Universidad de Concepción
Dirección de Postgrado
Facultad de Ciencias Físicas y Matemáticas
Programa de Doctorado en Ciencias Aplicadas
con Mención en Ingeniería Matemática

MÉTODOS DE ELEMENTOS FINITOS MIXTOS PARA PROBLEMAS NO LINEALES EN BIOMEDICINA Y BIOLOGÍA

MIXED FINITE ELEMENT METHODS FOR NONLINEAR PROBLEMS IN BIOMEDICINE AND BIOLOGY

Tesis para optar al grado de Doctor en Ciencias
Aplicadas con mención en Ingeniería Matemática

Willian Armando Miranda Tobar
concepción-chile
2021

Profesor Guía: Gabriel N. Gatica
CI^{2}MA y Departamento de Ingeniería Matemática
Universidad de Concepción, Chile

Cotutor: Eligio Colmenares
Departamento de Ciencias Básicas
Universidad del Bío-Bío, Chillán, Chile

Cotutor: Daniel Hurtado
IIBM y Departamento de Ingeniería Estructural y Geotécnica
Pontificia Universidad Católica de Chile, Chile

Mixed Finite Element Methods for Nonlinear Problems in Biomedicine and Biology

Willian Armando Miranda Tobar

Directores de Tesis: Gabriel N. Gatica, Universidad de Concepción, Chile.
Eligio Colmenares, Universidad del Bío-Bío, Chillán, Chile.
Daniel Hurtado, Pontificia Universidad Católica de Chile, Chile.

Director de Programa: Raimund Bürger, Universidad de Concepción, Chile.

Comisión evaluadora

Prof. Juan Gabriel Calvo, Universidad de Costa Rica, Costa Rica.

Prof. Carlos Sing Long, Pontificia Universidad Católica de Chile, Chile.

Comisión examinadora

Firma: _____
Prof. Sergio Caucao, Universidad Católica de la Santísima Concepción, Chile.

Firma: _____
Prof. Eligio Colmenares, Universidad del Bío-Bío, Chillán, Chile.

Firma: _____
Prof. Gabriel N. Gatica, Universidad de Concepción, Chile.

Firma: _____
Prof. Daniel Hurtado, Pontificia Universidad Católica de Chile, Chile.

Firma: _____
Prof. Ricardo Ruiz Baier, Monash University, Australia.

Calificación: _____

Concepción, Agosto de 2021

Abstract

This thesis aims to develop the mathematical and numerical analysis of nonlinear coupled partial differential equations (PDE's)-based models that describe certain phenomena in Biology and Biomedicine encompassing generalized bioconvection and deformable image registration. More precisely, we introduce primal and mixed schemes based on finite elements for the aforementioned models, prove the solvability of the continuous and discrete problems, establish the corresponding error estimates, and present a variety of tests to validate the theoretical results and illustrate the performance of such methods including applied examples.

We begin with the bioconvective flows model, which describes the hydrodynamics of microorganisms in a culture fluid and takes place in several biological processes, including reproduction, infection, and the marine life ecosystem. The flows are governed by a Navier-Stokes type system coupled to a conservation equation that models the microorganisms concentration. The culture fluid is assumed to be viscous and incompressible with a concentration dependent viscosity. For the mathematical analysis, the model is rewritten in terms of a first-order system based on the introduction of the strain, the vorticity, and the pseudo-stress tensors in the fluid equations along with an auxiliary vector in the concentration equation. The resulting weak model is then augmented using appropriate redundant parameterized terms and rewritten as a fixed-point problem. Existence and uniqueness results for both the continuous and the discrete scheme are obtained under certain regularity assumptions combined with the Lax-Milgram theorem or the Babuška-Brezzi theory, and the Banach and Brouwer fixed-point theorems. Optimal a priori error estimates are also derived and confirmed via numerical examples.

Next, we address the study of a deformable image registration (DIR) model, which arises in numerous research fields as a solution to the combination or comparison of a series of images. Specifically, in Biomedicine, there is a need to detect changes in images obtained from the same subject over time, whereby the deformable image registration represents a powerful computational method for image analysis, with promising applications in the diagnosis of human disease. One important and recent application of DIR is the study of local lung tissue deformation from computed-tomography images of the thorax, which allows the early detection of damage induced by mechanical ventilation in the lung. In our case, for the first model studied in this part, which we will call extended deformable image registration problem, we propose a finite element method for its numerical approximation, proving well-posedness of the primal and dual-mixed continuous formulations, as well as of the associated Galerkin schemes. A priori error estimates and the corresponding rates of convergence are also established for both discrete methods. In addition, we provide numerical examples confronting our formulations with the standard ones.

Finally, in order to guarantee an appropriate convergence behavior of the discrete approximations

obtained by the aforementioned primal and mixed variational formulations of the image registration problem, we develop an a posteriori error analysis for both schemes in terms of residual estimators, which we prove to be reliable and efficient. Based on the latters, we implement adaptive mesh-refinement schemes for the formulations, confirm their properties and illustrate their applicability using medical brain images and binary images.

Resumen

Esta tesis tiene como objetivo desarrollar un análisis matemático y numérico de modelos basados en ecuaciones diferenciales parciales (PDE's) acopladas y no lineales que describen ciertos fenómenos en Biología y Biomedicina que abarcan la bioconvección generalizada y el registro de imágenes deformables. Más precisamente, introducimos esquemas primales y mixtos basados en elementos finitos para los modelos antes mencionados, probamos la solubilidad de los problemas continuos y discretos, establecemos las estimaciones de error correspondientes y presentamos una variedad de experimentos numéricos para validar los resultados teóricos e ilustrar el desempeño de tales métodos incluyendo ejemplos aplicados.

Iniciamos con el modelo de flujos bioconvectivos el cual describe la hidrodinámica de un cultivo de microorganismos y se usa para estudiar y entender diversos procesos biológicos tales como la reproducción, infecciones, y el ecosistema de la vida marina. Desde un punto de vista matemático, el problema está constituido por ecuaciones tipo Navier-Stokes para el movimiento del fluido acoplada a una ecuación de conservación para describir la hidrodinámica y la concentración de microorganismos, respectivamente. El cultivo se asume como un fluido viscoso e incompresible con una viscosidad dependiente de la concentración. Para el análisis matemático de este modelo, se reescribe en términos de un sistema de primer orden basado en la introducción de los tensores de esfuerzo, de vorticidad y de pseudo-estrés en las ecuaciones de fluidos junto con un vector auxiliar en la ecuación de concentración. La formulación débil resultante se aumenta utilizando términos parametrizados redundantes apropiados y lo reescribimos como un problema de punto fijo. La existencia y unicidad, tanto para el esquema continuo como para el discreto se obtienen bajo ciertos supuestos de regularidad combinados con el teorema de Lax-Milgram o la teoría de Babuška-Brezzi y los teoremas de punto fijo de Banach y Brouwer. También derivamos estimaciones de error a priori óptimas y que se ilustran a través de experimentos numéricos.

Luego, estudiamos un modelo de registro deformable de imágenes (DIR, por sus siglas en inglés), el cual surge en un gran número de campos de investigación como solución a la combinación o comparación de una serie de imágenes. Específicamente, en biomedicina, existe la necesidad de detectar cambios en imágenes obtenidas a partir de un mismo sujeto a través del tiempo, por lo cual, el registro deformable de ellas representa un poderoso método computacional para analizar imágenes biomédicas, con prometedoras aplicaciones en el diagnóstico en enfermedades humanas. Una aplicación importante y reciente de este problema es estudiar la deformación regional del tejido pulmonar a partir de imágenes de tomografía computarizada del tórax, lo cual permite la detección temprana del daño inducido por ventilación mecánica en el pulmón. En nuestro caso, para el primer modelo estudiado en esta parte, el cual llamaremos problema de registro deformable de imágenes extendido, proponemos un método de

elementos finitos para su aproximación numérica, probando que las formulaciones continuas primal y dual-mixta, así como de los esquemas de Galerkin asociados están bien puestos. También se establecen estimaciones de error a priori y las correspondientes tasas de convergencia para ambos métodos discretos. Adicionalmente, proporcionamos ejemplos numéricos que comparan nuestras formulaciones con la estándar.

Finalmente, con el fin de garantizar un comportamiento de convergencia adecuado de las aproximaciones discretas que se obtienen a través de las formulaciones variacionales primales y mixtas antes mencionadas para el problema de registro de imágenes, desarrollamos un análisis de error a posteriori para ambos esquemas en términos de estimadores residuales, que demostramos ser confiables y eficientes. Basados en estos últimos, implementamos esquemas adaptativos de refinamiento de malla para las formulaciones, confirmamos sus propiedades e ilustramos la aplicabilidad de éstos utilizando imágenes médicas cerebrales e imágenes sintéticas.

Agradecimientos

En primer lugar, todo mi agradecimiento al Prof. Gabriel N. Gatica, por la excelente calidad como docente, por su guía y apoyo como director de tesis, por sus consejos, paciencia, y entusiasmo durante todo mi periodo en el Doctorado y desarrollo de esta Tesis. Lo mucho que he aprendido durante este periodo es todo gracias a él.

Agradezco de igual manera a mis co-directores: Dr. Eligio Colmenares por toda su dedicación a este trabajo, por su constante apoyo y paciencia; agradezco también al Dr. Daniel Hurtado, por su paciencia y su apoyo incondicional, por su invitación para trabajar con él en el Instituto de Ingeniería Biológica y Médica UC, la amabilidad con la que me recibió y el entusiasmo con el que guió mi trabajo durante mi estadía.

Así mismo, mi agradecimiento a los investigadores que han colaborado para el desarrollo de esta tesis: Dr. Nicolas Barnafi (Politecnico di Milano, Italy) y Dr. Ricardo Ruiz-Baier (Monash University, Australia). Al Prof. Todd Arbogast (University of Texas at Austin, USA), por su apoyo y la hospitalidad mostrada durante mi pasantía.

Adicional, quiero agradecer de mis profesores en el programa: Rodolfo Rodríguez, Raimund Bürger, Rodolfo Araya, Leonardo Figueroa, y Mauricio Sepúlveda, por todas sus enseñanzas.

De igual forma, gracias por toda su ayuda y orientación, al personal administrativo del CI^{2}MA, del Departamento de Ingeniería Matemática, y de la Dirección de Postgrado de la Universidad de Concepción: Lorena Carrasco, Cecilia Leiva y Constaza Greig.

A todos mis compañeros del doctorado, de manera especial agadezco a mis compañeros de generación: Bryan, Paul, Rafael y Cristian, por todos los buenos momentos compartidos y el agrado de aprender con ellos.

Agradezco a las instituciones y proyectos que han financiado mis estudios e investigación: la Dirección de Postgrado; al Proyecto de la Red Doctoral en Ciencia, Tecnología y Ambiente REDOC.CTA Convenio UCO 1202; al proyecto CMM-BASAL AFB 17001; y a las becas de doctorado nacional de la Comisión Nacional de Ciencia y Tecnología (CONICYT), hoy Agencia Nacional de Investigación y Desarrollo (ANID), por el financiamiento mediante el Programa Formación de Capital Humano Avanzado (PFCHA/DOCTORADO NACIONAL/2019-21191204).

Finalmente, extiendo mi más profundo agradecimiento a mis padres y hermanos, gracias por su amor y constante aliento. También, a los familiares que de una u otra forma me han apoyado en este periodo.

Willian Miranda

Contents

List of Tables

List of Figures

Introduction

Most of the partial differential equations (PDEs) that model many natural phenomena in science and engineering are difficult or even impossible to solve analytically, so numerical methods are required to generate approximate solutions that allow a better understanding and description of such phenomena. Finite element methods are one of such techniques and have shown to be appropriate for a wide range of problems representing, in particular, a very powerful tool to obtain approximate solutions in finite dimensional spaces and to conduct computational simulations. In particular, the mixed finite element method is a technique used for numerically solving mathematical models in the form of systems of PDE's that involve several physically disparate quantities, which need to be approximated simultaneously. In some cases, one or several fields are introduced in the formulation of the problem because of its physical interest and they are usually related with some derivatives of the original unknown fields, or a combination of these.

According to the above, this thesis deals with mixed finite element methods for certain phenomena based on nonlinear coupled partial differential equations of special interest in Biology and Biomedicine that encompass generalized bioconvection and deformable image registration (see Section Model Problems below). For each of these models, we are particularly interested in:

- deriving suitable variational formulations based on mixed or primal-mixed approaches,
- establishing the existence and uniqueness of continuous weak solutions,
- proposing Galerkin schemes and analyzing their well-posedness,
- obtaining the corresponding solvability and convergence results,
- developing a posteriori error analysis, in some cases, and
- validating theoretical results and illustrating the performance of the schemes through essays and numerical simulations.

In the following two sections, we first describe the models we focus in this thesis and briefly discuss some of their applications. Then, we present the outline section in which we set the organization of the thesis and explain the mathematical and numerical focusing used for each model.

Model problems

In this thesis we address two problems that are generated in important areas of science and health, such as Biology and Biomedicine. The problem that we study in the biological area is known as the bioconvective fluid model, and the corresponding one to the Biomedicine area is called the deformable image registration (DIR) model, which are described next.

First, we focus on the bioconvective fluids problem [65, 71, 75, 78], represented by the following system of partial differential equations, describing the three-dimensional hydrodynamics of negatively geotactic micro-organisms in suspension in a viscous and incompressible culture fluid Ω,

$$\begin{gathered} -\mathbf{div}\,(\mu(\varphi)\mathbf{e}(\boldsymbol{u}))+(\nabla\boldsymbol{u})\,\boldsymbol{u}+\nabla\,p=\boldsymbol{f}-g\,(1+\gamma\varphi)\,\mathbf{i}_3\,, \quad \text{and} \quad \operatorname{div}\boldsymbol{u}\,=\,0 \qquad \text{in} \quad \Omega, \\ -\kappa\Delta\varphi+\boldsymbol{u}\cdot\nabla\varphi+U\frac{\partial\varphi}{\partial x_3}=0 \quad \text{in} \quad \Omega, \\ \boldsymbol{u}=\mathbf{0}\,, \quad \text{and} \quad \kappa\frac{\partial\varphi}{\partial\boldsymbol{\nu}}-\nu_3 U\varphi\,=\,0 \quad \text{on} \quad \Gamma, \qquad \text{and} \qquad \frac{1}{|\Omega|}\int_\Omega\varphi=\alpha, \end{gathered} \tag{1}$$

where the unknowns are the velocity $\boldsymbol{u}$, the pressure p, and the micro-organism concentration φ of the culture fluid which might affect the kinematic viscosity μ. Here, $\mathbf{e}(\boldsymbol{u})$ stands for the symmetric part of the velocity gradient, $\boldsymbol{f}$ refers to a volume-distributed external force, g is the gravitational force magnitude, $\mathbf{i}_3=(0,0,1)^{\mathrm{t}}$ is the vertical unitary vector, κ and U are constants associated to the diffusion rate and the mean velocity of upward swimming of the microorganisms, respectively, and $\gamma>0$ is a given constant depending on the micro-organisms density and the culture fluid density. The model (1) will be studied in Chapter 1. We also remark that the bioconvection phenomenon takes place in several biological processes, including reproduction, infection, and the marine life ecosystem [17, 67, 68, 78]. Some direct applications are related to bacterial research, microbiological cultures, separation of subpopulations of geotactic micro-organisms in lab experiments, and population control of plankton communities in the oceans, to name a few.

The deformable image registration (DIR) model concerns the problem of aligning a given set of images by means of a transformation that warps one or more of these images. Its formulation requires three main ingredients: (i) the transformation model, composed by a family of mappings that warp the target image into the reference image; (ii) the similarity measure, a function that measures the differences between the images; and (iii) the regularizer, which renders the problem well-posed. Specifically, consider a domain $\Omega\subset\mathrm{R}^{d=2,3}$, $R:\Omega\to\mathrm{R}$ the reference image and $T:\Omega\to\mathrm{R}$ the target image, where $R(\boldsymbol{x})$ and $T(\boldsymbol{x})$ denote the image intensity at point $\boldsymbol{x}$. Then, the objective of DIR is to find a transformation $\boldsymbol{u}:\Omega\to\mathrm{R}^d$, also known as the displacement field, that best aligns the images R and T, namely

$$T(\boldsymbol{x}+\boldsymbol{u}(\boldsymbol{x}))=R(\boldsymbol{x}) \quad \forall\,\boldsymbol{x}\in\Omega\,.$$

This problem is ill-posed in general, so one formulates it as a minimization problem by considering a family of deformations $\mathcal{V}$ (such that $\boldsymbol{u}\in\mathcal{V}$), a similarity measure $\mathcal{D}:\mathcal{V}\to\mathrm{R}$ (a functional which attains its minimum when the equality above holds), a regularizer $\mathcal{S}:\mathcal{V}\to\mathrm{R}$ (which provides smoothness to the problem), and a positive constant α (which balances $\mathcal{D}$ and $\mathcal{S}$). Putting everything together, the following minimization problem arises:

$$\min_{\boldsymbol{u}\in\mathcal{V}}\Big\{\alpha\mathcal{D}(\boldsymbol{u};R,T)+\mathcal{S}(\boldsymbol{u})\Big\}. \tag{2}$$

We call to the equation (2) the standard DIR. A extended version for DIR is formulated as follows: Let Q be the kernel of the adjoint operator induced by $\mathcal{S}$, which we assume to be non trivial and finite dimensional, splitting $\mathcal{V}=Q^{\perp}\oplus Q$, from which we recall the orthogonality condition, that is, if $\boldsymbol{u}\in Q^{\perp}$ then $\langle\boldsymbol{u},\boldsymbol{\rho}\rangle=0\ \forall\,\boldsymbol{\rho}\in Q$, also given a positive constant β, hinted to control $\boldsymbol{u}$ in Q. Then the

extended DIR version is formulated as the following minimization problem:

$$\min_{(\boldsymbol{u},\boldsymbol{\lambda})\in\mathcal{V}\times Q}\max_{\boldsymbol{\rho}\in Q}\left\{\alpha\mathcal{D}(\boldsymbol{u};R,T)+\mathcal{S}(\boldsymbol{u})+\langle\boldsymbol{u}-\boldsymbol{\lambda},\boldsymbol{\rho}\rangle+\frac{\beta}{2}\|\boldsymbol{\lambda}\|_{\mathcal{V}}^2\right\}. \tag{3}$$

Additional details for obtain (3) will be mentioned in Chapter 2. A common choice for the similarity measure is the sum of squares difference, i.e, the L^2 error that takes the form

$$\mathcal{D}(\boldsymbol{u};R,T):=\frac{1}{2}\int_\Omega (T(\boldsymbol{x}+\boldsymbol{u}(\boldsymbol{x}))-R(\boldsymbol{x}))^2,$$

where R and T are reference and target images, respectively. In this thesis we study the case of elastic DIR, in which $\mathcal{V}=\mathbf{H}^1(\Omega)$ and the regularizing term is taken to be the elastic deformation energy, defined by

$$S(\boldsymbol{u}):=\frac{1}{2}\int_\Omega \mathcal{C}\mathbf{e}(\boldsymbol{u}):\mathbf{e}(\boldsymbol{u}),$$

where

$$\mathbf{e}(\boldsymbol{u})=\frac{1}{2}\{\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{t}}\}\qquad\text{and}\qquad\mathcal{C}\boldsymbol{\tau}=\lambda\mathrm{tr}(\boldsymbol{\tau})\mathbb{I}+d\mu\boldsymbol{\tau}\quad\forall\boldsymbol{\tau}\in\mathbb{L}^2(\Omega),$$

are respectively, the infinitesimal strain tensor (symmetric component of the displacement field gradient) and the elasticity tensor for isotropic solids with the Lamé constants $\lambda,\mu>0$ characterizing the material. In this case, the associated Euler-Lagrange equations from (2) deliver the following strong problem: Find $\boldsymbol{u}$ such that

$$\begin{aligned}\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}))&=\alpha\nabla\mathcal{D}(\boldsymbol{u})&&\text{in}\quad\Omega,\\ \mathcal{C}\mathbf{e}(\boldsymbol{u})\boldsymbol{\nu}&=\mathbf{0}&&\text{on}\quad\partial\Omega.\end{aligned} \tag{4}$$

We observe that this problem presents a structure similar to that of a linear elasticity problem with a nonlinear load source. In turn, the associated Euler-Lagrange equations from (3) allow us to formulate the following problem with unknowns $\boldsymbol{u}$, and the rigid motions $\boldsymbol{\rho}$ y $\boldsymbol{\lambda}$

$$\begin{gathered}-\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}))+\boldsymbol{\rho}=-\alpha\nabla\mathcal{D}(\boldsymbol{u}),\qquad\boldsymbol{\lambda}=\Pi_Q\,\boldsymbol{u}\,,\qquad\boldsymbol{\rho}=\beta\,\boldsymbol{\lambda}\quad\text{in}\quad\Omega,\\ \mathcal{C}\mathbf{e}(\boldsymbol{u})\boldsymbol{\nu}=\mathbf{0}\quad\text{on}\quad\partial\Omega,\end{gathered} \tag{5}$$

where $\Pi_Q:\mathcal{V}\to Q$ is the orthogonal projection on Q. The extended DIR problem (5) will be analyzed in Chapter 2, whilst a posteriori error analysis for the standard DIR problem (4) will be developed in Chapter 3. It is important to remark that this model arises in a number of important applications, particularly in the field of medical imaging [85], for example in the study of lung regional deformation computed from tomography images of the thorax [29, 64], and problems related with the image registration of the human brain.

Outline of the thesis

This thesis is organized as follows. In **Chapter** 1, we extend the results obtained in [24] to analyze the solvability of the coupled system (6). We write the model as a first-order system of equations in which the resulting unknowns become the velocity and concentration along with the strain tensor, the vorticity tensor, a pseudo-stress tensor and a vectorial unknown depending on the fluid velocity,

the microorganism concentration and its gradient (introduced as auxiliary unknowns). After the variational formulation is derived, the problem is then augmented by using redundant parameterized Galerkin terms, which allows to set the problem in standard Hilbert spaces and, in turn, to circumvent any inf-sup compatibility condition between the involved spaces. Then the analysis is carried out using a fixed-point approach [36], combining the Lax-Milgram theorem with the classical Banach and Brouwer fixed-point theorems for stating the respective solvability of the continuous problem and the associated Galerkin scheme, under suitable regularity assumptions, a feasible choice of parameters and, in the discrete case, for any family of finite element subspaces. A Strang-type lemma, valid for linear problems, enables us to derive the corresponding Céa estimate and to provide optimal a priori error bounds for the Galerkin solution. The contents of this chapter gave rise to the following paper:

[34] E. Colmenares, G. N. Gatica and W. Miranda, *Analysis of an augmented fully-mixed finite element method for a bioconvective flows model*. Journal of Computational and Applied Mathematics, vol. 393, Art. Num. 113504, (2021).

In **Chapter** 2, we generalize the analysis presented in [13] to regularizers that may present a kernel, and to Lipschitz similarity measures. This is performed by splitting weakly the warping with respect to the kernel of the regularizer so that such kernel remains present in the formulation throughout the model, under the assumption of a relationship between the regularizer and the similarity measure. Then, we derive the new model and analyze its primal formulation at both continuous and discrete levels. The main results, which are obtained by using the Babuška-Brezzi theory and duality arguments, include well-posedness of the continuous and discrete formulations, a priori error estimates, and the respective rates of convergence. In addition, we introduce and analyze (using basically the same theoretical tools from the primal case) an extended dual-mixed formulation in the particular case of an elastic energy. The contents of this chapter gave rise to the following paper:

[15] N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda and R. Ruiz-Baier, *New primal and dual-mixed finite element methods for stable image registration with singular regularization*. Mathematical Models and Methods Applied Sciences, to appear (2021).

In **Chapter** 3, we develop an a posteriori error analysis for the variational formulations described in [13]. More precisely, we develop an reliable and efficient residual-based a posteriori error estimators, which allows us to establish appropriate adaptive methods to guarantee greater precision of the numerical approximations, and mainly the convergence of the Galerkin scheme in situations in which there are singularities or high gradients of the solution. Our theoretical results, make use of the standard tools, which include global inf-sup conditions, Helmholtz decompositions, and the approximation properties of the Raviart-Thomas and Clément interpolants for proving reliability of the estimator. In turn, localization techniques using bubble functions and inverse inequalities are employed to prove the corresponding efficiency estimates. This chapter is constituted by the following preprint:

[14] N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda and R. Ruiz-Baier, *A posteriori error estimates for primal and mixed finite element approximations of the deformable image registration problem*. Preprint 2018-50, Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción, Chile, (2018).

Throughout the chapters of this thesis, the theoretical results are illustrated through several numerical examples, that corroborate the accuracy of the numerical schemes. In addition, the computational implementations were obtained using the free access software for finite elements: FreeFem++ [59], FEniCS [4] and the illustrator ParaView.

Preliminary notations

Let $\Omega \subseteq \mathbb{R}^{d=2,3}$ a bounded domain with boundary $\Gamma := \partial\Omega$, and outward unit normal given by $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_d)^{\mathrm{t}}$. Standard notation will be adopted for Lebesgue spaces $\mathrm{L}^p(\Omega)$ and Sobolev spaces $\mathrm{H}^s(\Omega)$ with norm $\|\cdot\|_{s,\Omega}$, and semi-norm $|\cdot|_{s,\Omega}$. Given a generic scalar functional space A, we let $\mathbf{A}$ and $\mathbb{A}$ be its vectorial and tensor versions, respectively, and we denote by $\|\cdot\|$, with no subscripts, the natural norm of either an element or an operator in any product functional space. As usual, for any vector field $\boldsymbol{v} = (v_i)_{i=1,d}$, we set the gradient, divergence and, tensor product operators, as

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

Furthermore, given tensor fields $\boldsymbol{\tau} = (\tau_{ij})_{i,j=1,d}$ and $\boldsymbol{\zeta} = (\zeta_{ij})_{i,j=1,d}$, 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, and the deviatoric tensor, respectively, as

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

Finally, we recall the following Hilbert space

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

eqquiped with the usual norm

$$\|\boldsymbol{\tau}\|^2_{\mathbf{div},\Omega} := \|\boldsymbol{\tau}\|^2_{0,\Omega} + \|\mathbf{div}\, \boldsymbol{\tau}\|^2_{0,\Omega}.$$

Introducción

La mayoría de las ecuaciones diferenciales parciales (EDP's) que modelan una diversidad de fenómenos naturales en ciencia e ingeniería son difíciles o incluso imposibles de resolver analíticamente, por lo que se requieren métodos numéricos para generar soluciones aproximadas que permitan una mejor comprensión y descripción de dichos fenómenos. Los métodos de elementos finitos son una de esas técnicas y han demostrado ser apropiados para una amplia gama de problemas, representando en particular, una herramienta muy poderosa para obtener soluciones aproximadas en espacios de dimensión finita y para realizar simulaciones computacionales. En particular, el método de elementos finitos mixtos es una técnica utilizada para resolver numéricamente modelos matemáticos presentados en forma de sistemas de EDP's que involucran varias cantidades físicamente dispares, que requieran aproximarse simultáneamente. En algunos casos, una o más variables se introducen en la formulación del problema por su interés físico y suelen estar relacionadas con algunas derivadas de las incógnitas originales, o una combinación de estas.

De acuerdo con lo anterior, esta tesis trata sobre métodos mixtos de elementos finitos para ciertos fenómenos representados por medio de ecuaciones diferenciales parciales acopladas no lineales, que son de especial interés en Biología y Biomedicina y que engloban bioconvección generalizada y el registro de imágenes deformables (ver apartado Problemas modelo más adelante). Para cada uno de estos modelos, estamos particularmente interesados en:

- derivar formulaciones variacionales adecuadas basadas en enfoques mixtos o primarios-mixtos,
- establecer la existencia y la unicidad de soluciones débiles a nivel continuo,
- proponer esquemas de Galerkin y analizar su buena planteamiento,
- obtener correspondientes resultados de solubilidad y convergencia,
- desarrollar análisis de errores a posteriori, en algunos casos, y
- validar resultados teóricos e ilustrar el desempeño de los esquemas a través de ensayos y simulaciones numéricas.

En las dos secciones siguientes, primero describimos los modelos que en los que nos enfocamos en esta tesis y discutimos brevemente algunas de sus aplicaciones. A continuación, presentamos la organización de la tesis y explicamos el enfoque matemático y numérico utilizado para cada modelo.

Problemas modelo

En esta tesis abordamos dos problemas que se generan en importantes áreas de la ciencia y la salud, como lo son la Biología y la Biomedicina. El problema que estudiamos en el área biológica se

conoce como modelo de flujos bioconvectivos, y el correspondiente al área de Biomedicina es llamado problema de registro deformable de imágenes (DIR, por sus siglas en inglés), los cuales se describen a continuación.

Primero, estudiamos en el problema de fluidos bioconvectivos [65, 71, 75, 78], el cual se representa mediante el siguiente sistema de ecuaciones diferenciales parciales, y que describe la hidrodinámica de un grupo de microorganismos que tienen un comportamiento geotáctico negativo, es decir que se mueven contra la gravedad (tienden a nadar hacia arriba), suspendidos en un cultivo, fluido viscoso e imcopresible Ω,

$$\begin{aligned} -\mathbf{div}\,(\mu(\varphi)\mathbf{e}(\boldsymbol{u})) + (\nabla \boldsymbol{u})\,\boldsymbol{u} + \nabla p = \boldsymbol{f} - g\,(1+\gamma\varphi)\,\mathbf{i}_3\,, \quad \text{and} \quad \operatorname{div}\boldsymbol{u} = 0 \qquad \text{in} \quad \Omega, \\ -\kappa\Delta\varphi + \boldsymbol{u}\cdot\nabla\varphi + U\frac{\partial\varphi}{\partial x_3} = 0 \quad \text{in} \quad \Omega, \\ \boldsymbol{u} = \mathbf{0}\,, \quad \text{and} \quad \kappa\frac{\partial\varphi}{\partial\boldsymbol{\nu}} - \nu_3 U\varphi = 0 \quad \text{on} \quad \Gamma, \qquad \text{and} \qquad \frac{1}{|\Omega|}\int_\Omega \varphi = \alpha, \end{aligned} \tag{6}$$

donde la incógnitas son la velocidad $\boldsymbol{u}$, la presión p, y la concentración de microoganismos φ, la cual puede afectar a la viscosidad cinemática μ. Aquí, $\mathbf{e}(\boldsymbol{u})$ es el tensor de pequeñas deformaciones, $\boldsymbol{f}$ es una fuerza externa distribuida en el volumen, g es la magnitud de la fuerza de gravedad, $\mathbf{i}_3 = (0,0,1)^{\mathrm{t}}$ es el vector unitario vertical, y las constantes positivas: κ y U asociadas a la tasa de difusión y la velocidad promedio de natación ascendente de los microorganismos, respectivamente; γ que depende de la densidad de los microorganismos y el cultivo; α que asegura que ningún microorganismo pueda salir o entrar en el dominio físico. El modelo (6) será estudiado en el Capítulo 1. Señalamos además que el fenómeno de bioconvección aparece en varios procesos biológicos, incluida la reproducción, la infección, y el ecosistema de vida marina [17, 67, 68, 78]. Algunas aplicaciones directas están relacionadas con la investigación bacteriana, los cultivos microbiológicos, la separación de subpoblaciones de microorganismos geotácticos en experimentos de laboratorio y el control de la población de comunidades de plancton en los océanos, por nombrar algunas.

El modelo de registro deformable de imágenes (DIR) consiste en el problema de alinear un conjunto dado de imágenes mediante una transformación que deforma una o más de estas imágenes. Su formulación requiere de tres ingredientes principales: (i) el modelo de transformación, compuesto por una familia de mapeos que deforman la imagen objetivo en la imagen de referencia; (ii) la medida de similitud, función que mide las diferencias entre las imágenes; y (iii) el regularizador, que hace que el problema este bien planteado. Específicamente, considerando un dominio $\Omega \subset \mathbb{R}^{d=2,3}$, $R : \Omega \to \mathrm{R}$ la imagen de referencia y $T : \Omega \to \mathrm{R}$ la imagen objetivo, donde $R(\boldsymbol{x})$ y $T(\boldsymbol{x})$ denotan la intensidad de la imagen en el punto $\boldsymbol{x}$. Entonces, el objetivo del DIR es encontrar una transformación $\boldsymbol{u} : \Omega \to \mathrm{R}^d$, también conocida como campo de desplazamiento, que mejor se alinea las imágenes R y T, esto es

$$T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) = R(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Omega\,.$$

En general este problema está mal puesto, por lo que se formula como un problema de minimización considerando una familia de deformaciones $\mathcal{V}$ (tal que $\boldsymbol{u} \in \mathcal{V}$), una medida de similitud $\mathcal{D} : \mathcal{V} \to \mathrm{R}$ (un funcional que alcanza su mínimo cuando se cumple la igualdad anterior, un regularizador $\mathcal{S} : \mathcal{V} \to \mathrm{R}$ (el cual aporta suavidad al problema), y una constante positiva α (la cual equilibra $\mathcal{D}$ y $\mathcal{S}$). Escribiendo matemáticamente eso, se genera el siguiente problema de minimización:

$$\min_{\boldsymbol{u}\in\mathcal{V}} \Big\{ \alpha\mathcal{D}(\boldsymbol{u}; R, T) + \mathcal{S}(\boldsymbol{u}) \Big\}. \tag{7}$$

Llamaremos a la ecuación (7) la versión estándar del problema DIR. Una versión extendida de este problema se formula como sigue: Sea Q el kernel del operador adjunto inducido por $\mathcal{S}$, el cual se asume no trivial y de dimensión finita, se tiene la descomposición $\mathcal{V} = Q^{\perp} \oplus Q$, de la cual recordamos la condición de ortogonalidad es decir, si $\boldsymbol{u} \in Q^{\perp}$ entonces $\langle \boldsymbol{u}, \boldsymbol{\rho} \rangle = 0 \ \ \forall\, \boldsymbol{\rho} \in Q$, además dada una constante β, sugerida para controlar $\boldsymbol{u}$ en Q. Entonces la versión extendida DIR se formula como el siguiente problema de minimización:

$$\min_{(\boldsymbol{u},\boldsymbol{\lambda})\in\mathcal{V}\times Q} \max_{\boldsymbol{\rho}\in Q} \left\{ \alpha \mathcal{D}(\boldsymbol{u}; R, T) + \mathcal{S}(\boldsymbol{u}) + \langle \boldsymbol{u} - \boldsymbol{\lambda}, \boldsymbol{\rho} \rangle + \frac{\beta}{2} \|\boldsymbol{\lambda}\|_{\mathcal{V}}^2 \right\}. \tag{8}$$

Detalles adicionales para obtener (8) se mencionan en el Capítulo 2. Usualmente, la medida de similitud se elige como la suma de diferencia de cuadrados, dada por

$$\mathcal{D}(\boldsymbol{u}; R, T) := \frac{1}{2} \int_{\Omega} (T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x}))^2,$$

donde R y T son las imágenes de referencia y objetivo, respectivamente. En esta tesis nos ocupamos del caso elástico del DIR, en el cual $\mathcal{V} = \mathbf{H}^1(\Omega)$ y el término regularizador se elige como la energía de deformación elástica, definida por

$$S(\boldsymbol{u}) := \frac{1}{2} \int_{\Omega} \mathcal{C}\mathbf{e}(\boldsymbol{u}) : \mathbf{e}(\boldsymbol{u}),$$

donde

$$\mathbf{e}(\boldsymbol{u}) = \frac{1}{2}\{\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^{\mathrm{t}}\} \qquad \text{y} \qquad \mathcal{C}\boldsymbol{\tau} = \lambda \mathrm{tr}(\boldsymbol{\tau})\mathbb{I} + d\mu\boldsymbol{\tau} \quad \forall \boldsymbol{\tau} \in \mathbb{L}^2(\Omega),$$

son respectivamente, el tensor de deformación infinitesimal (componente simétrico del gradiente del campo de desplazamiento) y el tensor de elasticidad para sólidos isotrópicos con las constantes de Lamé $\lambda, \mu > 0$. En este caso, las ecuaciones de Euler-Lagrange asociadas a (7), nos proporcionan el siguiente problema: Encontrar $\boldsymbol{u}$ tal que

$$\begin{aligned} \mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u})) &= \alpha \nabla \mathcal{D}(\boldsymbol{u}) \qquad \text{in} \quad \Omega, \\ \mathcal{C}\mathbf{e}(\boldsymbol{u})\boldsymbol{\nu} &= \mathbf{0} \qquad\qquad\ \ \text{on} \quad \partial\Omega. \end{aligned} \tag{9}$$

Hacemos notar que este problema presenta una estructura similar al problema de elasticidad lineal con termino fuente no lineal. A su vez, las ecuaciones de Euler-Lagrange asociadas al problema extendido (8), nos permiten formular el siguiente problema con incógnitas $\boldsymbol{u}$, y los movimientos rígidos $\boldsymbol{\rho}$ y $\boldsymbol{\lambda}$

$$\begin{gathered} -\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u})) + \boldsymbol{\rho} = -\alpha \nabla \mathcal{D}(\boldsymbol{u}), \qquad \boldsymbol{\lambda} = \Pi_Q \, \boldsymbol{u}\,, \qquad \boldsymbol{\rho} = \beta\, \boldsymbol{\lambda} \quad \text{in} \quad \Omega, \\ \mathcal{C}\mathbf{e}(\boldsymbol{u})\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \partial\Omega, \end{gathered} \tag{10}$$

donde $\Pi_Q : \mathcal{V} \to Q$ es la proyección ortogonal en Q. La versión extendida (10) del problema DIR será analizada en el Capítulo 2, mientras que el análisis de error a posteriori para la versión estándar (9) será desarrollado en el Capítulo 3. Es importante señalar que este modelo tiene una serie de aplicaciones importantes, particularmente en el campo de las imágenes médicas [85], como por ejemplo en el estudio de la deformación regional del pulmón a partir de imágenes de tomografía del tórax [29, 64], y a problemas relacionados con el registro de imágenes del cerebro humano.

Organización de la tesis

Esta tesis está organizada como sigue. En el **Capítulo** 1, extendemos los resultados obtenidos en [24] para analizar la solubilidad del sistema acoplado (6). Primero, escribimos el modelo como un sistema de ecuaciones de primer orden en el cual las incógnitas resultantes son la velocidad y la concentración junto con el tensor de esfuerzo; la vorticidad; el tensor de pseudo esfuerzo y un vector desconocido que depende de la velocidad de fluido, la concentración de microorganismos y su gradiente. Después se obtiene la formulación variacional, el problema entonces es aumentado usando términos redundantes de Galerkin, lo cual nos permite establecer el problema en espacios estándar de Hilbert y, a su vez, evitar cualquier condición de compatibilidad inf-sup entre los espacios involucrados. Luego, el análisis es llevado a cabo usando una estrategia de punto-fijo [36], combinando el teorema de Lax-Milgram con los teoremas clásicos de Banach y punto-fijo de Brouwer para obtener la respectiva solubilidad del problema continuo y el esquema de Galerkin asociado, bajo supuestos adecuados de regularidad, una elección factible de parámetros y, en el caso discreto, para cualquier familia de subespacios de elementos finitos. Utilizando un lema de tipo Strang, válido para problemas lineales, derivamos la correspondiente estimación de Céa y provee cotas óptimas de error a priori para la solución de Galerkin. Los contenidos de este capítulo dieron lugar al siguiente artículo:

[34] E. Colmenares, G. N. Gatica and W. Miranda, *Analysis of an augmented fully-mixed finite element method for a bioconvective flows model*. Journal of Computational and Applied Mathematics, vol. 393, Art. Num. 113504, (2021).

En el **Capítulo** 2, generalizamos el análisis presentado en [13] para regularizadores que podrían presentar un kernel no trivial, y para medidas de similitud Lipschitz, lo cual se realiza separando débilmente la deformación con respecto al kernel del regularizador para que dicho kernel permanezca presente en la formulación a lo largo del modelo, bajo el supuesto de una relación entre el regularizador y la medida de similitud. Luego, derivamos el nuevo modelo y analizamos su formulación primal tanto en el caso continuo como discreto. Los principales resultados en este capítulo, los cuales se obtienen usando la teoría de Babuška-Brezzi y argumentos de dualidad, incluyen solubilidad de las formulaciones continua y discreta, estimaciones de error a priori y la respectiva tasa de convergencia. Adicionalmente, introducimos y analizamos (usando básicamente las mismas herramientas que en el caso primal) una formulación dual-mixta para el caso particular de energía elástica. Los contenidos de este capítulo dieron lugar al siguiente artículo:

[15] N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda and R. Ruiz-Baier, *New primal and dual-mixed finite element methods for stable image registration with singular regularization*. Mathematical Models and Methods Applied Sciences, to appear (2021).

En el **Capítulo** 3, desarrollamos un análisis de error a posteriori para las formulaciones variacionales descrita en [13]. Más precisamente, desarrollamos estimadores de error a posteriori confiables y eficientes basados en residuos, los cuales permiten establecer métodos adaptativos apropiados para garantizar mayor precisión de las aproximaciones numéricas, y principalmente la convergencia del esquema de Galerkin en situaciones en las que hay presencia de singularidades o bien altos gradientes de la solución. Para los resultados teóricos hacemos uso de herraminetas estándar, las cuales incluyen la condición inf-suf global, descomposiciones de Helmholtz, propiedades de aproximación de los

interpolantes de Raviart-Thomas y Clément para probar la confiabilidad. Por otro lado, técnicas de localización basadas en funciones burbuja y desigualdades inversa se utilizan para demostrar la correspondiente estimación de eficiencia. Este capítulo está constituido por la siguiente pre-publicación:

[14] N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda and R. Ruiz-Baier, *A posteriori error estimates for primal and mixed finite element approximations of the deformable image registration problem*. Preprint 2018-50, Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción, Chile, (2018).

A lo largo de los capítulos que conforman esta tesis, los resultados teóricos, son ilustrados a través de varios ejemplos numéricos, que corroboran la precisión de los esquemas numéricos. Además, las implementaciones computacionales de los métodos, se obtuvieron empleando las librerías de elementos finitos de acceso libre: FreeFem++ [59], FEniCS [4] y el ilustrador ParaView.

CHAPTER 1

Analysis of an augmented fully-mixed finite element method for a bioconvective flows model

1.1 Introduction

Bioconvective flows, or bioconvection, refers to a spontaneous flow and pattern formation due to the motion of a large number of upswimming micro-organisms as an innate behavioral response to a stimulus like gravity, light, oxygen, food, changes on temperature, or some combination of these. In a fluid of finite depth, upswimming means that cells accumulate near the top surface due to the gathering of micro-organisms, so the upper regions of the suspensions become denser than the lower, and when the density gradient is high enough, micro-organisms fall down; leading to an overturning convection [78].

By its nature, this phenomenon takes place in several biological processes, including reproduction, infection and the marine life ecosystem [68]. Some direct applications are related to bacterial research, microbiological cultures, separating swimming subpopulations of geotactic micro-organisms (whose movement is gravity-induced) in lab experiments, and controlling population of plankton communities in the oceans, to name a few. In addition, more recently, bioconvective flows have also been considered useful to medical, bioengineering and pharmaceutical applications [17, 67]. For instance, it can be used to configure new geometries of bioreactors, to improve the biofuel production and to enhance microfluidics mixing, which are often linked to several pharmaceutical and biotechnological experiments such as analyses of DNA or drugs, screening of patients and combinatorial synthesis.

A fluid dynamical model to describe bioconvection of geotactic microorganisms was introduced in [71] and [75], independently, from a biological and physical point of view. Using the Boussinesq approximation, the resulting model consists of a Navier-Stokes type system for describing the hydrodynamic of the culture fluid assumed to be viscous and incompressible, in terms of the velocity and the pressure, nonlinearly coupled to an advection-diffusion equation for the micro-organisms concentration, which comes from a cell conservation equation.

The mathematical analysis of this model was carried out in [65]. There, the authors prove existence of weak solutions by the Galerkin method, and existence of strong solutions by a semi-group approach along with the method of successive approximations, for both stationary and evolution problems.

Also, a positivity property of the concentration is shown there. Later, generalized models in which the effective viscosity depends on the concentration of the organisms are mathematically analyzed in [19], for initial conditions, and in [32], for periodic conditions and assuming that the viscosity is a concentration-dependent continuously differentiable function. In these works, uniqueness results of solutions are further given. Then in [40], the authors complement the results from [65] by addressing the problem of obtaining convergence rates for the error when using spectral Galerkin approximations of the problem with a constant viscosity.

First numerical simulations of bioconvection are developed in [27, 58] in two dimensions. Whilst in [27] the authors integrate the Navier-Stokes equations, they treat the cells as individuals moving points, instead of using the continuum cell conservation. In [58], the problem is solved integrating the incompressible Navier-Stokes equations and the cell conservation equation in a shallow box as a physical domain. To the best of our knowledge, [24] is one of the first finite element analysis for the bioconvection model. There, the problem is considered with concentration-dependent viscosity and the authors firstly improve the existence result from [32], by allowing the viscosity to be a continuous and bounded function. They then state existence and uniqueness results for the continuous and discrete problems, as well as a the convergence associated to the classical primal method based on finite elements; whose solvability requires an inf-sup compatibility condition. Additionally, although the analysis is carried out in two and three dimensions, they test the performance and accuracy of the numerical technique only in the 2d-case, including an example with data obtained from lab experiments. Here the Taylor-Hood finite element of second order is used for approximating the velocity and pressure, whereas piecewise quadratic polynomials are used for the concentration. Other numerical techniques developed for related models and their respective mathematical analysis are [38, 41, 42, 46, 52–54, 60, 70, 72, 86] and the references there in, which include gyrotactic, geotactic, oxitactic and chemotactic microorganisms modeling.

As a phenomenon from fluid dynamics, in certain applications some additional physically relevant variables such as the gradient of the fluid velocity or the gradient of the micro-organisms concentration might reveal specific mechanisms of the bioconvection, and hence become of primary interest. Whilst these variables could be obtained via numerical integration of the discrete solutions provided by standard methods, this certainly would lead to a loss of accuracy or deteriorate the expected convergence order. In light of this, the purpose of this work is to contribute with the construction, analysis and implementation of a new numerical technique based on mixed finite elements for simulating bioconvective flows of geotactic micro-organisms, allowing

(a) direct computation of physically relevant variables in the phenomena such as the velocity gradient, the vorticity, the shear stress tensor of the fluid and the micro-organisms concentration gradient,

(b) flexibility regarding the use of finite element subspaces, avoiding any inf-sup compatibility restriction,

(c) high-order approximations, and optimal-order a priori error estimates.

To that end, based on previous mixed methods developed for related problems [2, 3, 5, 26, 35, 37], we firstly re-write the original model as a first-order system of equations in which the resulting un-

knowns become the velocity and concentration (as primal variables) along with the strain tensor, the vorticity tensor, a pseudo-stress tensor and a vectorial unknown depending on the fluid velocity, the microorganism concentration and its gradient (introduced as auxiliary unknowns). After a variational formulation, the problem is then augmented by using redundant parameterized Galerkin terms, which allows to set the problem in standard Hilbert spaces and, in turn, to circumvent any inf-sup compatibility condition between the involved spaces. The analysis is then carried out by a fixed-point approach [36], combining the Lax-Milgram theorem with the classical Banach and Brouwer fixed-point theorems for stating the respective solvability of the continuous problem and the associated Galerkin scheme, under suitable regularity assumptions, a feasible choice of parameters and, in the discrete case, for any family of finite element subspaces. A Strang-type lemma, valid for linear problems, enables us to derive the corresponding Céa estimate and to provide optimal a priori error bounds for the Galerkin solution. In turn, the pressure can be recovered by a post-processed of the discrete solutions, preserving the same rate of convergence. Finally, numerical experiments are presented to illustrate the performance of the technique and confirming the expected orders.

We have organized the contents of this chapter as follows. In Section 1.2, we introduce the model problem, and the auxiliary variables in terms of which an equivalent first-order set of equations is obtained. Next, in Section 1.3, we derive the augmented mixed variational formulation and establish its well-posedness. The associated Galerkin scheme is introduced and analyzed in Section 1.4. In Section 1.5, we derive the corresponding Céa estimate and, finally, in Section 1.6 we present a couple of numerical examples illustrating the performance of our augmented fully-mixed finite element method.

1.2 The bioconvective flows model

In this section, we present the model problem, and define the auxiliary unknowns to be introduced into the respective continuous formulation. From [71, 75, 78], we consider the following system of partial differential equations, describing the three-dimensional hydrodynamics of negatively geotactic micro-organisms in suspension in a viscous and incompressible culture fluid Ω, given by

$$\begin{aligned} -\mathbf{div}\,(\mu(\varphi)\mathbf{e}(\boldsymbol{u}))+(\nabla\boldsymbol{u})\,\boldsymbol{u}+\nabla p &= \boldsymbol{f}-g\,(1+\gamma\varphi)\,\mathbf{i}_3\,, \quad \text{and} \quad \operatorname{div}\boldsymbol{u}\,=\,0 \qquad \text{in} \quad \Omega, \\ -\kappa\Delta\varphi+\boldsymbol{u}\cdot\nabla\varphi+U\frac{\partial\varphi}{\partial x_3} &= 0 \quad \text{in} \quad \Omega, \end{aligned} \tag{1.1}$$

that is, a set of coupled non-linear equations given by a Navier-Stokes type-system and an advection–diffusion equation, in the Boussinesq approximation framework, where the unknowns are the velocity $\boldsymbol{u}=(u_j)_{j=1,3}$, the pressure p and the micro-organism concentration φ of the culture fluid, and in the realistic case in which the micro-organisms concentration might affect the kinematic viscosity $\mu(\,\cdot\,)$.

Here, $\mathbf{e}(\boldsymbol{u})$ stands for the symmetric part ot the velocity gradient, defined as $\mathbf{e}(\boldsymbol{u})=\frac{1}{2}(\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{t}})$, $\boldsymbol{f}$ refers to a volume-distributed external force, g is the gravitational force magnitude, κ and U are constants associated to the diffusion rate and the mean velocity of upward swimming of the microorganisms, respectively, $\mathbf{i}_3=(0,0,1)^{\mathrm{t}}$ is the vertical unitary vector, and $\gamma:=\rho_0/\rho_m-1>0$, is a given constant depending on the micro-organisms density ρ_0 and the culture fluid density ρ_m. In turn, such as in [24] (cf. [32]), we assume that the viscosity $\mu(\,\cdot\,)$ is a Lipschitz continuous and bounded

from above and below function; that is, for some constants $L_\mu > 0$ and $\mu_1, \mu_2 > 0$, there hold

$$|\mu(s) - \mu(t)| \leq L_\mu |s - t|, \qquad \forall\, s, t \geq 0, \tag{1.2}$$

and

$$\mu_1 \leq \mu(s) \leq \mu_2, \qquad \forall s \geq 0. \tag{1.3}$$

We complete the system (1.1), with a non-slip condition for the velocity and a zero flux Robin-type condition for the micro-organisms on the boundary, that is

$$\boldsymbol{u} = \mathbf{0}, \qquad \text{and} \qquad \kappa \frac{\partial \varphi}{\partial \boldsymbol{\nu}} - \nu_3 U \varphi = 0 \qquad \text{on} \quad \Gamma, \tag{1.4}$$

as well as the total mass restriction

$$\frac{1}{|\Omega|} \int_\Omega \varphi = \alpha, \tag{1.5}$$

where α is a given positive constant, assuring that no micro-organisms are allowed to leave or enter the physical domain. Note that (1.5) is equivalent to

$$\int_\Omega (\varphi - \alpha) = 0,$$

and consequently, when setting the auxiliary concentration $\varphi_\alpha := \varphi - \alpha$, which satisfies $\int_\Omega \varphi_\alpha = 0$, and by introducing it into (1.1) and (1.4), we get

$$-\mathbf{div}\,(\mu(\varphi_\alpha + \alpha)\mathbf{e}(\boldsymbol{u})) + (\nabla \boldsymbol{u})\,\boldsymbol{u} + \nabla p = \boldsymbol{f}_\alpha - g\,(1 + \gamma\varphi_\alpha)\,\mathbf{i}_3 \quad \text{in} \quad \Omega,$$

$$\kappa \frac{\partial \varphi_\alpha}{\partial \boldsymbol{\nu}} - \nu_3 U(\varphi_\alpha + \alpha) = 0 \quad \text{on} \quad \Gamma,$$

where $\boldsymbol{f}_\alpha := \boldsymbol{f} - g\gamma\alpha\,\mathbf{i}_3$. Note that the rest of equations remains unchanged with φ_α in place of φ. Therefore, to simplify the notation and without confusion, we rename from now on $\varphi := \varphi_\alpha$ and $\boldsymbol{f} := \boldsymbol{f}_\alpha$, so that the original problem (1.1), (1.4) and (1.5), takes the form

$$\begin{gathered}
-\mathbf{div}\,(\mu(\varphi + \alpha)\mathbf{e}(\boldsymbol{u})) + (\nabla \boldsymbol{u})\,\boldsymbol{u} + \nabla p = \boldsymbol{f} - g\,(1 + \gamma\varphi)\,\mathbf{i}_3, \quad \text{and} \quad \operatorname{div} \boldsymbol{u} = 0 \qquad \text{in} \quad \Omega, \\
-\kappa \Delta \varphi + \boldsymbol{u} \cdot \nabla \varphi + U \frac{\partial \varphi}{\partial x_3} = 0 \quad \text{in} \quad \Omega, \quad \text{with} \quad \int_\Omega \varphi = 0, \\
\boldsymbol{u} = \mathbf{0} \quad \text{and} \quad \kappa \frac{\partial \varphi}{\partial \boldsymbol{\nu}} - \nu_3 U(\varphi + \alpha) = 0 \quad \text{on} \quad \Gamma.
\end{gathered} \tag{1.6}$$

From the first equation of (1.6), it is clear that uniqueness of an eventual pressure solution of this problem (see [55] or [76]) is ensured in the space

$$\mathrm{L}^2_0(\Omega) := \left\{ q \in \mathrm{L}^2(\Omega) : \quad \int_\Omega q = 0 \right\}.$$

Likewise, from the total mass condition on the auxiliary concentration (second equation of the second row in system (1.6)), we see that an eventual weak solution φ of (1.6) belongs to the space

$$\widetilde{\mathrm{H}}^1(\Omega) := \mathrm{H}^1(\Omega) \cap \mathrm{L}^2_0(\Omega) = \left\{ \psi \in \mathrm{H}^1(\Omega) : \quad \int_\Omega \psi = 0 \right\}, \tag{1.7}$$

which is a closed subspace of $\mathrm{H}^1(\Omega)$, and in which the norm and the seminorm are equivalent (result to be used in Lemma 1.2).

Next, in order to derive our fully-mixed formulation, we firstly need to rewrite (1.6) as a first-order system of equations. To this purpose, inspired by the approach from [26] (see also [2,3]), we introduce as additional unknowns the strain and vorticity tensors

$$\boldsymbol{t} := \mathbf{e}(\boldsymbol{u}) \qquad \text{and} \qquad \boldsymbol{\rho} = \frac{1}{2}\Big\{\nabla \boldsymbol{u} - (\nabla \boldsymbol{u})^{\mathrm{t}}\Big\} =: \nabla \boldsymbol{u} - \boldsymbol{t} \qquad \text{in} \qquad \Omega, \tag{1.8}$$

as well as the pseudo-stress tensor

$$\boldsymbol{\sigma} := \mu(\varphi + \alpha)\boldsymbol{t} - p\mathbb{I} - (\boldsymbol{u} \otimes \boldsymbol{u}) \quad \text{in} \quad \Omega. \tag{1.9}$$

Note that $\mathbf{div}(\boldsymbol{u} \otimes \boldsymbol{u}) = (\nabla \boldsymbol{u})\boldsymbol{u}$ when $\operatorname{div} \boldsymbol{u} = 0$ (incompressibility condition - second equation of first row in (1.6)). Thus, the first equation of (1.6) and the constitutive relation (1.9), gives the equilibrium equation

$$-\mathbf{div}(\boldsymbol{\sigma}) \,=\, \boldsymbol{f} - g\,(1 \,+\, \gamma\,\varphi)\mathbf{i}_3 \quad \text{in} \quad \Omega. \tag{1.10}$$

Again, from the incompressibility condition, we have that $\operatorname{tr}(\nabla \boldsymbol{u}) = 0$ and so $\operatorname{tr}(\boldsymbol{\rho}) = \operatorname{tr}(\boldsymbol{t}) = 0$. In particular, by taking deviatoric part from both sides of (1.9), we find that

$$\boldsymbol{\sigma}^{\mathrm{d}} = \mu(\varphi + \alpha)\boldsymbol{t} - (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathrm{d}} \quad \text{in} \quad \Omega\,, \tag{1.11}$$

and so the pressure can be eliminated from the system but, by taking trace from both sides of (1.9), we readily deduce that it can be recovered in terms of $\boldsymbol{\sigma}$ and $\boldsymbol{u}$ as

$$p = -\frac{1}{3}\operatorname{tr}(\boldsymbol{\sigma} + (\boldsymbol{u} \otimes \boldsymbol{u})) \qquad \text{in} \qquad \Omega\,. \tag{1.12}$$

As for the equation modeling the micro-organisms concentration, similarly to [37], we introduce as the new vectorial unknown that we call "pseudo-concentration" gradient

$$\boldsymbol{p} := \kappa \nabla \varphi - \varphi \boldsymbol{u} - U(\varphi + \alpha)\mathbf{i}_3 \qquad \text{in} \qquad \Omega, \tag{1.13}$$

so that, from the first equation of second row from (1.6), the incompressibility condition and the Robin condition for the concentration, we get

$$-\operatorname{div} \boldsymbol{p} \,=\, 0 \quad \text{in} \quad \Omega, \qquad \text{and} \quad \boldsymbol{p} \cdot \boldsymbol{\nu} \,=\, 0 \qquad \text{on} \qquad \Gamma\,. \tag{1.14}$$

Finally, gathering together (1.8), (1.10), (1.11), (1.13) and (1.14), we arrive at the following first-order system with unknowns $\boldsymbol{t}$, $\boldsymbol{\sigma}$, $\boldsymbol{\rho}$, $\boldsymbol{u}$, $\boldsymbol{p}$ and φ

$$\begin{gathered}
\boldsymbol{t} + \boldsymbol{\rho} \,=\, \nabla \boldsymbol{u}\,, \quad \boldsymbol{\sigma}^{\mathrm{d}} + (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathrm{d}} \,=\, \mu(\varphi + \alpha)\boldsymbol{t}\,, \quad -\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f} - g\,(1 + \gamma\varphi)\,\mathbf{i}_3 \quad \text{in} \quad \Omega \\
\kappa^{-1}\boldsymbol{p} + \kappa^{-1}\varphi\boldsymbol{u} + \kappa^{-1}U(\varphi + \alpha)\mathbf{i}_3 \,=\, \nabla\varphi\,, \quad -\operatorname{div} \boldsymbol{p} \,=\, 0 \quad \text{in} \quad \Omega \\
\boldsymbol{u} = \mathbf{0} \quad \text{and} \quad \boldsymbol{p} \cdot \boldsymbol{\nu} \,=\, 0 \quad \text{on} \quad \Gamma \\
\int_\Omega \operatorname{tr}(\boldsymbol{\sigma} + (\boldsymbol{u} \otimes \boldsymbol{u})) = 0 \quad \text{and} \quad \int_\Omega \varphi = 0\,.
\end{gathered} \tag{1.15}$$

Note that according to (1.12), the zero mean value restriction of the pressure on the domain is imposed via the first equation in the last row in (1.15). Also, notice that the incompressibility condition of the fluid is implicitly present through the equilibrium relation (1.10) and by stating that $\boldsymbol{t}$ is a trace-free tensor.

1.3 The continuous formulation

In this section we introduce and analyze the weak formulation of the system described by (1.15). To this end, in Section 1.3.1 we firstly deduce an augmented variational formulation of (1.15) and then in Section 1.3.2 we equivalently rewrite it as a fixed-point problem in terms of operators, which arise by decoupling the fluid equations and the concentration equation. Their well-definiteness and solvability are addressed through Sections 1.3.3 and 1.3.4.

1.3.1 The augmented fully-mixed variational formulation

We first recall (see, e.g., [49] or [55]) that there holds

$$\mathbb{H}(\mathbf{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega) \oplus \mathrm{R}\mathbb{I}, \tag{1.16}$$

where

$$\mathbb{H}_0(\mathbf{div};\Omega) := \left\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega) : \ \int_\Omega \mathrm{tr}\,\boldsymbol{\tau} = 0 \right\},$$

which means that any $\boldsymbol{\zeta} \in \mathbb{H}(\mathbf{div};\Omega)$, can be uniquely written in terms of its orthogonal projection, namely $\boldsymbol{\zeta}_0 \in \mathbb{H}_0(\mathbf{div},\Omega)$, as

$$\boldsymbol{\zeta} \ = \ \boldsymbol{\zeta}_0 \ + \ c\,\mathbb{I}\,, \qquad \text{where} \qquad c = \frac{1}{3\,|\Omega|}\int_\Omega \mathrm{tr}\,\boldsymbol{\zeta}\,.$$

In particular, using the first equation in the last row of (1.15), it is easy to see that an eventual solution $\boldsymbol{\sigma} \in \mathbb{H}(\mathbf{div};\Omega)$ of that system is given by

$$\boldsymbol{\sigma} \ = \ \boldsymbol{\sigma}_0 \ + \ c\mathbb{I} \qquad \text{with} \qquad \boldsymbol{\sigma}_0 \in \mathbb{H}_0(\mathbf{div};\Omega)\,, \quad \text{and} \quad c = -\frac{1}{3\,|\Omega|}\int_\Omega \mathrm{tr}\,(\boldsymbol{u}\otimes\boldsymbol{u})\,. \tag{1.17}$$

Then, since $\boldsymbol{\sigma}^{\mathrm{d}} = \boldsymbol{\sigma}_0^{\mathrm{d}}$ and $\mathbf{div}\,\boldsymbol{\sigma}^{\mathrm{d}} = \mathbf{div}\,\boldsymbol{\sigma}_0^{\mathrm{d}}$, it follows that the equations in (1.15) remain unchanged when replacing there $\boldsymbol{\sigma}_0$ in place of $\boldsymbol{\sigma}$. This fact along with (1.17) allows us to reduce the problem by only looking for $\boldsymbol{\sigma}_0$. According to that, and for simplifying the notation, we set from now on $\boldsymbol{\sigma} := \boldsymbol{\sigma}_0 \in \mathbb{H}_0(\mathbf{div};\Omega)$.

