcal{H}_2'}$.

Lemma 4.2.3. Assume that $\mathbf{u}_D \in \mathbf{H}^1(\Gamma)$. Then, there exists a positive constant C, independent of h, such that

$$\|\mathcal{R}_2\|_{\mathcal{H}_2'} \leq C \left\{ \left(\sum_{K\in\mathcal{T}_h} \widetilde{\Theta}_K^2 \right)^{1/2} + \left(\sum_{K\in\mathcal{T}_h} \Theta_{3,K}^4 \right)^{1/4} \right\} . \tag{4.34}$$

where $\Theta_{3,K}$ is defined in (4.4), and

$$\begin{aligned} \widetilde{\Theta}_K^2 \; &:= \; h_K^2 \left\| \mathbf{rot}\,(\mathbf{D}_h + \boldsymbol{\gamma}_h) \right\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h(K)} h_e \left\| [\![\big(\mathbf{D}_h + \boldsymbol{\gamma}_h\big)\mathbf{s}]\!] \right\|_{0,e}^2 \\ &+ \sum_{e\in\mathcal{E}_h(K)\cap\mathcal{E}_h(\Gamma)} h_e \left\| \nabla\mathbf{u}_D\,\mathbf{s} - \big(\mathbf{D}_h + \boldsymbol{\gamma}_h\big)\mathbf{s} \right\|_{0,e}^2 . \end{aligned}$$

Proof. We proceed as in [42, Lemma 3.6]. In fact, according to (4.31), we begin by estimating $\mathcal{R}_2(\widehat{\boldsymbol{\zeta}})$. Let us first observe that, for each $e \in \mathcal{E}_h$, the identity (A.12) and the fact that $\mathbf{u}_h|_e \in \mathbf{P}_k(e)$, yield $\displaystyle\int_e \widehat{\boldsymbol{\zeta}}\boldsymbol{\nu} \cdot \mathbf{u}_h = 0$. Hence, locally integrating by parts the second term in (4.32), we readily obtain

$$\mathcal{R}_2(\widehat{\boldsymbol{\zeta}}) \;=\; -\sum_{K\in\mathcal{T}_h} \int_K \left\{ \nabla\mathbf{u}_h - \big(\mathbf{D}_h + \boldsymbol{\gamma}_h\big) \right\} : \widehat{\boldsymbol{\zeta}} - \sum_{e\in\mathcal{E}_h(\Gamma)} \int_e \big(\mathbf{u}_D - \mathbf{u}_h\big) \cdot \widehat{\boldsymbol{\zeta}}\boldsymbol{\nu} \,.$$

Thus, applying the Hölder inequality along with the approximation properties of $\boldsymbol{\Pi}_h^k$ (cf. (A.17)–(A.18) in Lemma A.2.1) with $p = 4/3$ and $l = 0$, and the stability estimate from (4.29), we get

$$\begin{aligned} \left|\mathcal{R}_2(\widehat{\boldsymbol{\zeta}})\right| \leq \widehat{C}_1 \Bigg\{ &\sum_{K\in\mathcal{T}_h} h_K^4 \left\| \nabla\mathbf{u}_h - \big(\mathbf{D}_h + \boldsymbol{\gamma}_h\big) \right\|_{0,4;K}^4 \\ &+ \sum_{e\in\mathcal{E}_h(\Gamma)} h_e \left\| \mathbf{u}_D - \mathbf{u}_h \right\|_{0,4;e}^4 \Bigg\}^{1/4} \|\boldsymbol{\tau}\|_{\mathbf{div}_{4/3};\Omega} \,. \end{aligned} \tag{4.35}$$

Next, we estimate $\mathcal{R}_2(\underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}}))$ (cf. (4.33)). In fact, regarding its second term, a suitable boundary integration by parts formula (cf. [71, eq. (3.35) in Lemma 3.5]) yields

$$\langle \underline{\mathbf{curl}}\,(\widehat{\boldsymbol{\xi}})\boldsymbol{\nu}, \mathbf{u}_D \rangle_\Gamma \;=\; -\,\langle \nabla\mathbf{u}_D\,\mathbf{s}, \widehat{\boldsymbol{\xi}} \rangle_\Gamma \,. \tag{4.36}$$

In turn, locally integrating by parts the first term of $\mathcal{R}_2(\underline{\mathbf{curl}}(\widehat{\boldsymbol{\xi}}))$, we get

$$\begin{aligned}\int_\Omega \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big) : \underline{\mathbf{curl}}(\widehat{\boldsymbol{\xi}}) \quad &= \sum_{K\in\mathcal{T}_h} \int_K \mathbf{rot}\,(\mathbf{D}_h + \boldsymbol{\gamma}_h)\cdot\widehat{\boldsymbol{\xi}} \\ - \sum_{e\in\mathcal{E}_h(\Omega)} \int_e [\![\Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big)\mathbf{s}]\!] \cdot \widehat{\boldsymbol{\xi}} &- \sum_{e\in\mathcal{E}_h(\Gamma)} \int_e \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big)\mathbf{s}\cdot\widehat{\boldsymbol{\xi}}\,,\end{aligned}$$

which together with (4.36), the Cauchy–Schwarz inequality, the approximation properties of $\mathcal{I}_h$ (cf. Lemma A.2.3), and again the stability estimate from (4.29), implies

$$\begin{aligned}\Big|\mathcal{R}_2(\underline{\mathbf{curl}}(\widehat{\boldsymbol{\xi}}))\Big| \,\le\, \widehat{C}_2 \Bigg\{ \sum_{K\in\mathcal{T}_h} h_K^2 \,\Big\|\mathbf{rot}\,(\mathbf{D}_h + \boldsymbol{\gamma}_h)\Big\|_{0,K}^2 \,+\, \sum_{e\in\mathcal{E}_h(\Omega)} h_e \,\Big\|[\![\Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big)\mathbf{s}]\!]\Big\|_{0,e}^2 \\ + \sum_{e\in\mathcal{E}_h(\Gamma)} h_e \,\Big\|\nabla\mathbf{u}_D\,\mathbf{s} - \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h\Big)\mathbf{s}\Big\|_{0,e}^2 \Bigg\}^{1/2} \|\boldsymbol{\tau}\|_{\mathbf{div}_{4/3};\Omega}\,.\end{aligned} \tag{4.37}$$

Finally, it is easy to see that (4.28), (4.29), (4.35), and (4.37) give (4.34), which ends the proof. □

We end this section by stressing that the reliability estimate (4.11) (cf. Theorem 4.2.1) follows by bounding each one of the terms $\|\mathcal{R}_1\|_{\mathcal{H}_1'}$, $\|\mathcal{R}_2\|_{\mathcal{H}_2'}$, and $\|\mathcal{R}_3\|_{\mathcal{Q}'}$, in Lemma 4.2.1 by the corresponding upper bounds derived in Lemmas 4.2.2 and 4.2.3, and considering the definition of the global estimator Θ (cf. (4.1)).

4.2.2 Efficiency

We now aim to establish the efficiency estimate of Θ (cf. (4.1)). For this purpose, we will make extensive use of the notations and results from Appendix A.3, and the original system of equations given by (3.19), which is recovered from the mixed continuous formulation (3.35) by choosing suitable test functions and integrating by parts backwardly the corresponding equations. The following theorem is the main result of this section.

Theorem 4.2.2. There exists a positive constant $C_{\mathtt{eff}}$, independent of h, such that

$$C_{\mathtt{eff}}\,\Theta \,+\, \mathtt{h.o.t} \,\le\, \|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,+\, \|p - p_h\|_{0,\Omega}\,, \tag{4.38}$$

where $\mathtt{h.o.t}$ stands eventually for one or several terms of higher order.

Throughout this section we assume, without loss of generality, that $\mathbf{u}_D$ is piecewise polynomial. Otherwise, if it is not, but it is sufficiently smooth, one proceeds similarly to [72, Section 6.2], so that higher order terms given by the error arising from a suitable polynomial approximation of this function appear in (4.38). This possibility explains the expression `h.o.t.` in (4.38).

We begin deriving the efficiency estimate (4.38) by first addressing $\Theta_{1,K}$ and the first two terms of $\Theta_{2,K}$ (cf. (4.2), (4.3)).

Lemma 4.2.4. For each $K \in \mathcal{T}_h$ there hold

$$\|\mathbf{f} + \mathbf{div}(\boldsymbol{\sigma}_h)\|_{0,4/3;K} \,\leq\, \|\mathbf{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{0,4/3;K} \tag{4.39}$$

$$\text{and} \quad \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathtt{t}}\|_{0,K} \,\leq\, 2\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}\,. \tag{4.40}$$

In addition, there exists a positive constant C, independent of h, such that

$$\begin{aligned} &\Big\|\eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho(\mathbf{u}_h \otimes \mathbf{u}_h)^{\mathtt{d}}\Big\|_{0,K} \\ &\quad\leq\, C\Big\{\|\mathbf{D} - \mathbf{D}_h\|_{0,K} \,+\, \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K} \,+\, \|\mathbf{u} - \mathbf{u}_h\|_{0,4;K} \,+\, \|p - p_h\|_{0,K}\Big\}\,. \end{aligned} \tag{4.41}$$

Proof. First, in order to show (4.39) and (4.40), it suffices to recall that $\mathbf{f} = -\,\mathbf{div}(\boldsymbol{\sigma})$ and $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathtt{t}}$ in Ω (cf. (3.19)). In turn, for the proof of (4.41), we first use the identity $\eta(p, |\mathbf{D}|)\,\mathbf{D} - \boldsymbol{\sigma}^{\mathtt{d}} - \rho\,(\mathbf{u} \otimes \mathbf{u})^{\mathtt{d}} = \mathbf{0}$ in Ω (cf. (3.19)) and triangle inequality, to deduce

$$\begin{aligned} &\Big\|\eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho(\mathbf{u}_h \otimes \mathbf{u}_h)^{\mathtt{d}}\Big\|_{0,K} \\ &\quad\leq\, \Big\|\eta\big(p, |\mathbf{D}|\big)\mathbf{D} - \eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h\Big\|_{0,K} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K} + \rho\,\|\mathbf{u} \otimes \mathbf{u} - \mathbf{u}_h \otimes \mathbf{u}_h\|_{0,K}\,, \end{aligned} \tag{4.42}$$

where, adding and subtracting $\eta\big(p, |\mathbf{D}_h|\big)\mathbf{D}_h$ in the first term on the right-hand side of (4.42), and using the Lipschitz continuity estimates [43, eqs. (4.8) and (4.11)], we find that there exists positive constants $L_{\mathcal{A}}, L_\eta$, such that

$$\begin{aligned} &\Big\|\eta\big(p, |\mathbf{D}|\big)\mathbf{D} - \eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h\Big\|_{0,K} \\ &\quad\leq\, \Big\|\eta\big(p, |\mathbf{D}|\big)\mathbf{D} - \eta\big(p, |\mathbf{D}_h|\big)\mathbf{D}_h\Big\|_{0,K} \,+\, \Big\|\Big\{\eta\big(p, |\mathbf{D}_h|\big) - \eta\big(p_h, |\mathbf{D}_h|\big)\Big\}\mathbf{D}_h\Big\|_{0,K} \\ &\quad\leq\, L_{\mathcal{A}}\,\|\mathbf{D} - \mathbf{D}_h\|_{0,K} + L_\eta\,\|p - p_h\|_{0,K}\,. \end{aligned} \tag{4.43}$$

In turn, proceeding as in (4.23) in combination with the fact that $\|\mathbf{u}\|_{0,4;K}$ and $\|\mathbf{u}_h\|_{0,4;K}$ are bounded by $\|\mathbf{u}\|_{0,4;\Omega}$ and $\|\mathbf{u}_h\|_{0,4;\Omega}$, respectively, with $\mathbf{u} \in \mathrm{W}(\delta)$ and $\mathbf{u}_h \in \mathrm{W}(\delta_{\mathtt{d}})$, there holds

$$\begin{aligned}\|\mathbf{u}\otimes\mathbf{u}-\mathbf{u}_h\otimes\mathbf{u}_h\|_{0,K} &\le n^{1/2}\Big(\|\mathbf{u}\|_{0,4;K}+\|\mathbf{u}_h\|_{0,4;K}\Big)\|\mathbf{u}-\mathbf{u}_h\|_{0,4;K}\\ &\le n^{1/2}\Big(\|\mathbf{u}\|_{0,4;\Omega}+\|\mathbf{u}_h\|_{0,4;\Omega}\Big)\|\mathbf{u}-\mathbf{u}_h\|_{0,4;K}\\ &\le n^{1/2}\Big(\delta+\delta_{\mathtt{d}}\Big)\|\mathbf{u}-\mathbf{u}_h\|_{0,4;K}\,.\end{aligned}\tag{4.44}$$

Finally, replacing back (4.43) and (4.44) into (4.42) we obtain (4.41) and conclude the proof. □

We remark that the local efficiency estimates for the remaining terms in the definition of Θ (cf. (4.1)) have already been established in the literature. These estimates are derived using the localization technique based on triangle-bubble and edge-bubble functions (cf. (A.22) and Lemma A.3.4), together with the local inverse inequality (cf. (A.23)) and the discrete trace inequality (cf. (A.24)). For completeness, we state the following result.

Lemma 4.2.5. There exist positive constants C_i, $i\in\{1,\dots,5\}$, all independent of h, such that

a) $$h_K^4\Big\|\nabla\mathbf{u}_h-\Big(\mathbf{D}_h+\boldsymbol{\gamma}_h\Big)\Big\|_{0,4;K}^4 \le C_1\Big\{\|\mathbf{u}-\mathbf{u}_h\|_{0,4;K}^4+h_K^2\|\mathbf{D}-\mathbf{D}_h\|_{0,K}^4+h_K^2\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_h\|_{0,K}^4\Big\}\quad\forall\,K\in\mathcal{T}_h\,,$$

b) $$h_e\,\|\mathbf{u}_D-\mathbf{u}_h\|_{0,4;e}^4 \le C_2\,\Big\{\|\mathbf{u}-\mathbf{u}_h\|_{0,4;K_e}^4+h_{K_e}^2\,\|\mathbf{D}-\mathbf{D}_h\|_{0,K_e}^4+h_{K_e}^2\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_h\|_{0,K_e}^4\Big\}\quad\forall\,e\in\mathcal{E}_h(\Gamma)\,,$$

c) $$h_K^2\,\Big\|\mathbf{rot}\,(\mathbf{D}_h+\boldsymbol{\gamma}_h)\Big\|_{0,K}^2\le C_3\,\Big\{\|\mathbf{D}-\mathbf{D}_h\|_{0,K}^2+\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_h\|_{0,K}^2\Big\}\quad\forall\,K\in\mathcal{T}_h\,,$$

d) $$h_e\,\Big\|[\![(\mathbf{D}_h+\boldsymbol{\gamma}_h)\mathbf{s}]\!]\Big\|_{0,e}^2\le C_4\,\Big\{\|\mathbf{D}-\mathbf{D}_h\|_{0,\omega_e}^2+\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_h\|_{0,\omega_e}^2\Big\}\quad\forall\,e\in\mathcal{E}_h(\Omega)\,,$$

e) $$h_e\,\Big\|\nabla\mathbf{u}_D\,\mathbf{s}-\Big(\mathbf{D}_h+\boldsymbol{\gamma}_h\Big)\mathbf{s}\Big\|_{0,e}^2\le C_5\,\Big\{\|\mathbf{D}-\mathbf{D}_h\|_{0,K_e}^2+\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_h\|_{0,K_e}^2\Big\}\quad\forall\,e\in\mathcal{E}_h(\Gamma)\,,$$

where K_e is the triangle of $\mathcal{T}_h$ having e as an edge, whereas ω_e denotes the union of the two elements of $\mathcal{T}_h$ sharing the edge e.

