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 DE DIFUSIÓN ACOPLADOS EN MECÁNICA

MIXED FINITE ELEMENT METHODS FOR COUPLED DIFFUSION PROBLEMS IN MECHANICS

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

Bryan Andrés Gómez Vargas
concepción-chile
2019

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

Cotutor: Ricardo Ruiz Baier
Mathematical Institute
University of Oxford, United Kingdom

MIXED FINITE ELEMENT METHODS FOR COUPLED DIFFUSION PROBLEMS IN MECHANICS

Bryan Andrés Gómez Vargas

Directores de Tesis: Gabriel N. Gatica, Universidad de Concepción, Chile.
Ricardo Ruiz Baier, University of Oxford, United Kingdom.

Director de Programa: Rodolfo Rodríguez, Universidad de Concepción, Chile.

Comisión evaluadora

Prof. Johnny Guzman, Brown University, USA.

Prof. Sarvesh Kumar, Indian Institute of Space Science and Technology, India.

Prof. Paolo Zunino, Politecnico di Milano, Italy.

Comisión examinadora

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

Firma: _____
Prof. Michael Karkulik, Universidad Técnica Federico Santa María, Chile.

Firma: _____
Prof. Ricardo Oyarzúa, Universidad del Bío-Bío, Chile.

Firma: _____
Prof. Manuel Solano, Universidad de Concepción, Chile.

Calificación: _____

Concepción, 19 de Diciembre de 2019

Abstract

The aim of this thesis is to develop new mixed finite element methods for generating approximate solutions to problems governed by coupled systems of partial differential equations arising in the modelling of fluid and solid mechanics. In particular, we focus on two models: stress-assisted diffusion and a phase change framework. Due to the scarce information concerning mathematical and numerical analysis for these specific models, in this thesis we propose to establish well-posed finite element approaches in order to obtain existence and uniqueness of the solution. Thus, for the mathematical and numerical analysis, we introduce primal and mixed schemes, and then, by using classical techniques and results, we prove the solvability of the continuous and discrete problems, and establish the corresponding error estimates. In turn, for all problems mentioned above, numerical experiments validate the theory. Moreover, several tests illustrate the applicability of these schemes, including the simulation of microscopic electrode damage in lithium ion batteries, phase change in a cuboid cavity, and melting of solid materials.

We begin with the mathematical and numerical analysis of a coupled elasticity-diffusion system modeling the transport phenomena and chemical interactions within a solid. The coupling is introduced with a so-called stress-enhanced diffusion framework, where the propagation of the species is affected by stresses generated by solid motion. The system is fomulated in terms of stress, displacement and rotation for the elasticity equations, whereas concentration is used for the diffusion problem. For the mathematical analysis, two variational formulations are proposed, namely mixed-primal and augmented mixed-primal approaches. The solvability of the resulting coupled formulations is established by combining fixed-point arguments, regularity estimates, Babuška-Brezzi theory and the Lax-Milgram lemma. We then construct corresponding Galerkin discretisations based on adequate finite element spaces, and derive optimal a priori error estimates.

Next, we analyse the model presented above, but we do so based on a fully-mixed formulation. Based on regularity considerations, an augmented mixed formulation for the diffusion problem is proposed, whereas the classical stress-displacement-rotation mixed formulation is used for the elasticity equations. The resulting Galerkin scheme yields an augmented fully-mixed finite element method employing Arnold-Falk-Winther elements for elasticity, and a Raviart-Thomas in conjunction with a piecewise polynomial triplet for the mixed-diffusion equations. The well-known Schauder and Brouwer fixed point theorems are utilised to establish the existence of solutions of the continuous and discrete formulations, respectively. Then, sufficiently small data allow us to prove uniqueness of solution and to derive optimal a priori error estimates.

In addition, a posteriori error analysis and adaptive computations in two dimensions are further carried out for the aforementioned mixed-primal and fully mixed approaches. For the analysis of

the reliability of the residual-based a posteriori error indicators, we proceed using continuous global inf-sup conditions that come from the well-posedness of the continuous problem, together with stable Helmholtz decompositions, and approximation properties of the involved interpolation operators. Also, we use localisation techniques through edge- and face-bubble functions as well as inverse and discrete trace inequalities, in an adequately modified context, to derive the efficiency of the estimators.

Next, we address the modelling of phase change in Boussinesq-type models within porous media. A finite element method is proposed for its numerical approximation, where the properties of stability, existence and uniqueness of the continuous and discrete equations are established using classical techniques for nonlinear evolutive problems, such as Galerkin's method, Gronwall's inequality and Brouwer's fixed point theorem. Next, we test the performance of the method using a classical benchmark for air convection, where the scaled viscosity is one, there is no porosity, and no enthalpy terms. Then, we simulate the melting of a material, where the phase change is incorporated using either viscosity or porosity as then main effect producing the interface movement.

Finally, we present two new augmented variational formulations for a stationary phase change problem, namely, mixed-primal and fully-mixed formulations. Taking advantage of the regularity assumed for the velocity, we do not introduce here the rotation as an additional unknown, which is one of the novelties of this part. Thus, the main unknowns associated with the method are: the pseudostress, strain rate and velocity for the Navier-Stokes-Brinkman equations, whereas temperature, normal heat flux on the boundary, and an auxiliary unknown are introduced for the energy conservation equation. We prove solvability of both continuous and discrete problems, and derive the corresponding error analysis.

Resumen

El objetivo de esta tesis es desarrollar nuevos métodos de elementos finitos mixtos para generar soluciones aproximadas a problemas acoplados que se rigen por sistemas de ecuaciones diferenciales parciales, los cuales surgen en la mecánica de fluidos y sólidos. En particular, nos enfocamos en dos modelos: difusión asistida por esfuerzo y un problema de cambio de fase. Debido a la poca información matemática y numérica relacionada con este tipo específico de problemas, en esta tesis proponemos establecer aproximaciones bien puestas de elementos finitos, con la intención de obtener existencia y unicidad de la solución. Así, para el análisis matemático y numérico, introducimos esquemas mixtos y primales, y entonces, usando técnicas y resultados clásicos, probamos la solubilidad de los problemas continuos y discretos, y establecemos las estimaciones de error correspondientes. A su vez, para todos los problemas mencionados anteriormente, se presentan experimentos numéricos que validan la teoría propuesta. Además, se presenta una variedad de ejemplos aplicados de interés, los cuales incluyen: la simulación del daño de electrodos microscópicos en baterías de iones de litio, cambio de fase en una cavidad cuboide, y derretimiento de un material sólido.

Comenzamos con el análisis matemático y numérico de un sistema acoplado regido por las ecuaciones de elasticidad-difusión, el cual, modela los fenómenos de transporte y las interacciones químicas dentro de un sólido. El acoplamiento se introduce por medio de la difusión asistida por esfuerzo, donde la propagación de especies se ve afectada debido a los esfuerzos generados por el movimiento del sólido. El sistema se formula en términos del esfuerzo, desplazamiento y rotación para las ecuaciones de elasticidad, mientras que la concentración es usada para el problema de difusión. Para el análisis matemático, se proponen dos formulaciones variacionales, las cuales llamamos: aproximaciones mixta-primal y completamente mixta aumentada. La solubilidad de las formulaciones resultantes se establece combinando argumentos de punto fijo, estimaciones de regularidad, teoría de Babuška-Brezzi y lema de Lax-Milgram. Luego, construimos las correspondientes discretizaciones de Galerkin basadas en espacios de elementos finitos adecuados y derivamos estimaciones de error a priori óptimas.

A continuación, analizamos el modelo presentado anteriormente, pero ahora basados en una formulación completamente mixta. Por razones de regularidad, proponemos aquí una formulación mixta aumentada para el problema de difusión, mientras que la clásica formulación mixta de esfuerzo, desplazamiento y rotación se utiliza para las ecuaciones de elasticidad. El esquema de Galerkin resulta en un método aumentado completamente mixto de elementos finitos, el cual utiliza los elementos Arnold-Falk-Winther para la elasticidad, y un triplete dado por Raviart-Thomas en conjunto con elementos polinomiales a trozos para la ecuación mixta de difusión. Los clásicos teoremas de punto fijo de Schauder y Brouwer se utilizan para establecer la existencia de solución, tanto para la formulación continua como para la discreta. Luego, bajo el supuesto de dato pequeño, nos es posible demostrar

unicidad de la solución y obtener estimaciones de error a priori óptimas.

Adicionalmente, análisis de error a posteriori y adaptatividad computacional son desarrollados para las formulaciones mixta-primal y completamente mixta mencionadas anteriormente. Para el análisis de la confiabilidad de los indicadores de error basados en términos residuales, procedemos usando la condición inf-sup continua, la cual viene dada de la solubilidad del problema continuo, en conjunto con descomposiciones estables de Helmholtz, donde aprovechamos las propiedades de aproximación de los operadores de interpolación. Además, utilizamos técnicas de localización basada en funciones burbuja sobre triángulos y lados, desigualdades inversas y una desigualdad de trazas discreta, para derivar la eficiencia de los estimadores.

Por otro lado, trabajamos con un modelo de cambio de fase del tipo Boussinesq dentro de medios porosos. Proponemos un método de elementos finitos para su aproximación numérica, donde, las propiedades de estabilidad, existencia y unicidad de las formulaciones continuas y discretas son establecidas aplicando técnicas clásicas para problemas evolutivos no lineales, tales como: el método de Galerkin, la desigualdad de Gronwall y el teorema del punto fijo de Brouwer. Luego, probamos el rendimiento del método utilizando un problema clásico de referencia para la convección del aire, donde la viscosidad escalada es uno, no hay porosidad, ni términos de entalpía. En segundo lugar, simulamos el derretimiento de un material sólido, donde el cambio de fase se incorpora de dos maneras alternativas: ya sea, usando viscosidad o porosidad como principales efectos que producen el movimiento de la interfaz.

Finalmente, cerramos esta tesis presentando dos nuevas formulaciones variacionales aumentadas para un problema estacionario de cambio de fase, las cuales llamamos: formulaciones mixta-primal y totalmente mixta. Aprovechando la regularidad asumida para la velocidad, no necesitamos introducir aquí la rotación como una nueva incógnita, lo cual es una de las novedades de esta tesis. Así, las principales incógnitas asociadas a nuestro método son: el pseudo-esfuerzo, la tensión y la velocidad para las ecuaciones de de Navier-Stokes-Brinkman, mientras que la temperatura, el flujo de calor normal en la frontera y una incógnita auxiliar son introducidas para la ecuación de conservación de energía. Probamos la solubilidad de los problemas continuos y discretos, y derivamos el análisis de error correspondiente.

Agradecimientos

En primer lugar, agradezco a Dios por haberme permitido llegar hasta aquí, por darme fuerzas cuando más las necesité e ilusión cuando se iba esfumando. En segundo lugar, a mi esposa María Amalia, sin ella jamás hubiera iniciado este camino y mucho menos culminarlo con éxito. Sus incontables consejos, palabras de ánimo y abrazos, fueron mi motivación diaria. Gracias de verdad por toda su paciencia y lucha a mi lado, ha sido la mejor compañera en toda esta aventura vivida. Por todo lo anterior, y por todo lo que posiblemente olvidé mencionar, esta tesis se la dedico a usted. A mis papás, Egidio y Dinorah, por siempre creer en mí, y estar pendientes de cada paso dado, han sido un baluarte fundamental en todo este proceso. A mis familias Gómez-Vargas y Salazar-Alvarado, gracias a todos por apoyarme constantemente, estar siempre para mí cuando los he necesitado y llorar conmigo cuando ha sido necesario.

A mi tutor de tesis, el Dr. Gabriel N. Gatica, por haber creído en mí y aceptarme como su estudiante. Por todo su tiempo, sus consejos y enseñanzas académicas en todos estos años. Por estar siempre pendiente de mi avance y motivarme a dar lo mejor de mí. Su entusiasmo por la docencia e investigación me han enseñado muchísimo en todo este proceso.

A mi co-tutor de tesis, el Dr. Ricardo Ruiz Baier, quien ha sido fundamental en todo este proceso. Gracias por toda su paciencia, su apoyo incondicional, sus constantes consejos y su disposición siempre que fuese necesaria. Por toda la hospitalidad mostrada en cada una de las dos pasantías realizadas; estaré siempre agradecido por toda la amabilidad y cariño mostrado, tanto a mí, como a mi esposa.

A mis estimados compañeros de generación, Cristian, Paul, Rafa y Willian, por haberme acompañado en toda esta aventura vivida. Por apoyarme cuando los necesité y convertirse en grandes amigos para mí. Ha sido un gusto aprender, compartir y crecer con cada uno de ustedes.

A todos mis amigos y amigas que me han acompañado en estos años, tanto aquellos que me abrieron las puertas de sus casas y me enseñaron lo lindo de ser chileno, como aquellos extranjeros que sin dudarlo me han acogido como un compatriota más: Miranda, Manuel, Andrea Claudia, Javier, Silvia, don Héctor, doña Lilian, Oscar, Amaidy, Jéssica, Daviel, Maray, Ariadna, Fabián, Yisley.

A los compañeros y compañeras del doctorado, con los cuales he compartido grandes momentos: Joaquín, Daniel, Eduardo, Sergio, Paulo, Iván, Mauricio, Yissedt, Adrián, Néstor, Yolanda, Juan Paulo y Romel. Al personal administrativo del CI^{2}MA y del DIM, especialmente a la señoras Lorena Carrasco y Cecilia Leiva, por todas sus gestiones y buen trato siempre. Además, agradezco inmensamente tanto al profesor Raimund Bürger, como al profesor Rodolfo Rodríguez por todo su apoyo y amistad durante sus gestiones como directores del programa de doctorado.

Un agradecimiento especial a la Sección de Matemática de la Sede de Occidente, por creer siempre en

mí y apoyarme para realizar mis estudios doctorales. En particular, a los profesores Carlos Márquez, Carlos Ulate, Patricia Maroto y Carlos Bonilla por mostrarme todas las oportunidades existentes, desconocidas para mí en aquel momento.

A mi hermano Juan Gabriel Gómez, a Bolívar Ramírez y a Jesús Rodríguez, por la confianza brindada sin pedir nada a cambio.

Finalmente, pero no por eso menos importante, agradezco profundamente a la Dirección de Postgrado, de la Universidad de Concepción (UdeC), a la Comisión Nacional de Ciencia y Tecnología (CONICYT), y a la Universidad de Costa Rica (UCR), por haber financiado mis estudios doctorales, así como al Centro de Investigación en Ingeniería Matemática (CI^{2}MA) por brindarme un espacio para trabajar durante mis estudios.

Bryan Andrés Gómez Vargas

Contents

Abstract iii

Resumen v

Agradecimientos vii

Contents ix

List of Tables xiii

List of Figures xv

Introduction 1

Introducción 7

1 Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems 12

1.1 Introduction . . . 12

1.2 A model for stress-assisted diffusion in elastic solids . . . 14

1.3 The mixed-primal formulation . . . 15

1.3.1 The continuous setting . . . 15

1.3.2 Fixed-point approach and well-posedness of the uncoupled problems . . . 17

1.3.3 Solvability of the fixed-point equation . . . 20

1.4 A mixed-primal Galerkin scheme . . . 22

1.4.1 The mixed-primal discrete formulation . . . 22

1.4.2 Discrete fixed-point approach . . . 23

1.4.3 Solvability of the discrete fixed-point equation . . . 24

1.4.4 Specific finite element subspaces . . . 25
1.4.5 A priori error analysis . . . 27
1.5 An augmented mixed-primal formulation . . . 30
1.5.1 The continuous setting . . . 30
1.5.2 The discrete scheme . . . 34
1.5.3 A priori error analysis . . . 35
1.6 Numerical results . . . 37

2 Formulation and analysis of fully-mixed methods for stress-assisted diffusion problems 44

2.1 Introduction . . . 44
2.2 The model problem . . . 45
2.3 Weak formulation and solvability analysis . . . 46
2.3.1 The mixed-mixed formulation . . . 46
2.3.2 A fixed-point approach . . . 48
2.3.3 Solvability analysis of the fixed-point equation . . . 52
2.4 The Galerkin scheme and well-posedness of the discrete problem . . . 54
2.5 Error analysis for the proposed Galerkin method . . . 57
2.6 Numerical results . . . 61

3 A posteriori error analysis of mixed finite element methods for stress-assisted diffusion problems 68

3.1 Introduction . . . 68
3.2 The stress-assisted diffusion problem . . . 69
3.2.1 Governing equations . . . 69
3.3 Continuous and discrete mixed formulations . . . 70
3.3.1 Mixed-primal approach . . . 70
3.3.2 Fully-mixed approach . . . 72
3.4 Residual-based a posteriori error estimators . . . 74
3.4.1 Preliminaries . . . 74
3.4.2 A posteriori error analysis for the mixed-primal scheme . . . 76
3.4.3 A posteriori error analysis for the fully-mixed scheme . . . 85
3.5 Numerical results . . . 90

4 Stability and finite element approximation of phase change models for natural convection in porous media **98**

4.1 Introduction . . . 98

4.2 Phase-change Boussinesq models . . . 99

4.2.1 Main assumptions and model equations . . . 99

4.2.2 Enthalpy-porosity models for phase change . . . 100

4.2.3 Enthalpy-viscosity models for phase change . . . 101

4.2.4 Relationship with the rheology of suspended particles . . . 102

4.3 Analysis of Boussinesq phase change models . . . 102

4.3.1 Weak formulation . . . 102

4.3.2 Stability analysis . . . 104

4.4 Two families of finite element schemes . . . 107

4.4.1 A conforming method in primal formulation . . . 107

4.4.2 A mixed-primal finite element method . . . 111

4.4.3 Consistent linearisation . . . 114

4.5 Numerical verification . . . 115

4.5.1 Experimental convergence for the semidiscrete and fully discrete methods . . . 115

4.5.2 Benchmark test: natural convection of air . . . 116

4.6 Examples using phase-change models . . . 119

4.6.1 Simulating the melting of N-octadecane . . . 119

4.6.2 Changing the size of the mushy region and the jump nonlinearity . . . 120

4.6.3 Flow patterns in a local element . . . 122

5 New mixed finite element methods for natural convection with phase-change in porous media **125**

5.1 Introduction . . . 125

5.2 The model problem . . . 126

5.3 The mixed-primal approach . . . 127

5.3.1 The continuous formulation . . . 128

5.3.2 Solvability analysis . . . 131

5.3.3 The Galerkin scheme . . . 136

5.3.4 A priori error analysis . . . 139

5.3.5 Specific finite element subspaces . . . 142

5.4 The fully-mixed approach . . . 143
5.4.1 The continuous formulation . . . 143
5.4.2 Solvability analysis . . . 144
5.4.3 The Galerkin scheme . . . 147
5.4.4 A priori error analysis . . . 148
5.4.5 Specific finite element subspaces . . . 149
5.5 Numerical tests . . . 150
5.5.1 Preliminary notations . . . 150
5.5.2 Tests for the mixed-primal scheme . . . 150
5.5.3 Tests for the fully-mixed scheme . . . 153

Conclusions, summary and future work **159**

Conclusiones, sumario y trabajo futuro **163**

References **168**

List of Tables

1.1 Example 1: Degrees of freedom, meshsizes, errors, rates of convergence, and number of Picard iterations for the mixed-primal PEERS-P_1 and augmented $\mathbf{RT}_k - \mathbf{P}_{k+1} - \mathbb{P}_k - \mathrm{P}_{k+1}$ approximations of the coupled problem with $k = 0, 1$, and using $\nu = 0.4$ and $\kappa_2 = 0.5\mu, \kappa_4 = \mu$. In the first block of the table, the displacement error is measured in the $\mathbf{L}^2-$norm. 37

1.2 Example 1: Error history produced using a higher Poisson ratio $\nu = 0.49999$ and setting $\kappa_2 = \kappa_4 = 0.001\mu$. In the first block of the table, the displacement error is measured in the $\mathbf{L}^2-$norm. 39

1.3 Example 3: Experimental error history against a reference (fine mesh) solution, and number of Picard iterations per refinement level. Lowest-order augmented method. . 43

2.1 Example 1: Convergence history and Picard iteration count for the augmented $\mathbf{BDM}_{k+1} - \mathbf{P}_k - \mathbb{P}_k - \mathbf{RT}_k - \mathbf{P}_k - \mathrm{P}_{k+1}$ approximations with $k = 0, 1$. Here N stands for the number of degrees of freedom associated to the each triangulation $\mathcal{T}_h$. 61

3.1 Example 1: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity index for the first- and second-order mixed-primal finite element methods. . 92

3.2 Example 1: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity indexes for the first- and second-order augmented fully-mixed finite element methods. 92

3.3 Example 2: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity index for the lowest-order mixed-primal finite element method. 94

3.4 Example 2: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity indexes for the lowest-order augmented fully-mixed finite element method. . 95

4.1 Error history (errors on a sequence of successively refined grids, experimental convergence rates, and Newton iteration count at each refinement level) associated to the spatial discretisation using the finite element spaces (4.4.1) with $k = 1$ and $k = 2$. . . . 116

4.2 Time discretisation errors produced with a BDF2 method on different timestep resolutions, convergence rates, and average number of Newton iterations. 117

4.3 Error history (errors on a sequence of successively refined grids, experimental convergence rates, and Newton iteration count at each refinement level) associated to the spatial discretisation using the mixed-primal formulation from Section 4.4.2, using a lowest-order scheme with $k = 0$. 117

4.4 Average Nusselt number (4.5.3) and maximum velocities on the midplanes attained at $(0.5, y_\infty)$ and $(x_\infty, 0.5)$, computed for different values of the Rayleigh number and compared with respect to reference values from [66]. 119

5.1 Example 5.2.1. Convergence history for $k = 0, 1$. 151

5.2 Example 5.3.1. Convergence history and Picard iteration count for $k = 0, 1$. 155

5.3 Example 5.3.2. Convergence history and Picard iteration count for $k = 0, 1$. 157

List of Figures

1.1 Example 1: $\mathbf{RT}_0-\mathbf{P}_1-\mathbb{P}_0-\mathrm{P}_1$ approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), relevant component of the rotation tensor $\boldsymbol{\rho}_h$ (c), and concentration of the diffusive substance ϕ_h (d); using $\nu = 0.4$. All fields are plotted on the deformed domain. 39

1.2 Example 2: Approximate solutions (stress components, displacement magnitude with directions, rotation, and concentration) using a lowest order PEERS-Lagrange scheme displayed on the undeformed domain (a); and individual errors computed with respect to a reference solution (b). 40

1.3 Example 2: Concentration profiles of the diffusive substance ϕ_h plotted on the deformed domain, for different values of the additional diffusivity constants. 41

1.4 Example 3: Augmented mixed-primal approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), rotation tensor magnitude $|\boldsymbol{\rho}_h|$ (c), and concentration of the diffusive substance ϕ_h (d); all plotted on the deformed domain and showing the undeformed, skeleton mesh. 42

1.5 Example 3: Iteration count produced when varying the coupling parameters defining the concentration-dependent body load and displacement-dependent source (a), and the stress-assisted diffusivity parameters (b). 43

2.1 Example 1: Lowest-order approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), relevant component of the rotation $\boldsymbol{\rho}_h$ (c), gradient of concentration $|\boldsymbol{t}_h|$ (d), diffusive flux $|\widetilde{\boldsymbol{\sigma}}_h|$ (e), and solute concentration ϕ_h (f). All fields are plotted on the deformed domain 62

2.2 Example 2. Approximation of different functions $\vartheta(\boldsymbol{\sigma})$ varying the $\boldsymbol{\sigma}_{11}$ component (a), normalised $\mathbf{L}^2$-norm for $\boldsymbol{\sigma}_h$ and $\boldsymbol{t}_h$, $\boldsymbol{\ell}^\infty$-norm for $\boldsymbol{u}_h$ and number of Picard iterations needed for different values of β with E=100 (b), and normalized $\mathbf{L}^2$-norm for $\boldsymbol{\sigma}_h, \boldsymbol{t}_h$ and $\widetilde{\boldsymbol{\sigma}}_h$ for five different values of α (c) 64

2.3 Example 2. Schematic representation of domain boundaries on a secondary particle silicone anode (a), lowest-order approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (b), displacement magnitude $|\boldsymbol{u}_h|$ (c), rotation components (d,e,f), concentration gradient $|\boldsymbol{t}_h|$ (g), diffusive flux $|\widetilde{\boldsymbol{\sigma}}_h|$ (h), and solute concentration ϕ_h (i). 65

2.4 Example 3: Geometry for a perforated cylindrical particle (a), approximate displacement magnitude (b), magnitude of the rows of the approximate Cauchy stress (c,d,e), concentration gradient (f), diffusive flux (g), and concentration (h) shown on a clipped geometry. 66

3.1 Example 1: Approximation of the stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), rotation magnitude $|\boldsymbol{\rho}_h|$ (c), diffusive flux magnitude $|\widetilde{\boldsymbol{\sigma}}_h|$ (d), concentration of the diffusive substance ϕ_h (e), and concentration gradient magnitude $|\boldsymbol{t}_h|$ (f), by using the lowest-order augmented fully-mixed scheme with adaptive refinement according to $\widetilde{\boldsymbol{\Theta}}$. 93

3.2 Example 2: From left to right, three snapshots of successively refined meshes according to the indicators $\boldsymbol{\Theta}$ (a,b,c), $\widetilde{\boldsymbol{\Theta}}$ (d,e,f), and $\widehat{\boldsymbol{\Theta}}$ (g,h,i). 96

3.3 Example 2: Plot of the total error *versus* the number of degrees of freedom N associated with the uniform mesh refinement and adaptive algorithms according to $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$ (a); and approximate stress magnitude (b), rotation magnitude (c), displacement magnitude (d), and solute concentration (e) computed using the lowest-order scheme where mesh adaptation is done via the estimator $\boldsymbol{\Theta}$ after eight steps of refinement. 97

4.1 Velocity, pressure, and temperature profiles for the 2D differentially heated cavity (a,b,c respectively) and comparisons to the benchmark data in [154] (d and e). 118

4.2 Example 3. Comparison between the melting of N-octadecane using enthalpy-viscosity (a,b,c) and enthalpy-porosity models (d,e,f). 120

4.3 Example 4: Contour plots of the phase change depending on the size of the mushy region on enthalpy-viscosity models (a), or on the nonlinearity of the regularisation in enthalpy-porosity models (b). 121

4.4 Example 4. Velocity components, pressure, temperature, and adapted mesh at $t = 160$, using an enthalpy-porosity model with $n = 5$. 122

4.5 Example 5. Velocity and pressure profiles on a local element. 123

4.6 Temperature profiles at three different times of the phase change dynamics using (4.2.10). 124

5.1 Example 5.2.1. Lowest-order approximate solutions: (a)-(c) pseudostress entries, (d) displacement magnitude, (e) strain rate, (f) postprocessed pressure, (g) temperature, (h) effective viscosity, and (i) effective porosity fields. 152

5.2 Example 5.2.1. Errors associated with the mixed-primal approximation *versus* DoFs for $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathrm{P}_1 - \mathrm{P}_0$ and $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathrm{P}_2 - \mathrm{P}_1$ finite elements (left and right, respectively). 153

5.3 Example 5.2.2. Computed solutions with the lowest-order mixed-primal scheme. (a) pseudo-stress magnitude, (b) velocity magnitude, (c) temperature. 154

5.4 Example 5.3.1. Errors associated with the fully-mixed approximation *versus* DoFs for $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ and $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2$ finite elements (left and right, respectively). 156

5.5 Example 5.3.2. Lowest-order approximate solutions: (a)-(b) relevant components of the strain rate, (c) pseudostress magnitude, (d) displacement magnitude, (e) postprocessed pressure, and (f) temperature 158

Introduction

Mathematical models given in terms of partial differential equations (PDEs) can be found in areas such as physics, geology, biology, medicine, engineering, to name a few. A major part of these models concern the coupling between classical PDEs, which allows us to understand the behaviour of the involved unknowns when interacting with each other. The coupling can occur in a variety of manners; in particular, we are interested in models in which the coupling arises from the dependency of the constitutive equation or source term on the unknowns coming from other models. There is a large number of these models, including transport problems through viscous flow in porous media [14, 15, 41], phase-change models [44, 64, 101, 134, 155, 158, 165], stress-assisted diffusion problems [6, 83, 125, 144], among others.