In addition, by their definitions, we introduce the following spaces for the strain tensor $\boldsymbol{t}$ and the vorticity $\boldsymbol{\rho}$, respectively,

$$\mathbb{L}^2_{\mathrm{tr}}(\Omega) := \left\{ \boldsymbol{r} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{r}^{\mathrm{t}} = \boldsymbol{r} \quad \text{and} \quad \mathrm{tr}\,(\boldsymbol{r}) = 0 \right\}, \quad \text{and} \quad \mathbb{L}^2_{\mathrm{skew}}(\Omega) := \left\{ \boldsymbol{\eta} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\eta}^{\mathrm{t}} = -\boldsymbol{\eta} \right\}.$$

Also, the boundary condition for $\boldsymbol{p}$ on Γ (see third row in (1.15)) suggests the introduction of the functional space

$$\mathbf{H}_\Gamma(\mathrm{div};\Omega) := \left\{ \boldsymbol{q} \in \mathbf{H}(\mathrm{div};\Omega) : \ \boldsymbol{q}\cdot\boldsymbol{\nu} = 0 \quad \text{on} \quad \Gamma \right\}.$$

Now, multiplying the first equation in (1.15) by a test function $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$, integrating by parts, using the Dirichlet condition for $\boldsymbol{u}$, and the identity $\boldsymbol{t} : \boldsymbol{\tau} = \boldsymbol{t} : \boldsymbol{\tau}^{\mathrm{d}}$ (since $\boldsymbol{t}$ is trace-free), we get

$$\int_\Omega \boldsymbol{t} : \boldsymbol{\tau}^{\mathrm{d}} + \int_\Omega \boldsymbol{u}\cdot\mathbf{div}\,\boldsymbol{\tau} + \int_\Omega \boldsymbol{\rho} : \boldsymbol{\tau} = 0 \quad \forall\,\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega).$$

Next, testing the second equation from first row in (1.15) with $\boldsymbol{r} \in \mathbb{L}^2_{\rm tr}(\Omega)$, we obtain

$$\int_\Omega \boldsymbol{\sigma}^{\rm d} : \boldsymbol{r} + \int_\Omega (\boldsymbol{u} \otimes \boldsymbol{u})^{\rm d} : \boldsymbol{r} = \int_\Omega \mu(\varphi + \alpha) \boldsymbol{t} : \boldsymbol{r} \quad \forall\, \boldsymbol{r} \in \mathbb{L}^2_{\rm tr}(\Omega).$$

In turn, the equilibrium relation associated to $\boldsymbol{\sigma}$ (third equation from first row in (1.15)) is written as

$$-\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\sigma} = \int_\Omega \big(\boldsymbol{f} - g\,(1 + \gamma\varphi)\,\mathbf{i}_3\big) \cdot \boldsymbol{v} \quad \forall\, \boldsymbol{v} \in \mathbf{L}^2(\Omega),$$

whereas the symmetry of the pseudo-stress tensor is weakly imposed through the identity

$$-\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\eta} = 0 \quad \forall\, \boldsymbol{\eta} \in \mathbb{L}^2_{\rm skew}(\Omega).$$

As for the equations associated to the micro-organisms concentration (second row from (1.15)), we firstly multiply the respective constitutive relation by a function $\boldsymbol{q} \in \mathbf{H}_\Gamma(\mathrm{div}; \Omega)$ and, after integrating by parts, we find

$$\kappa^{-1} \int_\Omega \boldsymbol{p} \cdot \boldsymbol{q} + \kappa^{-1} \int_\Omega \varphi \boldsymbol{u} \cdot \boldsymbol{q} + \kappa^{-1} \int_\Omega U(\varphi + \alpha) \mathbf{i}_3 \cdot \boldsymbol{q} = -\int_\Omega \varphi\, \mathrm{div} \boldsymbol{q} \quad \forall\, \boldsymbol{q} \in \mathbf{H}_\Gamma(\mathrm{div}; \Omega)\,,$$

and the equilibrium relation for the concentration is weakly expressed as

$$-\int_\Omega \psi\, \mathrm{div}\, \boldsymbol{p} = 0 \quad \forall\, \psi \in \mathrm{L}^2(\Omega).$$

In this way, we arrive at first instance to the mixed formulation: Find $\boldsymbol{t} \in \mathbb{L}^2_{\rm tr}(\Omega)$, $\boldsymbol{\sigma} \in \mathbb{H}_0(\mathbf{div}; \Omega)$, $\boldsymbol{\rho} \in \mathbb{L}^2_{\rm skew}(\Omega)$, $\boldsymbol{p} \in \mathbf{H}_\Gamma(\mathrm{div}; \Omega)$, and $\boldsymbol{u}$, φ in suitable spaces to be specified below, such that

$$\begin{aligned}
\int_\Omega \boldsymbol{t} : \boldsymbol{\tau}^{\rm d} + \int_\Omega \boldsymbol{u} \cdot \mathbf{div}\, \boldsymbol{\tau} + \int_\Omega \boldsymbol{\rho} : \boldsymbol{\tau} &= 0, \\
\int_\Omega \mu(\varphi + \alpha) \boldsymbol{t} : \boldsymbol{r} - \int_\Omega \boldsymbol{\sigma}^{\rm d} : \boldsymbol{r} - \int_\Omega (\boldsymbol{u} \otimes \boldsymbol{u})^{\rm d} : \boldsymbol{r} &= 0, \\
-\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\sigma} - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\eta} &= \int_\Omega (\boldsymbol{f} - g\,(1 + \gamma\varphi)\,\mathbf{i}_3) \cdot \boldsymbol{v}, \\
\kappa^{-1} \int_\Omega \boldsymbol{p} \cdot \boldsymbol{q} + \int_\Omega \varphi\, \mathrm{div} \boldsymbol{q} + \kappa^{-1} \int_\Omega \varphi \boldsymbol{u} \cdot \boldsymbol{q} &= -\kappa^{-1} \int_\Omega U(\varphi + \alpha) \mathbf{i}_3 \cdot \boldsymbol{q}, \\
-\int_\Omega \psi\, \mathrm{div}\, \boldsymbol{p} &= 0,
\end{aligned} \tag{1.18}$$

for all $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega)$, $\boldsymbol{r} \in \mathbb{L}^2_{\rm tr}(\Omega)$, $(\boldsymbol{\eta}, \boldsymbol{v}) \in \mathbb{L}^2_{\rm skew}(\Omega) \times \mathbf{L}^2(\Omega)$, $\boldsymbol{q} \in \mathbf{H}_\Gamma(\mathrm{div}; \Omega)$, and $\psi \in \mathrm{L}^2(\Omega)$. Note that the third terms on the left-hand side of the second and fourth equations in (1.18) require a suitable regularity for both unknowns $\boldsymbol{u}$ and φ. Indeed, by applying Cauchy-Schwarz and Hölder inequalities, and then the continuous injections $i : \mathrm{H}^1(\Omega) \to \mathrm{L}^4(\Omega)$ and $\boldsymbol{i} : \mathbf{H}^1(\Omega) \to \mathbf{L}^4(\Omega)$ (see e.g. [1] or [80]), we deduce that there exist positive constants $c_1(\Omega) := \|i\|\, \|\boldsymbol{i}\|$ and $c_2(\Omega) := \|\boldsymbol{i}\|^2$, such that

$$\left| \int_\Omega \varphi \boldsymbol{u} \cdot \boldsymbol{q} \right| \leq c_1(\Omega) \|\varphi\|_{1,\Omega} \|\boldsymbol{u}\|_{1,\Omega} \|\boldsymbol{q}\|_{0,\Omega} \qquad \forall\, \varphi \in \mathrm{H}^1(\Omega), \forall\, \boldsymbol{u} \in \mathbf{H}^1(\Omega), \forall\, \boldsymbol{q} \in \mathbf{L}^2(\Omega), \tag{1.19}$$

and

$$\left|\int_\Omega (\boldsymbol{u}\otimes\boldsymbol{w})^{\mathrm{d}} : \boldsymbol{r}\right| \leq c_2(\Omega)\|\boldsymbol{u}\|_{1,\Omega}\|\boldsymbol{w}\|_{1,\Omega}\|\boldsymbol{r}\|_{0,\Omega} \qquad \forall\, \boldsymbol{u},\boldsymbol{w}\in \mathbf{H}^1(\Omega), \forall\, \boldsymbol{r}\in \mathbb{L}^2(\Omega). \tag{1.20}$$

In light of the above, and in order to be able to set the variational formulation (1.18) in a framework on standard Hilbert spaces for both the velocity and concentration, we propose to seek $\boldsymbol{u}\in\mathbf{H}^1_0(\Omega)$ and $\varphi\in\widetilde{\mathrm{H}}^1(\Omega)$, and so their respective test spaces. In turn, similarly as in [26, Section 2] (see also [3,37]), we additionally augment (1.18) by incorporating the following redundant Galerkin terms coming from the constitutive and equilibrium equations,

$$\begin{aligned}
\kappa_1\int_\Omega (\mathbf{e}(\boldsymbol{u})-\boldsymbol{t}) : \mathbf{e}(\boldsymbol{v}) = 0 \qquad & \forall\, \boldsymbol{v}\in\mathbf{H}^1_0(\Omega)\,,\\
\kappa_2\int_\Omega \mathbf{div}\,\boldsymbol{\sigma}\cdot\mathbf{div}\,\boldsymbol{\tau} = -\kappa_2\int_\Omega (\boldsymbol{f}-g\,(1+\gamma\varphi)\,\mathbf{i}_3)\cdot\mathbf{div}\,\boldsymbol{\tau} \qquad & \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\
\kappa_3\int_\Omega \Big\{\boldsymbol{\sigma}^{\mathrm{d}}+(\boldsymbol{u}\otimes\boldsymbol{u})^{\mathrm{d}}-\mu(\varphi+\alpha)\boldsymbol{t}\Big\} : \boldsymbol{\tau}^{\mathrm{d}} = 0 \qquad & \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\
\kappa_4\int_\Omega \big\{\boldsymbol{\rho}-\big(\nabla\boldsymbol{u}-\mathbf{e}(\boldsymbol{u})\big)\big\} : \boldsymbol{\eta} = 0 \qquad & \forall\boldsymbol{\eta}\in\mathbb{L}^2_{\mathrm{skew}}(\Omega)\,,
\end{aligned} \tag{1.21}$$

and

$$\begin{aligned}
\kappa_5\int_\Omega \big\{\nabla\varphi-\kappa^{-1}\boldsymbol{p}-\kappa^{-1}\varphi\boldsymbol{u}-\kappa^{-1}U(\varphi+\alpha)\mathbf{i}_3\big\}\cdot\nabla\psi = 0 \quad & \forall\,\psi\in\widetilde{\mathrm{H}}^1(\Omega),\\
\kappa_6\int_\Omega \mathrm{div}\,\boldsymbol{p}\,\mathrm{div}\,\boldsymbol{q} = 0 \quad & \forall\,\boldsymbol{q}\in\mathbf{H}_\Gamma(\mathrm{div};\Omega),
\end{aligned} \tag{1.22}$$

where $(\kappa_1,\kappa_2,\kappa_3,\kappa_4,\kappa_5,\kappa_6)$ is a vector of positive parameters to be specified later in Section 1.3.3. Hence, letting

$$\underline{\boldsymbol{t}} := (\boldsymbol{t},\boldsymbol{\sigma},\boldsymbol{\rho})\in\mathbb{H} := \mathbb{L}^2_{\mathrm{tr}}(\Omega)\times\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega),$$

where $\mathbb{H}$ is endowed with the natural norm

$$\|\underline{\boldsymbol{r}}\|_{\mathbb{H}} := \Big\{\|\boldsymbol{r}\|^2_{0,\Omega}+\|\boldsymbol{\tau}\|^2_{\mathbf{div},\Omega}+\|\boldsymbol{\eta}\|^2_{0,\Omega}\Big\}^{1/2} \quad \forall\,\underline{\boldsymbol{r}} := (\boldsymbol{r},\boldsymbol{\tau},\boldsymbol{\eta})\in\mathbb{H}\,,$$

and adding up (1.18) with (1.21) and (1.22), we arrive at the following augmented fully-mixed formulation for the bioconvective flow problem1: Find $(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi)\in\mathbb{H}\times\mathbf{H}^1_0(\Omega)\times\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$ such that

$$\begin{aligned}
\mathbf{A}_\varphi((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v}))+\mathbf{B}_{\boldsymbol{u}}((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) &= F_\varphi(\underline{\boldsymbol{r}},\boldsymbol{v}) \quad \forall\,(\underline{\boldsymbol{r}},\boldsymbol{v})\in\mathbb{H}\times\mathbf{H}^1_0(\Omega),\\
\widetilde{\mathbf{A}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi))+\widetilde{\mathbf{B}}_{\boldsymbol{u}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) &= \widetilde{F}_\varphi(\boldsymbol{q},\psi) \quad \forall\,(\boldsymbol{q},\psi)\in\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega),
\end{aligned} \tag{1.23}$$

where, given $\phi\in\widetilde{\mathrm{H}}^1(\Omega)$ and $\boldsymbol{w}\in\mathbf{H}^1_0(\Omega)$, $\mathbf{A}_\phi$, $\mathbf{B}_{\boldsymbol{w}}$, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}_{\boldsymbol{w}}$ are the bilinear forms defined, respectively, as

$$\begin{aligned}
\mathbf{A}_\phi((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) &:= \int_\Omega \mu(\phi+\alpha)\boldsymbol{t} : \big(\boldsymbol{r}-\kappa_3\boldsymbol{\tau}^{\mathrm{d}}\big)+\int_\Omega \boldsymbol{\sigma}^{\mathrm{d}} : \big(\kappa_3\boldsymbol{\tau}^{\mathrm{d}}-\boldsymbol{r}\big)+\int_\Omega \boldsymbol{t} : \boldsymbol{\tau}^{\mathrm{d}}\\
&+\int_\Omega \big(\boldsymbol{u}+\kappa_2\mathbf{div}\,\boldsymbol{\sigma}\big)\cdot\mathbf{div}\,\boldsymbol{\tau}-\int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\sigma}+\int_\Omega \boldsymbol{\rho} : \boldsymbol{\tau}-\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\eta}\\
&+\kappa_1\int_\Omega \big(\mathbf{e}(\boldsymbol{u})-\boldsymbol{t}\big) : \mathbf{e}(\boldsymbol{v})+\kappa_4\int_\Omega \big\{\boldsymbol{\rho}-\big(\nabla\boldsymbol{u}-\mathbf{e}(\boldsymbol{u})\big)\big\} : \boldsymbol{\eta}\,,
\end{aligned} \tag{1.24}$$

$$\mathbf{B}_{\boldsymbol{w}}((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) := \int_\Omega (\boldsymbol{u}\otimes\boldsymbol{w})^{\mathrm{d}} : \left(\kappa_3\boldsymbol{\tau}^{\mathrm{d}} - \boldsymbol{r}\right), \tag{1.25}$$

$$\begin{aligned}\widetilde{\mathbf{A}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) := \kappa^{-1}\int_\Omega \boldsymbol{p}\cdot\left(\boldsymbol{q}-\kappa_5\nabla\psi\right) + \int_\Omega \left(\varphi + \kappa_6\,\mathrm{div}\boldsymbol{p}\right)\mathrm{div}\boldsymbol{q}\\ -\int_\Omega \psi\,\mathrm{div}\,\boldsymbol{p} + \kappa_5\int_\Omega \nabla\varphi\cdot\nabla\psi,\end{aligned} \tag{1.26}$$

and

$$\widetilde{\mathbf{B}}_{\boldsymbol{w}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) := \kappa^{-1}\int_\Omega \varphi\boldsymbol{w}\cdot\left(\boldsymbol{q}-\kappa_5\nabla\psi\right), \tag{1.27}$$

for all $(\underline{\boldsymbol{t}},\boldsymbol{u})$, $(\underline{\boldsymbol{r}},\boldsymbol{v}) \in \mathbb{H}\times\mathbf{H}^1_0(\Omega)$ and for all $(\boldsymbol{p},\varphi)$, $(\boldsymbol{q},\psi)\in \mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$. In turn, given $\phi\in\widetilde{\mathrm{H}}^1(\Omega)$, F_ϕ and $\widetilde{F}_\phi$ are the bounded linear functionals given by

$$F_\phi(\underline{\boldsymbol{r}},\boldsymbol{v}) := \int_\Omega (\boldsymbol{f} - g\,(1+\gamma\phi)\,\mathbf{i}_3)\cdot\left(\boldsymbol{v}-\kappa_2\mathbf{div}\,\boldsymbol{\tau}\right) \qquad \forall\,(\underline{\boldsymbol{r}},\boldsymbol{v})\in\mathbb{H}\times\mathbf{H}^1_0(\Omega), \tag{1.28}$$

and

$$\widetilde{F}_\phi(\boldsymbol{q},\psi) := -\kappa^{-1}\int_\Omega U(\phi+\alpha)\mathbf{i}_3\cdot\left(\boldsymbol{q}-\kappa_5\nabla\psi\right) \qquad \forall\,(\boldsymbol{q},\psi)\in\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega). \tag{1.29}$$

1.3.2 The fixed point approach

Now, we proceed similarly as in [36] (see also [26,37]) and rewrite (1.23) as an equivalent fixed-point equation in terms of a certain operator $\mathbf{T}$ to be defined below. Firstly, we set $\mathbf{H} := \mathbf{H}^1_0(\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$ and start by introducing the operator $\mathbf{S}:\mathbf{H}\longrightarrow\mathbb{H}\times\mathbf{H}^1_0(\Omega)$ by

$$\mathbf{S}(\boldsymbol{w},\phi) := \Big(\left(\mathbf{S}_1(\boldsymbol{w},\phi),\mathbf{S}_2(\boldsymbol{w},\phi),\mathbf{S}_3(\boldsymbol{w},\phi)\right),\,\mathbf{S}_4(\boldsymbol{w},\phi)\Big) = (\underline{\boldsymbol{t}},\boldsymbol{u}) \qquad \forall\,(\boldsymbol{w},\phi)\in\mathbf{H}, \tag{1.30}$$

where, given $(\boldsymbol{w},\phi)\in\mathbf{H}$, $(\underline{\boldsymbol{t}},\boldsymbol{u})$ is the unique solution to the problem1: Find $(\underline{\boldsymbol{t}},\boldsymbol{u})\in\mathbb{H}\times\mathbf{H}^1_0(\Omega)$ such that

$$\mathbf{A}_\phi((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) + \mathbf{B}_{\boldsymbol{w}}((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) = F_\phi(\underline{\boldsymbol{r}},\boldsymbol{v}) \qquad \forall\,(\underline{\boldsymbol{r}},\boldsymbol{v})\in\mathbb{H}\times\mathbf{H}^1_0(\Omega)\,. \tag{1.31}$$

In addition, we also introduce the operator $\widetilde{\mathbf{S}}:\mathbf{H}\longrightarrow\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$ defined as

$$\widetilde{\mathbf{S}}(\boldsymbol{w},\phi) := \left(\widetilde{\mathbf{S}}_1(\boldsymbol{w},\phi),\widetilde{\mathbf{S}}_2(\boldsymbol{w},\phi)\right) = (\boldsymbol{p},\varphi) \qquad \forall\,(\boldsymbol{w},\phi)\in\mathbf{H}, \tag{1.32}$$

where, given $(\boldsymbol{w},\phi)\in\mathbf{H}$, $(\boldsymbol{p},\varphi)$ is the unique solution to the problem1: Find $(\boldsymbol{p},\varphi)\in\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$ such that

$$\widetilde{\mathbf{A}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) + \widetilde{\mathbf{B}}_{\boldsymbol{w}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) = \widetilde{F}_\phi(\boldsymbol{q},\psi) \quad \forall\,(\boldsymbol{q},\psi)\in\mathbf{H}_\Gamma(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega). \tag{1.33}$$

Having introduced the auxiliary mappings $\mathbf{S}$ and $\widetilde{\mathbf{S}}$, we now define the operator $\mathbf{T}:\mathbf{H}\longrightarrow\mathbf{H}$ as

$$\mathbf{T}(\boldsymbol{w},\phi) := \Big(\mathbf{S}_4(\boldsymbol{w},\phi),\widetilde{\mathbf{S}}_2\big(\mathbf{S}_4(\boldsymbol{w},\phi),\phi\big)\Big) \quad \forall(\boldsymbol{w},\phi)\in\mathbf{H}\,, \tag{1.34}$$

and realize that (1.23) can be rewritten as the fixed point problem1: Find $(\boldsymbol{u},\varphi)\in\mathbf{H}$ such that

$$\mathbf{T}(\boldsymbol{u},\varphi) = (\boldsymbol{u},\varphi). \tag{1.35}$$

In this way, through the following sections we study the conditions under which the operator $\mathbf{T}$ is well-defined, has a fixed point and when it is unique.

1.3.3 Well-definiteness of the fixed point operator

In what follows we show that $\mathbf{T}$ is well-defined. Notice that it suffices to prove that the uncoupled problems (1.31) and (1.33) defining $\mathbf{S}$ and $\widetilde{\mathbf{S}}$, respectively, are well-posed. To state the solvability of (1.31), we start studying the stability properties of the forms $\mathbf{A}_\phi$ and $\mathbf{B}_{\boldsymbol{w}}$ and the functional F_ϕ (cf. (1.24), (1.25) and (1.28), respectively). Firstly, given $\phi \in \widetilde{\mathrm{H}}^1(\Omega)$, from the Cauchy-Schwarz inequality we find that there exists a positive constant, denoted by $\|\mathbf{A}_\phi\|$, and depending on μ_2 (cf. (1.3)) and the parameters $\kappa_1, \kappa_2, \kappa_3, \kappa_4$, such that

$$|\mathbf{A}_\phi((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{r}}, \boldsymbol{v}))| \leq \|\mathbf{A}_\phi\| \, \|(\underline{\boldsymbol{t}}, \boldsymbol{u})\| \, \|(\underline{\boldsymbol{r}}, \boldsymbol{v})\| \qquad \forall\, (\underline{\boldsymbol{t}}, \boldsymbol{u}),\, (\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega). \tag{1.36}$$

Also, given $\boldsymbol{w} \in \mathbf{H}^1_0(\Omega)$, from the estimation (1.20) we have that

$$|\mathbf{B}_{\boldsymbol{w}}((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{r}}, \boldsymbol{v}))| \leq c_2(\Omega)(1 + \kappa_3^2)^{1/2} \|\boldsymbol{w}\|_{1,\Omega} \|\boldsymbol{u}\|_{1,\Omega} \|(\underline{\boldsymbol{r}}, \boldsymbol{v})\| \qquad \forall\, (\underline{\boldsymbol{t}}, \boldsymbol{u}),\, (\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega). \tag{1.37}$$

It then follows from (1.36) and (1.37) that there exists a positive constant, denoted by $\|\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}\|$, and depending on μ_2, $\kappa_1, \kappa_2, \kappa_3, \kappa_4$, $c_2(\Omega)$, and $\|\boldsymbol{w}\|_{1,\Omega}$, such that

$$|\big(\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}\big)((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{r}}, \boldsymbol{v}))| \leq \|\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}\| \, \|(\underline{\boldsymbol{t}}, \boldsymbol{u})\| \, \|(\underline{\boldsymbol{r}}, \boldsymbol{v})\| \qquad \forall\, (\underline{\boldsymbol{t}}, \boldsymbol{u}),\, (\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega)\,. \tag{1.38}$$

Regarding the ellipticity of $\mathbf{A}_\phi$ we proceed similarly to [26, Lemma 3.1]. So, we use the bounds for $\mu(\,\cdot\,)$ (cf. (1.3)), the Cauchy-Schwarz and Young inequalities (with $\delta_1, \delta_2, \delta_3 > 0$), and subsequently the Korn inequality and the Poincaré inequality (see [81, Théorème 1.2-5]) with constant c_p, to deduce that there exists $\alpha(\Omega) > 0$ satisfying

$$\mathbf{A}_\phi((\underline{\boldsymbol{r}}, \boldsymbol{v}), (\underline{\boldsymbol{r}}, \boldsymbol{v})) \geq \alpha(\Omega) \, \|(\underline{\boldsymbol{r}}, \boldsymbol{v})\|^2 \qquad \forall\, (\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega), \tag{1.39}$$

where

$$\alpha(\Omega) := \min\Big\{\alpha_1(\Omega), \alpha_3(\Omega), c_p\alpha_4(\Omega), \alpha_5(\Omega)\Big\}, \tag{1.40}$$

with

$$\alpha_1(\Omega) := \mu_1 - \frac{\kappa_3\mu_2}{2\delta_1} - \frac{\kappa_1}{2\delta_2}, \quad \alpha_2(\Omega) := \min\left\{\kappa_3\left(1 - \frac{\mu_2\delta_1}{2}\right), \frac{\kappa_2}{2}\right\},$$

$$\alpha_3(\Omega) := \min\left\{c_3(\Omega)\alpha_2(\Omega), \frac{\kappa_2}{2}\right\}, \quad \alpha_4(\Omega) := \frac{\kappa_1}{2}\left(1 - \frac{\delta_2}{2}\right) - \frac{\kappa_4}{4\delta_3}, \quad \text{and} \quad \alpha_5(\Omega) := \kappa_4\left(1 - \frac{\delta_3}{2}\right),$$

and $c_3(\Omega) > 0$ (see [49, Lemma 2.3] for details) is such that

$$c_3(\Omega)\|\boldsymbol{\tau}\|^2_{0,\Omega} \leq \|\boldsymbol{\tau}^{\mathrm{d}}\|^2_{0,\Omega} + \|\mathbf{div}\,\boldsymbol{\tau}\|^2_{0,\Omega} \qquad \forall \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega)\,.$$

In turn, the positivity of $\alpha(\Omega)$ is ensured as long as the constants α_i in (1.40) are positive, which gives the following feasible ranges for the parameters $(\kappa_i)_{1\leq i\leq 4}$,

$$0 < \kappa_1 < 2\delta_2\Big(\mu_1 - \frac{\mu_2\kappa_3}{2\delta_1}\Big), \quad \kappa_2 > 0\,, \quad 0 < \kappa_3 < \frac{2\delta_1\mu_1}{\mu_2} \quad \text{and} \quad 0 < \kappa_4 < 2\delta_3\kappa_1\Big(1 - \frac{\delta_2}{2}\Big) \tag{1.41}$$

with

$$0 < \delta_1 < \frac{2}{\mu_2} \quad \text{and} \quad 0 < \delta_2\,,\, \delta_3 < 2\,. \tag{1.42}$$

Next, combining (1.37) with (1.39), we have that for all $(\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega)$ there holds

$$(\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}})((\underline{\boldsymbol{r}}, \boldsymbol{v}), (\underline{\boldsymbol{r}}, \boldsymbol{v})) \geq \Big\{\alpha(\Omega) - c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{w}\|_{1,\Omega}\Big\}\|(\underline{\boldsymbol{r}}, \boldsymbol{v})\|^2 \geq \frac{\alpha(\Omega)}{2}\|(\underline{\boldsymbol{r}}, \boldsymbol{v})\|^2, \quad (1.43)$$

provided $c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{w}\|_{1,\Omega} \leq \dfrac{\alpha(\Omega)}{2}$. Therefore, the ellipticity of the form $\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}$ is ensured with the constant $\dfrac{\alpha(\Omega)}{2} > 0$, independent of $\boldsymbol{w}$, by requiring $\|\boldsymbol{w}\|_{1,\Omega} \leq r_0$, with

$$r_0 := \frac{\alpha(\Omega)}{2c_2(\Omega)(1+\kappa_3^2)^{1/2}}. \quad (1.44)$$

Finally, the functional F_ϕ (with $\phi \in \widetilde{\mathrm{H}}^1(\Omega)$, given) is clearly linear in $\mathbb{H} \times \mathbf{H}^1_0(\Omega)$, and using Cauchy-Schwarz inequality, we conclude with $M_{\mathbf{S}} := (1+\kappa_2^2)^{1/2}$, that

$$\|F_\phi\| \leq M_{\mathbf{S}}\Big\{\|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\|\phi\|_{0,\Omega}\Big)\|\boldsymbol{g}\|_{\infty,\Omega}\Big\}. \quad (1.45)$$

where $\boldsymbol{g} := g\,\mathbf{i}_3 \in \mathbf{L}^\infty(\Omega)$. The foregoing analysis essentially gives us conditions for the well-posedness of the uncoupled problem (1.31) or, equivalently, the well definition of the operator $\mathbf{S}$ (cf. (1.30)). This is summarized in the following result.

Lemma 1.1. *Let $r_0 > 0$ given by (1.44) and let $r \in (0, r_0)$. Assume that $\kappa_1 \in \Big(0, 2\delta_2\Big(\mu_1 - \frac{\kappa_3\mu_2}{2\delta_1}\Big)\Big)$, $\kappa_2 > 0$, $\kappa_3 \in \Big(0, \frac{2\delta_1\mu_1}{\mu_2}\Big)$, and $\kappa_4 \in \Big(0, 2\delta_3\kappa_1\Big(1 - \frac{\delta_2}{2}\Big)\Big)$, with $\delta_1 \in \Big(0, \frac{2}{\mu_2}\Big)$, and $\delta_2, \delta_3 \in (0,2)$. Then, for each $(\boldsymbol{w}, \phi) \in \mathbf{H}$ such that $\|\boldsymbol{w}\|_{1,\Omega} \leq r$, there exist a unique solution $(\underline{\boldsymbol{t}}, \boldsymbol{u}) = \mathbf{S}(\boldsymbol{w}, \phi) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega)$ to problem (1.31) and a positive constant $c_{\mathbf{S}} > 0$, independent of $(\boldsymbol{w}, \phi)$, such that*

$$\|\mathbf{S}(\boldsymbol{w}, \phi)\| = \|(\underline{\boldsymbol{t}}, \boldsymbol{u})\| \leq c_{\mathbf{S}}\Big\{\|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\|\phi\|_{0,\Omega}\Big)\|\boldsymbol{g}\|_{\infty,\Omega}\Big\}. \quad (1.46)$$

Proof. It follows from the estimates (1.38), (1.43) and (1.45) and a straightforward application of the Lax-Milgram Theorem (see e.g. [49, Theorem 1.1]), and the respective continuous dependence result gives the a priori estimate (1.46) with $c_{\mathbf{S}} := \dfrac{2M_{\mathbf{S}}}{\alpha(\Omega)}$. In turn, the ranges for the parameters are stated according to (1.41)-(1.42), guaranteeing the positivity of the ellipticity constant $\alpha(\Omega)$. ∎

Next, we concentrate in proving that problem (1.33) is well posed or, in other words, that the operator $\widetilde{\mathbf{S}}$ (cf. (1.32)) is well-defined. The following lemma establishes this result.

Lemma 1.2. *Assume that $\kappa_5 \in (0, 2\widetilde{\delta})$, with $\widetilde{\delta} \in (0, 2\kappa)$, and $\kappa_6 > 0$. Then, there exists a positive constant $\widetilde{r}_0$ (see (1.52) below) such that for all $r \in (0, \widetilde{r}_0)$ and $(\boldsymbol{w}, \phi) \in \mathbf{H}$ with $\|\boldsymbol{w}\|_{1,\Omega} \leq r$, the problem (1.33) has a unique solution $(\boldsymbol{p}, \varphi) := \widetilde{\mathbf{S}}(\boldsymbol{w}, \phi) \in \mathbf{H}_\Gamma(\mathrm{div}; \Omega) \times \widetilde{\mathrm{H}}^1(\Omega)$. Moreover, there exists a constant $c_{\widetilde{\mathbf{S}}} > 0$, independent of $(\boldsymbol{w}, \phi)$, satisfying*

$$\|\widetilde{\mathbf{S}}(\boldsymbol{w}, \phi)\| = \|(\boldsymbol{p}, \varphi)\| \leq c_{\widetilde{\mathbf{S}}}\,\kappa^{-1}\,U\Big\{\alpha|\Omega|^{1/2} + \|\phi\|_{0,\Omega}\Big\}. \quad (1.47)$$

Proof. For a given $\boldsymbol{w} \in \mathbf{H}^1_0(\Omega)$, we firstly observe from (1.26) and (1.27) that $\widetilde{\mathbf{A}} + \widetilde{\mathbf{B}}_{\boldsymbol{w}}$ is clearly a bilinear form. Also, from the Cauchy-Schwarz inequality we have that

$$|\widetilde{\mathbf{A}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi))| \,\leq\, \|\widetilde{\mathbf{A}}\|\,\|(\boldsymbol{p},\varphi)\|\,\|(\boldsymbol{q},\psi)\|\,,$$

where $\|\widetilde{\mathbf{A}}\|$ depends on κ, κ_5 and κ_6, and from the estimate (1.19) we get

$$|\widetilde{\mathbf{B}}_{\boldsymbol{w}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi))\,| \,\leq\, \kappa^{-1}(1+\kappa_5^2)^{1/2}c_1(\Omega)\|\boldsymbol{w}\|_{1,\Omega}\|\varphi\|_{1,\Omega}\|(\boldsymbol{q},\psi)\|\,, \tag{1.48}$$

for all $(\boldsymbol{p},\varphi)$, $(\boldsymbol{q},\psi) \in \mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$. Then, by gathering the foregoing estimates, we find that there exists a positive constant, which we denote by $\|\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{w}}\|$, only depending on $\kappa, \kappa_5, \kappa_6$ and $c_1(\Omega)$, such that

$$|\widetilde{\mathbf{A}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) + \widetilde{\mathbf{B}}_{\boldsymbol{w}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi))| \leq \|\widetilde{\mathbf{A}} + \widetilde{\mathbf{B}}_{\boldsymbol{w}}\|\,\|(\boldsymbol{p},\varphi)\|\,\|(\boldsymbol{q},\psi)\|,$$

for all $(\boldsymbol{p},\varphi)$, $(\boldsymbol{q},\psi) \in \mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$. Likewise, from the definition of the bilinear form $\widetilde{\mathbf{A}}$ (cf. (1.26)), we have that

$$\widetilde{\mathbf{A}}((\boldsymbol{q},\psi),(\boldsymbol{q},\psi)) \,=\, \kappa^{-1}\|\boldsymbol{q}\|^2_{0,\Omega} \,-\, \kappa^{-1}\,\kappa_5\int_{\Omega}\boldsymbol{q}\cdot\nabla\psi \,+\, \kappa_6\,\|\mathrm{div}\boldsymbol{q}\|^2_{0,\Omega} \,+\, \kappa_5\,|\psi|^2_{1,\Omega},$$

and hence, using the Cauchy-Schwarz inequality and the Young inequality with $\widetilde{\delta} > 0$, we obtain for all $(\boldsymbol{q},\psi) \in \mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$ that

$$\widetilde{\mathbf{A}}((\boldsymbol{q},\psi),(\boldsymbol{q},\psi)) \,\geq\, \kappa^{-1}\left(1-\frac{\kappa_5}{2\widetilde{\delta}}\right)\|\boldsymbol{q}\|^2_{0,\Omega} \,+\, \kappa_6\,\|\mathrm{div}\boldsymbol{q}\|^2_{0,\Omega} \,+\, \kappa_5\left(1-\frac{\kappa^{-1}\,\widetilde{\delta}}{2}\right)|\psi|^2_{1,\Omega}. \tag{1.49}$$

In this way, recalling that the norm and semi-norm are equivalent in the space $\widetilde{\mathrm{H}}^1(\Omega)$ (cf. 1.7), we apply the generalized Poincaré inequality with constant $\widetilde{c}_p$ to the last term in (1.49) (see [48, Teorema 9.13]), and define the constants

$$\widetilde{\alpha}_1(\Omega) := \min\left\{\kappa^{-1}\left(1-\frac{\kappa_5}{2\widetilde{\delta}}\right),\kappa_6\right\} \quad \text{and} \quad \widetilde{\alpha}_2(\Omega) := \kappa_5\left(1-\frac{\kappa^{-1}\,\widetilde{\delta}}{2}\right),$$

which are positive thanks to the hypotheses on $\widetilde{\delta}, \kappa_5$ and κ_6, to obtain

$$\widetilde{\mathbf{A}}((\boldsymbol{q},\psi),(\boldsymbol{q},\psi)) \,\geq\, \widetilde{\alpha}(\Omega)\|(\boldsymbol{q},\psi)\|^2 \qquad \forall(\boldsymbol{q},\psi)\in\mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega), \tag{1.50}$$

with $\widetilde{\alpha}(\Omega) := \min\{\widetilde{\alpha}_1(\Omega), \widetilde{c}_p\widetilde{\alpha}_2(\Omega)\}$, which shows that $\widetilde{\mathbf{A}}$ is elliptic. Therefore, combining (1.48) and (1.50), we deduce that for all $(\boldsymbol{q},\psi) \in \mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$, there holds

$$(\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{w}})((\boldsymbol{q},\psi),(\boldsymbol{q},\psi)) \geq \left\{\widetilde{\alpha}(\Omega) - \kappa^{-1}(1+\kappa_5^2)^{1/2}c_1(\Omega)\|\boldsymbol{w}\|_{1,\Omega}\right\}\|(\boldsymbol{q},\psi)\|^2 \geq \frac{\widetilde{\alpha}(\Omega)}{2}\|(\boldsymbol{q},\psi)\|^2, \tag{1.51}$$

whenever $\kappa^{-1}(1+\kappa_5^2)^{1/2}c_1(\Omega)\|\boldsymbol{w}\|_{1,\Omega} \leq \dfrac{\widetilde{\alpha}(\Omega)}{2}$. Thus, the ellipticity of $\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{w}}$ with constant $\dfrac{\widetilde{\alpha}(\Omega)}{2}$, independent of $\boldsymbol{w}$, is ensured by requiring $\|\boldsymbol{w}\|_{1,\Omega} \leq \widetilde{r}_0$, with

$$\widetilde{r}_0 := \frac{\widetilde{\alpha}(\Omega)}{2\kappa^{-1}(1+\kappa_5^2)^{1/2}c_1(\Omega)}. \tag{1.52}$$

Next, it is easy to see from (1.29) that the functional $\widetilde{F}_\phi$ is bounded with

$$\|\widetilde{F}_\phi\| \,\leq\, \kappa^{-1}\,U\,(1+\kappa_5^2)^{1/2}\left\{\alpha|\Omega|^{1/2}+\|\phi\|_{0,\Omega}\right\}. \tag{1.53}$$

Summing up, and owing to the hypotheses on κ_5 and κ_6, we have proved that for any sufficiently small $\boldsymbol{w}\in\mathbf{H}^1_0(\Omega)$, the bilinear form $\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{w}}$ and the functional $\widetilde{F}_\phi$ satisfy the hypotheses of the Lax-Milgram Theorem, which guarantees the well-posedness of (1.33) and the a priori estimate (1.47) with $c_{\widetilde{\mathbf{S}}}:=\dfrac{2}{\widetilde{\alpha}(\Omega)}(1+\kappa_5^2)^{1/2}$. ■

At this point, we remark that, for computational purposes, the constants $\alpha(\Omega)$ and $\widetilde{\alpha}(\Omega)$ yielding the ellipticity of $\mathbf{A}_\phi+\mathbf{B}_{\boldsymbol{w}}$ and $\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{w}}$, respectively, can be maximized by taking the parameters δ_1, δ_2, δ_3, κ_1,κ_3, κ_4, $\widetilde{\delta}$ and κ_5 as the middle points of their feasible ranges, and by choosing κ_2 and κ_6 so that they maximize the minima defining $\alpha_2(\Omega)$ and $\widetilde{\alpha}_1(\Omega)$, respectively. More precisely, we take

$$\begin{gathered}\delta_1=\frac{1}{\mu_2},\quad \delta_2=\delta_3=1,\quad \kappa_3=\frac{\delta_1\mu_1}{\mu_2}=\frac{\mu_1}{\mu_2^2},\quad \kappa_1=\delta_2\left(\mu_1-\frac{\kappa_3\mu_2}{2\delta_1}\right)=\frac{\mu_1}{2},\\ \kappa_4=\delta_3\kappa_1\left(1-\frac{\delta_2}{2}\right)=\frac{\mu_1}{4},\quad \kappa_2=2\kappa_3\left(1-\frac{\mu_2\delta_1}{2}\right)=\frac{\mu_1}{\mu_2^2},\quad \widetilde{\delta}=\kappa,\\ \kappa_5=\widetilde{\delta}=\kappa,\quad \text{and}\quad \kappa_6=\kappa^{-1}\left(1-\frac{\kappa_5}{2\widetilde{\delta}}\right)=\frac{\kappa^{-1}}{2}\,.\end{gathered} \tag{1.54}$$

The explicit values of the stabilization parameters κ_i, $i\in\{1,\ldots,6\}$, given above will be employed in Section 1.6 for the corresponding numerical examples.

1.3.4 Solvability analysis of the fixed point equation

Having proved the well-posedness of the uncoupled problems (1.31) and (1.33), which ensures that the operators $\mathbf{S}$, $\widetilde{\mathbf{S}}$ and $\mathbf{T}$ are well defined, we now aim to establish the existence of a unique fixed point of the operator $\mathbf{T}$. For this purpose, in what follows we will verify the hypothesis of the Banach fixed-point theorem (see, e.g. [30, Theorem 3.7-1]). We begin with the following result.

Lemma 1.3. *Suppose that the parameters κ_i, $i\in\{1,\ldots,6\}$, satisfy the conditions required by Lemmas 1.1 and 1.2. Given $r\in(0,\min\{r_0,\widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (1.44) and (1.52), respectively, we let W_r be the closed ball in $\mathbf{H}$ defined by*

$$W_r:=\left\{(\boldsymbol{w},\phi)\in\mathbf{H}:\ \ \|(\boldsymbol{w},\phi)\|\leq r\right\}, \tag{1.55}$$

and assume that the data satisfy

$$c_{\mathbf{S}}\left\{\|\boldsymbol{f}\|_{0,\Omega}+\left(|\Omega|^{1/2}+\gamma\,r\right)\|\boldsymbol{g}\|_{\infty,\Omega}\right\}+c_{\widetilde{\mathbf{S}}}\,\kappa^{-1}\,U\left\{\alpha|\Omega|^{1/2}+r\right\}\leq r, \tag{1.56}$$

with $c_{\mathbf{S}}$ and $c_{\widetilde{\mathbf{S}}}$ as in (1.46) and (1.47), respectively. Then $\mathbf{T}(W_r)\subseteq W_r$.

Proof. Given $(\boldsymbol{w},\phi)\in W_r$, and so $\|\boldsymbol{w}\|_{1,\Omega}\leq r_0$, it follows from Lemma 1.1 that there exists a unique $\boldsymbol{u}=\mathbf{S}_4(\boldsymbol{w},\phi)\in\mathbf{H}^1_0(\Omega)$ solution to problem (1.31) and it satisfies the a priori estimate (1.46). In turn, if the data satisfies (1.56), we have that $\|\mathbf{S}_4(\boldsymbol{w},\phi)\|_{1,\Omega}\leq\widetilde{r}_0$, and according to Lemma 1.2 there exists

a unique $\varphi = \widetilde{\mathbf{S}}_2(\mathbf{S}_4(\boldsymbol{w},\phi),\phi) \in \widetilde{\mathrm{H}}^1(\Omega)$ solution to (1.33) with $\boldsymbol{w} := \mathbf{S}_4(\boldsymbol{w},\phi)$. As a consequence, from the definition of the operator $\mathbf{T}$ (cf. (1.34)), $\exists!\,(\boldsymbol{u},\varphi) = (\mathbf{S}_4(\boldsymbol{w},\phi),\widetilde{\mathbf{S}}_2(\mathbf{S}_4(\boldsymbol{w},\phi),\phi)) = \mathbf{T}(\boldsymbol{w},\phi)$ and from (1.46) and (1.47)

$$\begin{aligned}\|(\boldsymbol{u},\varphi)\| &\leq \|\mathbf{S}_4(\boldsymbol{w},\phi)\|_{1,\Omega} + \|\widetilde{\mathbf{S}}_2\big(\mathbf{S}_4(\boldsymbol{w},\phi),\phi\big)\|_{1,\Omega},\\ &\leq c_{\mathbf{S}}\left\{\|\boldsymbol{f}\|_{0,\Omega} + \big(|\Omega|^{1/2} + \gamma\,\|\phi\|_{0,\Omega}\big)\,\|\boldsymbol{g}\|_{\infty,\Omega}\right\} + c_{\widetilde{\mathbf{S}}}\,\kappa^{-1}\,U\left\{\alpha|\Omega|^{1/2} + \|\phi\|_{0,\Omega}\right\}.\end{aligned}$$

The results then follows using that $\|\phi\|_{0,\Omega} \leq r$ and the assumption on the data (1.56). ■

Next, we establish two lemmas that will be useful to derive conditions under which the operator $\mathbf{T}$ is continuous. To this end, in a similar way to [5, Section 3.3] and [26, Section 3.3], we introduce the following regularity hypotheses on the operator $\mathbf{S}$. From now on, we suppose that $\boldsymbol{f} \in \mathbf{H}^{\delta}(\Omega)$, for some $\delta \in (1/2,1)$ and that for each $(\boldsymbol{w},\phi) \in \mathbf{H}$ with $\|\boldsymbol{w}\|_{1,\Omega} \leq r$, $r > 0$ given, there holds

$$\mathbf{S}(\boldsymbol{w},\phi) \in \Big(\big(\mathbb{L}^2_{\mathrm{tr}}(\Omega)\cap\mathbb{H}^{\delta}(\Omega)\big)\times\big(\mathbb{H}_0(\mathbf{div};\Omega)\cap\mathbb{H}^{\delta}(\Omega)\big)\times\big(\mathbb{L}^2_{\mathrm{skew}}(\Omega)\cap\mathbb{H}^{\delta}(\Omega)\big)\Big)\times\big(\mathbf{H}^1_0\cap\mathbf{H}^{1+\delta}(\Omega)\big),$$

and

$$\begin{aligned}&\|\mathbf{S}_1(\boldsymbol{w},\phi)\|_{\delta,\Omega} + \|\mathbf{S}_2(\boldsymbol{w},\phi)\|_{\delta,\Omega} + \|\mathbf{S}_3(\boldsymbol{w},\phi)\|_{\delta,\Omega} + \|\mathbf{S}_4(\boldsymbol{w},\phi)\|_{1+\delta,\Omega}\\ &\qquad\leq \widehat{C}_{\mathbf{S}}\Big\{\|\boldsymbol{f}\|_{\delta,\Omega} + \Big(|\Omega|^{1/2} + \gamma\,\|\phi\|_{0,\Omega}\Big)\|\boldsymbol{g}\|_{\infty,\Omega}\Big\},\end{aligned} \tag{1.57}$$

where $\widehat{C}_{\mathbf{S}}$ is a constant independent of $(\boldsymbol{w},\phi)$. The aforementioned range for δ will become clear in the proof of Lemma 1.4 and Lemma 1.6 below, in which we will require to suitably control an expression involving the norm of $\boldsymbol{t} = \mathbf{S}_1(\boldsymbol{w},\phi)$ in some $\mathbb{L}^{2p}-$space by the respective norm in the $\mathbb{H}^{\delta}-$space, so that it then can be bounded by data using the a priori estimate (1.57).

Lemma 1.4. *Let $r \in (0,r_0)$, with r_0 given by (1.44). Then, for all $(\boldsymbol{w},\phi)$, $(\widetilde{\boldsymbol{w}},\widetilde{\phi}) \in \mathbf{H}$ such that $\|\boldsymbol{w}\|_{1,\Omega}$, $\|\widetilde{\boldsymbol{w}}\|_{1,\Omega} \leq r$, there exists a positive constant $C_{\mathbf{S}}$, depending on the parameters κ_2, κ_3, the constant $c_2(\Omega)$ (cf. (1.20)), the ellipticity constant $\alpha(\Omega)$ of the bilinear form $\mathbf{A}_{\phi}$ (cf. (1.39)) and δ (cf. (1.57)), such that*

$$\begin{aligned}\|\mathbf{S}(\boldsymbol{w},\phi) - \mathbf{S}(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| \leq C_{\mathbf{S}}\Big\{&L_{\mu}\,\|\mathbf{S}_1(\boldsymbol{w},\phi)\|_{\delta,\Omega}\|\phi-\widetilde{\phi}\|_{\mathrm{L}^{3/\delta}(\Omega)}\\ &+\|\mathbf{S}_4(\boldsymbol{w},\phi)\|_{1,\Omega}\|\boldsymbol{w}-\widetilde{\boldsymbol{w}}\|_{1,\Omega} + \gamma\,\|\boldsymbol{g}\|_{\infty,\Omega}\|\phi-\widetilde{\phi}\|_{0,\Omega}\Big\},\end{aligned} \tag{1.58}$$

where L_{μ} and γ are given by (1.2) and (1.1), respectively.

Proof. Given $(\boldsymbol{w},\phi)$, $(\widetilde{\boldsymbol{w}},\widetilde{\phi}) \in \mathbf{H}$ with $\|\boldsymbol{w}\|_{1,\Omega}$, $\|\widetilde{\boldsymbol{w}}\|_{1,\Omega} \leq r$, let $(\underline{\boldsymbol{t}},\boldsymbol{u}) := \mathbf{S}(\boldsymbol{w},\phi)$ and $(\widetilde{\underline{\boldsymbol{t}}},\widetilde{\boldsymbol{u}}) := \mathbf{S}(\widetilde{\boldsymbol{w}},\widetilde{\phi})$ be the corresponding solutions to the problem (1.31), respectively. Firstly, from the bilinearity of the forms $\mathbf{A}_{\phi}$ and $\mathbf{B}_{\boldsymbol{w}}$, it is observed that

$$\begin{aligned}\big(\mathbf{A}_{\widetilde{\phi}} + \mathbf{B}_{\widetilde{\boldsymbol{w}}}\big)((\underline{\boldsymbol{t}},\boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}},\widetilde{\boldsymbol{u}}),(\underline{\boldsymbol{r}},\boldsymbol{v})) &= -\big(\mathbf{A}_{\phi} - \mathbf{A}_{\widetilde{\phi}}\big)((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v}))\\ &\quad - \mathbf{B}_{\boldsymbol{w}-\widetilde{\boldsymbol{w}}}((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) + \Big(F_{\phi} - F_{\widetilde{\phi}}\Big)(\underline{\boldsymbol{r}},\boldsymbol{v}) \qquad \forall\,(\underline{\boldsymbol{r}},\boldsymbol{v}) \in \mathbb{H}\times\mathbf{H}^1_0(\Omega)\,,\end{aligned} \tag{1.59}$$

where we can notice that

$$\big(\mathbf{A}_{\phi} - \mathbf{A}_{\widetilde{\phi}}\big)((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{r}},\boldsymbol{v})) = \int_{\Omega}\Big\{\mu(\phi+\alpha) - \mu(\widetilde{\phi}+\alpha)\Big\}\,\boldsymbol{t} : \{\boldsymbol{r} - \kappa_3\boldsymbol{\tau}^{\mathrm{d}}\}\,, \tag{1.60}$$

and

$$\left(F_\phi - F_{\widetilde{\phi}}\right)(\underline{\boldsymbol{r}}, \boldsymbol{v}) \;=\; -\int_\Omega \gamma(\phi - \widetilde{\phi})\boldsymbol{g}\cdot\{\boldsymbol{v} - \kappa_2 \mathbf{div}\,\boldsymbol{\tau}\}\,. \tag{1.61}$$

Thus, using the ellipticity of $\mathbf{A}_{\widetilde{\phi}} + \mathbf{B}_{\widetilde{\boldsymbol{w}}}$ (cf. (1.43)) and then the identities (1.59), (1.60) and (1.61) with $(\underline{\boldsymbol{r}}, \boldsymbol{v}) = (\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})$, we find that

$$\begin{aligned}
&\frac{\alpha(\Omega)}{2}\|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\|^2 \leq \big(\mathbf{A}_{\widetilde{\phi}} + \mathbf{B}_{\widetilde{\boldsymbol{w}}}\big)\big((\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}}), (\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\big),\\
&= -\big(\mathbf{A}_{\phi} - \mathbf{A}_{\widetilde{\phi}}\big)\big((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\big) - \mathbf{B}_{\boldsymbol{w}-\widetilde{\boldsymbol{w}}}\big((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\big) + \big(F_\phi - F_{\widetilde{\phi}}\big)\big((\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\big),\\
&= -\int_\Omega \Big\{\mu(\phi+\alpha) - \mu(\widetilde{\phi}+\alpha)\Big\}\boldsymbol{t} : \Big\{(\boldsymbol{t} - \widetilde{\boldsymbol{t}}) - \kappa_3(\boldsymbol{\sigma}^{\mathrm{d}} - \widetilde{\boldsymbol{\sigma}}^{\mathrm{d}})\Big\} - \mathbf{B}_{\boldsymbol{w}-\widetilde{\boldsymbol{w}}}\big((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\big)\\
&\qquad - \int_\Omega \gamma(\phi - \widetilde{\phi})\boldsymbol{g}\cdot\{(\boldsymbol{u} - \widetilde{\boldsymbol{u}}) - \kappa_2 \mathbf{div}\,(\boldsymbol{\sigma} - \widetilde{\boldsymbol{\sigma}})\}\,.
\end{aligned}$$

Now, applying Cauchy-Schwarz and Hölder inequalities, the Lipschitz continuity of $\mu(\,\cdot\,)$ (cf. (1.2)), and the estimate (1.20), we obtain

$$\begin{aligned}
\frac{\alpha(\Omega)}{2}\|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\|^2 \leq \Big\{&L_\mu(1+\kappa_3^2)^{1/2}\|\boldsymbol{t}\|_{\mathbb{L}^{2p}(\Omega)}\|\phi - \widetilde{\phi}\|_{\mathrm{L}^{2q}(\Omega)}\\
&+ c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}\|_{1,\Omega}\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_{1,\Omega} + \gamma(1+\kappa_2^2)^{1/2}\|\boldsymbol{g}\|_{\infty,\Omega}\|\phi - \widetilde{\phi}\|_{0,\Omega}\Big\}\,\|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\widetilde{\underline{\boldsymbol{t}}}, \widetilde{\boldsymbol{u}})\|\,,
\end{aligned} \tag{1.62}$$

where $p, q \in [1, +\infty)$ are such that $1/p + 1/q = 1$. Next, according to the additional regularity assumed in (1.57), and recalling that the Sobolev embedding theorem (cf. [1, Theorem 4.12] or [80, Theorem 1.3.4]) establishes the continuous injection $i_\delta : \mathrm{H}^\delta(\Omega) \to \mathrm{L}^{\delta^*}(\Omega)$ with boundedness constant $C_\delta > 0$, where

$$\delta^* := \frac{6}{3 - 2\delta}\,,$$

we then take p such that $2p = \delta^*$ to deduce that, on the one hand, since $\boldsymbol{t} := \mathbf{S}_1(\boldsymbol{w}, \phi)$

$$\|\boldsymbol{t}\|_{\mathbb{L}^{2p}(\Omega)} = \|\mathbf{S}_1(\boldsymbol{w}, \phi)\|_{\mathbb{L}^{2p}(\Omega)} \leq C_\delta\|\mathbf{S}_1(\boldsymbol{w}, \phi)\|_{\delta,\Omega}\,, \tag{1.63}$$

and, on the other hand, the respective conjugate index q is given by

$$2q \;=\; \frac{2p}{p-1} \;=\; \frac{3}{\delta}.$$

Finally, inequalities (1.62) and (1.63) together with the previous identity give (1.58) with constant $C_{\mathbf{S}} := \dfrac{2}{\alpha(\Omega)}\max\Big\{C_\delta(1+\kappa_3^2)^{1/2},\, c_2(\Omega)(1+\kappa_3^2)^{1/2},\, (1+\kappa_2^2)^{1/2}\Big\}$. ■

In turn, the following result establishes the Lipschitz-continuity of the operator $\widetilde{\mathbf{S}}$.

Lemma 1.5. *Let $r \in (0, \widetilde{r}_0)$, with $\widetilde{r}_0$ given by (1.52). Then, for all $(\boldsymbol{w}, \phi), (\widetilde{\boldsymbol{w}}, \widetilde{\phi}) \in \mathbf{H}$ such that $\|\boldsymbol{w}\|_{1,\Omega}, \|\widetilde{\boldsymbol{w}}\|_{1,\Omega} \leq r$, there exists a positive constant $C_{\widetilde{\mathbf{S}}}$, depending on the parameter κ_5, the ellipticity constant $\widetilde{\alpha}(\Omega)$ of the bilinear form $\widetilde{\mathbf{A}}$ (cf. (1.50)) and the constant $c_1(\Omega)$ (cf. (1.19)), such that*

$$\|\widetilde{\mathbf{S}}(\boldsymbol{w}, \phi) - \widetilde{\mathbf{S}}(\widetilde{\boldsymbol{w}}, \widetilde{\phi})\| \;\leq\; \kappa^{-1}\, C_{\widetilde{\mathbf{S}}}\Big\{U\|\phi - \widetilde{\phi}\|_{0,\Omega} + \|\widetilde{\mathbf{S}}_2(\boldsymbol{w}, \phi)\|_{1,\Omega}\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_{1,\Omega}\Big\}\,, \tag{1.64}$$

where κ is given in (1.1).

Proof. Given r and $(\boldsymbol{w},\phi)$, $(\widetilde{\boldsymbol{w}},\widetilde{\phi}) \in \mathbf{H}$ as in the hypothesis, let us denote $(\boldsymbol{p},\varphi) := \widetilde{\mathbf{S}}(\boldsymbol{w},\phi)$ and $(\widetilde{\boldsymbol{p}},\widetilde{\varphi}) := \widetilde{\mathbf{S}}(\widetilde{\boldsymbol{w}},\widetilde{\phi})$, that is, the respective solutions to problem (1.33) in $\mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$. Thus, from the bilinearity of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}_{\boldsymbol{w}}$ for any $\boldsymbol{w}$, we have that

$$(\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{w}}})((\boldsymbol{p},\varphi)-(\widetilde{\boldsymbol{p}},\widetilde{\varphi}),(\boldsymbol{q},\psi)) = -\widetilde{\mathbf{B}}_{\boldsymbol{w}-\widetilde{\boldsymbol{w}}}((\boldsymbol{p},\varphi),(\boldsymbol{q},\psi)) + \big(\widetilde{F}_{\phi}-\widetilde{F}_{\widetilde{\phi}}\big)(\boldsymbol{q},\psi),$$

for all $(\boldsymbol{q},\psi)\in\mathbf{H}_{\Gamma}(\mathrm{div};\Omega)\times\widetilde{\mathrm{H}}^1(\Omega)$. Hence, using the ellipticity of $\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{w}}}$ (cf. (1.51)) and the continuity of $\widetilde{\mathbf{B}}_{\boldsymbol{w}}$ (cf. (1.48)) and the definition of $\widetilde{F}_{\phi}$ (cf. 1.29), we obtain

$$\begin{aligned}\frac{\widetilde{\alpha}(\Omega)}{2}\|(\boldsymbol{p},\varphi)-(\widetilde{\boldsymbol{p}},\widetilde{\varphi})\|^2 &\leq -\widetilde{\mathbf{B}}_{\boldsymbol{w}-\widetilde{\boldsymbol{w}}}((\boldsymbol{p},\varphi),(\boldsymbol{p},\varphi)-(\widetilde{\boldsymbol{p}},\widetilde{\varphi})) + \big(\widetilde{F}_{\phi}-\widetilde{F}_{\widetilde{\phi}}\big)((\boldsymbol{p},\varphi)-(\widetilde{\boldsymbol{p}},\widetilde{\varphi})),\\ &\leq \Big\{\kappa^{-1}(1+\kappa_5^2)^{1/2}\Big(U\|\phi-\widetilde{\phi}\|_{0,\Omega}+c_1(\Omega)\|\boldsymbol{w}-\widetilde{\boldsymbol{w}}\|_{1,\Omega}\|\varphi\|_{1,\Omega}\Big)\Big\}\,\|(\boldsymbol{p},\varphi)-(\widetilde{\boldsymbol{p}},\widetilde{\varphi})\|\,.\end{aligned}$$

The result then follows with $C_{\widetilde{\mathbf{S}}} := \dfrac{2}{\widetilde{\alpha}(\Omega)}(1+\kappa_5^2)^{1/2}\max\{1,c_1(\Omega)\}$ and recalling that $\varphi=\widetilde{\mathbf{S}}_2(\boldsymbol{w},\phi)$.

■

As a consequence of the previous lemmas, we have the following result.

Lemma 1.6. *Given $r\in(0,\min\{r_0,\widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (1.44) and (1.52), respectively, we let W_r be the closed ball in $\mathbf{H}$ defined in (1.55) and assume that the data satisfy (1.56). Then, there holds*

$$\begin{aligned}&\|\mathbf{T}(\boldsymbol{w},\phi)-\mathbf{T}(\widetilde{\boldsymbol{w}},\widetilde{\phi})\|\\ &\quad\leq (1+\kappa^{-1})(1+L_{\mu})\,C_{\mathbf{T}}\Big\{\|\boldsymbol{f}\|_{\delta,\Omega}+\|\boldsymbol{f}\|_{0,\Omega}+\Big(|\Omega|^{1/2}+\gamma\Big)\|\boldsymbol{g}\|_{\infty,\Omega}+U\Big\}\|(\boldsymbol{w},\phi)-(\widetilde{\boldsymbol{w}},\widetilde{\phi})\|\,,\end{aligned}\tag{1.65}$$

for all $(\boldsymbol{w},\phi)$, $(\widetilde{\boldsymbol{w}},\widetilde{\phi})\in W_r$, where $C_{\mathbf{T}}$ is a positive constant depending on r, the constants $c_{\mathbf{S}}$, $C_{\mathbf{S}}$, $C_{\widetilde{\mathbf{S}}}$ (cf. (1.46), (1.58), (1.64)) and δ (cf. (1.57)).

Proof. Given $r\in(0,\min\{r_0,\widetilde{r}_0\})$, and $(\boldsymbol{w},\phi)$, $(\widetilde{\boldsymbol{w}},\widetilde{\phi})\in W_r$, from the definition of $\mathbf{T}$ (cf. (1.34)), the Lipschitz-continuity of $\widetilde{\mathbf{S}}$ (cf. (1.64)) and the a priori estimate given for $\widetilde{\mathbf{S}}$ (1.47) we note that

$$\begin{aligned}\|\mathbf{T}(\boldsymbol{w},\phi)-\mathbf{T}(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| &\leq \|\mathbf{S}_4(\boldsymbol{w},\phi)-\mathbf{S}_4(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| + \|\widetilde{\mathbf{S}}_2(\mathbf{S}_4(\boldsymbol{w},\phi),\phi)-\widetilde{\mathbf{S}}_2(\mathbf{S}_4(\widetilde{\boldsymbol{w}},\widetilde{\phi}),\widetilde{\phi})\|\\ &\leq \|\mathbf{S}_4(\boldsymbol{w},\phi)-\mathbf{S}_4(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| + \kappa^{-1}C_{\widetilde{\mathbf{S}}}\Big\{U\|\phi-\widetilde{\phi}\|_{0,\Omega}+\|\widetilde{\mathbf{S}}_2(\mathbf{S}_4(\boldsymbol{w},\phi),\phi)\|_{1,\Omega}\|\mathbf{S}_4(\boldsymbol{w},\phi)-\mathbf{S}_4(\widetilde{\boldsymbol{w}},\widetilde{\phi})\|_{1,\Omega}\Big\}\\ &\leq \kappa^{-1}C_{\widetilde{\mathbf{S}}}\,U\,\|\phi-\widetilde{\phi}\|_{0,\Omega} + (1+\kappa^{-1}C_{\widetilde{\mathbf{S}}}\,r)\|\mathbf{S}_4(\boldsymbol{w},\phi)-\mathbf{S}_4(\widetilde{\boldsymbol{w}},\widetilde{\phi})\|_{1,\Omega}\,,\end{aligned}$$

where in the last inequality we have used that the data satisfy (1.56) and so $\|\widetilde{\mathbf{S}}_2(\mathbf{S}_4(\boldsymbol{w},\phi),\phi)\|_{1,\Omega}\leq r$. Next, using the Lipschitz-continuity of $\mathbf{S}$ (cf. (1.58)) and then applying the estimates (1.46) and (1.57),

and the fact that $\|\phi\|_{1,\Omega} \leq r$, we get

$$\begin{aligned}\|\mathbf{T}(\boldsymbol{w},\phi) - \mathbf{T}(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| &\leq \kappa^{-1} C_{\widetilde{\mathbf{S}}}\, U\, \|\phi - \widetilde{\phi}\|_{0,\Omega} + (1 + \kappa^{-1} C_{\widetilde{\mathbf{S}}}\, r) C_{\mathbf{S}} \Big\{ L_\mu\, \|\mathbf{S}_1(\boldsymbol{w},\phi)\|_{\delta,\Omega} \|\phi - \widetilde{\phi}\|_{\mathrm{L}^{3/\delta}(\Omega)} \\ &\quad + \|\mathbf{S}_4(\boldsymbol{w},\phi)\|_{1,\Omega} \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_{1,\Omega} + \gamma\, \|\boldsymbol{g}\|_{\infty,\Omega} \|\phi - \widetilde{\phi}\|_{0,\Omega} \Big\} \\ &\leq \kappa^{-1} C_{\widetilde{\mathbf{S}}}\, U\, \|\phi - \widetilde{\phi}\|_{0,\Omega} + (1 + \kappa^{-1} C_{\widetilde{\mathbf{S}}} r) C_{\mathbf{S}} \Big\{ L_\mu \widehat{C}_{\mathbf{S}} \widetilde{C}_\delta \Big[\|\boldsymbol{f}\|_{\delta,\Omega} + \Big(|\Omega|^{1/2} + \gamma\, r \Big) \|\boldsymbol{g}\|_{\infty,\Omega} \Big] \|\phi - \widetilde{\phi}\|_{1,\Omega} \\ &\quad + c_{\mathbf{S}} \Big[\|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\, r \Big) \|\boldsymbol{g}\|_{\infty,\Omega} \Big] \|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_{1,\Omega} + \gamma \|\boldsymbol{g}\|_{\infty,\Omega} \|\phi - \widetilde{\phi}\|_{0,\Omega} \Big\}\,,\end{aligned}$$

where the multiplicative constant $\widetilde{C}_\delta$, appearing in the second term of the last inequality, stands for the boundedness constant of the continuous injection of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^{3/\delta}(\Omega)$. In this way, with

$$C(r) := \big(1 + r C_{\widetilde{\mathbf{S}}}\big)(1 + r) C_{\mathbf{S}}\,, \qquad C_{\mathbf{T},1} := \max\{\widehat{C}_{\mathbf{S}} \widetilde{C}_\delta, c_{\mathbf{S}}\} \quad \text{and} \quad C_{\mathbf{T},2} := 3 \max\{\widehat{C}_{\mathbf{S}} \widetilde{C}_\delta, c_{\mathbf{S}}, 1\}\,,$$

after performing some algebraic manipulations, we find that

$$\begin{aligned}\|\mathbf{T}(\boldsymbol{w},\phi) - \mathbf{T}(\widetilde{\boldsymbol{w}},\widetilde{\phi})\| \;\leq\; (1 + \kappa^{-1})(1 + L_\mu) \Big\{ C(r) \Big[C_{\mathbf{T},1} \Big(\|\boldsymbol{f}\|_{\delta,\Omega} + \|\boldsymbol{f}\|_{0,\Omega} \Big) \\ + C_{\mathbf{T},2} \Big(|\Omega|^{1/2} + \gamma \Big) \|\boldsymbol{g}\|_{\infty,\Omega} \Big] + C_{\widetilde{\mathbf{S}}} U \Big\} \|(\boldsymbol{w},\phi) - (\widetilde{\boldsymbol{w}},\widetilde{\phi})\|,\end{aligned}$$

and so (1.65) follows with $C_{\mathbf{T}} := \max\{C(r)\, C_{\mathbf{T},1},\, C(r)\, C_{\mathbf{T},2},\, C_{\widetilde{\mathbf{S}}}\}$.

■

We are now in a position to establish sufficient conditions for the existence and uniqueness of a fixed-point for our problem (1.35) (equivalently, the well-posedness of the variational problem (1.23)). Indeed, we have from Lemmas 1.1 and 1.2 that $\mathbf{T}$ is well-defined in any ball W_r, with $r \in (0, \min\{r_0, \widetilde{r}_0\})$, and if the data satisfy (1.56) then $\mathbf{T}(W_r) \subseteq W_r$ (cf. Lemma 1.3). Furthermore, Lemma 1.6 guarantees that $\mathbf{T}$ is Lipschitz-continuous. So, if the data is small enough so that

$$(1 + \kappa^{-1})(1 + L_\mu)\, C_{\mathbf{T}} \Big\{ \|\boldsymbol{f}\|_{\delta,\Omega} + \|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma \Big) \|\boldsymbol{g}\|_{\infty,\Omega} + U \Big\} \;<\; 1\,, \tag{1.66}$$

then $\mathbf{T}$ becomes a contraction. Therefore, the Banach fixed-point Theorem provides the existence of a unique fixed-point of $\mathbf{T}$; that is, a unique solution to the problem (1.35), or equivalently, to the variational problem (1.23). We have then shown the main result of this section, and we state it as follows.

Theorem 1.1. *Let W_r be the closed ball in $\mathbf{H} = \mathbf{H}^1_0(\Omega) \times \widetilde{\mathrm{H}}^1(\Omega)$ defined in (1.55). Suppose that the parameters κ_i, $i \in \{1, \ldots, 6\}$, satisfy the conditions required by Lemmas 1.4 and 1.5, that the estimate (1.57) holds and the data satisfy (1.56) and (1.66). Then, the augmented fully-mixed problem (1.23) has unique solution $(\underline{\boldsymbol{t}}, \boldsymbol{u}, \boldsymbol{p}, \varphi) \in \mathbb{H} \times \mathbf{H}^1_0(\Omega) \times \mathbf{H}_\Gamma(\mathrm{div}; \Omega) \times \widetilde{\mathrm{H}}^1(\Omega)$, with $(\boldsymbol{u}, \varphi) \in W_r$. Moreover, the following a priori estimates hold*

$$\|(\underline{\boldsymbol{t}}, \boldsymbol{u})\| \leq c_{\mathbf{S}} \Big\{ \|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\, \|\varphi\|_{0,\Omega} \Big) \|\boldsymbol{g}\|_{\infty,\Omega} \Big\}\,,$$

and

$$\|(\boldsymbol{p}, \varphi)\| \leq c_{\widetilde{\mathbf{S}}}\, \kappa^{-1}\, U \Big\{ \alpha |\Omega|^{1/2} + \|\varphi\|_{0,\Omega} \Big\}.$$

with $c_{\mathbf{S}}$ and $c_{\widetilde{\mathbf{S}}}$ are given as in Lemmas 1.1 and 1.2, respectively.

We point out here that in practice micro-organisms are slightly denser than water and so the parameter $\gamma = \rho_0/\rho_m - 1$ is small. Then, the data restrictions (1.56) and (1.66) are equivalent to require the diffusion rate κ to be sufficiently large while the average velocity of upward swimming U and the physical domain Ω to be sufficiently small. Hence, Theorem 1.1 essentially states that our augmented fully-mixed formulation provides unique solutions to the Bioconvection problem for suspensions with viscous culture fluid, large diffusion rate, and slowly upswimming micro-organisms in small containers, similarly to the primal method for bioconvection proposed in [24].