Proof. The estimate a) follows directly from the proof of [41, Lemma 3.15], replacing $\mathbf{t}_h$ therein with $\mathbf{D}_h+\boldsymbol{\gamma}_h$, while b) is given in [41, Lemma 3.16]. For c) and d), we refer to [73, Lemmas 4.3 and 4.4]. Finally, the proof of e) follows the same arguments as those in [74, Lemma 4.15]. □

We conclude this section by noting that the proof of (4.38) (cf. Theorem 4.2.2) follows directly from Lemmas 4.2.4 and 4.2.5 and summing the local efficiency estimates over all $K \in \mathcal{T}_h$. Further details are omitted.

4.3 Numerical results

This section serves to illustrate the performance and accurancy of the proposed mixed finite element scheme (3.85) along with the reliability and efficiency properties of the *a posteriori* error estimator Θ (cf. (4.1)) derived in Section 4.2. In what follows, we refer to the corresponding sets of finite element subspaces generated by $\ell = \{0,1\}$ as simply PEERS$_\ell$ and AFW$_\ell$ based discretizations (cf. (3.125), (3.126)). The numerical methods have been implemented using the open source finite element library `FEniCS` [52]. Regarding the implementation of the Newton-type iterative method associated with (3.85) (see [43, steps (1)-(3) in Section 7] for details), the iterations are terminated once the relative error of the entire coefficient vectors between two consecutive iterates, namely $\mathbf{coeff}^m$ and $\mathbf{coeff}^{m+1}$, is sufficiently small, that is,

$$\frac{\|\mathbf{coeff}^{m+1} - \mathbf{coeff}^m\|_{\mathtt{DOF}}}{\|\mathbf{coeff}^{m+1}\|_{\mathtt{DOF}}} \leq \mathsf{tol}\,,$$

where $\|\cdot\|_{\mathtt{DOF}}$ stands for the usual Euclidean norm in $\mathrm{R}^{\mathtt{DOF}}$ with $\mathtt{DOF}$ denoting the total number of degrees of freedom defining the finite element subspaces $\mathcal{H}_{1,h}$, $\widetilde{\mathcal{H}}_{2,h}$, $\mathcal{Q}_{1,h}$, and $\mathcal{Q}_{2,h}$ (cf. (3.125), (3.126)), and tol is a fixed tolerance chosen as $\mathsf{tol} = 1\mathrm{E}-06$.

The global error and the effectivity index associated to the global estimator Θ (cf. (4.1)) are denoted, respectively, by

$$\mathsf{e}(\vec{\mathbf{t}}) \,:=\, \mathsf{e}(\mathbf{D}) + \mathsf{e}(\boldsymbol{\sigma}) + \mathsf{e}(\mathbf{u}) + \mathsf{e}(\boldsymbol{\gamma}) + \mathsf{e}(p) \quad \text{and} \quad \mathtt{eff}(\Theta) := \frac{\mathsf{e}(\vec{\mathbf{t}})}{\Theta}\,,$$

where

$$\mathsf{e}(\mathbf{D}) \,:=\, \|\mathbf{D} - \mathbf{D}_h\|_{0,\Omega}\,, \quad \mathsf{e}(\boldsymbol{\sigma}) \,:=\, \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div}_{4/3};\Omega}\,, \quad \mathsf{e}(\mathbf{u}) \,:=\, \|\mathbf{u} - \mathbf{u}_h\|_{0,4;\Omega}\,,$$

$$\mathsf{e}(\boldsymbol{\gamma}) \,:=\, \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,\Omega}\,, \quad \text{and} \quad \mathsf{e}(p) \,:=\, \|p - p_h\|_{0,\Omega}\,.$$

Moreover, using the fact that $\mathtt{DOF}^{-1/n} \cong h$, the respective experimental rates of convergence are computed as

$$\mathsf{r}(\diamond) \,:=\, -n\,\frac{\log(\mathsf{e}(\diamond)/\widehat{\mathsf{e}}(\diamond))}{\log(\mathtt{DOF}/\widehat{\mathtt{DOF}})} \quad \text{for each } \diamond \in \Big\{\mathbf{D}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\gamma}, p, \vec{\mathbf{t}}\Big\}\,,$$

where $\mathtt{DOF}$ and $\widehat{\mathtt{DOF}}$ denote the total degrees of freedom associated to two consecutive triangulations with errors $\mathsf{e}(\diamond)$ and $\widehat{\mathsf{e}}(\diamond)$, respectively. We stress that, for the sake of simplicity and clarity of presentation, in the examples considered below we only report errors and rates of convergence for the most physically relevant unknowns, namely $\boldsymbol{\sigma}$, $\mathbf{u}$, p, and $\vec{\mathbf{t}} = (\vec{\mathbf{D}}, p)$. We recall that the reliability and efficiency of the global estimator Θ (cf. (4.11), (4.38)) are with respect to the full error in $\vec{\mathbf{t}}$, and therefore we are particularly interested in the behavior of this error.

The examples to be considered in this section are described next, for which we consider the regularized viscosity $\eta(\varrho, \omega)$ defined by (3.11). In the first three examples, for the sake of simplicity, we take $\mu_s = 0.1$, $\mu_d = 1$, $I_0 = 1$, $d = 1$ and $\rho = 1$. In addition, it is easy to see for these examples that the boundary data $\mathbf{u}_D := \mathbf{u}|_\Gamma$ satisfy the required regularity $\mathbf{u}_D \in \mathbf{H}^1(\Gamma)$ since the given exact solutions $\mathbf{u}$ are sufficiently regular. In turn, the null mean value of $\mathrm{tr}(\boldsymbol{\sigma}_h)$ over Ω is fixed via a real Lagrange multiplier strategy.

Example 1 is used to corroborate the reliability and efficiency of the *a posteriori* error estimator Θ, whereas Examples 2, 3 and 4 are utilized to illustrate the behavior of the associated adaptive algorithm in 2D and 3D domains with and without manufactured solution, respectively, which applies the following procedure from [75]:

(1) Start with a coarse mesh $\mathcal{T}_h$ of $\overline{\Omega}$.

(2) Solve the Newton iterative method associated with (3.85) on the current mesh.

(3) Compute the local indicator Θ_K for each $K \in \mathcal{T}_h$, where

$$\Theta_K := \Theta_{1,K} + \Theta_{2,K} + \Theta_{3,K} \quad \text{(cf. (4.2), (4.3), (4.4))}\,.$$

(4) Check the stopping criterion and decide whether to finish or go to the next step.

(5) Use Plaza and Carey's algorithm [76] to refine each $K' \in \mathcal{T}_h$ satisfying

$$\Theta_{K'} \,\geq\, C_{\mathtt{PC}} \max\left\{\Theta_K : \quad K \in \mathcal{T}_h\right\} \text{ for some } C_{\mathtt{PC}} \in (0,1)\,.$$

(6) Define the resulting mesh as the current mesh, and go to step (2).

In particular, in the 2D Examples 2 and 4 below, we set $C_{\mathtt{PC}} = \{0.25, 0.1\}$ for $\ell = \{0, 1\}$, respectively, while in the 3D Example 3, we set $C_{\mathtt{PC}} = 0.5$.

Example 4.1: Accuracy assessment with a smooth solution in a square domain

We first focus on the accuracy of the mixed methods and the properties of the *a posteriori* error estimator through the effectivity index $\texttt{eff}(\Theta)$ under a quasi-uniform refinement strategy. We consider the square domain $\Omega := (0,1)^2$ and set the regularization parameter to $\varepsilon = 1\mathrm{E}-08$. The data $\mathbf{f}$ and $\mathbf{u}_D$ are adjusted so that a manufactured solution of (3.19) is given by the following smooth functions

$$\mathbf{u}(\mathbf{x}) \;=\; \begin{pmatrix} \sin(x_1)\cos(x_2) \\ -\cos(x_1)\sin(x_2) \end{pmatrix} \quad \text{and} \quad p(\mathbf{x}) \;=\; \exp(x_1 + x_2),$$

where $p \in \mathrm{L}^2_\kappa(\Omega)$, with $\kappa = (\exp(1)-1)^2$. Tables 4.1 and 4.2 shows the convergence history for a sequence of quasi-uniform mesh refinements for both PEERS$_\ell$ and AFW$_\ell$-based discretizations, corresponding to $\ell = 0$ and $\ell = 1$, respectively. The results are consistent with the theoretical bounds established in [43, Theorem 6.2]. In addition, we compute the global *a posteriori* error indicator Θ (cf. (4.1)) and assess its reliability and efficiency through the effectivity index. We observe that the estimator remains uniformly bounded throughout the refinement process.

PEERS$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
3314	0.177	14	5.5e-01	–	3.7e-02	–	2.0e-01	–	1.0e-00	–	1.1e-00	0.911
16634	0.079	12	2.4e-01	1.05	1.6e-02	1.06	7.8e-02	1.13	4.3e-01	1.06	5.3e-01	0.827
29522	0.059	12	1.8e-01	1.03	1.2e-02	1.03	5.8e-02	1.08	3.2e-01	1.04	4.0e-01	0.812
73874	0.037	11	1.1e-01	1.02	7.4e-03	1.01	3.6e-02	1.05	2.0e-01	1.03	2.5e-01	0.797
209282	0.022	9	6.5e-02	1.01	4.4e-03	1.01	2.1e-02	1.02	1.2e-01	1.02	1.5e-01	0.787
510602	0.014	8	4.2e-02	1.01	2.8e-03	1.00	1.3e-02	1.01	7.5e-02	1.01	9.6e-02	0.782

AFW$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
2369	0.177	14	2.8e-01	–	3.5e-02	–	1.6e-01	–	5.4e-01	–	5.4e-01	0.995
11809	0.079	11	1.2e-01	1.01	1.6e-02	1.01	7.3e-02	1.01	2.4e-01	1.01	2.4e-01	1.003
20929	0.059	10	9.3e-02	1.01	1.2e-02	1.01	5.5e-02	1.01	1.8e-01	1.01	1.8e-01	1.004
52289	0.037	9	5.9e-02	1.00	7.3e-03	1.00	3.4e-02	1.00	1.1e-01	1.00	1.1e-01	1.005
147969	0.022	7	3.5e-02	1.00	4.4e-03	1.00	2.0e-02	1.00	6.7e-02	1.00	6.7e-02	1.005
360801	0.014	6	2.2e-02	1.00	2.8e-03	1.00	1.3e-02	1.00	4.3e-02	1.00	4.3e-02	1.005

Table 4.1 [Example 4.1, $\ell = 0$] Number of degrees of freedom, meshsizes, Newton iteration count, errors, rates of convergence, global estimator, and effectivity index for the mixed approximations.

PEERS$_\ell$-based discretization with $\ell = 1$ and quasi-uniform refinement												
DOF	h	it	$\mathsf{e}(\boldsymbol{\sigma})$	$\mathsf{r}(\boldsymbol{\sigma})$	$\mathsf{e}(\mathbf{u})$	$\mathsf{r}(\mathbf{u})$	$\mathsf{e}(p)$	$\mathsf{r}(p)$	$\mathsf{e}(\vec{\mathbf{t}})$	$\mathsf{r}(\vec{\mathbf{t}})$	Θ	$\mathtt{eff}(\Theta)$
7010	0.177	10	1.2e-02	–	1.2e-03	–	4.5e-03	–	2.6e-02	–	4.6e-02	0.559
35210	0.079	7	2.4e-03	1.99	2.3e-04	2.01	8.9e-04	2.02	5.5e-03	1.93	9.5e-03	0.572
62498	0.059	7	1.3e-03	1.99	1.3e-04	2.01	5.0e-04	2.00	3.1e-03	1.96	5.4e-03	0.575
156410	0.037	6	5.4e-04	1.99	5.1e-05	2.00	2.0e-05	2.00	1.3e-03	1.97	2.2e-03	0.579
443138	0.022	4	1.9e-04	1.99	1.8e-05	2.00	7.0e-05	2.00	4.5e-04	1.98	7.7e-04	0.581
1081202	0.014	4	7.8e-05	2.00	7.3e-06	2.00	2.9e-05	2.00	1.9e-04	1.99	3.2e-04	0.583

AFW$_\ell$-based discretization with $\ell = 1$ and quasi-uniform refinement												
DOF	h	it	$\mathsf{e}(\boldsymbol{\sigma})$	$\mathsf{r}(\boldsymbol{\sigma})$	$\mathsf{e}(\mathbf{u})$	$\mathsf{r}(\mathbf{u})$	$\mathsf{e}(p)$	$\mathsf{r}(p)$	$\mathsf{e}(\vec{\mathbf{t}})$	$\mathsf{r}(\vec{\mathbf{t}})$	Θ	$\mathtt{eff}(\Theta)$
5473	0.177	7	6.1e-03	–	1.2e-03	–	4.3e-03	–	1.3e-02	–	2.1e-02	0.600
27433	0.079	5	1.2e-03	2.02	2.3e-04	2.01	8.6e-04	2.01	2.5e-03	2.02	4.3e-03	0.591
48673	0.059	5	6.7e-04	2.01	1.3e-04	2.01	4.8e-04	2.01	1.4e-03	2.01	2.4e-03	0.590
121753	0.037	4	2.7e-04	2.01	5.1e-05	2.01	1.9e-04	2.01	5.7e-04	2.01	9.6e-04	0.587
344833	0.022	3	9.4e-05	2.01	1.8e-05	2.00	6.8e-05	2.00	2.0e-04	2.01	3.4e-04	0.585
841201	0.014	3	3.8e-05	2.00	7.3e-06	2.00	2.8e-05	2.00	8.2e-05	2.00	1.4e-04	0.585

Table 4.2 [Example 4.1, $\ell = 1$] Number of degrees of freedom, meshsizes, Newton iteration count, errors, rates of convergence, global estimator, and effectivity index for the mixed approximations.