Since in general the governing equations of these models are difficult to solve analytically, we need to generate approximate solutions, using for instance, numerical methods. Here is where Numerical Analysis plays a major role, not only in producing approximations, but also studying under which conditions the systems of equations have solutions, and determining stability and convergence properties of the methods. Numerical Analysis constitutes a very active research topic, and it involves PDE theory, approximation tools, and scientific computing. In this context, finite element methods have been demonstrated to be a valuable tool, allowing to obtain approximate solutions in finite dimensional spaces, and generating tools to simulate problems of interest. In this thesis, finite element methods will be developed in both primal and mixed form, the latter being the main topic in the chapters presented later. In general, mixed methods propose variational formulations which are motivated by the introduction of additional unknowns of interest. Depending on the model, these unknowns can be: stress, rotation (vorticity in fluids), pseudostress, total pressure, heat flux, to name a few. We would like to make a special mention of models involving linear elasticity. Here, the main focus is on the dual unknown, which describes the stresses within a material, and on the primal unknown describing its deformations. Thus, if the material is elastic and incompressible, it is not possible to determine the stress only from the displacement without experiencing the so-called volumetric locking in our numerical computations. It is well-known that an additional unknown can assist in characterising better the volumetric changes of the material and the resulting method can capture better the solid motion when the material approaches the incompressibility limit. Contributions dealing with mixed formulations for linear elasticity include [23, 24, 79, 81]. For heat diffusion problems, where by introducing new unknowns, it is possible to obtain a better approximation for the gradient of the temperature than the one obtained from the approximation of the temperature when solving the corresponding primal formulation. Approaches that combine mixed formulations for incompressible flow with mixed formulations for diffusing quantities have been used mainly for natural convection problems, or Boussinesq-type models [9, 10, 57, 61, 76, 123].

According to the above discussion, the purpose of this thesis is the introduction of new discretisations for several coupled problems in continuum mechanics. In particular, we are interested in proposing coupled models in the context of solid and fluid mechanics, such as stress-assisted diffusion, and phase-change problems. Throughout the development of this thesis we will work in the derivation of the corresponding variational formulations based on both primal and mixed methods and then, study the mathematical properties (e.g. existence, uniqueness, stability and regularity of solutions), discrete schemes and error estimates of the proposed systems, by means of classical fixed point theory, Lax Milgram lemma, Babuška-Brezzi theory and Strang type-inequalities. Finally, we remark that in the following two sections, we present the models, some of their applications and respective references, and the outline of the thesis (which details the mathematical and numerical properties for each model).

Model problems

This thesis adresses coupling problems in two directions: solid and fluid mechanics. Regarding fluid mechanics, we focus on phase change problems [64, 134]. Here, the problems are modeled either as a viscous Newtonian fluid where the change of phase is encoded in the viscosity itself, or using a Brinkman-Boussinesq approximation where the solidification process influences the drag directly. Regarding solid and fluid mechanics, we focus on analysing the so-called stress-assisted diffusion problem [6, 125], representing diffusion-deformation processes where the stress acts as a coupling variable.

We begin by exhibiting the following system of partial differential equations

$$\begin{aligned} \boldsymbol{\sigma} &= \lambda \operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u})\,, && -\ \mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}(\phi), \\ \widetilde{\boldsymbol{\sigma}} &= \vartheta(\boldsymbol{\sigma})\nabla\phi\,, && -\ \operatorname{div}\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}), \end{aligned} \tag{1}$$

which describes balance laws governing the motion of an elastic solid occupying the domain $\Omega \subseteq \mathrm{R}^n$ and a diffusing solute interacting with it. We note that system (1) describes the constitutive relations inherent to linear elastic materials, conservation of linear momentum, the constitutive description of diffusive fluxes, and the mass transport of the diffusive substance, respectively. Here, the local concentration of species is represented by ϕ; $\boldsymbol{\sigma}$ is the Cauchy solid stress; $\boldsymbol{u}$ is the displacement field; $\boldsymbol{\varepsilon}(\boldsymbol{u})$ is the infinitesimal strain tensor; $\widetilde{\boldsymbol{\sigma}}$ is the diffusive flux; $\lambda, \mu > 0$ are the Lamé constants; $\vartheta : \mathrm{R}^{n\times n} \to \mathrm{R}^{n\times n}$ is a tensorial diffusivity function; $\boldsymbol{f} : \mathrm{R} \to \mathrm{R}^n$ is a vector field of body loads (depending on the species concentration), and $g : \mathrm{R}^n \to \mathrm{R}$ denotes a nonlinear source term depending locally on the solid displacement. We remark that stress-enhanced diffusion effects constitute the main mechanism in many applicative problems [19, 53, 65, 117, 118, 131, 141, 162], including diffusion of boron and arsenic in silicon, hydrogen diffusion in metals, voiding of aluminum conductor lines in integrated circuits, strain-aging measurements in iron, sorption in fibre-reinforced polymeric materials, drying of liquid paint layers, gels and general-purpose solute penetration, anisotropy of cardiac dynamics, and several other effects. In particular, in Section 2.6 we will focus on the simulation of microscopic electrode damage in lithium ion batteries [19, 50, 94, 110, 136].

We point out that the main difficulty of the analysis of this model lies in the stress-dependence of the diffusivity tensor ϑ. Thus, for the corresponding mathematical analysis in Chapters 1 and 2, we will need to apply suitable regularity estimates [27, 93, 121] in order to obtain the existence and uniqueness of solution at continuous level. Finally, it is important to mention that up to our knowledge, mixed

formulations specifically tailored for the stress-assisted diffusion processes are not yet available from the literature.

Regarding on formulations for fluids, we present a time-dependent model involving a homogeneous isotropic porous structure which occupies a spatial domain Ω, and which is saturated with an incompressible viscous fluid. In this way, the governing equations are written in terms of the velocity $\boldsymbol{u}(t)$, the pressure $p(t)$, and the temperature $\theta(t)$ as follows:

$$\begin{aligned} \partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} - \frac{1}{\mathrm{Re}}\,\mathbf{div}\big[2\mu(\theta)\boldsymbol{\varepsilon}(\boldsymbol{u})\big] + \nabla p + \eta(\theta)\boldsymbol{u} &= f(\theta)\boldsymbol{k}, \\ \mathrm{div}\,\boldsymbol{u} &= 0, \\ \partial_t\theta + \boldsymbol{u}\cdot\nabla\theta - \frac{1}{C\mathrm{Pr}}\,\mathrm{div}(\kappa\nabla\theta) + \partial_t s + \boldsymbol{u}\cdot\nabla s &= 0, \end{aligned} \tag{2}$$

and which state the conservation of momentum, mass, and energy with enthalpy, respectively. In this model, the fluid has kinematic viscosity ν, thermal expansion coefficient α, and nondimensional specific heat C. Moreover, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ represents here the strain rate tensor; the function $s(t)$ is the enthalpy; $\boldsymbol{k}$ stands for the unit vector pointing in the opposite direction to gravity, and η, μ are nonlinear functions of temperature that encode the permeability of the porous material and the viscosity of the fluid, respectively. We emphasize that $s(\theta)$ in (2) denotes the regularised enthalpy function and it accounts for the latent heat of fusion, i.e. the energy needed to change the phase of a material [151, 152].

It is important to remark that model (2) has many physical and industrial applications [64, 69, 99, 120, 150, 155], including ocean and atmosphere dynamics, design of double glass windows, ventilation devices, melting and solidification in the refining of metals, among others. Thus, in Section 4.5, we will apply our finite element approximation of a steady version of (2) to simulate the melting of a solid material, where the phase change is incorporated using either viscosity or porosity as main effects producing the interface movement. For the above, we will define the adimensional buoyancy force $f(\theta) = \mathrm{Ra}\theta(\mathrm{Pr}\,\mathrm{Re}^2)^{-1}$, where the Reynolds, Rayleigh and Prandtl numbers are defined by $\mathrm{Re} = \rho_{\mathrm{ref}} V_{\mathrm{ref}} L_{\mathrm{ref}} \mu^{-1}$, $\mathrm{Ra} = g\beta L_{\mathrm{ref}}(\theta_h - \theta_c)[\nu\alpha]^{-1}$ and $\mathrm{Pr} = \nu\alpha^{-1}$, respectively. Here g represents the gravity magnitude, $L_{\mathrm{ref}}, \rho_{\mathrm{ref}}, V_{\mathrm{ref}}$ are the reference length, density, and velocity defining the flow, and θ_h, θ_c are maximum and minimum temperatures.

Outline of the thesis

This thesis is organised as follows. In **Chapter 1** we analyse the solvability of the coupled system (1). The problem is formulated in terms of solid stress, rotation tensor, solid displacement, and concentration of the solute. Existence and uniqueness of weak solutions follow from adapting a fixed-point strategy decoupling linear elasticity from a generalised Poisson equation. We then construct mixed-primal and augmented mixed-primal Galerkin schemes based on adequate finite element spaces, for which we rigorously derive a priori error bounds. The contents of this chapter gave rise to the following paper:

[83] G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems*. Computer Methods in Applied Mechanics and Engineering, vol. 337, pp. 411–438, (2018).

Chapter 2 is devoted to the mathematical and numerical analysis of a mixed-mixed PDE system describing the stress-assisted diffusion of a solute into an elastic material. The equations of elastostatics are written in mixed form using stress, rotation and displacements, whereas the diffusion equation is also set in a mixed three-field form, solving for the solute concentration, for its gradient, and for the diffusive flux. This setting simplifies the treatment of the nonlinearity in the stress-assisted diffusion term. The analysis of existence and uniqueness of weak solutions of the coupled problem follows as a combination of Schauder and Banach fixed-point theorems together with the Babuška-Brezzi and Lax-Milgram theories. Concerning numerical discretisation, we propose two families of finite element methods, based on either PEERS or Arnold-Falk-Winther elements for elasticity, and a Raviart-Thomas and piecewise polynomial triplet approximating the mixed diffusion equation. We prove the well-posedness of the discrete problems, and derive optimal error bounds using a Strang inequality. The contents of this chapter originally appeared in the following paper:

[84] G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *Formulation and analysis of fully-mixed methods for stress-assisted diffusion problems.* Computers & Mathematics with Applications, vol. 77, 5, pp. 1312–1330, (2019).

In **Chapter 3**, we develop the a posteriori error analysis for the approaches presented in Chapters 1 and 2 concerning the stress-assisted diffusion of solutes in elastic materials. The systems are formulated in terms of stress, rotation and displacements for the elasticity equations, whereas the nonlinear diffusion is cast using either solute concentration (leading to a four-field mixed-primal formulation), or the triplet concentration - concentration gradient - and nonlinear diffusive flux (yielding the six-field fully-mixed variational formulation). In this chapter, we advocate the derivation of three efficient and reliable residual-based a posteriori error estimators focusing on the two-dimensional case. The proofs of reliability depend on adequately formulated inf-sup conditions in combination with a Helmholtz decomposition and they also rely on the local approximation features of Clément and Raviart-Thomas interpolations. The efficiency of the estimators is established by a modification of classical inverse inequalities together with localisation techniques based on edge- and triangle-bubble functions. The contents of this chapter will appear in the following work currently in preparation:

[20] G.N. Gatica, B. Gómez-Vargas and R. Ruiz-Baier, *A posteriori error analysis of mixed finite element methods for stress-assisted diffusion problems.* In preparation.

In **Chapter 4**, we study a phase change problem for non-isothermal incompressible viscous flows given by (2). The underlying continuum is modelled as a viscous Newtonian fluid where the change of phase is either encoded in the viscosity itself, or in the Brinkman-Boussinesq approximation where the solidification process influences the drag directly. We address these and other modelling assumptions and their consequences in the simulation of differentially heated cavity flows of diverse types. A second order finite element method for the primal formulation of the problem in terms of velocity, temperature, and pressure is constructed, and we provide conditions for its stability. The contents of this chapter gave rise to the following paper:

[158] J. Woodfield, M. Álvarez, B. Gomez-Vargas and R. Ruiz-Baier, *Stability and finite element approximation of phase change models for natural convection in porous media.* Journal of Computational and Applied Mathematics, vol. 360, pp. 117–137, (2019).

Chapter 5 is concerned with the mathematical and numerical analysis of a phase change problem for non-isothermal incompressible viscous flow given by a steady version of (2). The system is formulated in terms of pseudostress, strain rate and velocity for the Navier-Stokes-Brinkman equation, whereas temperature, normal heat flux on the boundary, and an auxiliary unknown are introduced for the energy conservation equation. In addition, and as one of the novelties of our approach, the symmetry of the pseudostress is imposed in an ultra-weak sense, thanks to which the usual introduction of the vorticity as an additional unknown is no longer needed. Then, for the mathematical analysis two variational formulations are proposed, namely mixed-primal and fully-mixed approaches, and the solvability of the resulting coupled formulations is established by combining fixed-point arguments, Sobolev embedding theorems and certain regularity assumptions. We then construct corresponding Galerkin discretisations based on adequate finite element spaces, and derive optimal a priori error estimates. The contents of this chapter originally appeared in the following paper:

[13] M. Álvarez, G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *New mixed finite element methods for natural convection with phase-change in porous media.* Journal of Scientific Computing, vol. 80, 1, pp. 141–174, (2019).

Finally, we would like to mention that throughout all the chapters, we present several numerical tests corroborating the accuracy of the numerical schemes as well as illustrating key properties of the models. In addition, all the computational implementations of the methods were obtained using the freely available finite element libraries; FEniCS [11], FreeFem++ [96], the open mesh generator Gmsh [90], and the illustrator Paraview [5].

Preliminary notations

Let us denote by $\Omega \subseteq \mathrm{R}^n$, $n \in \{2,3\}$ a given bounded domain with polyhedral boundary $\Gamma = \partial\Omega$, and denote by $\boldsymbol{\nu}$ the outward unit normal vector on the boundary. We will adopt a fairly standard notation for Lebesgue and Sobolev spaces: $\mathrm{L}^p(\Omega)$ and $\mathrm{H}^s(\Omega)$, respectively. Norms and seminorms for the latter will be written as $\|\cdot\|_{s,\Omega}$ and $|\cdot|_{s,\Omega}$. The space $\mathrm{H}^{1/2}(\Gamma)$ contains traces of functions of $\mathrm{H}^1(\Omega)$, and $\mathrm{H}^{-1/2}(\Gamma)$ denotes its dual. In general, the notation $\mathbf{M}$ and $\mathbb{M}$ will refer to vectorial and tensorial counterparts of a generic scalar functional space M. Furthermore, by

$$\|\boldsymbol{w}\|_{\infty,\Omega} := \max_{i=1,n}\{\|w_i\|_{\infty,\Omega}\}, \quad \text{and} \quad \|\psi\|_{1,\infty,\Omega} := \max_{\alpha\leq 1}\left(\operatorname{ess\,sup}_{x\in\Omega}|\partial^\alpha\psi(x)|\right),$$

we will denote norms for the Banach spaces $\mathbf{L}^\infty(\Omega)$ and $\mathrm{W}^{1,\infty}(\Omega)$, respectively. We also need to introduce spaces of functions defined on a bounded time interval $(0,T)$ and with values in a separable Hilbert space V, with norm $\|\cdot\|_V$. Thus, for a nonnegative integer m, and for $1 \leq p < \infty$, we denote by $\mathrm{L}^p(V)$ and $\mathrm{W}^{m,p}(V)$ the spaces of classes of functions $f : (0,T) \to V$ for which

$$\|f(t)\|^p_{\mathrm{L}^p(V)} := \int_0^T \|f\|^p_V \, dt < \infty \quad \text{and} \quad \|f(t)\|^p_{\mathrm{W}^{m,p}(V)} := \sum_{l=0}^{m} \left\|\partial^l u/\partial t^l\right\|^p_{\mathrm{L}^p(V)} < \infty.$$

As usual, for brevity we write $\partial_t f$ to denote $\partial f/\partial t$, $\mathbb{I}$ stands for the identity tensor in $\mathrm{R}^{n\times n}$, and $|\cdot|$ denotes both the Euclidean norm in R^n and the Frobenius norm in $\mathrm{R}^{n\times n}$. In turn, for any vector field

$\boldsymbol{v} = (v_i)_{i=1,n}$ we set the gradient, divergence and tensor product operators as

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

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

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

Finally, we recall the following Hilbert space equipped with its usual norm

$$\mathbb{H}(\mathbf{div};\Omega) := \left\{ \boldsymbol{\tau} \in \mathbb{L}^2(\Omega) : \quad \mathbf{div}\ \boldsymbol{\tau} \in \mathbf{L}^2(\Omega) \right\}, \qquad \|\boldsymbol{\tau}\|^2_{\mathbf{div};\Omega} := \|\boldsymbol{\tau}\|^2_{0,\Omega} + \|\,\mathbf{div}\ \boldsymbol{\tau}\|^2_{0,\Omega}\,.$$

Introducción

Modelos matemáticos dados en términos de ecuaciones diferenciales parciales (EDPs) pueden ser encontrados en áreas de la física, geología, biología, medicina, ingeniería, entre otras. Una parte importante de estos modelos, tiene que ver con aquellos en donde se realiza un acoplamiento entre sistemas clásicos de ecuaciones, lo cual nos permite conocer el comportamiento de las incógnitas involucradas al interactuar entre sí. El acoplamiento se puede dar de varias formas, en particular, estamos interesados en modelos en que el acoplamiento sea dado por medio de la dependencia en las ecuaciones constitutivas o términos fuente, de incógnitas provenientes de otros modelos. Existe una gran cantidad de este tipo de modelos, tal es el caso de problemas de transporte a través de flujos viscosos en medios porosos [14, 15, 41], modelos de cambio de fase [44, 64, 101, 134, 155, 158, 165], difusión asistida por esfuerzo [6, 83, 125, 144], entre otros.

Debido a que en general, las ecuaciones gobernantes de estos modelos tienen una alta complejidad para ser resueltas analíticamente, necesitamos generar soluciones aproximadas, usando para esto, métodos numéricos. Aquí, el Análisis Numérico juega un rol primordial, no solo construyendo aproximaciones, sino también estudiando bajo qué condiciones los sistemas de ecuaciones tienen solución, y determinando propiedades de estabilidad y convergencia de los métodos. Análisis Numérico constituye un tópico muy atractivo de investigación, el cual envuelve teoría de EDP, herramientas de aproximación, y computación científica. En este contexto, los métodos de elemento finito han demostrado ser un instrumento muy capaz, permitiendo obtener soluciones aproximadas en espacios de dimensión finita, y generando herramientas para simular problemas de interés. En esta tesis, los métodos de elemento finito serán desarrollados tanto de forma primal, como mixta, siendo este último el principal enfoque en los capítulos presentados posteriormente. Generalmente, dichos métodos, proponen formulaciones variacionales que son motivadas por la introducción de variables adicionales de interés. Dependiendo del tipo de modelo que se esté abordando, dichas variables pueden ser por ejemplo: el esfuerzo, la rotación (vorticidad en fluidos), pseudo-esfuerzo, presión total, flujo de calor, entre otros. Un apartado especial a modelos que involucran elasticidad lineal, ya que aquí, el interés principal podría concentrarse en la variable dual que describe los esfuerzos dentro de un material, además de la variable primal que describe las deformaciones del mismo. Así, si el material es elástico e incompresible, no es posible determinar el esfuerzo solo a partir del desplazamiento, sin tener que obligadamente experimentar el llamado bloqueo volumétrico en nuestro esquema numérico. Es bien sabido que una incógnita adicional puede permitir una mejor caracterización de los cambios volumétricos del material, además de que, el método resultante, puede capturar mejor los movimientos del sólido cuando el material se aproxima al límite de incompresibilidad. Así, algunas referencias al respecto de formulaciones mixtas para elasticidad son [23, 24, 79, 81]. Para problemas de calor difusivos, donde al introducir nuevas incógnitas, es posible obtener una mejor aproximación del gradiente de temperatura

que la que uno encontraría a partir de la aproximación de la temperatura misma obtenida al resolver la formulación primal correspondiente. Aproximaciones que combinan formulaciones mixtas para fluidos incompresibles con formulaciones mixtas para cantidades difusivas, han sido principalmente usadas para problemas de convección natural, o aproximaciones del tipo Boussinesq [9, 10, 57, 61, 76, 123].

De acuerdo a lo anterior, la propuesta de esta tesis es la introducción de nuevas discretizaciones para problemas en mecánica del medio continuo, siendo el enfoque mixto, la principal herramienta que se utilizará. Más precisamente, estamos interesados en proponer modelos acoplados en el contexto de mecánica de fluidos y sólidos, tales como difusión asistida por esfuerzo, y problemas de cambio de fase. A través del desarrollo de esta tesis, se trabajará en la derivación de formulaciones variacionales primales y mixtas, y entonces, estudiaremos las propiedades matemáticas (por ejemplo, existencia, unicidad, estabilidad y regularidad de las soluciones), esquemas discretos y estimaciones de error para los sistemas propuestos, por medio de teoría clásica de punto fijo, combinada con el lema de Lax Milgram, la teoría de Babuška-Brezzi e inecuaciones del tipo Strang. En los siguientes dos apartados, se presentarán los modelos trabajados, algunas de sus aplicaciones más significativas y sus respectivas referencias, y la organización de la tesis (en donde se detalla el enfoque matemático y numérico utilizado para cada modelo).

Problemas modelo

Esta tesis se enfoca en problemas acoplados en dos direcciones: mecánica de sólidos y de fluidos. Respecto a mecánica de fluidos, nos enfocamos en problemas de cambio de fase [64, 134]. Aquí los problemas son modelados ya sea, como un fluido Newtoniano viscoso, donde el cambio de fase aparece en la viscosidad misma, o usando una aproximación del tipo Brinkman-Boussinesq donde el proceso de solidificación tiene influencia directa en el arrastre. Respecto a mecánica de sólidos y fluidos, nos enfocamos en analizar un problema que representa procesos de difusión-deformación, donde el esfuerzo actúa como una variable de acoplamiento, así llamado: problemas de difusión asistido por esfuerzo [6, 125].

Empezamos mostrando el siguiente sistema de ecuaciones diferenciales parciales

$$\begin{aligned} \boldsymbol{\sigma} &= \lambda \operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u})\,, \quad && -\ \mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}(\phi), \\ \widetilde{\boldsymbol{\sigma}} &= \vartheta(\boldsymbol{\sigma})\nabla\phi\,, && -\ \operatorname{div}\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}), \end{aligned} \tag{1}$$

el cual describe, las leyes de balance que generan el movimiento de un sólido que ocupa un dominio $\Omega \subseteq \mathrm{R}^n$ y la difusión de un soluto interactuando con él. Notamos que el sistema (1) describe las relaciones constitutivas de materiales elásticos, conservación lineal de momento, la descripción constitutiva de flujos difusivos, y el transporte de masa de una sustancia difusiva, respectivamente. Aquí, la concentración local de especies está representada por ϕ, $\boldsymbol{\sigma}$ es el esfuerzo sólido de Cauchy, $\boldsymbol{u}$ es el campo desplazamiento, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ es el tensor de pequeñas deformaciones, $\widetilde{\boldsymbol{\sigma}}$ es el flujo difusivo, $\lambda, \mu > 0$ son las constantes de Lamé, $\vartheta : \mathrm{R}^{n\times n} \to \mathrm{R}^{n\times n}$ es una función tensorial de difusividad, $\boldsymbol{f} : \mathrm{R} \to \mathrm{R}^n$ es un campo vectorial (el cual depende de la concentración de especies), y $g : \mathrm{R}^n \to \mathrm{R}$ denota un término fuente no lineal que depende localmente del desplazamiento del sólido. Señalamos que, efectos de la difusión asistida por esfuerzo constituyen el principal mecanismo en muchos problemas aplicados [19, 53, 65, 117, 118, 131, 141, 162], entre los cuales podemos mencionar, la difusión de boro y arsénico

en silicio, difusión de hidrógeno en metales, anulación de líneas conductoras de aluminio en circuitos integrados, mediciones de envejecimiento por tensión en hierro, sorción en materiales poliméricos reforzados con fibra, secado de capas de pintura líquida, geles y penetración de solutos de uso general, anisotropía de dinámicas cardíacas, entre otros. En particular, en la Sección 2.6, nos enfocaremos en la simulación del daño microscópico de electrodos en baterias de litio [19, 50, 94, 110, 136].

Enfatizamos que la principal dificultad en el análisis de este modelo, radica en la dependencia del esfuerzo en el tensor de difusividad ϑ. Así, para el correspondiente análisis matemático en los Capítulos 1 and 2, necesitaremos aplicar estimaciones de regularidad adecuadas [27, 93, 121], con la intención de obtener existencia y unicidad de la solución a nivel continuo. Finalmente, es importante mencionar que hasta donde es de nuestro conocimiento, formulaciones mixtas para procesos de difusión asistido por esfuerzo, es algo que no está presente en la literatura.

Correspondiente a formulaciones para fluidos, presentamos un modelo temporal relacionado con una estructura isotrópica porosa homogénea, la cual ocupa un dominio espacial Ω, y que además está saturada con fluido viscoso incompresible. De esta manera, las ecuaciones de este modelo son escritas en términos de la velocidad $\boldsymbol{u}(t)$, la presión $p(t)$, y la temperatura $\theta(t)$ como sigue:

$$\begin{aligned} \partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} - \frac{1}{\mathrm{Re}}\,\mathbf{div}\big[2\mu(\theta)\boldsymbol{\varepsilon}(\boldsymbol{u})\big] + \nabla p + \eta(\theta)\boldsymbol{u} &= f(\theta)\boldsymbol{k}, \\ \mathrm{div}\,\boldsymbol{u} &= 0, \\ \partial_t\theta + \boldsymbol{u}\cdot\nabla\theta - \frac{1}{C\mathrm{Pr}}\,\mathrm{div}(\kappa\nabla\theta) + \partial_t s + \boldsymbol{u}\cdot\nabla s &= 0, \end{aligned} \tag{2}$$

las cuales representan, la continuidad de momento, masa, y energía con entalpía, respectivamente. En este modelo, el fluido tiene viscosidad cinemática ν, coeficiente de expansión térmica α, y calor específico adimensional C. Más aún, $\boldsymbol{\varepsilon}(\boldsymbol{u})$ representa aquí la tensión, la función $s(t)$ es la entalpía, $\boldsymbol{k}$ denota el vector unitario que apunta en dirección contraria a la gravedad, η, μ son funciones no lineales de la temperatura que involucran la permeabilidad del material poroso y la viscosidad del fluido, respectivamente. Enfatizamos que $s(\theta)$ en (2), denota la función de entalpía regularizada y es quien toma en cuenta el calor de fusión latente, es decir, la energía necesaria para el cambio de fase de un material [151, 152].