1.4 The Galerkin Scheme

We here present and analyze the Galerkin scheme of the augmented fully-mixed formulation (1.23). In Section 1.4.1, after introducing the finite element spaces in which the discretization is based, we set the discrete problem and adapt the same strategy from Section 1.3.2 to equivalently write it as a fixed-point equation. The respective solvability analysis will be then address in Section 1.4.2 by adapting the results for the continuous case obtained in Sections 1.3.3 and 1.3.4.

1.4.1 The discrete framework

As usual, given a shape–regular triangulation $\mathcal{T}_h$ of $\overline{\Omega}$ made up of tetrahedra K of diameter h_K, we define the meshsize $h := \max\{h_K : K \in \mathcal{T}_h\}$. Furthermore, for any $k \geq 0$ and for each $K \in \mathcal{T}_h$, let $\mathrm{P}_k(K)$ (resp. $\widetilde{\mathrm{P}}_k(K)$) be the space of polynomial functions on K of degree $\leq k$ (resp. $= k$), and with the same notations from Section 1.1, we define the local Raviart-Thomas space of order k as

$$\mathbf{RT}_k(K) := \mathbf{P}_k(K) \oplus \widetilde{\mathrm{P}}_k(K)\mathbf{x}\,,$$

where $\mathbf{x}$ is a generic vector in R^3. Thus, we introduce the following finite element spaces for approximating $\boldsymbol{t}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$, respectively,

$$\mathbb{H}_h^{\boldsymbol{t}} := \Big\{ \boldsymbol{r}_h \in \mathbb{L}^2_{\mathrm{tr}}(\Omega) : \qquad \boldsymbol{r}_h|_K \in \mathbb{P}_k(K)\,, \quad \forall\, K \in \mathcal{T}_h \Big\},$$

$$\mathbb{H}_h^{\boldsymbol{\sigma}} := \Big\{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \qquad \mathbf{c}^{\mathrm{t}}\boldsymbol{\tau}_h\big|_K \in \mathbf{RT}_k(K)\,, \quad \forall\, \mathbf{c} \in \mathrm{R}^3\,, \quad \forall\, K \in \mathcal{T}_h \Big\},$$

$$\mathbb{H}_h^{\boldsymbol{\rho}} := \Big\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \qquad \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K)\,, \quad \forall\, K \in \mathcal{T}_h \Big\},$$

and for approximating $\boldsymbol{u}$, $\boldsymbol{p}$ and φ, respectively, we define

$$\mathbf{H}_h^{\boldsymbol{u}} := \Big\{ \boldsymbol{v}_h \in \mathbf{C}(\overline{\Omega}) : \qquad \boldsymbol{v}_h|_K \in \mathbf{P}_{k+1}(K)\,, \quad \forall\, K \in \mathcal{T}_h\,, \quad \boldsymbol{v}_h = 0 \text{ on } \Gamma \Big\},$$

$$\mathbf{H}_h^{\boldsymbol{p}} := \Big\{ \boldsymbol{q}_h \in \mathbf{H}_{\Gamma}(\mathrm{div};\Omega) : \qquad \boldsymbol{q}_h|_K \in \mathbf{RT}_k(K)\,, \quad \forall\, K \in \mathcal{T}_h \Big\},$$

$$\mathrm{H}_h^{\varphi} := \left\{ \psi_h \in \mathrm{C}(\overline{\Omega}) : \qquad \psi_h|_K \in \mathrm{P}_{k+1}(K)\,, \quad \forall\, K \in \mathcal{T}_h\,, \quad \text{and} \quad \int_\Omega \psi_h = 0 \right\}.$$

That is, trace-free and skew-symmetric tensors in the space $\mathbb{P}_k^{disc}$ (or simply $\mathbb{P}_0$ when $k = 0$) of discontinuous piecewise polynomials tensors of degree $\leq k$, are used for approximating the strain

tensor $\boldsymbol{t}$ and the vorticity $\boldsymbol{\rho}$, respectively, Raviart-Thomas elements of degree k for approximating the pseudo-stress $\boldsymbol{\sigma}$ and the pseudo-concentration gradient $\boldsymbol{p}$, whereas the components of the velocity $\boldsymbol{u}$ and the concentration φ are approximating by using the Lagrange space of piecewise polynomials of degree $k+1$ (with zero-mean value for φ).

Then, letting $\mathbb{H}_h := \mathbb{H}_h^{\boldsymbol{t}} \times \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$ and $\underline{\boldsymbol{t}}_h := (\boldsymbol{t}_h, \boldsymbol{\sigma}_h, \boldsymbol{\rho}_h)$, $\underline{\boldsymbol{r}}_h := (\boldsymbol{r}_h, \boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbb{H}_h$, the Galerkin scheme of (1.23) reads: Find $(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h, \boldsymbol{p}_h, \varphi_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$ such that

$$\begin{array}{rcll} \mathbf{A}_{\varphi_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) + \mathbf{B}_{\boldsymbol{u}_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) & = & F_{\varphi_h}(\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) & \forall\, (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}, \\ \widetilde{\mathbf{A}}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) + \widetilde{\mathbf{B}}_{\boldsymbol{u}_h}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) & = & \widetilde{F}_{\varphi_h}(\boldsymbol{q}_h, \psi_h) & \forall\, (\boldsymbol{q}_h, \psi_h) \in \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}. \end{array} \tag{1.67}$$

Similarly to the continuous case, we now rewrite (1.67) as a fixed-point problem in terms of operators arising by decoupling the system. Indeed, adapting the approach from Section (1.3.2), we firstly define $\mathbf{H}_h := \mathbf{H}_h^{\boldsymbol{u}} \times \mathrm{H}_h^{\varphi}$ and introduce the operator $\mathbf{S}_h : \mathbf{H}_h \longrightarrow \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}$ as

$$\mathbf{S}_h(\boldsymbol{w}_h, \phi_h) := \Big(\big(\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h), \mathbf{S}_{2,h}(\boldsymbol{w}_h, \phi_h), \mathbf{S}_{3,h}(\boldsymbol{w}_h, \phi_h) \big), \mathbf{S}_{4,h}(\boldsymbol{w}_h, \phi_h) \Big) = (\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h),$$

for all $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$, where, for $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$ given, $(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h)$ is the unique solution to the discrete version of the problem (1.31), namely: Find $(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}$ such that

$$\mathbf{A}_{\phi_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) + \mathbf{B}_{\boldsymbol{w}_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) = F_{\phi_h}(\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \quad \forall\, (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}, \tag{1.68}$$

where the bilinear forms $\mathbf{A}_{\phi_h}$ (with ϕ_h in place of ϕ) and $\mathbf{B}_{\boldsymbol{w}_h}$ (with $\boldsymbol{w}_h$ in place of $\boldsymbol{w}$), and the functional F_{ϕ_h} (with ϕ_h instead of ϕ) are defined as in (1.24), (1.25) and (1.28), respectively. Secondly, we define the operator $\widetilde{\mathbf{S}}_h : \mathbf{H}_h \longrightarrow \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$ as

$$\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h) := \big(\widetilde{\mathbf{S}}_{1,h}(\boldsymbol{w}_h, \phi_h), \widetilde{\mathbf{S}}_{2,h}(\boldsymbol{w}_h, \phi_h) \big) = (\boldsymbol{p}_h, \varphi_h) \quad \forall\, (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h,$$

where, for $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$ given, $(\boldsymbol{p}_h, \varphi_h)$ stands for the unique solution to the discrete version of problem (1.33), that is: Find $(\boldsymbol{p}_h, \varphi_h) \in \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$ such that

$$\widetilde{\mathbf{A}}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) + \widetilde{\mathbf{B}}_{\boldsymbol{w}_h}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) = \widetilde{F}_{\phi_h}(\boldsymbol{q}_h, \psi_h) \quad \forall\, (\boldsymbol{q}_h, \psi_h) \in \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}, \tag{1.69}$$

where the bilinear forms $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}_{\boldsymbol{w}_h}$ (with $\boldsymbol{w}_h$ in place of $\boldsymbol{w}$), and the functional $\widetilde{F}_{\phi_h}$ (with ϕ_h instead of ϕ) are defined as in (1.26), (1.27), and (1.29), respectively. Hence, by introducing the operator $\mathbf{T}_h : \mathbf{H}_h \longrightarrow \mathbf{H}_h$ as

$$\mathbf{T}_h(\boldsymbol{w}_h, \phi_h) := \Big(\mathbf{S}_{4,h}(\boldsymbol{w}_h, \phi_h), \widetilde{\mathbf{S}}_{2,h}\big(\mathbf{S}_{4,h}(\boldsymbol{w}_h, \phi_h), \phi_h \big) \Big) \quad \forall (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h\,,$$

we realize that solving (1.67) is equivalent to seeking for a fixed-point of the operator $\mathbf{T}_h$, that is: Find $(\boldsymbol{u}_h, \varphi_h) \in \mathbf{H}_h$ such that

$$\mathbf{T}_h(\boldsymbol{u}_h, \varphi_h) = (\boldsymbol{u}_h, \varphi_h). \tag{1.70}$$

1.4.2 Solvability analysis

Here we study the solvability of the fixed-point equation (1.70) by adapting the analysis from Sections 1.3.3 and 1.3.4. We remark in advance that most of the proofs are almost verbatim from

the analogues results at continuous level, and hence we omit the details in those cases. To begin with, using the same arguments from Lemmas 1.1 and 1.2, we firstly state conditions under which the discrete problems (1.68) and (1.69) are well-posed, and therefore the operators $\mathbf{S}_h$ and $\widetilde{\mathbf{S}}_h$ are well-defined.

Lemma 1.7. *Suppose that the parameters κ_i, $i \in \{1,\ldots,4\}$, satisfy the conditions required by Lemma 1.1. Then, for each $r \in (0, r_0)$, with r_0 given by (1.44), and for each $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$ such that $\|\boldsymbol{w}_h\|_{1,\Omega} \leq r$, the problem (1.68) has a unique solution $(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h) = \mathbf{S}_h(\boldsymbol{w}_h, \phi_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}$. Moreover, with the same constant $c_{\mathbf{S}} > 0$ from (1.46), which is independent of $(\boldsymbol{w}_h, \phi_h)$, there holds*

$$\|\mathbf{S}_h(\boldsymbol{w}_h, \phi_h)\| = \|(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h)\| \leq c_{\mathbf{S}} \left\{ \|\boldsymbol{f}\|_{0,\Omega} + \left(|\Omega|^{1/2} + \gamma \|\phi_h\|_{0,\Omega} \right) \|\boldsymbol{g}\|_{\infty,\Omega} \right\}.$$

Lemma 1.8. *Suppose that the parameters κ_i, $i \in \{5, 6\}$, satisfy the conditions required by Lemma 1.2. Then, for each $\widetilde{r} \in (0, \widetilde{r}_0)$, $\widetilde{r}_0$ given by (1.52), and for each $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$ such that $\|\boldsymbol{w}_h\|_{1,\Omega} \leq \widetilde{r}$, the problem (1.69) has a unique solution $(\boldsymbol{p}_h, \varphi_h) \in \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$. Moreover, with the same constant $c_{\widetilde{\mathbf{S}}} > 0$ from (1.47), which is independent of $(\boldsymbol{w}_h, \phi_h)$, there holds*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h)\| = \|(\boldsymbol{p}_h, \varphi_h)\| \leq c_{\widetilde{\mathbf{S}}}\, \kappa^{-1}\, U \left\{ \alpha |\Omega|^{1/2} + \|\phi_h\|_{0,\Omega} \right\}.$$

Now we state the solvability of the fixed-point equation (1.70) by verifying the hypotheses of the Brouwer fixed-point Theorem (cf. [30, Theorem 9.9-2]). On the one hand, as a straightforward combination of Lemmas 1.7 and 1.8, we begin by establishing the discrete version of Lemma 1.3.

Lemma 1.9. *Given $r \in (0, \min\{r_0, \widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (1.44) and (1.52), respectively, we let $W_{r,h}$ be the closed ball in $\mathbf{H}_h$ defined by*

$$W_{r,h} := \left\{ (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h : \quad \|(\boldsymbol{w}_h, \phi_h)\| \leq r \right\}, \tag{1.71}$$

and assume that the data satisfy (1.56). Then $\mathbf{T}_h(W_{r,h}) \subseteq W_{r,h}$.

On the other hand, we focus now on the Lipschitz continuity of the operators $\mathbf{S}_h$ and $\widetilde{\mathbf{S}}_h$. Regarding $\mathbf{S}_h$, the discrete version of Lemma 1.4 is provided next. Here, we particularly notice in advance that the additional regularity assumption (1.57) employed there to suitably bound $\boldsymbol{t}$ in the $\mathbb{L}^{2p}-$norm by some $\mathbb{H}^{\delta}-$norm can not be applied at the present discrete context to bound $\boldsymbol{t}_h$. On the contrary, we will utilize a $\mathrm{L}^4 - \mathrm{L}^4 - \mathrm{L}^2$ argument (Hölder inequality) to bound the respective term in which it is involved and then make use of the fact that $\boldsymbol{t}_h \in \mathbb{H}_h^{\boldsymbol{t}}$, and so their components are piecewise polynomials (see at the beginning of Section 1.4.1).

Lemma 1.10. *Let $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in \mathbf{H}_h$ such that $\|\boldsymbol{w}_h\|_{1,\Omega}$, $\|\widetilde{\boldsymbol{w}}_h\|_{1,\Omega} \leq r$, for any $r \in (0, r_0)$ with r_0 given by (1.44). Then, there exists a positive constant $C_{\mathbf{S}_h}$, depending on κ_2, κ_3, $c_2(\Omega)$, and $\alpha(\Omega)$, but independent of h, such that*

$$\begin{aligned} \|\mathbf{S}_h(\boldsymbol{w}_h, \phi_h) - \mathbf{S}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \leq C_{\mathbf{S}_h} \Big\{ & L_\mu \, \|\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h)\|_{\mathbb{L}^4(\Omega)} \|\phi_h - \widetilde{\phi}_h\|_{\mathrm{L}^4(\Omega)} \\ & + \|\mathbf{S}_{4,h}(\boldsymbol{w}_h, \phi_h)\|_{1,\Omega} \|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega} + \gamma \, \|\boldsymbol{g}\|_{\infty,\Omega} \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} \Big\}. \end{aligned} \tag{1.72}$$

Proof. The proof is almost verbatim to that one of Lemma 1.4. Indeed, it suffices to see that when applying the Hölder inequality with $p = q = 2$, the estimate (1.62) becomes

$$\begin{aligned}\frac{\alpha(\Omega)}{2}\|(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h) - (\widetilde{\underline{\boldsymbol{t}}}_h, \widetilde{\boldsymbol{u}}_h)\| \leq \Big\{ & L_\mu (1+\kappa_3^2)^{1/2} \|\boldsymbol{t}_h\|_{\mathbb{L}^4(\Omega)} \|\phi_h - \widetilde{\phi}_h\|_{\mathrm{L}^4(\Omega)} \\ & + c_2(\Omega)(1+\kappa_3^2)^{1/2} \|\boldsymbol{u}_h\|_{1,\Omega} \|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega} + \gamma (1+\kappa_2^2)^{1/2} \|\boldsymbol{g}\|_{\infty,\Omega} \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} \Big\} .\end{aligned} \tag{1.73}$$

Since elements of $\mathbb{H}_h^{\boldsymbol{t}}$ are piecewise polynomials by components we have that $\|\boldsymbol{t}_h\|_{\mathbb{L}^4(\Omega)} < +\infty$, and using the fact that $\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h) = \boldsymbol{t}_h$, the inequality (1.73) immediately yields the estimate (1.72) with $C_{\mathbf{S}_h} := \dfrac{2}{\alpha(\Omega)} \max\Big\{(1+\kappa_3^2)^{1/2},\, c_2(\Omega)(1+\kappa_3^2)^{1/2},\, (1+\kappa_2^2)^{1/2}\Big\}$, which is clearly independent of h. ∎

Following the same arguments used in the proof Lemma 1.5, we directly have the following result regarding the operator $\widetilde{\mathbf{S}}_h$.

Lemma 1.11. *Let $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in \mathbf{H}_h$ such that $\|\boldsymbol{w}_h\|_{1,\Omega}$, $\|\widetilde{\boldsymbol{w}}_h\|_{1,\Omega} \leq \widetilde{r}$, for any $r \in (0, \widetilde{r}_0)$, with r_0 given by (1.52). Then, with the same constant $C_{\widetilde{\mathbf{S}}}$ provided by Lemma 1.5, there holds*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h) - \widetilde{\mathbf{S}}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \leq \kappa^{-1} C_{\widetilde{\mathbf{S}}} \Big\{ U \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} + \|\widetilde{\mathbf{S}}_{2,h}(\boldsymbol{w}_h, \phi_h)\|_{1,\Omega} \|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega} \Big\} . \tag{1.74}$$

As a result of the previous two lemmas, we can state the Lipschitz-continuity of the operator $\mathbf{T}_h$, which constitutes the discrete version of Lemma 1.6.

Lemma 1.12. *Let $r \in (0, \min\{r_0, \widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (1.44) and (1.52), respectively, let $W_{r,h}$ be the closed ball in $\mathbf{H}_h$ defined in (1.71) and assume that the data satisfy (1.56). Then, there exists a constant $C_{\mathbf{T}_h} > 0$, that depends on r and other constants but is independent of h, such that*

$$\begin{aligned}\|\mathbf{T}_h(\boldsymbol{w}_h, \phi_h) - \mathbf{T}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \leq (1+\kappa^{-1})(1+L_\mu) C_{\mathbf{T}_h} \Big\{ & \|\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h)\|_{\mathbb{L}^4(\Omega)} \\ & + \|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\Big) \|\boldsymbol{g}\|_{\infty,\Omega} + U \Big\} \|(\boldsymbol{w}, \phi) - (\widetilde{\boldsymbol{w}}, \widetilde{\phi})\| ,\end{aligned} \tag{1.75}$$

for all $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in W_{r,h}$.

Proof. Proceeding as in the proof of Lemma 1.6, but using (1.72) and (1.74) instead (1.58) and (1.64), respectively, the continuous injection of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^4(\Omega)$ with constant $\widetilde{C}$, and then the a priori estimate provided by Lemma 1.7, we find that

$$\begin{aligned}&\|\mathbf{T}_h(\boldsymbol{w}_h, \phi_h) - \mathbf{T}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \\ &\quad \leq \kappa^{-1} C_{\widetilde{\mathbf{S}}}\, U \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} + (1 + \kappa^{-1} C_{\widetilde{\mathbf{S}}}\, r) C_{\mathbf{S}_h} \Big\{ L_\mu \widetilde{C} \|\mathbf{S}_1(\boldsymbol{w}_h, \phi_h)\|_{\mathbb{L}^4(\Omega)} \|\phi_h - \widetilde{\phi}_h\|_{1,\Omega} \\ &\qquad + c_{\mathbf{S}} \Big[\|\boldsymbol{f}\|_{0,\Omega} + \Big(|\Omega|^{1/2} + \gamma\, r\Big) \|\boldsymbol{g}\|_{\infty,\Omega} \Big] \|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega} + \gamma \|\boldsymbol{g}\|_{\infty,\Omega} \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} \Big\} .\end{aligned}$$

Then, after performing algebraic manipulations and defining

$$\widetilde{C}(r) := \big(1 + r C_{\widetilde{\mathbf{S}}}\big)(1+r) C_{\mathbf{S}_h} , \qquad \widetilde{C}_{\mathbf{T},1} := \max\{\widetilde{C}, c_{\mathbf{S}}\} \quad \text{and} \quad \widetilde{C}_{\mathbf{T},2} := 2\max\{c_{\mathbf{S}}, 1\} ,$$

the results follows with $C_{\mathbf{T}_h} := \max\{\widetilde{C}(r)\, C_{\mathbf{T},1},\, \widetilde{C}(r)\, C_{\mathbf{T},2},\, C_{\widetilde{\mathbf{S}}}\}$, which is independent of h because so the constants $\widetilde{C}(r)$, $C_{\mathbf{T},1}$, $C_{\mathbf{T},2}$ and $C_{\widetilde{\mathbf{S}}}$ are. ∎

The previous lemma provides the continuity required by the Brouwer fixed-point theorem, in the convex and compact set $W_{r,h} \subset \mathbf{H}_h$. Therefore, we have essentially proved the following result.

Theorem 1.2. *Suppose that the parameters κ_i, $i \in \{1,\ldots,6\}$, satisfy the conditions required by Lemmas 1.4 and 1.5. Let $W_{r,h}$ be the closed ball in $\mathbf{H}_h$ defined in (1.71) and assume that the data satisfy (1.56). Then, the Galerkin scheme (1.67) has at least one solution $(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h, \boldsymbol{p}_h, \varphi_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$, with $(\boldsymbol{u}_h, \varphi_h) \in W_{r,h}$, and the following a priori estimates hold*

$$\|(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h)\| \leq c_{\mathbf{S}} \left\{ \|\boldsymbol{f}\|_{0,\Omega} + \left(|\Omega|^{1/2} + \gamma \|\varphi_h\|_{0,\Omega} \right) \|\boldsymbol{g}\|_{\infty,\Omega} \right\},$$

and

$$\|(\boldsymbol{p}_h, \varphi_h)\| \leq c_{\widetilde{\mathbf{S}}}\, \kappa^{-1}\, U \left\{ \alpha |\Omega|^{1/2} + \|\varphi_h\|_{0,\Omega} \right\},$$

with $c_{\mathbf{S}}$ and $c_{\widetilde{\mathbf{S}}}$ as in Lemmas 1.1 and 1.2, respectively.

We end this section by remarking that the lack of suitable estimates for $\|\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h)\|_{\mathbb{L}^4(\Omega)}$ (similar to [26, Section 4.2]) stops us of trying to use (1.75) to derive a condition on data so that $\mathbf{T}_h$ becomes a contraction. This is the reason why in the previous theorem we can only guarantee the existence of a discrete solution. In turn, as we commented after Theorem 1.1 for the continuous case, the previous result states that our augmented fully-mixed scheme provides existence of discrete solutions to the bioconvection problem whenever the data satisfy the condition (1.56), that is, for suspensions with viscous culture fluid, large diffusion rate, and slowly upswimming micro-organisms in small containers, similarly to the classical finite element method for bioconvection that was constructed in [24].

1.5 A priori error analysis

In this section, we undertake the error analysis for the Galerkin scheme (1.67) associated to the problem (1.23). To that end, we will deduce the corresponding Céa estimate as well as the respective theoretical convergence rates according to the approximation properties of the discrete spaces introduced in Section 1.4.1. To begin with, we let

$$(\underline{\boldsymbol{t}}, \boldsymbol{u}, \boldsymbol{p}, \varphi) \in \mathbb{H} \times \mathbf{H}_0^1(\Omega) \times \mathbf{H}_\Gamma(\mathrm{div};\Omega) \times \widetilde{\mathrm{H}}^1(\Omega) \quad \text{with} \quad (\boldsymbol{u}, \varphi) \in W_r\,,$$

and

$$(\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h, \boldsymbol{p}_h, \varphi_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi} \quad \text{with} \quad (\boldsymbol{u}_h, \varphi_h) \in W_{r,h}\,,$$

be solutions to the problems (1.23) and (1.67), respectively. Therefore, we have that

$$\begin{aligned} \mathbf{A}_\varphi((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{r}}, \boldsymbol{v})) + \mathbf{B}_{\boldsymbol{u}}((\underline{\boldsymbol{t}}, \boldsymbol{u}), (\underline{\boldsymbol{r}}, \boldsymbol{v})) &= F_\varphi(\underline{\boldsymbol{r}}, \boldsymbol{v}) \quad \forall\, (\underline{\boldsymbol{r}}, \boldsymbol{v}) \in \mathbb{H} \times \mathbf{H}_0^1(\Omega), \\ \mathbf{A}_{\varphi_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) + \mathbf{B}_{\boldsymbol{u}_h}((\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h), (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)) &= F_{\varphi_h}(\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \quad \forall\, (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}, \end{aligned} \tag{1.76}$$

and

$$\begin{aligned} \widetilde{\mathbf{A}}((\boldsymbol{p}, \varphi), (\boldsymbol{q}, \psi)) + \widetilde{\mathbf{B}}_{\boldsymbol{u}}((\boldsymbol{p}, \varphi), (\boldsymbol{q}, \psi)) &= \widetilde{F}_\varphi(\boldsymbol{q}, \psi) \quad \forall\, (\boldsymbol{q}, \psi) \in \mathbf{H}_\Gamma(\mathrm{div};\Omega) \times \widetilde{\mathrm{H}}^1(\Omega), \\ \widetilde{\mathbf{A}}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) + \widetilde{\mathbf{B}}_{\boldsymbol{u}_h}((\boldsymbol{p}_h, \varphi_h), (\boldsymbol{q}_h, \psi_h)) &= \widetilde{F}_{\varphi_h}(\boldsymbol{q}_h, \psi_h) \quad \forall\, (\boldsymbol{q}_h, \psi_h) \in \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}. \end{aligned} \tag{1.77}$$

Because of the structure of the systems (1.76) and (1.77), in what follows we apply the well-known Strang lemma for elliptic variational problems (see [83, Theorem 11.1]) in order to derive an upper bound for the total error $\|(\underline{\boldsymbol{t}}, \boldsymbol{u}, \boldsymbol{p}, \varphi) - (\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h, \boldsymbol{p}_h, \varphi_h)\|$. We recall this auxiliary result as follows.

Lemma 1.13. *Let V be a Hilbert space, $F \in V'$, and $A : V \times V \to \mathbb{R}$ be a bounded and V-elliptic bilinear form. In addition, let $\{V_h\}_{h>0}$ be a sequence of finite dimensional subspaces of V, and for each $h > 0$ consider a bounded bilinear form $A : V_h \times V_h \to \mathbb{R}$ and a functional $F_h \in V_h'$. Assume that the family $\{A_h\}_{h>0}$ is uniformly elliptic, that is, there exists a constant $\widetilde{\alpha} > 0$, independent of h, such that*

$$A_h(v_h,\, v_h) \,\geq\, \widetilde{\alpha}\, \|v_h\|_V^2\,, \quad \forall\, v_h \in V_h\,, \quad \forall\, h > 0.$$

In turn, let $u \in V$ and $u_h \in V_h$ such that

$$A(u,\, v) = F(v)\,, \quad \forall\, v \in V\,, \quad \textit{and} \quad A_h(u_h,\, v_h) = F_h(v_h)\,, \quad \forall\, v_h \in V_h.$$

Then, for each $h > 0$ there holds

$$\|u - u_h\|_V \;\leq C_{ST} \left\{ \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{|F(w_h) - F_h(w_h)|}{\|w_h\|_V} \right.$$
$$\left. + \inf_{\substack{v_h \in V_h \\ v_h \neq 0}} \left(\|u - v_h\|_V + \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{|A(v_h,\, w_h) - A_h(v_h,\, w_h)|}{\|w_h\|_V} \right) \right\},$$

where $C_{ST} := \widetilde{\alpha}^{-1} \max\{1,\, \|A\|\}$.

In that follows, we denote as usual

$$\mathrm{dist}\big((\underline{\boldsymbol{t}}, \boldsymbol{u}),\, \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}\big) := \inf_{(\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h) \,\in\, \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}} \|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\underline{\boldsymbol{r}}_h, \boldsymbol{v}_h)\|\,,$$

and

$$\mathrm{dist}\big((\boldsymbol{p}, \varphi),\, \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}\big) := \inf_{(\boldsymbol{q}_h, \psi_h) \,\in\, \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}} \|(\boldsymbol{p}, \varphi) - (\boldsymbol{q}_h, \psi_h)\|\,.$$

The following lemma provides a preliminary estimate for the error $\|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h)\|$ associated to the system (1.76).

Lemma 1.14. *Let $C_{ST} := \dfrac{2}{\alpha(\Omega)} \max\{1, \|\mathbf{A}_\varphi + \mathbf{B}_{\boldsymbol{u}}\|\}$, where $\alpha(\Omega)$ is the constant yielding the ellipticity of $\mathbf{A}_\varphi + \mathbf{B}_{\boldsymbol{u}}$ (cf. (1.43) and Lemma 1.1). Then, there holds*

$$\|(\underline{\boldsymbol{t}}, \boldsymbol{u}) - (\underline{\boldsymbol{t}}_h, \boldsymbol{u}_h)\| \;\leq\; C_{ST} \Big\{ \Big(1 + 2\|\mathbf{A}_\varphi\| + c_2(\Omega)(1 + \kappa_3^2)^{1/2} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} \Big)\, \mathrm{dist}\big((\underline{\boldsymbol{t}}, \boldsymbol{u}),\, \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}\big)$$
$$+ \gamma(1 + \kappa_2^2)^{1/2} \|\boldsymbol{g}\|_{\infty,\Omega} \|\varphi - \varphi_h\|_{0,\Omega} \;+\; L_\mu C_\delta (1 + \kappa_3^2)^{1/2} \|\boldsymbol{t}\|_{\delta,\Omega} \|\varphi - \varphi_h\|_{\mathrm{L}^{3/\delta}(\Omega)} \tag{1.78}$$
$$+ c_2(\Omega)(1 + \kappa_3^2)^{1/2} \|\boldsymbol{u}\|_{1,\Omega} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} \Big\}\,.$$

Proof. Since $(\boldsymbol{u},\varphi)\in W_r$ and $(\boldsymbol{u}_h,\varphi_h)\in W_{r,h}$, Lemma 1.1 and Lemma 1.7 guarantee that the bilinear forms $\mathbf{A}_\varphi+\mathbf{B}_{\boldsymbol{u}}$ and $\mathbf{A}_{\varphi_h}+\mathbf{B}_{\boldsymbol{u}_h}$ are $\mathbb{H}\times\mathbf{H}^1_0(\Omega)-$elliptic and $\mathbb{H}_h\times\mathbf{H}^{\boldsymbol{u}}_h-$elliptic $(\forall\, h>0)$, respectively, with the same constant $\frac{\alpha(\Omega)}{2}$ (see (1.43)). Also, F_φ and F_{φ_h} are clearly both linear and bounded functionals. Therefore the system (1.76) satisfies the hypotheses of Strang's lemma and thus, a direct application of the Lemma 1.13 to the specific context (1.76) with

$$A:=\mathbf{A}_\varphi+\mathbf{B}_{\boldsymbol{u}}\,,\quad \{A_h\}_{h>0}:=\{\mathbf{A}_{\varphi_h}+\mathbf{B}_{\boldsymbol{u}_h}\}_{h>0}\,,\quad F:=F_\varphi\,,\quad \text{and}\quad F_h:=F_{\varphi_h}\,,$$

yields

$$\begin{aligned}\|(\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h)\| \;\leq\; C_{ST}\left\{\left\|\big(F_\varphi-F_{\varphi_h}\big)\big|_{\mathbb{H}_h\times\mathbf{H}^{\boldsymbol{u}}_h}\right\|+\inf_{\substack{(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\in\mathbb{H}_h\times\mathbf{H}^{\boldsymbol{u}}_h\\(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\neq\mathbf{0}}}\left(\|(\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\|\right.\right.\\ \left.\left.+\sup_{\substack{(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\in\mathbb{H}_h\times\mathbf{H}^{\boldsymbol{u}}_h\\(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\neq\mathbf{0}}}\frac{\big|(\mathbf{A}_\varphi-\mathbf{A}_{\varphi_h})((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))+\mathbf{B}_{\boldsymbol{u}-\boldsymbol{u}_h}((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))\big|}{\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|}\right)\right\},\end{aligned}\tag{1.79}$$

where $C_{ST}:=\dfrac{2}{\alpha(\Omega)}\max\{1,\|\mathbf{A}_\varphi+\mathbf{B}_{\boldsymbol{u}}\|\}$. Now, from the estimate (1.61), observe that the first term at the right-hand side of (1.79) can be bounded as

$$\left\|\big(F_\varphi-F_{\varphi_h}\big)\big|_{\mathbb{H}_h\times\mathbf{H}^{\boldsymbol{u}}_h}\right\|\;\leq\;\gamma\,(1+\kappa_2^2)^{1/2}\,\|\boldsymbol{g}\|_{\infty,\Omega}\|\varphi-\varphi_h\|_{0,\Omega}\,.\tag{1.80}$$

To estimate the supremum in (1.79), on the one hand, we first conveniently add and subtract $(\underline{\boldsymbol{t}},\boldsymbol{u})$ in the first component of the bilinear form $\mathbf{A}_\varphi-\mathbf{A}_{\varphi_h}$ to find

$$\begin{aligned}(\mathbf{A}_\varphi-\mathbf{A}_{\varphi_h})((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)) \;&=\; \mathbf{A}_\varphi((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)-(\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))\\ &+(\mathbf{A}_\varphi-\mathbf{A}_{\varphi_h})((\underline{\boldsymbol{t}},\boldsymbol{u}),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))+\mathbf{A}_{\varphi_h}((\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))\,.\end{aligned}\tag{1.81}$$

Now, we apply (1.36) to estimate the first and third terms at the right-hand side of (1.81), whereas the the second term is estimated by proceeding similarly to the derivation of (1.62) combined with (1.63), which gives

$$\begin{aligned}\big|(\mathbf{A}_\varphi-\mathbf{A}_{\varphi_h})((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))\big| \;&\leq\; 2\|\mathbf{A}_\varphi\|\,\|(\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\|\,\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|\\ &+L_\mu C_\delta(1+\kappa_3^2)^{1/2}\|\boldsymbol{t}\|_{\delta,\Omega}\|\varphi-\varphi_h\|_{\mathrm{L}^{3/\delta}(\Omega)}\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|.\end{aligned}$$

On the other hand, for estimating the term that involves $\mathbf{B}_{\boldsymbol{u}-\boldsymbol{u}_h}$, we apply (1.37) with $\boldsymbol{w}=\boldsymbol{u}-\boldsymbol{u}_h$,

$$\begin{aligned}\big|\mathbf{B}_{\boldsymbol{u}-\boldsymbol{u}_h}((\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h),(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h))\big| \;&\leq\; c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\|\boldsymbol{v}_h\|_{1,\Omega}\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|\\ &\leq\; c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\|(\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\|\,\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|\\ &+\;c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\|\boldsymbol{u}\|_{1,\Omega}\|(\underline{\boldsymbol{s}}_h,\boldsymbol{z}_h)\|\,,\end{aligned}\tag{1.82}$$

where the last inequality arises after adding and subtracting $\boldsymbol{u}$ in the term $\|\boldsymbol{v}_h\|_{1,\Omega}$, using triangle inequality and then bounding $\|\boldsymbol{u}-\boldsymbol{v}_h\|_{1,\Omega}$ by $\|(\underline{\boldsymbol{t}},\boldsymbol{u})-(\underline{\boldsymbol{r}}_h,\boldsymbol{v}_h)\|$. Finally, by replacing (1.80), (1.81) and (1.82) back into (1.79), we get (1.78). ∎

Concerning the error $\|(\boldsymbol{p},\varphi)-(\boldsymbol{p}_h,\varphi_h)\|$ associated to the concentration equations (1.77), we have the following result.

Lemma 1.15. *Let* $\widetilde{C}_{ST}:=\dfrac{2}{\widetilde{\alpha}(\Omega)}\max\{1,\|\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{u}}\|\}$*, where* $\widetilde{\alpha}(\Omega)$ *is the constant yielding the ellipticity of* $\widetilde{\mathbf{A}}+\widetilde{\mathbf{B}}_{\boldsymbol{u}}$ *(cf. (1.51) and Lemma 1.2). Then, there holds*

$$\begin{aligned}\|(\boldsymbol{p},\varphi)-(\boldsymbol{p}_h,\varphi_h)\| \,\leq\, \widetilde{C}_{ST}\Big\{&\Big(1+\kappa^{-1}c_1(\Omega)(1+\kappa_5^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\mathrm{dist}\big((\boldsymbol{p},\varphi),\,\mathbf{H}_h^{\boldsymbol{p}}\times\mathrm{H}_h^{\varphi}\big)\\&+\kappa^{-1}c_1(\Omega)(1+\kappa_5^2)^{1/2}\|\varphi\|_{1,\Omega}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}+\kappa^{-1}U(1+\kappa_5^2)^{1/2}\|\varphi-\varphi_h\|_{0,\Omega}\Big\}.\end{aligned}\tag{1.83}$$

Proof. It follows from a slight modification of the proof of [35, Lemma 5.3] which makes use of Lemma 1.13. There, the consistency error associated to the functional in the Strang estimate vanishes, but this does not happen in the present case with $\widetilde{F}_{\varphi}-\widetilde{F}_{\varphi_h}$. We simply bound this term similarly as in the proof of Lemma 1.5. We omit further details. ∎

We now combine the two previous lemmas to derive an a priori estimate for the total error $\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi)-(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\|$. Indeed, by gathering together the estimates (1.78) and (1.83), we get

$$\begin{aligned}\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi)-(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\| \,\leq\,& C_{ST}L_\mu C_\delta(1+\kappa_3^2)^{1/2}\|\boldsymbol{t}\|_{\delta,\Omega}\|\varphi-\varphi_h\|_{\mathrm{L}^{3/\delta}(\Omega)}\\&+\Big(C_{ST}c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}\|_{1,\Omega}\,+\,\widetilde{C}_{ST}\kappa^{-1}c_1(\Omega)(1+\kappa_5^2)^{1/2}\|\varphi\|_{1,\Omega}\Big)\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\\&+\Big(C_{ST}\gamma(1+\kappa_2^2)^{1/2}\|\boldsymbol{g}\|_{\infty,\Omega}\,+\,\widetilde{C}_{ST}\kappa^{-1}U(1+\kappa_5^2)^{1/2}\Big)\|\varphi-\varphi_h\|_{0,\Omega}\\&+C_{ST}\Big(1+2\|\mathbf{A}_\varphi\|+c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\mathrm{dist}\big((\underline{\boldsymbol{t}},\boldsymbol{u}),\,\mathbb{H}_h\times\mathbf{H}_h^{\boldsymbol{u}}\big)\\&+\widetilde{C}_{ST}\Big(1+\kappa^{-1}c_1(\Omega)(1+\kappa_5^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\mathrm{dist}\big((\boldsymbol{p},\varphi),\,\mathbf{H}_h^{\boldsymbol{p}}\times\mathrm{H}_h^{\varphi}\big)\,.\end{aligned}$$

The first term of the right-hand side of the foregoing inequality is estimated by using (1.57) to bound $\|\boldsymbol{t}\|_{\delta,\Omega}$, and the continuous injection of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^{3/\delta}(\Omega)$ to get $\|\varphi-\varphi_h\|_{\mathrm{L}^{3/\delta}(\Omega)}\leq\widetilde{C}_\delta\|\varphi-\varphi_h\|_{1,\Omega}$. In turn, in the second term, we use that $(\boldsymbol{u},\varphi)\in W_r$ to bound $\|\boldsymbol{u}\|_{1,\Omega}$ and $\|\varphi\|_{1,\Omega}$ by r. In this way, after performing some algebraic manipulations, we can assert that

$$\begin{aligned}\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi)-(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\| \,\leq\,& \mathbf{C}(\boldsymbol{f},\boldsymbol{g},\kappa,\mu,\gamma,U,r,|\Omega|)\,\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi)-(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\|\\&+C_{ST}\Big(1+2\|\mathbf{A}_\varphi\|+c_2(\Omega)(1+\kappa_3^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\mathrm{dist}\big((\underline{\boldsymbol{t}},\boldsymbol{u}),\,\mathbb{H}_h\times\mathbf{H}_h^{\boldsymbol{u}}\big)\\&+\widetilde{C}_{ST}\Big(1+\kappa^{-1}c_1(\Omega)(1+\kappa_5^2)^{1/2}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\mathrm{dist}\big((\boldsymbol{p},\varphi),\,\mathbf{H}_h^{\boldsymbol{p}}\times\mathrm{H}_h^{\varphi}\big)\,,\end{aligned}\tag{1.84}$$

where $\mathbf{C}(\boldsymbol{f},\boldsymbol{g},\kappa,\mu,\gamma,U,|\Omega|)$ is a constant, depending only on data, r and $|\Omega|$, but is independent of h, defined by

$$\mathbf{C}(\boldsymbol{f},\boldsymbol{g},\kappa,\mu,\gamma,U,r,|\Omega|)\,:=\,\max\big\{\,\mathbf{C}_1(\boldsymbol{f},\boldsymbol{g},\mu,\gamma,r,|\Omega|)\,,\mathbf{C}_2(\kappa,r)\,,\mathbf{C}_3(\boldsymbol{g},\kappa,\gamma,U)\,\big\}\,,\tag{1.85}$$

with

$$\mathbf{C}_1(\boldsymbol{f},\boldsymbol{g},\mu,\gamma,r,|\Omega|)\,:=\,L_\mu\,C_1\,\big\{\|\boldsymbol{f}\|_{\delta,\Omega}\,+\,\big(|\Omega|^{1/2}+\gamma\,r\big)\|\boldsymbol{g}\|_{\infty,\Omega}\big\}\,,\quad\mathbf{C}_2(\kappa,r)\,:=\,r\,C_2\,(\kappa^{-1}+1)\,,$$

$$\text{and}\quad\mathbf{C}_3(\boldsymbol{g},\kappa,\gamma,U)\,:=\,C_3\,\big(\gamma\|\boldsymbol{g}\|_{\infty,\Omega}\,+\kappa^{-1}U\big)\,,$$

where

$$C_1 := C_{ST}\, C_\delta\, \widetilde{C}_\delta\, \widehat{C}_{\mathbf{S}}\, (1+\kappa_3^2)^{1/2}\,, \quad C_2 := C_{ST}\, c_2(\Omega)\, (1+\kappa_3^2)^{1/2} \,+\, \widetilde{C}_{ST}\, c_1(\Omega)\, (1+\kappa_5^2)^{1/2}\,,$$
$$\text{and} \quad C_3 := C_{ST}\, (1+\kappa_2^2)^{1/2} \,+\, \widetilde{C}_{ST}\, (1+\kappa_5^2)^{1/2}\,.$$

Note that the constants multiplying the distances $\mathrm{dist}\big((\underline{\boldsymbol{t}},\boldsymbol{u}),\, \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}\big)$ and $\mathrm{dist}\big((\boldsymbol{p},\varphi),\, \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}\big)$ are both controlled by other constants, parameters, and data only because so $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}$ does, according to Theorem 1.1. Consequently, we are in position to establish the following result providing the complete Céa estimate.

Theorem 1.3. *Let $r \in (0, \min\{r_0, \widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (1.44) and (1.52), respectively, and $(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi) \in \mathbb{H} \times \mathbf{H}_0^1(\Omega) \times \mathbf{H}_\Gamma(\mathrm{div};\Omega) \times \widetilde{\mathrm{H}}^1(\Omega)$ and $(\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h) \in \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}$, with $(\boldsymbol{u},\varphi) \in W_r$ and $(\boldsymbol{u}_h,\varphi_h) \in W_{r,h}$, be solutions to the problems (1.23) and (1.67), respectively. Assume that the data, r and Ω are such that the constant defined by (1.85) satisfies*

$$\mathbf{C}(\boldsymbol{f},\boldsymbol{g},\kappa,\mu,\gamma,U,r,|\Omega|) \,\leq\, \frac{1}{2}\,. \tag{1.86}$$

Then, there exists a positive constant C, depending only on parameters, data and other constants, all of them independent of h, such that

$$\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi) - (\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\| \,\leq C\Big\{\mathrm{dist}\big((\underline{\boldsymbol{t}},\boldsymbol{u}),\, \mathbb{H}_h \times \mathbf{H}_h^{\boldsymbol{u}}\big) + \mathrm{dist}\big((\boldsymbol{p},\varphi),\, \mathbf{H}_h^{\boldsymbol{p}} \times \mathrm{H}_h^{\varphi}\big)\Big\}. \tag{1.87}$$

Proof. It follows by using the hypothesis (1.86) into the estimate (1.84) and the fact that $\boldsymbol{u}$ and $\boldsymbol{u}_h$ are both bounded by r and so $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega} \leq 2r$. ∎

Finally, we complete our a priori error analysis stating the corresponding convergence rate of our Galerkin scheme (1.67).

Theorem 1.4. *In addition to the hypotheses of Theorems 1.1, 1.2 and 1.3, assume that there exists $s > 0$ such that $\boldsymbol{t} \in \mathbb{H}^s(\Omega)$, $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega)$ with $\mathbf{div}\,\boldsymbol{\sigma} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{\rho} \in \mathbb{H}^s(\Omega)$, $\boldsymbol{u} \in \mathbf{H}^{s+1}(\Omega)$, $\boldsymbol{p} \in \mathbf{H}^s(\Omega)$ with $\mathrm{div}\,\boldsymbol{p} \in \mathrm{H}^s(\Omega)$, and $\varphi \in \mathrm{H}^{s+1}(\Omega)$. Then, there exists $C > 0$, independent of h, such that*

$$\|(\underline{\boldsymbol{t}},\boldsymbol{u},\boldsymbol{p},\varphi) - (\underline{\boldsymbol{t}}_h,\boldsymbol{u}_h,\boldsymbol{p}_h,\varphi_h)\| \,\leq\, Ch^{\min\{s,k+1\}}\Big\{\Big\}.\ \|\boldsymbol{t}\|_{s,\Omega}+\|\boldsymbol{\sigma}\|_{s,\Omega}+\|\mathbf{div}\,\boldsymbol{\sigma}\|_{s,\Omega}+\|\boldsymbol{\rho}\|_{s,\Omega} \qquad +\|\boldsymbol{u}\|_{s+1,\Omega}+\|\boldsymbol{p}\|_{s,\Omega}+\| \tag{1.88}$$

Proof. It follows directly from the Céa estimate (1.87) and standard approximation properties of the discrete spaces $\mathbb{H}_h^{\boldsymbol{t}}$, $\mathbb{H}_h^{\boldsymbol{\sigma}}$, $\mathbb{H}_h^{\boldsymbol{\rho}}$, $\mathbf{H}_h^{\boldsymbol{u}}$, $\mathbf{H}_h^{\boldsymbol{p}}$ and H_h^{φ} (see [21, 49], for instance). ∎

Now, regarding the postprocessing of additional variables, on the one hand, we recall the orthogonal decomposition for the pseudostress tensor provided in (1.16), and then the modified equation for the continuous pressure (1.12) becomes

$$p = -\frac{1}{3}\mathrm{tr}(\boldsymbol{\sigma} + c\mathbb{I} + (\boldsymbol{u}\otimes\boldsymbol{u})), \quad \text{with} \quad c := -\frac{1}{3|\Omega|}\int_\Omega \mathrm{tr}(\boldsymbol{u}\otimes\boldsymbol{u}). \tag{1.89}$$

Thus, according to (1.89), we define our discrete approximation of the pressure as

$$p_h = -\frac{1}{3}\mathrm{tr}(\boldsymbol{\sigma}_h + c_h\mathbb{I} + (\boldsymbol{u}_h \otimes \boldsymbol{u}_h)), \quad \text{with} \quad c_h := -\frac{1}{3|\Omega|}\int_\Omega \mathrm{tr}(\boldsymbol{u}_h \otimes \boldsymbol{u}_h), \tag{1.90}$$

which yields

$$p - p_h = \frac{1}{3}\mathrm{tr}\Big\{(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}) + (\boldsymbol{u}_h \otimes \boldsymbol{u}_h - \boldsymbol{u} \otimes \boldsymbol{u})\Big\} + (c_h - c).$$

On the other hand, such as in [37], it is not difficult to see that the relation (1.13) gives also the chance to compute the discrete concentration gradient through the formulae

$$\nabla\varphi_h \;=\; \kappa^{-1}\boldsymbol{p}_h + \kappa^{-1}\varphi_h\boldsymbol{u}_h + \kappa^{-1}U(\varphi_h + \alpha)\mathbf{i}_3\,. \tag{1.91}$$

Therefore, similarly to [23, Section 4], we easily deduce that there exist constants $C, \widetilde{C} > 0$, independent of h, such that

$$\begin{aligned} \|p - p_h\|_{0,\Omega} &\leq C\Big\{\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}\Big\}, \\ \|\nabla\varphi - \nabla\varphi_h\|_{0,\Omega} &\leq \widetilde{C}\Big\{\|\boldsymbol{p} - \boldsymbol{p}_h\|_{\mathrm{div};\Omega} + \|\varphi - \varphi_h\|_{1,\Omega} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}\Big\}, \end{aligned} \tag{1.92}$$

and so the convergence rates of the postprocessed variables, in the L^2-norm, coincide with those provided by (1.88) (cf. Theorem 1.4).

1.6 Numerical results

This section presents a few numerical examples to illustrate the performance of our augmented fully-mixed formulation (1.67) and to support the respective convergence theoretical results for the primary and postprocessed variables predicted by Theorem 1.4 and the estimates (1.92), respectively. The fixed-point problem (1.70) has been implemented through a Picard iteration on a **FreeFem++** code (cf. [59]) and the resulting algebraic linear systems have been solved with the direct linear solver UMFPACK (see [39]). As an initial solution, we have simply taken $(\boldsymbol{u}_h^{(0)}, \varphi_h^{(0)}) = (\mathbf{0}, 0)$ to construct, on each step m, the entire solution vector

$$\mathbf{sol}^{(m)} = (\boldsymbol{t}_h^{(m)}, \boldsymbol{\sigma}_h^{(m)}, \boldsymbol{\rho}_h^{(m)}, \boldsymbol{u}_h^{(m)}, \boldsymbol{p}_h^{(m)}, \varphi_h^{(m)}) \quad \text{for all} \quad m \geq 1\,.$$

As a stopping criteria, we have prescribed a fixed tolerance $\mathtt{tol} = 1\mathrm{E}{-8}$ to finish the iterative technique when either a maximum number of iterations is reached or the relative error between two consecutive iterations, let us say $\mathbf{sol}^{(m)}$ and $\mathbf{sol}^{(m+1)}$, satisfies

$$\frac{||\mathbf{sol}^{(m+1)} - \mathbf{sol}^{(m)}||_{\ell^2}}{||\mathbf{sol}^{(m+1)}||_{\ell^2}} < \mathtt{tol}\,,$$

where $||\cdot||_{\ell^2}$ stands for the Euclidean ℓ^2-norm in R^N with N denoting the total number of degrees of freedom defined by the finite element family $(\mathbb{H}_h^{\boldsymbol{t}}, \mathbb{H}_h^{\boldsymbol{\sigma}}, \mathbb{H}_h^{\boldsymbol{\rho}}, \mathbf{H}_h^{\boldsymbol{u}}, \mathbf{H}_h^{\boldsymbol{p}}, \mathrm{H}_h^{\varphi})$ specified in Section 1.4.1.

The individual errors associated to the primary unknowns are denoted and defined by

$$\mathtt{e}(\boldsymbol{t}) := \|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,\Omega}, \quad \mathtt{e}(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}, \quad \mathtt{e}(\boldsymbol{\rho}) := \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\Omega},$$

$$\mathtt{e}(\boldsymbol{u}) := \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}, \quad \mathtt{e}(\boldsymbol{p}) := \|\boldsymbol{p} - \boldsymbol{p}_h\|_{\mathrm{div};\Omega}, \quad \text{and} \quad \mathtt{e}(\varphi) := \|\varphi - \varphi_h\|_{1,\Omega}\,,$$

and the errors associated to the postprocessed variables (cf. (1.90) and (1.91)) are given, respectively, as

$$\mathtt{e}(p) := \|p - p_h\|_{0,\Omega} \quad \text{and} \quad \mathtt{e}(\nabla\varphi) := \|\nabla\varphi - \nabla\varphi_h\|_{0,\Omega}\,.$$

We also let $\mathtt{e}_{\mathtt{prim}}$ and $\mathtt{e}_{\mathtt{post}}$ be the total errors related to the primary and post-processed variables, respectively, that is,

$$\mathtt{e}_{\mathtt{prim}} := \big\{\mathtt{e}(\boldsymbol{t})^2 + \mathtt{e}(\boldsymbol{\sigma})^2 + \mathtt{e}(\boldsymbol{\rho})^2 + \mathtt{e}(\boldsymbol{u})^2 + \mathtt{e}(\boldsymbol{p})^2 + \mathtt{e}(\varphi)^2\big\}^{1/2} \quad \text{and} \quad \mathtt{e}_{\mathtt{post}} := \big\{\mathtt{e}(p)^2 + \mathtt{e}(\nabla\varphi)^2\big\}^{1/2}.$$

Following the same notation, we denote $\mathtt{r}(\,\cdot\,)$, $\mathtt{r}_{\mathtt{prim}}$ and $\mathtt{r}_{\mathtt{post}}$ as the individual experimental convergence rate associated to each variable, and the total convergence rates of the primary unknowns and post-processed variables, respectively, that is

$$\mathtt{r}(\,\cdot\,) := \frac{\log(\mathtt{e}(\,\cdot\,)/\mathtt{e}'(\,\cdot\,))}{\log(h/h')}, \quad \mathtt{r}_{\mathtt{prim}} := \frac{\log(\mathtt{e}_{\mathtt{prim}}/\mathtt{e}'_{\mathtt{prim}})}{\log(h/h')} \quad \text{and} \quad \mathtt{r}_{\mathtt{post}} := \frac{\log(\mathtt{e}_{\mathtt{post}}/\mathtt{e}'_{\mathtt{post}})}{\log(h/h')},$$

where h and h' denote two consecutive meshsizes with errors $\mathtt{e}$ and $\mathtt{e}'$.

1.6.1 Example 1: Accuracy assessment in 2D

In our first example we study the accuracy of the method in 2D by manufacturing an exact solution of a corresponding modification of problem (1.6) and considering a non concentration-dependent viscosity. More precisely, the expressions $\mathbf{i}_3$, $\frac{\partial\varphi}{\partial x_3}$, and ν_3 are replaced in (1.6) by $\mathbf{i}_2 := (0,1)$, $\frac{\partial\varphi}{\partial x_2}$, and ν_2, respectively. Then, we consider the square $\Omega := (-1,1)^2$ and the data

$$\mu(x_1,x_2) = 1 + \sin^2(x_1), \quad U = 0.01, \quad \gamma = 0.5, \qquad \kappa = 1, \quad \alpha = 0.5 \quad \text{and} \quad \boldsymbol{g} = (0,1)^t\,. \tag{1.93}$$

It follows that $\mu_1 = 1$ and $\mu_2 = 2$ (cf. (1.3)), and hence the stabilization parameters κ_i, $(i = 1,\ldots,6)$, are chosen as in (1.54) and in accordance to Lemmas 1.1 and 1.2, that is

$$\kappa_1 = \frac{\mu_1}{2}, \quad \kappa_2 = \kappa_3 = \frac{\mu_1}{\mu_2^2}, \quad \kappa_4 = \frac{\mu_1}{4}, \quad \kappa_5 = \kappa, \quad \text{and} \quad \kappa_6 = \frac{\kappa^{-1}}{2}\,. \tag{1.94}$$

The terms on the right-hand sides are adjusted in such a way that the exact solutions are given by the smooth functions

$$\boldsymbol{u}(x_1,x_2) = \begin{pmatrix} 2\pi\cos(\pi x_2)\sin(\pi x_2)\sin^2(\pi x_1) \\ -2\pi\cos(\pi x_1)\sin(\pi x_1)\sin^2(\pi x_2) \end{pmatrix}, \quad p(x_1,x_2) = -5x_1\sin(x_2)\,,$$

and

$$\varphi(x_1,x_2) = \vartheta\, exp\left(\frac{U}{\kappa}x_2\right) - \alpha\,, \quad \text{where } \vartheta \in \mathbb{R} \text{ is taken so that} \quad \int_\Omega \varphi = 0.$$

Note that the homogeneous Dirichlet condition for the velocity, the Robin-type boundary condition for the concentration, the incompressibbility condition of the fluid, and the zero-mean value restriction for both the pressure and the concentration are satisfied by the above functions.

Values of errors and corresponding convergence rates associated to the approximations with the finite element families $\mathbb{P}_0 - \mathbb{RT}_0 - \mathbb{P}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ and $\mathbb{P}_1^{disc} - \mathbb{RT}_1 - \mathbb{P}_1^{disc} - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2$

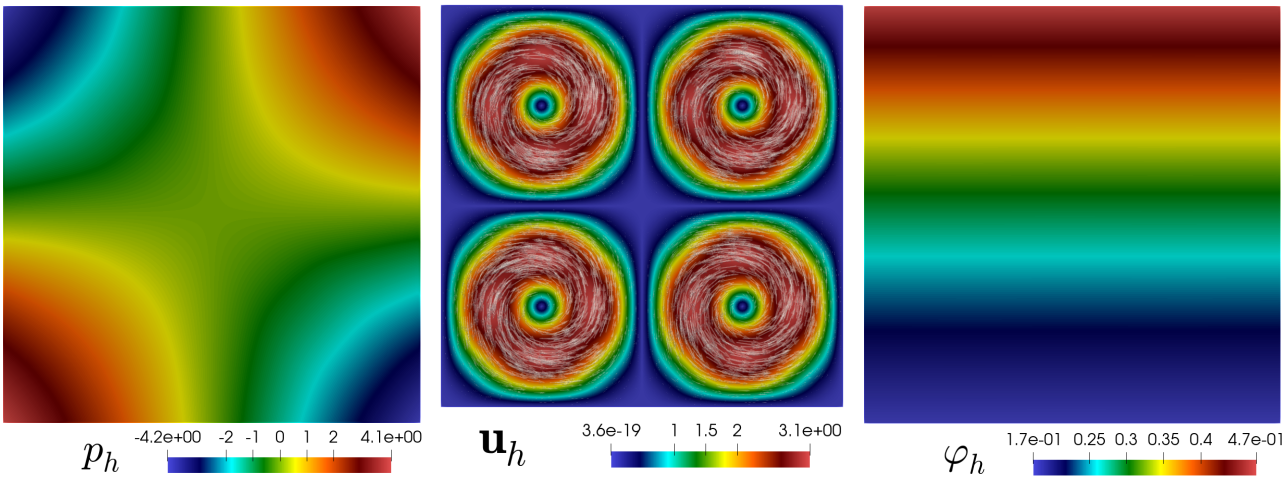


Figure 1.1: Example 1: Approximated pressure, velocity magnitude, and concentration obtained with the fully-mixed method using $k = 0$ and $N = 873843$ degrees of freedom.

corresponding to approximations of order $k = 0$ and $k = 1$, respectively, are reported in Table 1.1. There, we observe that the convergence rates are linear (in the case $k = 0$) and quadratic (in the case $k = 1$) with respect to h for all the main unknowns in their respective norms, as well as the post-processed variables in the L^2-norm. Also, it is observed that the errors decay faster when increasing the approximation order from $k = 0$ to $k = 1$. In particular, this behavior can be observed from the values related to the total convergence rates $\mathtt{r}_{\mathtt{prim}}$ and $\mathtt{r}_{\mathtt{post}}$ for the primary and the variables obtained by post-processing. Our findings are in agreement with the theoretical error bounds predicted from Theorem 1.4 and the estimates (1.92). On the other hand, we mention that 8 and 9 Picard steps were required to reach the prescribed tolerance $\mathtt{tol}$ = 1E-08 in the cases $k = 0$ and $k = 1$, respectively. The approximation of the velocity magnitude, the pressure and concentration are depicted in Figure 1.1 computed with our fully-mixed method on a mesh with $N = 873843$ degrees of freedom and $k = 0$.

1.6.2 Example 2: Accuracy assessment in 3D with concentration-dependent viscosity

In this example we focus on testing the accuracy of our method in the three-dimensional setting and considering the viscosity as a concentration-dependent function. To that end, we define the manufactured exact solution in the cube $\Omega := (0,1)^3$ as

$$\boldsymbol{u}(x_1,x_2,x_3) = \begin{pmatrix} 4x_1^2x_2x_3(x_3-1)(x_2-1)(x_2-x_3)(x_1-1)^2 \\ -4x_1x_2^2x_3(x_2-1)^2(x_3-1)(x_1-1)(x_1-x_3) \\ 4x_1x_2x_3^2(x_3-1)^2(x_2-1)(x_1-1)(x_1-x_2) \end{pmatrix},$$

$$p(x_1,x_2,x_3) = \cos(\pi x_1)\cos(x_2)\cos(x_3)\,,$$

and, similarly as in the first example, the auxiliary exact concentration satisfying the Robin-type boundary condition takes the form

$$\varphi(x_1,x_2,x_3) = \vartheta\, exp\left(\frac{U}{\kappa}x_3\right) - \alpha\,, \quad \text{where } \vartheta \in \mathbb{R} \text{ is taken so that } \int_\Omega \varphi = 0.$$

Fully–mixed $\mathbb{P}_0 - \mathbb{RT}_0 - \mathbb{P}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1\,(k=0)$ scheme											
$\mathtt{e}(\boldsymbol{t})$	$\mathtt{r}(\boldsymbol{t})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\rho})$	$\mathtt{r}(\boldsymbol{\rho})$	$\mathtt{e}(\boldsymbol{u})$	$\mathtt{r}(\boldsymbol{u})$	$\mathtt{e}(\boldsymbol{p})$	$\mathtt{r}(\boldsymbol{p})$	$\mathtt{e}(\varphi)$	$\mathtt{r}(\varphi)$
2.4590	–	17.755	–	2.1535	–	4.4929	–	0.1792	–	0.1728	–
1.2280	1.0018	8.9033	0.9958	1.1132	0.9520	2.2405	1.0038	0.0898	0.9968	0.0869	0.9917
0.8925	0.9976	6.4778	0.9943	0.8147	0.9759	1.6285	0.9974	0.0654	0.9912	0.0633	0.9906
0.7010	0.9980	5.0906	0.9958	0.6419	0.9850	1.2792	0.9976	0.0514	0.9953	0.0498	0.9928
0.5607	1.0098	4.0729	1.0085	0.5144	1.0012	1.0232	1.0097	0.0411	1.0112	0.0398	1.0117
0.3924	0.9928	2.8513	0.9919	0.3606	0.9881	0.7162	0.9923	0.0288	0.9892	0.0279	0.9881
0.3567	1.0724	2.5921	1.0715	0.3279	1.0687	0.6510	1.0731	0.0262	1.0637	0.0253	1.0998
	$\mathtt{e}(p)$	$\mathtt{r}(p)$	$\mathtt{e}(\nabla\varphi)$	$\mathtt{r}(\nabla\varphi)$	$\mathtt{e}_{\mathtt{prim}}$	$\mathtt{r}_{\mathtt{prim}}$	$\mathtt{e}_{\mathtt{post}}$	$\mathtt{r}_{\mathtt{post}}$	h	N	It
	1.5750	–	0.1728	–	18.606	–	1.5845	–	0.0884	18819	8
	0.7723	1.0281	0.0869	0.9917	9.3301	0.9958	0.7772	1.0277	0.0442	74499	8
	0.5587	1.0122	0.0633	0.9906	6.7884	0.9943	0.5623	1.0119	0.0321	140451	8
	0.4378	1.0076	0.0497	0.9995	5.3347	0.9957	0.4406	1.0075	0.0252	227139	8
	0.3497	1.0169	0.0398	1.0044	4.2682	1.0085	0.3520	1.0158	0.0202	354483	8
	0.2444	0.9966	0.0279	0.9881	2.9881	0.9918	0.2460	0.9964	0.0141	722403	8
	0.2222	1.0706	0.0253	1.0998	2.7164	1.0716	0.2236	1.0710	0.0129	873843	8

Fully–mixed $\mathbb{P}_1^{disc} - \mathbb{RT}_1 - \mathbb{P}_1^{disc} - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2\,(k=1)$ scheme											
$\mathtt{e}(\boldsymbol{t})$	$\mathtt{r}(\boldsymbol{t})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\rho})$	$\mathtt{r}(\boldsymbol{\rho})$	$\mathtt{e}(\boldsymbol{u})$	$\mathtt{r}(\boldsymbol{u})$	$\mathtt{e}(\boldsymbol{p})$	$\mathtt{r}(\boldsymbol{p})$	$\mathtt{e}(\varphi)$	$\mathtt{r}(\varphi)$
0.3204	–	2.2831	–	0.2628	–	0.5630	–	0.0214	–	0.0205	–
0.1812	1.9853	1.2892	1.9905	0.1523	1.9001	0.3186	1.9830	0.0121	1.9859	0.0117	1.9533
0.0808	1.9898	0.5745	1.9907	0.0695	1.9322	0.1422	1.9868	0.0054	1.9871	0.0052	1.9972
0.0455	1.9988	0.3234	2.0013	0.0395	1.9679	0.0801	1.9990	0.0031	1.9330	0.0030	1.9158
0.0359	2.0168	0.2556	2.0023	0.0313	1.9803	0.0633	2.0033	0.0024	2.1782	0.0023	2.2613
0.0239	2.0105	0.1712	1.9805	0.0208	2.0194	0.0423	1.9919	0.0016	2.0036	0.0015	2.1122
	$\mathtt{e}(p)$	$\mathtt{r}(p)$	$\mathtt{e}(\nabla\varphi)$	$\mathtt{r}(\nabla\varphi)$	$\mathtt{e}_{\mathtt{prim}}$	$\mathtt{r}_{\mathtt{prim}}$	$\mathtt{e}_{\mathtt{post}}$	$\mathtt{r}_{\mathtt{post}}$	h	N	It
	0.2185	–	0.0205	–	2.3879	–	0.2195	–	0.1178	35139	9
	0.1236	1.9843	0.0117	1.9533	1.3490	1.9889	0.1242	1.9841	0.0884	62211	9
	0.0550	1.9942	0.0052	1.9972	0.6014	1.9897	0.0552	1.9943	0.0589	139395	9
	0.0309	2.0082	0.0030	1.9158	0.3386	2.0007	0.0310	2.0073	0.0442	247299	9
	0.0244	2.0100	0.0023	2.2613	0.2676	2.0024	0.0245	2.0123	0.0393	312771	9
	0.0163	1.9935	0.0015	2.1122	0.1792	1.9822	0.0164	1.9945	0.0321	466755	9

Table 1.1: Example 1: Convergence history for the fully-mixed approximation of the Bioconvection problem with $k = 0$ (first and second panel) and $k = 1$ (third and fourth panel)

.

Next, the viscosity is taken as a concentration-dependent function defined as

$$\mu(\varphi) = 1 + \sin^2(\varphi),$$

satisfying (1.2) and (1.3) with $\mu_1 = 1$ and $\mu_2 = 2$. The rest of data and stabilization parameters are taken as in (1.93) and (1.94).