Example 4.2: Adaptivity in a 2D L-shaped domain

The second example is aimed at testing the features of adaptive mesh refinement after the *a posteriori* error estimator Θ (cf. (4.1)). We consider a 2D L-shaped domain $\Omega := (0,1)^2 \setminus (0.5,1)^2$ and the regularization parameter as $\varepsilon = 1\mathrm{E}-08$. The data $\mathbf{f}$ and $\mathbf{u}_D$ are chosen so that the exact solution is given by

$$\mathbf{u}(\mathbf{x}) \;=\; \begin{pmatrix} \sin(\pi\, x_1)\, \cos(\pi\, x_2) + x_2 \\ -\cos(\pi\, x_1)\, \sin(\pi\, x_2) + x_1 \end{pmatrix} \quad \text{and} \quad p(\mathbf{x}) \;=\; 18 - 10 \exp\left(\frac{-0.001}{r(\mathbf{x})}\right),$$

with $r(\mathbf{x}) := (x_1 - 0.51)^2 + (x_2 - 0.51)^2$. Notice that the pressure field exhibits high gradients near the vertex $(0.5, 0.5)$. Tables 4.3 and 4.4, together with Figure 4.1, summarize the convergence behavior of the mixed methods applied to a sequence of quasi-uniform and adaptively refined triangulations of the domain. Suboptimal convergence rates are observed in the quasi-uniform case. In contrast, adaptive refinement guided by the *a posteriori* error indicator Θ leads to optimal rates and stable effectivity indices for both PEERS$_\ell$ and AFW$_\ell$-based discretizations with $\ell = \{0, 1\}$. The adaptive strategy significantly enhances the efficiency of the method, enabling high-quality approximations at reduced computational cost. For $\ell = 0$, solutions with improved accuracy in terms of $\mathsf{e}(\vec{\mathbf{t}})$ are obtained using approximately 60% of the degrees of freedom required by the final quasi-uniform mesh. This reduction is significant, especially considering the challenges posed by the nonlinearities involved in the model. This efficiency is further enhanced for $\ell = 1$, where accurate solutions are obtained using only approximately 10% of the degrees of freedom, highlighting the substantial advantage of the adaptive approach in this case. Figure 4.2 displays the initial mesh and some approximate solutions computed with the adaptive PEERS$_1$-based method, using Θ, on a mesh with $706,301$ degrees of freedom and $13,061$ triangles. These results confirm that the pressure exhibits strong variations in the contraction region. Additionally, Figure 4.3 shows examples of adapted meshes for the mixed methods when $\ell = 1$. As expected, the refinement is concentrated near the reentrant corner of the 2D L-shaped domain, revealing the indicator's ability to effectively localize the singularity.

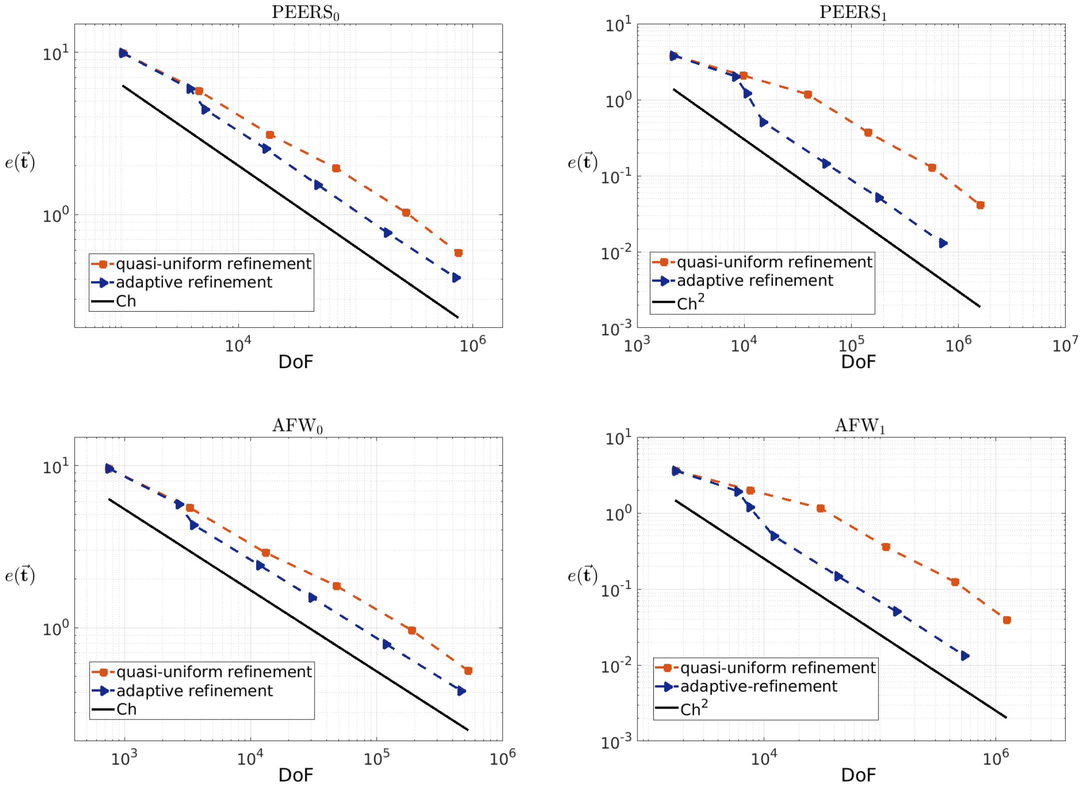


Figure 4.1 [Example 4.2] Log-log plot of $\mathrm{e}(\vec{\mathbf{t}})$ vs. DOF for quasi-uniform/adaptive refinements for PEERS$_\ell$ and AFW$_\ell$-based discretizations with $\ell = \{0, 1\}$ (top and bottom plots, respectively).

PEERS$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
1028	0.280	15	8.0e-00	–	1.9e-01	–	7.0e-01	–	9.9e-00	–	9.8e-00	1.009
4601	0.141	18	4.9e-00	0.64	9.0e-02	0.97	3.4e-01	0.97	5.8e-00	0.72	5.7e-00	1.016
18491	0.071	15	2.7e-00	0.86	4.4e-02	1.04	1.8e-01	0.94	3.1e-00	0.89	3.1e-00	1.022
67811	0.038	13	1.7e-00	0.71	2.3e-02	0.99	9.8e-02	0.91	1.9e-00	0.74	1.9e-00	1.023
267785	0.019	12	9.2e-01	0.91	1.1e-02	1.01	5.0e-02	0.96	1.0e-00	0.92	1.0e-00	1.020
752408	0.011	11	5.2e-01	1.12	6.8e-03	1.00	2.9e-02	1.06	5.8e-01	1.11	5.7e-01	1.018

PEERS$_\ell$-based discretization with $\ell = 0$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
1028	15	8.0e-00	–	1.9e-01	–	7.0e-01	–	9.9e-00	–	9.8e-00	1.009
3857	17	5.1e-00	0.68	1.0e-01	0.90	3.6e-01	1.02	6.0e-00	0.77	6.0e-00	0.999
5189	17	3.7e-00	2.11	8.9e-02	0.92	2.7e-01	1.92	4.5e-00	1.98	4.5e-00	0.979
16997	14	2.2e-00	0.90	5.0e-02	0.98	1.4e-01	1.15	2.6e-00	0.94	2.6e-00	0.967
47183	14	1.3e-00	1.03	3.3e-02	0.82	8.6e-02	0.91	1.5e-00	1.00	1.6e-00	0.967
184580	13	6.6e-01	0.98	1.6e-02	1.03	4.2e-02	1.05	7.8e-01	1.00	8.1e-01	0.962
710489	12	3.5e-01	0.94	8.1e-03	1.03	2.2e-02	0.96	4.1e-01	0.95	4.2e-01	0.966

AFW$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
745	0.280	19	7.9e-00	–	1.8e-01	–	8.0e-01	–	9.6e-00	–	8.7e-00	1.106
3285	0.141	19	4.7e-00	0.69	9.0e-02	0.96	3.3e-01	1.18	5.5e-00	0.75	5.2e-00	1.061
13117	0.071	18	2.5e-00	0.90	4.4e-02	1.04	1.5e-01	1.16	2.9e-00	0.93	2.8e-00	1.039
47997	0.038	17	1.6e-00	0.70	2.3e-02	0.99	8.1e-02	0.93	1.8e-00	0.73	1.8e-00	1.032
189285	0.019	17	8.7e-01	0.91	1.1e-02	1.01	4.2e-02	0.94	9.7e-01	0.92	9.4e-01	1.030
531593	0.011	16	4.9e-01	1.12	6.8e-03	1.00	2.4e-02	1.09	5.4e-01	1.11	5.3e-01	1.028

AFW$_\ell$-based discretization with $\ell = 0$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
745	19	7.9e-00	–	1.8e-01	–	8.0e-01	–	9.6e-00	–	8.7e-00	1.106
2685	19	4.9e-00	0.74	9.9e-02	0.95	3.6e-01	1.25	5.8e-00	0.80	5.4e-00	1.064
3517	19	3.6e-00	2.32	9.1e-02	0.63	2.4e-01	2.83	4.3e-00	2.16	4.1e-00	1.052
11729	18	2.1e-00	0.91	4.8e-02	1.07	1.0e-01	1.46	2.4e-00	0.96	2.4e-00	1.026
30457	18	1.3e-00	0.98	3.3e-02	0.76	5.7e-02	1.20	1.5e-00	0.96	1.5e-00	1.015
118453	17	6.8e-01	0.96	1.7e-02	1.03	2.9e-02	1.02	7.9e-01	0.97	7.8e-01	1.016
462749	15	3.5e-01	0.96	8.3e-03	1.02	1.5e-02	0.98	4.1e-01	0.97	4.0e-01	1.017

Table 4.3 [Example 4.2, $\ell = 0$] Comparison of the mixed approximations with quasi-uniform and adaptive refinements for the $\mu(I)$-rheology model.

Example 4.3: Adaptivity in a 3D L-shaped domain

Here, we replicate the Example 4.2 in a three-dimensional setting but now considering the 3D L-shaped domain $\Omega = (0,1) \times (0,0.5) \times (0,1) \setminus (0.5,1) \times (0,0.5) \times (0.5,1)$, the regularization parameter as $\varepsilon = 1\mathrm{E}-06$, and the manufactured exact solutions given

PEERS$_\ell$-based discretization with $\ell = 1$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
2171	0.280	16	3.4e-00	–	2.4e-02	–	2.4e-01	–	3.8e-00	–	3.9e-00	0.979
9734	0.141	14	1.9e-00	0.75	5.6e-03	1.95	9.1e-02	1.28	2.1e-00	0.80	2.1e-00	1.021
39143	0.071	12	1.1e-00	0.77	1.3e-03	2.13	3.2e-02	1.50	1.2e-00	0.82	1.2e-00	1.028
143573	0.038	9	3.5e-01	1.80	3.5e-04	1.99	1.3e-02	1.39	3.7e-01	1.78	3.6e-01	1.032
567023	0.019	7	1.2e-01	1.53	8.8e-05	2.00	4.1e-03	1.66	1.3e-01	1.54	1.3e-01	1.034
1593242	0.011	5	3.9e-02	2.22	3.1e-05	2.00	1.3e-03	2.28	4.1e-02	2.22	4.0e-02	1.029

PEERS$_\ell$-based discretization with $\ell = 1$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
2171	16	3.4e-00	–	2.4e-02	–	2.4e-01	–	3.8e-00	–	3.9e-00	0.979
8267	14	1.9e-00	0.90	6.3e-03	2.00	9.7e-02	1.32	2.0e-00	0.95	2.0e-00	1.024
10547	14	1.2e-00	4.04	6.2e-03	0.08	3.9e-02	7.56	1.2e-00	4.17	1.3e-00	0.965
14948	13	4.6e-01	5.25	5.4e-03	0.83	1.9e-02	4.17	5.1e-01	4.98	5.8e-01	0.880
57371	11	1.3e-01	1.87	1.4e-03	2.05	5.4e-03	1.85	1.5e-01	1.87	1.6e-01	0.891
179354	9	4.7e-02	1.80	3.6e-04	2.33	1.9e-03	1.87	5.2e-02	1.82	5.6e-02	0.918
706301	7	1.2e-02	2.00	9.1e-05	2.02	4.7e-04	2.01	1.3e-02	2.00	1.4e-02	0.916

AFW$_\ell$-based discretization with $\ell = 1$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
1702	0.280	15	3.3e-00	–	2.4e-02	–	2.3e-01	–	3.6e-00	–	3.5e-00	1.027
7597	0.141	13	1.9e-00	0.76	5.5e-03	1.95	9.0e-02	1.27	2.0e-00	0.81	1.9e-00	1.035
30490	0.071	11	1.1e-00	0.73	1.3e-03	2.12	3.1e-02	1.56	1.2e-00	0.77	1.1e-00	1.033
111760	0.038	8	3.5e-01	1.82	3.5e-04	1.99	1.3e-02	1.37	3.6e-01	1.80	3.5e-01	1.033
441202	0.019	6	1.2e-01	1.54	8.8e-05	2.01	4.0e-03	1.66	1.3e-01	1.54	1.2e-01	1.028
1239529	0.011	5	3.8e-02	2.23	3.1e-05	2.00	1.2e-03	2.30	3.9e-02	2.24	3.9e-02	1.020

AFW$_\ell$-based discretization with $\ell = 1$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
1702	15	3.3e-00	–	2.4e-02	–	2.3e-01	–	3.6e-00	–	3.5e-00	1.027
5893	14	1.8e-00	0.96	7.4e-03	1.88	9.5e-02	1.45	1.9e-00	1.01	1.9e-00	1.020
7456	13	1.1e-00	3.98	7.3e-03	0.04	3.4e-02	8.73	1.2e-00	4.10	1.2e-00	0.967
12022	13	4.7e-01	3.72	5.6e-03	1.19	1.3e-02	4.06	5.0e-01	3.65	5.4e-01	0.931
43087	11	1.4e-01	1.91	1.5e-03	2.09	4.1e-03	1.81	1.5e-01	1.91	1.6e-01	0.934
137791	9	4.7e-02	1.84	3.9e-04	2.28	1.5e-03	1.69	5.1e-02	1.84	5.3e-02	0.951
534541	6	1.2e-02	1.98	9.7e-05	2.04	4.1e-04	1.96	1.3e-02	1.98	1.4e-02	0.954

Table 4.4 [Example 4.2, $\ell = 1$] Comparison of the mixed approximations with quasi-uniform and adaptive refinements for the $\mu(I)$-rheology model.

by

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} \sin(x_1)\cos(x_2)\cos(x_3) \\ -2\,\cos(x_1)\sin(x_2)\cos(x_3) \\ \cos(x_1)\cos(x_2)\sin(x_3) \end{pmatrix} \quad \text{and} \quad p(\mathbf{x}) = 80 - 40\exp\left(\frac{-0.0001}{r(\mathbf{x})}\right),$$

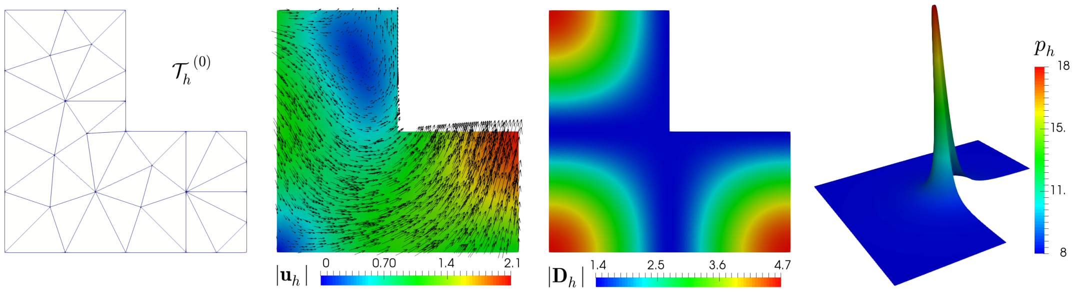


Figure 4.2 [Example 4.2] Initial mesh, computed magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

with $r(\mathbf{x}) := (x_1 - 0.505)^2 + (x_3 - 0.505)^2$. The convergence history for a set of quasi-uniform and adaptive mesh refinements using both PEERS$_0$ and AFW$_0$-based discretizations is shown in Table 4.5, along with Figure 4.4. We observe a considerable increase in the number of degrees of freedom in the PEERS$_0$-based scheme compared to the AFW$_0$ one. For this reason, and due to computational limitations, we report results for only four meshes in the case of the PEERS$_0$-based discretization. This is mainly explained by the fact that the symmetric part of the velocity gradient is approximated using $\mathbb{P}_3(\Omega)$ and $\mathbb{P}_1(\Omega)$, respectively. Nevertheless, in both cases we observe disturbed convergence under quasi-uniform refinement and optimal convergence rates when using adaptive refinement guided by the *a posteriori* error estimator Θ (cf. (4.1)). The initial mesh and some approximate solutions computed using the adaptive AFW$_0$-based scheme (driven by Θ), with $775,808$ degrees of freedom and $13,724$ tetrahedra, are displayed in Figure 4.5. Snapshots of three meshes generated via Θ are shown in Figure 4.6, where an incipient clustering of elements around the contraction region can be observed.