Es importante observar que (2) tiene muchas aplicaciones físicas e industriales [64, 69, 99, 120, 150, 155], las cuales incluyen, dinámicas oceánicas y atmosféricas, diseño de ventanas de doble vidrio, dispositivos de ventilación, derretimiento y solidificación en el refinamiento de metales, entre otros. Así, en la Sección 4.5, aplicaremos nuestra aproximación de elementos finitos de una versión estacionaria de (2), a la simulación del derretimiento de un material sólido, donde el cambio de fase es incorporado de dos formas alternativas: usando la viscosidad o la porosidad, como efectos principales que producen el movimiento de la interfaz. Para lo anterior, definiremos la fuerza de flotación $f(\theta) = \mathrm{Ra}\theta(\mathrm{Pr}\,\mathrm{Re}^2)^{-1}$, donde los números de Reynolds, Rayleigh y Prandtl, son definidos como $\mathrm{Re} = \rho_{\mathrm{ref}} V_{\mathrm{ref}} L_{\mathrm{ref}} \mu^{-1}$, $\mathrm{Ra} = g\beta L_{\mathrm{ref}}(\theta_h - \theta_c)[\nu\alpha]^{-1}$ y $\mathrm{Pr} = \nu\alpha^{-1}$, respectivamente. Aquí g representa la magnitud de la gravedad, $L_{\mathrm{ref}}, \rho_{\mathrm{ref}}, V_{\mathrm{ref}}$ son las longitudes de referencia, densidad, y velocidad que define el fluido, y θ_h, θ_c son las temperaturas máximas y mínimas.

Organización de la tesis

Esta tesis está organizada como sigue. En el **Capítulo 1**, analizamos la solubilidad del sistema acoplado (1). El problema es formulado en términos del esfuerzo, tensor de rotación, desplazamiento del sólido y concentración de soluto. Existencia y unicidad de la solución débil se obtiene de adaptar una estrategia de punto fijo, la cual desacopla la ecuación de elasticidad y una ecuación generalizada de Poisson. Basados entonces en espacios de elementos finitos adecuados, construimos esquemas de Galerkin del tipo mixto-primal y mixto-primal aumentado, para los cuales derivamos rigurosamente cotas de error a priori. Los contenidos de este capítulo dieron lugar al siguiente artículo:

[83] G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems*. Computer Methods in Applied Mechanics and Engineering, vol. 337, pp. 411–438, (2018).

El **Capítulo 2** es dedicado al análisis matemático y numérico de un sistema de ecuaciones diferenciales parciales mixto-mixto, el cual describe la difusión asistida por esfuerzo de un soluto dentro del un material elástico. Las ecuaciones de elasticidad son escritas en forma mixta utilizando esfuerzo, rotación y desplazamientos, mientras que la ecuación de difusión es también trabajada de forma mixta dependiente de tres campos, concentración de soluto, su gradiente y el flujo difusivo. Esta estructura simplifica el tratamiento de la no linealidad en el término difusivo. El análisis de la existencia y unicidad de la solución del problema acoplado, es obtenido de combinar los teoremas de punto fijo de Schauder y Banach, junto con las teorías de Babuška-Brezzi y Lax-Milgram. Respecto a las discretizaciones numéricas, proponemos dos familias de elementos finitos: elementos PEERS y Arnold-Falk-Winther para elasticidad, y un triplete que involucra Raviart-Thomas y elementos polinomiales a trozos para aproximar la ecuación mixta de difusión. Probamos que el problema discreto está bien puesto, y derivamos cotas de error óptimas utilizando la desigualdad de Strang. Los contenidos de este capítulo aparecen en el siguiente artículo:

[84] G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *Formulation and analysis of fully-mixed methods for stress-assisted diffusion problems*. Computers & Mathematics with Applications, vol. 77, 5, pp. 1312–1330, (2019).

En el **Capítulo 3**, desarrollamos análisis de error a posteriori para las aproximaciones presentadas en los Capítulos 1 and 2, los cuales tiene que ver con la difusión asistida por esfuerzo de solutos en materiales elásticos. Los sistemas son formulados en términos del esfuerzo, rotaciones, y desplazamientos para las ecuaciones de elasticidad, mientras que la difusión no lineal es representada usando, ya sea, concentración de soluto (derivando una formulación mixta-primal en cuatro campos), o el triplete concentración-gradiente de la concentración-y flujo difusivo no lineal (derivando una formulación completamente mixta en seis campos). En este capítulo, recomendamos derivar tres estimadores de error a posteriori confiables y eficientes del tipo residual, enfocándonos en el caso bidimensional. Las pruebas de confiabilidad dependen de las condiciones inf-sup en combinación con la descomposición de Helmholtz y las aproximaciones locales de los operadores de interpolación de Clément y Raviart-Thomas. La eficiencia de los estimadores es establecida aplicando modificaciones a las desigualdades inversas clásicas en conjunto con técnicas de localización basadas en funciones burbuja sobre triángulos y lados. Los contenidos de este capítulo aparecerán en el siguiente trabajo, el cual se encuentra actualmente en preparación:

[20] G.N. Gatica, B. Gómez-Vargas and R. Ruiz-Baier, *A posteriori error analysis of mixed finite element methods for stress-assisted diffusion problems*. En preparación.

En el **Capítulo 4**, estudiamos el problema de cambio de fase para fluidos viscosos no isotérmicos incompresibles dado por (2). El problema es modelado como un fluido viscoso Newtoniano, donde el cambio de fase se presenta por medio de la viscosidad misma, o por medio de una aproximación del tipo Brinkman-Boussinesq, donde el proceso de solidificación tiene influencia directa en el arrastre. Abordamos estos y otros supuestos del modelo y sus consecuencias, en la simulación de flujos de cavidad de diversos tipos. Un método de segundo orden de elementos finitos primales es construido en términos de la velocidad, temperatura y presión, y proporcionamos sus condiciones de estabilidad. Los contenidos de este capítulo dieron lugar al siguiente artículo:

[158] J. Woodfield, M. Álvarez, B. Gomez-Vargas and R. Ruiz-Baier, *Stability and finite element approximation of phase change models for natural convection in porous media*. Journal of Computational and Applied Mathematics, vol. 360, pp. 117-137, (2019).

El **Capítulo 5** está dedicado al análisis matemático y numérico de un problema estacionario de cambio de fase para fluidos viscosos no isotérmicos incompresibles, dado por una versión estacionaria de (2). El sistema es formulado en términos del pseudo-esfuerzo, tensión y velocidad para las ecuaciones de Navier-Stokes-Brinkman, mientras que temperatura, flujo de calor normal sobre la frontera, y una incógnita auxiliar, son introducidas para la ecuación de conservación de energía. Adicionalmente, y como una de las novedades de nuestra aproximación, la simetría del pseudo-esfuerzo es impuesta es un sentido ultra débil, lo cual permite que la introducción de la vorticidad no sea necesaria. Para el análisis matemático, dos formulaciones variacionales son propuestas: mixta-primal y completamente mixta, cuya solubilidad es establecida combinando argumentos de punto fijo, teoremas de inclusiones de Sobolev y algunos supuestos de regularidad. Construimos las correspondientes discretizaciones de Galerkin basadas en espacios de elementos finitos adecuados, y derivamos estimaciones de error a priori.

[13] M. Álvarez, G.N. Gatica, B. Gomez-Vargas and R. Ruiz-Baier, *New mixed finite element methods for natural convection with phase-change in porous media*. Journal of Scientific Computing, vol. 80, 1, pp. 141–174, (2019).

Finalmente, mencionamos que a través de todos los capítulos, presentamos ensayos numéricos que corroboran la precisión de los esquemas numéricos, además de ilustrar las principales propiedades de los modelos. Adicionalmente, todas las implementaciones computacionales de los métodos, se obtuvieron empleando las librerías de elementos finitos de acceso libre; FEniCS [11], FreeFem++ [96], el generador de mallas de código abierto Gmsh [90], y el ilustrador Paraview [5].

CHAPTER 1

Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems

1.1 Introduction

This first chapter is motivated by the mathematical and numerical investigation of stress-enhanced diffusion processes in deformable solids. Starting from the early works by e.g. Truesdell [144], Podstrigach [125], or Aifantis [6], a number of applicative studies and different models have been developed. Many of these contributions have focused on the modelling of hydrogen diffusion in metals [141], damage of electrodes in lithium ion batteries [19], sorption in fibre-reinforced polymeric materials [131], drying of liquid paint layers [65], gels and general-purpose solute penetration [98, 159], anisotropy of cardiac dynamics [53], and several other effects. Irrespective of the specific interaction under consideration, the assumptions in these models convey that the species diffuses on the elastic medium obeying a Fickean law enriched with additional contributions arising from local effects by exerted stresses.

Although there exist numerous advances on the modelling considerations for stress-assisted and strain-assisted diffusion problems, their counterparts from the viewpoint of mathematical and numerical analysis are still far behind. A few punctual references include the study of plane steady solutions [97], asymptotic analysis [63, 65], and the very recent general well-posedness theory for static and transient problems in a primal formulation, developed in [109]. Our goal at this stage is to focus on a simple stationary problem that represents the main ingredients of diffusion-deformation interaction models where the Cauchy stress acts as a coupling variable. We will concentrate on the regime of linear elasticity, and we will further assume that there are no additional nonlinearities in the diffusion process other than the coupling through stresses. In turn, it is supposed that the diffusing species affects the motion of the solid skeleton through external forces, constituting a two-way coupled system.

Apart from stress and displacement, the elasticity equations will incorporate the tensor of solid rotations as supplementary field variable, serving to impose symmetry of the Cauchy stress. This approach has been exploited in several mixed formulations for elastostatics [22, 80, 86], and in our case has particular importance as the stress influences directly the diffusion process. In contrast, we will use a primal formulation for the diffusion equation. Then, following a similar approach to the one employed in [14] and [15], the existence and uniqueness of weak solutions to the coupled system will be established invoking the Lax-Milgram lemma, the Babuška-Brezzi theory, suitable regularity estimates, and fixed-

point arguments permitting us to decouple the solid mechanics from the generalised Poisson problem. Nevertheless, while there are in fact certain similarities with [14] and [15], it is important to remark that the problems involved deal with very different models and that there are substantial differences between the respective analyses. In particular, in [14] and [15] it was needed to assume, without proof, a regularity result, whereas in the present work the regularity estimates that are required for the analysis are either proved or available in the literature. Also, in [14] and [15] the authors were able to show existence of solution for sufficiently small data only whereas in the present work this assumption is not required for that purpose. More specifically, Schauder's fixed-point theorem will yield existence of weak solutions, whereas Banach's fixed-point theorem (in combination with assumptions on the data) will give uniqueness of solution. Additionally, the Sobolev embedding and Rellich-Kondrachov compactness theorems will constitute essential tools in the analysis of the continuous problem. In turn, the regularity estimates needed for the uncoupled elasticity and diffusion problems will be adapted from those appearing in [27] and [75], respectively. Even if these results are valid provided one restricts the analysis to convex domains in two spatial dimensions, our computational tests indicate that this requirement may only be technical.

Regarding the numerical approximation of the problem, we propose two families of finite element discretisations: one that will follow the same mixed-primal character as in the continuous case, and a second one that utilises augmentation of the elasticity problem through redundant Galerkin contributions in order to achieve conformity and well-definiteness of appropriate terms. As a consequence, the resulting augmented scheme allows more flexibility in the choice of the finite element subspaces for the aforementioned problem. In addition, the Brouwer fixed-point theorem will be utilised to establish existence of solutions to the associated Galerkin schemes. In this context, the recent theory leading to the well-posedness of Stokes-transport coupled systems developed in [14, 15] will be modified accordingly. The convergence analysis in each case will be conducted using a blend of a Strang-type argument, Céa estimates, and the approximation properties of specific finite element spaces. To the best of our knowledge, the results presented in this chapter constitute the first rigorous analysis of continuous and discrete mixed formulations for stress-assisted diffusion problems. The structure of the chapter is as follows. Required definitions and preliminary notation are recalled in the remainder of this section, where we also present the governing equations in strong form together with main assumptions on the model. The weak formulation stated in mixed-primal form, as well as its solvability analysis, are provided in Section 1.3. We then provide a mixed-primal Galerkin method and derive existence of discrete solution along with the corresponding a priori error estimates in Section 1.4. Section 1.5 is dedicated to the derivation and analysis of an augmented mixed-primal formulation in continuous form, a suitable discretisation, and the derivation of error bounds. We then present a set of numerical examples in Section 1.6 that illustrate the accuracy and applicability of the proposed numerical schemes

1.2 A model for stress-assisted diffusion in elastic solids

The following system of partial differential equations describes balance laws governing the motion of an elastic solid occupying the domain Ω and a diffusing solute interacting with it:

$$\begin{aligned} \boldsymbol{\sigma} &= \lambda \operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u})\,, \quad && -\operatorname{\mathbf{div}}\boldsymbol{\sigma} = \boldsymbol{f}(\phi), \\ \widetilde{\boldsymbol{\sigma}} &= \widetilde{\vartheta}(\boldsymbol{\varepsilon}(\boldsymbol{u}))\nabla\phi\,, && -\operatorname{div}\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}), \end{aligned} \tag{1.2.1}$$

where ϕ represents the local concentration of species; $\boldsymbol{\sigma}$ is the Cauchy solid stress; $\boldsymbol{u}$ is the displacement field; $\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\mathrm{t}}\right)$ is the infinitesimal strain tensor (symmetrised gradient of displacements); $\widetilde{\boldsymbol{\sigma}}$ is the diffusive flux; $\lambda, \mu > 0$ are the Lamé constants (dilation and shear moduli) characterising the properties of the material; $\widetilde{\vartheta} : \mathrm{R}^{n\times n} \to \mathrm{R}^{n\times n}$ is a tensorial diffusivity function; $\boldsymbol{f} : \mathrm{R} \to \mathrm{R}^n$ is a vector field of body loads (which will depend on the species concentration), and $g : \mathrm{R}^n \to \mathrm{R}$ denotes an additional source term depending locally on the solid displacement. Specific requirements on these functions will be given below. We note that system (1.2.1) describes the constitutive relations inherent to linear elastic materials, conservation of linear momentum, the constitutive description of diffusive fluxes, and the mass transport of the diffusive substance, respectively. It also assumes that diffusive time scales are much lower than those of the elastic wave propagation, justifying the static character of the system (*cf.* [109]).

Hooke's law [81, eq. (2.36)] asserts that $\mathcal{C}^{-1}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\boldsymbol{u})$, where $\mathcal{C}^{-1}$ is the fourth order compliance tensor. This relation allows us to recast the strain-dependent diffusivity $\widetilde{\vartheta}(\boldsymbol{\varepsilon}(\boldsymbol{u}))$ as a *stress-dependent* diffusivity $\vartheta(\boldsymbol{\sigma}) := \widetilde{\vartheta}(\mathcal{C}^{-1}\boldsymbol{\sigma})$. Throughout this chapter we will suppose that ϑ is of class C^1 and uniformly positive definite, meaning that there exists $\vartheta_0 > 0$ such that

$$\vartheta(\boldsymbol{\tau})\boldsymbol{w}\cdot\boldsymbol{w} \geq \vartheta_0|\boldsymbol{w}|^2 \quad \forall\, \boldsymbol{w} \in \mathrm{R}^n, \quad \forall\, \boldsymbol{\tau} \in \mathrm{R}^{n\times n}. \tag{1.2.2}$$

We will also require uniform boundedness and Lipschitz continuity: there exist positive constants ϑ_1, ϑ_2 and L_ϑ, such that

$$\vartheta_1 \leq |\vartheta(\boldsymbol{\tau})| \leq \vartheta_2, \quad |\vartheta(\boldsymbol{\tau}) - \vartheta(\boldsymbol{\zeta})| \leq L_\vartheta|\boldsymbol{\tau} - \boldsymbol{\zeta}| \quad \forall\, \boldsymbol{\tau}, \boldsymbol{\zeta} \in \mathrm{R}^{n\times n}. \tag{1.2.3}$$

Similar assumptions will be placed on the load and source functions $\boldsymbol{f}$ and g: we suppose that there exist positive constants f_1, f_2, L_f, g_1, g_2 and L_g, such that

$$f_1 \leq |\boldsymbol{f}(s)| \leq f_2, \quad |\boldsymbol{f}(s) - \boldsymbol{f}(t)| \leq L_f|s - t| \quad \forall\, s, t \in \mathrm{R}, \tag{1.2.4}$$

$$g_1 \leq g(\boldsymbol{w}) \leq g_2, \quad |g(\boldsymbol{v}) - g(\boldsymbol{w})| \leq L_g|\boldsymbol{v} - \boldsymbol{w}| \quad \forall\, \boldsymbol{v}, \boldsymbol{w} \in \mathrm{R}^n. \tag{1.2.5}$$

Moreover, for each $\gamma \in (0, 1)$, there exists a constant $C_\gamma > 0$, such that $g(\boldsymbol{w}) \in \mathrm{H}^\gamma(\Omega)$ for each $\boldsymbol{w} \in \mathrm{H}^\gamma(\Omega)$ and

$$\|g(\boldsymbol{w})\|_{\gamma,\Omega} \leq C_\gamma \|\boldsymbol{w}\|_{\gamma,\Omega}\,. \tag{1.2.6}$$

An additional assumption is that for every $\phi \in \mathrm{H}^1(\Omega)$, we have $\boldsymbol{f}(\phi) \in \mathbf{H}^1(\Omega)$. Finally, given $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{1/2}(\Gamma)$, the following Dirichlet boundary conditions complement (1.2.1): $\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}}$ and $\phi = 0$ on Γ. Thus, we arrive at the following coupled system:

$$\begin{aligned} \boldsymbol{\sigma} &= \lambda \operatorname{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad && \text{and} \quad -\operatorname{\mathbf{div}}\boldsymbol{\sigma} = \boldsymbol{f}(\phi) \text{ in } \Omega, \quad \boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} \text{ on } \Gamma, \\ \widetilde{\boldsymbol{\sigma}} &= \vartheta(\boldsymbol{\sigma})\nabla\phi && \text{and} \quad -\operatorname{div}\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}) \text{ in } \Omega, \quad \phi = 0 \text{ on } \Gamma. \end{aligned} \tag{1.2.7}$$

Examples of specific constitutive relations for the tensor diffusivity in terms of stress appearing in the relevant literature include exponential functions of the volumetric stress for lithiation of batteries [94], simple polynomial relationships for biological materials [53], or Carreau-type laws

$$\vartheta(\boldsymbol{\sigma}) = C_0 \exp(-\mathrm{tr}\,\boldsymbol{\sigma})\mathbb{I}, \quad \vartheta(\boldsymbol{\sigma}) = C_0\,\mathbb{I} + C_1\boldsymbol{\sigma} + C_2\boldsymbol{\sigma}^2, \quad \vartheta(\boldsymbol{\sigma}) = (C_0 + C_1(1-|\boldsymbol{\sigma}|^2)^{-1/2})\mathbb{I},$$

respectively. Regarding the concentration-dependent body load we cite linear dependences modelling isotropic swelling in composite materials [103], saturation-based descriptions for viscous layers [65], or concentration gradient modulations for single-cell mechanics [132], adopting the form

$$\boldsymbol{f}(\phi) = \boldsymbol{C}\phi, \quad \boldsymbol{f}(\phi) = \boldsymbol{C}(1-\phi)^{m-1}, \quad \boldsymbol{f}(\phi) = C_0\nabla\phi,$$

respectively, where $\boldsymbol{C} \in \mathrm{R}^n, m > 1$.

1.3 The mixed-primal formulation

In this section we derive a mixed-primal variational formulation for (1.2.7) and verify the hypotheses of Schauder's fixed-point theorem, implying existence of weak solutions. In turn, an application of Banach's fixed-point theorem will be employed to prove uniqueness of solution under the assumption of adequately small data.

1.3.1 The continuous setting

The present treatment follows closely those in [81, 14]. First we note that Hooke's law can be recast in terms of the rotation tensor as follows

$$\mathcal{C}^{-1}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\boldsymbol{u}) = \nabla\boldsymbol{u} - \boldsymbol{\rho}, \quad \text{where} \quad \boldsymbol{\rho} := \frac{1}{2}(\nabla\boldsymbol{u} - \nabla\boldsymbol{u}^{\mathrm{t}}),$$

and we observe that $\boldsymbol{\rho} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) := \{\boldsymbol{\eta} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\eta} + \boldsymbol{\eta}^{\mathrm{t}} = 0\}$. The weak form associated with the first row of (1.2.7) eventually reads: find $(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) \in \mathbf{H}(\mathrm{div};\Omega) \times (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega))$ such that

$$\begin{aligned} a(\boldsymbol{\sigma},\boldsymbol{\tau}) + b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})) &= G(\boldsymbol{\tau}) \qquad \forall\ \boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega), \\ b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta})) &= F_\phi(\boldsymbol{v},\boldsymbol{\eta}) \quad \forall\ (\boldsymbol{v},\boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \end{aligned} \tag{1.3.1}$$

where the bilinear forms $a : \mathbf{H}(\mathrm{div};\Omega) \times \mathbf{H}(\mathrm{div};\Omega) \to \mathbb{R}$ and $b : \mathbf{H}(\mathrm{div};\Omega) \times (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)) \to \mathbb{R}$ are specified as

$$a(\boldsymbol{\zeta},\boldsymbol{\tau}) := \frac{1}{2\mu}\int_\Omega \boldsymbol{\zeta} : \boldsymbol{\tau} - \frac{\lambda}{2\mu(n\,\lambda + 2\mu)}\int_\Omega \mathrm{tr}(\boldsymbol{\zeta})\,\mathrm{tr}(\boldsymbol{\tau}), \tag{1.3.2}$$

$$b(\boldsymbol{\tau},(\boldsymbol{v},\boldsymbol{\eta})) := \int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\tau} + \int_\Omega \boldsymbol{\eta} : \boldsymbol{\tau}, \tag{1.3.3}$$

for $\boldsymbol{\zeta},\boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega)$ and $(\boldsymbol{v},\boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)$. In turn, the functionals $F_\phi \in \mathbf{H}(\mathrm{div};\Omega)'$ and $G \in (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega))'$ are given by

$$G(\boldsymbol{\tau}) := \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}}\rangle_\Gamma \qquad \text{and} \qquad F_\phi(\boldsymbol{v},\boldsymbol{\eta}) := -\int_\Omega \boldsymbol{f}\,(\phi)\cdot\boldsymbol{v}\,, \tag{1.3.4}$$

for $(\boldsymbol{\tau},(\boldsymbol{v},\boldsymbol{\eta})) \in \mathbf{H}(\mathrm{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))$, where $\boldsymbol{\nu}$ denotes from now on the unit outward normal on Γ, and $\langle\cdot,\cdot\rangle_\Gamma$ stands for the duality pairing of $\mathbf{H}^{-1/2}(\Gamma)$ and $\mathbf{H}^{1/2}(\Gamma)$ with respect to the inner product in $\mathbf{L}^2(\Gamma)$.

From (1.3.2) and (1.3.3) it follows that, for any $(\boldsymbol{\tau},(\boldsymbol{v},\boldsymbol{\eta}))\in\mathbf{H}(\mathrm{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))$, there holds

$$a(\mathbb{I},\boldsymbol{\tau}) = \frac{1}{n\,\lambda+2\mu}\int_\Omega \mathrm{tr}(\boldsymbol{\tau}) \qquad \text{and} \qquad b(\mathbb{I},(\boldsymbol{v},\boldsymbol{\eta})) = 0. \tag{1.3.5}$$

Algebraic manipulations then show that the bilinear form a can be recast as

$$a(\boldsymbol{\zeta},\boldsymbol{\tau}) = \frac{1}{2\,\mu}\int_\Omega \boldsymbol{\zeta}^{\mathrm{d}}\colon\boldsymbol{\tau}^{\mathrm{d}} + \frac{1}{n(n\,\lambda+2\mu)}\int_\Omega \mathrm{tr}(\boldsymbol{\zeta})\,\mathrm{tr}(\boldsymbol{\tau}) \qquad \forall\,\boldsymbol{\zeta},\boldsymbol{\tau}\in\mathbf{H}(\mathrm{div};\Omega).$$

On the other hand, we recall from [39] that $\mathbf{H}(\mathrm{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega)\oplus\mathrm{R}\mathbb{I}$, where

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

that is, for each $\boldsymbol{\tau}\in\mathbf{H}(\mathrm{div};\Omega)$ there exist unique

$$\boldsymbol{\tau}_0 := \boldsymbol{\tau} - \left\{\frac{1}{n|\Omega|}\int_\Omega \mathrm{tr}(\boldsymbol{\tau})\right\}\mathbb{I}\in\mathbb{H}_0(\mathbf{div};\Omega) \quad \text{and} \quad d := \frac{1}{n|\Omega|}\int_\Omega \mathrm{tr}(\boldsymbol{\tau})\in\mathrm{R},$$

such that $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\mathbb{I}$. In particular, we obtain from the first row of (1.2.7) that

$$\mathrm{tr}\,(\boldsymbol{\sigma}) = (n\lambda+2\mu)\,\mathbf{div}\,\boldsymbol{u},$$

which yields $\boldsymbol{\sigma} = \boldsymbol{\sigma}_0 + c\mathbb{I}$, where

$$\boldsymbol{\sigma}_0\in\mathbb{H}_0(\mathbf{div};\Omega) \quad \text{and} \quad c := \frac{n\lambda+2\mu}{n|\Omega|}\int_\Gamma \boldsymbol{u}_{\mathrm{D}}\cdot\boldsymbol{\nu}.$$

Then, replacing $\boldsymbol{\sigma}$ by the expression $\boldsymbol{\sigma}_0 + c\mathbb{I}$ in (1.3.1), applying (1.3.5) and denoting from now on the remaining unknown $\boldsymbol{\sigma}_0\in\mathbb{H}_0(\mathbf{div};\Omega)$ simply by $\boldsymbol{\sigma}$, we find that the mixed variational formulation for the elasticity problem (*cf.* first row of (1.2.7)) reduces to: find $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\in\mathbb{H}_0(\mathbf{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))$ such that

$$\begin{aligned} a(\boldsymbol{\sigma},\boldsymbol{\tau}) + b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})) &= G(\boldsymbol{\tau}) \qquad \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\ b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta})) &= F_\phi(\boldsymbol{v},\boldsymbol{\eta}) \quad \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega). \end{aligned} \tag{1.3.6}$$

On the other hand, the boundary condition for ϕ indicates the appropriate trial and test space

$$\mathrm{H}^1_0(\Omega) := \left\{\psi\in\mathrm{H}^1(\Omega): \quad \psi = 0 \ \text{ on } \ \Gamma\right\},$$

and Poincaré's inequality implies that there exists $c_p>0$, depending only on Ω and Γ, such that

$$\|\psi\|_{1,\Omega}\le c_p|\psi|_{1,\Omega} \qquad \forall\,\psi\in\mathrm{H}^1_0(\Omega). \tag{1.3.7}$$

We can then deduce a primal formulation for the diffusion equation: find $\phi\in\mathrm{H}^1_0(\Omega)$ such that

$$A_{\boldsymbol{\sigma}}(\phi,\psi) = G_{\boldsymbol{u}}(\psi) \quad \forall\,\psi\in\mathrm{H}^1_0(\Omega), \tag{1.3.8}$$

where

$$A_{\boldsymbol{\sigma}}(\phi,\psi) := \int_\Omega \vartheta(\boldsymbol{\sigma})\nabla\phi\cdot\nabla\psi \qquad \forall\,\phi,\psi\in \mathrm{H}^1_0(\Omega), \tag{1.3.9}$$

$$G_{\boldsymbol{u}}(\psi) := \int_\Omega g(\boldsymbol{u})\,\psi \qquad \forall\,\psi\in \mathrm{H}^1_0(\Omega). \tag{1.3.10}$$

In this way, the mixed-primal formulation for (1.2.7) consists in (1.3.6) and (1.3.8), that is: find $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}),\phi)\in \mathbb{H}_0(\mathbf{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))\times \mathrm{H}^1_0(\Omega)$, such that

$$\begin{aligned} a(\boldsymbol{\sigma},\boldsymbol{\tau})+b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})) &= G(\boldsymbol{\tau}) && \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\ b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta})) &= F_\phi(\boldsymbol{v},\boldsymbol{\eta}) && \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega),\\ A_{\boldsymbol{\sigma}}(\phi,\psi) &= G_{\boldsymbol{u}}(\psi) && \forall\,\psi\in\mathrm{H}^1_0(\Omega). \end{aligned} \tag{1.3.11}$$

1.3.2 Fixed-point approach and well-posedness of the uncoupled problems

In this section, we proceed similarly as in [14] and utilise a fixed-point strategy to prove that (1.3.11) is uniquely solvable. We first set $H := \mathbb{H}_0(\mathbf{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))$ and let $\mathbf{S}:\mathrm{H}^1_0(\Omega)\to H$ be the operator defined by

$$\mathbf{S}(\phi) := (\mathbf{S}_1(\phi),(\mathbf{S}_2(\phi),\mathbf{S}_3(\phi))) := (\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})) \quad \forall\phi\in\mathrm{H}^1_0(\Omega),$$

where, for a given ϕ, the triple $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))$ is the unique solution of (1.3.6). In turn, let $\widetilde{\mathbf{S}}:\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega)\to\mathrm{H}^1_0(\Omega)$ be the operator defined by

$$\widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u}) := \phi \qquad \forall\,(\boldsymbol{\sigma},\boldsymbol{u})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega),$$

where ϕ is the unique solution of (1.3.8), for a given pair $(\boldsymbol{\sigma},\boldsymbol{u})$. Then, we define the map $\mathbf{T}:\mathrm{H}^1_0(\Omega)\to\mathrm{H}^1_0(\Omega)$ as

$$\mathbf{T}(\phi) := \widetilde{\mathbf{S}}(\mathbf{S}_1(\phi),\mathbf{S}_2(\phi)) \quad \forall\phi\in\mathrm{H}^1_0(\Omega),$$

and one readily realises that solving (1.3.11) is equivalent to seeking a fixed point of the solution operator $\mathbf{T}$, that is: find $\phi\in\mathrm{H}^1_0(\Omega)$ such that

$$\mathbf{T}(\phi)=\phi. \tag{1.3.12}$$

The following technical lemma will serve to establish solvability of (1.3.6) for a given ϕ.