For this example, we consider the finite element spaces introduced in Section 1.4.1 with $k = 0$. The convergence history is summarized in Table 1.2 and it is observed there that the total error decay is of order $\mathcal{O}(h)$ for the primary unknowns and the postprocessed variables as predicted by Theorem 1.4 and the estimates (1.92). In particular, 4 Picard steps were required to achieve the prescribed tolerance `tol` = 1E-08. Next, in Figure 1.2 we display the streamlines, the component $\boldsymbol{\rho}_{12,h}$ of the vorticity tensor and the concentration profile φ in the first panel, whereas in the second panel are depicted the component $\boldsymbol{t}_{11,h}$ of the shear stress tensor, the component $\boldsymbol{\sigma}_{23,h}$ of the pseudo-stress tensor and the concentration gradient vector field $\nabla\varphi_h$ obtained with $k = 0$ and $N = 1403428$ degrees of freedom.

1.6.3 Example 3: Accuracy assessment with no manufactured analytical solution

In this example we aim to illustrate the accuracy of our method by considering a case in which the exact solution is unknown in the two-dimensional setting. As in the previous example, we consider the viscosity as the concentration-dependent function given by $\mu(\varphi) = 1 + \sin^2(\varphi)$, satisfying (1.2) and (1.3) with $\mu_1 = 1$ and $\mu_2 = 2$. Here, the source term is taken as $\boldsymbol{f} = \mathbf{0}$, and the data are given by

$$U = 0.01, \quad \gamma = 0.1, \qquad \kappa = 0.08, \quad \alpha = 0.3 \quad \text{and} \quad \boldsymbol{g} = (0, 9.8)^t,$$

in terms of which the parameters κ_i, $(i = 1, \ldots, 6)$ are defined according to (1.94). The boundary conditions are imposed as in (1.4) In Table 1.3, we summarize the convergence history for a sequence of uniform triangulations, considering a $\mathbb{P}_0 - \mathbb{RT}_0 - \mathbb{P}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ approximation. We mention that the errors and the convergence rates of are computed by considering the discrete solution obtained with a finer mesh ($N = 822,774$) as the exact solution. It is observed that the rate of convergence $O(h)$ is attained by all the primary and post processed unknowns as well as the total convergence rates in agreement with Theorem 1.4 and the estimates (1.92). Additionally, in Figure 1.3 we display the approximation of the velocity components whereas in Figure 1.4, we illustrate, in a 3D view, the pressure (left) and the concentration (right) scalar fields and observe there that p_h has a linear behavior differently than φ_h. All the figures presented there were obtained with $N = 181,203$ degrees of freedom.

Fully–mixed $\mathbb{P}_0 - \mathbb{RT}_0 - \mathbb{P}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1\,(k=0)$ scheme											
$\mathtt{e}(\boldsymbol{t})$	$\mathtt{r}(\boldsymbol{t})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\rho})$	$\mathtt{r}(\boldsymbol{\rho})$	$\mathtt{e}(\boldsymbol{u})$	$\mathtt{r}(\boldsymbol{u})$	$\mathtt{e}(\boldsymbol{p})$	$\mathtt{r}(\boldsymbol{p})$	$\mathtt{e}(\varphi)$	$\mathtt{r}(\varphi)$
0.0817	–	0.6974	–	0.0446	–	0.0916	–	0.4489	–	1.0850	–
0.0676	0.4672	0.4764	0.9398	0.0419	0.1544	0.0647	0.8576	0.4365	0.0692	0.8401	0.6039
0.0550	0.7174	0.3580	0.9941	0.0371	0.4209	0.0452	1.2500	0.3301	0.9713	0.6096	1.1154
0.0389	0.8569	0.2360	1.0274	0.0290	0.6089	0.0252	1.4391	0.1918	1.3385	0.3433	1.4156
0.0297	0.9306	0.1750	1.0400	0.0234	0.7546	0.0161	1.5537	0.1240	1.5178	0.2160	1.6113
0.0240	0.9623	0.1388	1.0375	0.0194	0.8314	0.0113	1.5853	0.0875	1.5611	0.1478	1.6981
0.0200	0.9807	0.1150	1.0355	0.0165	0.8813	0.0085	1.5883	0.0658	1.5629	0.1077	1.7414
0.0151	0.9861	0.0856	1.0239	0.0127	0.9166	0.0054	1.5436	0.0426	1.5095	0.0654	1.7349
0.0121	0.9924	0.0682	1.0169	0.0103	0.9465	0.0039	1.4727	0.0310	1.4282	0.0447	1.7019
0.0101	0.9933	0.0567	1.0105	0.0086	0.9611	0.0030	1.3985	0.0242	1.3510	0.0331	1.6428
0.0093	1.0021	0.0523	1.0152	0.0080	0.9762	0.0027	1.3583	0.0218	1.3048	0.0291	1.6082
	$\mathtt{e}(p)$	$\mathtt{r}(p)$	$\mathtt{e}(\nabla\varphi)$	$\mathtt{r}(\nabla\varphi)$	$\mathtt{e}_{\mathtt{prim}}$	$\mathtt{r}_{\mathtt{prim}}$	$\mathtt{e}_{\mathtt{post}}$	$\mathtt{r}_{\mathtt{post}}$	h	N	It
	0.2231	–	1.0680	–	0.7096	–	0.2233	–	0.7071	972	4
	0.1496	0.9845	0.8146	0.6680	0.4874	0.9263	0.1499	0.9837	0.4714	3064	4
	0.1091	1.0995	0.5879	1.1342	0.3669	0.9875	0.1092	1.0099	0.3536	7028	4
	0.0674	1.1858	0.3303	1.4215	0.2423	1.0236	0.0675	1.1864	0.2357	22792	4
	0.0474	1.2239	0.2076	1.6150	0.1798	1.0375	0.0475	1.2247	0.1768	53604	4
	0.0362	1.2111	0.1421	1.6967	0.1426	1.0358	0.0362	1.2120	0.1414	103724	4
	0.0291	1.1919	0.1037	1.7373	0.1182	1.0342	0.0292	1.1926	0.1179	178132	4
	0.0201	1.1502	0.0631	1.7257	0.0880	1.0230	0.0209	1.1508	0.0884	419012	4
	0.0163	1.1107	0.0447	1.5443	0.0702	1.0163	0.0163	1.1110	0.0707	814644	4
	0.0134	1.0808	0.0322	1.7988	0.0583	1.0101	0.0134	1.0812	0.0589	1403428	4
	0.0123	1.0732	0.0284	1.5868	0.0538	1.0149	0.0123	1.0735	0.0129	1782252	4

Table 1.2: Example 2: Convergence history for the fully-mixed approximation of the three-dimensional Bioconvection problem with concentration-dependent viscosity and using approximation order $k = 0$

.

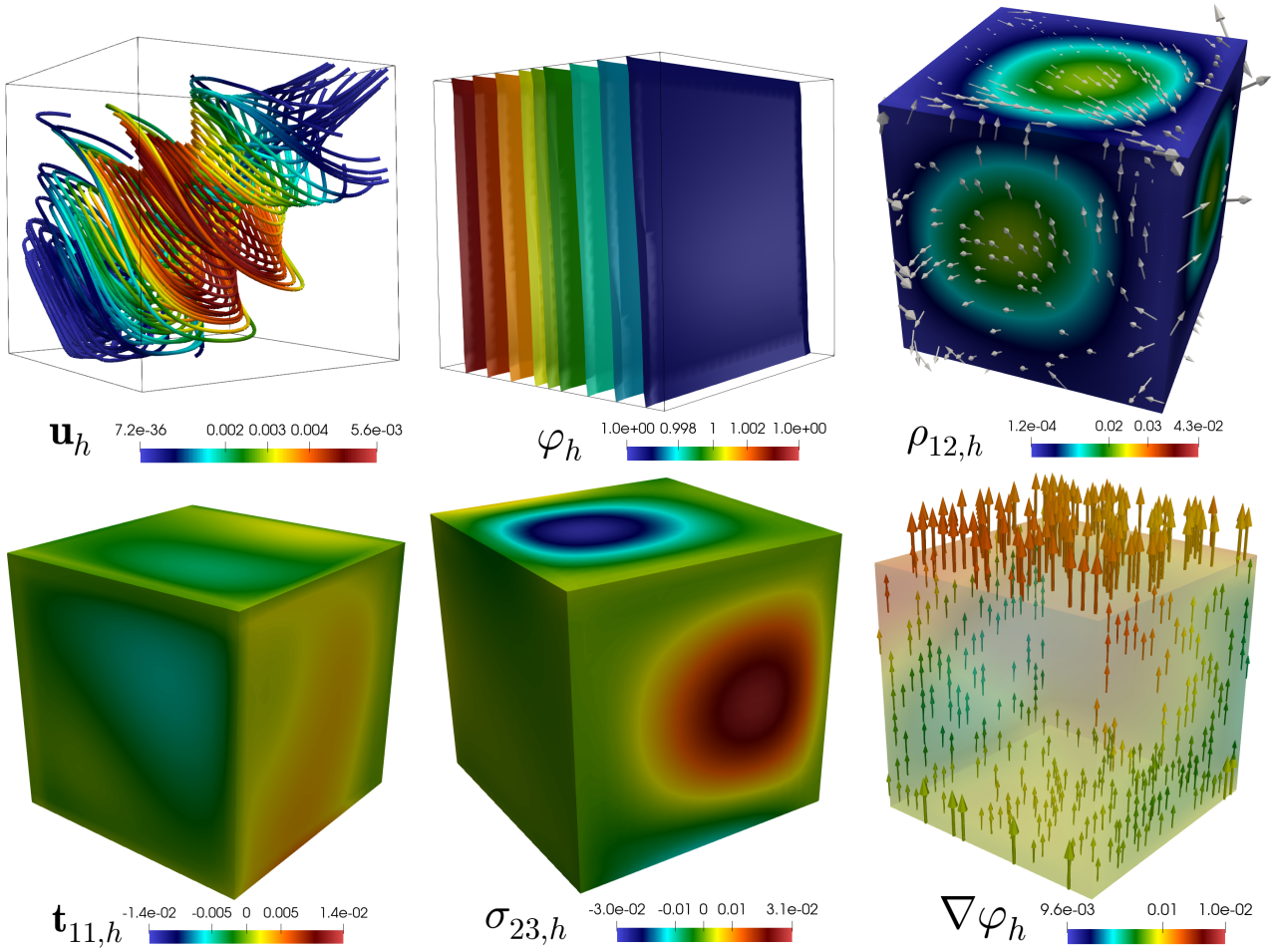


Figure 1.2: Example 2: Streamlines, concentration profile φ_h, and component $\boldsymbol{\rho}_{12,h}$ of the vorticity tensor (first panel), and the component $\boldsymbol{t}_{11,h}$ of the shear stress tensor, component $\boldsymbol{\sigma}_{23,h}$ of the pseudo-stress tensor and concentration gradient $\nabla\varphi_h$ obtained with the fully-mixed method for the Bioconvection problem using $k = 0$ and $N = 1403428$ degrees of freedom.

Fully–mixed $\mathbb{P}_0 - \mathbb{RT}_0 - \mathbb{P}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1\,(k=0)$ scheme											
$\mathtt{e}(\boldsymbol{t})$	$\mathtt{r}(\boldsymbol{t})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\rho})$	$\mathtt{r}(\boldsymbol{\rho})$	$\mathtt{e}(\boldsymbol{u})$	$\mathtt{r}(\boldsymbol{u})$	$\mathtt{e}(\boldsymbol{p})$	$\mathtt{r}(\boldsymbol{p})$	$\mathtt{e}(\varphi)$	$\mathtt{r}(\varphi)$
0.8054	–	0.9424	–	1.0043	–	0.3207	–	0.0044	–	0.7832	–
0.4911	0.7137	0.6119	0.6230	0.5652	0.8294	0.2062	0.6372	0.0026	0.7738	0.3074	1.3492
0.3845	0.8510	0.4995	0.7057	0.4456	0.8265	0.1564	0.9609	0.0020	0.8716	0.2119	1.2932
0.2689	0.8815	0.3550	0.8421	0.3166	0.8429	0.1019	1.0566	0.0014	0.8740	0.1355	1.1028
0.2051	0.9418	0.2695	0.9579	0.2455	0.8838	0.0766	0.9920	0.0010	1.0412	0.0989	1.0945
0.1831	0.9634	0.2380	1.0557	0.2197	0.9435	0.0678	1.0362	0.0009	1.1338	0.0871	1.0788
0.1488	1.0336	0.1966	0.9522	0.1781	1.0460	0.0525	1.2744	0.0007	1.3073	0.0689	1.1680
0.1290	1.1170	0.1690	1.1833	0.1522	1.2293	0.0461	1.0169	0.0005	1.4173	0.0582	1.3202
	$\mathtt{e}(p)$	$\mathtt{r}(p)$	$\mathtt{e}(\nabla\varphi)$	$\mathtt{r}(\nabla\varphi)$	$\mathtt{e}_{\mathtt{prim}}$	$\mathtt{r}_{\mathtt{prim}}$	$\mathtt{e}_{\mathtt{post}}$	$\mathtt{r}_{\mathtt{post}}$	h	N	It
	0.3211	–	0.0963	–	1.8060	–	0.3352	–	0.2357	2739	10
	0.1414	1.1828	0.0500	0.9449	1.0354	0.8026	0.1500	1.1599	0.1179	10659	12
	0.1089	0.9031	0.0372	1.0279	0.8156	0.8295	0.1152	0.9165	0.0884	18819	13
	0.0674	1.1862	0.0209	1.4220	0.5721	0.8745	0.0706	1.2088	0.0589	42051	13
	0.0481	1.2233	0.0159	0.9504	0.4366	0.9397	0.0500	1.1976	0.0442	74499	13
	0.0405	1.3406	0.0138	1.2028	0.3881	1.0001	0.0428	1.3265	0.0393	94179	13
	0.0330	1.0205	0.0102	1.5062	0.3163	1.0201	0.0345	1.0668	0.0321	140451	13
	0.0291	0.9838	0.0088	1.1549	0.2718	1.1847	0.0304	0.9985	0.0283	181203	13

Table 1.3: Example 3: Convergence history for the fully-mixed approximation of a two-dimensional Bioconvection problem with no manufactured analytical solution and with concentration-dependent viscosity, using approximation order $k = 0$

.

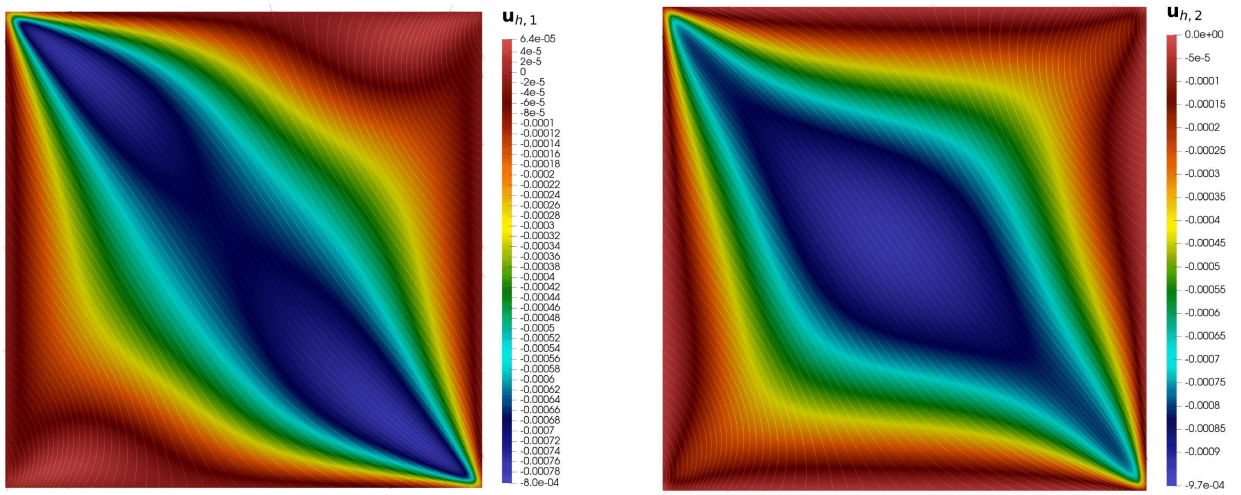


Figure 1.3: Example 3: Horizontal and vertical components $\boldsymbol{u}_{h,1}$ and $\boldsymbol{u}_{h,2}$ (left and right, respectively) of the velocity vector field obtained with the fully-mixed method for the Bioconvection problem with no manufactured analytical solution and with concentration-dependent viscosity using $k = 0$ and $N = 181,203$ degrees of freedom.

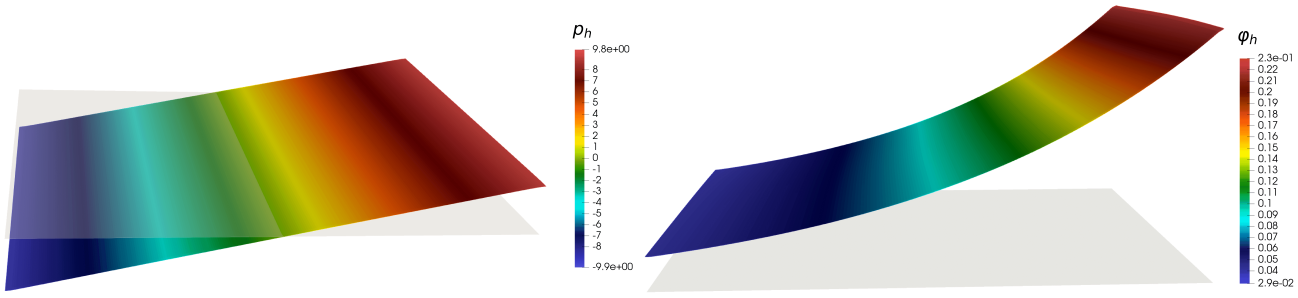


Figure 1.4: Example 3: Pressure p_h and concentration φ_h (left and right, respectively) obtained with the fully-mixed method for the Bioconvection problem with no manufactured analytical solution and with concentration-dependent viscosity using $k = 0$ and $N = 181,203$ degrees of freedom.

CHAPTER 2

New primal and dual-mixed finite element methods for stable image registration with singular regularization

2.1 Introduction

Deformable image registration (DIR) is a challenging process where a given set of images are aligned by means of a transformation that warps one or more of these images. It arises in numerous applications and particularly in medical imaging [85]. Its formulation requires three inputs: a transformation model (composed by a family of mappings that warp the target images), a function that measures the differences between images known as similarity measure, and a regularizer that renders the problem well-posed. In addition to the many variants of these components, different modeling approaches exist, between which we highlight: traditional variational minimization [61, 74], L^2-optimal mass transport [57, 93] (which does not require regularization), and level-set modeling [87]. The solution strategy in general considers the incorporation of an auxiliary time variable, which can be seen as a semi-implicit formulation of the proximal point algorithm [84] recently extended to a more general class of proximal operators by using Forward-backward splitting [45]. Also, machine learning techniques have been recently developed for the solution of this problem, which do not depend on the existence of ground truth solutions and support large deformations [12]. This last work proved competitive against the well-established software ANTs [11].

For a more mathematical explanation of DIR, let us now consider a domain $\Omega \subset \mathrm{R}^{d=2,3}$, and two fields $R : \Omega \to \mathrm{R}$ and $T : \Omega \to \mathrm{R}$ referred to as *reference* and *target* images, where $R(\boldsymbol{x})$ and $T(\boldsymbol{x})$ denote the *image intensity* at point $\boldsymbol{x}$. Then, the objective of DIR is to find a mapping of T onto R by means of a warping $\boldsymbol{u}$ such that–ideally–it holds that

$$T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) = R(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Omega\,. \tag{2.1}$$

This problem is ill-posed in general, so one formulates it as a minimization problem by considering a *similarity measure* $\mathcal{D}$ (a functional which attains its minimum when (2.1) holds), a regularizer $\mathcal{S}$ (which provides smoothness to the problem), a family of deformations $\mathcal{V}$ (such that $\boldsymbol{u} \in \mathcal{V}$) and a positive constant α (which balances $\mathcal{D}$ and $\mathcal{S}$). Putting everything together, the following minimization problem arises:

$$\min_{\boldsymbol{v} \in \mathcal{V}} \Big\{ \alpha \mathcal{D}(\boldsymbol{v}) + \mathcal{S}(\boldsymbol{v}) \Big\}. \tag{2.2}$$

The choices of $\mathcal{V}$ and $\mathcal{S}$ are not independent. For example, it would not make sense to consider $\mathcal{V} = \boldsymbol{L}^2(\Omega)$ together with a regularizer $\mathcal{S}(\boldsymbol{u}) = \int_\Omega |\nabla \boldsymbol{u}|^2 \, dx$ which penalizes steep gradients, as $\mathcal{S}$ would not be well-defined in all $\mathcal{V}$. It is common practice to consider $\mathcal{S}$ to be a quadratic term of the form $\mathcal{S}(\boldsymbol{v}) = \frac{1}{2} a(\boldsymbol{v}, \boldsymbol{v})$, where a is a suitable bounded bilinear form. One common example is given by considering the L^2 error as a similarity measure together with the $\boldsymbol{H}_0^1$ norm as a regularizer (equivalently, using $a(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \nabla \boldsymbol{u} \cdot \nabla \boldsymbol{v}$), which yields the following problem:

$$\min_{\boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega)} \left\{ \alpha \int_\Omega |T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x})|^2 \, dx + \int_\Omega |\nabla \boldsymbol{u}|^2 \, dx \right\}, \tag{2.3}$$

with first order conditions given by: Find $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = -\langle \nabla \mathcal{D}(\boldsymbol{u}), \boldsymbol{v} \rangle = -\int_\Omega \nabla T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}))(T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x})) \, dx \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega),$$

where $\nabla \mathcal{D}$ stands for the Fréchet derivative of $\mathcal{D}$. Further details and examples beyond this brief overview can be found in [74].

The present chapter has been mainly motivated by the study of lung regional deformation computed from tomography images of the thorax [29,64]. However, as we will illustrate later on, it is also applicable to related problems such as the image registration of the human brain. The optimal warping, $\boldsymbol{u}$, can be interpreted as a displacement field, from which the gradient $\nabla \boldsymbol{u}$ can be calculated to obtain the strain tensor required to characterize the continuum mechanics framework. The study of deformation from one side has revealed the lungs to present a highly heterogeneous and anisotropic behaviour [8,63], thus providing new deformation-based markers to understand lung diseases [28, 82]. The proposal of the optical flow formulation by Horn & Schunk [61] gave origin to much mathematical analysis at the continuous level, with an increasing interest towards the discrete analysis in an algorithm-specific fashion in [79], in the optimal-control setting within a more classical Galerkin framework [69], and more recently the variational problem was tackled in its primal and mixed formulation in [13].

In fact, the mixed finite element method (MFEM) is a well-established technique which allows to incorporate unknowns of physical interest, such as stress and rotation, and also delivers locking-free schemes in the context of incompressible elasticity (see, e.g., [21, 49]). It also introduces additional difficulties: (i) the new variables increase the dimension of the numerical scheme, making its computational solution more expensive, (ii) the mixed formulation may now possess a saddle-point structure, which results in linear systems of equations that are harder to solve numerically and (iii) only discrete spaces that satisfy the required inf-sup conditions grant a stable scheme, therefore restricting the choices for approximations and also demanding more attention in the analysis of the finite element scheme. For a mixed formulation of DIR with elastic regularization and a target image with Lipschitz gradient, it has been shown that classical existence of solutions is independent of the regularization parameter in the primal case. Furthermore, both primal and mixed schemes give existence and uniqueness for a sufficiently small regularization, and PEERS elements, as well as BDM-$\mathbb{P}_0$ for stress-displacement, are inf-sup stable [13]. In addition, the drawback mentioned in (iii) is alternatively overcome in [13] by using an augmented mixed variational formulation whose discrete analysis does not require the verification of any inf-sup condition, and hence arbitrary finite element subspaces can be employed. More precisely, in this last work a complete numerical analysis of the method was presented, in the particular case of an elastic regularizer and a sum-of-squared-differences similarity

measure with Neumann boundary conditions. Using such conditions is usually physically desirable, as other ones present artificial stress accumulation on the boundaries, thus yielding the difficulty of non-uniqueness to iterative schemes.

In this chapter we aim to generalize the analysis presented in [13] to regularizers that may present a kernel, and to Lipschitz similarity measures. This is performed by splitting weakly the warping with respect to the kernel of the regularizer so that such kernel remains present in the formulation throughout the model, under the assumption of a relationship between the regularizer and the similarity measure commonly known in the inverse problems community as *source condition* [92]. Numerical experiments validating our aforedescribed extended model and showing how it compares to a more traditional formulation are also presented.

The rest of the chapter is organized as follows. In Section 2.2 we derive the new model and analyze its primal formulation at both continuous and discrete levels. The main results, which are obtained by using the Babuška-Brezzi theory and duality arguments, include well-posedness of the continuous and discrete formulations, a priori error estimates, and the respective rates of convergence. Then, in Section 2.3 we introduce and analyze, using basically the same theoretical tools from Section 2.2, an extended dual-mixed formulation in the particular (though very common and useful) case of an elastic energy. Next, in Section 2.4 we explain how to use the traditional time regularization to implement the methods, and provide a suitable bound of the time step guaranteeing convergence. In Section 2.5 we present several numerical experiments illustrating convergence, the capability of the methods to capture translations and rotations, the effect of the added degrees of freedom, the advantage of using the dual-mixed approach in the quasi-incompressible case, and the application to the image registration of the human brain.

2.2 Extended primal formulation in abstract form

In this section we derive an abstract extended model and analyze its continuous and discrete primal formulations.

2.2.1 Setting of the problem

As briefly commented in the Introduction, our problem is posed in the following framework: a Hilbert space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$, a similarity measure $\mathcal{D} : \mathcal{V} \to \mathrm{R}$, a symmetric bounded bilinear form $a : \mathcal{V} \times \mathcal{V} \to \mathrm{R}$ acting as the regularizer, and a positive scalar α. Then, we look for minimizers of the following problem:

$$\min_{v\in\mathcal{V}} \left\{\alpha\mathcal{D}(v) + \frac{1}{2}a(v,v)\right\}. \tag{2.4}$$

The first order conditions yield the following nonlinear problem: Find $u \in \mathcal{V}$ such that

$$a(u,v) = \alpha\, F_u(v) \quad \forall v \in \mathcal{V}\,, \tag{2.5}$$

where, given $w \in \mathcal{V}$, $F_w : \mathcal{V} \to \mathrm{R}$ is the linear functional defined as

$$F_w(v) \;:=\; -\langle \nabla\mathcal{D}(w), v\rangle \qquad \forall\, v \in \mathcal{V}\,, \tag{2.6}$$

which is clearly bounded with $\|F_w\|_{\mathcal{V}'} = \|\nabla\mathcal{D}(w)\|_{\mathcal{V}}$. Next, denoting by Q the kernel of the adjoint of the bounded operator induced by a, which we assume to be non trivial and finite dimensional, and splitting $\mathcal{V}$ as $Q^\perp \oplus Q$, we can rewrite (2.4) equivalently as

$$\min_{(v,\eta)\in Q^\perp\times Q}\left\{\alpha\mathcal{D}(v+\eta)+\frac{1}{2}a(v,v)\right\},$$

and then impose the condition $v\in Q^\perp$ as $\langle v,\xi\rangle = 0 \quad \forall\,\xi\in Q$, to obtain

$$\min_{(v,\eta)\in\mathcal{V}\times Q}\max_{\xi\in Q}\left\{\alpha\mathcal{D}(v+\eta)+\frac{1}{2}a(v,v)+\langle v,\xi\rangle\right\}. \tag{2.7}$$

Finally, to avoid having the nonlinear term $\mathcal{D}$ in more than one equation, we perform the change of variables $v\leftarrow v+\eta$, whence (2.7) becomes

$$\min_{(v,\eta)\in\mathcal{V}\times Q}\max_{\xi\in Q}\left\{\alpha\mathcal{D}(v)+\frac{1}{2}a(v,v)+\langle v-\eta,\xi\rangle\right\}. \tag{2.8}$$

In this formulation a is not elliptic, which gives difficulties in proving the well-posedness of the weak problem. If we consider the form (2.7) with solution $(u,\lambda)\in\mathcal{V}\times Q$, we get that $F_{u+\lambda}(\xi)=0$ for all ξ in Q, which is fully nonlinear and does not give the required control over λ, but on the other hand, form (2.8) gives rise to a non invertible linear operator. This hints the requirement of controlling the component of u in Q, for which, given a positive constant β, we consider the problem

$$\min_{(v,\eta)\in\mathcal{V}\times Q}\max_{\xi\in Q}\left\{\alpha\mathcal{D}(v)+\frac{1}{2}a(v,v)+\langle v-\eta,\xi\rangle+\frac{\beta}{2}\|\eta\|_{\mathcal{V}}^2\right\}. \tag{2.9}$$

We call (2.9) the extended formulation of (2.4). Equivalently, this setting can be obtained by splitting $\mathcal{V}$ in the Euler-Lagrange equations (2.5). First write them as finding $(u,\lambda)\in\mathcal{V}\times Q$ such that

$$\begin{aligned} a(u,v) &= \alpha\,F_u(v) && \forall\, v\in Q^\perp,\\ \langle\lambda,\xi\rangle &= \langle u,\xi\rangle && \forall\,\xi\in Q, \end{aligned} \tag{2.10}$$

and then impose the weak orthogonality by adding a Lagrange multiplier ρ together with the compact perturbation $\beta\langle\lambda,\eta\rangle$ to obtain the extended weak form: Find $(u,\lambda,\rho)\in\mathcal{V}\times Q\times Q$ such that

$$\begin{aligned} a(u,v)+\beta\langle\lambda,\eta\rangle+\langle v-\eta,\rho\rangle &= \alpha F_u(v) && \forall\,(v,\eta)\in\mathcal{V}\times Q,\\ \langle u-\lambda,\xi\rangle &= 0 && \forall\,\xi\in Q. \end{aligned} \tag{2.11}$$

The extended formulation presents two advantages:

- The standard formulation gives rise to a nonlinear compatibility condition for the solution u, namely $0=F_u(\xi) \quad \forall\,\xi\in Q$, which arises after testing (2.5) against elements in Q. Thus, the new variable λ does not affect the compatibility condition. The existence of functions such that this holds is known as the source condition, and is usually stated in the inverse problems community as $\partial\mathcal{D}\perp Q$ [92], which we assume true throughout the chapter.

- Fixed-point schemes arising from such problems impose an undesired orthogonality to the solution, which we refer to as kernel locking. If we let u_n in $\mathcal{V}$ be a previous solution, we get the fixed-point problem of finding u_{n+1} in $\mathcal{V}$ such that

$$a(u_{n+1}, v) = F_{u_n}(v) \qquad \forall v \in \mathcal{V}\,.$$

  This problem does not have a unique solution, so it is common in practice to choose u_{n+1} such that $u_{n+1} \perp Q$. The orthogonal space is closed, and thus if the sequence $\{u_n\}_n$ converges to a solution u, such solution is also orthogonal to Q.

The interpretation of λ in the overall context of the problem is crucial to understand the extent to which it regularizes the problem. For it we first focus on the nonlinear compatibility condition $F_u(\xi) = 0$, also written as $\Pi_Q \nabla \mathcal{D}(u) = 0$, where $\Pi_Q : \mathcal{V} \to Q$ is the orthogonal projection on Q. This condition rises naturally from the extended formulation, and thus it is a necessary condition for the differentiability of $\mathcal{D}$. If the functional does not comply with it, then it is unrelated to a variational principle, so we can add a function $\tilde{\lambda} = \Pi_Q \nabla \mathcal{D}(u) \in Q$ to (2.11) –without λ– such that the compatibility condition holds, that is

$$\begin{aligned} a(u, v) + \langle v - \eta, \rho \rangle &= \alpha(F_u - \tilde{\lambda})(v) \qquad && \forall\, (v, \eta) \in \mathcal{V} \times Q\,, \\ \langle u - \lambda, \xi \rangle &= 0 && \forall\, \xi \in Q\,. \end{aligned} \tag{2.12}$$

We can see that λ indeed takes the desired values by testing the first equation with $v = \eta \in Q$, which gives $F_u(\eta) = \langle \tilde{\lambda}, \eta \rangle \quad \forall\, \eta \in Q$. Note that the same holds if we take the term $\langle \tilde{\lambda}, v \rangle$ to the left hand side and replace it with $\langle \tilde{\lambda}, \eta \rangle$, which means that the compatibilized problem is equivalent to (2.11) if we take $\beta \tilde{\lambda} = \lambda$. In what follows, we show that such choice gives a well posed problem with many numerical advantages, for which we will make the following assumptions

(A1) There exist two positive constants $\widetilde{c}_a$ and $\widetilde{C}_a$ such that

$$\widetilde{c}_a \|v\|_{\mathcal{V}}^2 \le a(v, v) \quad \forall\, v \in Q^{\perp}\,, \qquad \text{and} \qquad |a(w, v)| \le \widetilde{C}_a \|w\|_{\mathcal{V}} \|v\|_{\mathcal{V}} \quad \forall\, w,\, v \in \mathcal{V}\,.$$

(A2) There exists a positive constant $L_{\mathcal{D}}$ and a space $\mathcal{W}$ containing $\mathcal{V}$, such that the embedding $i_{\mathcal{W}} : \mathcal{V} \hookrightarrow \mathcal{W}$ is compact and there holds

$$\|\nabla \mathcal{D}(z_1) - \nabla \mathcal{D}(z_2)\|_{\mathcal{V}} \le L_{\mathcal{D}} \|z_1 - z_2\|_{\mathcal{W}} \quad \forall\, z_1,\, z_2 \in \mathcal{V}\,.$$

(A3) There exists a positive constant $M_{\mathcal{D}}$ such that $\|\nabla \mathcal{D}(w)\|_{\mathcal{V}} \le M_{\mathcal{D}}$ for all $w \in \mathcal{V}$.

2.2.2 Analysis of the continuous formulation

We now show that the extended problem (2.11) has at least one solution, which is stable with respect to the data. For this, we first set the product space $H := \mathcal{V} \times Q$, and let $A : H \times H \to \mathrm{R}$ and $B : H \times Q \to \mathrm{R}$ be the bilinear forms involved in (2.11), that is

$$A((w, \vartheta), (v, \eta)) := a(w, v) + \beta \langle \vartheta, \eta \rangle \qquad \forall\, (w, \vartheta),\, (v, \eta) \in H\,, \tag{2.13}$$

and

$$B((v,\eta),\xi) := \langle v-\eta,\xi\rangle \qquad \forall\,(v,\eta)\in H\,,\quad \forall\,\xi\in Q\,. \tag{2.14}$$

In addition, for each $z\in\mathcal{V}$, we denote by $G_z : H\to \mathrm{R}$ the linear functional given by (cf. (2.6))

$$G_z(v,\eta) := \alpha\, F_z(v) \qquad \forall\,(v,\eta)\in H\,. \tag{2.15}$$

Note here that A, B, and G_z are bounded. In fact, considering the corresponding euclidean norm for the product space H, and denoting the constants $\|A\| := \max\{\widetilde{C}_a,\beta\}$ (cf. (A1)) and $\|B\| := \sqrt{2}$, we easily find, using the Cauchy-Schwarz inequality, that

$$|\,A((w,\vartheta),(v,\eta))\,| \le \|A\|\,\|(w,\vartheta)\|_H\,\|(v,\eta)\|_H \quad\text{and}\quad |\,B((v,\eta),\xi)\,| \le \|B\|\,\|(v,\eta)\|_H\,\|\xi\|_{\mathcal{V}}$$

for all (w,ϑ), $(v,\eta)\in H$, $\forall\,\xi\in Q$. In turn, it is clear from the above definition of G_z and the fact that $F_z\in\mathcal{V}'$ (cf. (2.6)) that $G_z\in H'$ and $\|G_z\| = \alpha\,\|F_z\| = \alpha\,\|\nabla D(z)\|$. According to the previous notations, (2.11) can be rewritten as: Find $\big((u,\lambda),\rho\big)\in H\times Q$ such that

$$\begin{aligned} A\big((u,\lambda),(v,\eta)\big) + B\big((v,\eta),\rho\big) &= G_u(v,\eta) \qquad \forall\,(v,\eta)\in H\,,\\ B\big((u,\lambda),\xi\big) &= 0 \qquad\qquad \forall\,\xi\in Q\,. \end{aligned} \tag{2.16}$$

Then, we introduce the operator $T:\mathcal{V}\to\mathcal{V}$ defined by $T(z) := \widetilde{u}$ for each $z\in\mathcal{V}$, where $\widetilde{u}\in\mathcal{V}$ is the first component of the solution to the problem: Find $\big((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}\big)\in H\times Q$ such that

$$\begin{aligned} A\big((\widetilde{u},\widetilde{\lambda}),(v,\eta)\big) + B\big((v,\eta),\widetilde{\rho}\big) &= G_z(v,\eta) \qquad \forall\,(v,\eta)\in H\,,\\ B\big((\widetilde{u},\widetilde{\lambda}),\xi\big) &= 0 \qquad\qquad \forall\,\xi\in Q\,. \end{aligned} \tag{2.17}$$

We stress here that solving (2.16) is equivalent to seeking a fixed point of T, that is: Find $u\in\mathcal{V}$ such that $T(u) = u$. In the following lemma we show that, for any $z\in\mathcal{V}$, the linear problem (2.17) is well-posed, whence the operator T is well-defined.

Lemma 2.1. *Given z in $\mathcal{V}$, there exists a unique $\big((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}\big)\in H\times Q$ solution to (2.17). Moreover, there exists a positive constant C_T, independent of $\big((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}\big)$, such that the following a priori estimate holds*

$$\|T(z)\|_{\mathcal{V}} \le \|\big((\widetilde{u},\widetilde{\lambda}),\widetilde{\rho}\big)\|_{H\times Q} \le C_T\,\|G_z\|_{H'} = \alpha\, C_T\,\|\nabla\mathcal{D}(z)\|_{\mathcal{V}}\,. \tag{2.18}$$

Proof. In what follows we apply the Babuška-Brezzi theory (cf. [49, Chapter 2]). To this end, we first let N be the kernel of the operator induced by B, that is

$$N = \Big\{(v,\eta)\in H:\quad B((v,\eta),\xi) = 0\quad \forall\,\xi\in Q\Big\}\,,$$

which, according to (2.14), yields $N = \Big\{(v,\eta)\in H:\quad \eta = \varPi_Q v\Big\}$. Then, given $(v,\eta) = (v,\varPi_Q v)\in N$, we split $v = v^{\perp} + \eta \in Q^{\perp}\oplus Q$ and use assumption (A1) to obtain

$$A\big((v,\eta),(v,\eta)\big) = a(v^{\perp},v^{\perp}) + \beta\,\|\eta\|^2_{\mathcal{V}} \ge \widetilde{c}_a\,\|v^{\perp}\|^2_{\mathcal{V}} + \beta\,\|\eta\|^2_{\mathcal{V}} \ge c_a\,\|(v,\eta)\|^2_H\,, \tag{2.19}$$

with $c_a := \min\{\widetilde{c}_a, \frac{\beta}{2}\}$, which gives the N-ellipticity of A. On the other hand, given an arbitrary $\xi \in Q$, we easily see that

$$\sup_{\substack{(v,\eta)\in H\\(v,\eta)\neq(0,0)}} \frac{B((v,\eta),\xi)}{\|(v,\eta)\|_H} \geq \frac{B((0,-\xi),\xi)}{\|(0,-\xi)\|_H} = c_b\,\|\xi\|_{\mathcal{V}}\,, \tag{2.20}$$

with $c_b = 1$, which proves the continuous inf-sup condition for B. In this way, a straightforward application of [49, Theorem 2.3] implies the existence of a unique solution to (2.17) and the corresponding stability estimate (2.18) with a constant C_T depending on c_a, c_b, and $\|A\|$. ■

Now, given $r > 0$, we let $\bar{B}(r)$ be the closed ball of $\mathcal{V}$ centered at the origin with radius r. Then, as a consequence of the previous lemma, we have the following additional result.

Lemma 2.2. *Let $L_{\mathcal{D}}$, $M_{\mathcal{D}}$, and C_T be the constants specified in* (A2), (A3), *and Lemma* 2.1, *respectively, and define $r_0 := \alpha\, C_T\, M_{\mathcal{D}}$. Then, there hold $T(\mathcal{V}) \subseteq \bar{B}(r_0)$ and*

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha\, C_T\, L_{\mathcal{D}}\, \|z_1 - z_2\|_{\mathcal{W}} \qquad \forall\, z_1,\, z_2 \in \mathcal{V}\,. \tag{2.21}$$

Proof. Given $z \in \mathcal{V}$, it readily follows from (2.18) and (A3) that $\|T(z)\|_{\mathcal{V}} \leq \alpha\, C_T\, M_{\mathcal{D}} := r_0$, which proves the required inclusion for T. In turn, the fact that (2.17) is a linear problem guarantees that, given z_1, $z_2 \in \mathcal{V}$, the difference $T(z_1) - T(z_2)$ is the first component of the unique solution of (2.17) when G_z is replaced there by the functional $G_{z_1} - G_{z_2}$. Thus, from the stability estimate (2.18) again, and the Lipschitz-continuity provided by (A2), we deduce that

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha\, C_T\, \|\nabla\mathcal{D}(z_1) - \nabla\mathcal{D}(z_2)\|_{\mathcal{V}} \leq \alpha\, C_T\, L_{\mathcal{D}}\, \|z_1 - z_2\|_{\mathcal{W}}\,,$$

which completes the proof. ■

Having established the above properties of T, we are now in position to provide the main result of this section.

Theorem 2.1. *Let r_0 be the radius defined in the statement of Lemma* 2.2. *Then, problem* (2.16) *admits at least one solution $\big((u,\lambda),\rho\big) \in H \times Q$, with $u \in \bar{B}(r_0)$. Moreover, under the additional assumption $\alpha\, C_T\, L_{\mathcal{D}}\, \|i_{\mathcal{W}}\| < 1$, this solution is unique.*

Proof. We begin by noticing from Lemma 2.2 that certainly $T(\bar{B}(r_0)) \subseteq \bar{B}(r_0)$. Next, it is easy to see from the Lipschitz continuity of T (cf. (2.21)) and the compactness of the embedding $i_{\mathcal{W}} : \mathcal{V} \to \mathcal{W}$ (cf. (A2)) that $\overline{T(\bar{B}(r_0))}$ is compact. Hence, Schauder's fixed-point theorem (cf. [30, Theorem 9.12-1(b)]) implies the existence of a fixed point $u \in \bar{B}(r_0)$ for T, and hence of a solution $\big((u,\lambda),\rho\big) \in H \times Q$ to problem (2.16). Furthermore, it also follows from (2.21) and (A2) that

$$\|T(z_1) - T(z_2)\|_{\mathcal{V}} \leq \alpha\, C_T\, L_{\mathcal{D}}\, \|i_{\mathcal{W}}\|\, \|z_1 - z_2\|_{\mathcal{V}} \qquad \forall\, z_1,\, z_2 \in \mathcal{V}\,,$$

whence the uniqueness in $\mathcal{V}$ is imposed by forcing T to be a contraction and then using the Banach fixed-point theorem, which happens precisely when $\alpha\, C_T\, L_{\mathcal{D}}\, \|i_{\mathcal{W}}\| < 1$. ■

2.2.3 Analysis of the discrete scheme

In this section we consider the Galerkin scheme approximating the solutions of (2.16), establish its well-posedness, derive the associated Céa estimate, and provide the corresponding rates of convergence. For this purpose, we now let $\{\mathcal{V}_h\}_{h>0}$ be a sequence of finite dimensional subspaces of $\mathcal{V}$, where $h>0$ is an index thought as a characteristic meshsize. Then, bearing in mind that Q is finite dimensional, and defining $H_h := \mathcal{V}_h \times Q$, our discrete extended problem reduces to: Find $\big((u_h,\lambda_h),\rho_h\big) \in H_h \times Q$ such that

$$\begin{array}{rcll} A\big((u_h,\lambda_h),(v_h,\eta_h)\big) + B\big((v_h,\eta_h),\rho_h\big) & = & G_{u_h}(v_h,\eta_h) & \forall\,(v_h,\eta_h)\in H_h\,, \\ B\big((u_h,\lambda_h),\xi_h\big) & = & 0 & \forall\,\xi_h \in Q\,. \end{array} \tag{2.22}$$

In turn, we introduce the discrete operator $T_h : \mathcal{V}_h \to \mathcal{V}_h$ given by $T(z_h) := \widetilde{u}_h \quad \forall\, z_h \in \mathcal{V}_h$, where $\widetilde{u}_h$ is the first component of the solution $\big((\widetilde{u}_h,\widetilde{\lambda}_h),\widetilde{\rho}_h\big) \in H_h \times Q$ to (2.22) with G_{z_h} instead of G_{u_h}, that is:

$$\begin{array}{rcll} A\big((\widetilde{u}_h,\widetilde{\lambda}_h),(v_h,\eta_h)\big) + B\big((v_h,\eta_h),\widetilde{\rho}_h\big) & = & G_{z_h}(v_h,\eta_h) & \forall\,(v_h,\eta_h)\in H_h\,, \\ B\big((\widetilde{u}_h,\widetilde{\lambda}_h),\xi_h\big) & = & 0 & \forall\,\xi_h \in Q\,. \end{array} \tag{2.23}$$

As for the continuous case, we emphasize here that solving (2.22) is equivalent to finding $u_h \in \mathcal{V}_h$ such that $T_h(u_h) = u_h$. We start our discrete analysis by proving the well-posedness of (2.23), thus confirming that T_h is well-defined.

Lemma 2.3. *Given $z_h \in \mathcal{V}_h$, there exists a unique $\big((\widetilde{u}_h,\widetilde{\lambda}_h),\widetilde{\rho}_h\big) \in H_h \times Q$ solution to (2.23). Moreover, with the same constant C_T from Lemma 2.1, there holds*

$$\|T_h(z_h)\|_{\mathcal{V}} \,\leq\, \|\big((\widetilde{u}_h,\widetilde{\lambda}_h),\widetilde{\rho}_h\big)\|_{H\times Q} \,\leq\, C_T\,\|G_{z_h}\|_{H'} \,=\, \alpha\,C_T\,\|\nabla\mathcal{D}(z_h)\|_{\mathcal{V}} \,\leq\, \alpha\,C_T\,M_{\mathcal{D}} \,=:\, r_0\,. \tag{2.24}$$

Proof. The proof is analogous to the one shown for the well posedness of problem (2.17) (cf. Lemma 2.1). In fact, we first observe that the discrete kernel N_h of B becomes

$$N_h = \Big\{(v_h,\eta_h)\in H_h : \quad \eta_h \,=\, \Pi_Q v_h\Big\}\,,$$

which is clearly contained in N, and hence the N_h-ellipticity of A follows from that of N, and certainly with the same ellipticity constant c_a. In turn, given $\xi_h \in Q$, the discrete inf-sup condition for B is obtained as in (2.20) by bounding below the involved supremum with $(v_h,\eta_h) = (0,-\xi_h)$, which yields the same resulting constant c_b. In this way, applying now the discrete version of the Babuška-Brezzi theory (cf. [49, Theorem 2.4]), and using from (A3) that $\|\nabla\mathcal{D}(z_h)\| \leq M_{\mathcal{D}}$, we conclude the proof. ∎

Next, given $r>0$, we let $\bar{B}_h(r)$ be the closed ball of $\mathcal{V}_h$ centered at the origin with radius r. Then, the main result concerning the solvability of (2.22), which summarizes the discrete analogues of Lemma 2.2 and Theorem 2.1, is established as follows.

Theorem 2.2. *The discrete problem (2.22) has at least one solution $\big((u_h,\lambda_h),\rho_h\big) \in H_h \times Q$, with $u_h \in \bar{B}_h(r_0)$. Moreover, under the assumption $\alpha\,C_T\,L_{\mathcal{D}}\,\|i_W\| < 1$, this solution is unique.*

Proof. We first notice from (2.24) (cf. Lemma 2.3) that $T_h(\mathcal{V}_h) \subseteq \bar{B}_h(r_0)$, which obviously yields, in particular, $T_h\big(\bar{B}_h(r_0)\big) \subseteq \bar{B}_h(r_0)$. In addition, proceeding as in the proofs of Lemma 2.2 and Theorem 2.1, but certainly using now the linear character of problem (2.23), and employing the stability estimate (2.24), the assumption (A2), and the boundedness of i_W, we easily find that

$$\|T_h(z_{1,h}) - T_h(z_{2,h})\|_{\mathcal{V}} \leq \alpha\, C_T\, L_{\mathcal{D}}\, \|i_W\|\, \|z_{1,h} - z_{2,h}\|_{\mathcal{V}} \qquad \forall\, z_{1,h},\, z_{2,h} \in \mathcal{V}_h\,. \tag{2.25}$$

In this way, the fact that $\bar{B}_h(r_0)$ is clearly a compact and convex subset of $\mathcal{V}_h$, the continuity of $T_h : \bar{B}_h(r_0) \to \bar{B}_h(r_0)$, and a straightforward application of Brouwer's theorem (cf. [30, Theorem 9.9-2]) implies the existence of a fixed point $u_h \in \bar{B}_h(r_0)$ for T_h, and therefore of a solution $\big((u_h, \lambda_h), \rho_h\big) \in H_h \times Q$ to (2.22). Finally, uniqueness in $\mathcal{V}_h$ follows again by forcing T_h to be a contraction. ■

Having proved the existence of solutions for the discrete and continuous problems, we now provide the Céa estimate for the corresponding error. In what follows, given a subspace X_h of a generic Banach space $\big(X, \|\cdot\|_X\big)$, we set

$$\mathrm{dist}(x, X_h) := \inf_{x_h \in X_h} \|x - x_h\|_X \qquad \forall\, x \in X\,.$$

Theorem 2.3. *Assume that* $\alpha\, C_T\, L_{\mathcal{D}}\, \|i_W\| \leq 1 - \delta$, *with* $\delta \in]0,1[$, *and let* $\big((u,\lambda),\rho\big) \in H \times Q$ *and* $\big((u_h,\lambda_h),\rho_h\big) \in H_h \times Q$ *be the unique solutions of* (2.16) *and* (2.22), *respectively. Then, there exists a positive constant* $\widehat{C}$, *depending only on* c_a, c_b, $\|A\|$, *and* $\|B\|$, *and hence independent of* h, *such that*

$$\|\big((u,\lambda),\rho\big) - \big((u_h,\lambda_h),\rho_h\big)\|_{H\times Q} \leq \delta^{-1}\, \widehat{C}\, \mathrm{dist}(u, \mathcal{V}_h)\,. \tag{2.26}$$

Proof. Let $\big((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h\big) \in H_h \times Q$ be the resulting unique solution of the discrete scheme (2.22) when the functional G_{u_h} is replaced there by G_u. In this way, $\big((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h\big) \in H_h \times Q$ constitutes a conforming Galerkin approximation of the unique solution $\big((u,\lambda),\rho\big) \in H \times Q$ to (2.16), and hence the Céa estimate provided by the discrete Babuška-Brezzi theory (cf. [49, Theorems 2.5 and 2.6]) gives the existence of a positive constant $\widehat{C}$, depending only on c_a, c_b, $\|A\|$, and $\|B\|$, such that

$$\|\big((u,\lambda),\rho\big) - \big((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h\big)\|_{H\times Q} \leq \widehat{C}\, \mathrm{dist}\big(\big((u,\lambda),\rho\big), H_h \times Q\big) = \widehat{C}\, \mathrm{dist}\big(u, \mathcal{V}_h\big)\,, \tag{2.27}$$

where the last equality arises from the fact that λ and ρ belong to Q. On the other hand, the linear character of the discrete problem (2.23) readily implies that the difference $\big((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h\big) - \big((u_h,\lambda_h),\rho_h\big)$ is the unique solution of it when G_{z_h} is replaced there by $G_u - G_{u_h}$, and therefore, the a priori estimate (2.24) and the assumption (A2) yield

$$\begin{aligned} \|\big((\widehat{u}_h, \widehat{\lambda}_h), \widehat{\rho}_h\big) - \big((u_h,\lambda_h),\rho_h\big)\| &\leq C_T\, \|G_u - G_{u_h}\|_{H'} \\ &= \alpha\, C_T\, \|\nabla D(u) - \nabla D(u_h)\|_{\mathcal{V}} \leq \alpha\, C_T\, L_{\mathcal{D}}\, \|i_W\|\, \|u - u_h\|_{\mathcal{V}}\,. \end{aligned} \tag{2.28}$$

Finally, the required estimate (2.26) follows easily from triangle inequality, (2.27), (2.28), and the hypothesis $\alpha\, C_T\, L_{\mathcal{D}}\, \|i_W\| \leq 1 - \delta$. ■

We end this section by stressing that the main assumption in Theorem 2.3 is handled by choosing a particular value of δ. Certainly, the closer to 1, the smaller the constant $\delta^{-1}\,\widehat{C}$ in the Céa estimate, but then the hypothesis $\alpha\,C_T\,L_{\mathcal{D}}\,\|i_W\| \leq 1-\delta$, with $1-\delta$ approaching 0, is more demanding on the constants involved. Conversely, the closer to 0, the hypothesis is less restrictive, but then the constant in the Céa estimate blows up. According to the above, it seems more reasonable to consider the midpoint of the range for δ, that is $\delta = 1/2$, which yields the assumption $\alpha\,C_T\,L_{\mathcal{D}}\,\|i_W\| \leq 1/2$, and the corresponding Céa estimate

$$\|\big((u,\lambda),\rho\big) - \big((u_h,\lambda_h),\rho_h\big)\|_{H\times Q} \leq 2\,\widehat{C}\,\mathrm{dist}(u,\mathcal{V}_h)\,. \tag{2.29}$$

2.2.4 The rates of convergence

For the sake of exposition and clearness, we now assume $\mathcal{V} = \mathbf{H}^1(\Omega) := [\mathrm{H}^1(\Omega)]^2$, which is precisely the case of the application to an elastic energy that we report later on in Section 2.5. In there, the unknown u of the abstract analyses from Sections 2.2.1, 2.2.2, 2.2.3, and 2.4, becomes the respective displacement vector $\boldsymbol{u}$ of the elastic material.

Now, let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of $\bar{\Omega}$ made of triangles K with diameter h_K, and define the meshsize $h := \max\{h_K : \quad K \in \mathcal{T}_h\}$, which also acts as the index of $\mathcal{T}_h$. Then, given an integer $k \geq 1$, we denote by $\mathbf{P}_k(K) := [\mathrm{P}_k(K)]^2$ the space of polynomial vectors of degree $\leq k$ on K, introduce the Lagrange finite element subspace of $\mathcal{V}$ of order k

$$\mathcal{V}_h := \Big\{\boldsymbol{v}_h \in \mathbf{H}^1(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\Big\}\,, \tag{2.30}$$

and let $\mathcal{L}_h : \mathbf{C}(\bar{\Omega}) := [C(\bar{\Omega})]^2 \to \mathcal{V}_h$ be its associated interpolation operator. It is well-known that there holds the following approximation property (cf. [20]):

$(\mathbf{AP}_h^{\mathbf{u}})$ for each $m \in \{1,\ldots,k+1\}$ there exists a positive constant C_m such that

$$\mathrm{dist}(\boldsymbol{v},\mathcal{V}_h) \leq \|\boldsymbol{v} - \mathcal{L}_h(\boldsymbol{v})\|_{1,\Omega} \leq C_m\,h^{m-1}\,|\boldsymbol{v}|_{m,\Omega} \qquad \forall\,\boldsymbol{v} \in \mathbf{H}^m(\Omega) := [\mathrm{H}^m(\Omega)]^2\,. \tag{2.31}$$

Then, as a straightforward consequence of Theorem 2.3, (2.29), and $(\mathbf{AP}_h^{\mathbf{u}})$, and analogously to [13], we obtain the following convergence result.

Theorem 2.4. *Assume that* $\alpha\,C_T\,L_{\mathcal{D}}\,\|i_W\| \leq 1/2$, *and let* $\big((\mathbf{u},\lambda),\rho\big) \in H\times Q$ *and* $\big((\mathbf{u}_h,\lambda_h),\rho_h\big) \in H_h\times Q$ *be the unique solutions of* (2.16) *and* (2.22), *respectively. In addition, suppose that* $\mathbf{u} \in \mathbf{H}^m(\Omega)$, *for some* $m \in \{1,\ldots,k+1\}$. *Then, there holds*

$$\|\big((\mathbf{u},\lambda),\rho\big) - \big((\mathbf{u}_h,\lambda_h),\rho_h\big)\|_{H\times Q} \leq 2\,\widehat{C}\,C_m\,h^{m-1}\,|\mathbf{u}|_{m,\Omega}\,. \tag{2.32}$$

Furthermore, in what follows we apply usual duality arguments to derive the rate of convergence for the error $\mathbf{u}-\mathbf{u}_h$, but measured in the weaker norm $\|\cdot\|_{0,\Omega}$. For this purpose, we now simplify the writing of the vector versions of (2.16) and (2.22) by introducing the bilinear form arising after adding the expressions on the left-hand side of either one, that is we let $\mathcal{A} : \big(H\times Q\big)\times\big(H\times Q\big) \longrightarrow \mathrm{R}$ be defined as

$$\mathcal{A}\big((\vec{\mathbf{w}},\chi),(\vec{\boldsymbol{v}},\xi)\big) := A(\vec{\mathbf{w}},\vec{\boldsymbol{v}}) + B(\vec{\boldsymbol{v}},\chi) + B(\vec{\mathbf{w}},\xi)\,,$$

for all $\vec{\mathbf{w}} := (\mathbf{w}, \vartheta)$, $\vec{\boldsymbol{v}} := (\boldsymbol{v}, \eta) \in H := \mathcal{V} \times Q$, for all χ, $\xi \in Q$. In this way, (2.16) and (2.22) can be rewritten, respectively, as: Find $(\vec{\mathbf{u}}, \rho) := \big((\mathbf{u}, \lambda), \rho\big) \in H \times Q$ such that

$$\mathcal{A}\big((\vec{\mathbf{u}}, \rho), (\vec{\boldsymbol{v}}, \xi)\big) \,=\, G_{\mathbf{u}}(\vec{\boldsymbol{v}}) \qquad \forall\, (\vec{\boldsymbol{v}}, \xi) := \big((\boldsymbol{v}, \eta), \xi\big) \in H \times Q\,, \tag{2.33}$$

and: Find $(\vec{\mathbf{u}}_h, \rho_h) := \big((\mathbf{u}_h, \lambda_h), \rho_h\big) \in H_h \times Q$ such that

$$\mathcal{A}\big((\vec{\mathbf{u}}_h, \rho_h), (\vec{\boldsymbol{v}}_h, \xi_h)\big) \,=\, G_{\mathbf{u}_h}(\vec{\boldsymbol{v}}_h) \qquad \forall\, (\vec{\boldsymbol{v}}_h, \xi_h) := \big((\boldsymbol{v}_h, \eta_h), \xi_h\big) \in H_h \times Q\,. \tag{2.34}$$

Note that $\mathcal{A}$ is obviously bounded with a corresponding constant $\|\mathcal{A}\|$ depending on $\|A\|$ and $\|B\|$.

Next, we let $(\vec{\mathbf{w}}, \chi) := \big((\mathbf{w}, \vartheta), \chi\big) \in H \times Q$ be the unique solution, guaranteed by Lemma 2.1 and the symmetry of $\mathcal{A}$, of the continuous problem

$$\mathcal{A}\big((\vec{\boldsymbol{v}}, \xi), (\vec{\mathbf{w}}, \chi)\big) \,=\, \int_\Omega (\mathbf{u} - \mathbf{u}_h) \cdot \boldsymbol{v} \qquad \forall\, (\vec{\boldsymbol{v}}, \xi) := \big((\boldsymbol{v}, \eta), \xi\big) \in H \times Q\,, \tag{2.35}$$

and consider the following regularity assumption:

($\mathbf{RA}^{\mathbf{w}}$) there holds $\mathbf{w} \in \mathrm{H}^2(\Omega)$ and there exists a positive constant $C_{\mathtt{reg}}$, independent of $\mathbf{w}$ and h, such that

$$\|\mathbf{w}\|_{2,\Omega} \,\leq\, C_{\mathtt{reg}}\, \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}\,. \tag{2.36}$$

In addition, throughout the rest of the section we assume $\mathcal{W} = \mathbf{L}^2(\Omega)$ in (A2). Then, we are able to prove the following result, which establishes an extra $O(h)$ for the rate of convergence of $\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}$.

Theorem 2.5. *In addition to the hypotheses of Theorem 2.3 with $\delta = 1/2$, assume ($\mathbf{RA}^{\mathbf{w}}$) and that $\alpha\, L_{\mathcal{D}}\, (C_2 + 1)\, C_{\mathtt{reg}} \,\leq\, 1/2$. Then, there exists a positive constant C_0, depending only on $\|\mathcal{A}\|$, $\widehat{C}$, C_2 (cf. (2.31)), and $C_{\mathtt{reg}}$ (cf. (2.36)), and hence independent of h, such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \,\leq\, C_0\, h\, \mathrm{dist}(\mathbf{u}, \mathcal{V}_h)\,. \tag{2.37}$$

In particular, if $\mathbf{u} \in \mathbf{H}^m(\Omega)$, with $m \in \{1, \ldots, k+1\}$, there holds

$$\|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega} \,\leq\, \widetilde{C}_0\, h^m\, |\mathbf{u}|_{m,\Omega}\,, \tag{2.38}$$

with $\widetilde{C}_0 \,:=\, C_m\, C_0$.

Proof. We begin by taking $(\vec{\boldsymbol{v}}, \xi) = (\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h)$ in (2.35), which yields

$$\|\mathbf{u} - \mathbf{u}_h\|^2_{0,\Omega} \,=\, \mathcal{A}\big((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi)\big)\,,$$

and by recalling from the Sobolev embedding theorem that $\mathbf{H}^2(\Omega) \,\subseteq\, \mathbf{C}(\bar{\Omega})$, which implies, according to ($\mathbf{RA}^{\mathbf{w}}$), that $\mathbf{w} \in \mathbf{C}(\bar{\Omega})$. Thus, adding and subtracting $(\vec{\mathbf{w}}_h, \chi_h) := \big((\mathcal{L}_h(\mathbf{w}), \vartheta), \chi\big) \in H_h \times Q$ in the second component of $\mathcal{A}$, and then using (2.33), (2.34), and the definition of the functional $G_{\mathbf{z}}$ (cf. vector version of (2.6) and (2.15)), we obtain from the foregoing equation

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|^2_{0,\Omega} \,&=\, \mathcal{A}\big((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)\big) + \mathcal{A}\big((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}_h, \chi_h)\big) \\
&=\, \mathcal{A}\big((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)\big) + G_{\mathbf{u}}(\vec{\mathbf{w}}_h) - G_{\mathbf{u}_h}(\vec{\mathbf{w}}_h) \\
&=\, \mathcal{A}\big((\vec{\mathbf{u}}, \rho) - (\vec{\mathbf{u}}_h, \rho_h), (\vec{\mathbf{w}}, \chi) - (\vec{\mathbf{w}}_h, \chi_h)\big) + \alpha\, \langle \nabla \mathcal{D}(\mathbf{u}_h) - \nabla \mathcal{D}(\mathbf{u}), \mathbf{w}_h \rangle\,.
\end{aligned} \tag{2.39}$$

Next, employing now the boundedness of $\mathcal{A}$, the assumption (A2), the estimate (2.29), the approximation property (2.31) for $\mathcal{L}_h$, and the regularity bound (2.36), we deduce from (2.39) that

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}^2 &\leq \|\mathcal{A}\|\,\|(\vec{\mathbf{u}},\rho)-(\vec{\mathbf{u}}_h,\rho_h)\|\,\|\mathbf{w}-\mathcal{L}_h(\mathbf{w})\|_{1,\Omega} + \alpha\,L_{\mathcal{D}}\,\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,\|\mathbf{w}_h\|_{1,\Omega}\\ &\leq \|\mathcal{A}\|\,2\,\widehat{C}\,\mathrm{dist}(\mathbf{u},\mathcal{V}_h)\,C_2\,h\,|\mathbf{w}|_{2,\Omega} + \alpha\,L_{\mathcal{D}}\,\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,\|\mathbf{w}_h\|_{1,\Omega} \qquad (2.40)\\ &\leq C\,h\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,\mathrm{dist}(\mathbf{u},\mathcal{V}_h) + \alpha\,L_{\mathcal{D}}\,\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,\|\mathbf{w}_h\|_{1,\Omega}\,,\end{aligned}$$

with $C := 2\,\|\mathcal{A}\|\,\widehat{C}\,C_2\,C_{\mathtt{reg}}$, which yields

$$\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega} \leq C\,h\,\mathrm{dist}(\mathbf{u},\mathcal{V}_h) + \alpha\,L_{\mathcal{D}}\,\|\mathbf{w}_h\|_{1,\Omega}\,. \qquad (2.41)$$

In turn, applying again (2.31) and (2.36), and assuming for sake of simplicity that $h \leq 1$, we find that

$$\|\mathbf{w}_h\|_{1,\Omega} \leq \|\mathbf{w}-\mathbf{w}_h\|_{1,\Omega} + \|\mathbf{w}\|_{1,\Omega} \leq \big(C_2\,h+1\big)\,\|\mathbf{w}\|_{2,\Omega} \leq (C_2+1)C_{\mathtt{reg}}\,\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,,$$

which, replaced back into (2.41), leads to (2.37) with $C_0 = 2C$. Finally, it is straightforward to see that (2.31) and (2.37) imply (2.38), which completes the proof. ■

As a particular case of (2.38), we notice that for $k=1$ and $\mathbf{u}\in\mathbf{H}^2(\Omega)$ there holds the error estimate $\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega} \leq \widetilde{C}_0\,h^2\,|\mathbf{u}|_{2,\Omega}$, that is $\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega} = O(h^2)$. This rate of convergence will be illustrated below in Section 2.5 with some numerical results.

2.3 Extended mixed formulation and application to elastic energies

In this section we present and analyse a dual-mixed formulation of problem (2.16) in the particular case of an elastic energy. In this regard, we find it important to remark in advance that the setting and analysis to be considered and developed, respectively, in what follows, do not correspond to a straightforward application of those from Sections 2.2.1 and 2.2.2, which basically refer to a primal formulation, but to a modification of them yielding the associated extended mixed approach to be employed here. Still, the point of departure for this novel model is the use of an elastic regularizer with Neumann boundary conditions, which presents a non-trivial kernel.