Example 4.4: Fluid flow through a 2D cavity with two circular obstacles

Inspired by Example 3.7, we finally focus on studying the behavior of the regularized $\mu(I)$-rheology model for granular materials in fluid flow through a 2D cavity with two circular obstacles, without employing a manufactured solution. More precisely, we

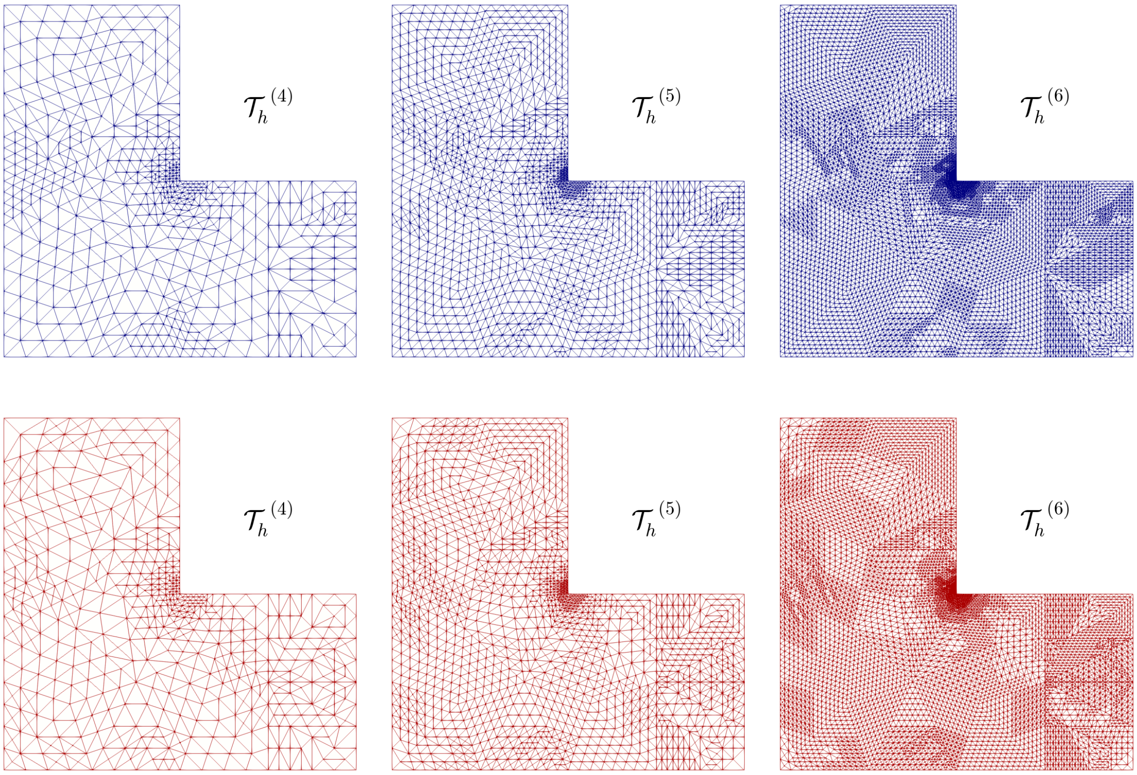


Figure 4.3 [Example 4.2] Three snapshots of adapted meshes according to the indicator Θ for PEERS$_1$ and AFW$_1$-based discretizations (top and bottom plots, respectively).

consider the domain $\Omega = (0,1)^2 \setminus \Omega_c$, where

$$\Omega_c = \Big\{(x_1,x_2):\ (x_1-1/2)^2+(x_2-1/3)^2 < 0.1^2\Big\} \\ \cup\Big\{(x_1,x_2):\ (x_1-1/2)^2+(x_2-2/3)^2 < 0.1^2\Big\},$$

with boundary Γ, whose part around the circles is given by $\Gamma_c = \partial\Omega_c$. The model parameters are chosen as $\mu_s = 0.36, \mu_d = 0.91, I_0 = 0.73, d = 0.05, \rho = 2500$, and the regularization factor is $\varepsilon = 1\mathrm{E}-03$. Notice that the relation between the diameter of the particles d and the width of the cavity is $1:20$, whereas the radius of both circular obstacles is double that of d. The mean value of p is fixed as $\kappa = 100$, no presence of gravity is assumed, that is, $\mathbf{f} = \mathbf{0}$, and the boundaries conditions are

$$\mathbf{u} = (0.2\,x_2 - 0.1, 0)^{\mathrm{t}} \quad \text{on} \quad \Gamma\setminus\Gamma_c \quad \text{and} \quad \mathbf{u} = \mathbf{0} \quad \text{on} \quad \Gamma_c\,.$$

In particular, we impose that flows cannot go in nor out through Γ_c, whereas at the top and bottom of the domain flows are faster in opposite direction. In Figure 4.7,

PEERS$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
32744	0.522	18	1.3e+01	–	9.8e-02	–	2.1e-00	–	1.6e+01	–	1.2e+01	1.271
296142	0.207	16	6.5e-00	0.97	4.1e-02	1.18	8.7e-01	1.19	7.5e-00	1.00	6.4e-00	1.181
605245	0.164	16	5.8e-00	0.49	3.2e-02	1.03	6.9e-01	0.98	6.6e-00	0.55	5.6e-00	1.168
1651385	0.114	16	5.2e-00	0.31	2.3e-02	1.03	4.9e-01	1.03	5.8e-00	0.39	5.0e-00	1.149

PEERS$_\ell$-based discretization with $\ell = 0$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
32744	18	1.3e+01	–	9.8e-02	–	2.1e-00	–	1.6e+01	–	1.2e+01	1.271
106606	17	7.2e-00	1.55	7.4e-02	0.70	9.8e-01	1.90	8.4e-00	1.59	7.2e-00	1.157
374390	17	5.2e-00	0.78	6.2e-02	0.43	4.7e-01	1.78	5.8e-00	0.88	5.3e-00	1.084
935833	17	3.8e-00	1.05	4.1e-02	1.36	2.5e-01	2.08	4.1e-00	1.12	3.9e-00	1.061

AFW$_\ell$-based discretization with $\ell = 0$ and quasi-uniform refinement												
DOF	h	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
10911	0.522	11	1.3e+01	–	9.8e-02	–	2.0e-00	–	1.5e+01	–	1.2e+01	1.267
94997	0.207	10	6.4e-00	1.00	4.1e-02	1.20	8.6e-01	1.20	7.3e-00	1.03	6.3e-00	1.173
193678	0.164	10	5.7e-00	0.51	3.2e-02	1.04	6.8e-01	1.00	6.4e-00	0.57	5.5e-00	1.154
525096	0.114	10	5.1e-00	0.34	2.3e-02	1.04	4.8e-01	1.03	5.6e-00	0.41	5.0e-00	1.126
1595337	0.079	10	4.4e-00	0.39	1.6e-02	1.04	3.3e-01	1.05	4.7e-00	0.44	4.3e-00	1.105

AFW$_\ell$-based discretization with $\ell = 0$ and adaptive refinement via Θ											
DOF	it	e($\boldsymbol{\sigma}$)	r($\boldsymbol{\sigma}$)	e($\mathbf{u}$)	r($\mathbf{u}$)	e(p)	r(p)	e($\vec{\mathbf{t}}$)	r($\vec{\mathbf{t}}$)	Θ	eff(Θ)
10911	11	1.3e+01	–	9.8e-02	–	2.0e-00	–	1.5e+01	–	1.2e+01	1.267
34300	11	7.0e-00	1.64	7.4e-02	0.72	9.7e-01	1.93	8.2e-00	1.66	7.2e-00	1.140
114721	11	5.0e-00	0.86	6.2e-02	0.44	4.6e-01	1.88	5.6e-00	0.96	5.3e-00	1.043
314569	10	3.6e-00	1.01	3.9e-02	1.41	2.3e-01	2.02	3.9e-00	1.09	3.8e-00	1.013
775808	10	2.6e-00	1.08	2.8e-02	1.14	1.5e-01	1.40	2.8e-00	1.10	2.8e-00	1.002

Table 4.5 [Example 4.3, $\ell = 0$] Comparison of the mixed approximations with quasi-uniform and adaptive refinements for the $\mu(I)$-rheology model.

we display the initial mesh, the computed magnitude of the velocity and symmetric part of the velocity gradient, and pressure field, which were built using the mixed PEERS$_0$-based scheme on a mesh with $23,390$ triangle elements (actually representing $597,375$ DOF) obtained via Θ (cf. (4.1)). Similarly to [43, Example 3 in Section 7], we observe higher velocities along the top and bottom boundaries, moving rightward and leftward, respectively, as anticipated. Additionally, a circulation phenomenon emerges near the lateral boundaries, driven by the fact that the fluid cannot enter or exit through the circular obstacles. Most of the variations in both the pressure field and the magnitude of the symmetric part of the velocity gradient tensor are concentrated around the circular obstacles. Notably, between the obstacles and in some central regions of the domain, the magnitude of the symmetric part of the velocity gradient is either zero or nearly so, indicating zones where the original viscosity η (cf. [43,

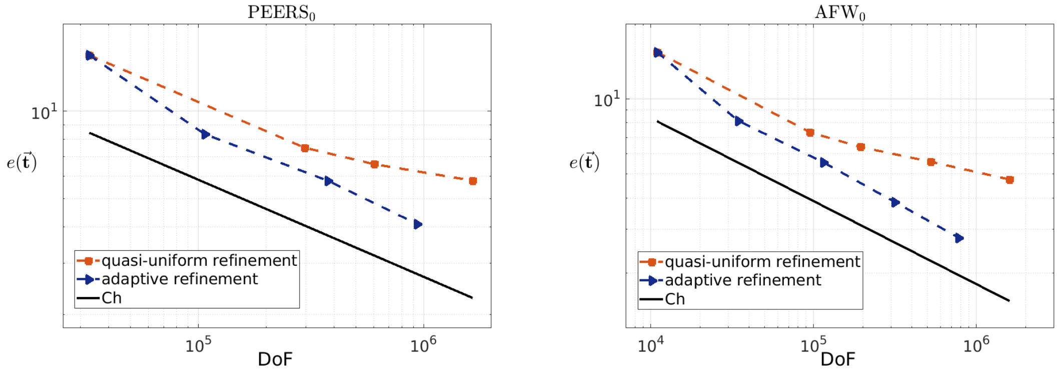


Figure 4.4 [Example 4.3] Log-log plot of $\mathrm{e}(\vec{\mathbf{t}})$ vs. `DOF` for quasi-uniform/adaptive refinements for PEERS_0 and AFW_0-based discretizations (left and right plots, respectively).

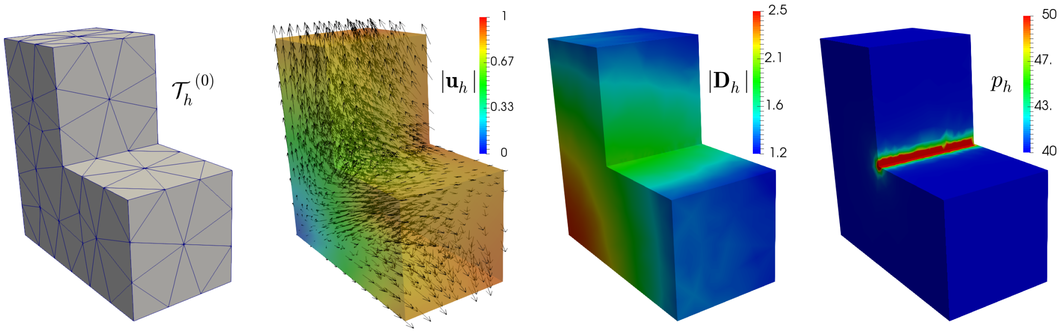


Figure 4.5 [Example 4.3] Initial mesh, computed magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.

eq. (2.9)]) becomes singular and the granular flow remains static. This behavior is consistent with the velocity field and is properly handled by the mixed formulations using the regularized viscosity (3.11). The results align with those reported in [43], now incorporating an adaptive mesh refinement strategy driven by the *a posteriori* error indicator Θ. Snapshots of some of the adapted meshes are shown in Figure 4.8, where we can clearly observe refinement concentrated around the obstacles and in regions where the velocity gradient vanishes or is nearly zero. This confirms that the indicator Θ successfully identifies both the singular zones and the areas with large solution variations, as intended.

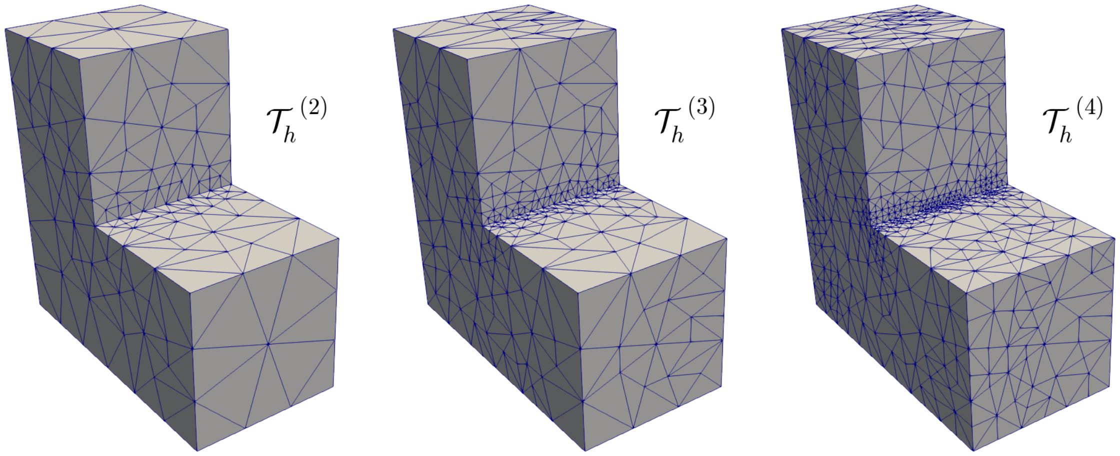


Figure 4.6 [Example 4.3] Three snapshots of adapted meshes according to the indicator Θ for the AFW_0-based discretization.