Lemma 1.3.1. *There exists $c_1>0$ such that*

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

Proof. See [81, Lemma 2.3]. □

We now proceed to show that the uncoupled problems defined by $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ are well-posed.

Lemma 1.3.2. *For each $\phi\in\mathrm{H}^1_0(\Omega)$ the problem (1.3.6) has a unique solution $\mathbf{S}(\phi):=(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\in H$. Moreover, there exists $c_{\mathbf{S}}>0$ independent of ϕ, such that*

$$\|\mathbf{S}(\phi)\|_H=\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_H\le c_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right\}. \tag{1.3.13}$$

Proof. Along the lines of [81, Section 2.4.3.1], we first observe that

$$|a(\boldsymbol{\zeta},\boldsymbol{\tau})| \leq \frac{1}{\mu}\,\|\boldsymbol{\zeta}\|_{\mathbf{div},\Omega}\,\|\boldsymbol{\tau}\|_{\mathbf{div},\Omega} \qquad \forall\,\boldsymbol{\zeta},\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),$$

proving that $\mathbf{A}:\mathbb{H}_0(\mathbf{div};\Omega)\to\mathbb{H}_0(\mathbf{div};\Omega)$, the operator induced by a, is bounded with $\|\mathbf{A}\|\leq\frac{1}{\mu}$. In turn we define the operator induced by the bilinear form b as $\mathbf{B}:\mathbb{H}_0(\mathbf{div};\Omega)\to\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega)$, with

$$\mathbf{B}(\boldsymbol{\tau}) := \Big(\mathbf{div}\,\boldsymbol{\tau},\frac{1}{2}(\boldsymbol{\tau}-\boldsymbol{\tau}^{\mathrm{t}})\Big) \qquad \forall\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega), \tag{1.3.14}$$

from which one readily has that $\|\mathbf{B}\|\leq 1$. Next, from (1.3.14) we deduce that

$$V := N(\mathbf{B}) = \big\{\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega):\quad \mathbf{div}\,\boldsymbol{\tau}=0\quad\text{in}\quad\Omega,\quad \boldsymbol{\tau}=\boldsymbol{\tau}^{\mathrm{t}}\quad\text{in}\quad\Omega\big\}.$$

Consequently, using Lemma 1.3.1, we find that

$$a(\boldsymbol{\tau},\boldsymbol{\tau}) \geq \frac{1}{2\mu}\|\boldsymbol{\tau}^{\mathrm{d}}\|^2_{0,\Omega} \geq \frac{c_1}{2\mu}\,\|\boldsymbol{\tau}\|^2_{0,\Omega} = \alpha\,\|\boldsymbol{\tau}\|^2_{\mathbf{div},\Omega} \qquad \forall\boldsymbol{\tau}\in V, \tag{1.3.15}$$

thus showing that a is V-elliptic with ellipticity constant $\alpha_1 := \dfrac{c_1}{2\mu}$. On the other hand, the surjectivity of $\mathbf{B}$ follows exactly as in [81, Sect. 2.4.3.1]. Finally, from (1.3.4), we find that the functionals G and F_ϕ are bounded with

$$\|G\| \leq \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \quad\text{and}\quad \|F_\phi\| \leq f_2|\Omega|^{1/2}. \tag{1.3.16}$$

Therefore, a straightforward application of the Babuška-Brezzi theory [81, Thm. 2.3] guarantees that, for each $\phi\in\mathrm{H}^1_0(\Omega)$, problem (1.3.6) has a unique solution $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\in H$, and there holds

$$\|\mathbf{S}(\phi)\|_H = \|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_H \leq c_{\mathbf{S}}\Big\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\Big\},$$

where $c_{\mathbf{S}}$ is a constant depending on α_1,μ and the inf-sup constant associated with the bilinear form b. □

The following result asserts the unique solvability of (1.3.8).

Lemma 1.3.3. *For each $(\boldsymbol{\sigma},\boldsymbol{u})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega)$, the problem (1.3.8) has a unique solution $\phi := \widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u})\in\mathrm{H}^1_0(\Omega)$. Moreover, there exists a constant $r>0$ depending on c_p, ϑ_0, g_2 and Ω (cf.(1.3.7), (1.2.2), (1.2.5)), such that*

$$\|\widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u})\|_{1,\Omega} = \|\phi\|_{1,\Omega} \leq r. \tag{1.3.17}$$

Proof. We note from (1.3.9) that $A_{\boldsymbol{\sigma}}$ is a bilinear form. Next, from (1.2.3) and (1.3.9), we deduce that

$$|A_{\boldsymbol{\sigma}}(\phi,\psi)| \leq \vartheta_2\,\|\phi\|_{1,\Omega}\,\|\psi\|_{1,\Omega} \qquad \forall\,\phi,\psi\in\mathrm{H}^1_0(\Omega),$$

which gives $\|A_{\boldsymbol{\sigma}}\|\leq\vartheta_2$, and thus $A_{\boldsymbol{\sigma}}$ is bounded independently of $\boldsymbol{\sigma}$ and $\boldsymbol{u}$. Furthermore, from (1.2.2) and the estimate (1.3.7), for each $\phi\in\mathrm{H}^1_0(\Omega)$, we find that

$$A_{\boldsymbol{\sigma}}(\phi,\phi) = \int_\Omega \vartheta(\boldsymbol{\sigma})\nabla\phi\cdot\nabla\phi \geq \frac{\vartheta_0}{c_p^2}\,\|\phi\|^2_{1,\Omega}, \tag{1.3.18}$$

which proves that $A_{\boldsymbol{\sigma}}$ is $\mathrm{H}^1_0(\Omega)$-elliptic with constant $\alpha_2 := \frac{\vartheta_0}{c_p^2}$, independently of $\boldsymbol{\sigma}$ and $\boldsymbol{u}$ as well. Now, using (1.2.5), (1.3.10) and applying Cauchy-Schwarz's inequality, we deduce that

$$|G_{\boldsymbol{u}}(\psi)| \leq g_2|\Omega|^{1/2}\,\|\psi\|_{0,\Omega} \quad \forall\, \psi \in \mathrm{H}^1_0(\Omega), \tag{1.3.19}$$

which implies that $G_{\boldsymbol{u}} \in \mathrm{H}^1_0(\Omega)'$ and $\|G_{\boldsymbol{u}}\| \leq g_2|\Omega|^{1/2}$. Thus, a straightforward application of the Lax-Milgram Lemma (see, *e.g.* [81], Thm. 1.1) proves that for each $(\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathbb{H}_0(\mathbf{div};\Omega) \times \mathbf{L}^2(\Omega)$, problem (1.3.8) has a unique solution $\phi := \widetilde{\mathbf{S}}(\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathrm{H}^1_0(\Omega)$. Moreover, the corresponding continuous dependence on the data is formulated as

$$\|\phi\|_{1,\Omega} \leq r,$$

where

$$r := \frac{c_p^2}{\vartheta_0} g_2|\Omega|^{1/2}\,. \tag{1.3.20}$$

□

The next step consists in deriving regularity estimates for the problems defining $\mathbf{S}$ and $\widetilde{\mathbf{S}}$. The following theorem (which summarizes the respective analysis in [27]) is particularly crucial in the treatment for the operator $\mathbf{S}$.

Theorem 1.3.4. *Given a convex polygonal domain $\Omega \subseteq \mathrm{R}^2$ and $\mathbf{F} \in \mathbf{L}^2(\Omega)$, we let $\boldsymbol{u}$ be the solution of the elasticity problem*

$$\begin{aligned} \mu\Delta\boldsymbol{u} + (\mu+\lambda)\nabla(\nabla\cdot\boldsymbol{u}) &= \mathbf{F} \quad \text{in} \quad \Omega, \\ \boldsymbol{u} &= \mathbf{0} \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where the Lamé moduli are bounded as $\mu \in [\mu_1, \mu_2]$ and $\lambda \in [0, \infty)$, with fixed constants $\mu_1, \mu_2 > 0$. Then, there exists $\gamma > 0$ such that whenever $\mathbf{F} \in \mathbf{H}^{\gamma}(\Omega)$, there holds $\boldsymbol{u} \in \mathbf{H}^{2+\gamma}(\Omega)$ and

$$\|\boldsymbol{u}\|_{2+\gamma,\Omega} \leq \widetilde{C}_1\,\|\mathbf{F}\|_{\gamma,\Omega}\,,$$

with a constant $\widetilde{C}_1$ independent of the Lamé coefficients.

According to Theorem 1.3.4, in what follows we should probably concentrate in the case where Ω is a convex polygonal domain and $n = 2$. Nevertheless, it is easy to see that, assuming the regularity provided by this theorem, the forthcoming analysis and all the associated results hold even for the non-convex or 3D cases. We then recall that $\boldsymbol{f}(\psi) \in \mathbf{H}^1(\Omega)$ for each $\psi \in \mathrm{H}^1_0(\Omega)$, and suppose from now on that $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{3/2+\gamma}(\Omega)$. Then, applying the theorem or the respective assumption, and recalling from the constitutive equation that the regularities of the unknowns are connected, we immediately find that $\mathbf{S}(\psi) \in \mathbb{H}_0(\mathbf{div};\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega) \times \mathbf{H}^{2+\gamma}(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega)$.

In turn, for the operator $\widetilde{\mathbf{S}}$, we invoke [92, Remark (a)] and [75, Thm. 3.12], and observe that, for a given pair $(\boldsymbol{\zeta}, \boldsymbol{w}) := (\mathbf{S}_1(\psi), \mathbf{S}_2(\psi)) \in \mathbb{H}_0(\mathbf{div};\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega) \times \mathbf{H}^{2+\gamma}(\Omega)$ (which denote the first and second components of the unique solution produced by the operator $\mathbf{S}$), the hypothesis given by relation (1.2.6) implies in particular that $g(\boldsymbol{w}) \in \mathrm{H}^{\gamma}(\Omega)$. If one further assumes that the coefficients $\vartheta(\boldsymbol{\zeta})_{ij}$ are in $C^{1+\gamma}(\overline{\Omega})$, then elliptic regularity results (cf. [93], [121]) guarantee that $\phi := \widetilde{\mathbf{S}}(\boldsymbol{\zeta}, \boldsymbol{w}) \in \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^{2+\gamma}(\Omega)$, and we conclude that there exists a constant $\widetilde{C}_2 > 0$ such that

$$\|\widetilde{\mathbf{S}}(\boldsymbol{\zeta}, \boldsymbol{w})\|_{2+\gamma,\Omega} = \|\phi\|_{2+\gamma,\Omega} \leq \widetilde{C}_2\,\|g(\boldsymbol{w})\|_{\gamma,\Omega}\,. \tag{1.3.21}$$

On the other hand, the Sobolev embedding theorem (*cf.* [2], Thm. 4.12, [107], Thm. A.5) gives the continuous injection $i_\gamma : \mathrm{H}^{2+\gamma}(\Omega) \longrightarrow C^1(\overline{\Omega})$, with boundedness constant $\widetilde{C}_\gamma$. Then, using the aforementioned continuous injection and applying (1.3.21), we deduce that

$$\|\widetilde{\mathbf{S}}(\boldsymbol{\zeta},\boldsymbol{w})\|_{1,\infty,\Omega} = \|\phi\|_{1,\infty,\Omega} \leq \widetilde{C}_\gamma \, \|\phi\|_{2+\gamma,\Omega} \leq \widetilde{C}_\gamma \widetilde{C}_2 \, \|g(\boldsymbol{w})\|_{\gamma,\Omega} \, . \tag{1.3.22}$$

Finally, using (1.2.6) and (1.3.13), we find that

$$\|\widetilde{\mathbf{S}}(\boldsymbol{\zeta},\boldsymbol{w})\|_{1,\infty,\Omega} = \|\phi\|_{1,\infty,\Omega} \leq C_\infty \, c_{\mathbf{S}} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\} , \tag{1.3.23}$$

where C_∞ is a positive constant depending on C_γ, $\widetilde{C}_\gamma$ and $\widetilde{C}_2$ (*cf.* (1.2.6), (1.3.21), (1.3.22)).

1.3.3 Solvability of the fixed-point equation

In this section we address the solvability analysis of the fixed-point equation (1.3.12). To this end, we will verify the hypotheses of the Schauder fixed-point theorem (see, *e.g.* [54, Thm. 9.12-1(b)]).

Lemma 1.3.5. *Let $r > 0$ be the constant from (1.3.20) (cf. proof of Lemma 1.3.3). Then, for the closed ball $W := \left\{ \phi \in \mathrm{H}^1_0(\Omega) : \ \|\phi\|_{1,\Omega} \leq r \right\}$, it holds that $\mathbf{T}(W) \subseteq W$.*

Proof. It suffices to recall the definition of $\mathbf{T}$ (*cf.* Section 1.3.2), and simply apply estimate (1.3.17). □

Lemma 1.3.6. *There exists $C_{\mathbf{S}} > 0$ depending on μ, L_f, α (cf.(1.2.1), (1.2.4), (1.3.15)) and the inf-sup constant of b, such that*

$$\|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_H \leq C_{\mathbf{S}} \, \|\phi - \varphi\|_{0,\Omega} \quad \forall \, \phi, \varphi \in \mathrm{H}^1_0(\Omega). \tag{1.3.24}$$

Proof. Given $\phi, \varphi \in \mathrm{H}^1_0(\Omega)$, we let $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})), (\boldsymbol{\zeta},(\boldsymbol{w},\boldsymbol{\chi})) \in H$ be two solutions to (1.3.6), corresponding to ϕ and φ, respectively. That is, $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})) = \mathbf{S}(\phi)$ and $(\boldsymbol{\zeta},(\boldsymbol{w},\boldsymbol{\chi})) = \mathbf{S}(\varphi)$. We then invoke the linearity of the forms a and b to deduce (using both formulations arising from (1.3.6)) that

$$\begin{aligned} a(\boldsymbol{\sigma}-\boldsymbol{\zeta},\boldsymbol{\tau}) + b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{w},\boldsymbol{\chi})) &= 0 && \forall \, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega), \\ b(\boldsymbol{\sigma}-\boldsymbol{\zeta},(\boldsymbol{v},\boldsymbol{\eta})) &= (F_\phi - F_\varphi)(\boldsymbol{v},\boldsymbol{\eta}) && \forall \, (\boldsymbol{v},\boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega) \end{aligned} \, . \tag{1.3.25}$$

From (1.3.4), we readily note that $\|F_\phi - F_\varphi\| \leq L_f \, \|\phi - \varphi\|_{0,\Omega}$. Consequently, and similarly to the proof of Lemma 1.3.2, the Babuška-Brezzi theory implies that for each $\phi, \varphi \in \mathrm{H}^1_0(\Omega)$, problem (1.3.25) has a unique solution $(\boldsymbol{\sigma}-\boldsymbol{\zeta},(\boldsymbol{u}-\boldsymbol{w},\boldsymbol{\rho}-\boldsymbol{\chi})) \in H$, as well as the continuous dependence on the data

$$\|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_H = \|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})) - (\boldsymbol{\zeta},(\boldsymbol{w},\boldsymbol{\chi}))\|_H \leq C_{\mathbf{S}} \, \|\phi - \varphi\|_{0,\Omega} \, ,$$

which gives (1.3.24) and concludes the proof. □

The following result is a consequence of Lemma 1.3.6.

Lemma 1.3.7. *Assume that $C_{\mathbf{S}}$ is as in* Lemma 1.3.6. *Then, for each $\phi, \varphi \in \mathrm{H}^1_0(\Omega)$, there holds*

$$\|\mathbf{T}(\phi) - \mathbf{T}(\varphi)\|_{1,\Omega} \leq \frac{1}{\alpha_2} \, C_{\mathbf{S}} \left\{ L_g + L_\vartheta \, \|\mathbf{T}(\varphi)\|_{1,\infty,\Omega} \right\} \|\phi - \varphi\|_{0,\Omega} \, . \tag{1.3.26}$$

Proof. Firstly we recall that $\mathbf{T}(\phi) = \widetilde{\mathbf{S}}(\mathbf{S}_1(\phi), \mathbf{S}_2(\phi))$ and $\mathbf{T}(\varphi) = \widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)) \quad \forall\, \phi, \varphi \in \mathrm{H}^1_0(\Omega)$. In view of unifying the notation throughout the chapter, we apply the following renaming

$$(\boldsymbol{\sigma}, \boldsymbol{u}) := (\mathbf{S}_1(\phi), \mathbf{S}_2(\phi)) \quad \text{and} \quad (\boldsymbol{\zeta}, \boldsymbol{w}) := (\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)),$$

where $(\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\zeta}, \boldsymbol{w}) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{L}^2(\Omega)$. In addition, we let $\widetilde{\phi} := \widetilde{\mathbf{S}}(\boldsymbol{\sigma}, \boldsymbol{u})$ and $\widetilde{\varphi} := \widetilde{\mathbf{S}}(\boldsymbol{\zeta}, \boldsymbol{w})$, that is

$$A_{\boldsymbol{\sigma}}(\widetilde{\phi}, \widetilde{\psi}) = G_{\boldsymbol{u}}(\widetilde{\psi}) \quad \text{and} \quad A_{\boldsymbol{\zeta}}(\widetilde{\varphi}, \widetilde{\psi}) = G_{\boldsymbol{w}}(\widetilde{\psi}) \quad \forall\, \widetilde{\psi} \in \mathrm{H}^1_0(\Omega).$$

Adding and subtracting appropriate terms, and appealing to the ellipticity of $A_{\boldsymbol{\sigma}}$, we readily find that

$$\begin{aligned} \alpha_2 \|\widetilde{\phi} - \widetilde{\varphi}\|^2_{1,\Omega} &\leq A_{\boldsymbol{\sigma}}(\widetilde{\phi}, \widetilde{\phi} - \widetilde{\varphi}) - A_{\boldsymbol{\sigma}}(\widetilde{\varphi}, \widetilde{\phi} - \widetilde{\varphi}) \\ &= (G_{\boldsymbol{u}} - G_{\boldsymbol{w}})(\widetilde{\phi} - \widetilde{\varphi}) + (A_{\boldsymbol{\zeta}} - A_{\boldsymbol{\sigma}})(\widetilde{\varphi}, \widetilde{\phi} - \widetilde{\varphi}). \end{aligned} \tag{1.3.27}$$

Next we use (1.3.9), (1.3.10), we apply Cauchy-Schwarz's inequality, and exploit the assumptions (1.2.3) and (1.2.5), to obtain the bounds

$$\begin{aligned} |(G_{\boldsymbol{u}} - G_{\boldsymbol{w}})(\widetilde{\phi} - \widetilde{\varphi})| \, &= \left| \int_\Omega (g(\boldsymbol{u}) - g(\boldsymbol{w}))(\widetilde{\phi} - \widetilde{\varphi}) \right| \\ &\leq L_g \, \|\boldsymbol{u} - \boldsymbol{w}\|_{0,\Omega} \, \|\widetilde{\phi} - \widetilde{\varphi}\|_{0,\Omega} \, , \end{aligned} \tag{1.3.28}$$

and

$$\begin{aligned} |(A_{\boldsymbol{\zeta}} - A_{\boldsymbol{\sigma}})(\widetilde{\varphi}, \widetilde{\phi} - \widetilde{\varphi})| \, &= \left| \int_\Omega (\vartheta(\boldsymbol{\zeta}) - \vartheta(\boldsymbol{\sigma})) \nabla \widetilde{\varphi} \cdot \nabla(\widetilde{\phi} - \widetilde{\varphi}) \right| \\ &\leq L_\vartheta \|\nabla \widetilde{\varphi}\|_{\infty,\Omega} \, \|\boldsymbol{\sigma} - \boldsymbol{\zeta}\|_{0,\Omega} \, |\widetilde{\phi} - \widetilde{\varphi}|_{1,\Omega} \, . \end{aligned} \tag{1.3.29}$$

We then observe that the inequalities (1.3.27)-(1.3.29) imply that

$$\|\widetilde{\phi} - \widetilde{\varphi}\|_{1,\Omega} \leq \frac{1}{\alpha_2} \left\{ L_g \, \|\boldsymbol{u} - \boldsymbol{w}\|_{0,\Omega} + L_\vartheta \, \|\widetilde{\varphi}\|_{1,\infty,\Omega} \, \|\boldsymbol{\sigma} - \boldsymbol{\zeta}\|_{0,\Omega} \right\}. \tag{1.3.30}$$

Next, according to the definitions given at the beginning of the proof, we can rewrite (1.3.30) as

$$\begin{aligned} &\|\widetilde{\mathbf{S}}(\mathbf{S}_1(\phi), \mathbf{S}_2(\phi)) - \widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)\|_{1,\Omega} \\ &\qquad \leq \frac{1}{\alpha_2} \left\{ L_g \, \|\mathbf{S}_2(\phi) - \mathbf{S}_2(\varphi)\|_{0,\Omega} + L_\vartheta \|\widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi))\|_{1,\infty,\Omega} \, \|\mathbf{S}_1(\phi) - \mathbf{S}_1(\varphi)\|_{0,\Omega} \right\}. \end{aligned} \tag{1.3.31}$$

It is important to note here that the term $\|\widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi))\|_{1,\infty,\Omega}$ is bounded for each $\varphi \in \mathrm{H}^1_0(\Omega)$, thanks to (1.3.23). In this way, we are in a position to prove the Lipschitz continuity of $\mathbf{T}$. In fact, from (1.3.24) and (1.3.31) we find that

$$\begin{aligned} \|\mathbf{T}(\phi) - \mathbf{T}(\varphi)\|_{1,\Omega} &= \|\widetilde{\mathbf{S}}(\mathbf{S}_1(\phi), \mathbf{S}_2(\phi)) - \widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)\|_{1,\Omega} \\ &\leq \frac{1}{\alpha_2} \left\{ L_g \, \|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_H + L_\vartheta \|\mathbf{T}(\varphi)\|_{1,\infty,\Omega} \, \|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_H \right\} \\ &\leq \frac{1}{\alpha_2} \, C_{\mathbf{S}} \left\{ L_g + L_\vartheta \, \|\mathbf{T}(\varphi)\|_{1,\infty,\Omega} \right\} \|\phi - \varphi\|_{0,\Omega} \, , \end{aligned}$$

which gives (1.3.26) and completes the proof. □

Lemma 1.3.8. *Let W be as in* Lemma 1.3.5. *Then, $\mathbf{T} : W \to W$ is continuous and $\overline{\mathbf{T}(W)}$ is compact.*

Proof. It follows analogously to the proof of [14, Lemma 3.12], and it is a consequence of the Rellich-Kondrachov compactness Theorem [2, Thm. 6.3] in combination with (1.3.23), and the fact that every bounded sequence in a Hilbert space has a weakly convergent subsequence. □

The main result of this section is stated next.

Theorem 1.3.9. *The mixed-primal problem* (1.3.11) *has at least one solution* $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}),\phi)\in H\times \mathrm{H}^1_0(\Omega)$ *satisfying the bounds*

$$\|\phi\|_{1,\Omega}\leq r \tag{1.3.32}$$

and

$$\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_H\leq c_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right\}. \tag{1.3.33}$$

Moreover, if the data is such that

$$\frac{1}{\alpha_2}C_{\mathbf{S}}\left\{L_g+L_\vartheta C_\infty c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}<1, \tag{1.3.34}$$

then the solution ϕ *is unique in* W.

Proof. Thanks to Lemmas 1.3.5 and 1.3.8, the existence of solution is merely an application of the Schauder fixed-point theorem. In turn, the estimates (1.3.32) and (1.3.33) follow from Lemmas 1.3.3 and 1.3.2, respectively. Furthermore, given another solution $\varphi\in W$ of (1.3.12), the estimate in (1.3.23) confirms (1.3.34) as a sufficient condition for concluding, together with (1.3.26), that $\phi=\varphi$. □

As announced in the Introduction, we notice here that, differently from the analysis in [14] and [15], the existence result provided by Theorem 1.3.9 does not require the data to be sufficiently small. However, we point out that the existence of the fourth component ϕ of the solution is restricted to the ball $W:=\left\{\phi\in \mathrm{H}^1_0(\Omega):\ \ \|\phi\|_{1,\Omega}\leq r\right\}$, whose radius r depends on the data ϑ_0 and g_2 (cf. (1.3.20)).

1.4 A mixed-primal Galerkin scheme

In this section we define a first numerical approximation associated with (1.3.11). We derive general hypotheses on the finite-dimensional subspaces defining the Galerkin finite element method, and ensuring that the discrete problem is indeed well-posed. Existence of solutions will follow by means of Brouwer's fixed-point theorem, and we will derive adequate *a priori* error estimates.