2.3.1 Setting of the problem

Let $\mathcal{C}:\mathbb{L}^2(\Omega)\longrightarrow\mathbb{L}^2(\Omega)$ be the Hooke operator defined by

$$\mathcal{C}\boldsymbol{\tau} := \lambda_s\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbb{I} + 2\mu_s\,\boldsymbol{\tau} \qquad \forall\,\boldsymbol{\tau}\in\mathbb{L}^2(\Omega)\,, \qquad (2.42)$$

where λ_s and μ_s are the associated Lamé parameters, and let $\varepsilon(\mathbf{u}) := \frac{1}{2}\Big\{(\nabla\mathbf{u})+(\nabla\mathbf{u})^{\mathtt{t}}\Big\}$ be the strain rate tensor, also known as the symmetric component of $\nabla\mathbf{u}$. Then, letting $\mathcal{V} := \mathbf{H}^1(\Omega)$, the bilinear form a from Section 2.2 is defined as

$$a(\mathbf{w},\boldsymbol{v}) := \int_\Omega \mathcal{C}\varepsilon(\mathbf{w}) : \varepsilon(\boldsymbol{v}) \qquad \forall\,\mathbf{w},\,\boldsymbol{v}\in\mathcal{V}\,, \qquad (2.43)$$

and its kernel Q is given by the subspace of $\mathcal{V}$ determined by the rigid motions, that is

$$Q := \left\langle \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right\} \right\rangle . \tag{2.44}$$

Next, we introduce the auxiliary unknown $\boldsymbol{\sigma} := \mathcal{C}\,\varepsilon(\mathbf{u})$, and observe that there holds

$$\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathtt{t}} \quad \text{and} \quad \mathcal{C}^{-1}\,\boldsymbol{\sigma} = \nabla \mathbf{u} - \boldsymbol{\Phi} \quad \text{in} \quad \Omega\,, \tag{2.45}$$

where the rotation $\boldsymbol{\Phi} := \frac{1}{2}\left\{(\nabla \mathbf{u}) - (\nabla \mathbf{u})^{\mathtt{t}}\right\}$ is considered as a further unknown as well. In addition, we look for rigid motions $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ such that

$$-\,\mathbf{div}\,\boldsymbol{\sigma} + \boldsymbol{\rho} = -\alpha\,\nabla \mathcal{D}(\mathbf{u})\,, \quad \boldsymbol{\lambda} = \varPi_Q\,\mathbf{u}\,, \quad \text{and} \quad \boldsymbol{\rho} = \beta\,\boldsymbol{\lambda} \quad \text{in} \quad \Omega\,, \tag{2.46}$$

where α and β are the analogue parameters from Section 2.2, and incorporate the Neumann boundary condition

$$\boldsymbol{\sigma}\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \varGamma\,. \tag{2.47}$$

We now proceed to derive the variational formulation of (2.45), (2.46), and (2.47). In fact, recalling that the definition of $\mathbb{H}(\mathbf{div};\Omega)$ was provided in Section 2.1, we first define the spaces

$$\mathbb{H}_0(\mathbf{div};\Omega) := \left\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \varGamma \right\},$$

and

$$\mathbb{L}^2_{\mathtt{skew}}(\Omega) := \left\{ \boldsymbol{\Psi} \in \mathbb{L}^2(\Omega) : \quad \boldsymbol{\Psi}^{\mathtt{t}} = -\,\boldsymbol{\Psi} \right\},$$

noting in advance that $\boldsymbol{\sigma}$ and $\boldsymbol{\Phi}$ will be sought in $\mathbb{H}_0(\mathbf{div};\Omega)$ and $\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, respectively. Thus, performing the tensor inner product of the second equation in (2.45) with an arbitrary $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$, integrating by parts, and using the boundary condition that holds for $\boldsymbol{\tau}$, we obtain

$$\int_\Omega \mathcal{C}^{-1}\,\boldsymbol{\sigma} : \boldsymbol{\tau} + \int_\Omega \boldsymbol{\Phi} : \boldsymbol{\tau} + \int_\Omega \mathbf{u} \cdot \mathbf{div}\,\boldsymbol{\tau} = 0 \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)\,. \tag{2.48}$$

In addition, testing the first and third equations in (2.46) against $\boldsymbol{v} \in \mathbf{L}^2(\Omega)$ and $\boldsymbol{\xi} \in Q$, respectively, and rewriting the second equation in (2.46) as the equivalent orthogonality condition, we find that

$$\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\,\boldsymbol{\sigma} - \int_\Omega \boldsymbol{\rho} \cdot \boldsymbol{v} = \alpha \int_\Omega \nabla \mathcal{D}(\mathbf{u}) \cdot \boldsymbol{v} \quad \forall\, \boldsymbol{v} \in \mathbf{L}^2(\Omega)\,, \tag{2.49}$$

$$\int_\Omega (\boldsymbol{\rho} - \beta\,\boldsymbol{\lambda}) \cdot \boldsymbol{\xi} = 0 \qquad \forall\, \boldsymbol{\xi} \in Q\,, \tag{2.50}$$

and

$$\int_\Omega (\boldsymbol{\lambda} - \mathbf{u}) \cdot \boldsymbol{\eta} = 0 \qquad \forall\, \boldsymbol{\eta} \in Q\,. \tag{2.51}$$

Finally, the symmetry of $\boldsymbol{\sigma}$ (first equation in (2.45)) is imposed weakly as

$$\int_\Omega \boldsymbol{\Psi} : \boldsymbol{\sigma} = 0 \qquad \forall\, \boldsymbol{\Psi} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega)\,. \tag{2.52}$$

Therefore, incorporating (2.51) into (2.48), and adding (2.49), (2.50), and (2.52), we arrive at the following dual-mixed variational formulation of (2.45) - (2.47): Find $\vec{\boldsymbol{\sigma}} := (\boldsymbol{\sigma}, \boldsymbol{\rho}) \in \mathbf{H} := \mathbb{H}_0(\mathbf{div}; \Omega) \times Q$ and $\vec{\mathbf{u}} := (\mathbf{u}, \boldsymbol{\Phi}, \boldsymbol{\lambda}) \in \mathbf{Q} := \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathtt{skew}}(\Omega) \times Q$, such that

$$\begin{aligned} \mathbf{a}(\vec{\boldsymbol{\sigma}}, \vec{\boldsymbol{\tau}}) + \mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{u}}) &= 0 \qquad && \forall\, \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}\,, \\ \mathbf{b}(\vec{\boldsymbol{\sigma}}, \vec{\boldsymbol{v}}) - \mathbf{c}(\vec{\mathbf{u}}, \vec{\boldsymbol{v}}) &= \alpha\, \mathbf{F}_{\mathbf{u}}(\vec{\boldsymbol{v}}) \qquad && \forall\, \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}\,, \end{aligned} \tag{2.53}$$

where $\mathbf{a} : \mathbf{H} \times \mathbf{H} \to \mathrm{R}$, $\mathbf{b} : \mathbf{H} \times \mathbf{Q} \to \mathrm{R}$, and $\mathbf{c} : \mathbf{Q} \times \mathbf{Q} \to \mathrm{R}$, are the bilinear forms defined as

$$\mathbf{a}(\vec{\boldsymbol{\zeta}}, \vec{\boldsymbol{\tau}}) := \int_\Omega \mathcal{C}^{-1}\, \boldsymbol{\zeta} : \boldsymbol{\tau}\,, \tag{2.54}$$

$$\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}}) := \int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\tau} + \int_\Omega \boldsymbol{\Psi} : \boldsymbol{\tau} + \int_\Omega (\boldsymbol{\xi} - \boldsymbol{v}) \cdot \boldsymbol{\eta}\,, \tag{2.55}$$

and

$$\mathbf{c}(\vec{\mathbf{w}}, \vec{\boldsymbol{v}}) := \beta \int_\Omega \boldsymbol{\vartheta} \cdot \boldsymbol{\xi}\,, \tag{2.56}$$

for all $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta}, \chi)$, $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}$, for all $\vec{\mathbf{w}} := (\mathbf{w}, \boldsymbol{\Upsilon}, \boldsymbol{\vartheta})$, $\vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}$. In turn, given $\vec{\mathbf{w}} := (\mathbf{w}, \boldsymbol{\Upsilon}, \boldsymbol{\vartheta}) \in \mathbf{Q}$, the linear functional $\mathbf{F}_{\mathbf{w}} : \mathbf{Q} \to \mathbb{R}$ is defined by

$$\mathbf{F}_{\mathbf{w}}(\vec{\boldsymbol{v}}) := \int_\Omega \nabla \mathcal{D}(\mathbf{w}) \cdot \boldsymbol{v} \qquad \forall\, \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}\,. \tag{2.57}$$

At this point we stress that $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are all bounded bilinear forms with respect to the usual norms of the product spaces $\mathbf{H}$ and $\mathbf{Q}$, that is

$$\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}} := \Big\{ \|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} + \|\boldsymbol{\eta}\|^2_{0,\Omega} \Big\}^{1/2} \qquad \forall\, \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}\,,$$

and

$$\|\vec{\boldsymbol{v}}\|_{\mathbf{Q}} := \Big\{ \|\boldsymbol{v}\|^2_{0,\Omega} + \|\boldsymbol{\Psi}\|^2_{0,\Omega} + \|\boldsymbol{\xi}\|^2_{0,\Omega} \Big\}^{1/2} \qquad \forall\, \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}\,.$$

Moreover, $\mathbf{a}$ and $\mathbf{c}$ are both symmetric and positive semi-definite, that is

$$\mathbf{a}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{\tau}}) \geq 0 \quad \forall\, \vec{\boldsymbol{\tau}} \in \mathbf{H} \quad \text{and} \quad \mathbf{c}(\vec{\boldsymbol{v}}, \vec{\boldsymbol{v}}) \geq 0 \quad \forall\, \vec{\boldsymbol{v}} \in \mathbf{Q}\,. \tag{2.58}$$

In addition, it is clear that $\mathbf{F}_{\mathbf{w}}$ is bounded for each $\mathbf{w} \in \mathbf{L}^2(\Omega)$.

2.3.2 Analysis of the continuous formulation

In order to study the solvability of (2.53), and similarly to the analysis in Section 2.2.2, we now introduce the operator $\mathbf{T} : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ defined by $\mathbf{T}(\mathbf{z}) := \underline{\mathbf{u}}$ for each $\mathbf{z} \in \mathbf{L}^2(\Omega)$, where $\vec{\underline{\boldsymbol{\sigma}}} := (\underline{\boldsymbol{\sigma}}, \underline{\boldsymbol{\rho}}) \in \mathbf{H}$ and $\vec{\underline{\mathbf{u}}} := (\underline{\mathbf{u}}, \underline{\boldsymbol{\Phi}}, \underline{\boldsymbol{\lambda}}) \in \mathbf{Q}$ are such that

$$\begin{aligned} \mathbf{a}(\vec{\underline{\boldsymbol{\sigma}}}, \vec{\boldsymbol{\tau}}) + \mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\underline{\mathbf{u}}}) &= 0 \qquad && \forall\, \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}\,, \\ \mathbf{b}(\vec{\underline{\boldsymbol{\sigma}}}, \vec{\boldsymbol{v}}) - \mathbf{c}(\vec{\underline{\mathbf{u}}}, \vec{\boldsymbol{v}}) &= \alpha\, \mathbf{F}_{\mathbf{z}}(\vec{\boldsymbol{v}}) \qquad && \forall\, \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}\,. \end{aligned} \tag{2.59}$$

We remark here that solving (2.53) is equivalent to seeking a fixed point of $\mathbf{T}$, that is: Find $\mathbf{u} \in \mathbf{L}^2(\Omega)$ such that $\mathbf{T}(\mathbf{u}) = \mathbf{u}$. The following abstract result will allow us to show below that, given $\mathbf{z} \in \mathbf{L}^2(\Omega)$, the linear problem (2.59) is well-posed, thus confirming that the operator $\mathbf{T}$ is well-defined.

Theorem 2.6. *Let* $\mathbf{H}$ *and* $\mathbf{Q}$ *be real Hilbert spaces, and let* $\mathbf{a} : \mathbf{H} \times \mathbf{H} \to \mathrm{R}$, $\mathbf{b} : \mathbf{H} \times \mathbf{Q} \to \mathrm{R}$, *and* $\mathbf{c} : \mathbf{Q} \times \mathbf{Q} \to \mathrm{R}$ *be bounded bilinear forms with induced bounded linear operators* $\mathbf{A} : \mathbf{H} \to \mathbf{H}'$, $\mathbf{B} : \mathbf{H} \to \mathbf{Q}'$, $\mathbf{B}^{\mathtt{t}} : \mathbf{Q} \to \mathbf{H}'$, *and* $\mathbf{C} : \mathbf{Q} \to \mathbf{Q}'$, *defined, respectively, by the identities*

$$\mathbf{A}(\boldsymbol{\zeta})(\boldsymbol{\tau}) := \mathbf{a}(\boldsymbol{\zeta}, \boldsymbol{\tau}) \qquad \forall\, \boldsymbol{\zeta},\, \boldsymbol{\tau} \in \mathbf{H}\,,$$

$$\mathbf{B}(\boldsymbol{\tau})(\boldsymbol{v}) = \mathbf{B}^{\mathtt{t}}(\boldsymbol{v})(\boldsymbol{\tau}) := \mathbf{b}(\boldsymbol{\tau}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{\tau} \in \mathbf{H},\ \forall\, \boldsymbol{v} \in \mathbf{Q}\,,$$

$$\mathbf{C}(\mathbf{w})(\boldsymbol{v}) := \mathbf{c}(\mathbf{w}, \boldsymbol{v}) \qquad \forall\, \mathbf{w},\, \boldsymbol{v} \in \mathbf{Q}\,.$$

In turn, let $\mathbf{K} = N(\mathbf{B})$ *and* $\mathbf{V} = N(\mathbf{B}^{\mathtt{t}})$, *and assume the following hypotheses:*

i) $\mathbf{a}$ *and* $\mathbf{c}$ *are symmetric and positive semi-definite.*

ii) $\mathbf{a}$ *is* $\mathbf{K}$*-elliptic, that is there exists a positive constant* $\alpha_{\mathbf{K}}$ *such that*

$$\mathbf{a}(\boldsymbol{\tau}, \boldsymbol{\tau}) \,\geq\, \alpha_{\mathbf{K}}\, \|\boldsymbol{\tau}\|^2_{\mathbf{H}} \qquad \forall\, \boldsymbol{\tau} \in \mathbf{K}\,.$$

iii) $R(\mathbf{B})$ *is closed, that is there exists a positive constant* $\beta_{\mathbf{B}}$ *such that*

$$\sup_{\substack{\boldsymbol{\tau} \in \mathbf{H} \\ \boldsymbol{\tau} \neq \mathbf{0}}} \frac{\mathbf{b}(\boldsymbol{\tau}, \boldsymbol{v})}{\|\boldsymbol{\tau}\|_{\mathbf{H}}} \,\geq\, \beta_{\mathbf{B}}\, \|\boldsymbol{v}\|_{\mathbf{Q}} \qquad \forall\, \boldsymbol{v} \in \mathbf{V}^{\perp}\,,$$

or equivalently

$$\sup_{\substack{\boldsymbol{v} \in \mathbf{Q} \\ \boldsymbol{v} \neq \mathbf{0}}} \frac{\mathbf{b}(\boldsymbol{\tau}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\mathbf{Q}}} \,\geq\, \beta_{\mathbf{B}}\, \|\boldsymbol{\tau}\|_{\mathbf{H}} \qquad \forall\, \boldsymbol{\tau} \in \mathbf{K}^{\perp}\,.$$

iv) $\mathbf{c}$ *is* $\mathbf{V}$*-elliptic, that is there exists a positive constant* $\gamma_{\mathbf{V}}$ *such that*

$$\mathbf{c}(\boldsymbol{v}, \boldsymbol{v}) \,\geq\, \gamma_{\mathbf{V}}\, \|\boldsymbol{v}\|^2_{\mathbf{Q}} \qquad \forall\, \boldsymbol{v} \in \mathbf{V}\,.$$

Then, for each pair $(\mathbf{F}, \mathbf{G}) \in \mathbf{H}' \times \mathbf{Q}'$ *there exists a unique* $(\boldsymbol{\sigma}, \mathbf{u}) \in \mathbf{H} \times \mathbf{Q}$ *solution to*

$$\begin{aligned} \mathbf{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \,+\, \mathbf{b}(\boldsymbol{\tau}, \mathbf{u}) \;&=\; \mathbf{F}(\boldsymbol{\tau}) \qquad \forall\, \boldsymbol{\tau} \in \mathbf{H}\,, \\ \mathbf{b}(\boldsymbol{\sigma}, \boldsymbol{v}) \,-\, \mathbf{c}(\mathbf{u}, \boldsymbol{v}) \;&=\; \mathbf{G}(\boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in \mathbf{Q}\,. \end{aligned} \tag{2.60}$$

In addition, there exists a positive constant C, *depending only on* $\alpha_{\mathbf{K}}$, $\beta_{\mathbf{B}}$, $\gamma_{\mathbf{V}}$, $\|\mathbf{A}\|$, *and* $\|\mathbf{C}\|$, *such that*

$$\|\boldsymbol{\sigma}\|_{\mathbf{H}} \,+\, \|\mathbf{u}\|_{\mathbf{Q}} \,\leq\, C \left\{ \|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'} \right\}.$$

Proof. See [18, Theorem 4.3.1]. ∎

We now apply Theorem 2.6 to show the well-posedness of (2.59), and hence the well-definiteness of the associated operator $\mathbf{T}$. To this end, we first rewrite the bilinear form $\mathbf{b}$ (cf. (2.55)) as

$$\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}}) \,:=\, \int_{\Omega} \boldsymbol{v} \cdot \{\mathbf{div}\, \boldsymbol{\tau} - \boldsymbol{\eta}\} + \int_{\Omega} \boldsymbol{\Psi} : \boldsymbol{\tau} + \int_{\Omega} \boldsymbol{\xi} \cdot \boldsymbol{\eta}\,, \tag{2.61}$$

for all $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H}$, for all $\vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q}$, from which we deduce that the null space of its induced operator $\mathbf{B} : \mathbf{H} \to \mathbf{Q}'$ is given by

$$\mathbf{K} = N(\mathbf{B}) := \left\{ \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H} : \quad \mathbf{div}\, \boldsymbol{\tau} - \boldsymbol{\eta} = 0, \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}, \quad \text{and} \quad \boldsymbol{\eta} = \mathbf{0} \right\},$$

which yields

$$\mathbf{K} = \left\{ \vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{\eta}) \in \mathbf{H} : \quad \mathbf{div}\, \boldsymbol{\tau} = \mathbf{0}, \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}, \quad \text{and} \quad \boldsymbol{\eta} = \mathbf{0} \right\}. \tag{2.62}$$

Similarly, looking at the original definition (2.55) of $\mathbf{b}$, we readily find that

$$\mathbf{V} = N(\mathbf{B}^{\mathtt{t}}) := \left\{ \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q} : \quad \int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\tau} + \int_\Omega \boldsymbol{\Psi} : \boldsymbol{\tau} = 0 \right.$$
$$\left. \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \quad \text{and} \quad \boldsymbol{\xi} = \Pi_Q \boldsymbol{v} \right\},$$

from which, rewriting the expression involving $\boldsymbol{\tau}$ in the distributional sense, we are lead to

$$\mathbf{V} = \left\{ \vec{\boldsymbol{v}} := (\boldsymbol{v}, \boldsymbol{\Psi}, \boldsymbol{\xi}) \in \mathbf{Q} : \quad \boldsymbol{\Psi} = \nabla \boldsymbol{v} \quad \text{in} \quad \mathcal{D}'(\Omega) \quad \text{and} \quad \boldsymbol{\xi} = \Pi_Q \boldsymbol{v} \right\}.$$

Moreover, the fact that $\nabla \boldsymbol{v} = \boldsymbol{\Psi} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega)$ implies that $\varepsilon(\boldsymbol{v}) = \mathbf{0}$, that is $\boldsymbol{v}$ lies in the subspace of rigid motions Q, and therefore $\mathbf{V} \subseteq \mathbf{V}_0$, where

$$\mathbf{V}_0 := \left\{ \vec{\mathbf{q}} := (\mathbf{q}, \nabla \mathbf{q}, \mathbf{q}) \in \mathbf{Q} : \quad \mathbf{q} \in Q \right\}. \tag{2.63}$$

Conversely, it is easy to see that, given $\vec{\mathbf{q}} \in \mathbf{V}_0$, there holds $\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\mathbf{q}}) = 0$ for all $\vec{\boldsymbol{\tau}} \in \mathbf{H}$ (see also (2.67) below), which shows that $\mathbf{V}_0 \subseteq \mathbf{V}$, and hence $\mathbf{V} = \mathbf{V}_0$.

We now aim to show the $\mathbf{K}$-ellipticity of $\mathbf{a}$, for which we first state two preliminary results that are based on the decomposition $\mathbb{H}(\mathbf{div}; \Omega) := \widetilde{\mathbb{H}}(\mathbf{div}; \Omega) \oplus \mathrm{R}\mathbb{I}$, where

$$\widetilde{\mathbb{H}}(\mathbf{div}; \Omega) := \left\{ \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega) : \quad \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) = 0 \right\}.$$

In fact, we have the following lemmas, in which we use that for each $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega)$ there exist unique $\boldsymbol{\tau}_0 \in \widetilde{\mathbb{H}}(\mathbf{div}; \Omega)$ and $d \in \mathrm{R}$ such that $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\,\mathbb{I} \in \mathbb{H}(\mathbf{div}; \Omega)$.

Lemma 2.4. *There exists a positive constant c_1, depending only on Ω, such that*

$$\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega} + \|\mathbf{div}(\boldsymbol{\tau})\|^2_{0,\Omega} \geq c_1 \|\boldsymbol{\tau}_0\|^2_{0,\Omega} \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega). \tag{2.64}$$

Proof. See [21, Proposition 3.1 of Chapter IV] or [49, Lemma 2.3]. ∎

Lemma 2.5. *There exists a positive constant c_2, depending only on Ω, such that*

$$\|\boldsymbol{\tau}_0\|^2_{\mathbf{div};\Omega} \geq c_2 \|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega). \tag{2.65}$$

Proof. See [47, Lemma 2.2] or [49, Lemma 2.5]. ∎

Then, the announced result for $\mathbf{a}$ is established as follows.

Lemma 2.6. *There exists a constant $\alpha_{\mathbf{K}} > 0$, independent of the Lamé parameter λ_s, such that*

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) \geq \alpha_{\mathbf{K}} \|\vec{\boldsymbol{\tau}}\|^2_{\mathbf{H}} \qquad \forall\, \vec{\boldsymbol{\tau}} \in \mathbf{K}\,.$$

Proof. We begin by recalling from [49, Section 2.4.3] that in the present 2D case the inverse $\mathcal{C}^{-1}$ of the Hooke tensor $\mathcal{C}$ becomes

$$\mathcal{C}^{-1}\boldsymbol{\tau} = \frac{1}{2\mu_s}\,\boldsymbol{\tau} - \frac{\lambda_s}{4\mu_s(\lambda_s+\mu_s)}\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbb{I} \qquad \forall\, \boldsymbol{\tau} \in \mathbb{L}^2(\Omega)\,,$$

which, after some algebraic manipulations, yields (cf. [49, eqs. (2.48) and (2.52)])

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) = \int_\Omega \mathcal{C}^{-1}\boldsymbol{\tau} : \boldsymbol{\tau} = \frac{1}{2\mu_s}\,\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega} + \frac{1}{4(\lambda_s+\mu_s)}\,\|\mathrm{tr}(\boldsymbol{\tau})\|^2_{0,\Omega} \geq \frac{1}{2\mu_s}\,\|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{0,\Omega}$$

for all $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}$. In particular, given $\vec{\boldsymbol{\tau}} \in \mathbf{K}$, that is $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$ such that $\mathbf{div}(\boldsymbol{\tau}) = 0$ and $\boldsymbol{\tau} = \boldsymbol{\tau}^{\mathtt{t}}$, it follows from the foregoing inequality and straightforward applications of Lemmas 2.4 and 2.5, that

$$\mathbf{a}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{\tau}}) \geq \frac{c_1}{2\mu_s}\,\|\boldsymbol{\tau}_0\|^2_{0,\Omega} = \frac{c_1}{2\mu_s}\,\|\boldsymbol{\tau}_0\|^2_{\mathbf{div};\Omega} \geq \frac{c_1 c_2}{2\mu_s}\,\|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} = \frac{c_1 c_2}{2\mu_s}\,\|\vec{\boldsymbol{\tau}}\|^2_{\mathbf{H}}\,,$$

which completes the proof with the constant $\alpha_{\mathbf{K}} := \frac{c_1 c_2}{2\mu_s}$. ∎

A preliminary continuous inf-sup condition for the bilinear form $\mathbf{b}$ (cf. (2.55)), in which the space $\mathbf{V}_0$ as such (cf. (3.19)) plays a key role, is established next.

Lemma 2.7. *There exists a positive constant $\beta_{\mathbf{B}}$, independent of the Lamé parameters, such that*

$$\sup_{\substack{\vec{\boldsymbol{\tau}}\in\mathbf{H}\\ \vec{\boldsymbol{\tau}}\neq\mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}}\,\mathrm{dist}(\vec{\boldsymbol{v}},\mathbf{V}_0) \qquad \forall\, \vec{\boldsymbol{v}} \in \mathbf{Q}\,. \tag{2.66}$$

Proof. While we already know that $\mathbf{V}_0 = \mathbf{V}$, the inclusion $\mathbf{V}_0 \subseteq \mathbf{V} = N(\mathbf{B}^{\mathtt{t}})$ suffices to realize that (2.66) trivially holds for $\vec{\boldsymbol{v}} \in \mathbf{V}_0$, and therefore in what follows we prove for $\vec{\boldsymbol{v}} := (\boldsymbol{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}\setminus\mathbf{V}_0$. Indeed, given an arbitrary $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}$, we first use the orthogonal decomposition $\boldsymbol{v} = (\boldsymbol{v}-\Pi_Q\boldsymbol{v})+\Pi_Q\boldsymbol{v} \in Q^\perp \oplus Q$, and then integrate by parts the expression $\int_\Omega \Pi_Q\boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\tau}$, to deduce from (2.55) that there holds

$$\mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\boldsymbol{v}}) := \int_\Omega (\boldsymbol{v}-\Pi_Q\boldsymbol{v})\cdot\mathbf{div}\,\boldsymbol{\tau} + \int_\Omega (\boldsymbol{\Psi}-\nabla\Pi_Q\boldsymbol{v}) : \boldsymbol{\tau} + \int_\Omega (\boldsymbol{\xi}-\Pi_Q\boldsymbol{v})\cdot\boldsymbol{\eta}\,. \tag{2.67}$$

Next, we proceed as in the proof of [50, Lemma 3.4]. In fact, assuming that $\boldsymbol{v}-\Pi_Q\boldsymbol{v} \neq \mathbf{0}$, we let $\boldsymbol{\zeta} := \varepsilon(\mathbf{z})$ in Ω, where $\mathbf{z} \in \mathbf{H}^1(\Omega)$ is the unique solution, up to an element in Q, of the problem

$$\mathbf{div}(\varepsilon(\mathbf{z})) = \boldsymbol{v}-\Pi_Q\boldsymbol{v} \quad \text{in} \quad \Omega\,, \quad \varepsilon(\mathbf{z})\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \Gamma\,. \tag{2.68}$$

Note that the compatibility condition required by this Neumann problem is satisfied thanks to the orthogonality relation $\int_\Omega(\boldsymbol{v}-\Pi_Q\boldsymbol{v})\cdot\mathbf{q} = 0 \;\; \forall\,\mathbf{q} \in Q$. Thus, it is clear that $\boldsymbol{\zeta} \in \mathbb{H}_0(\mathbf{div};\Omega)$ with $\mathbf{div}(\boldsymbol{\zeta}) = \boldsymbol{v}-\Pi_Q\boldsymbol{v}$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}^{\mathtt{t}}$ in Ω. In addition, the corresponding continuous dependence result

for (2.68) guarantees the existence of a positive constant C_N, independent of $\boldsymbol{v} - \Pi_Q \boldsymbol{v}$, such that $\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega} \leq C_N \, \|\boldsymbol{v} - \Pi_Q \boldsymbol{v}\|_{0,\Omega}$. In this way, defining $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta}, \mathbf{0}) \in \mathbf{H}$, it readily follows that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} = \frac{\|\boldsymbol{v} - \Pi_Q \boldsymbol{v}\|^2_{0,\Omega}}{\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega}} \geq \frac{1}{C_N} \, \|\boldsymbol{v} - \Pi_Q \boldsymbol{v}\|_{0,\Omega} \, . \tag{2.69}$$

In turn, if $\boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v} \neq \mathbf{0}$, a slight variation of the proof of [51, Lemma 4.4] allows us to show that there exists $\boldsymbol{\zeta} \in \mathbb{H}_0(\mathbf{div}; \Omega)$ such that $\frac{1}{2} \left(\boldsymbol{\zeta} - \boldsymbol{\zeta}^{\mathtt{t}} \right) = \boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}$ and $\|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega} \leq c_N \, \|\boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}\|_{0,\Omega}$, with a positive constant c_N, independent of $\boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}$. Hence, setting $\vec{\boldsymbol{\zeta}} := (\boldsymbol{\zeta}, \mathbf{0}) \in \mathbf{H}$, we see that

$$\begin{aligned} \sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} &= \frac{\|\boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}\|^2_{0,\Omega} + \int_\Omega (\boldsymbol{v} - \Pi_Q \boldsymbol{v}) \cdot \mathbf{div} \, \boldsymbol{\zeta}}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} \\ &\geq \frac{1}{c_N} \, \|\boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}\|_{0,\Omega} \, - \, \|\boldsymbol{v} - \Pi_Q \boldsymbol{v}\|_{0,\Omega} \, . \end{aligned} \tag{2.70}$$

Furthermore, assuming that $\boldsymbol{\xi} - \Pi_Q \boldsymbol{v} \neq \mathbf{0}$, we define $\vec{\boldsymbol{\zeta}} := (\mathbf{0}, \boldsymbol{\xi} - \Pi_Q \boldsymbol{v}) \in \mathbf{H}$ and readily observe that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\zeta}}\|_{\mathbf{H}}} = \|\boldsymbol{\xi} - \Pi_Q \boldsymbol{v}\|_{0,\Omega} \, . \tag{2.71}$$

In this way, since at least one of the components of $(\boldsymbol{v} - \Pi_Q \boldsymbol{v}, \boldsymbol{\Psi} - \nabla \Pi_Q \boldsymbol{v}, \boldsymbol{\xi} - \Pi_Q \boldsymbol{v})$ does not vanish, which follows from the fact that $\vec{\boldsymbol{v}} \notin \mathbf{V}_0$, a suitable linear combination of (2.69), (2.70), and (2.71) implies the existence of a positive constant $\beta_{\mathbf{B}}$, depending on C_N and c_N, such that

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}} \, \left\| \vec{\boldsymbol{v}} - \left(\Pi_Q \boldsymbol{v}, \nabla \Pi_Q \boldsymbol{v}, \Pi_Q \boldsymbol{v} \right) \right\|_{\mathbf{Q}} \, . \tag{2.72}$$

Finally, (2.72) and the fact that $\left(\Pi_Q \boldsymbol{v}, \nabla \Pi_Q \boldsymbol{v}, \Pi_Q \boldsymbol{v} \right) \in \mathbf{V}_0$ yield (2.66) and complete the proof. ■

We remark here that the inf-sup condition (2.66) provides an alternative proof of the inclusion $\mathbf{V} \subseteq \mathbf{V}_0$, and hence of the identity $\mathbf{V} = \mathbf{V}_0$. In fact, for each $\vec{\boldsymbol{v}} \in \mathbf{V}$ there necessarily holds, due to (2.66), $\mathrm{dist}(\vec{\boldsymbol{v}}, \mathbf{V}_0) = 0$, which is obviously equivalent to saying $\vec{\boldsymbol{v}} \in \mathbf{V}_0$. Furthermore, as a direct corollary of Lemma 2.7, we now state the continuous inf-sup condition for $\mathbf{b}$ required by item iii) of Theorem 2.6.

Lemma 2.8. *With the same constant $\beta_{\mathbf{B}}$ from Lemma 2.7 there holds*

$$\sup_{\substack{\vec{\boldsymbol{\tau}} \in \mathbf{H} \\ \vec{\boldsymbol{\tau}} \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}, \vec{\boldsymbol{v}})}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}}} \geq \beta_{\mathbf{B}} \, \|\vec{\boldsymbol{v}}\|_{\mathbf{Q}} \qquad \forall \, \vec{\boldsymbol{v}} \in \mathbf{V}^{\perp} \, . \tag{2.73}$$

Proof. It suffices to use in (2.66) that $\mathrm{dist}(\vec{\boldsymbol{v}}, \mathbf{V}_0) = \mathrm{dist}(\vec{\boldsymbol{v}}, \mathbf{V}) = \|\vec{\boldsymbol{v}}\|_{\mathbf{Q}}$ for all $\vec{\boldsymbol{v}} \in \mathbf{V}^{\perp}$. ■

Next, having in mind that $\mathbf{V} = \mathbf{V}_0$ (cf. (3.19)), we prove the $\mathbf{V}$-ellipticity of the bilinear form $\mathbf{c}$ (cf. (2.56)).

Lemma 2.9. *There exists a positive constant* $\gamma_{\mathbf{V}}$ *such that*

$$\mathbf{c}(\vec{\boldsymbol{v}},\vec{\boldsymbol{v}}) \,\geq\, \gamma_{\mathbf{V}}\,\|\vec{\boldsymbol{v}}\|^2_{\mathbf{Q}} \qquad \forall\,\vec{\boldsymbol{v}}\in\mathbf{V}\,.$$

Proof. Given $\vec{\boldsymbol{v}} := (\mathbf{q},\nabla\mathbf{q},\mathbf{q}) \in \mathbf{V}$ (cf. (3.19)), it follows from (2.56) and the fact that all the norms in Q are equivalent, that there exists a positive constant c_E, depending only on Q, such that

$$\mathbf{c}(\vec{\boldsymbol{v}},\vec{\boldsymbol{v}}) \,=\, \beta\,\|\mathbf{q}\|^2_{0,\Omega} \,\geq\, \frac{\beta}{2}\Big\{\|\mathbf{q}\|^2_{0,\Omega} \,+\, c_E\,\|\mathbf{q}\|^2_{1,\Omega}\Big\} \,\geq\, \gamma_{\mathbf{V}}\,\|\vec{\boldsymbol{v}}\|^2_{\mathbf{Q}} \qquad \forall\,\vec{\boldsymbol{v}}\in\mathbf{V}\,,$$

with $\gamma_{\mathbf{V}} \,=\, \frac{\beta}{2}\,\min\big\{1,c_E\big\}$. ∎

Hence, thanks to (2.58), and Lemmas 2.6, 2.8, and 2.9, we are able to prove the following result.

Lemma 2.10. *For each pair* $(\mathbf{F},\mathbf{G}) \in \mathbf{H}'\times\mathbf{Q}'$ *there exist unique* $\vec{\underline{\boldsymbol{\sigma}}} := (\underline{\boldsymbol{\sigma}},\underline{\boldsymbol{\rho}}) \in \mathbf{H}$ *and* $\vec{\underline{\mathbf{u}}} := (\underline{\mathbf{u}},\underline{\boldsymbol{\Phi}},\underline{\boldsymbol{\lambda}}) \in \mathbf{Q}$ *such that*

$$\begin{aligned} \mathbf{a}(\vec{\underline{\boldsymbol{\sigma}}},\vec{\boldsymbol{\tau}}) \,+\, \mathbf{b}(\vec{\boldsymbol{\tau}},\vec{\underline{\mathbf{u}}}) \,&=\, \mathbf{F}(\vec{\boldsymbol{\tau}}) \qquad \forall\,\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau},\boldsymbol{\eta}) \in \mathbf{H}\,, \\ \mathbf{b}(\vec{\underline{\boldsymbol{\sigma}}},\vec{\boldsymbol{v}}) \,-\, \mathbf{c}(\vec{\underline{\mathbf{u}}},\vec{\boldsymbol{v}}) \,&=\, \mathbf{G}(\vec{\boldsymbol{v}}) \qquad \forall\,\vec{\boldsymbol{v}} := (\boldsymbol{v},\boldsymbol{\Psi},\boldsymbol{\xi}) \in \mathbf{Q}\,. \end{aligned} \tag{2.74}$$

Moreover, there exists a positive constant $\underline{C}$*, depending only on* $\alpha_{\mathbf{K}}$*,* $\beta_{\mathbf{B}}$*,* $\gamma_{\mathbf{V}}$*, and the norms of the operators induced by* $\mathbf{a}$ *and* $\mathbf{b}$*, such that*

$$\|(\vec{\underline{\boldsymbol{\sigma}}},\vec{\underline{\mathbf{u}}})\|_{\mathbf{H}\times\mathbf{Q}} \,\leq\, \underline{C}\,\Big\{\|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'}\Big\}\,. \tag{2.75}$$

Proof. It follows from a straightforward application of Theorem 2.6. ∎

Next, given an arbitrary $\mathbf{z}\in\mathbf{L}^2(\Omega)$, we consider the particular pair $(\mathbf{F},\mathbf{G}) := (\mathbf{0},\alpha\,\mathbf{F}_{\mathbf{z}}) \in \mathbf{H}'\times\mathbf{Q}'$, and conclude, thanks to Lemma 2.10, that the problem defining $\mathbf{T}(\mathbf{z})$ (cf. (2.59)) is well-posed, thus confirming that the operator $\mathbf{T} : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ is well-defined. Moreover, by noticing from (2.57) that $\|\alpha\,\mathbf{F}_{\mathbf{z}}\|_{\mathbf{Q}'} = \alpha\,\|\nabla\mathcal{D}(\mathbf{z})\|_{0,\Omega}$, we deduce from (2.75) that there holds

$$\|\mathbf{T}(\mathbf{z})\|_{0,\Omega} \,\leq\, \|(\vec{\underline{\boldsymbol{\sigma}}},\vec{\underline{\mathbf{u}}})\|_{\mathbf{H}\times\mathbf{Q}} \,\leq\, \underline{C}\,\alpha\,\|\nabla\mathcal{D}(\mathbf{z})\|_{0,\Omega} \qquad \forall\,\mathbf{z}\in\mathbf{L}^2(\Omega)\,. \tag{2.76}$$

The Lipschitz-continuity of the operator $\mathbf{T}$ is established in the following lemma.

Lemma 2.11. *Assume* (A2) *and let* $\underline{C}$ *be the constant provided by the continuous dependence estimate* (2.75)*. Then, there holds*

$$\|\mathbf{T}(\mathbf{z}_1) - \mathbf{T}(\mathbf{z}_2)\|_{0,\Omega} \,\leq\, \alpha\,\underline{C}\,L_{\mathcal{D}}\,\|\mathbf{z}_1-\mathbf{z}_2\|_{0,\Omega} \qquad \forall\,\mathbf{z}_1,\,\mathbf{z}_2\in\mathbf{L}^2(\Omega)\,.$$

Proof. We proceed analogously to [13, Lemma 11]. In this way, given $\mathbf{z}_j \in \mathbf{L}^2(\Omega)$, $j\in\big\{1,2\big\}$, we let $\vec{\underline{\boldsymbol{\sigma}}}_j := (\underline{\boldsymbol{\sigma}}_j,\underline{\boldsymbol{\rho}}_j) \in \mathbf{H}$ and $\vec{\underline{\mathbf{u}}}_j := (\underline{\mathbf{u}}_j,\underline{\boldsymbol{\Phi}}_j,\underline{\boldsymbol{\lambda}}_j) \in \mathbf{Q}$ be the unique solution to (2.59) with $\mathbf{z}=\mathbf{z}_j$, so that $\mathbf{T}(\mathbf{z}_j) = \underline{\mathbf{u}}_j$. Subtracting the respective rows of the resulting systems (2.59), we easily find that $(\vec{\underline{\boldsymbol{\sigma}}}_1 - \vec{\underline{\boldsymbol{\sigma}}}_2,\vec{\underline{\mathbf{u}}}_1 - \vec{\underline{\mathbf{u}}}_2) \in \mathbf{H}\times\mathbf{Q}$ is solution of (2.74) with $\mathbf{F} := 0$ and $\mathbf{G} := \alpha\,\big(\mathbf{F}_{\mathbf{z}_1} - \mathbf{F}_{\mathbf{z}_2}\big)$, and hence the corresponding estimate (2.75) and the Lipschitz continuity of $\nabla\mathcal{D}$ (cf. (A2)) yield

$$\begin{aligned} \|\mathbf{T}(\mathbf{z}_1) - \mathbf{T}(\mathbf{z}_2)\|_{0,\Omega} \,&\leq\, \|\vec{\underline{\mathbf{u}}}_1 - \vec{\underline{\mathbf{u}}}_2\|_{\mathbf{Q}'} \,\leq\, \underline{C}\,\|\alpha(\mathbf{F}_{\mathbf{z}_1} - \mathbf{F}_{\mathbf{z}_2})\|_{\mathbf{Q}'} \\ &=\, \underline{C}\,\alpha\,\|\nabla\mathcal{D}(\mathbf{z}_1) - \nabla\mathcal{D}(\mathbf{z}_2)\|_{0,\Omega} \,\leq\, \underline{C}\,\alpha\,L_{\mathcal{D}}\,\|\mathbf{z}_1-\mathbf{z}_2\|_{0,\Omega}\,, \end{aligned}$$

which finishes the proof. ∎

We are now in position to establish the existence of a unique fixed-point for the operator $\mathbf{T}$, or equivalently, the well-possedness of problem (2.53). More precisely, we have the following result.

Theorem 2.7. *Assume* (A2), (A3) *and* $\alpha\,\underline{C}\,L_{\mathcal{D}} < 1$. *Then, the mixed problem* (2.53) *has a unique solution* $(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) \in \mathbf{H} \times \mathbf{Q}$. *Moreover, the following a priori estimate holds*

$$\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\,\alpha\, M_D\,.$$

Proof. It follows straightforwardly from Lemma 2.11 and the present hypothesis involving the constants α, $\underline{C}$, and $L_{\mathcal{D}}$ that $\mathbf{T}$ is a contraction, and hence the classical Banach theorem implies the existence of a unique fixed point of $\mathbf{T}$. Equivalently, the mixed problem (2.53) has a unique solution $(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) \in \mathbf{H} \times \mathbf{Q}$, which, according to the estimate (2.76) and the assumption (A3), satisfies

$$\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}})\|_{\mathbf{H}\times\mathbf{Q}} \leq \underline{C}\,\alpha\,\|\nabla\mathcal{D}(\mathbf{u})\|_{0,\Omega} \leq \underline{C}\,\alpha\, M_{\mathcal{D}}\,,$$

thus completing the proof. ∎

2.3.3 Analysis of the discrete scheme

In this section we introduce and analyze a Galerkin scheme for problem (2.53). As in Section 2.2.4, we first let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of $\bar{\Omega}$ made of triangles K with diameter h_K, and define the meshsize $h := \max\{h_K : \quad K \in \mathcal{T}_h\}$, which also serves as the index of $\mathcal{T}_h$. In turn, we recall that, given a non-negative integer k, $\mathrm{P}_k(K)$ stands for the space of polynomials of degree $\leq k$ on K, whose vector and tensor versions are denoted by $\mathbf{P}_k(K)$ and $\mathbb{P}_k(K)$, respectively. Then, noting that certainly the space of rigid motions Q is already of finite dimension, we propose next two possible sets of finite element subspaces of $\mathbb{H}_0(\mathbf{div};\Omega)$, $\mathbf{L}^2(\Omega)$, and $\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, which, in order to make clear the unknowns they are approximating, are denoted by $\mathrm{H}_h^{\boldsymbol{\sigma}}$, $\mathrm{H}_h^{\mathbf{u}}$ and $\mathrm{H}_h^{\boldsymbol{\Phi}}$, respectively. The first choice, employed in [13, Section 4.2] and [14, Section 3.4] for previous related results, consists of the Brezzi-Douglas-Marini (BDM) space of order 1 for the stress (cf. [22]) and the rest as in [10, Theorem 7.2], that is

$$\begin{aligned}
\mathrm{H}_h^{\boldsymbol{\sigma}} &:= \left\{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbb{P}_1(K) \quad \forall\, K \in \mathcal{T}_h\right\},\\
\mathrm{H}_h^{\mathbf{u}} &:= \left\{\boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_0(K) \quad \forall\, K \in \mathcal{T}_h\right\},\\
\mathrm{H}_h^{\boldsymbol{\Phi}} &:= \left\{\boldsymbol{\Psi}_h := \begin{pmatrix} 0 & \psi_h \\ -\psi_h & 0 \end{pmatrix} \in \mathbb{L}^2_{\mathtt{skew}}(\Omega) : \quad \psi_h|_K \in \mathrm{P}_0(K) \quad \forall\, K \in \mathcal{T}_h\right\},
\end{aligned} \tag{2.77}$$

In addition, we also consider the classical PEERS space of order 0, originally introduced in [9] for linear elasticity as well, which is given by

$$\begin{aligned}
\mathrm{H}_h^{\boldsymbol{\sigma}} &:= \left\{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \ \boldsymbol{\tau}_{h,i}|_K \in \mathbf{RT}_0(K) \oplus \mathrm{P}_0(K)\,\mathrm{curl}^{\mathtt{t}} b_K \ \forall\, i \in \{1,2\},\ \forall\, K \in \mathcal{T}_h\right\},\\
\mathrm{H}_h^{\mathbf{u}} &:= \left\{\boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_0(K) \quad \forall\, K \in \mathcal{T}_h\right\},\\
\mathrm{H}_h^{\boldsymbol{\Phi}} &:= \left\{\boldsymbol{\Psi}_h := \begin{pmatrix} 0 & \psi_h \\ -\psi_h & 0 \end{pmatrix} \in \mathbb{C}(\bar{\Omega}) : \quad \psi_h|_K \in \mathrm{P}_1(K) \quad \forall\, K \in \mathcal{T}_h\right\},
\end{aligned} \tag{2.78}$$

where $\boldsymbol{\tau}_{h,i}$ denotes the ith row of $\boldsymbol{\tau}_h$, $\mathbf{RT}_0(K)$ is the local Raviart-Thomas space of order 0 (cf. [21], [49]), b_K is the usual cubic bubble function on K, and $\mathrm{curl}^{\mathtt{t}} b_K = \left(\frac{\partial b_K}{\partial x_2}, -\frac{\partial b_K}{\partial x_1}\right)$. Nevertheless, for stability purposes to be discussed later on (see Lemma 2.12 below), we need that the space of rigid motions Q be contained in the finite element subspace approximating $\mathbf{u}$, reason why we now enrich this space with the $\mathbf{P}_1(\Omega)$-component of Q, thus yielding the introduction of

$$\widetilde{\mathrm{H}}_h^{\mathbf{u}} := \mathrm{H}_h^{\mathbf{u}} \oplus \left\langle \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right\rangle . \tag{2.79}$$

Then, letting $\mathbf{H}_h := \mathrm{H}_h^{\boldsymbol{\sigma}} \times Q$ and $\mathbf{Q}_h := \widetilde{\mathrm{H}}_h^{\mathbf{u}} \times \mathrm{H}_h^{\boldsymbol{\Phi}} \times Q$, the Galerkin scheme of (2.53) reads: Find $\vec{\boldsymbol{\sigma}}_h := (\boldsymbol{\sigma}_h, \boldsymbol{\rho}_h) \in \mathbf{H}_h$ and $\vec{\mathbf{u}}_h := (\mathbf{u}_h, \boldsymbol{\Phi}_h, \boldsymbol{\lambda}_h) \in \mathbf{Q}_h$ such that

$$\begin{aligned} \mathbf{a}(\vec{\boldsymbol{\sigma}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{u}}_h) &= 0 \qquad && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h\,, \\ \mathbf{b}(\vec{\boldsymbol{\sigma}}_h, \vec{\boldsymbol{v}}_h) - \mathbf{c}(\vec{\mathbf{u}}_h, \vec{\boldsymbol{v}}_h) &= \alpha\, \mathbf{F}_{\mathbf{u}_h}(\vec{\boldsymbol{v}}_h) \qquad && \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h\,. \end{aligned} \tag{2.80}$$

Analogously to the analysis from Section 2.3.2, we now introduce the discrete operator $\mathbf{T}_h : \widetilde{\mathrm{H}}_h^{\mathbf{u}} \to \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ defined by $\mathbf{T}_h(\mathbf{z}_h) := \underline{\mathbf{u}}_h$ for each $\mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$, where $\vec{\underline{\boldsymbol{\sigma}}}_h := (\underline{\boldsymbol{\sigma}}_h, \underline{\boldsymbol{\rho}}_h) \in \mathbf{H}_h$ and $\vec{\underline{\mathbf{u}}}_h := (\underline{\mathbf{u}}_h, \underline{\boldsymbol{\Phi}}_h, \underline{\boldsymbol{\lambda}}_h) \in \mathbf{Q}_h$ satisfy

$$\begin{aligned} \mathbf{a}(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\underline{\mathbf{u}}}_h) &= 0 \qquad && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h\,, \\ \mathbf{b}(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{v}}_h) - \mathbf{c}(\vec{\underline{\mathbf{u}}}_h, \vec{\boldsymbol{v}}_h) &= \alpha\, \mathbf{F}_{\mathbf{z}_h}(\vec{\boldsymbol{v}}_h) \qquad && \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h\,. \end{aligned} \tag{2.81}$$

As for the continuous problem, it is easy to see that solving (2.80) is equivalent to looking for a fixed point of $\mathbf{T}_h$, that is: Find $\mathbf{u}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ such that $\mathbf{T}_h(\mathbf{u}_h) = \mathbf{u}_h$, for whose solvability analysis we need to show first that $\mathbf{T}_h$ is well-defined, equivalently that (2.81) is well-posed. For this purpose, in what follows we apply Theorem 2.6 to the discrete setting provided by the spaces $\mathbf{H}_h$ and $\mathbf{Q}_h$, the bilinear forms $\mathbf{a}|_{\mathbf{H}_h \times \mathbf{H}_h}$ and $\mathbf{b}|_{\mathbf{H}_h \times \mathbf{Q}_h}$, and the discrete kernels of $\mathbf{B}$ and $\mathbf{B}^{\mathtt{t}}$, which are given, respectively, by

$$\mathbf{K}_h := \left\{ \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \quad \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h) = 0 \quad \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h \right\}, \tag{2.82}$$

and

$$\mathbf{V}_h := \left\{ \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h : \quad \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h) = 0 \quad \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h \right\}. \tag{2.83}$$

Thus, employing the expression for $\mathbf{b}$ given by (2.61), we can redefine $\mathbf{K}_h$ as

$$\begin{aligned} \mathbf{K}_h := \Big\{ \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \quad & \int_\Omega \boldsymbol{v}_h \cdot \left\{ \mathbf{div}\, \boldsymbol{\tau}_h - \boldsymbol{\eta}_h \right\} = 0 \quad \forall\, \boldsymbol{v}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}\,, \\ & \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\tau}_h = 0 \quad \forall\, \boldsymbol{\Psi}_h \in \mathrm{H}_h^{\boldsymbol{\Phi}}\,, \quad \int_\Omega \boldsymbol{\xi}_h \cdot \boldsymbol{\eta}_h = 0 \quad \forall\, \boldsymbol{\xi}_h \in Q \Big\}\,, \end{aligned} \tag{2.84}$$

from which, noticing that the pair $(\mathrm{H}_h^{\boldsymbol{\sigma}}, \widetilde{\mathrm{H}}_h^{\mathbf{u}})$, taken either from (2.77) - (2.79) or (2.78) - (2.79), satisfies the inclusion $\mathbf{div}\, \mathrm{H}_h^{\boldsymbol{\sigma}} \subseteq \widetilde{\mathrm{H}}_h^{\mathbf{u}}$, it readily follows that

$$\mathbf{K}_h := \left\{ \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h : \quad \mathbf{div}\, \boldsymbol{\tau}_h = 0\,, \quad \boldsymbol{\eta}_h = 0\,, \quad \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\tau}_h = 0 \quad \forall\, \boldsymbol{\Psi}_h \in \mathrm{H}_h^{\boldsymbol{\Phi}} \right\}.$$

In this way, due to the first two identities characterizing $\mathbf{K}_h$ in the foregoing equation, we deduce that the $\mathbf{K}_h$-ellipticity of $\mathbf{a}$ can be proved exactly as we did for its $\mathbf{K}$-ellipticity, and hence with the same constant $\alpha_{\mathbf{K}} := \frac{c_1 c_2}{2\mu_s}$ from Lemma 2.6 there holds

$$\mathbf{a}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{\tau}}_h) \geq \alpha_{\mathbf{K}} \, \|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}^2 \qquad \forall\, \vec{\boldsymbol{\tau}}_h \in \mathbf{K}_h\,. \tag{2.85}$$

We now aim to establish the discrete analogue of Lemma 2.7, for which we first highlight that, thanks to the enriched space $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$ (cf. (2.79)), one guarantees that $\mathbf{V}_0$ (cf. (3.19)) is a subspace of $\mathbf{Q}_h$. Then, we have the following result.

Lemma 2.12. *There exists a positive constant $\widetilde{\beta}_{\mathbf{B}}$, independent of h, such that*

$$S_h(\vec{\boldsymbol{v}}_h) := \sup_{\substack{\vec{\boldsymbol{\tau}}_h \in \mathbf{H}_h \\ \vec{\boldsymbol{\tau}}_h \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h)}{\|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}} \geq \widetilde{\beta}_{\mathbf{B}} \, \mathrm{dist}(\vec{\boldsymbol{v}}_h, \mathbf{V}_0) \qquad \forall \, \vec{\boldsymbol{v}}_h \in \mathbf{Q}_h \, . \tag{2.86}$$

Proof. We proceed analogously to the proof of Lemma 2.7. However, because of the similarities involved, we simplify our reasoning by using the results already available along the proof of [50, Lemma 4.1], which in turn is an adaptation of the proof of [73, Theorem 4.5]. We begin by recalling from (2.67) that, given $\vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h$ and $\vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h$, we can rewrite $\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h)$ as

$$\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h) := \int_\Omega (\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h) \cdot \mathbf{div}\, \boldsymbol{\tau}_h + \int_\Omega (\boldsymbol{\Psi}_h - \nabla \boldsymbol{\Pi}_Q \boldsymbol{v}_h) : \boldsymbol{\tau}_h + \int_\Omega (\boldsymbol{\xi}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h) \cdot \boldsymbol{\eta}_h \, , \tag{2.87}$$

from which one easily deduces that $\mathbf{V}_0 \subseteq \mathbf{V}_h$, and hence (2.86) is trivially satisfied for $\vec{\boldsymbol{v}}_h \in \mathbf{V}_0$. According to this, it only remains to prove for $\vec{\boldsymbol{v}}_h \in \mathbf{Q}_h \setminus \mathbf{V}_0$. Indeed, if $\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h \neq \mathbf{0}$, we know from the first part of the proof of [50, Lemma 4.1] that there exists $\boldsymbol{\zeta}_h \in \mathrm{H}_h^{\boldsymbol{\sigma}}$ such that $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathcal{P}_h(\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h)$ and $\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega} \leq \widetilde{C}_N \, \|\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h\|_{0,\Omega}$, where $\mathcal{P}_h : \mathbf{L}^2(\Omega) \to \mathrm{H}_h^{\mathbf{u}}$ is the orthogonal projection, and $\widetilde{C}_N$ is a positive constant independent of h. In turn, decomposing $\boldsymbol{v}_h = \bar{\boldsymbol{v}}_h + \mathbf{q}_h$, with $\bar{\boldsymbol{v}}_h \in \mathrm{H}_h^{\mathbf{u}}$ and $\mathbf{q}_h \in \left\langle \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right\rangle$, we obtain $\boldsymbol{\Pi}_Q \boldsymbol{v}_h = \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h + \mathbf{q}_h$, and thus $\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h = \bar{\boldsymbol{v}}_h - \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h$. In particular, this latter identity obviously implies $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathcal{P}_h(\bar{\boldsymbol{v}}_h - \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h)$. Then, setting $\vec{\boldsymbol{\zeta}}_h := (\boldsymbol{\zeta}_h, \mathbf{0})$, using the original definition of $\mathbf{b}$ (cf. (2.55)), integrating by parts similarly as done for the derivation of (2.67), and applying the properties of the orthogonal projections $\mathcal{P}_h$ and $\boldsymbol{\Pi}_Q$, we find that

$$\begin{aligned}
\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\boldsymbol{v}}_h) &= \int_\Omega (\bar{\boldsymbol{v}}_h + \mathbf{q}_h) \cdot \mathbf{div}(\boldsymbol{\zeta}_h) + \int_\Omega \boldsymbol{\Psi}_h : \boldsymbol{\zeta}_h \\
&= \int_\Omega \bar{\boldsymbol{v}}_h \cdot \mathcal{P}_h(\bar{\boldsymbol{v}}_h - \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h) + \int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h) : \boldsymbol{\zeta}_h \\
&= \int_\Omega \bar{\boldsymbol{v}}_h \cdot (\bar{\boldsymbol{v}}_h - \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h) + \int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h) : \boldsymbol{\zeta}_h \\
&= \|\bar{\boldsymbol{v}}_h - \boldsymbol{\Pi}_Q \bar{\boldsymbol{v}}_h\|_{0,\Omega}^2 + \int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h) : \boldsymbol{\zeta}_h \, ,
\end{aligned}$$

which readily yields

$$\begin{aligned}
S_h(\vec{\boldsymbol{v}}_h) \geq \frac{\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\boldsymbol{v}}_h)}{\|\vec{\boldsymbol{\zeta}}_h\|_{\mathbf{H}}} &= \frac{\|\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h\|_{0,\Omega}^2 + \displaystyle\int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h) : \boldsymbol{\zeta}_h}{\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega}} \\
&\geq \frac{1}{\widetilde{C}_N} \|\boldsymbol{v}_h - \boldsymbol{\Pi}_Q \boldsymbol{v}_h\|_{0,\Omega} - \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega} \, .
\end{aligned} \tag{2.88}$$

Next, assuming that $\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h \neq \mathbf{0}$ and appealing now to the second half of the proof of [50, Lemma 4.1], there exists another $\boldsymbol{\zeta}_h \in \mathrm{H}_h^{\boldsymbol{\sigma}}$ such that $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathbf{0}$, $\displaystyle\int_\Omega (\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h) : \boldsymbol{\zeta}_h = \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega}^2$, and

$\|\boldsymbol{\zeta}_h\|_{\mathbf{div};\Omega} \leq \widetilde{c}_N \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega}$, where $\widetilde{c}_N$ is a positive constant independent of h. Hence, defining $\vec{\boldsymbol{\zeta}}_h := (\boldsymbol{\zeta}_h, \mathbf{0})$, and employing again (2.55), we obtain

$$\mathbf{b}(\vec{\boldsymbol{\zeta}}_h, \vec{\boldsymbol{v}}_h) = \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|^2_{0,\Omega} ,$$

which, similarly as before, gives

$$S_h(\vec{\boldsymbol{v}}_h) \geq \frac{1}{\widetilde{c}_N} \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega} . \tag{2.89}$$

In this way, a suitable linear combination of (2.88) and (2.89) implies the existence of a positive constant $\widetilde{\beta}_1$, depending only on $\widetilde{C}_N$ and $\widetilde{c}_N$, such that

$$S_h(\vec{\boldsymbol{v}}_h) \geq \widetilde{\beta}_1 \left\{ \|\boldsymbol{v}_h - \Pi_Q \boldsymbol{v}_h\|_{0,\Omega} + \|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega} \right\} . \tag{2.90}$$

In addition, proceeding exactly as for the derivation of (2.89), but now considering $\boldsymbol{\Psi}_h - \nabla \Pi_Q \boldsymbol{v}_h$ in place of $\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h$, and utilizing the expression (2.87) for $\mathbf{b}$, we are able to show that

$$S_h(\vec{\boldsymbol{v}}_h) \geq \frac{1}{\widehat{c}_N} \|\boldsymbol{\Psi}_h - \nabla \Pi_Q \boldsymbol{v}_h\|_{0,\Omega} , \tag{2.91}$$

with a positive constant $\widehat{c}_N$ independent of h. Furthermore, if $\boldsymbol{\xi}_h - \Pi_Q \boldsymbol{v}_h \neq \mathbf{0}$, we do as in the continuous case (cf. (2.71) in the proof of Lemma 2.7) and choose $\vec{\boldsymbol{\zeta}}_h := (\mathbf{0}, \boldsymbol{\xi}_h - \Pi_Q \boldsymbol{v}_h)$ to prove, according to (2.87), that

$$S_h(\vec{\boldsymbol{v}}_h) \geq \|\boldsymbol{\xi}_h - \Pi_Q \boldsymbol{v}_h\|_{0,\Omega} . \tag{2.92}$$

The rest of the proof follows analogously to the one of Lemma 2.7 by considering now the inequalities (2.90), (2.91), and (2.92), and after discarding the expression $\|\boldsymbol{\Psi}_h - \nabla \mathbf{q}_h\|_{0,\Omega}$ in the first one of them. We omit further details. ■

As a first straightforward consequence of (2.86) we have that $\mathbf{V}_h \subseteq \mathbf{V}_0$, and hence $\mathbf{V}_h = \mathbf{V}_0$. Moreover, since $\mathrm{dist}(\vec{\boldsymbol{v}}_h, \mathbf{V}_h) = \|\vec{\boldsymbol{v}}_h\|_{\mathbf{Q}}$ for all $\vec{\boldsymbol{v}}_h \in \mathbf{V}_h^{\perp}$, we conclude the discrete inf-sup condition for $\mathbf{b}$, that is

$$\sup_{\substack{\vec{\boldsymbol{\tau}}_h \in \mathbf{H}_h \\ \vec{\boldsymbol{\tau}}_h \neq \mathbf{0}}} \frac{\mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\boldsymbol{v}}_h)}{\|\vec{\boldsymbol{\tau}}_h\|_{\mathbf{H}}} \geq \widetilde{\beta}_{\mathbf{B}} \|\vec{\boldsymbol{v}}_h\|_{\mathbf{Q}} \qquad \forall\, \vec{\boldsymbol{v}}_h \in \mathbf{V}_h^{\perp} \cap \mathbf{Q}_h , \tag{2.93}$$

with certainly the same constant $\widetilde{\beta}_{\mathbf{B}}$ from Lemma 2.12. On the other hand, since the continuous and discrete kernels $\mathbf{V}$ and $\mathbf{V}_h$, respectively, coincide, the $\mathbf{V}_h$-ellipticity of the bilinear form $\mathbf{c}$ is already proved by Lemma 2.9.

Therefore, bearing in mind (2.85), (2.93), and Lemma 2.9, a straightforward application of Theorem 2.6 allows us to establish the following result.

Lemma 2.13. *For each pair* $(\mathbf{F}, \mathbf{G}) \in \mathbf{H}' \times \mathbf{Q}'$ *there exist unique* $\underline{\vec{\boldsymbol{\sigma}}}_h := (\underline{\boldsymbol{\sigma}}_h, \underline{\boldsymbol{\rho}}_h) \in \mathbf{H}_h$ *and* $\underline{\vec{\mathbf{u}}}_h := (\underline{\mathbf{u}}_h, \underline{\boldsymbol{\Phi}}_h, \underline{\boldsymbol{\lambda}}_h) \in \mathbf{Q}_h$ *such that*

$$\begin{aligned} \mathbf{a}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \underline{\vec{\mathbf{u}}}_h) &= \mathbf{F}(\vec{\boldsymbol{\tau}}_h) \qquad \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h , \\ \mathbf{b}(\underline{\vec{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{v}}_h) - \mathbf{c}(\underline{\vec{\mathbf{u}}}_h, \vec{\boldsymbol{v}}_h) &= \mathbf{G}(\vec{\boldsymbol{v}}_h) \qquad \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h . \end{aligned} \tag{2.94}$$

Moreover, there exists a positive constant $\widetilde{C}$, *depending only on* $\alpha_{\mathbf{K}}$, $\widetilde{\beta}_{\mathbf{B}}$, $\gamma_{\mathbf{V}}$, *and the norms of the operators induced by* $\mathbf{a}$ *and* $\mathbf{b}$, *such that*

$$\|(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\underline{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{C}\left\{\|\mathbf{F}\|_{\mathbf{H}'} + \|\mathbf{G}\|_{\mathbf{Q}'}\right\}. \tag{2.95}$$

Next, we proceed analogously to the continuous case (cf. (2.76) and the last part of Section 2.3.2) by applying now Lemma 2.13 to the pair of functionals $(\mathbf{F}, \mathbf{G}) := (\mathbf{0}, \alpha\, \mathbf{F}_{\mathbf{z}_h})$, with an arbitrary $\mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}$. In this way, we conclude that $\mathbf{T}_h : \widetilde{\mathrm{H}}_h^{\mathbf{u}} \to \widetilde{\mathrm{H}}_h^{\mathbf{u}}$ is well-posed, and that

$$\|\mathbf{T}_h(\mathbf{z}_h)\|_{0,\Omega} \leq \|(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\underline{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{C}\,\alpha\,\|\nabla\mathcal{D}(\mathbf{z}_h)\|_{0,\Omega} \qquad \forall\, \mathbf{z}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}.$$

Moreover, adopting the same arguments from Lemma 2.11, and employing the a priori estimate (2.95) and the Lipschitz-continuity of $\nabla\mathcal{D}$ (cf. (A2)), we arrive at the same property for the operator $\mathbf{T}_h$, that is

$$\|\mathbf{T}_h(\mathbf{z}_h) - \mathbf{T}_h(\mathbf{w}_h)\|_{0,\Omega} \leq \widetilde{C}\,\alpha\, L_{\mathcal{D}}\,\|\mathbf{z}_h - \mathbf{w}_h\|_{0,\Omega} \qquad \forall\, \mathbf{z}_h\,,\, \mathbf{w}_h \in \widetilde{\mathrm{H}}_h^{\mathbf{u}}.$$

Consequently, we are now in position to establish the well-posedness of our mixed finite element method (2.80), by appealing to its equivalence with the existence of a unique fixed point of $\mathbf{T}_h$, and applying again the respective Banach theorem. We omit further details and state the corresponding result as follows.