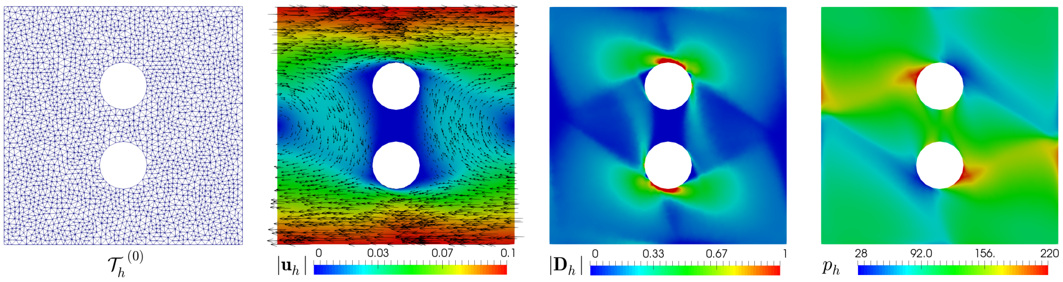
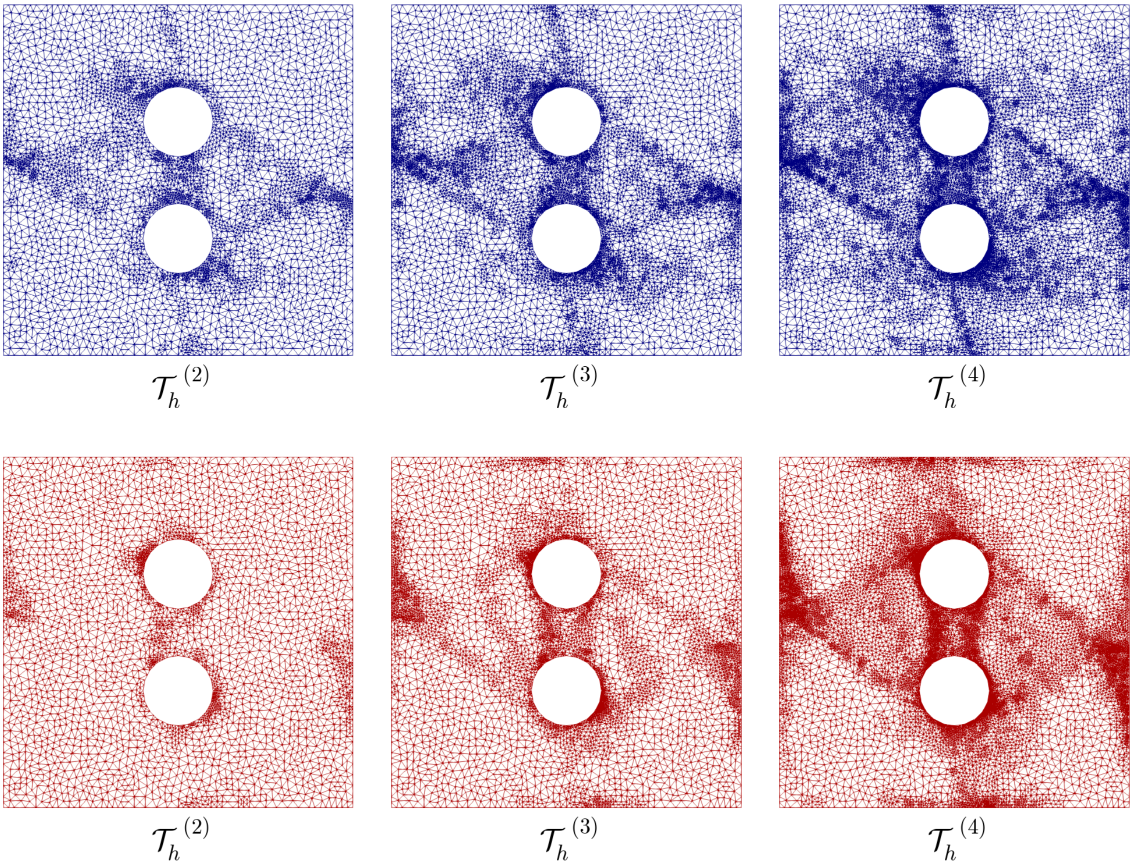


Figure 4.7 [Example 4.4] Initial mesh, computed magnitude of the velocity and symmetric part of the velocity gradient, and pressure field.



$\mathcal{T}_h^{(2)}$ $\mathcal{T}_h^{(3)}$ $\mathcal{T}_h^{(4)}$

$\mathcal{T}_h^{(2)}$ $\mathcal{T}_h^{(3)}$ $\mathcal{T}_h^{(4)}$

Figure 4.8 [Example 4.4] Three snapshots of adapted meshes according to the indicator Θ for PEERS$_0$ and AFW$_0$-based discretizations (top and bottom plots, respectively).

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, a mixed finite element method was developed and implemented for analyzing the rheological behavior of granular materials under the $\mu(I)$ model. The proposed method proved effective in capturing the inherent complexities of granular flow, particularly in regimes where transitions between solid and fluid behaviors are critical. The mixed approach allowed for the simultaneous incorporation of primary and secondary variables, such as pressure and velocity, ensuring greater accuracy and numerical stability in large deformation simulations.

One of the main challenges in working with the $\mu(I)$ model is its intrinsic nonlinearity. Unlike Newtonian fluids where viscosity is constant, in granular materials the friction (or effective viscosity) is also strongly influenced by normal pressure (or confining pressure). This coupling between dissipation and pressure leads to highly nonlinear equations that are significantly more challenging to solve numerically.

Traditional numerical schemes for incompressible fluids - such as pressure-correction projection methods, typically separate velocity and pressure calculations into distinct steps. However, this classical approach becomes inadequate for the $\mu(I)$ model, since dissipation explicitly depends on pressure, requiring more sophisticated techniques to ensure simulation accuracy and stability.

The solution approach treated the stress tensor as a new unknown in the system. This enabled formulating an explicit expression for pressure, using a fixed-point operator to resolve the stress tensor to pressure dependence. This strategy allowed partial decoupling of variables and made numerical implementation feasible.

The numerical implementation faced significant convergence challenges, particularly in regions with very low local deformation rates where system singularities caused pronounced numerical instabilities. The use of adaptive meshing allowed for more precise identification of these singularity regions and helped recover the lost convergence order in these areas. Additionally, the application of regularization techniques proved effective in addressing fundamental modeling issues. However, in domains with extensive singular regions, more intensive regularization was required. This behavior suggests that investigating alternative regularization schemes - potentially involving modified variational formulations or non-local operators - could not only improve numerical stability but also positively impact the continuous formulation of the problem itself.

Despite these difficulties, the developed method demonstrated great potential for practical applications in geotechnics, material processing, and mining industries. Future work could explore extending the method to include thermal effects and material heterogeneities. In summary, this thesis contributed to the advancement of computational modeling of granular materials, offering a robust and versatile tool for analyzing complex phenomena in granular rheology.

5.2 Future Work

The development of the mixed finite element method for $\mu(I)$ rheology presented in this thesis opens the way for several research directions and improvements. The main future lines of investigation include:

- **Influence of the regularization parameter:** Systematic investigation of the method's sensitivity to the regularization parameter. Studies could be conducted to assess how different values of this parameter affect the accuracy and stability of simulations, especially in highly nonlinear regimes. This would allow establishing guidelines for the appropriate choice of the parameter in different applied contexts.

- **Experimental validation:** Comparison of numerical results with experimental data, such as measurements of velocity profiles, pressure, and deformation in granular flows. This validation is crucial to consolidate the reliability of the method and ensure its accuracy in real scenarios.

- **Cross-validation with other numerical methods:** Comparison of the proposed method with other numerical approaches, such as finite difference methods or particle-based discretizations (e.g., DEM - *Discrete Element Method*). This cross-validation would allow identifying relative advantages and limitations of the developed method.

- **Temporal evolution of the problem:** Extension of the method to time-dependent problems, enabling the dynamic analysis of granular phenomena, such as the collapse of material columns or debris spreading in geophysical flows. The incorporation of transient terms in the model would allow studying the temporal evolution of these phenomena, capturing aspects such as shock wave propagation and the formation of complex flow patterns.
- **Multiphase flows:** Study of flows where granular material interacts with fluids or other phases. A relevant example is the simulation of the collapse of a column of particles immersed in a fluid, a classic problem that combines granular solid mechanics with hydrodynamics. Modeling these scenarios would require integrating the current method with governing equations for the fluid, such as the Navier-Stokes equations, and implementing phase-coupling techniques.
- **Application to real situations with complex geometry:** Adaptation of the method to handle irregular geometries and realistic boundary conditions, such as landslides, flows in silos, or industrial granular transport processes. This application would broaden the scope of the method, contributing to solving practical problems in engineering and geosciences.

In summary, future work could explore the temporal evolution of the problem, multiphase flows, the influence of the regularization parameter, experimental and numerical validations, and application to real-world scenarios. These advances would consolidate the method as a robust and versatile tool for computational modeling of granular materials in theoretical and practical contexts.

Bibliography

[1] P. A. Cundall and O. D. L. Strack. A discrete numerical model for granular assemblies. *Geotechnique*, 29:47–65, 1979.

[2] B. Andreotti, Y. Forterre, and O. Pouliquen. *Granular Media: Between Fluid and Solid*. Cambridge University Press, 2013.

[3] Food and Agriculture Organization of the United Nations. Agricultural production statistics 2010–2023. Technical Report 96, FAO, Rome, 2024.

[4] Louise Gallagher and Pascal Peduzzi. Sand and sustainability: Finding new solutions for environmental governance of global sand resources. Technical report, Université de Genève, 2019.

[5] S. B. Savage and K. Hutter. The motion of a finite mass of granular material down a rough incline. *Journal of Fluid Mechanics*, 199:177–215, 1989.

[6] GDR-MiDi-Group. On dense granular flows. *European Journal of Physics E*, 14: 341–365, 2004.

[7] P. Jop, Y. Forterre, and O. Pouliquen. A constitutive law for dense granular flows. *Nature*, 441(8):727–730, 2008.

[8] P. Y. Lagrée, L. Staron, and S. Popinet. The granular column collapse as a continuum: validity of a two-dimensional Navier–Stokes model with a $\mu(I)$-rheology. *Journal of Fluid Mechanics*, 686:378–408, 2011.

[9] L. Staron, P. Y. Lagrée, and S. Popinet. Continuum simulation of the discharge of the granular silo. *European Journal of Physics E*, 37:5, 2014.

[10] J. Chauchat and M. Médale. A three-dimensional numerical model for dense granular flows based on the $\mu(I)$ rheology. *Journal of Computational Physics*, 256: 696–712, 2014.

[11] A. Franci and M. Cremonesi. 3d regularized $\mu(I)$-rheology for granular flows simulation. *Journal of Computational Physics*, 378:257–277, 2019.

[12] G. C. Yang, S. C. Yang, L. Jing, C. Y. Kwok, and Y. D. Sobral. Efficient lattice Boltzmann simulation of free-surface granular flows with μ(I)-rheology. *Journal of Computational Physics*, 479:111956, 2023. ISSN 0021-9991.

[13] G. C. Yang, Y. J. Huang, Y. Lu, C. Y. Kwok, Y. D. Sobral, and Q. H. Yao. Frictional boundary condition for lattice Boltzmann modelling of dense granular flows. *Journal of Fluid Mechanics*, 973:A21, 2023. doi: 10.1017/jfm.2023.782.

[14] E. J. Hinch. *Think before you compute: a prelude to computational fluid dynamics*. Cambridge University Press, 2021.

[15] G. A. Benavides, S. Caucao, G. N. Gatica, and A.A. Hopper. A Banach spaces–based analysis of a new mixed–primal finite element method for a coupled flow–transport problem. *Computational Methods in Applied Mechanics and Engineering*, 371:113285, 2020.

[16] J. Camaño, C. García, and R. Oyarzúa. Analysis of a momentum conservative mixed-FEM for the stationary Navier-Stokes problem. *Numerical Methods Partial Differential Equations*, 37(5):2895–2923, 2021.

[17] S. Caucao, R. Oyarzúa, and S. Villa-Fuentes. A new mixed-FEM for steady-state natural convection models allowing conservation of momentum and thermal energy. *Calcolo*, 57(4):Paper No. 36, 2020.

[18] S. Caucao and I. Yotov. A Banach space mixed formulation for the unsteady Brinkman-Forchheimer equations. *IMA Journal of Numerical Analysis*, 41(4): 2708–2743, 2021.

[19] E. Colmenares, G. N. Gatica, and S. Moraga. A Banach spaces-based analysis of a new fully-mixed finite element method for the Boussinesq problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(5):1525–1568, 2020.

[20] E. Colmenares, G. N. Gatica, and J. C. Rojas. A Banach spaces-based mixed-primal finite element method for the coupling of Brinkman flow and nonlinear transport. *Calcolo*, 59(4):Paper No. 51, 2022.

[21] E. Colmenares and M. Neilan. Dual-mixed finite element methods for the stationary Boussinesq problem. *Computers and Mathematics with Applications*, 72(7):1828–1850, 2016.

[22] G. N. Gatica, R. Oyarzúa, R. Ruiz-Baier, and Y.D. Sobral. Banach spaces-based analysis of a fully-mixed finite element method for the steady-state model of fluidized beds. *Computers and Mathematics with Applications*, 84:244–276, 2021.

[23] M. Ainsworth and J.T. Oden. A posteriori error estimators for the Stokes and Oseen equations. *SIAM Journal on Numerical Analysis*, 34(1):228–245, 1997.

[24] A. Alonso. Error estimators for a mixed method. *Numerische Mathematik*, 74(4): 385–395, 1996.

[25] C. Carstensen. A posteriori error estimate for the mixed finite element method. *Mathematics of Computation*, 66(218):465–476, 1997.

[26] C. Carstensen and G. Dolzmann. A posteriori error estimates for mixed FEM in elasticity. *Numerische Mathematik*, 81(2):187–209, 1998.

[27] M. Lonsing and R. Verfürth. A posteriori error estimators for mixed finite element methods in linear elasticity. *Numerische Mathematik*, 97(4):757–778, 2004.

[28] S. Repin, S. Sauter, and A. Smolianski. Two-sided a posteriori error estimates for mixed formulations of elliptic problems. *SIAM Journal on Numerical Analysis*, 45 (3):928–945, 2007.

[29] J. T. Oden, W. Wu, and M. Ainsworth. An a posteriori error estimate for finite element approximations of the Navier-Stokes equations. *Computational Methods in Applied Mechanics and Engineering*, 111(1-2):185–202, 1994.

[30] R. Verfürth. A posteriori error estimators and adaptive mesh-refinement techniques for the Navier–Stokes equations. *Incompressible computational fluid dynamics: trends and advances*, pages 447–475, 2008.

[31] R. Verfürth. A posteriori error estimates for non-linear problems. finite element discretizations of elliptic equations. *SIAM Journal on Numerical Analysis*, 62 (206):445–475, 1994.

[32] M. Ainsworth and J.T. Oden. *A Posterori Error Estimation in Finite Element Analysis*. Wiley-Interscience [John Wiley & Sons], 2000.

[33] M. Farhloul, S. Nicaise, and L. Paquet. A priori and a posteriori error estimations for the dual mixed finite element method of the Navier-Stokes problem. *Numerical Methods for Partial Differential Equations*, 25(4):843–869, 2009.