1.4.1 The mixed-primal discrete formulation

Let $\mathcal{T}_h$ be a regular partition of $\overline{\Omega}$ into triangles K of diameter h_K, where $h:=\max\{h_K:\ K\in\mathcal{T}_h\}$ is the meshsize. Let us also consider arbitrary finite-dimensional subspaces

$$\mathbb{H}^{\boldsymbol{\sigma}}_h\subseteq\mathbb{H}_0(\mathbf{div};\Omega),\quad \mathbf{H}^{\boldsymbol{u}}_h\subseteq\mathbf{L}^2(\Omega),\quad \mathbb{H}^{\boldsymbol{\rho}}_h\subseteq\mathbb{L}^2_{\mathrm{skew}}(\Omega)\quad\text{and}\quad \mathrm{H}^{\phi}_h\subseteq\mathrm{H}^1_0(\Omega),$$

whose specification will be made clear later on, in Section 1.4.4. The corresponding Galerkin scheme can be already defined as: find $(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h),\phi_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\times\mathrm{H}_h^{\phi}$ such that

$$\begin{aligned} a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)+b(\boldsymbol{\tau}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) &= G(\boldsymbol{\tau}_h) && \forall\,\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}},\\ b(\boldsymbol{\sigma}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)) &= F_{\phi_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h) && \forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}},\\ A_{\boldsymbol{\sigma}_h}(\phi_h,\psi_h) &= G_{\boldsymbol{u}_h}(\psi_h) && \forall\,\psi_h\in\mathrm{H}_h^{\phi}. \end{aligned} \tag{1.4.1}$$

A discrete analogue to the fixed-point strategy from Section 1.3.2 will be presented in what follows.

1.4.2 Discrete fixed-point approach

Let us introduce the operator $\mathbf{S}_h:\mathrm{H}_h^{\phi}\to\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})$ defined by

$$\mathbf{S}_h(\phi_h):=(\mathbf{S}_{1,h}(\phi_h),(\mathbf{S}_{2,h}(\phi_h),\mathbf{S}_{3,h}(\phi_h)))=:(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\quad\forall\phi_h\in\mathrm{H}_h^{\phi},$$

where $(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)$ solves uniquely the problem

$$\begin{aligned} a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)+b(\boldsymbol{\tau}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) &= G(\boldsymbol{\tau}_h) && \forall\,\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}},\\ b(\boldsymbol{\sigma}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)) &= F_{\phi_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h) && \forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}, \end{aligned} \tag{1.4.2}$$

with F_{ϕ_h} defined in (1.3.4) with $\phi=\phi_h$. On the other hand, we define $\widetilde{\mathbf{S}}_h:\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\to\mathrm{H}_h^{\phi}$ as

$$\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h,\boldsymbol{u}_h):=\phi_h\qquad\forall\,(\boldsymbol{\sigma}_h,\boldsymbol{u}_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}},$$

where ϕ_h is the unique solution of

$$A_{\boldsymbol{\sigma}_h}(\phi_h,\psi_h)=G_{\boldsymbol{u}_h}(\psi_h)\qquad\forall\,\psi_h\in\mathrm{H}_h^{\phi}, \tag{1.4.3}$$

with $A_{\boldsymbol{\sigma}_h}$ and $G_{\boldsymbol{u}_h}$ being defined by (1.3.9) with $\boldsymbol{\sigma}=\boldsymbol{\sigma}_h$ and (1.3.10) with $\boldsymbol{u}=\boldsymbol{u}_h$, respectively. Therefore, solving (1.4.1) is equivalent to find $\phi_h\in\mathrm{H}_h^{\phi}$ such that

$$\mathbf{T}_h(\phi_h)=\phi_h,$$

where the fixed-point operator is characterised by

$$\mathbf{T}_h:\mathrm{H}_h^{\phi}\to\mathrm{H}_h^{\phi},\qquad\mathbf{T}_h(\phi_h):=\widetilde{\mathbf{S}}_h(\mathbf{S}_{1,h}(\phi_h),\mathbf{S}_{2,h}(\phi_h))\quad\forall\phi_h\in\mathrm{H}_h^{\phi}.$$

The well-definition of $\mathbf{T}_h$ then hinges on the well-posedness of $\widetilde{\mathbf{S}}_h$ and $\mathbf{S}_h$. For the latter, we anticipate that further hypotheses on the discrete spaces $\mathbb{H}_h^{\boldsymbol{\sigma}},\mathbf{H}_h^{\boldsymbol{u}}$ and $\mathbb{H}_h^{\boldsymbol{\rho}}$ will be required. To this end, we now let V_h be the discrete kernel of b, that is

$$V_h:=\left\{\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}}\ :\ b(\boldsymbol{\tau}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))=0\quad\forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right\},$$

and assume the following discrete inf-sup conditions (which do hold for some finite element spaces, as those listed in Section 1.4.4):

[H.0] There exists a constant $\widehat{\alpha} > 0$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h \in V_h \\ \boldsymbol{\tau}_h \neq 0}} \frac{a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)}{\|\boldsymbol{\tau}_h\|_{\mathbf{div},\Omega}} \geq \widehat{\alpha} \, \|\boldsymbol{\sigma}_h\|_{\mathbf{div},\Omega} \quad \forall \, \boldsymbol{\sigma}_h \in V_h. \tag{1.4.4}$$

[H.1] There exists a constant $\widehat{\beta} > 0$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h \in \mathbf{H}_h^{\boldsymbol{\sigma}} \\ \boldsymbol{\tau}_h \neq 0}} \frac{b(\boldsymbol{\tau}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h))}{\|\boldsymbol{\tau}_h\|_{\mathbf{div},\Omega}} \geq \widehat{\beta} \, \|(\boldsymbol{v}_h, \boldsymbol{\eta}_h)\|_{\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)} \quad \forall \, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}. \tag{1.4.5}$$

Lemma 1.4.1. *For each $\phi_h \in \mathrm{H}_h^{\phi}$ the problem (1.4.2) has a unique solution $\mathbf{S}_h(\phi_h) := (\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h)) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times (\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}})$. Moreover, there exists $\widetilde{C} > 0$, depending on $\mu, \widehat{\alpha}, \widehat{\beta}$, but independent of ϕ_h, such that*

$$\|\mathbf{S}_h(\phi_h)\|_H = \|(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h))\|_H \leq \widetilde{C} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}.$$

Proof. It follows directly from the discrete Babuška-Brezzi theory [81, Thm. 2.4]. Indeed, the induced operators for the forms a and b are bounded on subspaces of the corresponding continuous spaces. Furthermore, the linear functional G restricted to $\mathbb{H}_h^{\boldsymbol{\sigma}}$ is bounded as indicated in (1.3.16), and for each $\phi_h \in \mathrm{H}_h^{\phi}$, the functional F_{ϕ_h} restricted to $\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$ is bounded as well. The remaining hypotheses are precisely **[H.0]** and **[H.1]**, and hence the proof is finished. □

Lemma 1.4.2. *Let $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$. Then, there exists a unique $\phi_h := \widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathrm{H}_h^{\phi}$ solution of (1.4.3). Moreover, with the same constant r provided by* Lemma 1.3.3*, there holds*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h)\|_{1,\Omega} = \|\phi_h\|_{1,\Omega} \leq r.$$

Proof. It suffices to note that for each $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, the operator $A_{\boldsymbol{\sigma}_h}$ is elliptic on H_h^{ϕ} with the same constant α_2 from the proof of Lemma 1.3.3, and that $G_{\boldsymbol{u}_h}$ restricted to H_h^{ϕ} is bounded as in (1.3.19). Hence, the result is a direct application of the Lax-Milgram Lemma. □

1.4.3 Solvability of the discrete fixed-point equation

The following steps verify the hypotheses of the Brouwer fixed-point theorem (see, *e.g.* [54, Thm. 9.9-2]).

Lemma 1.4.3. *For the closed ball $W_h := \left\{ \phi_h \in \mathrm{H}_h^{\phi} : \ \|\phi_h\|_{1,\Omega} \leq r \right\}$, we have that $\mathbf{T}_h(W_h) \subseteq W_h$.*

Proof. It is a straightforward consequence of Lemma 1.4.2. □

Lemma 1.4.4. *There exists $C > 0$ depending on $\mu, L_f, \widehat{\alpha}$ and $\widehat{\beta}$ (cf.(1.2.1), (1.2.4), (1.4.4), (1.4.5)) such that*

$$\|\mathbf{S}_h(\phi_h) - \mathbf{S}_h(\varphi_h)\|_H \leq C \, \|\phi_h - \varphi_h\|_{0,\Omega} \quad \forall \, \phi_h, \varphi_h \in \mathrm{H}_h^{\phi}.$$

Proof. It follows analogously to the proof of Lemma 1.3.6. □

Lemma 1.4.5. *For each* $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), (\boldsymbol{\zeta}_h, \boldsymbol{w}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, *there holds*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) - \widetilde{\mathbf{S}}_h(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)\|_{1,\Omega} \leq \frac{1}{\alpha_2} \left\{ L_g \, \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{0,\Omega} + L_\vartheta \|\nabla \widetilde{\mathbf{S}}_h(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)\|_{\infty,\Omega} \, \|\boldsymbol{\sigma}_h - \boldsymbol{\zeta}_h\|_{0,\Omega} \right\}. \tag{1.4.6}$$

Proof. Given $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), (\boldsymbol{\zeta}_h, \boldsymbol{w}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, we let $\phi_h := \widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h)$ and $\varphi_h := \widetilde{\mathbf{S}}_h(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)$. We then proceed similarly to the proof of Lemma 1.3.7 to obtain

$$\alpha_2 \, \|\phi_h - \varphi_h\|_{1,\Omega}^2 \leq \left\{ L_g \, \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{0,\Omega} + L_\vartheta \|\nabla \varphi_h\|_{\infty,\Omega} \, \|\boldsymbol{\sigma}_h - \boldsymbol{\zeta}_h\|_{0,\Omega} \right\} \|\phi_h - \varphi_h\|_{1,\Omega},$$

and realise that H_h^ϕ consists of piecewise polynomials (see Section 1.4.4) to conclude that $\|\nabla \varphi_h\|_{\infty,\Omega} < +\infty$, and hence (1.4.6) holds. □

The following result is a consequence of Lemmas 1.4.3, 1.4.4 and 1.4.5.

Lemma 1.4.6. *Let* C *be as in* Lemma 1.4.4. *Then, for all* $\phi_h, \varphi_h \in \mathrm{H}_h^\phi$, *there holds*

$$\|\mathbf{T}_h(\phi_h) - \mathbf{T}_h(\varphi_h)\|_{1,\Omega} \leq \frac{1}{\alpha_2} C \, (L_g + L_\vartheta \|\nabla \mathbf{T}_h(\varphi_h)\|_{\infty,\Omega}) \, \|\phi_h - \varphi_h\|_{0,\Omega}.$$

Proof. It follows after recalling that $\mathbf{T}_h(\phi_h) = \widetilde{\mathbf{S}}_h(\mathbf{S}_{1,h}(\phi_h), \mathbf{S}_{2,h}(\phi_h))$ for all $\phi_h \in \mathrm{H}_h^\phi$, and applying Lemmas 1.4.3, 1.4.4 and 1.4.5. □

Finally, thanks to Lemmas 1.4.3 and 1.4.6, a straightforward application of the aforementioned Brouwer fixed-point theorem implies the main result of this section, stated as follows.

Theorem 1.4.7. *The Galerkin scheme* (1.4.1) *has at least one solution* $(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h), \phi_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times (\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}) \times \mathrm{H}_h^\phi$. *Furthermore, there exists a positive constant* $\widetilde{C}$, *independent of the discretisation parameters, such that*

$$\|\phi\|_{1,\Omega} \leq r \quad \textit{and} \quad \|(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h))\|_H \leq \widetilde{C} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}.$$

1.4.4 Specific finite element subspaces

Given an integer $k \geq 0$, for each $K \in \mathcal{T}_h$ we let $\mathrm{P}_k(K)$ be the space of polynomial functions on K of degree $\leq k$ and recall the definition of the local Raviart-Thomas space of order k as $\mathbf{RT}_k(K) := \mathbf{P}_k(K) \oplus \mathrm{P}_k(K) \, \boldsymbol{x}$, where $\mathbf{P}_k(K) = [\mathrm{P}_k(K)]^2$, and $\boldsymbol{x}$ is the generic vector in R^2. In addition, we let b_K be the element bubble function defined as the unique polynomial in $\mathrm{P}_{k+1}(K)$ vanishing on ∂K with $\int_K b_K = 1$. Then, for each $K \in \mathcal{T}_h$ we consider the bubble space of order k, by

$$\mathbf{B}_k(K) := \mathrm{P}_k(K) \left(\frac{\partial b_K}{\partial x_2}, -\frac{\partial b_K}{\partial x_1} \right).$$

Appropriate finite element subspaces approximating the elasticity unknowns are as follows

$$\mathbb{H}_h^{\boldsymbol{\sigma}} := \{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div}; \Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbf{RT}_k(K) \oplus \mathbf{B}_k(K) \ \ \forall \, K \in \mathcal{T}_h \}, \tag{1.4.7}$$

$$\mathbf{H}_h^{\boldsymbol{u}} := \left\{ \boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \tag{1.4.8}$$

$$\mathbb{H}_h^{\boldsymbol{\rho}} := \left\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \quad \boldsymbol{\eta}_h \in \mathbf{C}(\Omega) \quad \text{and} \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}. \tag{1.4.9}$$

The discrete product space $\mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$ constitutes the classical PEERS elements introduced in [22] for the mixed finite element approximation of Dirichlet linear elasticity. In contrast, the approximation of the diffusion problem will be carried out using Lagrange finite elements of degree $\leq k+1$, that is

$$\mathrm{H}_h^{\phi} := \left\{ \psi_h \in \mathrm{C}(\Omega) \cap \mathrm{H}_0^1(\Omega) \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}. \tag{1.4.10}$$

Useful approximation properties of these spaces are listed as follows (see *e.g.* [39, 81]):

($\mathbf{AP}_h^{\boldsymbol{\sigma}}$) there exists $C > 0$, independent of h, such that for each $s \in (0, k+1]$, and for each $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega) \cap \mathbb{H}_0(\mathbf{div};\Omega)$ with $\mathbf{div}\,(\boldsymbol{\sigma}) \in \mathbf{H}^s(\Omega)$, there holds

$$\mathrm{dist}(\boldsymbol{\sigma}, \mathbb{H}_h^{\boldsymbol{\sigma}}) := \inf_{\boldsymbol{\tau}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}}} \|\boldsymbol{\sigma} - \boldsymbol{\tau}_h\|_{\mathbf{div},\Omega} \leq C h^s \left\{ \|\boldsymbol{\sigma}\|_{s,\Omega} + \|\mathbf{div}\,(\boldsymbol{\sigma})\|_{s,\Omega} \right\}.$$

($\mathbf{AP}_h^{\boldsymbol{u}}$) there exists $C > 0$, independent of h, such that for each $s \in (0, k+1]$, and for each $\boldsymbol{u} \in \mathbf{H}^s(\Omega)$, there holds

$$\mathrm{dist}(\boldsymbol{u}, \mathbf{H}_h^{\boldsymbol{u}}) := \inf_{\boldsymbol{v}_h \in \mathbf{H}_h^{\boldsymbol{u}}} \|\boldsymbol{u} - \boldsymbol{v}_h\|_{0,\Omega} \leq C h^s \|\boldsymbol{u}\|_{s,\Omega}.$$

($\mathbf{AP}_h^{\boldsymbol{\rho}}$) there exists $C > 0$, independent of h, such that for each $s \in (0, k+1]$, and for each $\boldsymbol{\rho} \in \mathbb{H}^s(\Omega)$, there holds

$$\mathrm{dist}(\boldsymbol{\rho}, \mathbb{H}_h^{\boldsymbol{\rho}}) := \inf_{\boldsymbol{\eta}_h \in \mathbb{H}_h^{\boldsymbol{\rho}}} \|\boldsymbol{\rho} - \boldsymbol{\eta}_h\|_{0,\Omega} \leq C h^s \|\boldsymbol{\rho}\|_{s,\Omega}.$$

($\mathbf{AP}_h^{\phi}$) there exists $C > 0$, independent of h, such that for each $s \in (0, k+1]$, and for each $\phi \in \mathrm{H}^{s+1}(\Omega)$, there holds

$$\mathrm{dist}(\phi, \mathrm{H}_h^{\phi}) := \inf_{\psi_h \in \mathrm{H}_h^{\phi}} \|\phi - \psi_h\|_{1,\Omega} \leq C h^s \|\phi\|_{s+1,\Omega}.$$

Next, we recall from [81, Sect. 4.5] that the discrete kernel of b is given by

$$V_h := \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}} : \quad \mathbf{div}\,\boldsymbol{\tau}_h = 0 \quad \text{in} \quad \Omega \quad \text{and} \quad \int_\Omega \boldsymbol{\eta}_h : \boldsymbol{\tau}_h = 0 \quad \forall\, \boldsymbol{\eta}_h \in \mathbb{H}_h^{\boldsymbol{\rho}} \right\},$$

and according to (1.3.15) and Lemma 1.3.1, the bilinear form a is V_h-elliptic, implying that [**H.0**] is satisfied. Concerning assumption [**H.1**] we have the following result, proven in [81, Sect. 4.5].

Lemma 1.4.8. *There exists $\widehat{\beta} > 0$ such that*

$$\sup_{\boldsymbol{\tau}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}} \setminus \{\mathbf{0}\}} \frac{b(\boldsymbol{\tau}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h))}{\|\boldsymbol{\tau}_h\|_{\mathbf{div},\Omega}} \geq \widehat{\beta} \, \|(\boldsymbol{v}_h, \boldsymbol{\eta}_h)\|_{\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)} \quad \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}.$$

1.4.5 A priori error analysis

Let $(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}),\phi)\in\mathbb{H}_0(\mathbf{div};\Omega)\times(\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega))\times\mathrm{H}^1_0(\Omega)$ with $\phi\in W$, and $(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h),\phi_h)\in\mathbb{H}^{\boldsymbol{\sigma}}_h\times(\mathbf{H}^{\boldsymbol{u}}_h\times\mathbb{H}^{\boldsymbol{\rho}}_h)\times\mathrm{H}^{\phi}_h$ with $\phi_h\in W_h$; be the solutions of (1.3.11) and (1.4.1), respectively. That is,

$$\begin{aligned} a(\boldsymbol{\sigma},\boldsymbol{\tau})+b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho}))&=G(\boldsymbol{\tau}) && \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\ b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta}))&=F_\phi(\boldsymbol{v},\boldsymbol{\eta}) && \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega),\\ a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)+b(\boldsymbol{\tau}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))&=G(\boldsymbol{\tau}_h) && \forall\,\boldsymbol{\tau}_h\in\mathbb{H}^{\boldsymbol{\sigma}}_h,\\ b(\boldsymbol{\sigma}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))&=F_{\phi_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h) && \forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}^{\boldsymbol{u}}_h\times\mathbb{H}^{\boldsymbol{\rho}}_h \end{aligned} \tag{1.4.11}$$

and

$$\begin{aligned} A_{\boldsymbol{\sigma}}(\phi,\psi)&=G_{\boldsymbol{u}}(\psi) && \forall\,\psi\in\mathrm{H}^1_0(\Omega),\\ A_{\boldsymbol{\sigma}_h}(\phi_h,\psi_h)&=G_{\boldsymbol{u}_h}(\psi_h) && \forall\,\psi_h\in\mathrm{H}^{\phi}_h. \end{aligned} \tag{1.4.12}$$

Next, we recall a generalised Strang inequality (*cf.* [127, Thm. 11.2]), to be applied in (1.4.11).

Lemma 1.4.9. *For Hilbert spaces X, Y, let $\boldsymbol{a}:X\times X\to\mathrm{R},\boldsymbol{b}:X\times Y\to\mathrm{R}$ be bounded bilinear forms and $F\in X',G\in Y'$ satisfying the hypotheses of the Babuška-Brezzi theory. Furthermore, let $\{X_h\}_{h>0}$ and $\{Y_h\}_{h>0}$ be sequences of finite-dimensional subspaces of X and Y, respectively, and suppose that $\boldsymbol{a},\boldsymbol{b}$ and $F_h\in X_h',G_h\in Y_h'$ satisfy the hypotheses of the discrete Babuška-Brezzi theory uniformly on X_h and Y_h, that is, there exist positive constants $\overline{\alpha}$ and $\overline{\beta}$ independent of h, such that*

$$\sup_{\substack{\boldsymbol{\phi}_h\in V_h\\ \boldsymbol{\phi}_h\neq 0}}\frac{\boldsymbol{a}(\boldsymbol{\psi}_h,\boldsymbol{\phi}_h)}{\|\boldsymbol{\phi}_h\|_X}\geq\overline{\alpha}\,\|\boldsymbol{\psi}_h\|_X\quad\forall\,\boldsymbol{\psi}_h\in V_h\quad\text{and}\quad\sup_{\substack{\boldsymbol{\psi}_h\in X_h\\ \boldsymbol{\psi}_h\neq 0}}\frac{\boldsymbol{b}(\psi_h,\mu_h)}{\|\boldsymbol{\psi}_h\|_X}\geq\overline{\beta}\,\|\mu_h\|_Y\quad\forall\,\mu_h\in Y_h, \tag{1.4.13}$$

where V_h is the discrete kernel of $\boldsymbol{b}$. Then, there exists a constant C_{ST} dependent only on $\|\boldsymbol{a}\|,\|\boldsymbol{b}\|,\overline{\alpha}$ and $\overline{\beta}$ such that if $(\varphi,\lambda)\in X\times Y$ and $(\varphi_h,\lambda_h)\in X_h\times Y_h$ are solutions to

$$\begin{aligned} \boldsymbol{a}(\varphi,\psi)+\boldsymbol{b}(\psi,\lambda)&=F(\psi) && \forall\,\psi\in X,\\ \boldsymbol{b}(\varphi,\mu)&=G(\mu) && \forall\,\mu\in Y, \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{a}(\varphi_h,\psi_h)+\boldsymbol{b}(\psi_h,\lambda_h)&=F_h(\psi_h) && \forall\,\psi_h\in X_h,\\ \boldsymbol{b}(\varphi_h,\mu_h)&=G_h(\mu_h) && \forall\,\mu_h\in Y_h, \end{aligned}$$

respectively, then for each $h>0$, there holds

$$\begin{aligned} \|\varphi-\varphi_h\|_X+\|\lambda-\lambda_h\|_Y\leq C_{\mathrm{ST}}\Bigg\{&\inf_{\substack{\boldsymbol{\psi}_h\in X_h\\ \boldsymbol{\psi}_h\neq 0}}\|\varphi-\psi_h\|_X+\inf_{\substack{\boldsymbol{\mu}_h\in Y_h\\ \boldsymbol{\mu}_h\neq 0}}\|\lambda-\mu_h\|_Y\\ &+\sup_{\substack{\boldsymbol{\phi}_h\in X_h\\ \boldsymbol{\phi}_h\neq 0}}\frac{|F(\phi_h)-F_h(\phi_h)|}{\|\phi_h\|_X}+\sup_{\substack{\boldsymbol{\eta}_h\in Y_h\\ \boldsymbol{\eta}_h\neq 0}}\frac{|G(\eta_h)-G_h(\eta_h)|}{\|\eta_h\|_Y}\Bigg\}. \end{aligned}$$

In addition to the notations introduced in the approximation properties given in Section 1.4.4, we now define

$$\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right) := \inf_{(\boldsymbol{\tau}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})}\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\tau}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))\|_H\,,$$

or, equivalently,

$$\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right) := \mathrm{dist}(\boldsymbol{\sigma},\mathbb{H}_h^{\boldsymbol{\sigma}})+\mathrm{dist}(\boldsymbol{u},\mathbf{H}_h^{\boldsymbol{u}})+\mathrm{dist}\left(\boldsymbol{\rho},\mathbb{H}_h^{\boldsymbol{\rho}}\right).$$

The following lemma provides an estimate for $\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H$.

Lemma 1.4.10. *There exists $C_{\mathrm{ST}}>0$, depending on $\mu,\widehat{\alpha}$ and $\widehat{\beta}$ (cf. (1.2.1), (1.4.4), (1.4.5)), such that*

$$\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H\leq C_{\mathrm{ST}}\big\{\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)+L_f\,\|\phi-\phi_h\|_{0,\Omega}\big\}. \tag{1.4.14}$$

Proof. We clearly observe that (1.4.4) and (1.4.5) imply that the hypothesis (1.4.13) in Lemma 1.4.9 is satisfied. Then, a straightforward application of Lemma 1.4.9 to (1.4.11), readily gives

$$\begin{aligned}&\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H\\&\quad\leq C_{\mathrm{ST}}\left\{\|(F_\phi-F_{\phi_h})|_{\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}}\|+\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)\right\}.\end{aligned} \tag{1.4.15}$$

Next, and analogously to the proof of Lemma 1.3.6, we can assert that

$$\|(F_\phi-F_{\phi_h})|_{\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}}\|\leq L_f\,\|\phi-\phi_h\|_{0,\Omega}\,, \tag{1.4.16}$$

and finally, by replacing (1.4.16) back into (1.4.15), we get the desired result. □

Lemma 1.4.11. *Let α_2 be the ellipticity constant of the bilinear form $A_{\boldsymbol{\sigma}}$ (cf. (1.3.18)). Then, there holds*

$$\|\phi-\phi_h\|_{1,\Omega}\leq\frac{L_g}{\alpha_2}\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}+\left(1+\frac{\vartheta_2}{\alpha_2}\right)\mathrm{dist}(\phi,\mathrm{H}_h^{\phi})+\frac{L_\vartheta}{\alpha_2}\,\|\phi\|_{1,\infty,\Omega}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,. \tag{1.4.17}$$

Proof. We first observe by triangle inequality that

$$\|\phi-\phi_h\|_{1,\Omega}\leq\|\phi-\psi_h\|_{1,\Omega}+\|\phi_h-\psi_h\|_{1,\Omega}\quad\forall\,\psi_h\in\mathrm{H}_h^{\phi}. \tag{1.4.18}$$

Then, applying the ellipticity of $A_{\boldsymbol{\sigma}_h}$ and adding and subtracting the expression $G_{\boldsymbol{u}_h}(\phi_h-\psi_h)=A_{\boldsymbol{\sigma}_h}(\phi_h-\psi_h)$, (*cf.* (1.4.12)) we find that

$$\begin{aligned}\alpha_2\,\|\phi_h-\psi_h\|_{1,\Omega}^2&\leq A_{\boldsymbol{\sigma}_h}(\phi_h-\psi_h,\phi_h-\psi_h)\\&\leq|G_{\boldsymbol{u}_h}(\phi_h-\psi_h)-G_{\boldsymbol{u}}(\phi_h-\psi_h)|+|A_{\boldsymbol{\sigma}}(\phi,\phi_h-\psi_h)-A_{\boldsymbol{\sigma}_h}(\psi_h,\phi_h-\psi_h)|.\end{aligned} \tag{1.4.19}$$

Next, analogously to (1.3.28), we get

$$|G_{\boldsymbol{u}_h}(\phi_h-\psi_h)-G_{\boldsymbol{u}}(\phi_h-\psi_h)|\leq L_g\,\|\boldsymbol{u}_h-\boldsymbol{u}\|_{0,\Omega}\,\|\phi_h-\psi_h\|_{0,\Omega}\,. \tag{1.4.20}$$

In turn, adding and subtracting $\displaystyle\int_\Omega \vartheta(\boldsymbol{\sigma})\nabla\phi\cdot\nabla(\phi_h-\psi_h)$, and applying the upper bound of ϑ (*cf.* (1.2.3)), we arrive at

$$\begin{aligned}&|A_{\boldsymbol{\sigma}}(\phi,\phi_h-\psi_h)-A_{\boldsymbol{\sigma}_h}(\psi_h,\phi_h-\psi_h)|\\&\qquad\leq \vartheta_2|\phi-\psi_h|_{1,\Omega}|\phi_h-\psi_h|_{1,\Omega}+L_\vartheta\,\|\nabla\phi\|_{\infty,\Omega}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,|\phi_h-\psi_h|_{1,\Omega}.\end{aligned}\tag{1.4.21}$$

Thus, the inequalities (1.4.19), (1.4.20) and (1.4.21), imply that

$$\|\phi_h-\psi_h\|_{1,\Omega}\leq\frac{L_g}{\alpha_2}\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}+\frac{\vartheta_2}{\alpha_2}\,\|\phi-\psi_h\|_{1,\Omega}+\frac{L_\vartheta}{\alpha_2}\,\|\phi\|_{1,\infty,\Omega}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,.\tag{1.4.22}$$

Finally, replacing (1.4.22) back into (1.4.18) and taking the infimum on $\psi_h\in\mathrm{H}_h^\phi$, completes the proof. □

To derive the Céa estimation for the total error $\|\phi-\phi_h\|_{1,\Omega}+\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H$, we combine the inequalities provided by Lemmas 1.4.10 and 1.4.11. For sake of notational convenience we introduce the following constants

$$C_1:=\frac{L_g}{\alpha_2}C_{\mathrm{ST}},\qquad C_2:=\frac{L_\vartheta}{\alpha_2}C_\infty C_{\mathrm{ST}},\qquad C_3:=1+\frac{\vartheta_2}{\alpha_2}.\tag{1.4.23}$$

Hence, replacing the bound for $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$ and $\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}$ into (1.4.17), applying (1.3.23), and performing algebraic manipulations, we can deduce the bounds

$$\begin{aligned}\|\phi-\phi_h\|_{1,\Omega}&\leq C_1\left\{\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)+L_f\,\|\phi-\phi_h\|_{0,\Omega}\right\}+C_3\,\mathrm{dist}(\phi,\mathrm{H}_h^\phi)\\&\quad+C_2\,c_{\mathbf{S}}\big\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\big\}\big\{\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)+L_f\,\|\phi-\phi_h\|_{0,\Omega}\big\}\\&\leq\left\{C_1+C_2\,c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}\left\{\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)\right\}\\&\quad+L_f\left\{C_1+C_2\,c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}\|\phi-\phi_h\|_{1,\Omega}+C_3\,\mathrm{dist}\,(\phi,\mathrm{H}_h^\phi).\end{aligned}\tag{1.4.24}$$

Consequently, we can establish the following result which provides the complete Céa estimate.