Theorem 2.8. *Assume* (A2), (A3) *and* $\alpha\,\widetilde{C}\,L_{\mathcal{D}} < 1$. *Then, the discrete scheme* (2.80) *has a unique solution* $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$. *Moreover, the following a priori estimate holds*

$$\|(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{C}\,\alpha\, M_D\,.$$

2.3.4 A priori error analysis

Given $(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) \in \mathbf{H}\times\mathbf{Q}$ and $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$, the unique solutions of the continuous and discrete problems (2.53) and (2.80), respectively, we now aim to estimate the corresponding error given by $\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}}$. To this end, we first let $(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\underline{\mathbf{u}}}_h) \in \mathbf{H}_h \times \mathbf{Q}_h$ be the solution to (2.94) with $\mathbf{F} = \mathbf{0}$ and $\mathbf{G} = \alpha\,\mathbf{F}_{\mathbf{u}}$, equivalently the solution to (2.81) with $\mathbf{u}$ in place of $\mathbf{z}_h$, that is

$$\begin{aligned} \mathbf{a}(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\underline{\mathbf{u}}}_h) &= 0 && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h\,, \\ \mathbf{b}(\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{v}}_h) - \mathbf{c}(\vec{\underline{\mathbf{u}}}_h, \vec{\boldsymbol{v}}_h) &= \alpha\,\mathbf{F}_{\mathbf{u}}(\vec{\boldsymbol{v}}_h) && \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h\,, \end{aligned} \tag{2.96}$$

which certainly can be seen as the classical Galerkin approximation of (2.53). Then, invoking the corresponding Céa estimate (see, e.g., [18, Proposition 5.5.2.]), we have the preliminary estimate

$$\|(\vec{\boldsymbol{\sigma}}, \vec{\mathbf{u}}) - (\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\underline{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widehat{C}\left\{\mathrm{dist}(\vec{\boldsymbol{\sigma}}, \mathbf{H}_h) + \mathrm{dist}(\vec{\mathbf{u}}, \mathbf{Q}_h)\right\}, \tag{2.97}$$

where $\widehat{C}$ is a positive constant independent of h. Next, subtracting (2.96) from (2.80), we find that $(\vec{\boldsymbol{\sigma}}_h, \vec{\mathbf{u}}_h) - (\vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\underline{\mathbf{u}}}_h)$ solves

$$\begin{aligned} \mathbf{a}(\vec{\boldsymbol{\sigma}}_h - \vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{\tau}}_h) + \mathbf{b}(\vec{\boldsymbol{\tau}}_h, \vec{\mathbf{u}}_h - \vec{\underline{\mathbf{u}}}_h)) &= 0 && \forall\, \vec{\boldsymbol{\tau}}_h := (\boldsymbol{\tau}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h\,, \\ \mathbf{b}(\vec{\boldsymbol{\sigma}}_h - \vec{\underline{\boldsymbol{\sigma}}}_h, \vec{\boldsymbol{v}}_h) - \mathbf{c}(\vec{\mathbf{u}}_h - \vec{\underline{\mathbf{u}}}_h, \vec{\boldsymbol{v}}_h) &= \alpha\,(\mathbf{F}_{\mathbf{u}_h} - \mathbf{F}_{\mathbf{u}})(\vec{\boldsymbol{v}}_h) && \forall\, \vec{\boldsymbol{v}}_h := (\boldsymbol{v}_h, \boldsymbol{\Psi}_h, \boldsymbol{\xi}_h) \in \mathbf{Q}_h\,, \end{aligned}$$

and hence, thanks to the a priori estimate (2.95), the fact that $\|\mathbf{F}_{\mathbf{u}_h}-\mathbf{F}_{\mathbf{u}}\|_{\mathbf{Q}'} = \|\nabla\mathcal{D}(\mathbf{u}_h)-\nabla\mathcal{D}(\mathbf{u})\|_{0,\Omega}$ (cf. (2.57)), and the Lipschitz-continuity of $\nabla\mathcal{D}$ (cf. (A2)), there holds

$$\|(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)-(\underline{\vec{\boldsymbol{\sigma}}}_h,\underline{\vec{\mathbf{u}}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \widetilde{C}\,\alpha\,L_{\mathcal{D}}\,\|\mathbf{u}-\mathbf{u}_h\|_{0,\Omega}\,. \tag{2.98}$$

In this way, employing the triangle inequality together with the estimates (2.97) and (2.98), and then realizing that $\mathrm{dist}(\vec{\boldsymbol{\sigma}},\mathbf{H}_h)=\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})$ and that $\mathrm{dist}(\vec{\mathbf{u}},\mathbf{Q}_h)=\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})$, we get

$$\begin{aligned}\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} &\leq \widehat{C}\left\{\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})\right\}\\ &+\quad \widetilde{C}\,\alpha\,L_{\mathcal{D}}\,\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}}\,.\end{aligned}$$

The foregoing inequality readily implies the following main result.

Theorem 2.9. *Assume* (A2), (A3) *and that* $\widetilde{C}\,\alpha\,L_{\mathcal{D}} \leq 1-\delta$, *with* $\delta\in]0,1[$. *Then, there holds*

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq \delta^{-1}\,\widehat{C}\left\{\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})\right\}.$$

Exactly as remarked at the end of Section 2.2.3, we also stress here that the optimal value of δ is $1/2$, whence we obtain the assumption $\widetilde{C}\,\alpha\,L_{\mathcal{D}} \leq 1/2$ and the Céa estimate

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq 2\,\widehat{C}\left\{\mathrm{dist}(\boldsymbol{\sigma},\mathrm{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\mathbf{u},\widetilde{\mathrm{H}}_h^{\mathbf{u}})+\mathrm{dist}(\boldsymbol{\Phi},\mathrm{H}_h^{\boldsymbol{\Phi}})\right\}. \tag{2.99}$$

We end this section with the rates of convergence of our mixed finite element solution $(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)$, for which we first recall the approximation properties of the finite element subspaces involved (see [21]).

$(\mathbf{AP}_h^{\boldsymbol{\sigma}})$ there exists $C>0$, independent of h, such that for each $\boldsymbol{\tau}\in\mathbb{H}^1(\Omega)\cap\mathbb{H}_0(\mathbf{div};\Omega)$ with $\mathbf{div}(\boldsymbol{\tau})\in\mathbf{H}^1(\Omega)$ there holds

$$\mathrm{dist}(\boldsymbol{\tau},\mathrm{H}_h^{\boldsymbol{\sigma}}) \leq C\,h\left\{\|\boldsymbol{\tau}\|_{1,\Omega}+\|\mathbf{div}(\boldsymbol{\tau})\|_{1,\Omega}\right\}.$$

$(\mathbf{AP}_h^{\mathbf{u}})$ there exists $C>0$, independent of h, such that for each $\boldsymbol{v}\in\mathbf{H}^1(\Omega)$ there holds

$$\mathrm{dist}(\boldsymbol{v},\mathrm{H}_h^{\mathbf{u}}) \leq C\,h\,\|\boldsymbol{v}\|_{1,\Omega}\,.$$

$(\mathbf{AP}_h^{\boldsymbol{\Phi}})$ there exists $C>0$, independent of h, such that for each $\boldsymbol{\Psi}\in\mathbb{H}^1(\Omega)\cap\mathbb{L}^2_{\mathtt{skew}}(\Omega)$ there holds

$$\mathrm{dist}(\boldsymbol{\Psi},\mathrm{H}^{\boldsymbol{\Phi}}) \leq C\,h\,\|\boldsymbol{\Psi}\|_{1,\Omega}\,.$$

Note here that, while $(\mathbf{AP}_h^{\mathbf{u}})$ provides the approximation property of $\mathrm{H}_h^{\mathbf{u}}$, the fact that this space is contained in $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$ implies that $\mathrm{dist}(\boldsymbol{v},\widetilde{\mathrm{H}}_h^{\mathbf{u}}) \leq \mathrm{dist}(\boldsymbol{v},\mathrm{H}_h^{\mathbf{u}})$, and hence $(\mathbf{AP}_h^{\mathbf{u}})$ also serves to estimate the distance to $\widetilde{\mathrm{H}}_h^{\mathbf{u}}$. According to the above discussion, the main result of this section is stated as follows.

Theorem 2.10. *Assume* (A2), (A3) *and that* $\widetilde{C}\,\alpha\,L_{\mathcal{D}} \leq 1/2$. *In addition, suppose that the solution* $(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}}) := \big((\boldsymbol{\sigma},\boldsymbol{\rho}),(\mathbf{u},\boldsymbol{\Phi},\boldsymbol{\lambda})\big) \in \mathbf{H}\times\mathbf{Q}$ *of* (2.53) *verifies* $\boldsymbol{\sigma}\in\mathbb{H}^1(\Omega)$, $\mathbf{div}(\boldsymbol{\sigma})\in\mathbf{H}^1(\Omega)$, $\mathbf{u}\in\mathbf{H}^1(\Omega)$, *and* $\boldsymbol{\Phi}\in\mathbb{H}^1(\Omega)$. *Then, there exists a positive constant* C, *independent of* h, *such that*

$$\|(\vec{\boldsymbol{\sigma}},\vec{\mathbf{u}})-(\vec{\boldsymbol{\sigma}}_h,\vec{\mathbf{u}}_h)\|_{\mathbf{H}\times\mathbf{Q}} \leq C\,h\left\{\|\boldsymbol{\sigma}\|_{1,\Omega}+\|\mathbf{div}(\boldsymbol{\sigma})\|_{1,\Omega}+\|\mathbf{u}\|_{1,\Omega}+\|\boldsymbol{\Phi}\|_{1,\Omega}\right\}.$$

Proof. It is a simple consequence of the Céa estimate (2.99), the additional regularity assumptions on the solution, and the approximation properties $(\mathbf{AP}_h^{\boldsymbol{\sigma}})$, $(\mathbf{AP}_h^{\mathbf{u}})$, and $(\mathbf{AP}_h^{\boldsymbol{\Phi}})$. ■

2.4 Implementation of the methods

We now refer to the practical implementation of (2.11). The extension to (2.53) proceeds similarly. More precisely, in what follows we employ a fictional time variable in a gradient flow fashion to implement the solution of problem (2.11), thus rendering problem (2.4) convex for a sufficiently small time step. This means that, given a time step Δt, $k\in\mathrm{N}$, and a previous iteration u_k, we modify the extended problem (2.9) to obtain

$$\min_{(v,\eta)\in H}\max_{\xi\in Q}\left\{\alpha\mathcal{D}(v)+\frac{1}{2}a(v,v)+\langle v-\eta,\xi\rangle+\frac{\beta}{2}\|\eta\|^2+\frac{1}{2\Delta t}\|v-u_k\|_{\mathcal{V}}^2\right\}, \tag{2.100}$$

where we recall from Section 2.2.2 that $H=\mathcal{V}\times Q$. Then, the first order conditions of this problem are given by the following: Find $\big((u,\rho),\lambda\big)\in H\times Q$ such that

$$\begin{aligned}\langle u,v\rangle+\Delta t\,a(u,v)+\beta\Delta t\langle\lambda,\eta\rangle+\Delta t\langle v-\eta,\rho\rangle&=\alpha\Delta t F_{u_k}(v)+\langle u_k,v\rangle \quad \forall (v,\eta)\in H,\\ \langle u-\lambda,\xi\rangle&=0 \qquad\qquad\qquad\quad \forall \xi\in Q,\end{aligned} \tag{2.101}$$

where the nonlinear term is treated explicitly, and which is well-posed in virtue of Theorem 2.2. The resulting solution of (2.101) is then redenoted $\big((u_{k+1},\rho_{k+1}),\lambda_{k+1}\big)$. Our main modification to the classical time dependent scheme used to implement registration problems is that the extended variables prevent the orthogonality to the kernel of the adjoint operator. Now we establish a relationship between subsequent iterations to find a bound on the time step for stability.

Lemma 2.14. *Given an initial iteration* $\big((u_0,\rho_0),\lambda_0\big)\in H\times Q$ *and* $n\in\mathrm{N}$, *we let* $\big((u_n,\rho_n),\lambda_n\big)$ *and* $\big((u_{n+1},\rho_{n+1}),\lambda_{n+1}\big)$ *be the solutions of* (2.101) *with* $k=n-1$ *and* $k=n$, *respectively. In addition, let* $\widetilde{c}_a$ *be the ellipticity constant of the bilinear form* a *(cf.* (A1)*), and define* $\kappa_1(\Delta t):=\big(\frac{1}{\Delta t}+2\widetilde{c}_a-\alpha\big)$ *and* $\kappa_2(\Delta t):=\big(\frac{1}{\Delta t}+\alpha L_{\mathcal{D}}\big)$. *Then, there holds*

$$\kappa_1(\Delta t)\,\|u_{n+1}-u_n\|_{\mathcal{V}}^2 \leq \kappa_2(\Delta t)\,\|u_n-u_{n-1}\|_{\mathcal{V}}^2. \tag{2.102}$$

Proof. Subtracting the corresponding equations of the problems (2.101) yielding $\big((u_n,\rho_n),\lambda_n\big)$ and $\big((u_{n+1},\rho_{n+1}),\lambda_{n+1}\big)$, we obtain

$$\begin{aligned}&\frac{1}{\Delta t}\langle u_{n+1}-u_n,v\rangle+a(u_{n+1}-u_n,v)+\beta\langle\lambda_{n+1}-\lambda_n,\eta\rangle+\langle v-\eta,\rho_{n+1}-\rho_n\rangle\\ &\quad=\quad \alpha(F_{u_n}-F_{u_{n-1}})(v)+\frac{1}{\Delta t}\langle u_n-u_{n-1},v\rangle \qquad \forall\,(v,\eta)\in H,\end{aligned} \tag{2.103}$$

and

$$\langle u_{n+1} - u_n - \lambda_{n+1} + \lambda_n, \xi \rangle \;=\; 0 \qquad \forall\, \xi \in Q\,. \tag{2.104}$$

from which, testing (2.103) and (2.104) against $(v, \eta) = (u_{n+1} - u_n, \rho_{n+1} - \rho_n)$ and $\xi = \lambda_{n+1} - \lambda_n$, respectively, we deduce that

$$\begin{aligned} &\frac{1}{\Delta t}\|u_{n+1} - u_n\|_{\mathcal{V}}^2 + a(u_{n+1} - u_n, u_{n+1} - u_n) + \beta\|\lambda_{n+1} - \lambda_n\|_{\mathcal{V}}^2 \\ =\;& \alpha(F_{u_n} - F_{u_{n-1}})(u_{n+1} - u_n) + \frac{1}{\Delta t}\langle u_n - u_{n-1}, u_{n+1} - u_n\rangle\,. \end{aligned}$$

Next, using the ellipticity of a (cf. (A1)), the Lipschitz continuity of $\nabla\mathcal{D}$ (cf. (A2)), and Young's inequality, we arrive at

$$\left(\frac{1}{2\Delta t} + \widetilde{c}_a\right)\|u_{n+1} - u_n\|_{\mathcal{V}}^2 \;\le\; \frac{\alpha}{2}\|u_{n+1} - u_n\|^2 + \left(\frac{L_{\mathcal{D}}\alpha}{2} + \frac{1}{2\Delta t}\right)\|u_n - u_{n-1}\|^2\,,$$

which leads to the desired result after a minor algebraic rearrangment. ■

We stress here that the estimate (2.102) (cf. Lemma 2.14) becomes useless if $\kappa_1(\Delta t) \le 0$. According to it, we now provide a way to bound how small Δt should be in order to guarantee that $\kappa_1(\Delta t) > 0$.

Lemma 2.15. *Problem (2.100) is unconditionally stable in time, that is stable for any fixed time step Δt, if $\alpha < 2\widetilde{c}_a$. It is otherwise stable if $\Delta t < \frac{1}{\alpha - 2\widetilde{c}_a}$.*

Proof. We first observe that if $\alpha < 2\widetilde{c}_a$, then, independently of Δt, $\kappa_1(\Delta t)$ remains always strictly positive, bounded below precisely by $2\widetilde{c}_a - \alpha$. Otherwise, the strict positivity of $\kappa_1(\Delta t)$ is guaranteed only by imposing $\frac{1}{\Delta t} > \alpha - 2\widetilde{c}_a$. ■

Unfortunately, the previous scheme does not guarantee convergence for arbitrary α. Indeed, it is clear from (2.102) that in order to obtain $\|u_{n+1} - u_n\| \le \delta\,\|u_n - u_{n-1}\|$, with $\delta \in]0, 1[$, it suffices to require that $\kappa_2(\Delta t) < \kappa_1(\Delta t)$, which yields the condition $\alpha < \frac{2c_a}{L_{\mathcal{D}}+1}$. Alternatively, if we consider variable time steps, we can prove the following result.

Lemma 2.16. *Let $\left\{\Delta t^k\right\}_{k\in\mathrm{N}}$ be an arbitrary sequence of time steps, and given an initial iteration $\big((u_0, \rho_0), \lambda_0\big) \in H \times Q$ and $n \in \mathrm{N}$, we let $\big((u_n, \rho_n), \lambda_n\big)$ and $\big((u_{n+1}, \rho_{n+1}), \lambda_{n+1}\big)$ be the solutions of (2.101) with $(k, \Delta t) = (n-1, \Delta t^n)$ and $(k, \Delta t) = (n, \Delta t^{n+1})$, respectively. Then, there holds*

$$\kappa_1(\Delta t^{n+1})\,\|u_{n+1} - u_n\|_{\mathcal{V}}^2 \;\le\; \kappa_2(\Delta t^n)\,\|u_n - u_{n-1}\|_{\mathcal{V}}^2\,.$$

Consequently, under the assumption

$$\frac{1}{\Delta t^{n+1}} \;>\; \frac{1}{\Delta t^n} + \alpha(L_{\mathcal{D}} + 1) - 2\widetilde{c}_a\,,$$

the absolute step-wise error is strictly decreasing.

Proof. The derivation of the relationship between $\|u_{n+1}-u_n\|_{\mathcal{V}}^2$ and $\|u_n-u_{n-1}\|_{\mathcal{V}}^2$ is analogous to the one in the proof of Lemma 2.14, except for a minor modification. In fact, as time steps are different, the time derivatives gives rise to new terms which cancel out , that is

$$\left\langle \frac{u_{n+1}}{\Delta t^{n+1}} - \frac{u_n}{\Delta t^n}, u_{n+1}-u_n \right\rangle = \frac{1}{\Delta t^{n+1}}\|u_{n+1}-u_n\|^2 + \left(\frac{1}{\Delta t^{n+1}} - \frac{1}{\Delta t^n}\right)\langle u_n, u_{n+1}-u_n\rangle \ ,$$

and

$$\left\langle \frac{u_n}{\Delta t^{n+1}} - \frac{u_{n-1}}{\Delta t^n}, u_{n+1}-u_n \right\rangle = \frac{1}{\Delta t^n}\langle u_n-u_{n-1}, u_{n+1}-u_n\rangle + \left(\frac{1}{\Delta t^{n+1}} - \frac{1}{\Delta t^n}\right)\langle u_n, u_{n+1}-u_n\rangle \ .$$

Finally, the condition relating the subsequent time steps Δt^{n+1} and Δt^n is obtained by imposing $\kappa_2(\Delta t^n) < \kappa_1(\Delta t^{n+1})$. ■

The above formulation and its associated analysis apply straightforwardly to the mixed case, the only difference being that, while the H^1 inner product is employed in the regularizing terms for the primal case, the $\mathbf{L}^2$ one is utilized for the mixed approach.

2.5 Numerical examples

In this section we present several numerical examples to show the effectiveness of the proposed formulations. All tests were implemented with the FEniCS library [4]. For this, we will use in the primal case the same regularizer used in the mixed formulation, that is the bilinear form defined by (2.43), which arises from the Hooke law for elastic materials. Thus, as already announced at the beginning of Section 2.2.4, the abstract unknown u utilized in Sections 2.2.1, 2.2.2, 2.2.3, and 2.4, is rewritten here as $\boldsymbol{u}$ to denote the respective displacement vector. We consider the problem with null traction boundary conditions so that it kernel is given by the space of rigid motions Q (cf. (2.44)), and consider the similarity functional given by the squared error, i.e.:

$$\mathcal{D}(\mathbf{u}) = \int_\Omega (T(\boldsymbol{x}+\mathbf{u}(\boldsymbol{x})) - R(\boldsymbol{x}))^2 \, ,$$

where the maps $R, T : \Omega \to [0,1]$ denote the reference and target images respectively, and are such that the gradient $\nabla\mathcal{D}$ fulfills condition 2. In what follows we consider the domain $\Omega = (0,1)^2$, and all examples, except for the convergence one, use the classic time regularization scheme described in Section 2.4. Also, only in the real-case study we use the time-adaptivity strategy presented in Section 2.4. For the other examples, we used $\Delta t \propto \alpha^{-1}$ justified by Lemma 2.15, which does not account for the ellipticity constant of the problem but gives satisfactory results nonetheless. The Young modulus E and Poisson ratio ν are related to the Lamé parameters through $\lambda_s = \frac{E\nu}{(1+\nu)(1-2\nu)}$ and $\mu_s = \frac{E}{2(1+\nu)}$.

2.5.1 Example 1: Convergence

We consider the reference and target images

$$R(\boldsymbol{x}) \ = \ \exp\left(-20\|\boldsymbol{x} - 0.3(1,1)\|^2\right)$$

and

$$T(\boldsymbol{x}) \;=\; \exp\left(-20\|\boldsymbol{x} - 0.7(1,1)\|\right),$$

respectively, where $\boldsymbol{x} = (x_1, x_2)^{\mathtt{t}}$ with parameters $\mu_s = \lambda_s = \beta = 1$ and $\alpha = 0.1$. We define also the individual errors

$$\mathtt{e}_0(\boldsymbol{u}) := \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}\,, \quad \mathtt{e}_1(\boldsymbol{u}) := \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}\,, \quad \mathtt{e}_0(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\Omega}\,,$$

$$\mathtt{e}(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}\,, \quad \mathtt{e}(\mathbf{u}) := \|\mathbf{u} - \mathbf{u}_h\|_{0,\Omega}\,, \quad \text{and} \quad \mathtt{e}(\boldsymbol{\Phi}) := \|\boldsymbol{\Phi} - \boldsymbol{\Phi}_h\|_{0,\Omega}\,,$$

and the respective experimental rates of convergence

$$\mathtt{r}_0(\boldsymbol{u}) := \frac{\log\left(\mathtt{e}_0(\boldsymbol{u})/\mathtt{e}_0'(\boldsymbol{u})\right)}{\log\left(h/h'\right)}\,, \quad \mathtt{r}_1(\boldsymbol{u}) := \frac{\log\left(\mathtt{e}_1(\boldsymbol{u})/\mathtt{e}_1'(\boldsymbol{u})\right)}{\log\left(h/h'\right)}\,, \quad \mathtt{r}_0(\boldsymbol{\sigma}) := \frac{\log\left(\mathtt{e}_0(\boldsymbol{\sigma})/\mathtt{e}_0'(\boldsymbol{\sigma})\right)}{\log\left(h/h'\right)}\,,$$

$$\mathtt{r}(\boldsymbol{\sigma}) := \frac{\log\left(\mathtt{e}(\boldsymbol{\sigma})/\mathtt{e}'(\boldsymbol{\sigma})\right)}{\log\left(h/h'\right)}\,, \quad \mathtt{r}(\mathbf{u}) := \frac{\log\left(\mathtt{e}(\mathbf{u})/\mathtt{e}'(\mathbf{u})\right)}{\log\left(h/h'\right)}\,, \quad \mathtt{r}(\boldsymbol{\Phi}) := \frac{\log\left(\mathtt{e}(\boldsymbol{\Phi})/\mathtt{e}'(\boldsymbol{\Phi})\right)}{\log\left(h/h'\right)}\,,$$

where $\mathtt{e}$ y $\mathtt{e}'$, with and without subindex, denote in each case the errors of two consecutive triangulations with meshsizes given by h and h'.

We report the convergence results for the primal (2.11) and mixed (2.53) formulations in Tables 2.1 and 2.2, respectively with respect to a solution of higher resolution, where the mixed scheme is set with the BDM elements described in (2.77). We stress that this problem was solved with a low α, thus results from the point of view of registration are not satisfactory, but they help us to verify convergence, as it is theoretically established for small α without the time stabilization terms (see Section 2.4). In particular, the $O(h)$ and $O(h^2)$ rates of convergence for $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}$ and $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$, respectively, which are predicted by (2.32) (cf. Theorem 2.4) and (2.38) (cf. Theorem 2.5), are confirmed by the sixth and fourth columns of Table 2.1. Nevertheless, the convergence of the extended mixed scheme shown in Table 2.2 seems a bit slow for $\mathtt{e}_0(\boldsymbol{\sigma})$ and slightly oscillating for $\mathtt{e}(\mathbf{u})$, which could be originated by an insufficient number of degrees of freedom employed.

N_{dofs}	$h_{\max}$	$\mathtt{e}_0(\boldsymbol{u})$	$\mathtt{r}_0(\boldsymbol{u})$	$\mathtt{e}_1(\boldsymbol{u})$	$\mathtt{r}_1(\boldsymbol{u})$
56	3.536e-01	1.756e-03	–	1.959e-02	–
168	1.768e-01	5.669e-04	1.631	1.210e-02	0.695
584	8.839e-02	1.636e-04	1.793	6.253e-03	0.952
2184	4.419e-02	4.291e-05	1.931	3.147e-03	0.990
8456	2.210e-02	1.082e-05	1.988	1.575e-03	0.998
33288	1.105e-02	2.649e-06	2.030	7.878e-04	0.999

Table 2.1: Example 1: Errors and convergence rates for the primal extended scheme with $\alpha = 0.1$.

N_{dofs}	$h_{\max}$	$\mathtt{e}_0(\boldsymbol{\sigma})$	$\mathtt{r}_0(\boldsymbol{\sigma})$	$\mathtt{e}(\boldsymbol{\sigma})$	$\mathtt{r}(\boldsymbol{\sigma})$	$\mathtt{e}(\mathbf{u})$	$\mathtt{r}(\mathbf{u})$	$\mathtt{e}(\boldsymbol{\Phi})$	$\mathtt{r}(\boldsymbol{\Phi})$
95	7.071e-01	9.059e-03	–	8.327e-02	–	5.783e+00	–	6.327e+00	–
327	3.536e-01	3.304e-03	1.455	5.103e-02	0.706	2.194e-04	1.469	5.642e-04	1.345
1223	1.768e-01	1.285e-03	1.363	3.217e-02	0.666	1.126e-04	0.960	2.416e-04	1.224
4743	8.839e-02	3.924e-04	1.711	1.632e-02	0.979	5.731e-05	0.975	1.128e-04	1.098
18695	4.419e-02	1.262e-04	1.637	8.122e-03	1.006	3.129e-05	0.873	5.609e-05	1.008

Table 2.2: Example 1: Errors and convergence rates for the mixed extended scheme with $\alpha = 0.1$.

2.5.2 Example 2: To extend or not to extend

In this test we compare the results of the Neumann solver with and without extending the formulation, i.e. without the added degrees of freedom in Q and their corresponding terms to (2.11), which we call the standard formulation. The translation images are defined as in the convergence test, whereas the rotation images are given by

$$R(\boldsymbol{x}) = \varphi(\mathbb{S}\boldsymbol{x}) \quad \text{and} \quad T(\boldsymbol{x}) = \varphi(\mathbb{S}\mathbb{R}\boldsymbol{x})\,,$$

where

$$\mathbb{S} = \begin{bmatrix} 1 & 0 \\ 0 & a \end{bmatrix}, \quad \mathbb{R} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$

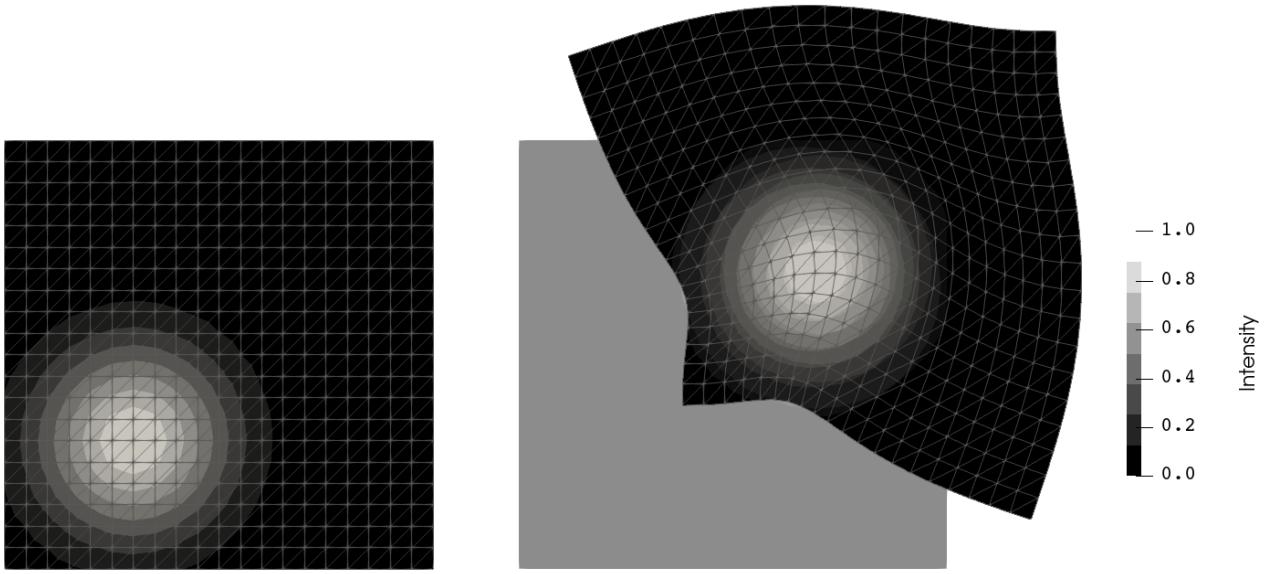
and the function $\varphi(\boldsymbol{x}) = \exp(-C|x|^2)$. The parameters used are given by $E = 10^3$, $\nu = 0.3$, $\alpha = 10^4$, $\Delta t = 0.1/\alpha$, $\beta = 1$, $C = 20$, $a = 0.4$, and the convergence criterion is given by a threshold on the similarity, so that the simulation stops when $\mathcal{D}(\mathbf{u}) \leq 0.01\mathcal{D}(\mathbf{0})$. In Figures 2.1 and 2.2, which display the reference image and the warped reference image with the target image in the background, we notice that both translations and rotations cannot be captured up to the required tolerance without extending the formulation. In this regard we stress that choosing a smaller Δt does not yield convergence in the non-extended scenario. This locking-like phenomenon is seen due to the choice of the convergence criterion, and indeed using another one such as the solution increments would yield convergence to a solution, albeit unsatisfactory.

	Formulation	Iterations	time [s]
Translation	Extended	64	3.516
	Standard	1000	–
Rotation	Extended	51	3.454
	Standard	1000	–

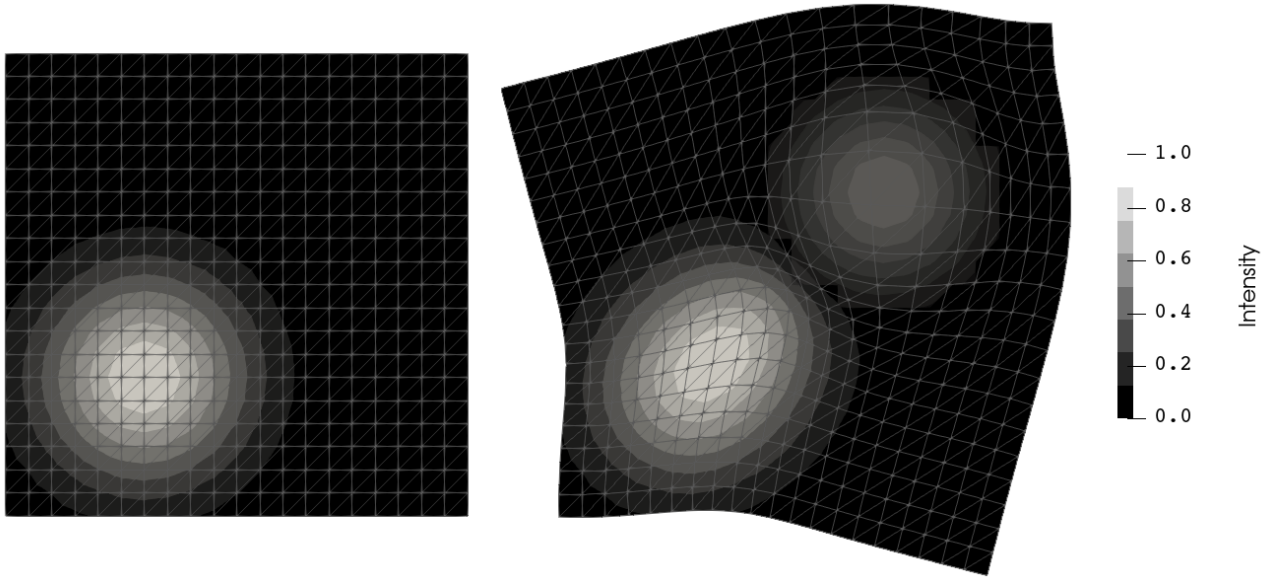
Table 2.3: Example 2: Extended vs. standard in terms of iterations and execution time on a personal computer.

2.5.3 Example 3: Translations in the quasi-incompressible case

In this test we register the translation images for the primal (2.11) and mixed (2.53) formulations, both with $E = 15$, $\nu = 0.4999$, $\alpha = 100$, $\Delta t = 0.1/\alpha$, $\beta = 1$, and time regularization terms



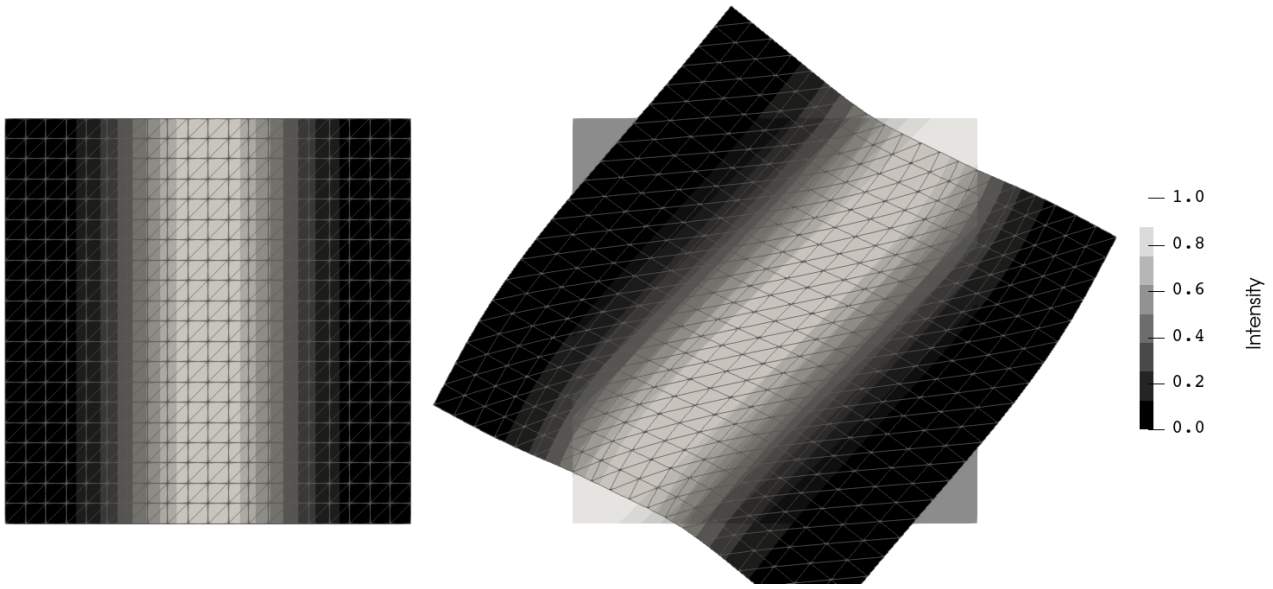
(a) Extended case



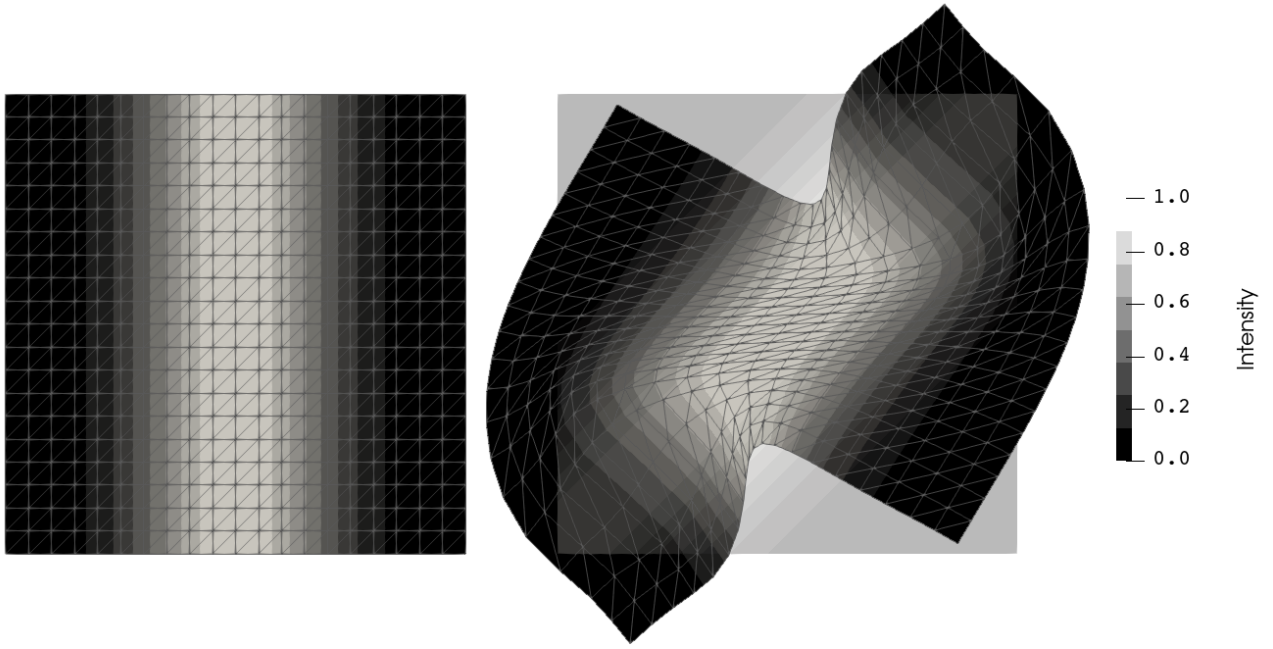
(b) Standard case

Figure 2.1: Example 2: Warped reference images in translation example. We present the reference image $R(\boldsymbol{x})$ in the first column and the deformed reference image $R \circ (\mathbb{I} + \boldsymbol{u}_h)^{-1}(\boldsymbol{x})$ with the target image $T(\vec{x})$ in the background in the second column.

were included and a tolerance of 10^{-8} for the absolute ℓ^∞ error between two subsequent steps was used. The results are reported in Figure 2.3, where the rigid motion components obtained were $\lambda = (0.386, 0.396, 0.022)$ for the primal case and $\lambda = (0.402, 0.381, -0.056)$ for the mixed one. As λ is a rigid motion, the first two components are translations in x and y, whereas the third one represents a rotation. The solution in this case presents no rotation and has by construction a translation of 0.4 in each axis, which is coherent with the results obtained. We highlight that the primal formulation



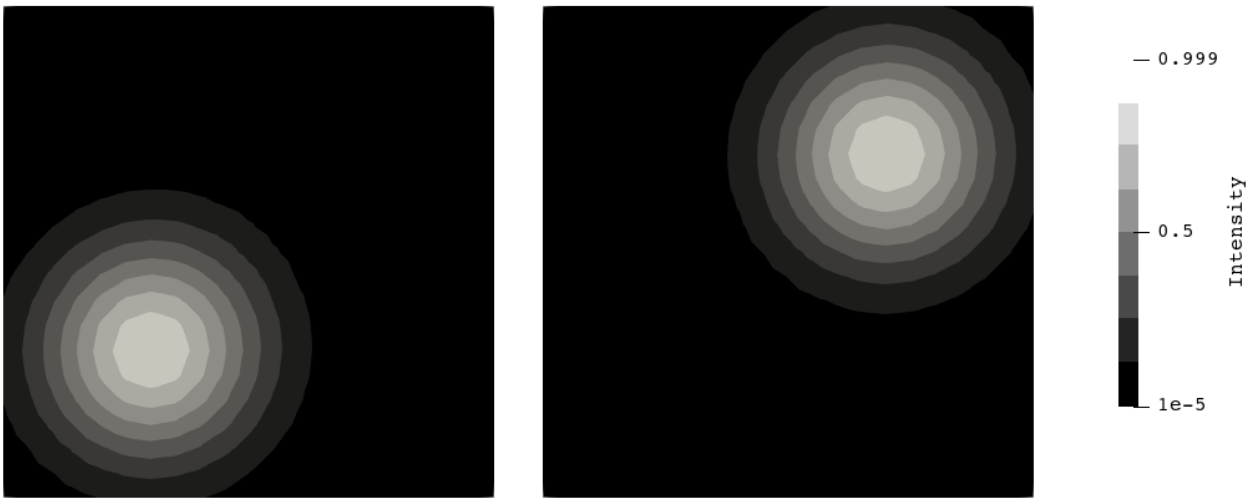
(a) Extended case



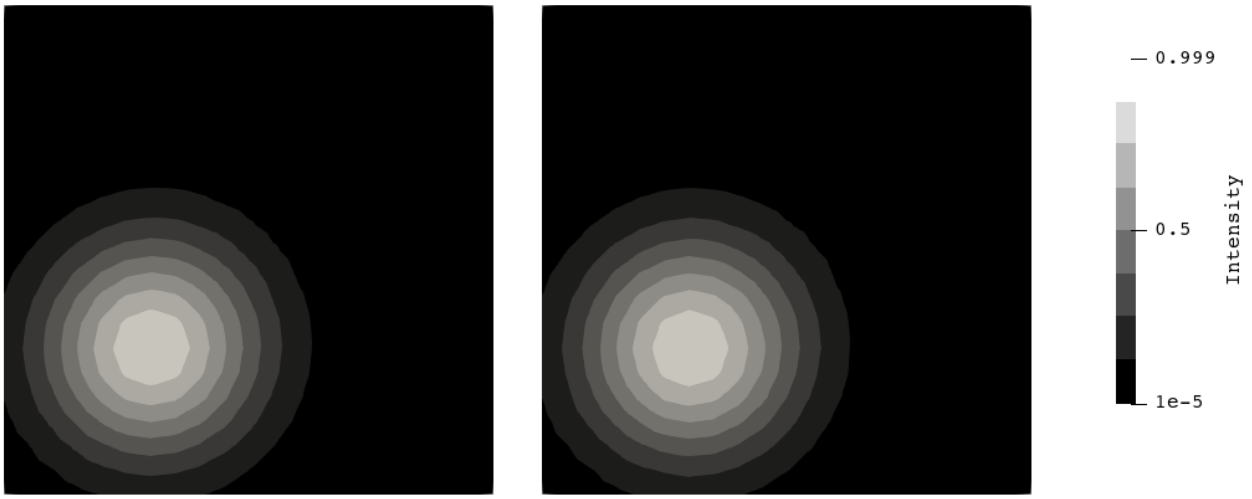
(b) Standard case

Figure 2.2: Example 2: Comparison warped reference images in rotation example. We present the reference image $R(\boldsymbol{x})$ in the first column and the deformed reference image $R \circ (\mathbb{I} + \boldsymbol{u}_h)^{-1}(\boldsymbol{x})$ with the target image $T(\vec{x})$ in the background in the second column.

took 213 iterations to achieve convergence, whereas the mixed one took 102. This difference is mainly due to the locking effects generated by $\nu \approx 0.5$ in the primal formulation, which are fully overcome by the mixed one.



(a) R and T, reference and target images.



(b) Warped target images $T \circ (\mathbb{I} + \boldsymbol{u}_h)$ for primal and mixed formulation.

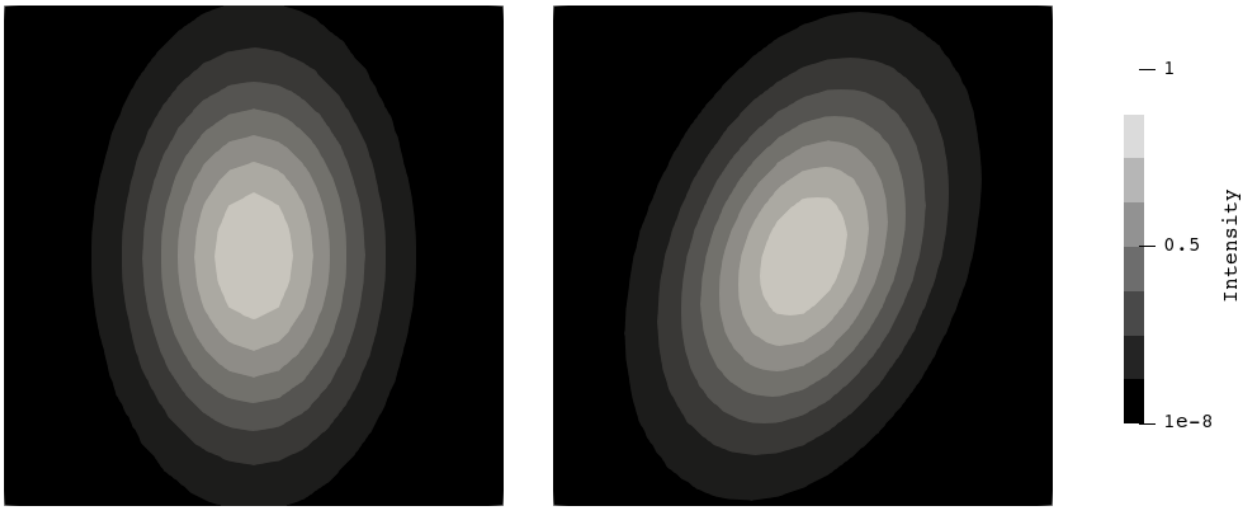
Figure 2.3: Example 3: Solutions of the primal and mixed formulations of the translation test.

2.5.4 Example 4: Rotations in the quasi-incompressible case

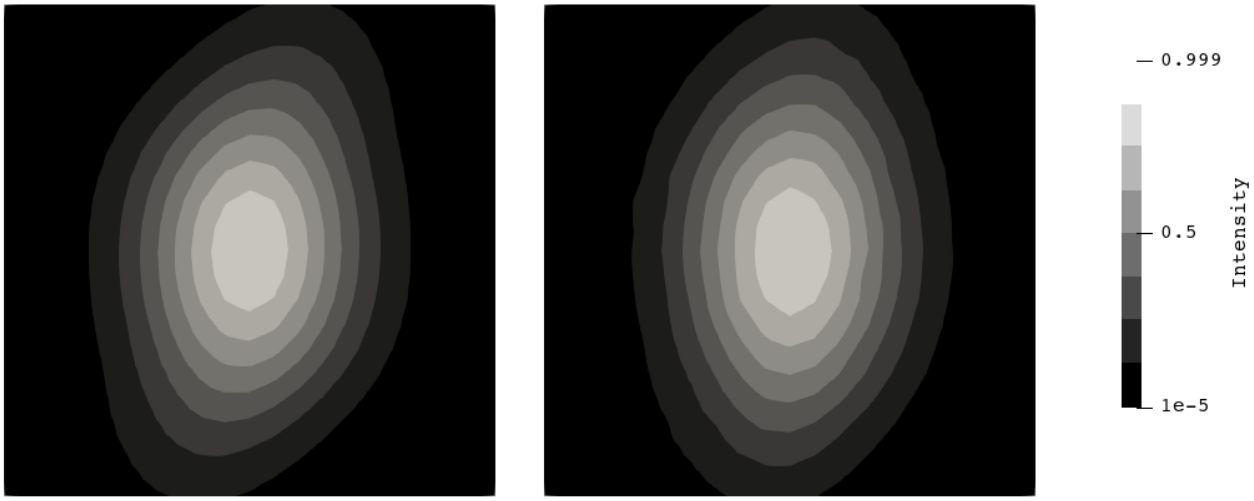
This test was performed for the same settings of the translation example but with the rotation images using $C = 20$ and $a = 0.4$. Results are reported in Figure 2.4, and the rigid motions obtained in this case are

$$\lambda = (-2.843\,10^{-4}, 3.120\,10^{-5}, -7.084\,10^{-2}) \quad \text{and} \quad \lambda = (6.798\,10^{-5}, 6.042\,10^{-4}, -1.476\,10^{-3})\,,$$

for the primal and mixed cases, respectively. We remark that we did not allow for more than 1000 iterations in time, which was achieved by the primal case still without reaching the required tolerance. The mixed one instead converged after 74 iterations, which is again explained by the superiority of the mixed formulation in the quasi-incompressible case.



(a) R and T, reference and target images.



(b) R and T, reference and target images.

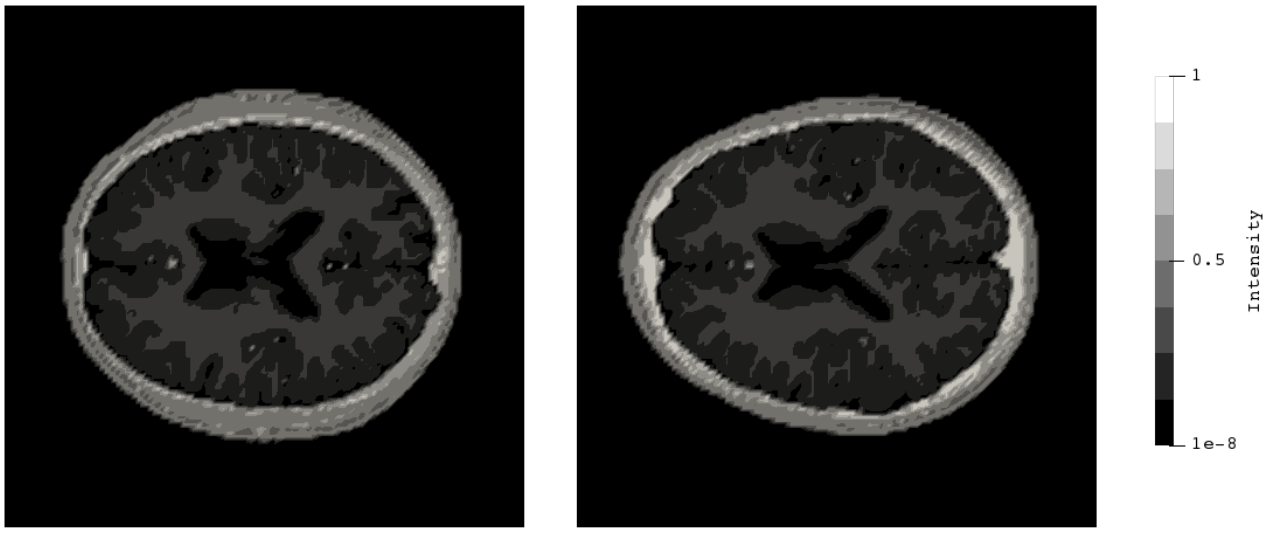
Figure 2.4: Example 4: Solutions of the primal and mixed formulations of the rotation test.

2.5.5 Example 5: Application to the image registration of the human brain

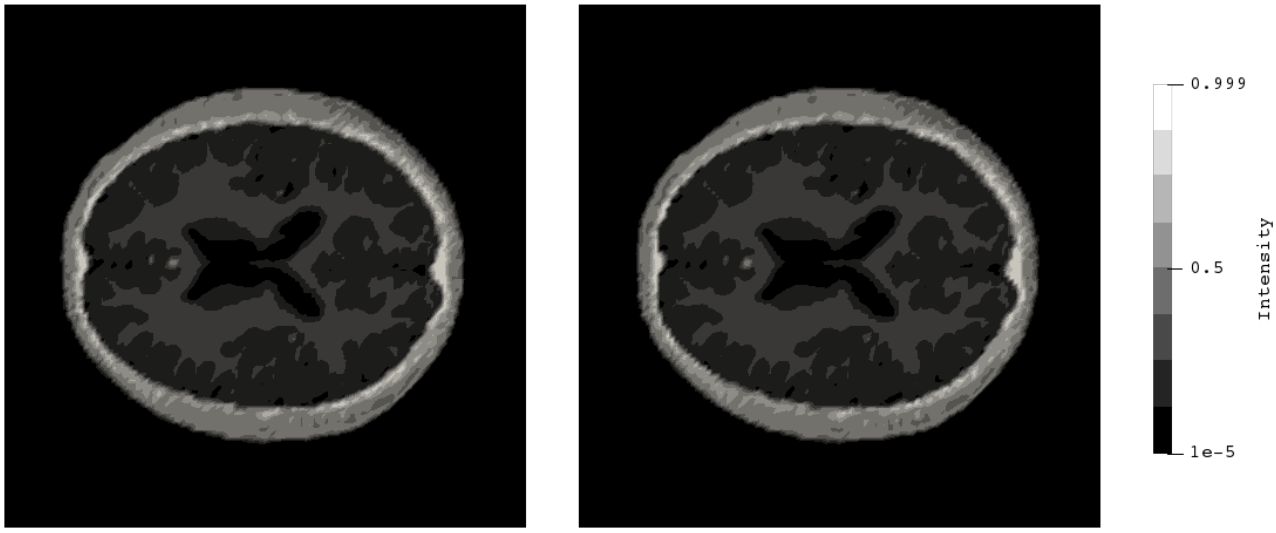
The real application is performed on brain images obtained in [33]. We use this case as well to test the condition on the time step given by

$$\Delta t^{n+1} < \frac{\Delta t^n}{1 + \Delta t^n(\alpha(L_{\mathcal{D}} + 1) - c_a)}. \tag{2.105}$$

Two important observations are in place for condition (2.105). One is that it guarantees the convergence of $\|u_{n+1} - u_n\|$, and not of $\|u_{n+1} - u_n\|/\Delta t_n$, which means that possibly the error performed by means of incorporating the time terms might not disappear. The second one is that it does not stall the simulation within a certain time. To see this, assume $\Delta t^0 = \tau = (\alpha(L_{\mathcal{D}} + 1) - c_a)^{-1}$. This choice gives $\Delta t^n < \tau/(n+1)$, and thus we can not insure that $\sum_n \Delta t^n < \infty$.



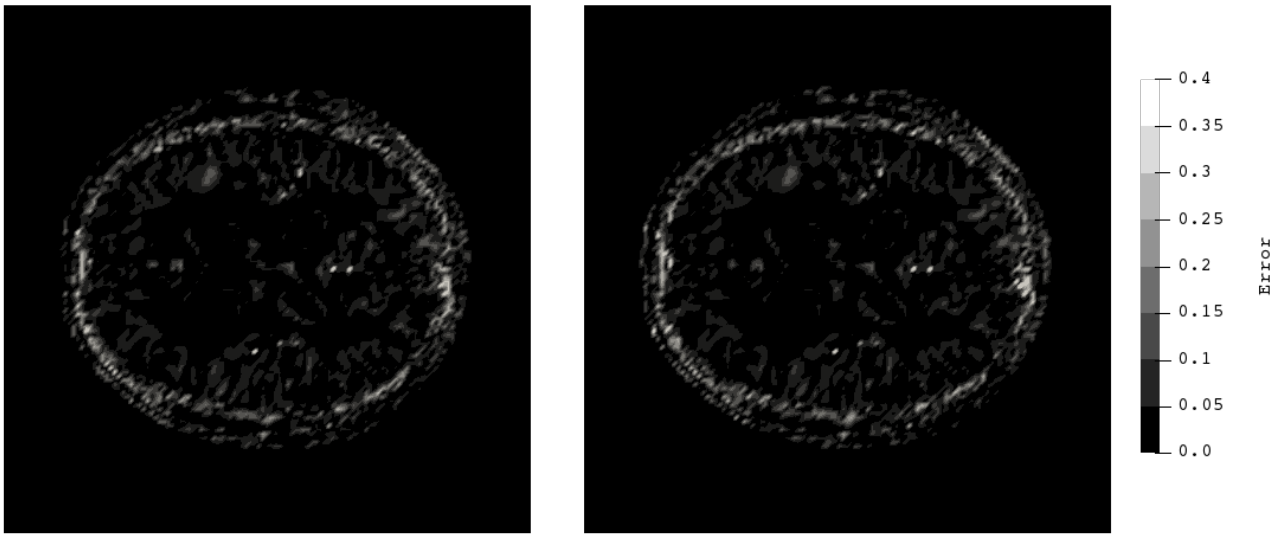
(a) Reference and template images.



(b) Primal and mixed formulation solutions.

Figure 2.5: Example 5: Results of registration for brain images scenario with $\alpha = 10^4$, $\beta = 1$.

In turn, for the simulations we use the elastic constants $E = 15$ and $\nu = 0.3$. For the others constants we consider $\alpha = 10^4, \beta = 1, \Delta t^0 = 0.01/\alpha$ and a tolerance of 10^{-6} for a domain with 128×128 elements. We report the outcome in Figures 2.5 and 2.6, that indicate sufficiently accurate results after convergence. To avoid an excessive reduction of the time step, we used (2.105) every ten iterations.



(a) $|T \circ (\mathbb{I} + \vec{u}) - R|$ for primal and mixed formulations.

Figure 2.6: Example 5: Results of registration for brain images scenario with $\alpha = 10^4$, $\beta = 1$.

CHAPTER 3

Adaptive mesh refinement in deformable image registration: A posteriori error estimates for primal and mixed formulations

3.1 Introduction

Deformable image registration (DIR) consists of aligning two images through a transformation that deforms one image onto the other. It arises in several applications, particularly in the medical imaging field [85]. Its mathematical formulation requires three objects: a transformation model, defined by a family of suitable mappings that warp the target image, a similarity measure, typically represented by a functional that quantifies the difference between images, and a regularizer, which renders the problem well-posed [74]. In addition to the many variants of these components, different modeling approaches exist, between which we highlight the traditional variational minimization [61,74], optimal mass transport [57] and level-set modeling [87]. The solution of the DIR problem typically considers incorporating an auxiliary time variable. This approach can be interpreted as a semi-implicit formulation of the proximal point algorithm [84], recently extended to a more general class of proximal operators by using forward-backward splitting [45]. The formulation of the optical flow problem put forward by Horn & Schunk [61] leads to a more rigorous mathematical analysis of the DIR problem continuous formulation, which is in contrast with the lack of rigorous numerical analysis of the discrete counterpart, recently developed in the variational formulation [79] in an algorithm-specific fashion and also in the optimal-control setting within a more classical Galerkin framework [69].

One active area of DIR application is the study of deformation in the lungs from the analysis of computed-tomography images of the thorax [29]. In this setting, the optimal warping u that solves the DIR problem can be interpreted as a displacement field, from which deformation metrics such as the strain tensor can be computed based on ∇u using the framework of continuum mechanics. The study of deformation using DIR has revealed that the lungs display a highly heterogeneous and anisotropic behavior [63]. Further, deformation metrics from the strain tensor recently proved very insightful in understanding certain pulmonary diseases and lung injury progression [28, 62, 82]. The prediction of strain measures from DIR is not without problems, as it has been shown that estimating the strain tensor from direct differentiation of the transformation mapping yields spurious numerical errors that can distort the physical meaning of the strain tensor [64]. This problem, together with an effort of providing a rigorous analysis of the Galerkin formulation of DIR, motivated the recent development of

primal and mixed continuous formulations and finite-element schemes [13]. This last work used null traction boundary conditions so as to avoid spurious stress. It relied on the mixed theory of linear elasticity problems with pure-traction conditions [50], which delivered *a priori* error estimates not only for the displacement solution but also for the stress and rotation fields in the mixed formulation. These analytical results provide a sound framework for the error assessment of stress and deformation estimates in DIR.

Depending on the amount of warping from the target to the reference images, the optimal warping u can typically result in localized regions with high variations. These localizations may not be accurately captured by the transformation model, which has motivated the development of adaptive refinement techniques in other areas of numerical analysis [91]. However, specific schemes developed for DIR remain understudied. One exception is the work of Haber *et al.* [56], where a finite-difference scheme was employed to solver the DIR problem, and a oct-tree strategy was used to improve the numerical solution by adaptive refinement. Another approach is the use of *ad-hoc* mesh-refinement techniques based on classical strategies in finite-element analysis for elasticity [77, 94]. While very useful, this approach does not directly extend to mixed formulations, and it lacks of a theoretical framework that can guarantee the numerical convergence of the scheme.

In this chapter, we propose *a posteriori* mesh-refinement scheme particularly tailored for primal and mixed formulations of the DIR problem. We start by constructing an optimal *a posteriori* error estimator Θ [91]. The estimator Θ is then decomposed into a sum of local error indicators θ_T that give a norm-wise equivalent of the error. The estimator Θ is said to be reliable (resp. efficient) if there exists $C_{\text{rel}} > 0$ (resp. $C_{\text{eff}} > 0$) independent of the mesh sizes such that

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

where $h.o.t.$ is a generic expression for denoting higher order terms. This estimator is designed to be effective in terms of computing cost, allowing to rapidly identify regions with large error that are candidates to local mesh refinement. At the same time, the use of the estimator prevents the refinement of areas where the error is small, delivering an efficient scheme for error reduction, which is in contrast to uniform refinement schemes. We validate the proposed mesh-refinement scheme and the associated theoretical results through applications on the registration of smooth and medical images, where the performance of the methods is assessed in terms on error measures and convergence rates.

We have organized the contents of this chapter as follows. In Section 3.2, we state the mathematical formulation of DIR, along with the similarity measure and regularizer considered in this work. In Section 3.3, we state the weak problems for the primal and mixed formulations of DIR, along with their corresponding Galerkin schemes. In Section 3.4, we develop *a posteriori* error indicators for the FE formulations, to then derive the corresponding theoretical bounds yielding reliability and efficiency of each estimator under reasonable assumptions. To demonstrate the applicability of the proposed methods, in Section 3.5 we apply the mesh-refinement scheme in the elastic registration of smooth and medical images, where we confirm the reliability and efficiency of the estimators, along with assessing their numerical performance.

3.2 Mathematical formulation of the deformable image registration problem

In this section we recall from [13, Section 2] the elastic deformable image registration model. Let $n \in \{2,3\}$ be the dimension of the images we are interested in analyzing, and let $\Omega \subseteq \mathbb{R}^n$ be a compact domain with Lipschitz boundary $\Gamma := \partial\Omega$. Let $R \in H^1(\Omega)$ be the reference image and $T \in H^1(\widetilde{\Omega})$ be the target image. The DIR problem consists in finding a transformation $\boldsymbol{u} : \Omega \to \mathbb{R}^n$, also known as the displacement field, that best aligns the images R and T, which is expressed as the variational problem (cf. [74])

$$\inf_{\boldsymbol{u}\in\mathcal{V}} \alpha\mathcal{D}[\boldsymbol{u}; R, T] + \mathcal{S}[\boldsymbol{u}], \tag{3.1}$$

where $\mathcal{V}$ is typically $\boldsymbol{H}^1(\Omega)$, $\mathcal{D} : V \to \mathbb{R}$ is the similarity measure between the images R and T, $\alpha > 0$ is a weighting constant, and $\mathcal{S} : V \to \mathbb{R}$ is a regularization term rendering the problem well-posed. A common choice for the similarity measure is the sum of squares difference, i.e, the L^2 error that takes the form

$$\mathcal{D}[\boldsymbol{u}; R, T] := \frac{1}{2}\int_\Omega (T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x}))^2.$$

For the case of elastic DIR, the regularizing term is commonly taken to be the elastic deformation energy, defined by

$$S[\boldsymbol{u}] := \frac{1}{2}\int_\Omega \mathcal{C}\mathbf{e}(\boldsymbol{u}) : \mathbf{e}(\boldsymbol{u}),$$

where

$$\mathbf{e}(\boldsymbol{u}) = \frac{1}{2}\{\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathrm{t}}\}$$

is the infinitesimal strain tensor, i.e., the symmetric component of the displacement field gradient, and $\mathcal{C}$ is the elasticity tensor for isotropic solids, that is

$$\mathcal{C}\boldsymbol{\tau} = \lambda\mathrm{tr}(\boldsymbol{\tau})\mathbb{I} + 2\mu\boldsymbol{\tau} \quad \forall\boldsymbol{\tau} \in \mathbb{L}^2(\Omega). \tag{3.2}$$

Assuming that (3.1) has at least one solution with sufficient regularity, the associated Euler-Lagrange equations deliver the following strong problem: Find $\boldsymbol{u}$ such that

$$\begin{aligned} \mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u})) &= \alpha\boldsymbol{f}_{\boldsymbol{u}} && \text{in} \quad \Omega, \\ \mathcal{C}\mathbf{e}(\boldsymbol{u})\boldsymbol{\nu} &= \mathbf{0} && \text{on} \quad \partial\Omega, \end{aligned} \tag{3.3}$$

where

$$\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{x}) = \big\{T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) - R(\boldsymbol{x})\big\}\nabla T(\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})) \quad \forall\boldsymbol{x} \in \Omega \ a.e. \tag{3.4}$$

We assume the following conditions on the nonlinear load term $\boldsymbol{f}_{\boldsymbol{u}}$:

$$\begin{aligned} |\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{x}) - \boldsymbol{f}_{\boldsymbol{v}}(\boldsymbol{x})| \le L_f|\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})| \quad & \forall\,\boldsymbol{x} \in \Omega \ a.e., \\ |\boldsymbol{f}_{\boldsymbol{u}}(\boldsymbol{x})| \le M_f \qquad\qquad\qquad\quad & \forall\,\boldsymbol{x} \in \Omega \ a.e., \end{aligned} \tag{3.5}$$

where L_f and M_f are positive constants.

3.3 Continuous and discrete weak formulations of DIR

In this section we summarize the continuous primal and mixed variational formulations of (3.3) derived in [13, Section 3] and [13, Section 4], respectively, and recall the respective solvability results.