[34] A. Allendes, E. Otarola, and A.J. Salgado. A posteriori error estimates for the stationary Navier–Stokes equations with Dirac measures. *SIAM Journal on Scientific Computing*, 42(3):A1860–A1884, 2020.

[35] G. Kanschat and D. Schötzau. Energy norm a posteriori error estimation for divergence-free discontinuous Galerkin approximations of the Navier–Stokes equations. *International Journal of Numerical Methods in Fluids*, 57(9):1093–1113, 2008.

[36] G. N. Gatica, R. Ruiz-Baier, and G. Tierra. A posteriori error analysis of an augmented mixed method for the Navier-Stokes equations with nonlinear viscosity. *Computers and Mathematics with Applications*, 72(9):2289–2310, 2016.

[37] J. Camaño, G. N. Gatica, R. Oyarzúa, and R. Ruiz-Baier. An augmented stress-based mixed finite element method for the steady state Navier-Stokes equations with nonlinear viscosity. *Numerical Methods for Partial Differential Equations*, 33 (5):1692–1725, 2017.

[38] S. Caucao, G. N. Gatica, and R. Oyarzúa. A posteriori error analysis of a fully-mixed formulation for the Navier–Stokes/Darcy coupled problem with nonlinear viscosity. *Computer Methods in Applied Mechanics and Engineering*, 315:943–971, 2017.

[39] J. Camaño, S. Caucao, R. Oyarzúa, and S. Villa-Fuentes. A posteriori error analysis of a momentum conservative Banach spaces based mixed-FEM for the Navier–Stokes problem. *Applied Numerical Mathematics*, 176:134–158, 2022.

[40] S. Caucao, R. Oyarzúa, and S. Villa-Fuentes. A posteriori error analysis of a momentum and thermal energy conservative mixed-FEM for the Boussinesq equations. *Calcolo*, 59(4):Paper No. 45, 40 pp., 2022.

[41] G. N. Gatica, C. Inzunza, R. Ruiz-Baier, and F. Sandoval. A posteriori error analysis of Banach spaces-based fully-mixed finite element methods for Boussinesq-type models. *Journal of Numerical Mathematics*, 30(4):325–356, 2022.

[42] S. Caucao, G. N. Gatica, and J.P. Ortega. A posteriori error analysis of a Banach spaces-based fully mixed FEM for double-diffusive convection in a fluid-saturated porous medium. *Computational Geosciences*, 27(2):289–316, 2023.

[43] S. Caucao, G. N. Gatica, S.R. Medrado, and Y.D. Sobral. Nonlinear twofold saddle point-based mixed finite element methods for a regularized $\mu(I)$-rheology model of granular materials. *Journal of Computational Physics*, 520:113462, 2025.

[44] Pierre-Arnaud Raviart and Jean-Marie Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* Masson, Paris, 1983.

[45] A. Ern and J.-L Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences.* Springer-Verlag, 2004.

[46] G. N. Gatica. *A Simple Introduction to the Mixed Finite Element Method. Theory and Applications.* SpringerBriefs in Mathematics. Springer, Cham, 2014.

[47] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, 1978.

[48] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Universitext. Springer New York, 2011.

[49] V. Girault and P.A. Raviart. *Finite Element Approximation of the Navier-Stokes Equations.* Lecture notes in mathematics. Springer-Verlag, 1979.

[50] G. N. Gatica. Solvability and Galerkin approximations of a class of nonlinear operator equations. *Journal for Analysis and its Applications*, 21(3):761–781, 2002.

[51] V. Girault and P.A. Raviart. *Finite element methods for Navier-Stokes equations: theory and algorithms*, volume 5. Springer-Verlag, 1986.

[52] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS project version 1.5. *Archives of Numerical Software*, 3:9–23, 2015.

[53] R. R. Kerswell. Dam break with Coulomb friction: A model for granular slumping? *Physics of Fluids*, 17(5):057101, 2005.

[54] T. C. Papanastasiou. Flows of materials with yield. *Journal of Rheology*, 31(5): 385–404, 1987.

[55] L. Jing, C. Y. Kwok, Y. F. Leung, and Y. D. Sobral. Characterization of base roughness for granular chute flows. *Physical Review E*, 94:052901, 2016.

[56] I. Bermúdez, C. I. Correa, G. N. Gatica, and J. P. Silva. A perturbed twofold saddle point-based mixed finite element method for the Navier-Stokes equations with variable viscosity. *Applied Numerical Mathematics*, 201:465–487, 2024.

[57] Y. H. Wu, J. M. Hill, and A. Yu. A finite element method for granular flow through a frictional boundary. *Communications in Nonlinear Science and Numerical Simulation*, 12(4):486–495, 2007.

[58] J. Galvis and M. Sarkis. Non-matching mortar discretization analysis for the coupling Stokes-Darcy equations. *Electronic Transactions on Numerical Analysis*, 26:350–384, 2007.

[59] G. N. Gatica, R. Oyarzúa, and F.-J. Sayas. A twofold saddle point approach for the coupling of fluid flow with nonlinear porous media flow. *IMA Journal of Numerical Analysis*, 32(3):845–887, 2012.

[60] S. Caucao, G. N. Gatica, and F. Sandoval. A fully-mixed finite element method for the coupling of the Navier-Stokes and Darcy-Forchheimer equations. *Numerical Methods Partial Differential Equations*, 37(3):2550–2587, 2021.

[61] G. N. Gatica and W. L. Wendland. Coupling of mixed finite elements and boundary elements for linear and nonlinear elliptic problems. *Applicable Analysis*, 63(1-2): 39–75, 1996.

[62] G. N. Gatica, N. Núñez, and R. Ruiz-Baier. New non-augmented mixed finite element methods for the Navier-Stokes-Brinkman equations using Banach spaces. *Journal of Numerical Mathematics*, 31(4):343–373, 2023.

[63] G. N. Gatica, N. Heuer, and S. Meddahi. On the numerical analysis of nonlinear twofold saddle point problems. *IMA Journal of Numerical Analysis*, 23(2):301–330, 2003.

[64] D. N. Arnold, F. Brezzi, and J. Douglas. PEERS: A new mixed finite element method for plane elasticity. *Japan Journal of Applied Mathematics*, 1:347–367, 1984.

[65] M. Lonsing and R. Verfürth. On the stability of BDMS and PEERS elements. *Numerische Mathematik*, 99(1):131–140, 2004.

[66] D. N. Arnold, R. S. Falk, and R. Winther. Mixed finite element methods for linear elasticity with weakly imposed symmetry. *Mathematics of Computation*, 76(260): 1699–1723, 2007.

[67] D. Boffi, F. Brezzi, and M. Fortin. *Mixed Finite Element Methods and Applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2013.

[68] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, 1991.

[69] L. Jing, J. M. Ottino, P. B. Umbanhowar, and R. M. Lueptow. Drag force in granular shear flows: regimes, scaling laws and implications for segregation. *Journal of Fluid Mechanics*, 948:A24, 2022.

[70] E. Creusé, M. Farhloul, and L. Paquet. A posteriori error estimation for the dual mixed finite element method for the p-Laplacian in a polygonal domain. *Computational Methods in Applied Mechanics and Engineering*, 196(25–28):2570–2582, 2007.

[71] C. Domínguez, G. N. Gatica, and S. Meddahi. A posteriori error analysis of a fully-mixed finite element method for a two-dimensional fluid-solid interaction problem. *Journal of Computational Mathematics*, 33(6):606–641, 2015.

[72] S. Caucao, D. Mora, and R. Oyarzúa. A priori and a posteriori error analysis of a pseudostress-based mixed formulation of the Stokes problem with varying density. *IMA Journal of Numerical Analysis*, 36(2):947–983, 2016.

[73] T.P. Barrios, G. N. Gatica, M. González, and N. Heuer. A residual based a posteriori error estimator for an augmented mixed finite element method in linear elasticity. *Mathematical Modelling and Numerical Analysis*, 40(5):843–869, 2006.

[74] G. N. Gatica, A. Márquez, and M.A. Sánchez. Analysis of a velocity-pressure-pseudostress formulation for the stationary Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 199(17-20):1064–1079, 2010.

[75] R. Verfurth. *A Review of Posteriori Error Estimation & Adaptive Mesh-Refinement Techniques*. Wiley, 1996.

[76] A. Plaza and G. F. Carey. Local refinement of simplicial grids based on the skeleton. *Applied Numerical Mathematics*, 32(2):195–218, 2000.

[77] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Modélisation Mathématique et Analyse Numérique*, 9:77–84, 1975.

[78] G. N. Gatica, L.F. Gatica, and F.A. Sequeira. A priori and a posteriori error analyses of a pseudostress-based mixed formulation for linear elasticity. *Computers and Mathematics with Applications*, 71(2):585–614, 2016.

[79] S. Caucao, G. N. Gatica, R. Oyarzúa, and F. Sandoval. Residual-based a posteriori error analysis for the coupling of the Navier–Stokes and Darcy–Forchheimer equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(2):659–687, 2021.

[80] S. Agmon. *Lectures on elliptic boundary value problems*. Van Nostrand, 1965.

[81] G. N. Gatica. A note on stable helmholtz decompositions in 3D. *Applicable Analysis*, 99(7):1110–1121, 2020.

Appendix A

The hypotheses on the viscosity

A.1 The hypotheses on the viscosity

In this appendix we refer to the regularized viscosity η and the corresponding fulfillment of the hypotheses (**H.1**), (**H.2**), and (**H.3**). We begin by recalling from (3.11) that

$$\eta(\varrho,\omega) \;:=\; \frac{a_1\,\varrho}{\omega+\varepsilon} \;+\; \frac{a_2\,\varrho}{a_3\,\sqrt{\varrho}+a_4\,\omega+\varepsilon} \qquad \forall\,(\varrho,\omega)\in \mathrm{R}^+\times\mathrm{R}^+\,, \tag{A.1}$$

so that

$$\frac{\partial}{\partial\omega}\eta(\varrho,\omega) \;=\; -\,\frac{a_1\,\varrho}{(\omega+\varepsilon)^2} \;-\; \frac{a_2\,a_4\,\varrho}{(a_3\,\sqrt{\varrho}+a_4\,\omega+\varepsilon)^2} \qquad \forall\,(\varrho,\omega)\in \mathrm{R}^+\times\mathrm{R}^+\,, \tag{A.2}$$

and then

$$\eta(\varrho,\omega) \;+\; \omega\frac{\partial}{\partial\omega}\eta(\varrho,\omega) \;=\; \frac{a_1\,\varrho\,\varepsilon}{(\omega+\varepsilon)^2} \;+\; \frac{a_2\,(a_3\,\sqrt{\varrho}+\varepsilon)\,\varrho}{(a_3\,\sqrt{\varrho}+a_4\,\omega+\varepsilon)^2} \qquad \forall\,(\varrho,\omega)\in \mathrm{R}^+\times\mathrm{R}^+\,. \tag{A.3}$$

Thus, in order to satisfy (**H.1**) and (**H.2**), we restrict the evaluation of η, as defined by (A.1), to a given rectangle $[\varrho_1,\varrho_2]\times[\omega_1,\omega_2]\subseteq \mathrm{R}^+\times\mathrm{R}^+$, so that η is extended by continuity outside this region, as illustrated in Figure A.1 below.

In this way, it is possible to accomplish the aforementioned hypotheses with positive constants η_1 and η_2, depending on ϱ_1, ϱ_2, ω_1, ω_2, ε, and the coefficients a_i, $i\in\Big\{1,...,4\Big\}$, defined in (3.10). Note also that, under this modification, one could even get rid of the parameter ε.

On the other hand, regarding (**H.3**), we show next that it is satisfied with a positive constant L_η depending only on the coefficients a_1, a_2, and a_4 (cf. (3.10)). Indeed, given

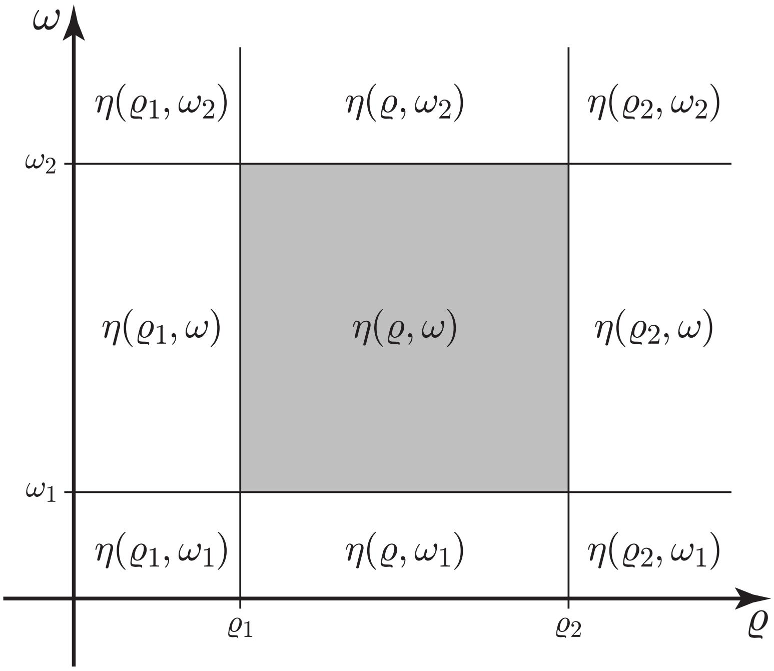


Figure A.1 Graphic representation of the modified version of the viscosity function η.

ϱ, χ, and ω in R$^+$, we first deduce from (A.1) and some algebraic manipulations, that

$$\Big\{\eta(\varrho,\omega) - \eta(\chi,\omega)\Big\}\,\omega$$
$$= \left\{\frac{a_1\,\omega}{\omega+\varepsilon} + \frac{a_2\,a_3\,\sqrt{\varrho}\,\sqrt{\chi}\,\omega}{(\sqrt{\varrho}+\sqrt{\chi})\,\mathbf{a}(\varrho,\omega,\varepsilon)\,\mathbf{a}(\chi,\omega,\varepsilon)} + \frac{a_2\,\omega\,(a_4\,\omega+\varepsilon)}{\mathbf{a}(\varrho,\omega,\varepsilon)\,\mathbf{a}(\chi,\omega,\varepsilon)}\right\}(\varrho-\chi)\,, \tag{A.4}$$

where

$$\mathbf{a}(\varrho,\omega,\varepsilon) := a_3\,\sqrt{\varrho} + a_4\,\omega + \varepsilon\,,$$

and analogously for $\mathbf{a}(\chi,\omega,\varepsilon)$. In order to bound the right-hand side of (A.4) we first observe that

$$\frac{a_1\,\omega}{\omega+\varepsilon} \le a_1\,. \tag{A.5}$$

Then, it is straightforward to show that

$$\frac{\sqrt{\varrho}}{\sqrt{\varrho}+\sqrt{\chi}} \le 1\,, \qquad \frac{a_3\,\sqrt{\chi}}{\mathbf{a}(\chi,\omega,\varepsilon)} = \frac{a_3\,\sqrt{\chi}}{a_3\,\sqrt{\chi}+a_4\,\omega+\varepsilon} \le 1\,, \quad \text{and}$$
$$\frac{a_2\,\omega}{\mathbf{a}(\varrho,\omega,\varepsilon)} = \frac{a_2\,a_4\,\omega}{a_4\,(a_3\,\sqrt{\varrho}+a_4\,\omega+\varepsilon)} \le \frac{a_2}{a_4}\,, \tag{A.6}$$

which yield

$$\frac{a_2\, a_3 \sqrt{\varrho}\,\sqrt{\chi}\,\omega}{(\sqrt{\varrho}+\sqrt{\chi})\,\mathbf{a}(\varrho,\omega,\varepsilon)\,\mathbf{a}(\chi,\omega,\varepsilon)} \;\le\; \frac{a_2}{a_4}\,. \tag{A.7}$$

In turn, it is readily seen that

$$\frac{a_4\,\omega+\varepsilon}{\mathbf{a}(\chi,\omega,\varepsilon)} \;=\; \frac{a_4\,\omega+\varepsilon}{a_3\sqrt{\chi}+a_4\,\omega+\varepsilon} \;\le\; 1\,,$$

which, along with the third inequality from (A.6), imply

$$\frac{a_2\,\omega\,(a_4\,\omega+\varepsilon)}{\mathbf{a}(\varrho,\omega,\varepsilon)\,\mathbf{a}(\chi,\omega,\varepsilon)} \;\le\; \frac{a_2}{a_4}\,. \tag{A.8}$$

Finally, employing (A.5), (A.7), and (A.8) in (A.4), we arrive at

$$\Big|\eta(\varrho,\omega)-\eta(\chi,\omega)\Big|\,\omega \;\le\; L_\eta\,|\varrho-\chi|\,, \tag{A.9}$$

where, using (3.10),

$$L_\eta \;:=\; a_1 \;+\; \frac{2\,a_2}{a_4} \;=\; \Big(2\,\mu_d-\mu_s\Big)\sqrt{2}\,,$$

thus proving (**H.3**).