Theorem 1.4.12. *Assume that the data satisfy*

$$L_f\left\{C_1+C_2\,c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}<\frac{1}{2}.\tag{1.4.25}$$

Then, there exist positive constants C_4 and C_5 independent of h, such that

$$\begin{aligned}&\|\phi-\phi_h\|_{1,\Omega}+\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H\\&\qquad\leq C_4\,\mathrm{dist}\,(\phi,\mathrm{H}_h^\phi)+C_5\,\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right).\end{aligned}\tag{1.4.26}$$

Proof. The estimate for $\|\phi-\phi_h\|_{1,\Omega}$ follows from (1.4.24) and (1.4.25), and the proof is complete after inserting the bound back into (1.4.14). □

Theorem 1.4.13. *In addition to the hypotheses of Theorems 1.3.9, 1.4.7 and 1.4.12, assume that there exists $s > 0$ such that $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega)$, $\mathbf{div}\,(\boldsymbol{\sigma}) \in \mathbf{H}^s(\Omega)$, $\boldsymbol{u} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{\rho} \in \mathbb{H}^s(\Omega)$ and $\phi \in \mathrm{H}^{1+s}(\Omega)$. Then, there exists $\widehat{C} > 0$, independent of h, such that, with the finite element subspaces defined by (1.4.7), (1.4.8), (1.4.9) and (1.4.10), there holds*

$$\begin{aligned} &\|\phi - \phi_h\|_{1,\Omega} + \|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})) - (\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H \\ &\qquad \leq \widehat{C} h^{\min\{s,k+1\}} \left\{ \|\boldsymbol{\sigma}\|_{s,\Omega} + \|\mathbf{div}\,\boldsymbol{\sigma}\|_{s,\Omega} + \|\boldsymbol{u}\|_{s,\Omega} + \|\boldsymbol{\rho}\|_{s,\Omega} + \|\phi\|_{1+s,\Omega} \right\}. \end{aligned} \tag{1.4.27}$$

Proof. It follows as a combination of the Céa estimate (1.4.26), and the approximation properties $(\mathbf{AP}^{\boldsymbol{\sigma}}_h)$, $(\mathbf{AP}^{\boldsymbol{u}}_h)$, $(\mathbf{AP}^{\boldsymbol{\rho}}_h)$ and $(\mathbf{AP}^{\phi}_h)$. □

1.5 An augmented mixed-primal formulation

In this section we follow the approach from previous works (see, e.g. [14, 79, 80] and the references therein) and put forward an augmented mixed-primal formulation for (1.2.7), which, as shown below, allows more freedom for choosing the finite element subspaces. We establish the augmented mixed-primal variational formulation of (1.2.1) and show that it is well-posed. Next, we define the corresponding Galerkin scheme, prove its solvability, introduce a specific mixed finite element method, and finally we establish the corresponding *a priori* error estimate.

1.5.1 The continuous setting

In order to increase flexibility in choosing discrete spaces for the approximation of the elasticity problem, we incorporate the following redundant terms in the variational formulation (1.3.6):

$$\begin{aligned} \kappa_1 \int_\Omega \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{C}^{-1}\boldsymbol{\sigma}\right) : \boldsymbol{\varepsilon}(\boldsymbol{v}) &= 0 && \forall\, \boldsymbol{v} \in \mathbf{H}^1(\Omega), \\ \kappa_2 \int_\Omega \mathbf{div}\,\boldsymbol{\sigma} \cdot \mathbf{div}\,\boldsymbol{\tau} = -\,\kappa_2 \int_\Omega \boldsymbol{f}(\phi) \cdot \mathbf{div}\,\boldsymbol{\tau} &&\quad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega), \\ \kappa_3 \int_\Omega (\boldsymbol{\rho} - (\nabla\,\boldsymbol{u} - \boldsymbol{\varepsilon}(\boldsymbol{u})) : \boldsymbol{\eta} = 0 &&\quad \forall\, \boldsymbol{\eta} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega), \\ \kappa_4 \int_\Gamma \boldsymbol{u} \cdot \boldsymbol{v} = \kappa_4 \int_\Gamma \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{v} &&\quad \forall\, \boldsymbol{v} \in \mathbf{H}^1(\Omega), \end{aligned} \tag{1.5.1}$$

where $(\kappa_1, \kappa_2, \kappa_3, \kappa_4)$ is a vector of positive parameters to be specified later on. It is important to observe here that the above terms now require that the displacement $\boldsymbol{u}$ live in $\mathbf{H}^1(\Omega)$.

Then, and alternatively to (1.3.6), we may consider the following augmented mixed formulation for the elasticity problem: find $(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) \in \mathbb{H}_0(\mathbf{div};\Omega) \times \mathbf{H}^1(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)$ such that

$$\widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) = \widetilde{F}_\phi(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \quad \forall (\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \in \mathbb{H}_0(\mathbf{div};\Omega) \times \mathbf{H}^1(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \tag{1.5.2}$$

where the multilinear form and the associated right hand side functional are defined as

$$\widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) := a(\boldsymbol{\sigma},\boldsymbol{\tau}) + b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})) - b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta})) + \kappa_1 \int_\Omega \left(\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathcal{C}^{-1}\boldsymbol{\sigma}\right) : \boldsymbol{\varepsilon}(\boldsymbol{v})$$

$$+ \kappa_2 \int_\Omega \mathbf{div}\,\boldsymbol{\sigma}\cdot\mathbf{div}\,\boldsymbol{\tau} + \kappa_3 \int_\Omega (\boldsymbol{\rho} - (\nabla\,\boldsymbol{u} - \boldsymbol{\varepsilon}(\boldsymbol{u})) : \boldsymbol{\eta} + \kappa_4 \int_\Gamma \boldsymbol{u}\cdot\boldsymbol{v}, \tag{1.5.3}$$

$$\widetilde{F}_\phi(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) := G(\boldsymbol{\tau}) - F_\phi(\boldsymbol{v},\boldsymbol{\eta}) - \kappa_2 \int_\Omega \boldsymbol{f}(\phi)\cdot\mathbf{div}\,\boldsymbol{\tau} + \kappa_4 \int_\Gamma \boldsymbol{u}_{\mathrm{D}}\cdot\boldsymbol{v}. \tag{1.5.4}$$

Hence, the augmented mixed-primal formulation for (1.2.7) reduces to (1.3.8) and (1.5.2), i.e.: find $(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho},\phi) \in \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega)\times\mathrm{H}^1_0(\Omega)$ such that

$$\begin{aligned} \widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) &= \widetilde{F}_\phi(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \quad \forall\,(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \in \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega), \\ A_{\boldsymbol{\sigma}}(\phi,\psi) &= G_{\boldsymbol{u}}(\psi) \qquad \forall\,\psi \in \mathrm{H}^1_0(\Omega). \end{aligned} \tag{1.5.5}$$

We proceed to adapt the approach from Sections 1.3.2 and 1.3.3. Since now $\boldsymbol{u}\in\mathbf{H}^1(\Omega)$, we can define

$$\mathbf{S} : \mathrm{H}^1_0(\Omega) \to \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega), \quad \mathbf{S}(\phi) := (\mathbf{S}_1(\phi),\mathbf{S}_2(\phi),\mathbf{S}_3(\phi)) := (\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),$$

where $(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})$ is the unique solution of (1.5.2) with a given $\phi\in\mathrm{H}^1_0(\Omega)$. In turn, we define the operator

$$\widetilde{\mathbf{S}} : \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega) \to \mathrm{H}^1_0(\Omega), \quad \widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u}) := \phi \qquad \forall\,(\boldsymbol{\sigma},\boldsymbol{u}) \in \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega),$$

where ϕ is the unique solution of (1.3.8) with the given $(\boldsymbol{\sigma},\boldsymbol{u})$. Next, the definition of $\mathbf{T}$ and the fixed-point strategy follow exactly as in Section 1.3.2. The analysis of $\widetilde{\mathbf{S}}$ can be therefore omitted.

The following lemma will be instrumental in showing the well-posedness of (1.5.2) for a given ϕ.

Lemma 1.5.1. *There exists $c_2>0$ such that*

$$\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{1,\Omega} + \|\boldsymbol{v}\|^2_{0,\Gamma} \geq c_2\,\|\boldsymbol{v}\|^2_{1,\Omega} \quad \forall\boldsymbol{v}\in\mathbf{H}^1(\Omega).$$

Proof. See [80, Lemma 3.1 and (3.9)]. □

Lemma 1.5.2. *Assume that $\kappa_1\in(0,4\delta\mu)$ and $\kappa_3\in\left(0,2c_2\kappa_1\widetilde{\delta}\left(1-\frac{\delta}{2}\right)\right)$ with $\delta,\widetilde{\delta}\in(0,2)$, and that $\kappa_2,\kappa_4>0$. Then, for each $\phi\in\mathrm{H}^1_0(\Omega)$, problem (1.5.2) has a unique solution $\mathbf{S}(\phi) := (\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})\in H := \mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega)$. Moreover, there exists $k_{\mathbf{S}}>0$, independent of ϕ, such that*

$$\|\mathbf{S}(\phi)\|_H = \|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})\|_H \leq k_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\} \quad \forall\phi\in\mathrm{H}^1_0(\Omega).$$

Proof. We first observe from (1.5.3) that B is a bilinear form. Next, applying Cauchy-Schwarz's inequality together with the trace theorem (with constant c_3), we can assert that

$$\begin{aligned} |\widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}))| &\leq \frac{1}{\mu}\|\boldsymbol{\sigma}\|_{0,\Omega}\|\boldsymbol{\tau}\|_{0,\Omega} + \|\boldsymbol{u}\|_{0,\Omega}\,\|\mathbf{div}\,\boldsymbol{\tau}\|_{0,\Omega} + \|\boldsymbol{\rho}\|_{0,\Omega}\,\|\boldsymbol{\tau}\|_{0,\Omega} + \|\boldsymbol{v}\|_{0,\Omega}\,\|\mathbf{div}\,\boldsymbol{\sigma}\|_{0,\Omega} \\ &+ \|\boldsymbol{\eta}\|_{0,\Omega}\,\|\boldsymbol{\sigma}\|_{0,\Omega} + \kappa_1\,\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{0,\Omega}\,\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{0,\Omega} + \frac{\kappa_1}{\mu}\,\|\boldsymbol{\sigma}\|_{0,\Omega}\,\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{0,\Omega} + \kappa_2\,\|\mathbf{div}\,\boldsymbol{\sigma}\|_{0,\Omega}\,\|\mathbf{div}\,\boldsymbol{\tau}\|_{0,\Omega} \\ &+ \kappa_3\,\|\boldsymbol{\rho}\|_{0,\Omega}\,\|\boldsymbol{\eta}\|_{0,\Omega} + \kappa_3|\boldsymbol{u}|_{1,\Omega}\,\|\boldsymbol{\eta}\|_{0,\Omega} + \kappa_3\,\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{0,\Omega}\,\|\boldsymbol{\eta}\|_{0,\Omega} + \kappa_4 c_3^2\,\|\boldsymbol{u}\|_{1,\Omega}\,\|\boldsymbol{v}\|_{1,\Omega}\,. \end{aligned}$$

It follows that there exists $\|\widetilde{B}\|>0$ depending on $\mu,\kappa_1,\kappa_2,\kappa_3,\kappa_4$ and c_3, such that

$$|\widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}))| \leq \|\widetilde{B}\|\,\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})\|_H\,\|(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})\|_H \quad \forall\,(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})\in H,$$

implying that $\widetilde{B}$ is bounded independently of $\phi \in \mathrm{H}^1_0(\Omega)$. The H-ellipticity analysis of $\widetilde{B}$ will be conducted as in the proof of [86, Thm. 3.1]. For each $(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \in H$, Young's inequality yields

$$\begin{aligned}\widetilde{B}((\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) &= \int_\Omega \mathcal{C}^{-1}\boldsymbol{\tau}:\boldsymbol{\tau} + \kappa_1 \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} - \kappa_1 \left\|\mathcal{C}^{-1}\boldsymbol{\tau}\right\|_{0,\Omega} \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{0,\Omega} + \kappa_2 \|\mathbf{div}\,\boldsymbol{\tau}\|^2_{0,\Omega} \\ &\quad + \kappa_3 \|\boldsymbol{\eta}\|^2_{0,\Omega} - \kappa_3 \|\nabla\boldsymbol{v} - \boldsymbol{\varepsilon}(\boldsymbol{v})\|_{0,\Omega} \|\boldsymbol{\eta}\|_{0,\Omega} + \kappa_4 \|\boldsymbol{v}\|^2_{0,\Gamma} \\ &= \int_\Omega \mathcal{C}^{-1}\boldsymbol{\tau}:\boldsymbol{\tau} - \frac{\kappa_1}{2\delta}\left\|\mathcal{C}^{-1}\boldsymbol{\tau}\right\|^2_{0,\Omega} + \kappa_1 \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} - \frac{\kappa_1\delta}{2}\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} + \kappa_2 \|\mathbf{div}\,\boldsymbol{\tau}\|^2_{0,\Omega} \\ &\quad + \kappa_3 \|\boldsymbol{\eta}\|^2_{0,\Omega} - \frac{\kappa_3}{2\widetilde{\delta}}\|\nabla\boldsymbol{v} - \boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} - \frac{\kappa_3\widetilde{\delta}}{2}\|\boldsymbol{\eta}\|^2_{0,\Omega} + \kappa_4 \|\boldsymbol{v}\|^2_{0,\Gamma},\end{aligned}$$

from which, taking $\delta, \widetilde{\delta}, \kappa_1, \kappa_2, \kappa_3, \kappa_4$ as stated in the hypotheses, applying Lemmas 1.3.1 and 1.5.1, and using the relation $\|\nabla\boldsymbol{v} - \boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} = |\boldsymbol{v}|^2_{1,\Omega} - \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega}$, we can deduce that

$$\begin{aligned}\widetilde{B}((\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) &\geq \frac{1}{2\mu}\left(1 - \frac{\kappa_1}{4\delta\mu}\right)\|\boldsymbol{\tau}^{\mathrm{d}}\|^2_{0,\Omega} + \kappa_2\|\mathbf{div}\,\boldsymbol{\tau}\|^2_{0,\Omega} + \kappa_1\left(1 - \frac{\delta}{2}\right)\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} \\ &\quad + \kappa_3\left(1 - \frac{\widetilde{\delta}}{2}\right)\|\boldsymbol{\eta}\|^2_{0,\Omega} - \frac{\kappa_3}{2\widetilde{\delta}}|\boldsymbol{v}|^2_{1,\Omega} + \kappa_4\|\boldsymbol{v}\|^2_{0,\Gamma} \\ &= \widetilde{\alpha}_2\|\boldsymbol{\tau}\|^2_{\mathbf{div},\Omega} + \left(c_2\widetilde{\alpha}_3 - \frac{\kappa_3}{2\widetilde{\delta}}\right)\|\boldsymbol{v}\|^2_{1,\Omega} + \kappa_3\left(1 - \frac{\widetilde{\delta}}{2}\right)\|\boldsymbol{\eta}\|^2_{0,\Omega},\end{aligned}$$

where $\widetilde{\alpha}_1 := \min\{\frac{1}{2\mu}\left(1 - \frac{\kappa_1}{4\delta\mu}\right), \frac{\kappa_2}{2}\}$, $\widetilde{\alpha}_2 := \min\{c_1\widetilde{\alpha}_1, \frac{\kappa_2}{2}\}$, and $\widetilde{\alpha}_3 := \min\{\kappa_1\left(1 - \frac{\delta}{2}\right), \kappa_4\}$. In this way, defining $\widetilde{\alpha} := \min\{\widetilde{\alpha}_2, c_2\widetilde{\alpha}_3 - \frac{\kappa_3}{2\widetilde{\delta}}, \kappa_3\left(1 - \frac{\widetilde{\delta}}{2}\right)\}$, which depends on $\mu, \delta, \widetilde{\delta}, \kappa_1, \kappa_2, \kappa_3, \kappa_4, c_1$ and c_2, we conclude that

$$\widetilde{B}((\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})) \geq \widetilde{\alpha}\,\|(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})\|^2_H \qquad \forall\,(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}) \in H. \tag{1.5.6}$$

Next, given $\phi \in \mathrm{H}^1_0(\Omega)$, we look at the functional $\widetilde{F}_\phi$, which is certainly linear. Similarly to the proof of [14, Lemma 3.4], there exists a positive constant $\|\widetilde{F}\|$ depending on κ_2, κ_4 and c_3, such that

$$|\widetilde{F}_\phi(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})| \leq \|\widetilde{F}\|\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\}\|(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})\|_H\,. \tag{1.5.7}$$

The foregoing inequality shows the boundedness of $\widetilde{F}_\phi$ with

$$\|\widetilde{F}_\phi\| \leq \|\widetilde{F}\|\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\}. \tag{1.5.8}$$

Finally, a straightforward application of the Lax-Milgram Lemma proves that for each $\phi \in \mathrm{H}^1_0(\Omega)$, problem (1.5.2) has a unique solution $\mathbf{S}(\phi) := (\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}) \in H$. Moreover, the corresponding continuous dependence result together with the estimates (1.5.6) and (1.5.7) give

$$\|\mathbf{S}(\phi)\|_H = \|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})\|_H \leq \frac{1}{\widetilde{\alpha}}\|\widetilde{F}_\phi\|_{H'} \leq k_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\},$$

with $k_{\mathbf{S}} := \dfrac{\|\widetilde{F}\|}{\widetilde{\alpha}}$, thus completing the proof. □

Lemma 1.5.3. *Let $\widetilde{\alpha}$ be the ellipticity constant provided in* Lemma 1.5.2. *Then, there exists $K_{\mathbf{S}} > 0$ depending on L_f, κ_2 and $\widetilde{\alpha}$ (cf. (1.2.4), (1.5.1), (1.5.6)), such that*

$$\|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_H \leq K_{\mathbf{S}}\,\|\phi - \varphi\|_{0,\Omega} \qquad \forall\,\phi,\varphi \in \mathrm{H}^1_0(\Omega). \tag{1.5.9}$$

Proof. We follow [14, Lemma 3.9], and fix $\phi, \varphi \in \mathrm{H}^1_0(\Omega)$. We then take $(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) = \mathbf{S}(\phi)$ and $(\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi}) = \mathbf{S}(\varphi)$, that is

$$\widetilde{B}((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}), (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta})) = \widetilde{F}_\phi(\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta}) \quad \text{and} \quad \widetilde{B}((\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi}), (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta})) = \widetilde{F}_\varphi(\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta}) \quad \forall (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta}) \in H.$$

Exploiting the ellipticity of $\widetilde{B}$ we readily get

$$\begin{aligned} \widetilde{\alpha} \, \|(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})\|^2_H &\leq \widetilde{B}((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}), (\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})) - \widetilde{B}((\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi}), (\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})) \\ &= (\widetilde{F}_\phi - \widetilde{F}_\varphi)((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})), \end{aligned} \tag{1.5.10}$$

and the definition of $\widetilde{F}_\phi$ in combination with Cauchy-Schwarz's inequality and (1.2.4) implies that

$$\begin{aligned} &|(\widetilde{F}_\phi - \widetilde{F}_\varphi)((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi}))| \\ &\quad = \left| \int_\Omega (\boldsymbol{f}(\phi) - \boldsymbol{f}(\varphi)) \cdot (\boldsymbol{u} - \boldsymbol{w}) - \kappa_2 \int_\Omega (\boldsymbol{f}(\phi) - \boldsymbol{f}(\varphi)) \cdot \mathbf{div}(\boldsymbol{\sigma} - \boldsymbol{\zeta}) \right| \\ &\quad \leq L_f (1 + \kappa_2^2)^{1/2} \, \|\phi - \varphi\|_{0,\Omega} \, \|(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})\|_H \end{aligned} \tag{1.5.11}$$

Back substitution of (1.5.11) into (1.5.10) then yields

$$\widetilde{\alpha} \, \|(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})\|^2_H \leq L_f (1 + \kappa_2^2)^{1/2} \, \|\phi - \varphi\|_{0,\Omega} \, \|(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) - (\boldsymbol{\zeta}, \boldsymbol{w}, \boldsymbol{\chi})\|_H \,,$$

which finally gives (1.5.9). □

Lemma 1.5.4. *Let W be the closed ball defined in* Lemma 1.3.5 *and $K_{\mathbf{S}}$ be as in* Lemma 1.5.3. *Then, for each $\phi, \varphi \in \mathrm{H}^1_0(\Omega)$, there holds*

$$\|\mathbf{T}(\phi) - \mathbf{T}(\varphi)\|_{1,\Omega} \leq \frac{1}{\alpha_2} K_{\mathbf{S}} \left(L_g + L_\vartheta \, \|\mathbf{T}(\varphi)\|_{1,\infty,\Omega} \right) \|\phi - \varphi\|_{0,\Omega} \,.$$

Proof. The definition of $\mathbf{T}$ together with Lemma 1.3.3 imply that $\mathbf{T}(W) \subseteq W$. The remainder of the proof proceeds exactly as the one of Lemma 1.3.7. □

Theorem 1.5.5. *The mixed-primal problem* (1.3.11) *has at least one solution $(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}, \phi) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{H}^1(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega) \times \mathrm{H}^1_0(\Omega)$, satisfying*

$$\|\phi\|_{1,\Omega} \leq r \quad \text{and} \quad \|(\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho})\|_H \leq k_{\mathbf{S}} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}.$$

Moreover, if the data satisfy

$$\frac{1}{\alpha_2} K_{\mathbf{S}} \left\{ L_g + L_\vartheta C_\infty \, k_{\mathbf{S}} \left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right) \right\} < 1,$$

then the solution ϕ is unique in W.

Proof. It follows as in the proof of Theorem 1.3.9. □

1.5.2 The discrete scheme

We begin by observing that, thanks to the ellipticity of the bilinear forms $\widetilde{B}$ and $A_{\boldsymbol{\sigma}}$, we can consider arbitrary finite dimensional-subspaces

$$\mathbb{H}_h^{\boldsymbol{\sigma}} \subseteq \mathbb{H}_0(\mathbf{div};\Omega), \quad \mathbf{H}_h^{\boldsymbol{u}} \subseteq \mathbf{H}^1(\Omega), \quad \mathbb{H}_h^{\boldsymbol{\rho}} \subseteq \mathbb{L}^2_{\mathrm{skew}}(\Omega) \quad \text{and} \quad \mathrm{H}_h^{\phi} \subseteq \mathrm{H}_0^1(\Omega),$$

for the augmented mixed-primal formulation. In particular, given an integer $k \geq 0$, we can define

$$\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{c}^{\mathrm{t}}\boldsymbol{\tau}_h|_K \in \mathbf{RT}_k(K) \ \ \forall\, \boldsymbol{c} \in \mathrm{R}^n, \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbf{H}_h^{\boldsymbol{u}} &:= \left\{ \boldsymbol{v}_h \in \mathbf{C}(\Omega) \quad \boldsymbol{v}_h|_K \in \mathbf{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbb{H}_h^{\boldsymbol{\rho}} &:= \left\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathrm{H}_h^{\phi} &:= \left\{ \psi_h \in \mathrm{C}(\Omega) \cap \mathrm{H}_0^1(\Omega) \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}.
\end{aligned} \tag{1.5.12}$$

Then, a Galerkin scheme for (1.5.5) reads: find $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h, \phi_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}} \times \mathrm{H}_h^{\phi}$ such that

$$\widetilde{B}((\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h), (\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\eta}_h)) = \widetilde{F}_{\phi_h}(\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\eta}_h) \quad \forall\, (\boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbb{H}_h^{\sigma} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}, \tag{1.5.13}$$

$$A_{\boldsymbol{\sigma}_h}(\phi_h, \psi_h) = G_{\boldsymbol{u}_h}(\psi_h) \qquad \forall\, \psi_h \in \mathrm{H}_h^{\phi}. \tag{1.5.14}$$

We can now proceed analogously to Section 1.5.1 and define a fixed-point scheme for the analysis of the coupled problem (1.5.13)-(1.5.14). For this purpose, we define $\mathbf{S}_h : \mathrm{H}_h^{\phi} \to \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$ as

$$\mathbf{S}_h(\phi_h) := (\mathbf{S}_{1,h}(\phi_h), \mathbf{S}_{2,h}(\phi_h), \mathbf{S}_{3,h}(\phi_h)) := (\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h) \quad \forall \phi_h \in \mathrm{H}_h^{\phi},$$

where the triple $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h)$ is the unique solution of (1.5.13), with $\widetilde{B}$ and $\widetilde{F}_{\phi_h}$ defined by (1.5.3) and (1.5.4), respectively, with $\phi = \phi_h$. In turn, the operators $\widetilde{\mathbf{S}}_h$ and $\mathbf{T}_h$ are defined as in Section 1.4.2.

As the analysis of the operator $\widetilde{\mathbf{S}}_h$ follows verbatim from Section 1.4.2, we can omit the details here. Concerning $\mathbf{S}_h$, we start by investigating the well-posedness of (1.5.13).