3.3.1 DIR primal formulation

The primal variational formulation for the registration problem reads: Find $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = \alpha F_{\boldsymbol{u}}(\boldsymbol{v}), \quad \boldsymbol{v} \in \boldsymbol{H}^1(\Omega), \tag{3.6}$$

where $a : \boldsymbol{H}^1(\Omega) \times \boldsymbol{H}^1(\Omega) \to \mathbb{R}$ is the bilinear form defined by

$$a(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \mathcal{C}\mathbf{e}(\boldsymbol{u}) : \mathbf{e}(\boldsymbol{v}) \qquad \forall\, \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{H}^1(\Omega), \tag{3.7}$$

and for every $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$, $F_{\boldsymbol{u}} : \boldsymbol{H}^1(\Omega) \to \mathbb{R}$ is the linear functional given by

$$F_{\boldsymbol{u}}(\boldsymbol{v}) := -\int_\Omega \boldsymbol{f}_{\boldsymbol{u}} \cdot \boldsymbol{v} \qquad \forall\, \boldsymbol{v} \in \boldsymbol{H}^1(\Omega).$$

By imposing the conditions (3.5), we can deduce the Lipschitz continuity and uniform boundedness properties for the functional $F_{\boldsymbol{u}}$, that is

$$\|F_{\boldsymbol{u}} - F_{\boldsymbol{v}}\|_{\boldsymbol{H}^1(\Omega)'} \leq L_F \|\boldsymbol{u} - \boldsymbol{v}\|_{0,\Omega} \qquad \forall\, \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{H}^1(\Omega), \tag{3.8}$$

and

$$\|F_{\boldsymbol{u}}\|_{\boldsymbol{H}^1(\Omega)'} \leq M_F \qquad \forall\, \boldsymbol{u} \in \boldsymbol{H}^1(\Omega),$$

respectively. We recall the results concerning the solvability of (3.6), as developed in [13, Section 3]. First, we define the following linear auxiliary problem: Given $\boldsymbol{z} \in \boldsymbol{H}^1(\Omega)$, find $\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = \alpha F_{\boldsymbol{z}}(\boldsymbol{v}), \quad \boldsymbol{v} \in \boldsymbol{H}^1(\Omega). \tag{3.9}$$

Since this problem does not have unisolvency, we modify it by imposing weak orthogonality to the rigid motions space, denoted by $\mathbb{RM}(\Omega)$ and defined as (see [20, Eq. 11.1.7])

$$\mathbb{RM}(\Omega) := \left\{ \boldsymbol{v} \in \boldsymbol{H}^1(\Omega) : \quad \mathbf{e}(\boldsymbol{v}) = 0 \right\}, \tag{3.10}$$

which guarantees unique solvability of (3.9) since $\mathbb{RM}(\Omega)$ is the null space of its solution operator. Defining

$$H := \mathbb{RM}(\Omega)^\perp = \left\{ \boldsymbol{v} \in \boldsymbol{H}^1(\Omega) : \quad \int_\Omega \boldsymbol{v} = \mathbf{0}, \quad \int_\Omega \operatorname{rot} \boldsymbol{v} = 0 \right\},$$

where $\operatorname{rot} \boldsymbol{v} = -\partial v_1/\partial x_2 + \partial v_2/\partial x_1$, for $\boldsymbol{v} = (v_1, v_2)^{\mathrm{t}}$, we consider the following restricted problem: Given $\boldsymbol{z} \in H$, find $\boldsymbol{u} \in H$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = \alpha F_{\boldsymbol{z}}(\boldsymbol{v}), \quad \boldsymbol{v} \in H. \tag{3.11}$$

Then, we have the following result:

Theorem 3.1. *Given $\boldsymbol{z} \in H$, problem (3.11) has a unique solution $\boldsymbol{u} \in H$, and there exists $C_p > 0$ such that*

$$\|\boldsymbol{u}\|_{1,\Omega} \leq \alpha C_p \|F_{\boldsymbol{z}}\|_{\boldsymbol{H}^1(\Omega)'}.$$

Proof. See [13, Theorem 2]. ■

We now define the operator $\widehat{\mathbf{T}} : H \to H$ given by $\widehat{\mathbf{T}}(\boldsymbol{z}) = \boldsymbol{u}$, where $\boldsymbol{u}$ is the unique solution to problem (3.11) and thus rewrite (3.6) as the fixed-point equation: Find $\boldsymbol{u} \in H$ such that

$$\widehat{\mathbf{T}}(\boldsymbol{u}) = \boldsymbol{u}. \tag{3.12}$$

The following result establishes the existence of solution to the fixed-point equation (3.12).

Theorem 3.2. *Under data assumptions (3.5), the operator $\widehat{\mathbf{T}}$ has at least one fixed point. Moreover, if $\alpha C_p L_F < 1$, the fixed point is unique.*

Proof. See [13, Theorem 3]. ■

3.3.2 DIR mixed formulation

In what follows we introduce a mixed variational formulation of (3.3). We begin by defining an auxiliary field as the skew symmetric component of the displacement field gradient

$$\boldsymbol{\rho} := \frac{1}{2}(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^{\mathrm{t}}).$$

We note that from a continuum mechanics perspective, $\boldsymbol{\rho}$ corresponds to the rotation tensor, which accounts for displacement gradients that do not induce deformation energy. We further define the auxiliary stress tensor field $\boldsymbol{\sigma} := \mathcal{C}\mathbf{e}(\boldsymbol{u})$. Further, we note that the constitutive relation (3.2) can be inverted (cf. [16] or [49]) as

$$\mathcal{C}^{-1}\boldsymbol{\sigma} = \frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{2\mu(2\mu + n\lambda)}\mathrm{tr}(\boldsymbol{\sigma})\mathbb{I}.$$

Then, the strong form of the mixed registration BVP associated with (3.3) becomes: Find $\boldsymbol{u}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ such that

$$\begin{aligned} \mathcal{C}^{-1}\boldsymbol{\sigma} &= \nabla \boldsymbol{u} - \boldsymbol{\rho} && \text{in} \quad \Omega, \\ \mathbf{div}(\boldsymbol{\sigma}) &= \alpha \boldsymbol{f}_{\boldsymbol{u}} && \text{in} \quad \Omega, \\ \boldsymbol{\sigma} &= \boldsymbol{\sigma}^{\mathrm{t}} && \text{in} \quad \Omega, \\ \boldsymbol{\sigma}\boldsymbol{\nu} &= \mathbf{0} && \text{on} \quad \partial\Omega. \end{aligned} \tag{3.13}$$

Introducing the spaces

$$\mathbb{H}_0(\mathbf{div}; \Omega) = \left\{\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega) : \ \gamma_{\boldsymbol{\nu}}\boldsymbol{\tau} = \mathbf{0}\right\},$$

and

$$\boldsymbol{Q} := \boldsymbol{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega),$$

where

$$\mathbb{L}^2_{\rm skew}(\Omega) := \{\boldsymbol{\eta} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\eta}^{\rm t} = -\boldsymbol{\eta}\},$$

and using a standard integration by parts procedure, the weak formulation of the mixed DIR problem (3.13) reads: Find $(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \boldsymbol{Q}$ such that

$$\begin{array}{rcll} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) & = & 0 & \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \\ b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) & = & \alpha F_{\boldsymbol{u}}(\boldsymbol{v}, \boldsymbol{\eta}) & \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}, \end{array} \tag{3.14}$$

where $a : \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbb{H}_0(\mathbf{div}; \Omega) \to \mathbb{R}$ and $b : \mathbb{H}_0(\mathbf{div}; \Omega) \times \boldsymbol{Q} \to \mathbb{R}$ are the bilinear forms defined by

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \int_\Omega \mathcal{C}^{-1}\boldsymbol{\sigma} : \boldsymbol{\tau} \quad \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \tag{3.15}$$

$$b(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta})) := \int_\Omega \boldsymbol{v} \cdot \mathbf{div}\boldsymbol{\tau} + \int_\Omega \boldsymbol{\eta} : \boldsymbol{\tau} \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega),\ \forall (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}. \tag{3.16}$$

In turn, given $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$, $F_{\boldsymbol{u}} : \boldsymbol{Q} \to \mathbb{R}$ is the linear functional defined by

$$F_{\boldsymbol{u}}(\boldsymbol{v}, \boldsymbol{\eta}) := \int_\Omega \boldsymbol{f}_{\boldsymbol{u}} \cdot \boldsymbol{v} \qquad \forall (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}\,.$$

In order to have unisolvency of (3.14), we define the auxiliary problem: Given $\boldsymbol{z} \in \boldsymbol{L}^2(\Omega)$, find $(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \boldsymbol{Q}$ such that

$$\begin{array}{rcll} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) & = & 0 & \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \\ b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) & = & \alpha F_{\boldsymbol{z}}(\boldsymbol{v}, \boldsymbol{\eta}) & \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}, \end{array} \tag{3.17}$$

which corresponds to a mixed formulation of the linear elasticity problem with pure traction boundary conditions. Since this problem does not yield unique solvability, we impose weak orthogonality to the rigid motions space $\mathbb{RM}(\Omega)$ (c.f. (3.10)). Defining $\boldsymbol{H} := \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbb{RM}(\Omega)$, we arrive at the following equivalent mixed variational formulation of (3.17): Given $\boldsymbol{z} \in \boldsymbol{L}^2(\Omega)$, find $((\boldsymbol{\sigma}, \boldsymbol{\chi}), (\boldsymbol{u}, \boldsymbol{\rho})) \in \boldsymbol{H} \times \boldsymbol{Q}$ such that

$$\begin{array}{rcll} A((\boldsymbol{\sigma}, \boldsymbol{\chi}), (\boldsymbol{\tau}, \boldsymbol{\xi})) + B((\boldsymbol{\tau}, \boldsymbol{\xi}), (\boldsymbol{u}, \boldsymbol{\rho})) & = & 0 & \forall\, (\boldsymbol{\tau}, \boldsymbol{\xi}) \in \boldsymbol{H}, \\ B((\boldsymbol{\sigma}, \boldsymbol{\chi}), (\boldsymbol{v}, \boldsymbol{\eta})) & = & \alpha F_{\boldsymbol{z}}(\boldsymbol{v}, \boldsymbol{\eta}) & \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}, \end{array} \tag{3.18}$$

where $A : \boldsymbol{H} \times \boldsymbol{H} \to \mathbb{R}$ and $B : \boldsymbol{H} \times \boldsymbol{Q} \to \mathbb{R}$ are the bilinear forms given by

$$\begin{aligned} A((\boldsymbol{\sigma}, \boldsymbol{\chi}), (\boldsymbol{\tau}, \boldsymbol{\xi})) &:= a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + \int_\Omega \boldsymbol{\chi} \cdot \boldsymbol{\xi} && \forall\, (\boldsymbol{\sigma}, \boldsymbol{\chi}), (\boldsymbol{\tau}, \boldsymbol{\xi}) \in \boldsymbol{H}, \\ B((\boldsymbol{\tau}, \boldsymbol{\xi}), (\boldsymbol{v}, \boldsymbol{\eta})) &:= b(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta})) + \int_\Omega \boldsymbol{\xi} \cdot \boldsymbol{v} && \forall\, ((\boldsymbol{\tau}, \boldsymbol{\xi}), (\boldsymbol{v}, \boldsymbol{\eta})) \in \boldsymbol{H} \times \boldsymbol{Q}. \end{aligned}$$

The following two lemmas are needed to establish the well-posedness of (3.18).

Lemma 3.1. *Let* $\boldsymbol{V} := \{(\boldsymbol{\tau}, \boldsymbol{\xi}) \in \boldsymbol{H} : \ B((\boldsymbol{\tau}, \boldsymbol{\xi}), (\boldsymbol{v}, \boldsymbol{\eta})) = 0,\ \forall (\boldsymbol{v}, \boldsymbol{\eta}) \in \boldsymbol{Q}\}$. *Then* $\boldsymbol{V} = V \times \{\mathbf{0}\}$, *with*

$$V := \{\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div}; \Omega) : \mathbf{div}\,\boldsymbol{\tau} = \mathbf{0} \ \textit{and}\ \boldsymbol{\tau} = \boldsymbol{\tau}^{\rm t} \ \text{in}\ \Omega\}, \tag{3.19}$$

and there exists $\widehat{\alpha} > 0$, *such that*

$$\widehat{\alpha}\|(\boldsymbol{\tau}, \boldsymbol{\xi})\|^2_{\boldsymbol{H}} \leq A((\boldsymbol{\tau}, \boldsymbol{\xi}), (\boldsymbol{\tau}, \boldsymbol{\xi})) \quad \forall\, (\boldsymbol{\tau}, \boldsymbol{\xi}) \in \boldsymbol{V}.$$

Proof. See [50, Lemma 3.3]. ■

Lemma 3.2. *There exists* $\widehat{\beta} > 0$*, such that*

$$\widehat{\beta}\|(\boldsymbol{v},\boldsymbol{\eta})\|_{\boldsymbol{Q}} \leq \sup_{\substack{(\boldsymbol{\tau},\boldsymbol{\xi})\in\boldsymbol{H}\\(\boldsymbol{\tau},\boldsymbol{\xi})\neq\boldsymbol{0}}} \frac{|B((\boldsymbol{\tau},\boldsymbol{\xi}),(\boldsymbol{v},\boldsymbol{\eta}))|}{\|(\boldsymbol{\tau},\boldsymbol{\xi})\|_{\boldsymbol{H}}} \quad \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\boldsymbol{Q}.$$

Proof. See [50, Lemma 3.4]. ■

The well-posedness of the variational formulation (3.18) is stated as follows.

Theorem 3.3. *There exists a unique solution* $((\boldsymbol{\sigma},\boldsymbol{\chi}),(\boldsymbol{u},\boldsymbol{\rho}))\in\boldsymbol{H}\times\boldsymbol{Q}$ *of (3.18). In addition,* $\boldsymbol{\chi}=\boldsymbol{0}$ *and there exist* $C_m>0$*, such that*

$$\|((\boldsymbol{\sigma},\boldsymbol{\chi}),(\boldsymbol{u},\boldsymbol{\rho}))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \leq \alpha C_m \|F_{\boldsymbol{z}}\|_{\boldsymbol{Q}'}.$$

Proof. See [50, Theorem 3.1]. ■

The treatment above allows us to define a fixed-point operator. Let $\boldsymbol{T}:\boldsymbol{L}^2(\Omega)\to\boldsymbol{L}^2(\Omega)$ given by $\boldsymbol{T}(\boldsymbol{z}):=\boldsymbol{u}\ \ \forall\boldsymbol{z}\in\boldsymbol{L}^2(\Omega)$, where $\boldsymbol{u}$ is the displacement component of the unique solution of problem (3.18), and so the mixed formulation (3.14) can be restated as: Find $\boldsymbol{u}\in\boldsymbol{L}^2(\Omega)$ such that

$$\boldsymbol{T}(\boldsymbol{u})=\boldsymbol{u}\,. \tag{3.20}$$

The following result establishes the existence of solution to the fixed-point problem (3.20):

Theorem 3.4. *Under data conditions (3.5) and assuming* $\alpha C_m L_F<1$*, there is a unique fixed point for (3.20). With this, the mixed formulation (3.14) has a unique solution* $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\in\mathbb{H}_0(\mathbf{div};\Omega)\times\boldsymbol{Q}$*. Furthermore*

$$\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_{\mathbb{H}_0(\mathbf{div};\Omega)\times\boldsymbol{Q}} \leq \alpha C_m M_F.$$

Proof. See [13, Theorem 12]. ■

3.3.3 The primal Galerkin finite-element scheme

Let $\boldsymbol{H}_h$ be a finite dimensional subspace of $\boldsymbol{H}^1(\Omega)$ and define $H_h:=\mathbb{RM}^{\perp}\cap\boldsymbol{H}_h$. Then the primal nonlinear discrete problem is: Find $\boldsymbol{u}_h\in H_h$ such that

$$a(\boldsymbol{u}_h,\boldsymbol{v}_h)=\alpha F_{\boldsymbol{u}_h}(\boldsymbol{v}_h),\quad \boldsymbol{v}_h\in H_h. \tag{3.21}$$

Analogously to the continuous case, we consider the auxiliary problem: Given $\boldsymbol{z}_h\in H_h$, find $\boldsymbol{u}_h\in H_h$ such that

$$a(\boldsymbol{u}_h,\boldsymbol{v}_h)=\alpha F_{\boldsymbol{z}_h}(\boldsymbol{v}_h),\quad \boldsymbol{v}_h\in H_h, \tag{3.22}$$

and also let $T_h:H_h\to H_h$ be the discrete operator given by $T_h(\boldsymbol{z}_h)=\boldsymbol{u}_h$, where $\boldsymbol{u}_h$ is the solution to problem (3.22). Considering the same data assumptions as in the continuous case, as well as the continuity and bound obtained before, we arrive at the following result.

Theorem 3.5. *Assume that data assumptions (3.5) hold. Then, the operator T_h has at least one fixed point. Moreover, if $\alpha C_p L_F < 1$, then such fixed point is unique.*

Proof. See [13, Theorem 5]. ■

3.3.4 The mixed Galerkin finite-element scheme

In this section we recall the Galerkin finite-element scheme for (3.14). First, let $\{\mathcal{T}_h\}_{h>0}$ be a regular family of triangulations of the polygonal region $\bar{\Omega}$ by triangles K of diameter h_K with global mesh size $h := \max\{h_K : K \in \mathcal{T}_h\}$, such that they are quasi-uniform around Γ. Let us consider finite dimensional subspaces $H_h^{\boldsymbol{\sigma}}$, $Q_h^{\boldsymbol{u}}$, and $Q_h^{\boldsymbol{\rho}}$ of $\mathbb{H}(\mathbf{div};\Omega)$, $\boldsymbol{L}^2(\Omega)$, and $\mathbb{L}^2_{\mathrm{skew}}(\Omega)$, respectively. Then we introduce the product spaces

$$\boldsymbol{H}_h := (H_h^{\boldsymbol{\sigma}} \cap \mathbb{H}_0(\mathbf{div};\Omega)\,) \times \mathbb{RM}, \qquad \boldsymbol{Q}_h := Q_h^{\boldsymbol{u}} \times Q_h^{\boldsymbol{\rho}},$$

and define the discrete version of (3.18): Given $\boldsymbol{z}_h \in Q_h^{\boldsymbol{u}}$, find $((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) \in \boldsymbol{H}_h \times \boldsymbol{Q}_h$ such that

$$\begin{array}{rcll} A((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{\tau}_h,\boldsymbol{\xi}_h)) + B((\boldsymbol{\tau}_h,\boldsymbol{\xi}_h),(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) & = & 0 & \forall\,(\boldsymbol{\tau}_h,\boldsymbol{\xi}_h) \in \boldsymbol{H}_h, \\ B((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{v}_h,\boldsymbol{\eta}_h)) & = & \alpha F_{\boldsymbol{z}_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h) & \forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h) \in \boldsymbol{Q}_h. \end{array} \tag{3.23}$$

The unique solvability and stability of (3.23), being the Galerkin scheme of a linear elasticity problem with pure traction boundary conditions, has already been established in [50, Theorem 4.1]. This allows us to define the discrete operator $\boldsymbol{T}_h : Q_h^{\boldsymbol{u}} \to Q_h^{\boldsymbol{u}}$ given by $\boldsymbol{T}_h(\boldsymbol{z}_h) := \boldsymbol{u}_h$, where $\boldsymbol{u}_h$ is the unique displacement from (3.23), and then we rewrite the discrete nonlinear problem as: Find $\boldsymbol{u}_h \in Q_h^{\boldsymbol{u}}$ such that

$$\boldsymbol{T}_h(\boldsymbol{u}_h) = \boldsymbol{u}_h. \tag{3.24}$$

Now we establish the well-posedness of problem (3.24).

Theorem 3.6. *Assuming (3.5) and $\alpha C_m L_F < 1$, the problem (3.24) has a unique solution $\boldsymbol{u}_h \in Q_h^{\boldsymbol{u}}$, which yields $((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) \in \boldsymbol{H}_h \times \boldsymbol{Q}_h$ the unique solution of (3.23) with $\boldsymbol{z}_h = \boldsymbol{u}_h$, which satifies*

$$\|((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_{\boldsymbol{H}\times\boldsymbol{Q}} \leq \alpha C_m M_F.$$

Proof. See [13, Theorem 14]. ■

3.4 Residual-based a posteriori error estimators

In this section we derive a reliable and efficient residual-based *a posteriori* error estimator for each one of the Galerkin finite-element schemes (3.21) and (3.23).

3.4.1 Preliminaries

We first let $\mathcal{E}_h$ be the set of all edges of the triangulation $\mathcal{T}_h$, and given $K \in \mathcal{T}_h$, we let $\mathcal{E}(K)$ be the set of its edges. Then we 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 $\mathcal{E}_h(\Gamma) := \{e \in \mathcal{E}_h : e \subseteq \Gamma\}$. Further, h_e stands for the length of a given edge e. Also, for each edge $e \in \mathcal{E}_h$ we fix a unit normal vector $\boldsymbol{\nu}_e := (\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)^{\mathrm{t}}$ and let $\boldsymbol{s}_e := (-\boldsymbol{\nu}_2, \boldsymbol{\nu}_1)^{\mathrm{t}}$ be the corresponding fixed unit tangential vector along e. However, when no confusion arises, we simple write $\boldsymbol{\nu}$ and $\boldsymbol{s}$ instead of $\boldsymbol{\nu}_e$ and $\boldsymbol{s}_e$, respectively. Now, let $\boldsymbol{\tau} \in \mathbb{L}^2(\Omega)$ such that $\boldsymbol{\tau}|_K \in \mathbb{C}(K)$ on each $K \in \mathcal{T}_h$. Then, given $e \in \mathcal{E}_h(\Omega)$, we denote by $[\boldsymbol{\tau}\ \boldsymbol{s}]$ and $[\boldsymbol{\tau}\ \boldsymbol{\nu}]$ the tangential and normal jumps of $\boldsymbol{\tau}$ across e, that is, $[\boldsymbol{\tau}\ \boldsymbol{s}] := (\boldsymbol{\tau}|_K - \boldsymbol{\tau}|_{K'})|_e \boldsymbol{s}$ and $[\boldsymbol{\tau}\ \boldsymbol{\nu}] := (\boldsymbol{\tau}|_K - \boldsymbol{\tau}|_{K'})|_e \boldsymbol{\nu}$, respectively, where K and K' are the triangles of $\mathcal{T}_h$ having e as a common edge. Additionally, given scalar, vector and tensor valued fields v, $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^{\mathrm{t}}$ and $\boldsymbol{\tau} := (\tau_{ij})_{1 \leq i,j \leq 2}$, respectively, we let

$$\mathbf{curl}(v) := \begin{pmatrix} \frac{\partial v}{\partial x_2} \\ -\frac{\partial v}{\partial x_1} \end{pmatrix}, \quad \underline{\mathbf{curl}}(\boldsymbol{\varphi}) := \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_2} & -\frac{\partial \varphi_1}{\partial x_1} \\ \frac{\partial \varphi_2}{\partial x_2} & -\frac{\partial \varphi_2}{\partial x_1} \end{pmatrix}, \quad \mathrm{curl}(\boldsymbol{\tau}) := \begin{pmatrix} \frac{\partial \tau_{12}}{\partial x_1} - \frac{\partial \tau_{11}}{\partial x_2} \\ \frac{\partial \tau_{22}}{\partial x_1} - \frac{\partial \tau_{21}}{\partial x_2} \end{pmatrix}.$$

Next, we collect a few preliminary definitions and results that we need in what follows. Given an integer $k \leq 0$ and $S \subseteq \mathbb{R}^2$, we let $P_k(S)$ be the space of polynomials of degree $\leq k$. Then, we let $I_h : H^1(\Omega) \to X_h$ be the usual Clément interpolation operator (cf. [31]), where

$$X_h := \{v_h \in C(\bar{\Omega}) : v_h|_K \in P_1(K), \quad \forall\, K \in \mathcal{T}_h\}.$$

The following lemma establishes the local approximation properties of I_h.

Lemma 3.3. *There exist constants $c_1, c_2 > 0$, independent of h, such that for all $v \in H^1(\Omega)$ there holds*

$$\|v - I_h(v)\|_{0,K} \leq c_1 h_K \|v\|_{0,\Delta(K)} \quad \forall\, K \in \mathcal{T}_h$$
$$\|v - I_h(v)\|_{0,e} \leq c_2 h_e^{1/2} \|v\|_{0,\Delta(e)} \quad \forall\, e \in \mathcal{E}_h(\Omega) \cup \mathcal{E}_h(\Gamma),$$

where $\Delta(K) := \cup\{K' \in \mathcal{T} : \ K' \cap K \neq \emptyset\}$ and $\Delta(e) := \cup\{K' \in \mathcal{T} : \ K' \cap e \neq \emptyset\}$.

Proof. See [31]. ∎

The main techniques involved below in the proof of efficiency include the localization technique based on element-bubble and edge-bubble functions. Given $K \in \mathcal{T}_h$ and $e \in \mathcal{E}(K)$, we let ψ_K and ψ_e be the usual triangle-bubble and edge-bubble functions [89, eqs. (1.5)-(1.6)], respectively, which satisfy:

(i) $\psi_K \in P_3(K)$, $\psi_K = 0$ on ∂K, $\mathrm{supp}(\psi_K) \subseteq K$, and $0 \leq \psi_K \leq 1$ in K,

(ii) $\psi_e \in P_2(K)$, $\psi_e = 0$ on ∂K, $\mathrm{supp}(\psi_e) \subseteq \omega_e$, and $0 \leq \psi_e \leq 1$ in ω_e,

where $\omega_e := \cup\{K' \in \mathcal{T}_h : \ e \in \mathcal{E}(K')\}$. Additional properties of ψ_K and ψ_e are collected in the following lemma (c.f. [88, Lemma 1.3], [89, Section 3.4] or [90, Section 4]).

Lemma 3.4. *Given $k \in \mathbb{N} \cup \{0\}$, there exist positive constants γ_1, γ_2, γ_3, γ_4 and γ_5, depending only on k and the shape regularity of the triangulations, such that for each $K \in \mathcal{T}_h$ and $e \in \mathcal{E}(K)$, there hold*

$$\begin{aligned}
\gamma_1 \|q\|_{0,K}^2 &\leq \left\|\psi_K^{1/2} q\right\|_{0,K}^2 && \forall\, q \in P_k(K),\\
\|\psi_K q\|_{1,K} &\leq \gamma_2 h_K^{-1} \|q\|_{0,K} && \forall\, q \in P_k(K),\\
\gamma_3 \|p\|_{0,e}^2 &\leq \left\|\psi_e^{1/2} p\right\|_{0,e}^2 && \forall\, p \in P_k(e),\\
\|\psi_e p\|_{1,\omega_e} &\leq \gamma_4 h_e^{-1/2} \|p\|_{0,e} && \forall\, p \in P_k(e),\\
\|\psi_e p\|_{0,\omega_e} &\leq \gamma_5 h_e^{1/2} \|p\|_{0,e} && \forall\, p \in P_k(e).
\end{aligned} \tag{3.25}$$

3.4.2 A posteriori error analysis for the primal finite-element scheme

We now derive a reliable and efficient residual-based *a posteriori* error estimator for (3.21). We draw ideas from [6,7] (see also the monograph [91]). Letting $\boldsymbol{u}_h \in H_h$ be the unique solution of (3.21), we define for each $K \in \mathcal{T}_h$ the *a posteriori* error indicator:

$$\Theta_K^2 := h_K^2 \left\|\alpha \boldsymbol{f}_{\boldsymbol{u}_h} - \mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h))\right\|_{0,K}^2 + \sum_{e \in \mathcal{E}(K) \cap \mathcal{E}_h(\Omega)} h_e \|[\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\boldsymbol{\nu}_e]\|_{0,e}^2 + \sum_{e \in \mathcal{E}(K) \cap \mathcal{E}_h(\Gamma)} h_e \|\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\boldsymbol{\nu}_e\|_{0,e}^2, \tag{3.26}$$

where, according to (3.4),

$$\boldsymbol{f}_{\boldsymbol{u}_h}\big|_K(\boldsymbol{x}) := \big\{T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x})) - R(\boldsymbol{x})\big\} \nabla T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x})) \quad \forall\, \boldsymbol{x} \in K,$$

and introduce the global *a posteriori* error estimator

$$\Theta := \left\{ \sum_{K \in \mathcal{T}_h} \Theta_K^2 \right\}^{1/2}.$$

The following theorem constitutes the main result of this section.

Theorem 3.7. *Let $\boldsymbol{u} \in H$ and $\boldsymbol{u}_h \in H_h$ be the solutions of (3.6) and (3.21), respectively, and assume that $\alpha C_p L_F < 1/2$. Then, there exist constants h_0, C_{rel}, $C_{\mathrm{eff}} > 0$, independent of h, such that for $h \leq h_0$ there holds*

$$C_{\mathrm{eff}} \Theta \leq \|\boldsymbol{u} - \boldsymbol{u}_h\|_H \leq C_{\mathrm{rel}} \Theta. \tag{3.27}$$

The reliability of the global *a posteriori* error estimator (upper bound in (3.27)) and the corresponding efficiency (lower bound in (3.27)) are established in Sections 3.4.2 and 3.4.2, respectively.

Reliability

The upper bound for (3.27) is established as follows.

Lemma 3.5. *Assume that $\alpha C_p L_F < 1/2$. Then, there exist h_0, $C_{\rm rel} > 0$, independent of h, such that for $h \leq h_0$ there holds*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_H \leq C_{\rm rel}\, \Theta.$$

Proof. Let us first define

$$\mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h) := \alpha F_{\boldsymbol{u}}(\boldsymbol{w} - \boldsymbol{w}_h) - a(\boldsymbol{u}_h, \boldsymbol{w} - \boldsymbol{w}_h) \quad \forall \boldsymbol{w}_h \in H_h.$$

As a consequence of the ellipticity of a (c.f (3.7)) with ellipticity constant $\bar{\alpha}$ (c.f. [20, Corollary 11.2.22]), we obtain the following condition

$$\bar{\alpha}\|\boldsymbol{v}\|_{1,\Omega} \leq \sup_{\substack{\boldsymbol{w}\in H\\ \boldsymbol{w}\neq \boldsymbol{0}}} \frac{a(\boldsymbol{v},\boldsymbol{w})}{\|\boldsymbol{w}\|_H} \quad \forall \boldsymbol{v} \in H.$$

In particular, for $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}_h \in H$, we notice from (3.6) and (3.21) that $a(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{w}_h) = 0 \ \forall \boldsymbol{w}_h \in H_h$, and hence we obtain $a(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{w}) = a(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{w} - \boldsymbol{w}_h) = \mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h)$, which yields

$$\bar{\alpha}\|\boldsymbol{u} - \boldsymbol{u}_h\|_H \leq \sup_{\substack{\boldsymbol{w}\in H\\ \boldsymbol{w}\neq \boldsymbol{0}}} \frac{\mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h)}{\|\boldsymbol{w}\|_H} \quad \forall \boldsymbol{w}_h \in H_h. \tag{3.28}$$

From the definition of $\mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h)$, integrating by parts on each $K \in \mathcal{T}_h$, and adding and subtracting a suitable term, we can write

$$\begin{aligned}
\mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h) =\ & \alpha F_{\boldsymbol{u}_h}(\boldsymbol{w} - \boldsymbol{w}_h) + \alpha F_{\boldsymbol{u}}(\boldsymbol{w} - \boldsymbol{w}_h) - a(\boldsymbol{u}_h, \boldsymbol{w} - \boldsymbol{w}_h) - \alpha F_{\boldsymbol{u}_h}(\boldsymbol{w} - \boldsymbol{w}_h),\\
=\ & \alpha \left\{F_{\boldsymbol{u}}(\boldsymbol{w} - \boldsymbol{w}_h) - F_{\boldsymbol{u}_h}(\boldsymbol{w} - \boldsymbol{w}_h)\right\} - \alpha \int_\Omega \boldsymbol{f}_{\boldsymbol{u}_h} \cdot (\boldsymbol{w} - \boldsymbol{w}_h) - \sum_{K\in\mathcal{T}_h} \int_K \mathcal{C}\mathbf{e}(\boldsymbol{u}_h) : \mathbf{e}(\boldsymbol{w} - \boldsymbol{w}_h),\\
=\ & \alpha\{(F_{\boldsymbol{u}} - F_{\boldsymbol{u}_h})(\boldsymbol{w} - \boldsymbol{w}_h)\} - \alpha \int_\Omega \boldsymbol{f}_{\boldsymbol{u}_h} \cdot (\boldsymbol{w} - \boldsymbol{w}_h)\\
& - \sum_{K\in\mathcal{T}_h} \left\{ - \int_K \mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)) \cdot (\boldsymbol{w} - \boldsymbol{w}_h) + \int_{\partial K} (\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\boldsymbol{\nu}_e) \cdot (\boldsymbol{w} - \boldsymbol{w}_h) \right\},\\
=\ & \alpha\{(F_{\boldsymbol{u}} - F_{\boldsymbol{u}_h})(\boldsymbol{w} - \boldsymbol{w}_h)\} + \sum_{K\in\mathcal{T}_h} \int_K (\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)) - \alpha \boldsymbol{f}_{\boldsymbol{u}_h}) \cdot (\boldsymbol{w} - \boldsymbol{w}_h)\\
& - \sum_{e\in\mathcal{E}_h(\Omega)} \int_e [(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\boldsymbol{\nu}_e)] \cdot (\boldsymbol{w} - \boldsymbol{w}_h) - \sum_{e\in\mathcal{E}_h(\Gamma)} \int_e (\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\boldsymbol{\nu}_e) \cdot (\boldsymbol{w} - \boldsymbol{w}_h).
\end{aligned} \tag{3.29}$$

Then, choosing $\boldsymbol{w}_h$ as the Clément interpolant of $\boldsymbol{w}$, that is $\boldsymbol{w}_h := I_h(\boldsymbol{w})$, the approximation properties of I_h (cf. Lemma 3.3) yield

$$\begin{aligned}
\|\boldsymbol{w} - \boldsymbol{w}_h\|_{0,K} \leq & c_1 h_K \|\boldsymbol{w}\|_{1,\Delta(K)},\\
\|\boldsymbol{w} - \boldsymbol{w}_h\|_{0,e} \leq & c_2 h_e \|\boldsymbol{w}\|_{1,\Delta(e)}.
\end{aligned} \tag{3.30}$$

In this way, applying the Cauchy-Schwarz inequality to each term (3.29), and making use of (3.30) together with the Lipschitz continuity of $F_{\boldsymbol{u}}$ (cf. (3.8)), we obtain

$$\begin{aligned}
\mathcal{R}_h(\boldsymbol{w} - \boldsymbol{w}_h) \ \leq\ & \alpha c_1 L_F h_K \|\boldsymbol{u} - \boldsymbol{u}_h\|_H \|\boldsymbol{w}\|_{1,\Delta(K)}\\
& + \widehat{C} \left\{ \sum_{K\in\mathcal{T}_h} \Theta_K^2 \right\}^{1/2} \left\{ \sum_{K\in\mathcal{T}_h} \|\boldsymbol{w}\|^2_{1,\Delta(K)} + \sum_{e\in\mathcal{E}_h(\Omega)} \|\boldsymbol{w}\|^2_{1,\Delta(e)} \right\}^{1/2},
\end{aligned}$$

where $\widehat{C}$ is a constant depending on c_1 and c_2 and Θ_K^2 defined by (3.26). Additionally using the fact that the number of triangles in $\Delta(K)$ and $\Delta(e)$ are bounded, we have

$$\sum_{K\in\mathcal{T}_h} \|\boldsymbol{w}\|^2_{1,\Delta(K)} \leq C_1\|\boldsymbol{w}\|^2_{1,\Omega} \qquad \text{and} \qquad \sum_{e\in\mathcal{E}_h(\Omega)} \|\boldsymbol{w}\|^2_{1,\Delta(e)} \leq C_2\|\boldsymbol{w}\|^2_{1,\Omega}$$

where C_1, C_2 are positive constant, and using that $\alpha C_p L_F \leq 1/2$, it follows that $h_0 := 1/(2c_1\alpha L_F)$, finally substituting in (3.28), we conclude that

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_H \leq C_{\mathrm{rel}}\,\Theta,$$

where C_{rel} is independent of h. ■

Efficiency

Now we focus on establish the lower bound in (3.27). We begin with the following lemma whose proof is a slight modification of [90, Section 6].

Lemma 3.6. *There exist constants $\eta_1,\eta_2,\eta_3>0$, independent of h, but depending on $\gamma_1,\gamma_2,\gamma_3,\gamma_4$ and γ_5 (c.f. (3.25)), such that for each $K\in\mathcal{T}_h$ there holds*

$$h_K\left\|\alpha\boldsymbol{f}_{\boldsymbol{u}_h}-\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h))\right\|_{0,K} \leq \eta_1\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,K},$$

$$h_e^{1/2}\|[\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\cdot\boldsymbol{\nu}_e]\|_{0,e} \leq \eta_2\left\{\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\omega_e}+\sum_{K\in\omega_e} h_K\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,K}\right\},$$

$$h_e^{1/2}\|\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)\cdot\boldsymbol{\nu}_e\|_{0,e} \leq \eta_3\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,K},$$

where $\omega_e := \cup\{K'\in\mathcal{T}_h:\ e\in\mathcal{E}(K')\}$.

Proof. Using the first inequality in (3.25), and let $R_K(\boldsymbol{u}_h) := \alpha\boldsymbol{f}_{\boldsymbol{u}_h}-\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h))$ we have

$$\|R_K(\boldsymbol{u}_h)\|^2_{0,K} \leq \gamma_1^{-1}\left\|\psi_K^{1/2}R_K(\boldsymbol{u}_h)\right\|^2_{0,K},$$

$$= \gamma_1^{-1}\int_K \psi_K R_K(\boldsymbol{u}_h)\{\alpha\boldsymbol{f}_{\boldsymbol{u}_h}-\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h))\},$$

$$= \gamma_1^{-1}\int_K \alpha\psi_K R_K(\boldsymbol{u}_h)\{\boldsymbol{f}_{\boldsymbol{u}_h}-\boldsymbol{f}_{\boldsymbol{u}}\}-\gamma_1^{-1}\int_K \psi_K R_K(\boldsymbol{u}_h)\{\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)-\mathcal{C}\mathbf{e}(\boldsymbol{u}))\},$$

$$= \gamma_1^{-1}\int_K \alpha\psi_K R_K(\boldsymbol{u}_h)\{\boldsymbol{f}_{\boldsymbol{u}_h}-\boldsymbol{f}_{\boldsymbol{u}}\}+\gamma_1^{-1}\int_K (\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)-\mathcal{C}\mathbf{e}(\boldsymbol{u}))\cdot\nabla(\psi_K R_K(\boldsymbol{u}_h)),$$

$$\leq \alpha\gamma_1^{-1}\|R_K(\boldsymbol{u}_h)\|_{0,K}\|\boldsymbol{f}_{\boldsymbol{u}_h}-\boldsymbol{f}_{\boldsymbol{u}}\|_{0,K}+\gamma_1^{-1}\gamma_2 h_K^{-1}\|\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)-\mathcal{C}\mathbf{e}(\boldsymbol{u})\|_{0,K}\|R_K(\boldsymbol{u}_h)\|_{0,K},$$

where, for the last inequality we used the inverse inequality (second relation in (3.25)). Next, we have

$$h_K\|R_K(\boldsymbol{u}_h)\|_{0,K} \leq \alpha h_K\gamma_1^{-1}\|\boldsymbol{f}_{\boldsymbol{u}_h}-\boldsymbol{f}_{\boldsymbol{u}}\|_{0,K}+\gamma_1^{-1}\gamma_2\|\mathcal{C}\mathbf{e}(\boldsymbol{u}_h)-\mathcal{C}\mathbf{e}(\boldsymbol{u})\|_{0,K},$$

now, using (3.5) and grouping terms, we conclude with $\eta_1>0$ independent of h, that

$$h_K\left\|\alpha\boldsymbol{f}_{\boldsymbol{u}_h}-\mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}_h))\right\|_{0,K} \leq \eta_1\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,K},$$

We omit further details and repeating arguments used for the remaining inequalities. ■

3.4.3 A posteriori error analysis for the mixed finite-element scheme

In this section we derive a reliable and efficient residual-based *a posteriori* error estimator for (3.23). Throughout the rest of this section we let $((\boldsymbol{\sigma},\boldsymbol{\chi}),(\boldsymbol{u},\boldsymbol{\rho}))\in \boldsymbol{H}\times\boldsymbol{Q}$ and $((\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\in \boldsymbol{H}_h\times\boldsymbol{Q}_h$ be the solutions of the continuous and discrete formulations (3.18) and (3.23), respectively. We introduce the global *a posteriori* error estimator

$$\Psi := \left\{\sum_{K\in\mathcal{T}_h}\Psi_K^2\right\}^{1/2},$$

where we define for each $K\in\mathcal{T}_h$

$$\begin{aligned}\Psi_K^2 := {} & \left\|\alpha\boldsymbol{f}_{\boldsymbol{u}_h}-\mathbf{div}\boldsymbol{\sigma}_h\right\|_{0,K}^2+\|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,K}^2+\|\boldsymbol{\chi}_h\|_{0,K}^2+h_K^2\|\mathrm{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h+\boldsymbol{\rho}_h)\|_{0,K}^2\\ & +h_K^2\|\mathcal{C}^{-1}\boldsymbol{\sigma}_h+\boldsymbol{\rho}_h\|_{0,K}^2+\sum_{e\in\mathcal{E}(K)\cap\mathcal{E}_h(\Omega)}h_e\|[(\mathcal{C}^{-1}\boldsymbol{\sigma}_h+\boldsymbol{\rho}_h)\mathbf{s}]\|_{0,e}^2\\ & +\sum_{e\in\mathcal{E}(K)\cap\mathcal{E}_h(\Gamma)}h_e\|(\mathcal{C}^{-1}\boldsymbol{\sigma}_h+\boldsymbol{\rho}_h)\mathbf{s}\|_{0,e}^2.\end{aligned}\tag{3.31}$$

The following theorem constitutes the main result of this section.

Theorem 3.8. *Assume that $\alpha C_m L_F<1/2$. Then, there exist $C_{\mathrm{rel}},C_{\mathrm{eff}}>0$ independent of h, such that*

$$C_{\mathrm{eff}}\Psi\le\|(\boldsymbol{\sigma},\boldsymbol{\chi})-(\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h)\|_{\boldsymbol{H}}+\|(\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_{\boldsymbol{Q}}\le C_{\mathrm{rel}}\Psi.\tag{3.32}$$

The reliability of the global error estimator (upper bound in (3.32)) and the corresponding efficiency (lower bound in (3.32)) are established in Sections 3.4.3 and 3.4.3, respectively.

Reliability

We begin by establishing a more general result due to Lemmas 3.1, 3.2 and Theorem 3.4, and that we will use to establish the upper bound in (3.32). This result we establish in the following theorem.

Theorem 3.9. *Given $\bar{F}\in\boldsymbol{H}'$ and $\bar{G}_{\boldsymbol{u}}\in\boldsymbol{Q}'$, there exists a unique $((\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}}),(\bar{\boldsymbol{u}},\bar{\boldsymbol{\rho}}))\in\boldsymbol{H}\times\boldsymbol{Q}$ such that*

$$\begin{aligned}A((\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}}),(\boldsymbol{\tau},\boldsymbol{\xi}))+B((\boldsymbol{\tau},\boldsymbol{\xi}),(\bar{\boldsymbol{u}},\bar{\boldsymbol{\rho}}))&=\bar{F}((\boldsymbol{\tau},\boldsymbol{\xi}))\qquad\forall(\boldsymbol{\tau},\boldsymbol{\xi})\in\boldsymbol{H},\\ B((\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}}),(\boldsymbol{v},\boldsymbol{\eta}))&=\bar{G}_{\boldsymbol{u}}((\boldsymbol{v},\boldsymbol{\eta}))\qquad\forall(\boldsymbol{v},\boldsymbol{\eta})\in\boldsymbol{Q}.\end{aligned}\tag{3.33}$$

In addition, there exists $C>0$, depending only on $\widehat{\alpha}$, $\widehat{\beta}$, $\|a\|$, and $\|b\|$, such that

$$\|(\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}})\|_{\boldsymbol{H}}+\|(\bar{\boldsymbol{u}},\bar{\boldsymbol{\rho}})\|_{\boldsymbol{Q}}\le C\{\|\bar{F}\|_{\boldsymbol{H}'}+\|\bar{G}_{\boldsymbol{u}}\|_{\boldsymbol{Q}'}\}.\tag{3.34}$$

To derive an upper bound for $\|(\boldsymbol{\sigma},\boldsymbol{\chi})-(\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h)\|_{\boldsymbol{H}}$ we consider the functional $S_h:\mathbb{H}(\mathbf{div};\Omega)\to\mathbb{R}$ defined by

$$S_h(\boldsymbol{\tau}):=a(\boldsymbol{\sigma}_h,\boldsymbol{\tau})+b(\boldsymbol{\tau},(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\qquad\forall\,\boldsymbol{\tau}\in\mathbb{H}(\mathbf{div};\Omega),\tag{3.35}$$

where a and b are the bilinear forms defined in (3.15) and (3.16), respectively, and let $S_h|_V$ be the restriction of S to V, the first component of the kernel $\mathbf{V}$ of B (cf. (3.19)) We note that $S_h(\boldsymbol{\tau}_h) = 0$ for each $\boldsymbol{\tau}_h \in H_h^{\boldsymbol{\sigma}}$.

Now, we make use of a particular problem of the form (3.33) with $\bar{F} \in \boldsymbol{H}'$ and $\bar{G}_{\boldsymbol{u}} \in \boldsymbol{Q}'$ defined by

$$\bar{F}\left((\boldsymbol{\tau},\boldsymbol{\xi})\right) := 0 \ \ \forall\, (\boldsymbol{\tau},\boldsymbol{\xi}) \in \boldsymbol{H} \quad \text{and} \quad \bar{G}_{\boldsymbol{u}}((\boldsymbol{v},\boldsymbol{\eta})) := B((\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h),(\boldsymbol{v},\boldsymbol{\eta})) \ \ \forall\, (\boldsymbol{v},\boldsymbol{\eta}) \in \boldsymbol{Q},$$

and let $((\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}}),(\bar{\boldsymbol{u}},\bar{\boldsymbol{\rho}})) \in \boldsymbol{H} \times \boldsymbol{Q}$ be the unique solution of this particular problem. We note that

$$\bar{G}_{\boldsymbol{u}}((\boldsymbol{v},\boldsymbol{\eta})) = \int_\Omega (\alpha \boldsymbol{f}_{\boldsymbol{u}} - \mathbf{div}\boldsymbol{\sigma}_h)\cdot \boldsymbol{v} - \int_\Omega \boldsymbol{\chi}_h \cdot \boldsymbol{v} - \int_\Omega \boldsymbol{\sigma}_h : \boldsymbol{\eta},$$

this conforming the definition of B and the second equation of (3.18). Adding and subtracting a suitable term we can rewrite the above equation as:

$$\bar{G}_{\boldsymbol{u}}((\boldsymbol{v},\boldsymbol{\eta})) = \int_\Omega \left(\alpha \boldsymbol{f}_{\boldsymbol{u}_h} - \mathbf{div}\boldsymbol{\sigma}_h\right)\cdot \boldsymbol{v} - \int_\Omega \boldsymbol{\chi}_h \cdot \boldsymbol{v} - \int_\Omega \boldsymbol{\sigma}_h : \boldsymbol{\eta} + \alpha \int_\Omega (\boldsymbol{f}_{\boldsymbol{u}} - \boldsymbol{f}_{\boldsymbol{u}_h})\cdot \boldsymbol{v}.$$

Applying Cauchy-Schwarz inequality and noting that $\boldsymbol{\sigma}_h : \boldsymbol{\eta} = \frac{1}{2}(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}) : \boldsymbol{\eta}$, together with the condition (3.5), we can establish

$$\|\bar{G}_{\boldsymbol{u}}\|_{\boldsymbol{Q}'} \leq C \left\{ \left\|\alpha \boldsymbol{f}_{\boldsymbol{u}_h} - \mathbf{div}\boldsymbol{\sigma}_h\right\|_{0,\Omega} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega} + \|\boldsymbol{\chi}_h\|_{0,\Omega} + \alpha L_F \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega} \right\},$$

by the previous estimate and the continuous dependence results (3.34), we have

$$\|(\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}})\|_{\boldsymbol{H}} \leq C \left\{ \left\|\alpha \boldsymbol{f}_{\boldsymbol{u}_h} - \mathbf{div}\boldsymbol{\sigma}_h\right\|_{0,\Omega} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega} + \|\boldsymbol{\chi}_h\|_{0,\Omega} + \alpha L_F \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega} \right\}. \tag{3.36}$$

Now, applying the triangle inequality we obtain

$$\|(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h)\|_{\boldsymbol{H}} \leq \|(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h) - (\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}})\|_{\boldsymbol{H}} + \|(\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}})\|_{\boldsymbol{H}}, \tag{3.37}$$

and hence, it remains to estimate $\|(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h) - (\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}})\|_{\boldsymbol{H}}$. First observe that $(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h) - (\bar{\boldsymbol{\sigma}},\bar{\boldsymbol{\chi}}) \in \boldsymbol{V}$, hence applying the ellipticity of A in $\boldsymbol{V}$ (cf. Lemma 3.1) and analogously to [43, Lemma 4.6], we obtain an estimate for this term that replacing together with (3.36) in (3.37), allows us to establish that

$$\|(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h)\|_{\boldsymbol{H}} \leq C \left\{ \|S_h|_V\|_{V'} + \left\|\alpha \boldsymbol{f}_{\boldsymbol{u}_h} - \mathbf{div}\boldsymbol{\sigma}_h\right\|_{0,\Omega} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega} + \|\boldsymbol{\chi}_h\|_{0,\Omega} + \alpha L_F \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega} \right\}. \tag{3.38}$$

To estimate $\|S_h|_V\|_{V'}$, (cf. (3.35)) in (3.38), we have the following result

Lemma 3.7. *There exists $C > 0$, such that*

$$\begin{aligned} \|S_h|_V\|_{V'} \leq C \Bigg\{ & h_K^2 \|\mathrm{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\|_{0,K}^2 + \sum_{e \in \mathcal{E}(K)\cap \mathcal{E}_h(\Omega)} h_e \|[(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}]\|_{0,e}^2 \\ & + \sum_{e \in \mathcal{E}(K)\cap \mathcal{E}_h(\Gamma)} h_e \|(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}\|_{0,e}^2 \Bigg\}. \end{aligned} \tag{3.39}$$

Proof. See [43, Lemma 4.7] for details. ■

From the above, the following lemma is configured.

Lemma 3.8. *Assume that $\alpha C_m L_F < 1/2$. Then, there exists $C > 0$ such that*

$$\|(\boldsymbol{\sigma},\boldsymbol{\chi}) - (\boldsymbol{\sigma}_h,\boldsymbol{\chi}_h)\|_{\boldsymbol{H}} \leq C \left\{ \sum_{K\in\mathcal{T}_h} \widetilde{\Psi}_K^2 \right\}^{1/2},$$

where

$$\widetilde{\Psi}_K^2 := h_K^2 \|\mathrm{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\|_{0,K}^2 + \sum_{e\in\mathcal{E}(K)\cap\mathcal{E}_h(\Omega)} h_e \|[(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}]\|_{0,e}^2 + \sum_{e\in\mathcal{E}(K)\cap\mathcal{E}_h(\Gamma)} h_e \|(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}\|_{0,e}^2$$
$$+ \|\alpha \boldsymbol{f}_{\boldsymbol{u}} - \mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega} + \|\boldsymbol{\chi}_h\|_{0,\Omega} + \alpha L_F \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}.$$

Proof. It follows straightforwardly from (3.38) and (3.39). ■

Now we proceed to obtain the corresponding upper bound for $\|(\boldsymbol{u},\boldsymbol{\rho}) - (\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_{\boldsymbol{Q}}$.

Lemma 3.9. *Assume that $\alpha C_m L_F < 1/2$. Then, there exists $C > 0$ such that*

$$\|(\boldsymbol{u},\boldsymbol{\rho}) - (\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_{\boldsymbol{Q}} \leq C \left\{ \sum_{K\in\mathcal{T}_h} \Psi_K^2 \right\}^{1/2},$$

where Ψ_K^2 is the local indicator defined in (3.31).

Proof. The proof follows directly from [43, Lemma 4.9] with small modifications. ■

The reliability of Ψ, is a straightforward consequence of Lemmas 3.8 and 3.9, assuming $\alpha C_m L_F < 1/2$.

Efficiency

In this section, we provide upper bounds depending on the actual errors for the seven terms defining the local indicator Ψ_K^2 (c.f. (3.31)). For this, analogously to [43, Section 4.3] we begin with the first three ones appearing there, more precisely, since $\mathbf{div}(\boldsymbol{\sigma}) = \alpha \boldsymbol{f}_{\boldsymbol{u}}$ in Ω, we have that

$$\|\alpha \boldsymbol{f}_{\boldsymbol{u}} - \mathbf{div}\boldsymbol{\sigma}_h\|_{0,K}^2 \leq \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div},K}^2.$$

Next, adding and subtracting $\boldsymbol{\sigma}$, and we use that $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathrm{t}}$ in Ω, we see that

$$\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,K}^2 \leq 4\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2.$$

Finally, since $\boldsymbol{\chi} = \mathbf{0}$, we obtain

$$\|\boldsymbol{\chi}_h\|_{0,K}^2 = \|\boldsymbol{\chi} - \boldsymbol{\chi}_h\|_{0,K}^2.$$

The upper bounds for the terms involving only the tensor $\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h$, are established in the following result.

Lemma 3.10. *There exist $C_1, C_2, C_3, C_4 > 0$, independent of h, such that for each $K \in \mathcal{T}_h$ there holds*

$$
\begin{aligned}
h_K^2 \|\mathrm{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\|_{0,K}^2 &\leq C_1 \left\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2 \right\} \\
\|\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h\|_{0,K}^2 &\leq C_2 \left\{ \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K}^2 + h_K^2 \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + h_K^2 \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2 \right\} \\
h_e \|[(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}]\|_{0,e}^2 &\leq C_3 \sum_{K \subseteq \omega_e} \left\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2 \right\} \\
\sum_{e \in \mathcal{E}_h(\Gamma)} h_e \|(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\boldsymbol{s}\|_{0,e}^2 &\leq C_4 \sum_{e \in \mathcal{E}_h(\Gamma)} \left\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2 \right\},
\end{aligned}
$$

where $\omega_e := \cup \{K' \in \mathcal{T}_h : \ e \in \mathcal{E}(K')\}$.

Proof. See [43, Section 4.3]. ∎

3.5 Applications and performance assessment

3.5.1 Numerical implementation

We now turn to the implementation of some numerical tests that confirm the predicted reliability and efficiency of the *a posteriori* error estimators (3.26) and (3.31). The DIR problem is in all cases restricted to images mapped to the unit square $\Omega = (0,1)^2$, and uniform triangular partitions are employed for all initial meshes. The discretization of the primal problem is done with continuous piecewise linear and continuous piecewise quadratic approximations for displacement. For the case of the mixed formulation we consider the lowest-order family of Brezzi-Douglas-Marini elements for the rows of the Cauchy stress tensor, and piecewise constant approximations of the entries of the displacement vector and the rotation tensor [49]. The Picard method is used to linearize the problem and we set a fixed tolerance of 1e-5 on the energy norm of the difference between two consecutive solutions. Unless otherwise specified, all linear solves related to the fixed-point iteration (in both primal and mixed formulations) are carried out with the stabilized bi-conjugated gradient method (BiCGStab) using an incomplete LU decomposition as preconditioner.

Mesh adaptation guided by the *a posteriori* error estimators is carried out by a classical conforming partitioning. No coarsening is applied (mainly due to the capabilities of the current version of the finite element library we use herein [4]). After computing locally the error indicators, we proceed to tag elements for refinement using the Dörfler strategy [44], where we mark sufficiently many elements so that one establishes equi-distribution of the error indicator mass, and then the diameter of each triangle in the new adapted mesh (contained in a generic element K on the initial grid) is set proportional to the diameter of the initial element times the ratio $\bar{\zeta}_h/\zeta_K$, where $\bar{\zeta}_h$ is the mean value of a generic error estimator ζ over the initial mesh (see for instance, [88]). In each of the accuracy tests below, these ratios are multiplied by a constant γ_{ratio} that is arbitrarily chosen so as to generate either a roughly similar number of degrees of freedom, or similar individual error magnitudes than in the case of uniform refinement. The density of the refinement process is tuned at will.

Let us also recall from [13] that the implementation of the fixed-point scheme includes an additional stabilization term associated with dynamic gradient flows, that essentially translates in having a

pseudo-time step in the Euler-Lagrange equations (3.3), that then read: knowing $\boldsymbol{u}^k$, for $k = 1, \ldots$, solve

$$\frac{\boldsymbol{u}^{k+1}}{\delta t} - \mathbf{div}(\mathcal{C}\mathbf{e}(\boldsymbol{u}^{k+1})) = \frac{\boldsymbol{u}^k}{\delta t} - \alpha \boldsymbol{f}_{\boldsymbol{u}^k}.$$

Further details can be found in [13, Appendix C]. Therefore the primal and mixed Galerkin methods, as well as the *a posteriori* error indicators Θ and Ψ are modified accordingly, and only affecting the residual terms associated with the momentum equation. The Picard iterations with pseudo time-stepping are located inside the adaptive refinement loop which consists in solving, estimating, marking and refining.

3.5.2 Example 1: Registration of smooth synthetic images

We assess the accuracy of the primal and mixed DIR methods using a smooth synthetic image under a smooth transformation. To this end, we define the reference image $R : [0,1]^2 \to \mathbb{R}$ by

$$R(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2).$$

We further define a manufactured displacement and the corresponding stress and rotation tensor fields by

$$\boldsymbol{u}(x_1, x_2) = \begin{pmatrix} 0.1\cos(\pi x_1)\sin(\pi x_2) + \frac{x_1^2(1-x_1)^2 x_2^2(1-x_2)^2}{2\lambda} \\ -0.1\sin(\pi x_1)\cos(\pi x_2) + \frac{x_1^3(1-x_1)^3 x_2^3(1-x_2)^3}{2\lambda} \end{pmatrix},$$

$$\boldsymbol{\sigma}(x_1, x_2) = \mathcal{C}\mathbf{e}(\boldsymbol{u}), \qquad \text{and} \qquad \boldsymbol{\rho}(x_1, x_2) = \frac{1}{2}(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^{\mathrm{t}}).$$

Then, we construct a synthetic target image via composition of the reference image and the inverse warping, namely $T = R \circ (\boldsymbol{id} + \boldsymbol{u})^{-1}$. An initial target in the fixed-point scheme is a perturbation of the reference image, that is $T_0(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi[x_2 + 0.01])$. These manufactured solutions satisfy the zero-traction boundary condition, and they are used to construct an additional body load (apart from $\boldsymbol{f}_{\boldsymbol{u}}$) that needs to be incorporated as right-hand side in the discrete problems, as well as in the residual term associated with the momentum conservation equation in the definition of the error indicators. The model parameters employed in this test are Young modulus $E = 1000$, Poisson ratio $\nu = 0.4$ (used to obtain the Lamé constants of the solid, $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$ and $\mu = \frac{E}{2+2\nu}$), a weight constant $\alpha = 100$, and pseudo time-step $\delta t = \alpha^{-2}$.

On sequences of uniformly or adaptive refined meshes, we solve the DIR problem with primal and mixed methods and compute (non-normalized) errors between the approximate and exact solutions in their natural norms, that is, for the primal method $\mathtt{e}_{\boldsymbol{u}} = \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}$; whereas for the mixed method $\mathtt{e}_{\boldsymbol{u}} = \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}$ and $\mathtt{e}_{\boldsymbol{\rho}} = \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\Omega}$, $\mathtt{e}_{\boldsymbol{\sigma}} = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div},\Omega}$. We also point out that in the case of adaptive mesh refinement, the experimental rates of convergence $\widehat{\mathtt{rate}}$ are computed differently than in the uniform case

$$\mathtt{rate} = \log(\mathtt{e}/\widehat{\mathtt{e}})[\log(h/\widehat{h})]^{-1}, \qquad \widehat{\mathtt{rate}} = -2\log(\mathtt{e}/\widehat{\mathtt{e}})[\log(\mathtt{DoF}/\widehat{\mathtt{DoF}})]^{-1},$$

where $\mathtt{e}$ and $\widehat{\mathtt{e}}$ denote errors produced on two consecutive meshes. These grids have respective mesh sizes h and h' (needed to compute the experimental order of convergence $\mathtt{rate}$), or they are associated with $\mathtt{DoF}$ and $\widehat{\mathtt{DoF}}$ degrees of freedom, respectively (in when computing $\widehat{\mathtt{rate}}$). In addition,

the effectivity index associated with the global estimators for the primal and mixed discretizations is computed as

$$\texttt{eff}(\Theta) = \frac{\lambda \mathsf{e}_{\boldsymbol{u}}}{\Theta}, \qquad \texttt{eff}(\Psi) = \frac{\left\{\mathsf{e}_{\boldsymbol{\sigma}}^2 + \mathsf{e}_{\boldsymbol{u}}^2 + \mathsf{e}_{\rho}^2\right\}^{1/2}}{\Psi},$$

where the additional scaling (with the dilation modulus λ) for the indicator Θ is motivated by the fact that the efficiency bound arising from the proof of Lemma 3.6 is proportional to λ due to the definition of the Hooke tensor $\mathcal{C}$. Such an explicit scaling is however not required for the *a posteriori* estimation in the mixed method.

In Figure 3.1(a,b) we show the reference image R_h and the resampled image $T_h = T(\boldsymbol{x}+\boldsymbol{u}_h(\boldsymbol{x}))$, and the panels (c-h) show examples of meshes adaptively refined guided by the estimators. We note that the primal method refines largely around the center of the domain. We also show in panels (i,j,k) the approximate solutions (the Frobenius norm of stress, displacement magnitude, and Frobenius norm of the rotation matrix) generated with the mixed method at the final refinement level.

The numerical convergence of the primal and mixed DIR methods are shown in Figure 3.2(a) and Figure 3.2(b), respectively. We observe that both methods do exhibit monotonic convergence. For this particular example, no major differences arise between the uniform and adaptive refinement schemes. Convergence rates for both methods are reported in Table Table 3.1, where we verify that optimal convergence are achieved with $O(h^k)$. No differences were observed in the number of Picard iterations required by the uniform and adaptive refinement strategies. The mixed DIR method also displays optimal convergence rates, see Table 3.2. We further note that the effectivity index values for the mixed scheme are roughly constant and close to 0.43, and that the convergence rate is not substantially improved by the adaptivity in this example.

(a) Primal method, uniform refinement					(b) Primal method, adaptive refinement					
k	DoF	h	rate	iter	k	DoF	$h_{\min}$	$\widehat{\texttt{rate}}$	$\texttt{eff}(\Theta)$	iter
1	53	0.3536	0.561	4	1	53	0.3536	1.123	0.6984	4
	165	0.1768	0.931	4		165	0.1768	1.123	0.7083	4
	581	0.0884	1.116	4		557	0.0884	1.123	0.6964	4
	2181	0.0442	1.082	4		2101	0.0442	1.041	0.6962	4
	8453	0.0221	1.030	4		8149	0.0221	1.022	0.6950	4
2	165	0.3536	1.649	4	2	165	0.3536	2.177	0.3599	4
	581	0.1768	1.945	4		581	0.1768	2.026	0.3675	4
	2181	0.0884	2.025	4		2181	0.0884	2.088	0.3602	4
	8453	0.0442	2.031	4		8453	0.0442	2.045	0.3546	4
	33285	0.0221	2.041	4		32933	0.0221	2.061	0.3457	4

Table 3.1: Example 1: Smooth synthetic image registration example. Error measures, convergence rates, and Picard iteration count for the approximate displacements $\boldsymbol{u}_h$ produced with the primal method (of polynomial degrees $k = 1$ and $k = 2$); and tabulated according to the resolution level. (a) Uniform mesh refinement, (b) adaptive mesh refinement based on error estimator Θ, with $\gamma_{\text{ratio}} = 0.1$, also displaying the rescaled effectivity index.

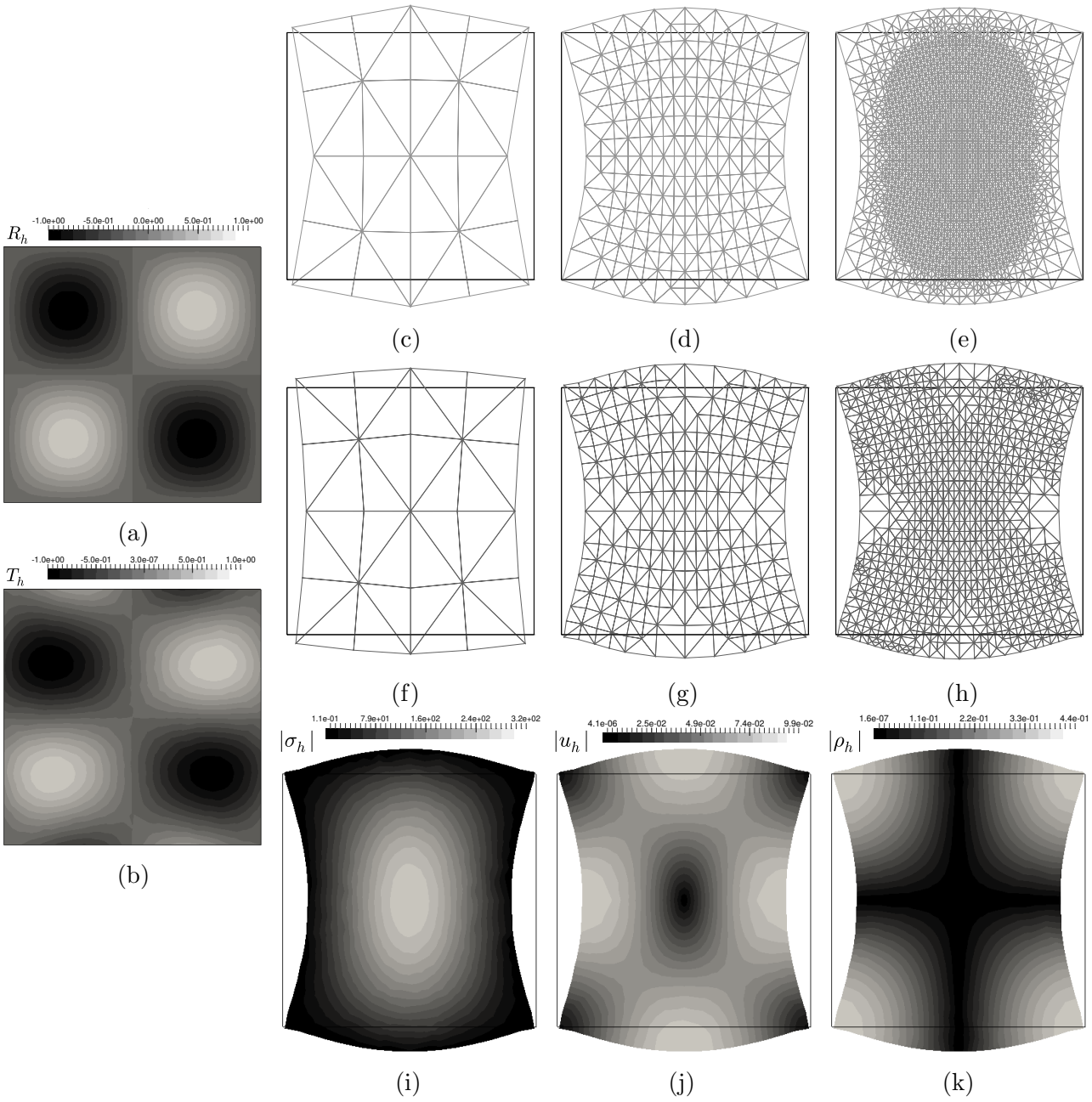


Figure 3.1: Example 1: Adaptive mesh refinement in the registration of a smooth synthetic images. (a,b) Projected fields of the reference R and composed $T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x}))$ images; (c,d,e) evolution of the mesh adaption for the primal scheme using the error indicator Θ; (f,g,h) evolution of the mesh adaption for the mixed scheme using the error indicator Ψ; (i,j,k) Stress, displacement and rotation norm fields predicted by the mixed scheme using mesh adaptivity.