A.2 Preliminaries for reliability

We begin by introducing useful notations to describe local information on elements and edges. For each $K \in \mathcal{T}_h$, let $\mathcal{E}(K)$ denote its set of edges, and let $\mathcal{E}_h$ be the set of all edges in $\mathcal{T}_h$, with corresponding diameters h_e. We further decompose $\mathcal{E}_h$ as $\mathcal{E}_h = \mathcal{E}_h(\Omega) \cup \mathcal{E}_h(\Gamma)$, where $\mathcal{E}_h(\Omega) := \{e \in \mathcal{E}_h : e \subseteq \Omega\}$ and $\mathcal{E}_h(\Gamma) := \{e \in \mathcal{E}_h : e \subseteq \Gamma\}$. For each $e \in \mathcal{E}_h$, we fix unit normal and tangential vectors, denoted by $\boldsymbol{\nu}_e := (\nu_1, \nu_2)^{\mathrm{t}}$ and $\mathbf{s}_e := (-s_2, s_1)^{\mathrm{t}}$, respectively. When no ambiguity arises, we will simply write $\boldsymbol{\nu}$ and $\mathbf{s}$. The usual jump operator $[\![\cdot]\!]$ across an internal edge $e \in \mathcal{E}_h(\Omega)$ is defined for a piecewise continuous tensor valued function $\boldsymbol{\zeta}$ as $[\![\boldsymbol{\zeta}]\!] := \boldsymbol{\zeta}|_K - \boldsymbol{\zeta}|_{K'}$, where K and K' are the elements of $\mathcal{T}_h$ sharing e. Finally, for a scalar field ϕ, a vector field $\mathbf{v} := (v_1, v_2)^t$, and a matrix-valued field $\boldsymbol{\tau} := (\tau_{ij})_{2\times 2}$, we define:

$$\mathbf{curl}\,(\phi) := \left(\frac{\partial \phi}{\partial x_2}, -\frac{\partial \phi}{\partial x_1}\right)^{\mathrm{t}}, \quad \underline{\mathbf{curl}}\,(\mathbf{v}) := \begin{pmatrix} \mathbf{curl}\,(v_1)^{\mathrm{t}} \\ \mathbf{curl}\,(v_2)^{\mathrm{t}} \end{pmatrix},$$

$$\mathrm{rot}\,(\mathbf{v}) := \frac{\partial v_1}{\partial x_2} - \frac{\partial v_2}{\partial x_1}, \quad \text{and} \quad \mathbf{rot}\,(\boldsymbol{\tau}) := \begin{pmatrix} \mathrm{rot}\,(\tau_{11}, \tau_{12}) \\ \mathrm{rot}\,(\tau_{21}, \tau_{22}) \end{pmatrix},$$

where the derivatives involved are taken in the distributional sense.

Let us now recall the main properties of the Raviart–Thomas and Clément interpolation operators (cf. [45], [77]). We begin by defining, for each $p \geq 2n/(n+2)$, the spaces

$$\mathbf{W}_p(\Omega) := \left\{ \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div}_p; \Omega) : \quad \boldsymbol{\tau}|_K \in \mathbf{W}^{1,p}(K), \quad \forall\, K \in \mathcal{T}_h \right\}, \tag{A.10}$$

and

$$\mathrm{RT}_\ell(\Omega) := \left\{ \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div}_p; \Omega) : \quad \boldsymbol{\tau}|_K \in \mathbf{RT}_\ell(K), \quad \forall\, K \in \mathcal{T}_h \right\}. \tag{A.11}$$

In addition, we let $\Pi_h^\ell : \mathbf{W}_p(\Omega) \to \mathrm{RT}_\ell(\Omega)$ be the Raviart–Thomas interpolation operator, which is characterized for each $\boldsymbol{\tau} \in \mathbf{W}_p(\Omega)$ by the identities (see, e.g. [45, Section 1.2.7])

$$\int_e \left(\Pi_h^\ell(\boldsymbol{\tau}) \cdot \boldsymbol{\nu}\right) \xi = \int_e (\boldsymbol{\tau} \cdot \boldsymbol{\nu})\, \xi \quad \forall\, \xi \in \mathrm{P}_k(e), \quad \forall \text{ edge or face } e \text{ of } \mathcal{T}_h, \tag{A.12}$$

when $k \geq 0$, and

$$\int_K \Pi_h^\ell(\boldsymbol{\tau}) \cdot \boldsymbol{\psi} = \int_K \boldsymbol{\tau} \cdot \boldsymbol{\psi} \quad \forall\, \boldsymbol{\psi} \in \mathbf{P}_{\ell-1}(K), \quad \forall\, K \in \mathcal{T}_h, \tag{A.13}$$

when $k \geq 1$. In turn, given $q > 1$ such that $1/p + 1/q = 1$, we let

$$\mathrm{P}_\ell(\Omega) \,:=\, \Big\{ v \in \mathrm{L}^q(\Omega) : \quad v|_K \,\in\, \mathrm{P}_\ell(K)\,, \quad \forall\, K \in \mathcal{T}_h \Big\}\,, \tag{A.14}$$

and recall from [45, Lemma 1.41] that there holds

$$\mathrm{div}(\Pi_h^\ell(\boldsymbol{\tau})) \,=\, \mathcal{P}_h^\ell(\mathrm{div}(\boldsymbol{\tau}))\,, \quad \forall\, \boldsymbol{\tau} \in \mathbf{W}_p(\Omega)\,, \tag{A.15}$$

where $\mathcal{P}_h^\ell : \mathrm{L}^2(\Omega) \to \mathrm{P}_\ell(\Omega)$ denotes the standard orthogonal projector with respect to the $\mathrm{L}^2(\Omega)$-inner product. This operator satisfies the following error estimate (see [45, Proposition 1.135]): there exists a positive constant C_0, independent of h, such that for $0 \leq l \leq \ell + 1$ and $1 \leq p \leq \infty$, the following holds

$$\|w - \mathcal{P}_h^\ell(w)\|_{0,p;\Omega} \,\leq\, C_0\, h^l\, \|w\|_{l,p;\Omega} \qquad \forall\, w \in \mathrm{W}^{l,p}(\Omega)\,. \tag{A.16}$$

We stress that $\mathcal{P}_h^\ell(w)|_K \,=\, \mathcal{P}_K^\ell(w|_K)\, \forall w \in \mathrm{L}^p(\Omega)$, where $\mathcal{P}_K^\ell \,:\, \mathrm{L}^p(K) \to \mathrm{P}_\ell(K)$ is corresponding local orthogonal projector. In addition, denoting by $\mathbf{P}_\ell(\Omega)$ the vector version of $\mathrm{P}_\ell(\Omega)$ (cf. (A.10)), we let $\boldsymbol{\mathcal{P}}_h^\ell : \mathbf{L}^2(\Omega) \to \mathbf{P}_\ell(\Omega)$ be the vector version of $\mathcal{P}_h^\ell$.

Next, we collect some approximation proprieties of Π_h^ℓ.

Lemma A.2.1. Given $p > 1$, there exist positive constants C_1, C_2, independent of h, such that for $0 \leq l \leq \ell$, and for each $K \in \mathcal{T}_h$, there holds

$$\|\boldsymbol{\tau} - \Pi_h^\ell(\boldsymbol{\tau})\|_{0,p;K} \,\leq\, C_1\, h_K^{l+1}\, |\boldsymbol{\tau}|_{l+1,p;K} \qquad \forall\, \boldsymbol{\tau} \in \mathbf{W}^{l+1,p}(K) \tag{A.17}$$

and

$$\|\boldsymbol{\tau}\cdot\boldsymbol{\nu} - \Pi_h^\ell(\boldsymbol{\tau})\cdot\boldsymbol{\nu}\|_{0,p;e} \,\leq\, C_2\, h_e^{1-1/p}\, |\boldsymbol{\tau}|_{1,p;K} \qquad \forall\, \boldsymbol{\tau} \in \mathbf{W}^{1,p}(K)\,, \quad \forall\, e \in \mathcal{E}_h(K)\,. \tag{A.18}$$

Proof. For the estimate (A.17) we refer to [41, Lemma 3.1], whereas the proof of (A.18) can be found in [39, Lemma 4.2]. □

Furthermore, denoting by $\mathbb{W}_p(\Omega)$ and $\mathbb{RT}_\ell(\Omega)$ the tensorial versions of $\mathbf{W}_p(\Omega)$ (cf. (A.10)) and $\mathbf{RT}_\ell(\Omega)$ (cf. (A.11)), respectively, we let $\boldsymbol{\Pi}_h^\ell : \mathbb{W}_p(\Omega) \to \mathbb{RT}_\ell(\Omega)$ be the operator Π_h^ℓ acting row-wise. Then, acording to decomposition (3.30), for each

$\boldsymbol{\tau} \in \mathbb{W}_p(\Omega)$ there holds

$$\boldsymbol{\Pi}_h^\ell(\boldsymbol{\tau}) \,:=\, \boldsymbol{\Pi}_{h,0}^\ell(\boldsymbol{\tau}) \,+\, c_0\,\mathbb{I}, \quad \text{with} \quad c_0 \,:=\, \frac{1}{n|\Omega|}\int_\Omega \operatorname{tr}\Big(\boldsymbol{\Pi}_h^\ell(\boldsymbol{\tau})\Big) \in \mathrm{R}$$

$$\text{and} \qquad \boldsymbol{\Pi}_{h,0}^\ell(\boldsymbol{\tau}) \,:=\, \boldsymbol{\Pi}_h^\ell(\boldsymbol{\tau}) - c_0\,\mathbb{I} \,\in \mathbb{RT}_\ell(\Omega) \cap \mathbb{H}_0(\mathbf{div}_{4/3};\Omega)\,.$$

Additional approximation properties of Π_h^ℓ and $\boldsymbol{\Pi}_h^\ell$, particularly those involving the div and **div** operators, can also be established using (A.15) and (A.16), along with their tensorial counterparts for $\boldsymbol{\Pi}_h^\ell$ and $\mathcal{P}_h^\ell$.

We now recall from [39, Lemma 4.4] a stable Helmholtz decomposition for the nonstandard Banach space $\mathbb{H}(\mathbf{div}_p;\Omega)$, which will be used in the forthcoming analysis for the particular case $p = 4/3$. More precisely, we state the following result:

Lemma A.2.2. Given $p \in (1,2)$, there exists a positive constant C_p such that for each $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}_p;\Omega)$ there exist $\boldsymbol{\zeta} \in \mathbb{W}^{1,p}(\Omega)$ and $\boldsymbol{\xi} \in \mathbf{H}^1(\Omega)$ satisfying

$$\boldsymbol{\tau} \,=\, \boldsymbol{\zeta} \,+\, \underline{\mathbf{curl}}\,(\boldsymbol{\xi}) \quad \text{in} \quad \Omega \quad \text{and} \quad \|\boldsymbol{\zeta}\|_{1,p;\Omega} + \|\boldsymbol{\xi}\|_{1,\Omega} \,\leq\, C_p\,\|\boldsymbol{\tau}\|_{\mathbf{div}_p;\Omega}\,. \tag{A.19}$$

On the other hand, let us define $\mathrm{X}_h := \Big\{ v_h \in \mathrm{C}(\overline{\Omega}) : \; v_h|_K \in \mathrm{P}_1(K) \quad \forall\, K \in \mathcal{T}_h \Big\}$ and denote by $\mathbf{X}_h$ its vector-valued counterpart. We consider the Clément interpolation operator $\mathcal{I}_h : \mathrm{H}^1(\Omega) \to \mathrm{X}_h$ and its vector version $\boldsymbol{\mathcal{I}}_h : \mathbf{H}^1(\Omega) \to \mathbf{X}_h$. Some local properties of $\mathcal{I}_h$, and consequently of $\boldsymbol{\mathcal{I}}_h$, corresponding to the particular case of [45, Lemma 1.127] with $m = 2$, $p = 2$, and $\ell = 1$, are established in the following lemma (cf. [77]).

Lemma A.2.3. There exist positive constants C_1 and C_2, such that for each $v \in H^1(\Omega)$ there hold

$$\|v - \mathcal{I}_h(v)\|_{0,K} \,\leq\, C_1\, h_K\, \|v\|_{1,\Delta(K)} \quad \forall\, K \in \mathcal{T}_h \tag{A.20}$$

and

$$\|v - \mathcal{I}_h(v)\|_{0,e} \,\leq\, C_2\, h_e^{1/2}\, \|v\|_{1,\Delta(e)} \quad \forall\, K \in \mathcal{E}_h\,, \tag{A.21}$$

where $\Delta(K) := \cup\Big\{K' \in \mathcal{T}_h : \; K' \cap K \neq \varnothing\Big\}$ and $\Delta(e) := \cup\Big\{K' \in \mathcal{T}_h : \; K' \cap e \neq \varnothing\Big\}$.

A.3 Preliminaries for efficiency

For the efficiency analysis of Θ (cf. (4.1)), we proceed as in [39, 41, 42, 73, 74, 74, 78, 79], and apply the localization technique based on bubble functions, along with inverse and discrete trace inequalities. For the former, given $K \in \mathcal{T}_h$, we let ψ_K be the usual element-bubble function (cf. [75, eq. (1.5)]), satisfying

$$\psi_K \in P_3(K), \quad \sup(\psi_K) \subseteq K, \quad \psi_K = 0 \quad \text{on} \quad \partial K \quad \text{and} \quad 0 \leq \psi_K \leq 1 \quad \text{in} \quad K\,. \tag{A.22}$$

The specific properties of ψ_K to be employed in what follows, are collected in the following lemma, for whose proof we refer to [75, Lemma 3.3].