Lemma 1.5.6. *Assume that* $\kappa_1 \in (0, 4\delta\mu)$ *and* $\kappa_3 \in \left(0, 2c_2\kappa_1\widetilde{\delta}\left(1 - \frac{\delta}{2}\right)\right)$ *with* $\delta, \widetilde{\delta} \in (0,2)$, *and that* $\kappa_2, \kappa_4 > 0$. *Then, for each* $\phi_h \in \mathrm{H}_h^{\phi}$ *the problem* (1.5.13) *has a unique solution* $\mathbf{S}(\phi_h) := (\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$. *Moreover, with the same constant* $k_{\mathbf{S}} > 0$ *provided by* Lemma 1.5.2, *there holds*

$$\|\mathbf{S}_h(\phi_h)\|_H = \|(\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h)\|_H \leq k_{\mathbf{S}} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2} \right\} \quad \forall \phi_h \in \mathrm{H}_h^{\phi}.$$

Proof. It suffices to note that for each $\phi_h \in \mathrm{H}_h^{\phi}$, the multilinear form $\widetilde{B}$ is elliptic on $\mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}$ with the same constant $\widetilde{\alpha}$ from Lemma 1.5.2 and that $\|\widetilde{F}_{\phi_h}\|_{(\mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}})'}$ is bounded as in (1.5.8) with ϕ_h in place of ϕ. Hence, the result follows from a direct application of the Lax-Milgram Lemma. □

We now provide the discrete analogues of Lemmas 1.5.3, 1.5.4 and Theorem 1.5.5, whose proofs, which are almost verbatim of the corresponding continuous ones, are omitted.

Lemma 1.5.7. *Let* $K_{\mathbf{S}}$ *be the constant provided by* Lemma 1.5.3. *Then, there holds*

$$\|\mathbf{S}_h(\phi_h) - \mathbf{S}_h(\varphi_h)\|_H \leq K_{\mathbf{S}} \|\phi_h - \varphi_h\|_{0,\Omega} \qquad \forall\, \phi_h, \varphi_h \in \mathbb{H}_h^{\phi}.$$

Lemma 1.5.8. *Let W_h be as in* Lemma 1.4.3. *Then*

$$\|\mathbf{T}_h(\phi_h)-\mathbf{T}_h(\varphi_h)\|_{1,\Omega}\leq\frac{1}{\alpha_2}K_{\mathbf{S}}\left(L_g+L_\vartheta\,\|\nabla\mathbf{T}_h(\varphi_h)\|_{\infty,\Omega}\right)\|\phi_h-\varphi_h\|_{0,\Omega}\quad\forall\,\phi_h,\varphi_h\in\mathrm{H}_h^\phi.$$

Theorem 1.5.9. *Let W_h be as in* Lemma 1.4.3. *Then, the Galerkin scheme* (1.5.13) – (1.5.14) *has at least one solution $(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h,\phi_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\times\mathrm{H}_h^\phi$, and there holds*

$$\|\phi_h\|_{1,\Omega}\leq r\quad\text{and}\quad\|(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H\leq k_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right\}.$$

1.5.3 A priori error analysis

The goal of this section is to derive an estimate for $\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H$, where $(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})$ and $(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)$ are the solutions to the problems

$$\begin{array}{rcll}\widetilde{B}((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta}))&=&\widetilde{F}_\phi(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})&\forall\,(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega),\\ \widetilde{B}((\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h),(\boldsymbol{\tau}_h,\boldsymbol{v}_h,\boldsymbol{\eta}_h))&=&\widetilde{F}_{\phi_h}(\boldsymbol{\tau}_h,\boldsymbol{v}_h,\boldsymbol{\eta}_h)&\forall\,(\boldsymbol{\tau}_h,\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}},\end{array}\tag{1.5.15}$$

respectively. For this purpose, we recall (again from [127]) a Strang-type lemma, which will be applied to (1.5.15).

Lemma 1.5.10. *Let H be a Hilbert space, $F\in H'$ and $\boldsymbol{a}:H\times H\to\mathrm{R}$ be a bounded and elliptic bilinear form. In addition, let $\{H_h\}_{h>0}$ be a sequence of finite dimensional subspaces of H and for each $h>0$ consider a bounded bilinear form $\boldsymbol{a}_h:H_h\times H_h\to\mathrm{R}$ and a functional $F_h\in H_h'$. Assume that the family $\{\boldsymbol{a}_h\}_{h>0}$ is uniformly elliptic, that is, there exists a constant $\alpha>0$, independent of h, such that*

$$\boldsymbol{a}_h(v_h,v_h)\geq\alpha\,\|v_h\|_H^2\quad\forall\,v_h\in H_h,\quad\forall\,h>0.$$

In turn, let $u\in H$ and $u_h\in H_h$ such that

$$\boldsymbol{a}(u,v)=F(v)\quad\forall\,v\in H\quad\text{and}\quad\boldsymbol{a}_h(u_h,v_h)=F_h(v_h)\quad\forall\,v_h\in H_h.$$

Then, for each $h>0$, there holds

$$\|u-u_h\|_H\leq\widetilde{C}_{\mathrm{ST}}\left\{\sup_{\substack{w_h\in H_h\\ w_h\neq0}}\frac{|F(w_h)-F_h(w_h)|}{\|w_h\|_H}+\inf_{\substack{v_h\in H_h\\ v_h\neq0}}\left(\|u-v_h\|_V+\sup_{\substack{w_h\in H_h\\ w_h\neq0}}\frac{|\boldsymbol{a}(v_h,w_h)-\boldsymbol{a}_h(v_h,w_h)|}{\|w_h\|_H}\right)\right\}.$$

where $\widetilde{C}_{\mathrm{ST}}:=\alpha^{-1}\max\{1,\|\boldsymbol{a}\|\}$.

Proof. See [127, Thm. 11.1]. □

As in Sections 1.4.4 and 1.4.5, we now set

$$\mathrm{dist}\left((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right):=\inf_{(\boldsymbol{\tau}_h,\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}}\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\tau}_h,\boldsymbol{v}_h,\boldsymbol{\eta}_h)\|_H\,,$$

or, equivalently

$$\mathrm{dist}\left((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right) := \mathrm{dist}(\boldsymbol{\sigma},\mathbb{H}_h^{\boldsymbol{\sigma}}) + \mathrm{dist}(\boldsymbol{u},\mathbf{H}_h^{\boldsymbol{u}}) + \mathrm{dist}\left(\boldsymbol{\rho},\mathbb{H}_h^{\boldsymbol{\rho}}\right),$$

where, having in mind that now $\mathbf{H}_h^{\boldsymbol{u}}\subseteq\mathbf{H}^1(\Omega)$, we set $\mathrm{dist}(\boldsymbol{u},\mathbf{H}_h^{\boldsymbol{u}}) := \inf\limits_{\boldsymbol{v}_h\in\mathbf{H}_h^{\boldsymbol{u}}}\|\boldsymbol{u}-\boldsymbol{v}_h\|_{1,\Omega}$. The other two distances are exactly as defined in Section 1.4.4.

Lemma 1.5.11. *Let $\widetilde{C}_{\mathrm{ST}} := \widetilde{\alpha}^{-1}\max\{1,\|\widetilde{B}\|\}$, where $\widetilde{\alpha}$ is the constant yielding the ellipticity of $\widetilde{B}$ (cf. (1.5.6)). Then, there holds*

$$\begin{aligned}\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H&\\ \leq \widetilde{C}_{\mathrm{ST}}\left\{\mathrm{dist}\left((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right)+L_f(1+\kappa_2^2)^{1/2}\|\phi-\phi_h\|_{0,\Omega}\right\}.&\end{aligned}\tag{1.5.16}$$

Proof. We note that the bilinear form $\widetilde{B}$ and the functionals $\widetilde{F}_\phi$ and $\widetilde{F}_{\phi_h}$ satisfy the hypotheses of Lemma 1.5.10. Then, a straightforward application of Lemma 1.5.10 to the context (1.5.15) gives

$$\begin{aligned}\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H&\\ \leq \widetilde{C}_{\mathrm{ST}}\left\{\|(\widetilde{F}_\phi-\widetilde{F}_{\phi_h})|_{\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}}\|+\mathrm{dist}\left((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right)\right\}.&\end{aligned}\tag{1.5.17}$$

Next, similarly as in the proof of Lemma 1.5.3, we deduce that

$$\|(\widetilde{F}_\phi-\widetilde{F}_{\phi_h})|_{\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}}\|\leq L_f(1+\kappa_2^2)^{1/2}\|\phi-\phi_h\|_{0,\Omega}\,.\tag{1.5.18}$$

Finally, by replacing (1.5.18) back into (1.5.17), we get (1.5.16) and the lemma follows. □

At this point, we realise that in the present context the estimate for $\|\phi-\phi_h\|_{1,\Omega}$ stays exactly as in (1.4.17). Consequently, the corresponding Céa estimate for the total error

$$\|\phi-\phi_h\|_{1,\Omega}+\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_H$$

is derived by combining (1.4.17) and (1.5.16). By virtue of the aforementioned, we can establish the analogues of Theorems 1.4.12 and 1.4.13, whose proofs are omitted.

Theorem 1.5.12. *Let C_1 and C_2 be the constants defined in (1.4.23), and assume that the data satisfy*

$$L_f(1+\kappa_2^2)^{1/2}\left\{C_1+C_2\,k_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}<\frac{1}{2}\,.$$

Then, there exist positive constants C_6 and C_7, independent of h, such that

$$\begin{aligned}\|\phi-\phi_h\|_{1,\Omega}+\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H&\\ \leq C_6\,\mathrm{dist}\,(\phi,\mathrm{H}_h^\phi)+C_7\,\mathrm{dist}\left((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho}),\mathbb{H}_h^{\boldsymbol{\sigma}}\times\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}}\right).&\end{aligned}$$

Theorem 1.5.13. *In addition to the hypotheses of Theorems 1.5.5, 1.5.9 and 1.5.12, assume that there exists $s>0$ such that $\boldsymbol{\sigma}\in\mathbb{H}^s(\Omega)$, $\mathbf{div}\,(\boldsymbol{\sigma})\in\mathbf{H}^s(\Omega)$, $\boldsymbol{u}\in\mathbf{H}^{1+s}(\Omega)$, $\boldsymbol{\rho}\in\mathbb{H}^s(\Omega)$ and $\phi\in\mathrm{H}^{1+s}(\Omega)$. Then, there exists $\widehat{C}>0$, independent of h, such that, with the finite element subspaces defined by (1.5.12), there holds*

$$\begin{aligned}\|\phi-\phi_h\|_{1,\Omega}+\|(\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\rho})-(\boldsymbol{\sigma}_h,\boldsymbol{u}_h,\boldsymbol{\rho}_h)\|_H&\\ \leq \widehat{C}h^{\min\{s,k+1\}}\left\{\|\boldsymbol{\sigma}\|_{s,\Omega}+\|\mathbf{div}\,\boldsymbol{\sigma}\|_{s,\Omega}+\|\boldsymbol{u}\|_{1+s,\Omega}+\|\boldsymbol{\rho}\|_{s,\Omega}+\|\phi\|_{1+s,\Omega}\right\}.&\end{aligned}\tag{1.5.19}$$

N	h	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e(ϕ)	$r(\phi)$	iter
Mixed-primal PEERS-Lagrange scheme with $k=0$										
129	0.7071	124.43	–	1.72e-2	–	5.49e-2	–	0.1125	–	4
457	0.3536	65.778	0.91	9.11e-3	0.92	2.87e-2	0.94	6.72e-2	0.74	5
1713	0.1768	33.305	0.98	4.61e-3	0.98	1.45e-2	0.98	3.81e-2	0.82	6
6625	0.0883	16.703	0.99	2.32e-3	0.99	7.26e-3	0.99	1.87e-2	1.02	6
26049	0.0441	8.3584	0.99	1.15e-3	0.99	3.63e-3	0.99	8.35e-3	1.16	6
103297	0.0221	4.1802	0.99	5.78e-4	0.99	1.81e-3	0.99	3.91e-3	1.09	6
Augmented scheme with $k=0$										
67	0.7071	132.53	–	0.1043	–	0.1120	–	0.1105	–	5
219	0.3536	70.733	0.91	0.0643	0.69	0.1036	0.61	0.0708	0.64	5
787	0.1768	35.492	0.99	0.0323	0.99	0.0789	0.93	0.0427	0.82	6
2979	0.0883	17.604	1.01	0.0157	1.04	0.0463	0.97	0.0230	0.99	6
11587	0.0441	8.7683	1.00	0.0077	1.01	0.0242	0.93	0.0108	1.08	6
45699	0.0221	4.3792	1.00	3.86e-3	1.00	0.0129	0.98	4.62e-3	1.23	6
Augmented scheme with $k=1$										
195	0.7071	38.856	–	0.0309	–	0.0169	–	0.0358	–	6
691	0.3536	10.373	1.90	0.0088	1.81	0.0074	1.49	0.0100	1.83	6
2595	0.1768	2.6473	1.97	0.0023	1.93	0.0029	1.53	0.0024	2.01	6
10051	0.0883	0.6637	1.99	0.0005	1.97	0.0009	1.67	0.0006	2.03	6
39555	0.0441	0.1658	2.00	0.0001	1.99	0.0002	1.86	0.0001	2.02	8
156931	0.0221	0.0414	2.00	3.72e-5	1.99	6.65e-5	1.94	3.68e-5	2.03	6

Table 1.1: Example 1: Degrees of freedom, meshsizes, errors, rates of convergence, and number of Picard iterations for the mixed-primal PEERS-P_1 and augmented $\mathbf{RT}_k - \mathbf{P}_{k+1} - \mathbb{P}_k - \mathrm{P}_{k+1}$ approximations of the coupled problem with $k = 0, 1$, and using $\nu = 0.4$ and $\kappa_2 = 0.5\mu, \kappa_4 = \mu$. In the first block of the table, the displacement error is measured in the $\mathbf{L}^2-$norm (table produced by the author).

1.6 Numerical results

In this section we provide a set of computational tests. The first one serves to illustrate the convergence rates anticipated by our previous analysis for the mixed-primal and the augmented Galerkin schemes, whereas the remaining examples address a few cases not covered by our analysis (mixed boundary conditions, non-convex domains, and the 3D case).

Example 1: Error history for a constructed solution in 2D. We consider (1.2.7) in the unit square $\Omega = (0,1)^2$ and propose exact solutions and coupling terms (tensorial diffusivity, body load,

and diffusive source) as follows

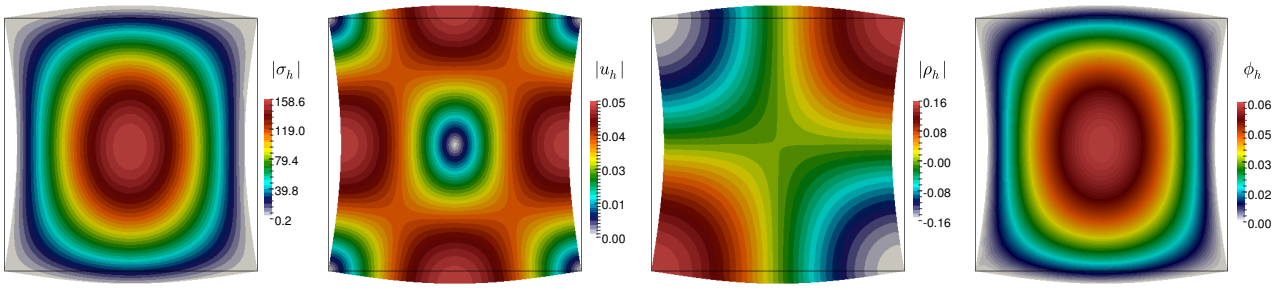
$$\boldsymbol{u} = \begin{pmatrix} d_1 \sin(\pi x_1)\cos(\pi x_2) + \frac{x_1^2}{2\lambda} \\ -d_1 \cos(\pi x_1)\sin(\pi x_2) + \frac{x_2^2}{2\lambda} \end{pmatrix}, \quad \boldsymbol{\sigma} = \lambda\,\mathrm{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}), \quad \boldsymbol{\rho} = \nabla u - \boldsymbol{\varepsilon}(\boldsymbol{u}),$$
$$\phi = x_1\,(1-x_2)\,x_2\,(1-x_2),\ \vartheta(\boldsymbol{\sigma}) = D_0\,\mathbb{I} + D_2\boldsymbol{\sigma}^2,\ \boldsymbol{f}(\phi) = d_2\begin{pmatrix}\cos^2(\phi)\\ -\sin(\phi)\end{pmatrix},\ g(\boldsymbol{u}) = d_2\left(1+\frac{1}{1+|\boldsymbol{u}|}\right). \tag{1.6.1}$$

These closed-form solutions satisfy the boundary conditions $\boldsymbol{u}_{\mathrm{D}} = \boldsymbol{u}$ on Γ and $\phi = 0$ on Γ. Moreover, the elasticity and diffusion equations are considered non-homogeneous and the extra source terms are chosen according to (1.6.1). This treatment does not compromise the continuous and discrete analyses, as the smoothness of the exact solution provides right-hand sides with terms in $\mathrm{L}^2(\Omega)$, thus only requiring a slight modification of the functionals in the variational formulation. We note that the forcing and source terms satisfy (1.2.4)-(1.2.5). Additionally, we pick out the following value to the model parameters: displacement and forcing term scalings $d_1 = 0.05$, $d_2 = 0.1$; Young's modulus $E = 1e3$; Poisson's ratio $\nu = 0.4$; the constants specifying ϑ given by $D_0 = 1.0$ and $D_2 = 0.1$, and the Lamé constants $\lambda = E\nu(1+\nu)^{-1}(1-2\nu)^{-1}$ and $\mu = E/(2+2\nu)$. We consider a heuristic value for Korn's constant (*cf.* Lemma 1.5.1) as $c_2 = 0.1$; and using the proof of Lemma 1.5.2, the stabilisation parameters assume the values $\delta = \widetilde{\delta} = 1$, $\kappa_1 = 2\mu$, $\kappa_2 = 0.5\mu$, $\kappa_3 = 0.1\mu$, and $\kappa_4 = \mu$. We generate a sequence of uniformly refined meshes and proceed to define errors and convergence rates as usual:

$$\mathrm{e}(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div},\Omega}\,,\ \mathrm{e}(\boldsymbol{u}) = \|\boldsymbol{u} - \boldsymbol{u}_h\|_{j,\Omega}\,,\ \mathrm{e}(\boldsymbol{\rho}) = \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\Omega}\,,\ \mathrm{e}(\phi) = \|\phi - \phi_h\|_{1,\Omega}\,,\ r(\cdot) = \frac{\log(\mathrm{e}(\cdot)/\widehat{\mathrm{e}}(\cdot))}{\log(h/\widehat{h})},$$

where e and $\widehat{\mathrm{e}}$ denote errors computed on two consecutive meshes of sizes h and $\widehat{h}$; and where $j = 0, 1$ will be used to measure the displacement error for the mixed-primal and augmented mixed-primal schemes, respectively.

On each refinement level we generate approximate solutions with the lowest-order PEERS-Lagrange elements indicated in Section 1.4.4, and also with the $\mathbf{RT}_k - \mathbf{P}_{k+1} - \mathbb{P}_k - \mathrm{P}_{k+1}$ scheme specified in Section 1.5.2, for $k = 0, 1$. The output of this error study is collected in Table 1.1 (where we tabulate errors, experimental convergence rates, and iteration count). We observe an asymptotic $O(h^{k+1})$ convergence for all individual errors (stress, displacement, rotation, and concentration), which agrees with the theoretical error bounds derived in Section 1.4.5 (*cf.* (1.4.27)) and Section 1.5.3 (*cf.* (1.5.19)). Around six Picard iterations are necessary to reach the prescribed tolerance Tol=1e-6 imposed on the $\ell^\infty-$norm of the total residual. At each fixed-point step the resulting linear systems were solved with the direct method SuperLU. For completeness, we also depict in Figure 1.1 the obtained numerical solutions computed with the lowest-order augmented method. We also mention that the proposed methods maintain their accuracy in the incompressibility limit. This is confirmed by replicating the same experimental analysis, now considering $\nu = 0.49999$. The error history for this case is displayed in Table 1.2, where we observe that the magnitude of errors and convergence rates are comparable to those in Table 1.1. However, if the stabilisation parameters are kept as in the first case, then the number of Picard iterations needed to achieve the prescribed tolerance for the augmented schemes is considerably higher. Similar iteration counts as those in the non-augmented case can be obtained with much smaller values of κ_2 and κ_4: here we choose $\kappa_2 = \kappa_4 = 0.001\mu$.



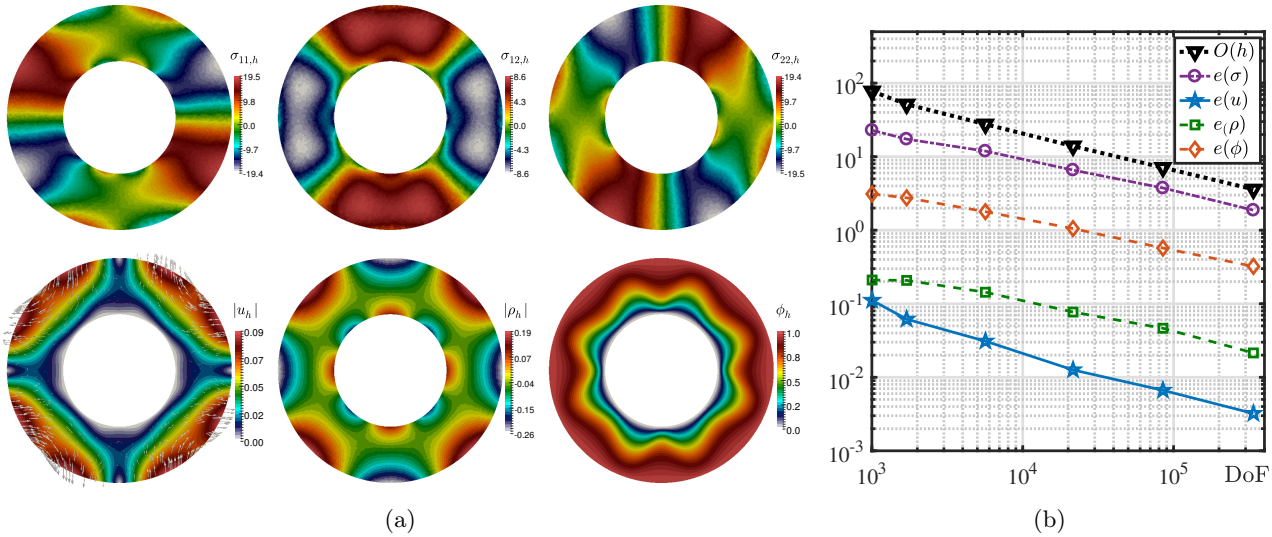
(a) (b) (c) (d)

Figure 1.1: Example 1: $\mathbf{RT}_0 - \mathbf{P}_1 - \mathbb{P}_0 - \mathrm{P}_1$ approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), relevant component of the rotation tensor $\boldsymbol{\rho}_h$ (c), and concentration of the diffusive substance ϕ_h (d); using $\nu = 0.4$. All fields are plotted on the deformed domain (figure produced by the author).

N	h	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e(ϕ)	$r(\phi)$	iter
Mixed-primal PEERS-Lagrange scheme with $k = 0$										
129	0.7071	9189.7	–	6.05e-2	–	0.1477	–	1.9940	–	6
457	0.3536	605.93	2.98	9.14e-3	2.72	2.88e-2	2.35	9.59e-2	4.37	6
1713	0.1768	30.604	4.30	4.61e-3	0.98	1.45e-2	0.99	3.67e-2	1.30	6
6625	0.0883	15.390	0.99	2.31e-3	0.99	7.26e-3	0.99	1.89e-2	0.96	6
26049	0.0441	7.7948	0.98	1.15e-3	0.99	3.63e-3	0.99	8.41e-3	1.17	6
103297	0.0221	3.9011	0.99	5.78e-4	0.99	1.81e-3	0.99	3.91e-3	1.10	6
Augmented scheme with $k = 0$										
67	0.7071	5525.5	–	1.6922	–	7.7691	–	0.1523	–	4
219	0.3536	853.17	5.02	0.1672	3.41	0.9461	4.14	8.05e-2	0.72	5
787	0.1768	33.563	4.62	7.50e-2	1.29	0.3937	1.25	3.75e-2	1.04	6
2979	0.0883	16.784	0.99	3.39e-2	1.04	0.1467	1.16	1.97e-2	0.92	6
11587	0.0441	8.2505	1.02	1.95e-2	0.93	7.43e-2	0.94	1.03e-2	0.94	6
45699	0.0221	4.0961	1.01	9.73e-3	0.99	3.73e-2	0.98	4.54e-3	1.18	6
Augmented scheme with $k = 1$										
195	0.7071	172.52	–	1.2010	–	1.4012	–	7.34e-2	–	10
691	0.3536	9.4288	3.94	2.33e-2	5.68	2.28e-2	5.93	1.84e-2	1.44	6
2595	0.1768	1.8711	2.59	2.36e-3	3.30	2.86e-3	2.99	4.19e-3	2.04	6
10051	0.0883	0.8375	2.14	5.90e-4	1.99	9.05e-4	1.69	7.26e-4	1.90	6
39555	0.0441	0.1559	2.24	1.48e-4	1.99	2.52e-4	1.84	1.49e-4	2.12	6
156931	0.0221	3.91e-2	1.99	3.72e-5	1.99	6.65e-5	1.92	3.79e-5	1.97	6

Table 1.2: Example 1: Error history produced using a higher Poisson ratio $\nu = 0.49999$ and setting $\kappa_2 = \kappa_4 = 0.001\mu$. In the first block of the table, the displacement error is measured in the $\mathbf{L}^2-$norm (table produced by the author).

Example 2: Convergence in a non-convex domain. The goal of this example is to observe



(a) (b)

Figure 1.2: Example 2: Approximate solutions (stress components, displacement magnitude with directions, rotation, and concentration) using a lowest order PEERS-Lagrange scheme displayed on the undeformed domain (a); and individual errors computed with respect to a reference solution (b) (figure produced by the author).

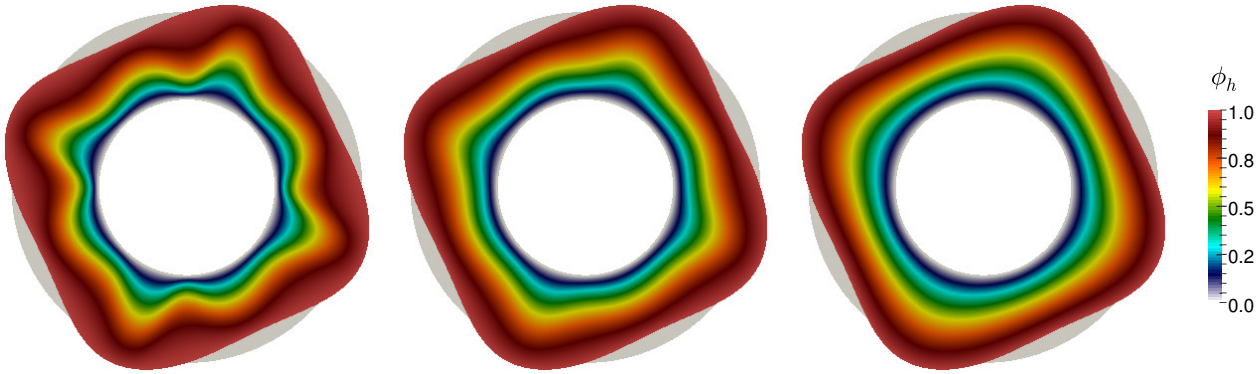
the behaviour of the numerical method producing solutions on a non-convex domain (we recall that convexity was required in the analysis of the fixed-point operators defining the coupled continuous problem). To this end we consider a ring-shaped membrane bounded by an outer circle of radius 1 and an inner circle of radius 0.5. Initial guesses for stress, displacement, and concentration are zero. Differently from Example 1, we now apply the following tensorial diffusivity, body load, source of species, and prescribed boundary displacement on the outer ring

$$\vartheta(\boldsymbol{\sigma}) = D_0\,\mathbb{I} + D_1\boldsymbol{\sigma} + D_2\boldsymbol{\sigma}^2,\ \boldsymbol{f}(\phi) = d_2\begin{pmatrix}\phi\\ \phi(1-\phi)\end{pmatrix},\ g(\boldsymbol{u}) = d_3|\boldsymbol{u}|,\ \boldsymbol{u}_{\mathrm{D}} = \begin{pmatrix} d_1\sin(\pi x_1)\cos(\pi x_2)\\ -d_1\cos(\pi x_1)\sin(\pi x_2)\end{pmatrix},$$

whereas on the inner ring the structure is clamped. We impose a concentration of 1 on the outer ring and zero on the inner boundary. The coefficients defining the problem assume the values $D_0 = d_1 = 0.1$, $D_1 = D_2 = 0.05$, $d_2 = 0.025$, $d_3 = -1$, $E = 100$ and $\nu = 0.33$, and the numerical solutions generated with the lowest-order PEERS-Lagrange scheme are presented in Figure 1.2(a).