3.5.3 Example 2: Registration of smooth synthetic images with high gradients

Next, we modify the closed-form displacement of Example 1 to produce higher gradients in the reference image and initial target image. To this end, we consider the following image and displacement

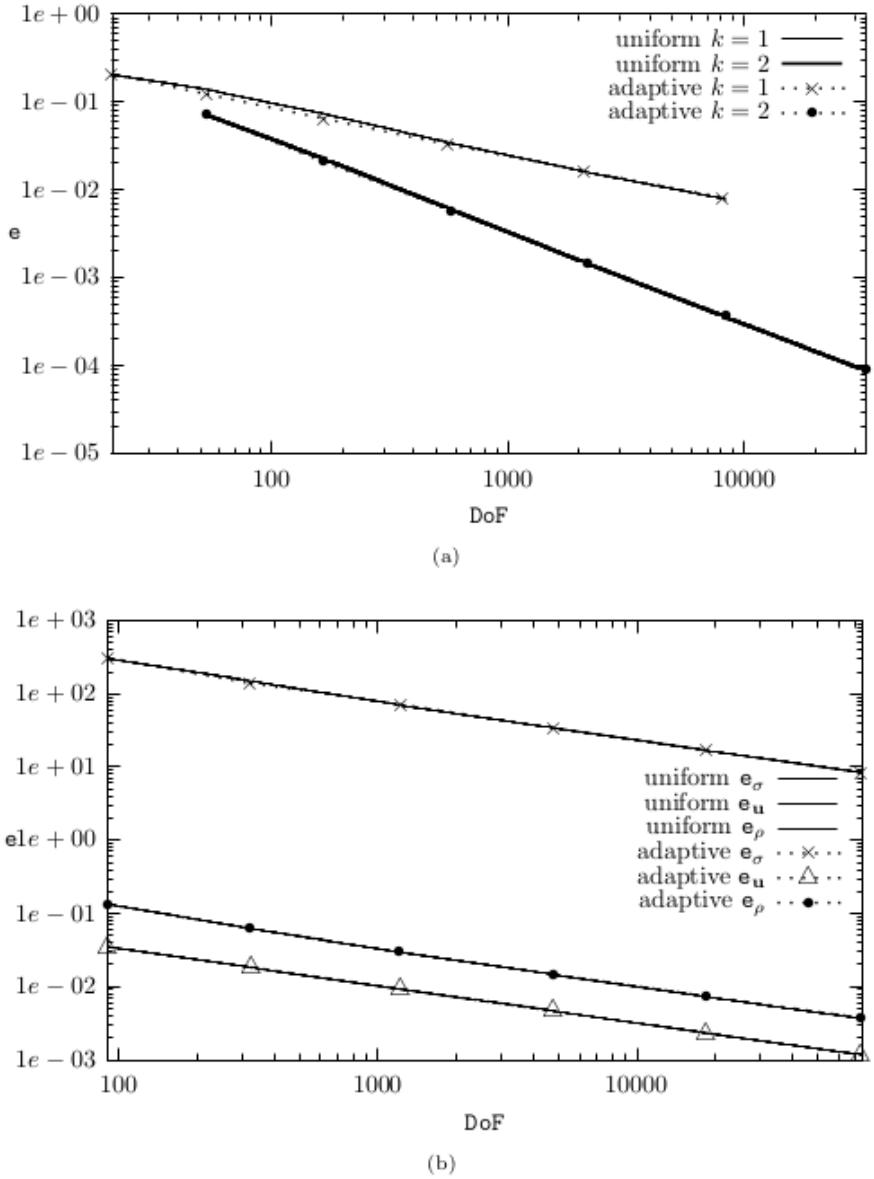


Figure 3.2: Example 1: Smooth synthetic image registration example. Error convergence with respect to the number of degrees of freedom for both (a) primal, and (b) mixed DIR formulations. Uniform refinement is shown in solid lines, while the adaptive refinement is shown in dotted lines.

field expressions:

$$\boldsymbol{u}(x_1,x_2)=\begin{pmatrix}0.1\cos(\pi x_1)\sin(\pi x_2)+\frac{x_1^2(1-x_1)^2x_2^2(1-x_2)^2}{2}\\ -0.1\sin(\pi x_1)\cos(\pi x_2)+\frac{x_1^3(1-x_1)^3x_2^3(1-x_2)^3}{2}\end{pmatrix},\quad R(x_1,x_2)=\frac{x_1x_2(x_1-1)(x_2-1)}{(x_1+0.01)^4+(x_2+0.01)^4},$$
$$T_0(x_1,x_2)=e^{-50[(x_1-0.2)^2+(x_2-0.2)^2]}.$$

All other remaining model parameters are kept the same as in Example 1.

(a) Mixed method, uniform refinement

DoF	h	$\texttt{rate}_{\boldsymbol{\sigma}}$	$\texttt{rate}_{\boldsymbol{u}}$	$\texttt{rate}_{\boldsymbol{\rho}}$	iter
323	0.3536	1.016	0.932	1.125	7
1219	0.1768	1.101	0.994	1.049	8
4739	0.0884	1.051	1.000	1.017	8
18691	0.0442	1.015	1.000	1.006	8
74243	0.0221	1.004	1.000	1.002	10

(b) Mixed method, adaptive refinement

DoF	$h_{\min}$	$\widehat{\texttt{rate}}_{\boldsymbol{\sigma}}$	$\widehat{\texttt{rate}}_{\boldsymbol{u}}$	$\widehat{\texttt{rate}}_{\boldsymbol{\rho}}$	$\texttt{eff}(\Psi)$	iter
323	0.3536	1.251	1.015	1.195	0.4281	8
1219	0.1768	1.005	1.037	1.114	0.4257	8
4739	0.08839	1.085	1.020	1.038	0.4219	8
18603	0.04419	1.030	1.004	0.998	0.4222	9
73635	0.0221	1.001	1.001	1.000	0.4217	9

Table 3.2: Example 1: Smooth synthetic image registration example. Convergence rates, and Picard iteration count for the approximate Cauchy stress, displacements, and rotation $\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h$ for the mixed formulations. (a) Uniform mesh refinement, (b) adaptive mesh refinement guided by Ψ, with $\gamma_{\text{ratio}} = 0.05$.

We show in Figure 3.3(a,b) synthetic images projected onto the space of piecewise linear and continuous functions, as well as a few adapted meshes produced with the indicators (c-h), where one sees that the agglomeration of vertices occurs not so much due to the high gradients of the synthetic images, but mainly because of the features in the solutions to the elasticity problem. Panels (i,j,k) have snapshots of approximate solutions generated with the mixed method after five steps of adaptive refinement, and plotted on the deformed domain. We note that, for the adaptive algorithm with $\gamma_{\text{ratio}} = 0.01$, the error indicator makes the refinement to be applied uniformly for the first three iterations, after which localized meshing takes place in certain regions of the domain.

Figure 3.4(a) shows the numerical convergence of the primal DIR method under uniform and adaptive refinement. We observe monotonic convergence for all displacement field error as the number of DoFs increases. A notable improvement in convergence is observed for the particular case of the adaptive refinement scheme using second-order element interpolations. Convergence rates for the primal DIR method using uniform and adaptive refinement are reported in Table 3.3, where we observe that the case of adaptive refinement using second-order elements results in convergence rates that reach $k = 2$, which is notoriously higher than the convergence rate of $k = 1.5$ reached by the primal method under uniform refinement. For the case of the mixed method, adaptive refinement always result in better convergence than uniform refinement for the displacement, stress and rotation fields, see Figure 3.4(b). Table 3.4 reports the convergence rates of the mixed method, where we note that the adaptive refinement always results in rates that are greater than those obtained under uniform refinement. Further, we observe that in systems with roughly similar number of DoFs, the number of Picard iterations needed to reach the tolerance are smaller in the case of adaptive refinement.

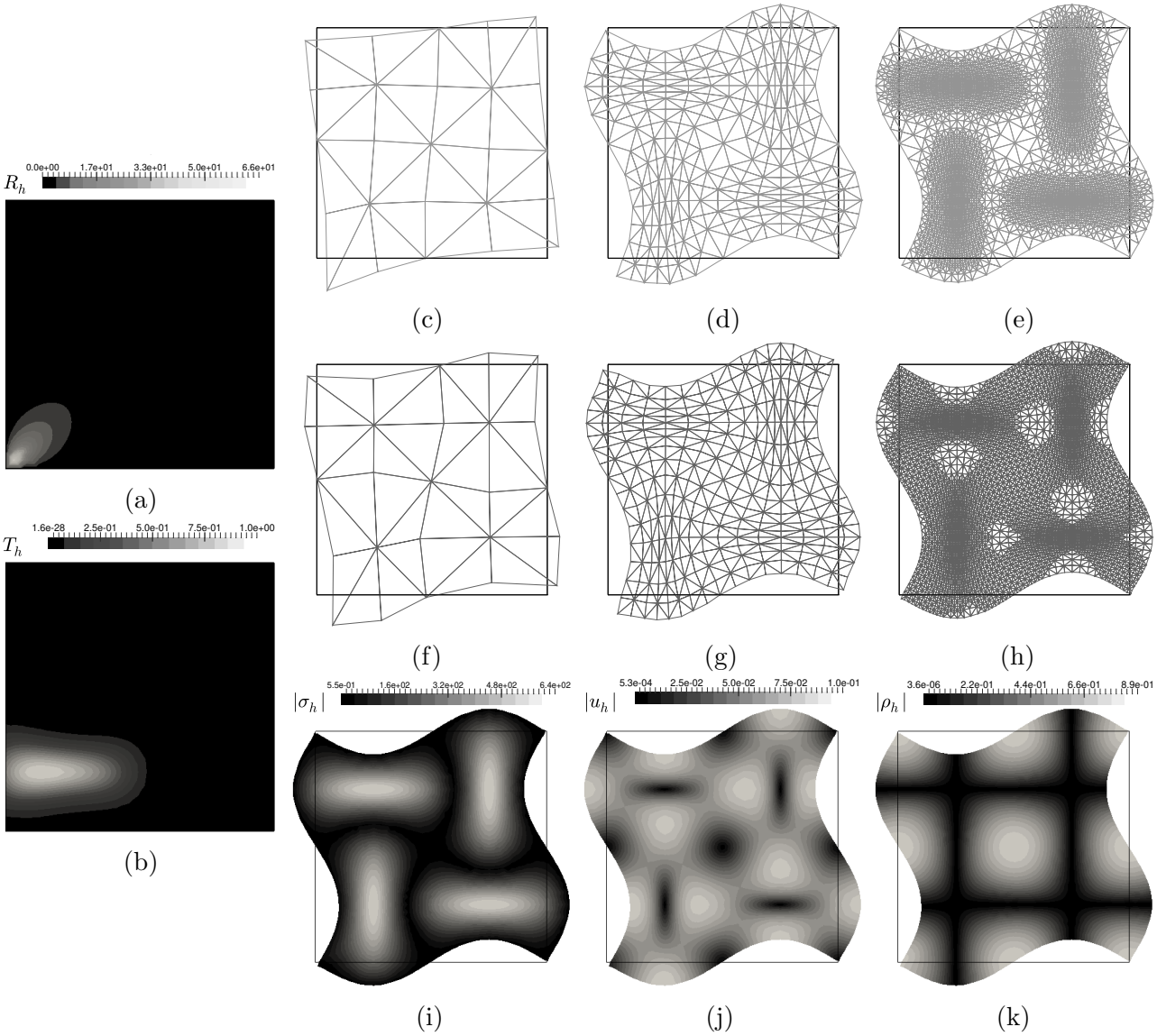


Figure 3.3: Example 2: Adaptive mesh refinement in the registration of smooth synthetic images with high gradients. (a,b) Reference image R and composed Target image $T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x}))$; (c,d,e) evolution of the mesh adaption for the primal DIR method using the error indicator Θ; (f,g,h) evolution of the mesh adaption for the mixed DIR method using the error indicator Ψ; (i,j,k) Stress, displacement and rotation norm fields predicted by the mixed scheme using mesh adaptivity.

3.5.4 Example 3: Registration of brain medical images

We now turn to the application of the adaptive primal and mixed DIR methods in the registration of medical images of human brains [33]. The reference and target images for the brain have dimensions 258×258 and the voxel resolution corresponds to 1 mm, see top panels in Figure 3.5. We proceed to solve the DIR problem using both primal and mixed adaptive schemes, starting from structured meshes with 32768 triangular elements. The elasticity parameters are set to $E = 15$, $\nu = 0.3$, the

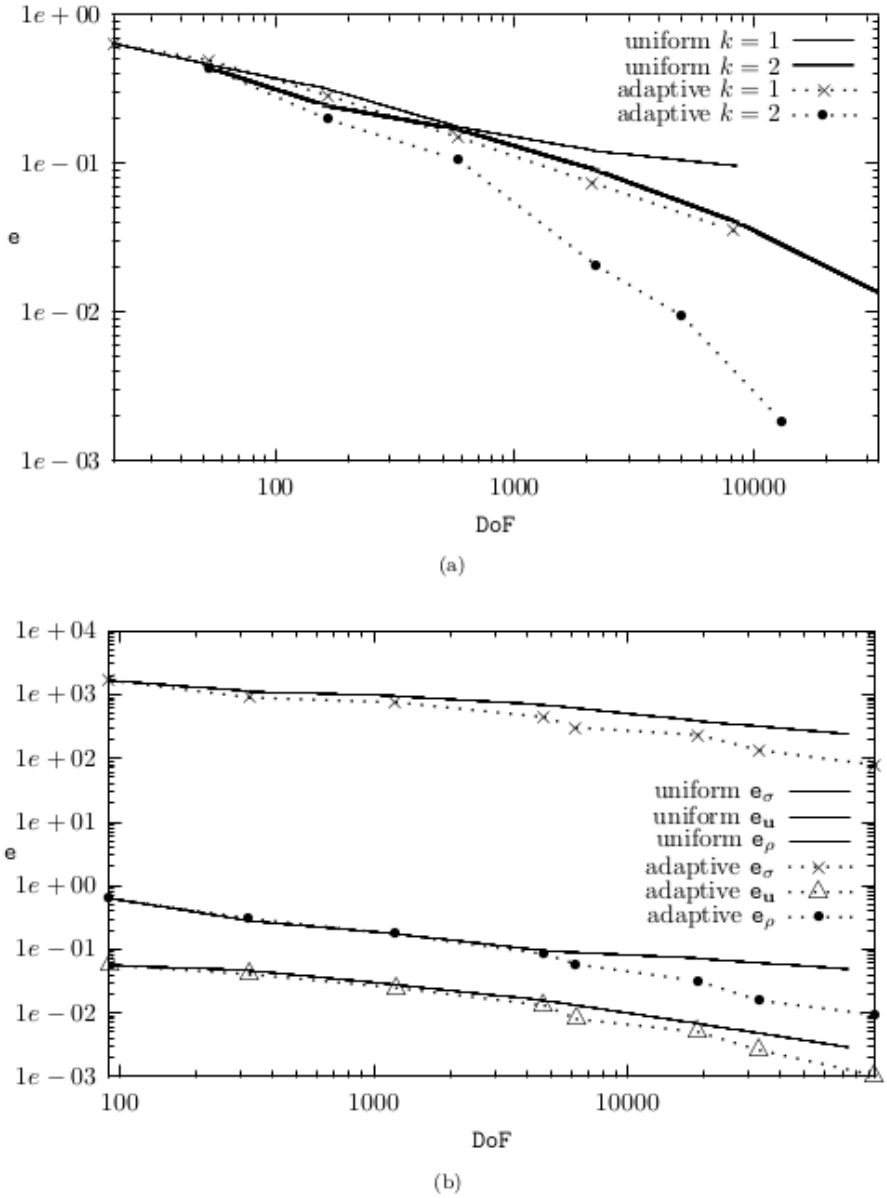


Figure 3.4: Example 2: Error convergence for (a) primal DIR method and (b) mixed DIR method under uniform and adaptive mesh refinement.

weight constant is $\alpha = 50$, and the pseudo timestep is $\delta t = 0.01/\alpha$. The tolerance for the Picard scheme is increased to 1e-04, and for the mixed method the refinement density proportion is ruled by the constant $\gamma_{\text{ratio}} = 0.1$. The primal method requires an average (over the number of mesh refinement steps, here assigned to 4) of 19 Picard steps to reach convergence, which is slightly larger for the mixed method (22 iterations). The first two plots on the middle row of Figure 3.5 depict the composed images $T \circ (\boldsymbol{id} + \boldsymbol{u}_h)$ generated with the primal and mixed methods, where we can notice very similar patterns in both cases. The two other figures on the right show the similarity between reference and warped images, $|R(\boldsymbol{x}) - T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x}))|$ resulting from both methods.

(a) Primal method, uniform refinement					(b) Primal method, adaptive refinement					
k	DoF	h	rate	iter	k	DoF	$h_{\min}$	$\widehat{\texttt{rate}}$	eff(Θ)	iter
1	53	0.3536	0.481	4	1	53	0.3536	0.617	0.8314	4
	165	0.1768	0.526	6		165	0.1768	1.277	0.8282	6
	581	0.0884	0.859	19		581	0.0884	1.037	0.8205	11
	2181	0.0442	0.793	24		2105	0.0442	1.084	0.8187	15
	8453	0.0221	0.620	28		8177	0.0221	1.099	0.8219	18
2	165	0.3536	0.844	4	2	165	0.3536	1.398	1.6422	4
	581	0.1768	0.529	6		581	0.1768	1.489	1.6926	6
	2181	0.0884	0.900	20		2181	0.0884	1.891	1.6360	12
	8453	0.0442	1.169	25		4959	0.0442	1.786	1.6799	15
	33285	0.0221	1.564	29		13129	0.0221	2.097	1.6492	18

Table 3.3: Example 2. Convergence rates, and Picard iteration count for the approximate displacements $\boldsymbol{u}_h$ produced with the first and second-order primal method; and tabulated according to the resolution level, under uniform (a) and adaptive mesh refinement guided by Θ, with $\gamma_{\text{ratio}} = 0.01$ ((b) also displaying the rescaled effectivity index).

(a) Mixed method, uniform refinement					
DoF	h	$\texttt{rate}_{\boldsymbol{\sigma}}$	$\texttt{rate}_{\boldsymbol{u}}$	$\texttt{rate}_{\boldsymbol{\rho}}$	iter
323	0.3536	0.552	0.278	1.174	7
1219	0.1768	0.278	0.728	0.674	13
4739	0.0884	0.443	0.846	0.882	28
18691	0.0442	0.741	1.183	0.658	45
74243	0.0221	0.598	1.235	0.606	50

(b) Mixed method, adaptive refinement							
DoF	$h_{\min}$	$h_{\max}$	$\widehat{\texttt{rate}}_{\boldsymbol{\sigma}}$	$\widehat{\texttt{rate}}_{\boldsymbol{u}}$	$\widehat{\texttt{rate}}_{\boldsymbol{\rho}}$	eff(Ψ)	iter
323	0.3536	0.3536	0.965	0.518	1.169	0.5333	5
1219	0.1768	0.1768	0.955	0.725	0.869	0.5272	8
4692	0.0742	0.1250	0.952	0.946	1.002	0.5188	19
6277	0.0264	0.1250	1.066	1.114	1.106	0.5139	21
18884	0.0107	0.0817	1.052	1.039	1.067	0.5205	24
32998	0.0051	0.0730	0.986	0.958	0.975	0.5216	30
94153	0.0020	0.0601	0.961	0.963	0.959	0.5210	30

Table 3.4: Example 2: Convergence rates, and Picard iteration count for the approximate Cauchy stress, displacements, and rotation $\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h$ produced with the lowest-order mixed method; and tabulated according to the resolution level, under uniform (a) and adaptive mesh refinement guided by Ψ, with $\gamma_{\text{ratio}} = 0.009$ ((b) also displaying the effectivity index).

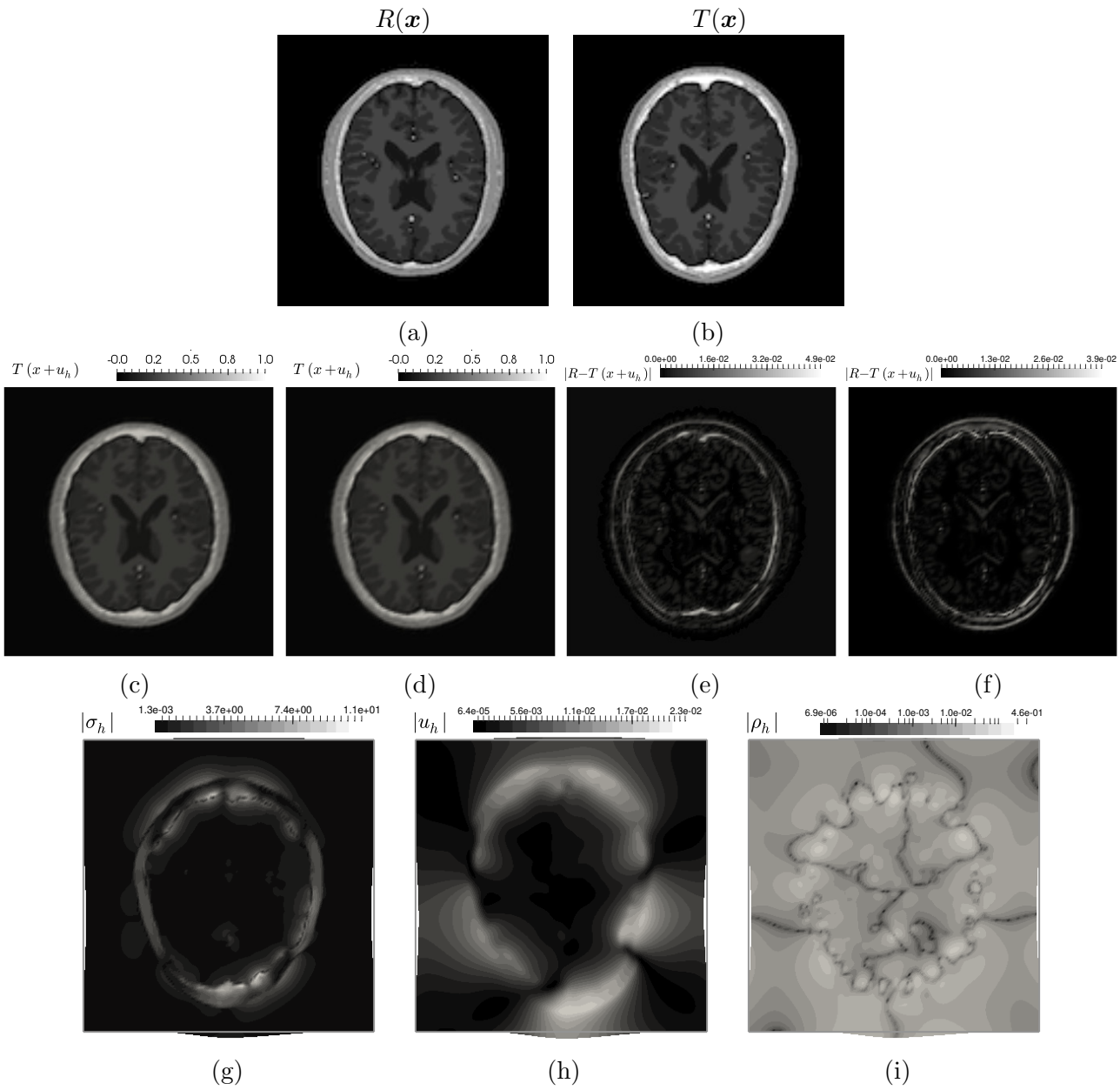


Figure 3.5: Example 3. Registration of brain medical images. (a) Reference image, (b) target image; (c,d) resampled (composed) (c,d) images from solutions using primal and mixed schemes, respectively; (e,f) similarity plots resulting from primal and mixed schemes, respectively; (g,h,i) stress, displacement and rotation norm fields resulting from the mixed DIR scheme using adaptive mesh refinement.

We also plot an example of a mesh obtained after four steps of adaptive refinement with the primal and mixed methods (see Figure 3.6). For illustration purposes we initiate the process from a coarse mesh of 8196 triangles (corresponding to a low resolution image of 64×64 pixels. Starting with images of higher resolution imply that the meshes obtained after adaptive refinement are too dense to be easily visualized). The figures exemplify the concentration of refinement near the skull, which is consistently

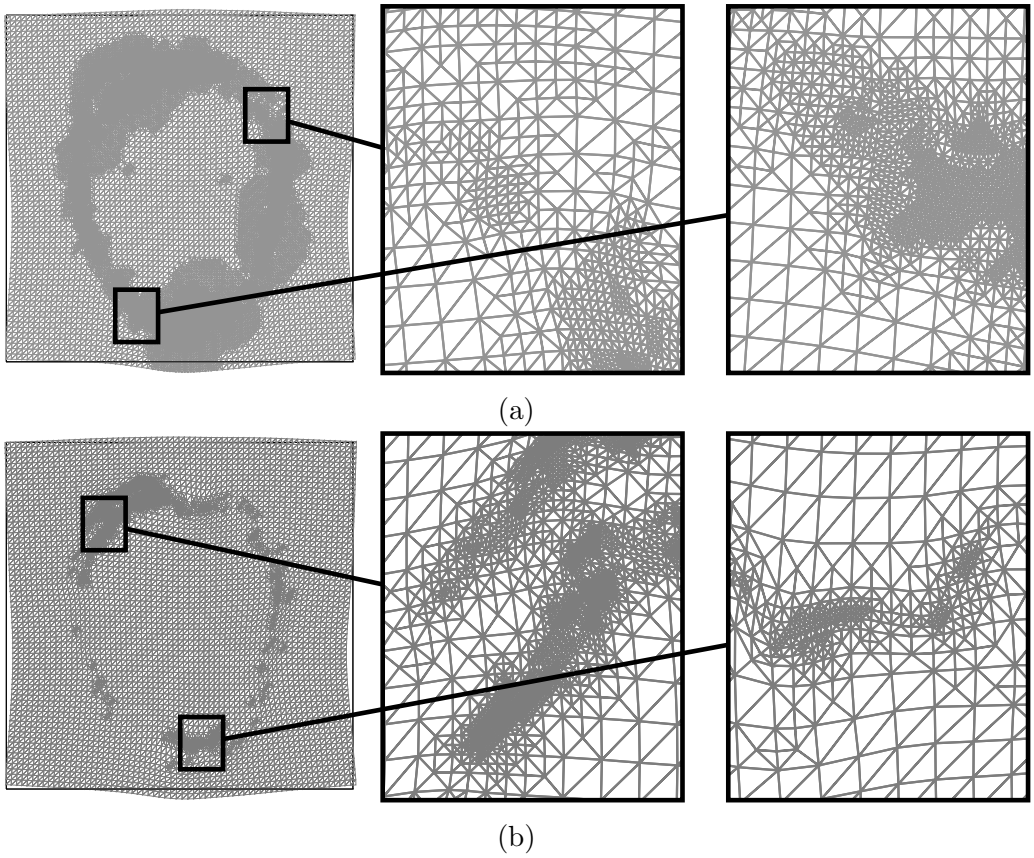


Figure 3.6: Example 3. Adaptive mesh refinement in the registration of brain medical images. (a) Mesh after four steps of adaptive refinement using the error indicator Θ for the primal DIR method; (b) Mesh after four steps of adaptive refinement using the error indicator Ψ for the mixed DIR method.

the zone with highest gradients in the reference and target images, as well as in stress and rotations (as inferred from panels (g,h,i) in Figure 3.5, where the Frobenius norm of the rotation tensor is plotted in log-scale for clarity). On the other hand, the displacements are, in comparison, rather smooth and they seem not to contribute substantially to the local error indicators.

In Table 3.5 we report information about the CPU time required in each step of the overall solution algorithm. We record the wall-time during the execution of the mixed and primal DIR methods, when starting from a coarse grid (representing 8715 DoFs for the primal method and 76573 DoFs for the mixed scheme) and in both cases applying five iterations of adaptive mesh refinement. An average of 17 fixed-point iterations are needed for the primal approximations and 25 for the mixed scheme.

3.5.5 Example 4: Registration of binary images under large deformation

The last example of application adressed in this study consists in a classic benchmark in DIR which introduces two important challenges. First, reference and target images are binary-composed, i.e. they have intensity values of either 0 or 1, which creates steep numerical gradients at the binary interface of order $1/h$. Thus, the images do not satisfy condition (3.5). Second, the deformation required for a

	refin. level	matrix assembly	solution computation	IO and residual	evaluation of estimator	marking and refinement
Primal method	1	0.101	0.075 (avg)	0.102	0.096	0.544
(total CPU time: 73.16)	2	0.099	0.163 (avg)	0.110	0.130	0.757
	3	0.162	0.312 (avg)	0.192	0.235	1.284
	4	0.489	1.127 (avg)	0.481	0.704	3.351
	5	0.853	2.093 (avg)	0.758	0.812	5.246
Mixed method	1	0.418	1.445 (avg)	0.101	0.099	0.530
(total CPU time: 997.83)	2	0.443	2.373 (avg)	0.109	0.141	0.668
	3	0.578	4.746 (avg)	0.135	0.154	0.719
	4	0.704	8.390 (avg)	0.204	0.237	1.298
	5	0.921	22.45 (avg)	0.439	0.304	2.616

Table 3.5: Example 3. CPU time (in [s]) of each step of the adaptive finite element method for the DIR problem, measured for the primal and mixed methods, starting from coarse meshes. The time associated with the solution of the linear systems is averaged over the number of inner Picard iterations.

satisfactory registration is large, so that the validity of the elastic potential is not clear from a physical viewpoint. We define the ball $B(\boldsymbol{x}, r) = \{\boldsymbol{x} \in \mathbb{R}^2 : |\boldsymbol{x}| \leq r\}$ to set the images as

$$R(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in B(0.5, 0.32) \cap [B(0.5, 0.16)]^c \cap [\{x_1 > 0.5\} \cap \{0.4 < x_2 < 0.6\}]^c, \\ 0 & \text{otherwise}, \end{cases}$$

$$T(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in B(0.5, 0.25), \\ 0 & \text{otherwise}. \end{cases}$$

Both methods consider quadrature rules of sixth order, with an initial mesh given by a unit square with 20 elements per side, which yields a total of 800 triangular elements. We consider the parameters $\alpha = 1000, E = 15, \nu = 0.3$ and set the pseudo timestep to $\delta t = h_{\min}^2/\alpha$ for the primal case and $\delta t = 0.01\, h_{\min}^2/\alpha$, where $h_{\min}$ is the minimum characteristic length of the mesh. This was motivated by a possible CFL condition on the timestep arising from the explicit treatment of the nonlinearity and proved effective during numerical tests. The convergence was set through the ℓ^∞ norm of the increment $|\boldsymbol{u}^k - \boldsymbol{u}^{k-1}|_{\ell^\infty}$ with a tolerance of $h_{\min}$, so that iterations stop when the displacement changes by less than the smallest element. A maximum number of 100 iterations was always achieved, following previous works adressing this problem [74]. Both the primal and mixed DIR problems for this example were solved in serial with the iterative scheme BiCGStab preconditioned with an incomplete LU factorization, using the default parameters available in FEniCS. The solution of the mixed DIR problem required a considerable numerical effort to converge to a solution that met the error criterion. To overcome this difficulty, we used at each refinement level the solution of the primal formulation as an initial solution for the mixed case, and then employed 5 iterations of the mixed formulation only. This was already implemented in [13] to substantially improve the registration of lung images in a mixed formulation.

	refin. level	matrix assembly	solution computation	IO and residual	evaluation of estimator	marking and refinement
Primal method	1	0.018	0.036 (avg)	0.024	0.017	0.010
(total CPU time: 261.83)	2	0.031	0.076 (avg)	0.029	0.023	0.023
	3	0.293	0.181 (avg)	0.049	0.054	0.081
	4	0.293	0.627 (avg)	0.130	0.140	0.202
	5	0.751	2.381 (avg)	0.338	0.466	0.799
Mixed method	1	0.04	0.051 (avg)	0.058	0.016	0.006
(total CPU time: 326.61)	2	0.248	0.112 (avg)	0.037	0.047	0.064
	3	0.777	0.361 (avg)	0.037	0.047	0.064
	4	2.781	1.256 (avg)	0.142	0.147	0.243
	5	10.725	5.104 (avg)	0.464	0.545	1.031

Table 3.6: Example 4. CPU time (in [s]) of each step of the adaptive finite element method for the DIR problem, measured for the primal and mixed methods, starting from coarse meshes. The time associated with the solution of the linear systems is averaged over the number of inner Picard iterations.

We report the solution with its components in Figure 3.7. In the first row we show the reference (a) and target (b) images, constructed as in [74], with the solution reported in the second row together with its absolute error $|R(\boldsymbol{x}) - T(\boldsymbol{x} + \boldsymbol{u}_h(\boldsymbol{x}))|$ in primal (c, e) and mixed (d, f) form. We note that the mixed DIR performs slightly worse than the primal DIR method, which is to be expected due to the lower order of approximation used. The last row shows the magnitude of all components of the solution, in both primal (j) and mixed (g,h,i) formulations. Also, in Figure 3.8 we present the refined mesh after three steps, where it can be observed how the mixed scheme yields a more localized refinement even though the amount of refined elements is the same in both schemes. Finally, we provide information on the CPU time required in each step of the overall solution algorithm in Table 3.6.

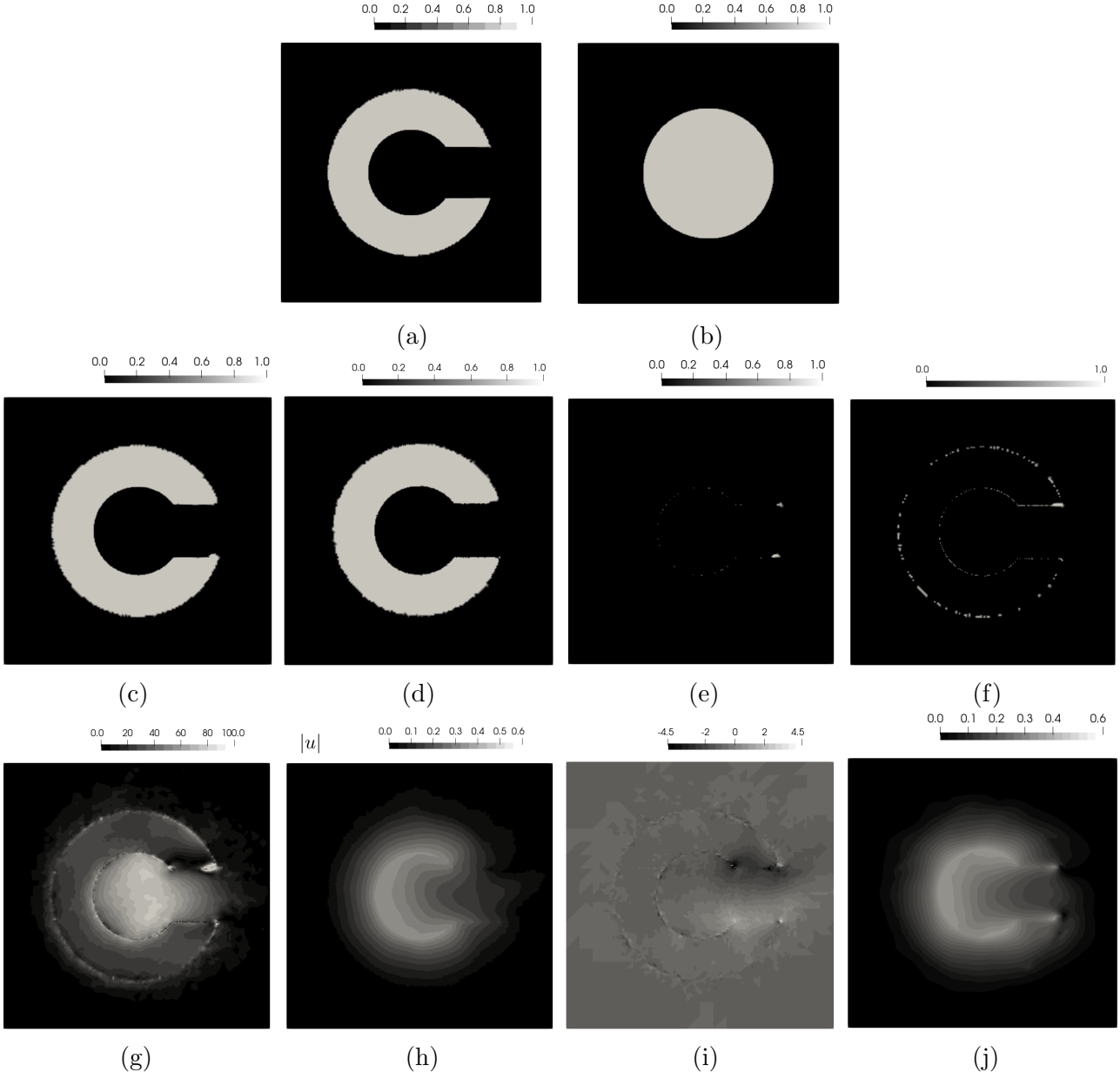
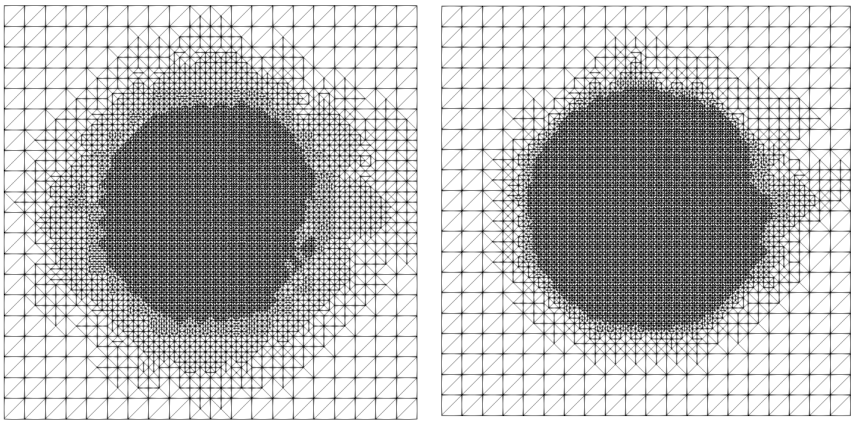


Figure 3.7: Example 4. Registration of binary images (O-C). (a) Reference image, (b) target image; (c,d) resampled (composed) images from solutions using primal and mixed schemes, respectively; (e,f) similarity images resulting from the primal and mixed methods, respectively; (g,h,i) stress, displacement and rotation norm fields using the adaptive mixed DIR method; (j) displacement norm field using the adaptive primal DIR method.



(a) (b)

Figure 3.8: Example 4. Registration of binary images. Mesh after three steps of adaptive refinement for (a) primal DIR problem, and (b) mixed DIR problem.

Conclusions

In this thesis we have developed primal and mixed finite element methods for a set of partial differential equations of physical interest in Biology and Biomedicine, more precisely, the bioconvective flows problem and deformable image registration problem. We have proved the solvability of the continuous and discrete problems as well as their convergence results, and we have also provided corresponding numerical examples and simulations. The main conclusions for each one of the models are:

1. We introduced a fully-mixed finite element method for the bioconvective flows problem. For convenience of the analysis, we introduced the strain tensor, vorticity, and pseudo-stress as additional unknowns (besides the pseudo-concentration gradient, the velocity, the pressure, and the concentration). This allows us to eliminate the pressure from the system, which is then recovered using an appropriate postprocessing formula, together with the concentration gradient. The original problem was reformulated as an augmented variational approach in the incompressible viscous fluid modelled by a Navier-Stokes type-system (with non-linear viscosity) coupled with an advection-diffusion equation. Then, through a fixed-point strategy together with sufficiently small data assumptions, the solvability analysis of both the continuous and discrete problems as well as its corresponding a priori estimate were developed. Finally, several numerical experiments were reported in order to validate the good performance of the method and confirm the corresponding order of convergence.

2. We presented a way to formulate deformable image registration problems with Neumann boundary conditions in a mathematically consistent way so as not to lose information from the images but still keeping all the degrees of freedom from the original problem in both primal and mixed formulations, the latter being particularly important in the quasi-incompressible case. This method presents clear advantages for capturing rigid motions, i.e translations and rotations. The results of well-posedness of the continuous and discrete formulations, a priori error estimates, and the respective rates of convergence, were obtained by using the Babuška-Brezzi theory and duality arguments.

3. We established an adaptive mesh-refinement scheme for the numerical solution of primal and mixed DIR problems. Our method hinges upon the development of *a posteriori* error estimators for both the primal and mixed finite-element formulations that are reliable and efficient, and at the same time, they are easily computed. These estimators allow for an optimal refinement of

the mesh in zones where the accuracy of the numerical approximation does not perform well. Thus, one distinctive feature of our work is the effectiveness of the mesh-adaption strategy, as they are justified on selectively reducing the local approximation error made by the finite-element schemes employed. This contrasts with current methods of mesh adaption employed in DIR problems, which either refine the discretization uniformly or rely on heuristic grounds to select regions that are refined. To assess the numerical performance of the proposed method, we employ uniform and adaptive mesh refinement to solve a DIR problem based on smooth synthetic images where the displacement solution is known in advance and to demonstrate the applicability of the method in medical images, we perform DIR on human brain images.

Future work

The methods developed and the results obtained in this thesis have motivated several ongoing and future projects. Some of them are described below:

1. A posteriori error analysis of the augmented fully-mixed formulation for bioconvective flows problem: We are interested in developing a posteriori error analysis for the method studied in Chapter 1 in order to improve its robustness in the context of problems involving complex geometries or solutions with high gradients.

2. Analysis of an augmented mixed-primal formulation for the bioconvective flow model: As an alternative to our fully-mixed method presented in Chapter 1, we are interested in studying a mixed formulation for the fluid (without considering the vorticity as an unknown of the system) and a primal formulation for the concentration equation.

3. Finite element methods for inelastic deformable image registration problem: We are interested in extending the results and techniques of Chapters 2 and 3 to the inelastic case. For this model, we consider the problem of elastoplasticity with internal hardening $\boldsymbol{\xi}$, in which we assume that the strain tensor $\mathbf{e}$ can be decomposed as $\mathbf{e}(\boldsymbol{u}) = \mathbf{e}^e(\boldsymbol{u}) + \mathbf{e}^p(\boldsymbol{u})$, where $\mathbf{e}^e$ and $\mathbf{e}^p$ are the elastic and plastic part of the strain tensor, respectively. The main unknowns of the model are the displacement $\boldsymbol{u}$, the plastic strain $\mathbf{e}^p$ and the internal hardening $\boldsymbol{\xi}$, whereas the equations reduce to

$$\begin{gathered} \operatorname{div} \boldsymbol{\sigma} + \alpha \boldsymbol{f}_{\boldsymbol{u}} = \mathbf{0}, \qquad \boldsymbol{\sigma} = \boldsymbol{C}(\mathbf{e}(\boldsymbol{u}) - \mathbf{e}^p), \qquad \boldsymbol{\chi} = -\boldsymbol{H}\boldsymbol{\xi} \quad \text{in} \quad \Omega, \\ \boldsymbol{u} = \mathbf{0}, \qquad \boldsymbol{\sigma}\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \partial\Omega, \end{gathered} \tag{3.40}$$

where $\boldsymbol{\chi}$ is the force conjugate to $\boldsymbol{\xi}$ and $\boldsymbol{H}$ represents a hardening modulus. Further, this model considers a flow law which governs the evolution of the plastic strain and internal hardening. The relevance of this kind of model is due to that certain human tissues, such as lung tissue, can exhibit plasticity behavior when they are subjected to stresses above their elastic range.

Conclusiones y trabajo futuro

Conclusiones

En esta tesis hemos desarrollado métodos de elementos finitos primales y mixtos para sistemas de ecuaciones diferenciales parciales de interés físico en Biología y Biomedicina, más precisamente, el problema de fluidos bioconvectivos y problema de registro deformable de imágenes. Hemos demostrado solubilidad de los problemas continuo y discreto, así como sus resultados de convergencia, para luego proporcionar ejemplos numéricos y simulaciones correspondientes. Las principales conclusiones para cada uno de los modelos son:

1. Introdujimos un método de elementos finitos completamente mixto para el problema de fluidos bioconvectivos. Por conveniencia del análisis, introdujimos el tensor de esfuerzo, la vorticidad, y el pseudo-estrés como incógnitas adicionales (además de la pseudo-concentración, la velocidad, la presión y la concentración). Esto nos permite eliminar la presión en el sistema, la cual es recuperada a través de una fórmula de postproceso adecuada, junto con el gradiente de la concentración. El problema original fue reformulado mediante un enfoque variacional aumentado para el fluido viscoso incompresible modelado por un sistema de ecuaciones de tipo Navier-Stokes (con viscosidad no lineal) acoplado con una ecuación de advección-difusión. Seguidamente, a través de una estrategia de punto fijo junto con supuestos de datos suficientemente pequeños, se desarrolló el análisis de solubilidad de los problemas continuo y discreto, con su estimación a priori correspondiente. Finalmente, se reportaron varios experimentos numéricos que validaron el buen desempeño del método y que confirmaron los órdenes de convergencia correspondientes.

2. Presentamos una manera de formular problemas de registro deformable de imágenes con condiciones de frontera Neumann de una manera matemáticamente consistente para para no perder información de las imágenes pero manteniendo todos los grados de libertad del problema original tanto en la formulación primal como mixta, siendo este último particularmente importante en el caso cuasi-incompresible. Este método presenta claras ventajas para capturar movimientos rígidos, es decir, traslaciones y rotaciones. Los resultados de solubilidad de las formulaciones continua y discreta, estimaciones de error a priori y la respectiva tasa de convergencia, fueron obtenidos usando la teoría de Babuška-Brezzi y argumentos de dualidad.

3. Establecimos un esquema de refinamiento adaptativo de malla para la solución numérica para los problemas de registro deformable de imágenes primal y mixto. Nuestro método depende del desarrollo de estimadores de error a posteriori, para las formulaciones de elementos finitos primal y mixta, los cuales son confiable y eficiente, y al mismo tiempo, se calculan fácilmente.

Estos estimadores permiten un refinamiento óptimo de la malla en zonas donde la precisión de la aproximación numérica no funciona bien. Así, una característica distintiva de nuestro trabajo es la efectividad de la estrategia de adaptación de la malla, ya que se justifica en la reducción selectiva del error de aproximación local de los esquemas de elementos finitos empleados. Esto contrasta con los métodos actuales de adaptación de malla empleados en los problemas de DIR, que refinan la discretización de manera uniforme o dependen de bases heurísticas para seleccionar regiones que se refinan. Para evaluar el desempeño numérico del método propuesto, empleamos un refinamiento de malla uniforme y adaptativo para resolver un problema de DIR basado en imágenes sintéticas suaves donde el desplazamiento se conoce de antemano y para demostrar la aplicabilidad del método en imágenes médicas, realizamos DIR en imágenes del cerebro humano.

Trabajo futuro

Los métodos desarrollados y los resultados obtenidos en esta tesis han motivado varios proyectos en proceso y a futuro. Algunos de ellos son descritos a continuación:

1. Análisis de error a posteriori para la formulación completamente mixta del problema de fluidos bioconvectivos: Estamos interesados en desarrollar el análisis de error a posteriori para el método estudiado en el Capítulo 1, y de esta forma, mejorar su solidez en el contexto de problemas que involucran geometrías complejas o soluciones con altos gradientes.

2. Análisis de una formulación mixta-primal aumentada para el modelo bioconvectivo: Como alternativa a nuestro método completamente mixto presentado en el Capítulo 1, nos interesa estudiar una formulación mixta para el fluido (sin considerar la vorticidad como incógnita del sistema) y una formulación primal para la ecuación de concentración.

3. Métodos de elementos finitos para el problema de registro deformable de imágenes inelástico: Estamos interesados en extender los contenidos de los Capítulos 2 y 3, al caso inelástico. Para este caso se estudia el problema de elastoplasticidad con endurecimiento interno $\boldsymbol{\xi}$, en el cual se asume que el tensor de esfuerzo $\mathbf{e}$ puede ser descompuesto como $\mathbf{e}(\boldsymbol{u}) = \mathbf{e}^e(\boldsymbol{u}) + \mathbf{e}^p(\boldsymbol{u})$, donde $\mathbf{e}^e$ y $\mathbf{e}^p$ representan la parte elástica y plástica del tensor de esfuerzo, respectivamente. Las incógnitas principales del modelo son el desplazamiento $\boldsymbol{u}$, la deformación plástica $\mathbf{e}^p$ y el endurecimiento interno $\boldsymbol{\xi}$, mientras que las ecuaciones se reducen a

$$\begin{aligned} \operatorname{div}\boldsymbol{\sigma} + \alpha \boldsymbol{f_u} = \mathbf{0}, \qquad \boldsymbol{\sigma} = \boldsymbol{C}(\mathbf{e}(\boldsymbol{u}) - \mathbf{e}^p), \qquad \boldsymbol{\chi} = -\boldsymbol{H}\boldsymbol{\xi} \quad \text{in} \quad \Omega, \\ \boldsymbol{u} = \mathbf{0}, \qquad \boldsymbol{\sigma}\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on} \quad \partial\Omega, \end{aligned} \tag{3.41}$$

donde $\boldsymbol{\chi}$ es la fuerza conjugada a $\boldsymbol{\xi}$ y $\boldsymbol{H}$ representa el módulo del endurecimiento. Adicionalmente, el modelo considera una ley de flujo que gobierna la evolución de las variables de deformación plástica y endurecimiento interno. La importancia de un modelo de este tipo, es debido a que ciertos tejidos humanos, como por ejemplo el tejido pulmonar, pueden presentar un comportamiento de plasticidad cuando están sometidos a tensiones por encima de su rango elástico.

References

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, Second edition. Pure and Applied Mathematics (Amsterdan), 140. Elsevier/Academic Press, Amsterdam, 2003.

[2] J. A. Almonacid and G. N. Gatica, *A fully-mixed finite element method for the n-dimensional Boussinesq problem with temperature-dependent parameters*, Computational Methods in Applied Mathematics, 20 (2020), pp. 187–213.

[3] J. A. Almonacid, G. N. Gatica, and R. Oyarzúa, *A new mixed finite element method for the n-dimensional Boussinesq problem with temperature-dependent viscosity*, Networks and Heterogeneous Media, 15 (2020), pp. 215–245.

[4] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wellsa, *The FEniCS project version 1.5*, Archive of Numerical Software, 3 (2015), pp. 9–23.

[5] M. Alvarez, G. N. Gatica, and R. Ruiz-Baier, *An augmented mixed-primal finite element method for a coupled flow-transport problem*, ESAIM: Mathematical Modelling and Numerical Analysis, 49 (2015), pp. 1399–1427.

[6] ——, *A posteriori error analysis for a viscous flow-transport problem*, ESAIM: Mathematical Modelling and Numerical Analysis, 50 (2016), pp. 1789–1816.

[7] ——, *A posteriori error estimation for an augmented mixed-primal method applied to sedimentation-consolidation systems*, Journal of Computational Physics, 367 (2018), pp. 332–346.

[8] R. E. Amelon, K. Cao, K. Ding, G. E. Christensen, J. M. Reinhardt, and M. L. Raghavan, *Three-dimensional characterization of regional lung deformation*, Journal of Biomechanics, 44 (2011), pp. 2489–2495.

[9] D. N. Arnold, F. Brezzi, and J. Douglas, *Peers: A new mixed finite element method for plane elasticity*, Japan Journal of Applied Mathematics, 1 (1984), pp. 347–367.

[10] D. N. Arnold, R. Falk, and R. Winther, *Mixed finite element methods for linear elasticity with weakly imposed symmetry*, Mathematics of Computation, 76 (2007), pp. 1699–1723.

[11] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, *A reproducible evaluation of ants similarity metric performance in brain image registration*, Neuroimage, 54 (2011), pp. 2033–2044.

[12] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, *An unsupervised learning model for deformable medical image registration*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2018), pp. 9252–9260.

[13] N. Barnafi, G. N. Gatica, and D. E. Hurtado, *Primal and mixed finite element methods for deformable image registration problems*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 2529–2567.

[14] N. Barnafi, G. N. Gatica, D. E. Hurtado, W. Miranda, and R. Ruiz-Baier, *A posteriori error estimates for primal and mixed finite element approximations of the deformable image registration problem*, Preprint 2018-50, Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción, Chile, (2018).

[15] ——, *New primal and dual-mixed finite element methods for stable image registration with singular regularization*, Mathematical Models and Methods in Applied Sciences, (2021).

[16] M. A. Barrientos, G. N. Gatica, and E. P. Stephan, *A mixed finite element method for nonlinear elasticity: two-fold saddle point approach and a-posteriori error estimate*, Numerische Mathematik, 91 (2002), pp. 197–222.

[17] M. A. Bees and O. A. Croze, *Mathematics for streamlined biofuel production from unicellular algae*, Biofuels, 5 (2014), pp. 53–65.

[18] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, 44. Springer, Heidelberg, 2013.

[19] J. L. Boldrini, M. A. Rojas-Medar, and M. D. Rojas-Medar, *Existence and uniqueness of stationary solutions to bioconvective flow equations*, Electronic Journal Differential Equations, 2013 (2013), pp. 1–15.

[20] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Method*, Third edition. Texts in Applied Mathematics, 15. Springer, New York, 2008.

[21] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, 15. Springer-Verlag, New York, 1991.

[22] F. Brezzi, J. D. Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Istituto di Analisi Numerica del Consiglio Nazionale delle Ricerche 435, Pavia, 1984.

[23] J. Camaño, G. N. Gatica, R. Oyarzúa, and G. Tierra, *An augmented mixed finite element method for the Navier–Stokes equations with variable viscosity*, SIAM Journal on Numerical Analysis, 54 (2015), pp. 1069–1092.

[24] Y. Cao and S. Chen, *Analysis and finite element method approximation of bioconvection flows with concentration dependent viscosity*, International Journal of Numerical Analysis and Modeling, 11 (2014), pp. 86–101.

[25] C. Carstensen and G. Dolzmann, *A posteriori error estimates for mixed FEM in elasticity*, Numerische Mathematik, 81 (1998), pp. 187–209.

[26] S. Caucao, G. N. Gatica, and R. Oyarzúa, *Analysis of an augmented fully-mixed formulation for the coupling of the Stokes and heat equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 52 (2018), pp. 1947–1980.

[27] S. Childress and R. Peyret, *A numerical study of two-dimensional convection by motile particles*, Journal de Mécanique, 15 (1976), pp. 753–779.

[28] S. Choi, E. A. Hoffman, S. E. Wenzel, M. H. Tawhai, Y. Yin, M. Castro, and C. L. Lin, *Registration-based assessment of regional lung function via volumetric CT images of normal subjects vs. severe asthmatics*, Journal of Applied Physiology, 115 (2013), pp. 730–742.

[29] G. E. Christensen, J. H. Song, W. Lu, I. E. Naqa, and D. A. Low, *Tracking lung tissue motion and expansion/compression with inverse consistent image registration and spirometry*, Medical Physics, 34 (2007), pp. 2155–2163.

[30] P. G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[31] P. Clément, *Approximation by finite element functions using local regularisation*, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, 9 (1975), pp. 77–84.

[32] B. Climent-Ezquerra, L. Friz, and M. A. Rojas-Medar, *Time-reproductive solutions for a bioconvective flow*, Annali di Matematica Pura ed Applicata, 192 (2013), pp. 763–782.

[33] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, *Design and construction of a realistic digital brain phantom*, IEEE Transactions on Medical Imaging, 17 (1998), pp. 463–468.

[34] E. Colmenares, G. N. Gatica, and W. Miranda, *Analysis of an augmented fully-mixed finite element method for a bioconvective flows model*, Journal of Computational and Applied Mathematics, 393 (2021), pp. 1–25.

[35] E. Colmenares, G. N. Gatica, and R. Oyarzúa, *Analysis of an augmented mixed-primal formulation for the stationary Boussinesq problem*, Numerical Methods for Partial Differential Equations, 32 (2016), pp. 445–478.

[36] ——, *Fixed point strategies for mixed variational formulations of the stationary Boussinesq problem*, Comptes Rendus Mathematique, 354 (2016), pp. 57–62.

[37] ——, *An augmented fully-mixed finite element method for the stationary Boussinesq problem*, Calcolo, 54 (2017), pp. 167–205.

[38] A. Coronel, L. Friz, I. Hess, and A. Tello, *A result on existence and uniqueness of stationary solutions for a bioconvective flow model*, Journal of Function Spaces, 2018 (2018), pp. 1–5.

[39] T. A. Davis, *Algorithm 832: UMFPACK V4.3–an unsymmetric-pattern multifrontal method*, ACM: Transactions on Mathematical Software, 30 (2004), pp. 196–199.

[40] R. de Aguiar, B. Climent-Ezquerra, M. A. Rojas-Medar, and M. D. Rojas-Medar, *On the convergence of Galerkin spectral methods for a bioconvective flow*, Journal of Mathematical Fluid Mechanics, 19 (2017), pp. 91–104.

[41] A. Decoene, A. Lorz, S. Martin, B. Maury, and M. Tang, *Simulation of self-propelled chemotactic bacteria in a Stokes flow*, ESAIM: Proceedings, 30 (2010), pp. 104–123.

[42] A. Decoene, S. Martin, and B. Maury, *Microscopic modelling of active bacterial suspensions*, Mathematical Modelling of Natural Phenomena, 6 (2011), pp. 98–129.

[43] Domínguez, G. N. Gatica, and A. Márquez, *A residual-based a posteriori error estimator for the plane linear elasticity problem with pure traction boundary conditions*, Journal of Computational and Applied Mathematics, 292 (2016), pp. 486–504.

[44] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM Journal on Numerical Analysis, 33 (1994), pp. 1106–1124.

[45] M. Ebner, M. Modat, S. Ferraris, S. Ourselin, and T. Vercauteren, *Forward-backward splitting in deformable image registration: A demons approach*, IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, (2018), pp. 1065–1069.

[46] L. Fauci, *A computational model of the fluid dynamics of undulatory and flagellar swimming*, American Zoologist, 36 (1996), pp. 599–607.

[47] G. N. Gatica, *Analysis of a new augmented mixed finite element method for linear elasticity allowing* $\mathbb{RT}_0-\mathbb{P}_1-\mathbb{P}_0$ *approximations*, ESAIM: Mathematical Modelling and Numerical Analysis, 40 (2006), pp. 1–28.

[48] ——, *Introducción al Análisis Funcional. Teoría y Aplicaciones*, Editorial Reverte, Barcelona Bogotá Buenos Aires Caracas México, 2014.

[49] ——, *A Simple Introduction to the Mixed Finite Element Method. Theory and Applications*, SpringerBriefs in Mathematics. Springer, Cham, Heidelberg New York Dordrecht London, 2014.

[50] G. N. Gatica, A. Márquez, and S. Meddahi, *A new dual-mixed finite element method for the plane linear elasticity problem with pure traction boundary conditions*, Computer Methods in Applied Mechanics and Engineering, 197 (2008), pp. 1115–1130.

[51] G. N. Gatica and W. L. Wendland, *Coupling of mixed finite elements and boundary elements for a hyperelastic interface problem*, SIAM Journal on Numerical Analysis, 34 (1997), pp. 2335–2356.

[52] S. Ghorai and N. Hill, *Development and stability of gyrotactic plumes in bioconvection*, Journal of Fluid Mechanics, 400 (1999), pp. 1–31.

[53] ——, *Periodic arrays of gyrotactic plumes in bioconvection*, Physics of Fluids, 12 (2000), pp. 5–22.

[54] ——, *Wavelengths of gyrotactic plumes in bioconvection*, Bulletin of Mathematical Biology, 62 (2000), pp. 429–450.

[55] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer Series in Computational Mathematics, 5. Springer-Verlag, Berlin, 1986.

[56] E. Haber, S. Heldmann, and J. Modersitzki, *Adaptive mesh refinement for nonparametric image registration*, SIAM Journal on Scientific Computing, 30 (2008), pp. 3012–3027.

[57] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, *Optimal mass transport for registration and warping*, International Journal of Computer Vision, 60 (2004), pp. 225–240.

[58] A. Harashima, M. Watanabe, and I. Fujishiro, *Evolution of bioconvection patterns in a culture of motile flagellates*, Physics of Fluids, 31, 764 (1988), pp. 764–775.

[59] F. Hecht, *New development in FreeFem++*, Journal of Numerical Mathematics, 20 (2012), pp. 251–265.

[60] M. Hopkins and L. Fauci, *A computational model of the collective fluid dynamics of motile micro-organisms*, Journal of Fluid Mechanic, 455 (2002), pp. 149–174.

[61] B. K. Horn and B. G. Schunc, *Determining optical flow*, Technical Report, Massachusetts Institute of Technology, Cambridge, MA, USA, (1980), pp. 1–7.

[62] D. E. Hurtado, B. Erranz, F. Lillo, M. Sarabia-Vallejos, P. Iturrieta, F. Morales, K. Blaha, T. Medina, F. Diaz, and P. Cruces, *Progression of regional lung strain and heterogeneity in lung injury : assessing the evolution under spontaneous breathing and mechanical ventilation*, Annals of Intensive Care, 10 (2020).

[63] D. E. Hurtado, N. Villarroel, C. Andrade, J. Retamal, G. Bugedo, and A. R. Bruhn, *Spatial patterns and frequency distributions of regional deformation in the healthy human lung*, Biomechanics and Modeling in Mechanobiology, 16 (2017), pp. 1413–1423.

[64] D. E. Hurtado, N. Villarroel, J. Retamal, G. Bugedo, and A. R. Bruhn, *Improving the accuracy of registration-based biomechanical analysis: A finite element approach to lung regional strain quantification*, IEEE Transactions on Medical Imaging, 35 (2016), pp. 580–588.

[65] Y. Kan-on, K. Narukawa, and Y. Teramoto, *On the equations of bioconvective flow*, Journal of Mathematics of Kyoto University, 32 (1992), pp. 135–153.

[66] M. Kuchta, K. A. Mardal, and M. Mortensen, *On the singular neumann problem in linear elasticity*, Numerical Linear Algebra with Applications, e2212 (2018).

[67] A. V. Kuznetsov, *The onset of bioconvection in a suspension of negatively geotactic microorganisms with high-frequency vertical vibration*, International Communications in Heat and Mass Transfer, 32 (2005), pp. 1119–1127.

[68] E. Lauga and T. R. Powers, *The hydrodynamics of swimming microorganisms*, Reports on Progress in Physics, 72 (2009), pp. 1–36.

[69] E. Lee and M. Gunzburger, *An optimal control formulation of an image registration problem*, Journal of Mathematical Imaging and Vision, 36 (2010), pp. 69–80.

[70] H. G. Lee and J. Kim, *Numerical investigation of falling bacterial plumes caused by bioconvection in a three-dimensional chamber*, European Journal of Mechanics - B/Fluids, 52 (2015), pp. 120–130.

[71] M. Levandowsky, W. S. Childress, S. H. Hutner, and E. A. Spiegel, *A mathematical model of pattern formation by swimming microorganisms*, Journal of Protozoology, 22 (1975), pp. 296–306.

[72] J.-G. Liu and A. Lorz, *A coupled chemotaxis-fluid model: global existence*, Annales de l'I.H.P. Analyse non linéaire, 28 (2011), pp. 643–652.

[73] M. Lonsing and R. Verfürth, *On the stability of BDMS and PEERS elements*, Numerische Mathematik, 99 (2004), pp. 131–140.

[74] J. Modersitzki, *Numerical Methods for Image Registration*, Numerical Mathematics and Scientific Computation, Oxford Science Publications, New York, 2004.

[75] Y. Moribe, *On the bioconvection of Tetrahymena pyriformis*, Master's thesis (in Japanese), Osaka University, 1973.

[76] R. Oyarzúa, T. Qin, and D. Schötzau, *An exactly divergence-free finite element method for a generalized Boussinesq problem*, IMA: Journal of Numerical Analysis, 34 (2014), pp. 1104–1135.

[77] A. Pawar, Y. Zhang, Y. Jia, X. Wei, T. Rabczuk, C. L. Chan, and C. Anitescu, *Adaptive FEM-based nonrigid image registration using truncated hierarchical B-splines*, Computers and Mathematics with Applications, 72 (2016), pp. 2028–2040.

[78] T. J. Pedley and J. O. Kessler, *Hydrodynamic phenomena in suspensions of swimming microorganisms*, Annual review of fluid mechanics, 24 (1992), pp. 313–358.

[79] C. Pöschl, J. Modersitzki, and O. Scherzer, *A variational setting for volume constrained image registration*, Inverse Problems and Imaging, 4 (2010), pp. 505–522.

[80] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Series in Computational Mathematics, 23. Springer-Verlag, Berlin, 1994.

[81] P.-A. Raviart and J.-M. Thomas, *Introduction à l'Analyse Numérique des Équations aux Dérivées Partielles*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1983.

[82] J. Retamal, D. E. Hurtado, N. Villarroel, A. Bruhn, G. Bugedo, M. B. P. Amato, E. L. V. Costa, G. Hedenstierna, A. Larsson, and J. B. Borges, *Does regional lung strain correlate with regional inflammation in acute respiratory distress syndrome during nonprotective ventilation? An experimental porcine study*, Critical Care Medicine, 46 (2018), pp. 591–599.

[83] J. E. Roberts and J.-M. Thomas, *Mixed and Hybrid Methods*, Handbook of numerical analysis, vol. II, 523–639, North-Holland, Amsterdam, 1991.

[84] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[85] A. Sotiras, C. Davatzikos, and N. Paragios, *Deformable medical image registration: A survey*, IEEE Transactions on Medical Imaging, 32 (2013), pp. 1153–1190.

[86] I. Tuval, L. Cisneros, C. Dombrowski, C. W. Wolgemuth, J. O. Kessler, and R. E. Goldstein, *Bacterial swimming and oxygen transport near contact lines*, Proceedings of the National Academy of Sciences USA, 102 (2005), pp. 2227–2282.

[87] G. Unal and G. Slabaugh, *Coupled PDEs for non-rigid registration and segmentation*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1 (2005), pp. 168–175.

[88] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, Journal of Computational and Applied Mathematics, 50 (1994), pp. 67–83.

[89] ——, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Stuttgart, 1996.

[90] ——, *A review of a posteriori error estimation techniques for elasticity problems*, Computer Methods in Applied Mechanics and Engineering, 176 (1999), pp. 419–440.

[91] ——, *A Posteriori Error Estimation Techniques for Finite Element Methods*, Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[92] C. R. Vogel, *Computational Methods for Inverse Problems. With a foreword by H. T. Banks*, Frontiers in Applied Mathematics, 23. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[93] J. Wlazlo, R. Fessler, R. Pinnau, N. Siedow, and O. Tse, *Elastic image registration with exact mass preservation*, Arxiv preprint arXiv:1609.04043, (2016).

[94] J. Zhang, J. Wang, X. Wang, and D. Feng, *The adaptive FEM elastic model for medical image registration*, Physics in Medicine & Biology, 59 (2013), pp. 97–118.

[95] B. Zitová and J. Flusser, *Image registration methods: a survey*, Image and Vision Computing, 21 (2003), pp. 977–1000.