Lemma A.3.4. Let ℓ be a non-negative integer, and $p, q \in (1, +\infty)$ conjugate to each other, that is, such that $1/p + 1/q = 1$, and $K \in \mathcal{T}_h$. then, there exist positive constants c_1, c_2, and c_3, independent of h and K, but depending on the shape-regularity of the triangulations (minimum angle condition) and ℓ, such that for each $u \in P_\ell(K)$ there hold

$$c_1 \|u\|_{0,p,K} \leq \sup_{0 \neq v \in P_\ell(K)} \frac{\displaystyle\int_K u\,\psi_K\,v}{\|v\|_{0,q,K}} \leq \|u\|_{0,p;K}$$

and

$$c_2 h_K^{-1}\, \|\psi_K\, u\|_{0,q;K} \leq \|\nabla(\psi_K\, u)\|_{0,q;K} \leq c_3 h_K^{-1}\, \|\psi_K\, u\|_{0,q;K}\,.$$

In turn, the aforementioned inverse inequality is stated as follows (cf. [45, Lemma 1.138]).

Lemma A.3.5. Let ℓ, l and m be non-negative integers such that $m \leq l$, and let $r, s \in [1, +\infty]$, and $K \in \mathcal{T}_h$. Then, there exists $c > 0$, independent of h, K, r and s, but depending on ℓ, l, m and the shape of the triangulations, such that

$$\|v\|_{l,r;K} \leq c h_K^{m-l+n(1/r-1/s)} \|v\|_{m,s;K} \quad \forall\, v \in \mathrm{P}_\ell(K)\,. \tag{A.23}$$

Finally, proceeding as in [80, Theorema 3.10], that is employing the usual scaling estimates with respect to a fixed reference element $\widehat{K}$, and applying the trace inequality in $\mathrm{W}^{1,p}(\widehat{K})$, for a given $p \in (1, +\infty)$, one is able to establish the following discrete trace inequality.

Lemma A.3.6. Let $p \in (1, +\infty)$. Then, there exists $c > 0$, depending only on the shape regularity of the triangulations, such that for each $K \in \mathcal{T}_h$ and $e \in \mathcal{E}_h(K)$, there holds

$$\|v\|_{0,p;e}^p \leq c \left\{ h_K^{-1}\, \|v\|_{0,p;K}^p + h_K^{p-1} |v|_{1,p;K}^p \right\} \quad \forall\, v \in \mathrm{W}^{1,p}(K)\,. \tag{A.24}$$

A.4 A posteriori error analysis: the 3D case

In this appendix, we extend the results from Section 4.2 to the three-dimensional version of (3.85). Similarly to the previous section, given a tetrahedron $K \in \mathcal{T}_h$, we denote by $\mathcal{E}_K$ the set of its faces and by $\mathcal{E}$ the set of all faces in the triangulation $\mathcal{T}_h$. We then define $\mathcal{E}_h = \mathcal{E}_h(\Omega) \cup \mathcal{E}_h(\Gamma)$, where $\mathcal{E}_h(\Omega) := \{e \in \mathcal{E}_h : e \subseteq \Omega\}$ and $\mathcal{E}_h(\Gamma) := \{e \in \mathcal{E}_h : e \subseteq \Gamma\}$. For each face $e \in \mathcal{E}_h$, we fix a unit normal vector $\boldsymbol{\nu}_e$. Given $\boldsymbol{\tau} = (\tau_{ij})_{3\times 3} \in \mathbb{L}^2(\Omega)$ such that $\boldsymbol{\tau}|_K \in \mathbb{C}(K)$ for each $K \in \mathcal{T}_h$, we define $[\![\boldsymbol{\tau} \times \boldsymbol{\nu}_e]\!]$ as the corresponding jump of the tangential trace across e. In other words, $[\![\boldsymbol{\tau} \times \boldsymbol{\nu}_e]\!] := (\boldsymbol{\tau}|_K - \boldsymbol{\tau}|_{K'}) \times \boldsymbol{\nu}_e$, where K and K' are the tetrahedra in $\mathcal{T}_h$ sharing e as a common face and

$$\boldsymbol{\tau} \times \boldsymbol{\nu}_e := \begin{pmatrix} (\tau_{11}, \tau_{12}, \tau_{13}) \times \boldsymbol{\nu}_e \\ (\tau_{21}, \tau_{22}, \tau_{23}) \times \boldsymbol{\nu}_e \\ (\tau_{31}, \tau_{32}, \tau_{33}) \times \boldsymbol{\nu}_e \end{pmatrix}.$$

From now on, when no confusion arises, we simply write $\boldsymbol{\nu}$ instead of $\boldsymbol{\nu}_e$, In the sequel we will also make use of the following differential operators

$$\mathbf{curl}\,(\mathbf{v}) = \nabla \times \mathbf{v} := \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right)^{\mathrm{t}},$$

and

$$\underline{\mathbf{curl}}\,(\boldsymbol{\tau}) := \begin{pmatrix} \mathbf{curl}\,(\tau_{11}, \tau_{12}, \tau_{13})^{\mathrm{t}} \\ \mathbf{curl}\,(\tau_{21}, \tau_{22}, \tau_{23})^{\mathrm{t}} \\ \mathbf{curl}\,(\tau_{31}, \tau_{32}, \tau_{33})^{\mathrm{t}} \end{pmatrix}.$$

In turn, we will also use the tensor version of the tangential curl operator $\mathbf{curl}_s$, denoted by $\underline{\mathbf{curl}}_s$, which is defined component-wise by $\mathbf{curl}_s$ (see [38, Section 3] for details).

We now set for each $K \in \mathcal{T}_h$ the local estimator

$$\begin{aligned} \Theta_{2,K}^2 := & \left\| \eta\big(p_h, |\mathbf{D}_h|\big)\mathbf{D}_h - \boldsymbol{\sigma}_h^{\mathtt{d}} - \rho\,(\mathbf{u}_h \otimes \mathbf{u}_h)^{\mathtt{d}} \right\|_{0,K}^2 + \left\| \boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}} \right\|_{0,K}^2 \\ & + h_K^2 \left\| \underline{\mathbf{curl}}\,(\mathbf{D}_h + \boldsymbol{\gamma}_h) \right\|_{0,K}^2 + \sum_{e \in \mathcal{E}_h(K) \cap \mathcal{E}(\Omega)} h_e \left\| [\![(\mathbf{D}_h + \boldsymbol{\gamma}_h) \times \boldsymbol{\nu}]\!] \right\|_{0,e}^2 \\ & + \sum_{e \in \mathcal{E}_h(K) \cap \mathcal{E}(\Gamma)} h_e \left\| \underline{\mathbf{curl}}_{\mathbf{s}}(\mathbf{u}_D) - (\mathbf{D}_h + \boldsymbol{\gamma}_h) \times \boldsymbol{\nu} \right\|_{0,e}^2, \end{aligned} \tag{A.25}$$

and the global *a posteriori* error estimator is defined as

$$\Theta \,=\, \left\{\sum_{K\in\mathcal{T}_h} \Theta_{1,K}^{4/3}\right\}^{3/4} \,+\, \left\{\sum_{K\in\mathcal{T}_h} \Theta_{2,K}^{2}\right\}^{1/2} \,+\, \left\{\sum_{K\in\mathcal{T}_h} \Theta_{3,K}^{4}\right\}^{1/4}\,, \tag{A.26}$$

where $\Theta_{1,K}^{4/3}$ and $\Theta_{3,K}^{4}$ are defined in (4.2) and (4.4), respectively. Accordingly, the corresponding reliability and efficiency estimates, which represent the analogues of Theorems 4.2.1 and 4.2.2, are stated as follows.

Theorem A.4.1. Assume that L_η and the radii δ and $\delta_{\mathtt{d}}$ satisfy (4.36), and that $\mathbf{u}_D$ is a piecewise polynomial. Then, there exist positive onstants $C_{\mathtt{eff}}$ and $C_{\mathtt{rel}}$, independent of h, such that

$$C_{\mathtt{eff}}\,\Theta \,+\, \mathtt{h.o.t} \,\leq\, \|\vec{\mathbf{D}} - \vec{\mathbf{D}}_h\|_{\mathcal{H}} \,+\, \|p - p_h\|_{0,\Omega} \,\leq\, C_{\mathtt{rel}}\,\Theta\,. \tag{A.27}$$

The proof of Theorem A.4.1 follows closely the analysis in Section 4.2, except for a few aspects that will be discussed below. Specifically, we first observe that the general *a posteriori* error estimate given in Lemma 4.2.1, as well as the upper bounds for $\|\mathcal{R}_1\|_{\mathcal{H}_1'}$ and $\|\mathcal{R}_3\|_{\mathcal{Q}'}$ (cf. (4.26), (4.27)), remain valid in 3D. Next, we follow [81, Theorem 3.2] to derive a 3D version of the Helmholtz decomposition for arbitrary polyhedral domains, as provided by Lemma A.2.2, with $p \in [6/5, 2)$ (cf. [39, Lemma 3.4]). The corresponding discrete Helmholtz decomposition and the functional $\mathcal{R}_2$ are then established and rewritten exactly as in (4.30) and (4.31). Furthermore, to derive the new upper bounds for $\|\mathcal{R}_2\|_{\mathcal{H}_2'}$ (cf. Lemma 4.2.3), we require the 3D analogue of the integration by parts formula on the boundary given in (4.36). In fact, using the identities from [51, Chapter I, 2.17, and Theorem 2.11], we deduce that in this case, the following holds

$$\langle \underline{\mathbf{curl}}\,(\boldsymbol{\xi})\boldsymbol{\nu}, \boldsymbol{\theta}\rangle_\Gamma \,=\, -\,\langle \underline{\mathbf{curl}}_{\,\mathbf{s}}(\boldsymbol{\theta}), \boldsymbol{\xi}\rangle_\Gamma\,, \qquad \forall\,\boldsymbol{\xi} \in \mathbb{H}^1(\Omega)\,, \quad \forall\,\boldsymbol{\theta} \in \mathbf{H}^{1/2}(\Gamma)\,. \tag{A.28}$$

In addition, the integration by parts formula on each tetrahedron $K \in \mathcal{T}_h$, which is used in the proof of the 3D analogues of Lemma 4.2.3, becomes (cf. [51, Chapter I, Theorem 2.11])

$$\int_K \underline{\mathbf{curl}}\,(\mathbf{q}) : \boldsymbol{\xi} \,-\, \int_K \mathbf{q} : \underline{\mathbf{curl}}\,(\boldsymbol{\xi}) \,=\, \langle \mathbf{q}\times\boldsymbol{\nu}, \boldsymbol{\xi}\rangle_{\partial K}\,, \quad \forall\,\mathbf{q} \in \mathbb{H}(\underline{\mathbf{curl}}\,;\Omega)\,, \quad \forall\,\boldsymbol{\xi} \in \mathbb{H}^1(\Omega)\,,$$

where $\langle\cdot,\cdot\rangle_{\partial K}$ denotes the duality pairing between $\mathbb{H}^{-1/2}(\partial K)$ and $\mathbb{H}^{1/2}(\partial K)$. As usual, $\mathbb{H}(\mathbf{curl}\,;\Omega)$ is the space of tensor fields in $\mathbb{L}^2(\Omega)$ whose $\underline{\mathbf{curl}}$ belongs to $\mathbb{L}^2(\Omega)$. We observe that, unlike in the 2D case, assuming $\mathbf{u}_D \in \mathbf{H}^1(\Gamma)$ is not necessary for the reliability analysis, since $\underline{\mathbf{curl}}_s$ is defined in $\mathbf{H}^{1/2}(\Gamma)$. Nevertheless, for computational purposes, in Section 4.3, we assume that $\mathbf{u}_D$ is sufficiently smooth, in which case $\underline{\mathbf{curl}}_s(\mathbf{u}_D)$ coincides with $\nabla \mathbf{u}_D \times \boldsymbol{\nu}$.

Finally, to prove the efficiency of Θ, we first observe that the term defining $\Theta_{1,K}^{4/3}$ (cf. (4.2)) and the first two terms defining $\Theta_{2,K}^2$ (cf. (4.3)) are estimated exactly as in the 2D case, following Lemma 4.2.4. For the remaining terms, we establish the following lemma.

Lemma A.4.7. Assume that $\mathbf{u}_D$ is piecewise polynomial. Then, there exist positive constants C_i for $i \in \{1, \ldots, 5\}$, all independent of h, such that

a) $$h_K^4 \left\| \nabla \mathbf{u}_h - \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h \Big) \right\|_{0,4;K}^4$$
$$\leq C_1 \left\{ \|\mathbf{u} - \mathbf{u}_h\|_{0,4;K}^4 + h_K^2 \|\mathbf{D} - \mathbf{D}_h\|_{0,K}^4 + h_K^2 \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,K}^4 \right\} \quad \forall\, K \in \mathcal{T}_h \,,$$

b) $$h_e \, \|\mathbf{u}_D - \mathbf{u}_h\|_{0,4;e}^4$$
$$\leq \; C_2 \left\{ \|\mathbf{u} - \mathbf{u}_h\|_{0,4;K_e}^4 + h_{K_e}^2 \, \|\mathbf{D} - \mathbf{D}_h\|_{0,K_e}^4 + h_{K_e}^2 \, \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,K_e}^4 \right\} \quad \forall\, e \in \mathcal{E}_h(\Gamma) \,,$$

c) $$h_K^2 \left\| \underline{\mathbf{curl}} \, (\mathbf{D}_h + \boldsymbol{\gamma}_h) \right\|_{0,K}^2 \leq C_3 \left\{ \|\mathbf{D} - \mathbf{D}_h\|_{0,K}^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,K}^2 \right\} \quad \forall\, K \, \in \mathcal{T}_h \,,$$

d) $$h_e \left\| [\![(\mathbf{D}_h + \boldsymbol{\gamma}_h) \times \boldsymbol{\nu}]\!] \right\|_{0,e}^2 \leq C_4 \left\{ \|\mathbf{D} - \mathbf{D}_h\|_{0,\omega_e}^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,\omega_e}^2 \right\} \quad \forall\, e \, \in \mathcal{E}_h(\Omega) \,,$$

e) $$h_e \left\| \underline{\mathbf{curl}}_{\mathbf{s}}(\mathbf{u}_D) - \Big(\mathbf{D}_h + \boldsymbol{\gamma}_h \Big) \times \boldsymbol{\nu} \right\|_{0,e}^2 \leq C_5 \left\{ \|\mathbf{D} - \mathbf{D}_h\|_{0,K_e}^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{0,K_e}^2 \right\}, \quad \forall\, e \in \mathcal{E}_h(\Gamma) \,,$$

where K_e is the tetrahedron in $\mathcal{T}_h$ having e as a face, whereas ω_e denotes the union of the two elements in $\mathcal{T}_h$ that share the face e.

Proof. For a), we refer again to [41, Lemma 3.15] by using now the local inverse inequality (A.23) with $n = 3$, whereas b) follows from [41, Lemma 3.16], (A.24) and the estimate in a). In addition, for the proof of c), we refer to [73, Lemma 4.3], while the proof of d) follows from [73, Lemma 4.4]. Finally, e) can be derived after a slight modification of the proof of [74, Lemma 4.15], along with the definition of $\underline{\mathbf{curl}}_{\mathbf{s}}$. □