In view of assessing the convergence of the lowest-order primal-mixed method, and in the absence of a closed-form expression for the solution of this problem, we consider a reference solution computed in a highly refined mesh (of around 50K elements) and proceed to compute approximate solutions on coarser meshes. The obtained errors (with respect to the reference solutions projected to each coarse mesh) and convergence rates are shown in Figure 1.2(b), where one sees that all fields exhibit an $O(h)$ accuracy, and note that the stress error is dominant. For all refinement levels the fixed-point algorithm took less than five iterations to converge.

We exploit the same setting to study the influence of different values for the additional diffusion parameters $D_1 = D_2$ (representing scenarios where the stress-assisted diffusion decreases in intensity). Figure 1.3 compares three different cases, where a substantial difference is observed in the generated



(a) $D_1 = D_2 = 0.05$ (b) $D_1 = D_2 = 0.001$ (c) $D_1 = D_2 = 0.0001$

Figure 1.3: Example 2: Concentration profiles of the diffusive substance ϕ_h plotted on the deformed domain, for different values of the additional diffusivity constants (figure produced by the author).

diffusion patterns. A similar effect as the one produced with very low values of D_1 and D_2 (the profiles in Figure 1.3(c) show a very smooth diffusion going uniformly from $\phi = 1$ on the outer circle, to $\phi = 0$ on the inner boundary) can be achieved by softening the material, prescribing a Young modulus of $E = 1$.

Example 3: Stress-assisted diffusion and experimental convergence on a 3D slab. In much the same way as in Example 2, here we will confirm that the other assumption in Theorem 1.3.4 (the restriction to two spatial dimensions) can be obviated at the implementation stage, and that it does not compromise the behaviour of the proposed methods. Focusing on an applicative test, let us regard a porous block occupying the domain $\Omega = (0, 250) \times (0, 250) \times (0, 50)$ and construct an unstructured tetrahedral mesh of 55K elements. The stress-dependent diffusivity is considered as in Example 2: $\vartheta(\boldsymbol{\sigma}) = D_0\,\mathbb{I} + D_1\boldsymbol{\sigma} + D_2\boldsymbol{\sigma}^2$, the concentration-dependent body load is $\boldsymbol{f}(\phi) = d_2(\phi, \phi, \phi(1-\phi))^{\mathrm{t}}$, and the displacement-dependent source is now $g(\boldsymbol{u}) = d_3 \operatorname{div} \boldsymbol{u}$. We will take the parameter values $D_0 = 0.5$, $D_1 = 0.025$, $D_2 = -0.015$, $d_2 = 0.1$, $d_3 = 0.25$, $E = 1e4$, and $\nu = 0.49$. Boundary conditions for the elasticity problem differ from the ones analysed in this chapter: The block is clamped on the surface $x_1 = 0$, a normal traction force is imposed on the surface $x_1 = 250$, $\boldsymbol{\sigma\nu} = 3/4\mu\,\boldsymbol{\nu}$, and zero normal stresses are considered elsewhere on the boundary, $\boldsymbol{\sigma\nu} = \mathbf{0}$. On the surface $x_1 = 0$ we fix the concentration $\phi = x_2(250 - x_2)x_3(50 - x_3)/(25 \cdot 125)^2$, we impose zero-flux boundary conditions on the face $x_1 = 250$, $\tilde{\boldsymbol{\sigma}} \cdot \boldsymbol{\nu} = 0$; and consider an homogeneous Dirichlet boundary condition for concentration on the remainder of $\partial\Omega$. Once again we consider the augmented mixed-primal method of lowest order, for which the penalisation constants adopt the values $\kappa_1 = 2\mu$, $\kappa_2 = 0.5\mu$, $\kappa_3 = 0.01\mu$, and $\kappa_4 = 1$. The linear systems encountered at each Picard step are solved with the GMRES method preconditioned with an incomplete LU factorisation. The computational results are summarised in Figure 1.4, indicating that stresses are concentrated on the corners of the boundaries where Dirichlet conditions are set for displacements, and rotations are higher in the vicinities of the rectangles at $x_1 = 0$ and $x_1 = 250$. For this case the Picard method takes eight iterations to converge.

We also assess the accuracy of the method through an experimental error analysis. Since, for this particular problem configuration, a closed form solution to (1.2.1) is not available, we produce an

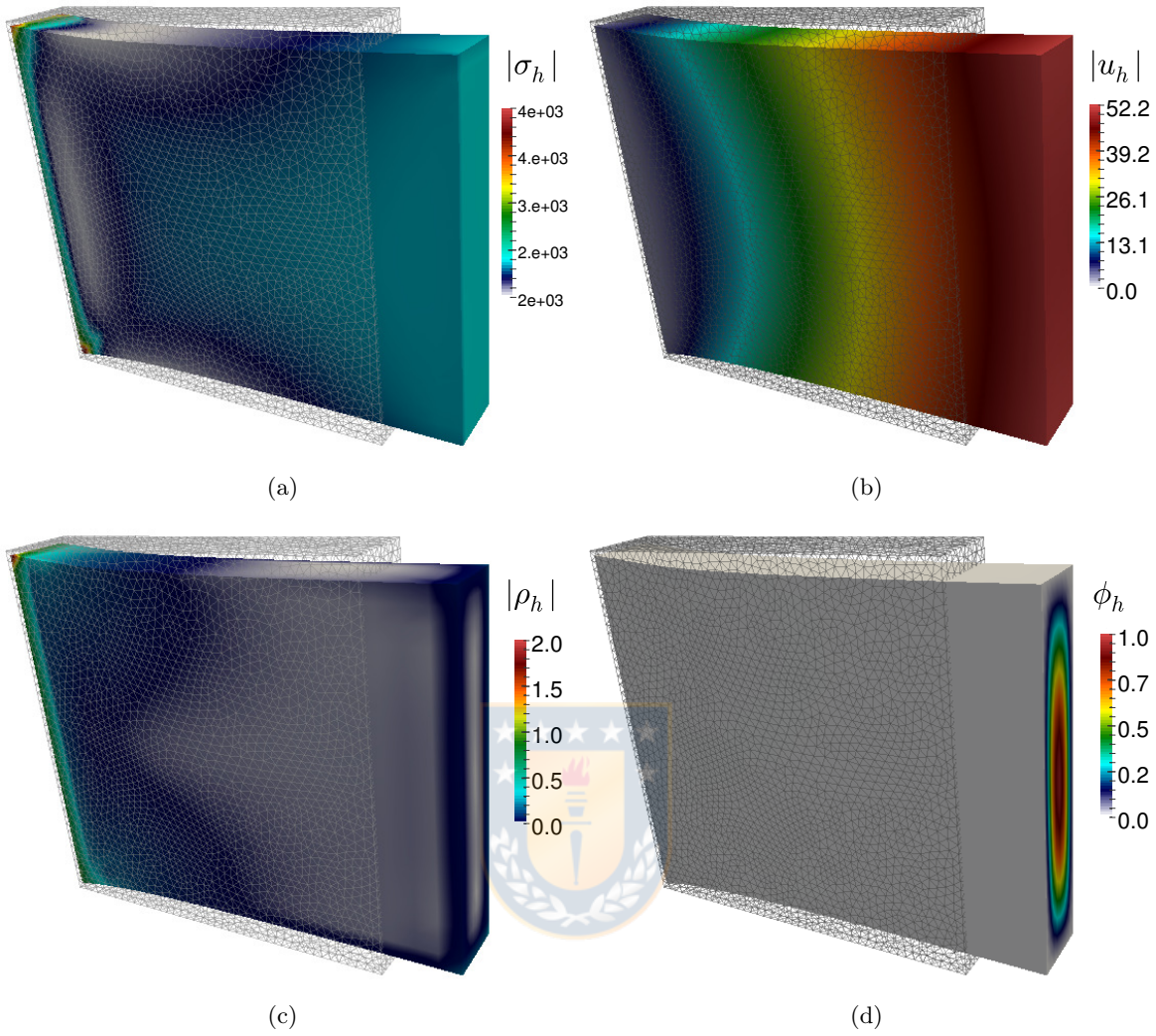


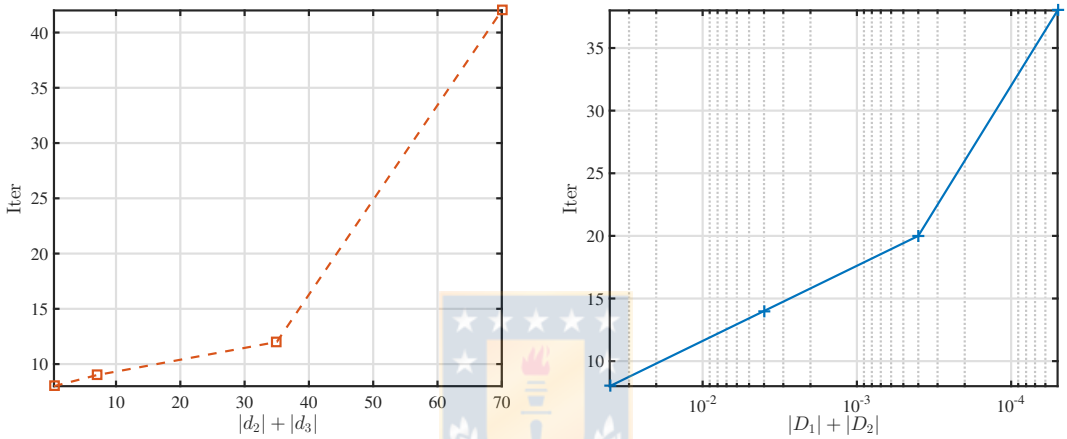
Figure 1.4: Example 3: Augmented mixed-primal approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), rotation tensor magnitude $|\boldsymbol{\rho}_h|$ (c), and concentration of the diffusive substance ϕ_h (d); all plotted on the deformed domain and showing the undeformed, skeleton mesh (figure produced by the author).

approximate solution using a highly refined mesh (of 290K elements) and consider it as a reference solution for error computation. We also generate a sequence of much coarser quasi-uniformly refined meshes (but not necessarily nested) on which we compute approximate solutions. The result of this error analysis is collected in Table 1.3. The observed convergence rates, here only presented for the lowest-order augmented scheme, approach the optimal values as the number of degrees of freedom increases. In addition, the fixed-point iteration count remains near the base case (of eight steps) for all levels of mesh refinement.

Next we investigate the effect of the stress-diffusion coupling (which is actually encoded in the magnitude of the parameters D_1, D_2 and d_2, d_3) on the performance of the fixed-point iteration count. We conduct six rounds of simulations, first fixing the tensorial diffusivity constants D_1, D_2 and increasing d_2, d_3; and then fixing d_2, d_3 and decreasing D_1, D_2 (large contributions from stresses will only

N	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e(ϕ)	$r(\phi)$	iter
487	1968.2	–	20.384	–	1.0379	–	0.2571	–	7
2837	639.81	0.64	6.9320	0.76	0.2825	0.93	0.0842	0.84	8
22156	204.12	0.79	2.1943	0.80	0.0873	0.92	0.0293	0.83	7
150109	70.451	0.95	0.7904	0.92	0.0315	0.79	0.0096	0.90	8
907803	25.298	0.94	0.2246	0.93	0.0102	0.89	0.0034	0.92	8

Table 1.3: Example 3: Experimental error history against a reference (fine mesh) solution, and number of Picard iterations per refinement level. Lowest-order augmented method (table produced by the author).



(a) with $D_1 = 0.025, D_2 = -0.015$ (b) with $d_2 = 0.1, d_3 = 0.25$

Figure 1.5: Example 3: Iteration count produced when varying the coupling parameters defining the concentration-dependent body load and displacement-dependent source (a), and the stress-assisted diffusivity parameters (b) (figure produced by the author).

increase diffusion, therefore making the generalised Poisson problem more stable). Figure 1.5 presents the response of the method in terms of number of fixed-point iterations needed to reach the tolerance Tol=1e-6. We observe that as the coupling terms depart from the base case, the solver performs a larger number of steps.

CHAPTER 2

Formulation and analysis of fully-mixed methods for stress-assisted diffusion problems

2.1 Introduction

We are interested in the mathematical and numerical study of a stationary problem representing diffusion-deformation processes where the stress acts as a coupling variable. So-called stress-assisted diffusion models (derived from thermodynamic principles and phenomenological arguments in e.g. [125, 6]) are relevant to numerous applications including diffusion of boron and arsenic in silicon [117], hydrogen diffusion in metals [141], voiding of aluminium conductor lines in integrated circuits [162], strain-aging measurements in iron [118], sorption in polymers [131], to name a few. Of special appeal to us is the study of microscopic electrode damage in lithium ion batteries [19, 50, 94, 136, 110]. When lithium diffuses into a secondary particle (an anode made of e.g. silicon), its core expands and its elastic response, also with that of neighboring particles and the surrounding electrolyte, modify the diffusive properties inside the medium. If the process is confined inside the anode, then the electric field is practically constant and the system may be described solely in terms of diffusion and stress.

Regarding the mathematical and numerical analysis of related models, the literature is rather scarce. Some recent references include homogenisation of concentration - electric potential systems [138], multi-scale analysis of the deterioration of binder in electrodes [78], and a general local-global well-posedness theory [109]. Differently from these approaches, in [83] we have recently proposed a mixed-primal formulation for stress-assisted diffusion. The model covers the linear elastic regime, it incorporates the rotation tensor as supplementary variable serving to impose stress symmetry in a weak manner; and this mixed problem is coupled with a primal formulation for diffusion. Here, in contrast, we consider an augmented mixed formulation for the diffusion equation. Similarly to [87], the concentration gradient and the diffusive flux are incorporated as auxiliary unknowns, which allows us to treat the stress-dependent diffusivity using a dual-mixed setting. In order to apply the regularity estimates from [83], we augment the formulation with redundant terms arising from a constitutive equation. Next, following the approach introduced in [14], we combine fixed-point arguments, regularity estimates, the Babuška-Brezzi theory, the Lax Milgram lemma, the Sobolev embedding and Rellich-Kondrachov theorems, and small data assumptions to establish existence and uniqueness of solution of the continuous problem. The solvability of the Galerkin scheme follows from the Brouwer fixed-point theorem and

properties of the finite element subspaces. Finally, the convergence analysis is conducted adapting Strang inequalities, Céa estimates, and using approximation properties of the finite element spaces.

The rest of the chapter is organised as follows. In Section 2.2 we describe required notation and functional spaces to be employed along the chapter. Then, we introduce the model problem and requirements on the specific constitutive functions. Next, in Section 2.3 we derive the augmented fully-mixed formulation and establish its well-posedness. The Galerkin scheme and the existence of discrete solution are then studied in Section 2.4. In addition, under similar assumptions we deduce error bounds in Section 2.5; and we close in Section 2.6 with a numerical example that confirms the theoretical rates of convergence, and a second test studying the applicability of the discrete formulation in the simulation of 3D microscopic lithiation processes.

2.2 The model problem

Let us consider the following system of PDEs, governing the diffusion of a solute interacting with the motion of an elastic solid occupying the domain Ω:

$$\boldsymbol{\sigma} = \lambda\,\mathrm{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbf{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{in} \quad \Omega\,, \qquad -\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}(\phi) \quad \text{in} \quad \Omega, \tag{2.2.1}$$

$$\widetilde{\boldsymbol{\sigma}} = \vartheta(\boldsymbol{\sigma})\nabla\phi \quad \text{in} \quad \Omega\,, \qquad -\,\mathrm{div}\,\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}) \quad \text{in} \quad \Omega, \tag{2.2.2}$$

$$\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} \quad \text{on} \quad \Gamma, \qquad \phi = \phi_{\mathrm{D}} \quad \text{on} \quad \Gamma. \tag{2.2.3}$$

Equations (2.2.1) state the constitutive relation and momentum balance for the elasticity equations, problem (2.2.2) defines the diffusion equation and diffusive flux, and (2.2.3) specifies the Dirichlet boundary conditions $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{1/2}(\Gamma)$ and $\phi_{\mathrm{D}} \in \mathrm{H}^{1/2}(\Gamma)$. The involved quantities and model parameters are the Cauchy solid stress $\boldsymbol{\sigma}$; the displacement field $\boldsymbol{u}$, the infinitesimal strain tensor $\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\mathrm{t}}\right)$; the Lamé constants $\lambda, \mu > 0$ characterizing the material; the diffusive flux $\widetilde{\boldsymbol{\sigma}}$; the solute concentration ϕ; the tensorial diffusivity $\vartheta : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$; the vector of body loads $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^n$, and a displacement-dependent source term $g : \mathbb{R}^n \to \mathbb{R}$. For the load, source, and diffusivity functions we will require uniform boundedness and Lipschitz continuity, that is there exist positive constants f_1, f_2, L_f, g_1, g_2, L_g, and $\vartheta_1, \vartheta_2, L_\vartheta$, such that

$$f_1 \leq |\boldsymbol{f}(s)| \leq f_2, \quad |\boldsymbol{f}(s) - \boldsymbol{f}(t)| \leq L_f|s-t| \quad \forall\, s,t \in \mathbb{R}, \tag{2.2.4}$$

$$g_1 \leq |g(\boldsymbol{w})| \leq g_2, \quad |g(\boldsymbol{v}) - g(\boldsymbol{w})| \leq L_g|\boldsymbol{v}-\boldsymbol{w}| \quad \forall\, \boldsymbol{v},\boldsymbol{w} \in \mathbb{R}^n, \tag{2.2.5}$$

$$\vartheta_1 \leq |\vartheta(\boldsymbol{\tau})| \leq \vartheta_2, \quad |\vartheta(\boldsymbol{\tau}) - \vartheta(\boldsymbol{\zeta})| \leq L_\vartheta|\boldsymbol{\tau}-\boldsymbol{\zeta}| \quad \forall\, \boldsymbol{\tau},\boldsymbol{\zeta} \in \mathbb{R}^{n\times n}. \tag{2.2.6}$$

Additionally, ϑ is of class C^1 and uniformly positive definite, the latter meaning that there exists $\vartheta_0 > 0$ such that

$$\vartheta(\boldsymbol{\tau})\boldsymbol{w}\cdot\boldsymbol{w} \geq \vartheta_0|\boldsymbol{w}|^2 \quad \forall\, \boldsymbol{w} \in \mathbb{R}^n, \quad \forall\, \boldsymbol{\tau} \in \mathbb{R}^{n\times n}\,. \tag{2.2.7}$$

Finally, we assume that $\boldsymbol{f}(\phi) \in \mathbf{H}^1(\Omega)$ for each $\phi \in \mathrm{H}^1(\Omega)$, and that for each $\gamma \in (0,1)$ there exists a constant $C_\gamma > 0$ such that $g(\boldsymbol{w}) \in \mathrm{H}^\gamma(\Omega)$ for each $\boldsymbol{w} \in \mathrm{H}^\gamma(\Omega)$ and

$$\|g(\boldsymbol{w})\|_{\gamma,\Omega} \leq C_\gamma\,\|\boldsymbol{w}\|_{\gamma,\Omega}\,. \tag{2.2.8}$$

2.3 Weak formulation and solvability analysis

In this section we derive an augmented fully-mixed variational formulation for (2.2.1)-(2.2.3) and propose a fixed-point strategy for its analysis. We show that the fixed-point operator is well-defined and apply the Schauder's theorem to prove existence of solution, whereas Banach fixed-point theorem will lead to uniqueness of solution under small data assumptions.

2.3.1 The mixed-mixed formulation

We begin by recalling from [39] that $\mathbf{H}(\mathrm{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega) \oplus \mathbb{R}\mathbf{I}$, with

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

which means that for each $\boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega)$ there exist unique

$$\boldsymbol{\tau}_0 := \boldsymbol{\tau} - \left\{ \frac{1}{n|\Omega|} \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) \right\} \mathbf{I} \in \mathbb{H}_0(\mathbf{div};\Omega) \quad \text{and} \quad d := \frac{1}{n|\Omega|} \int_\Omega \mathrm{tr}(\boldsymbol{\tau}) \in \mathbb{R},$$

such that $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\mathbf{I}$. Also, we define the space of skew-symmetric tensors as

$$\mathbb{L}^2_{\mathrm{skew}}(\Omega) := \{ \boldsymbol{\eta} \in \mathbb{L}^2(\Omega) : \ \boldsymbol{\eta} + \boldsymbol{\eta}^{\mathrm{t}} = 0 \}.$$

Then, proceeding as in [83, Section 2.1], we apply the Dirichlet boundary condition for displacements (first relation of (2.2.3)) and the aforementioned orthogonal decomposition to write the elasticity problem in weak form: find $(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) \in \mathbf{H}_1 := \mathbb{H}_0(\mathbf{div};\Omega) \times \left(\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega) \right)$ such that

$$\begin{aligned} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) &= G(\boldsymbol{\tau}) \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega), \\ b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) &= F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) \quad \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \end{aligned} \tag{2.3.1}$$

where $a : \mathbb{H}_0(\mathbf{div};\Omega) \times \mathbb{H}_0(\mathbf{div};\Omega) \to \mathbb{R}$ and $b : \mathbb{H}_0(\mathbf{div};\Omega) \times (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)) \to \mathbb{R}$ are bilinear forms defined as

$$a(\boldsymbol{\zeta}, \boldsymbol{\tau}) = \frac{1}{2\mu} \int_\Omega \boldsymbol{\zeta}^{\mathrm{d}} : \boldsymbol{\tau}^{\mathrm{d}} + \frac{1}{n(n\,\lambda + 2\mu)} \int_\Omega \mathrm{tr}(\boldsymbol{\zeta})\, \mathrm{tr}(\boldsymbol{\tau}), \quad b(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta})) := \int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\tau} + \int_\Omega \boldsymbol{\eta} : \boldsymbol{\tau},$$

for $\boldsymbol{\zeta}, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega)$ and $(\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)$. In turn, the functionals $F_\phi \in (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega))'$ and $G \in \mathbb{H}_0(\mathbf{div};\Omega)'$ are given by

$$G(\boldsymbol{\tau}) := \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma \qquad \text{and} \qquad F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) := -\int_\Omega \boldsymbol{f}\,(\phi) \cdot \boldsymbol{v}\,, \tag{2.3.2}$$

for $(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta})) \in \mathbf{H}_1$, where $\langle \cdot, \cdot \rangle_\Gamma$ stands for the duality pairing of $\mathbf{H}^{-1/2}(\Gamma)$ and $\mathbf{H}^{1/2}(\Gamma)$. Details on the derivation of the weak formulation (2.3.1) can be found in [81] as well as in [24].

In turn, defining the concentration gradient $\boldsymbol{t} := \nabla\phi$, we can recast the diffusion equation as

$$\begin{aligned} \widetilde{\boldsymbol{\sigma}} \ = \ \vartheta(\boldsymbol{\sigma})\,\boldsymbol{t} \quad \text{in} \quad \Omega, \quad \boldsymbol{t} = \nabla\phi \quad \text{in} \quad \Omega, \quad -\mathrm{div}\,\widetilde{\boldsymbol{\sigma}} \ = \ g(\boldsymbol{u}) \quad \text{in} \quad \Omega, \\ \phi = \phi_{\mathrm{D}} \quad \text{on} \quad \Gamma. \end{aligned} \tag{2.3.3}$$

We then test the three-field problem (2.3.3) against $\boldsymbol{s} \in \mathbf{L}^2(\Omega)$, $\widetilde{\boldsymbol{\tau}} \in \mathbf{H}(\mathrm{div};\Omega)$ and $\psi \in \mathrm{L}^2(\Omega)$. Integrating by parts the expression $\int_\Omega \nabla\phi\cdot\widetilde{\boldsymbol{\tau}}$ and using the Dirichlet boundary condition for ϕ (second equation in (2.2.3)), we arrive at the weak formulation: find $(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi) \in \mathbf{L}^2(\Omega)\times\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{L}^2(\Omega)$ such that

$$\begin{aligned}
\int_\Omega \vartheta(\boldsymbol{\sigma})\,\boldsymbol{t}\cdot\boldsymbol{s} \;-\; \int_\Omega \widetilde{\boldsymbol{\sigma}}\cdot\boldsymbol{s} &= 0 && \forall\,\boldsymbol{s}\in\mathbf{L}^2(\Omega),\\
\int_\Omega \widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{t} \;+\; \int_\Omega \phi\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} &= \langle\widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{\nu},\phi_{\mathrm{D}}\rangle_\Gamma && \forall\,\widetilde{\boldsymbol{\tau}}\in\mathbf{H}(\mathrm{div};\Omega),\\
-\int_\Omega \psi\,\mathrm{div}\,\widetilde{\boldsymbol{\sigma}} &= \int_\Omega \psi\,g(\boldsymbol{u}) && \forall\,\psi\in\mathrm{L}^2(\Omega).
\end{aligned} \tag{2.3.4}$$

In view of modifying the regularity properties of the coupled problem, we proceed to enrich the foregoing equations with the following residual terms:

$$\begin{aligned}
\kappa_1\int_\Omega \{\widetilde{\boldsymbol{\sigma}}-\vartheta(\boldsymbol{\sigma})\,\boldsymbol{t}\}\cdot\widetilde{\boldsymbol{\tau}} = 0 && \forall\,\widetilde{\boldsymbol{\tau}}\in\mathbf{H}(\mathrm{div};\Omega),\\
\kappa_2\int_\Omega \mathrm{div}\,\widetilde{\boldsymbol{\sigma}}\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} = -\,\kappa_2\int_\Omega g(\boldsymbol{u})\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} && \forall\,\widetilde{\boldsymbol{\tau}}\in\mathbf{H}(\mathrm{div};\Omega),\\
\kappa_3\int_\Omega \{\nabla\phi-\boldsymbol{t}\}\cdot\nabla\psi = 0 && \forall\,\psi\in\mathrm{H}^1(\Omega),\\
\kappa_4\int_\Gamma \phi\,\psi = \kappa_4\int_\Gamma \phi_{\mathrm{D}}\,\psi && \forall\,\psi\in\mathrm{H}^1(\Omega),
\end{aligned} \tag{2.3.5}$$

where κ_1,κ_2, κ_3 and κ_4 are positive parameters to be specified later on. We remark that the identities required in (2.3.5) are nothing but the constitutive and the equilibrium equations concerning $\widetilde{\boldsymbol{\sigma}}$, along with the relation defining $\boldsymbol{t}$, and the Dirichlet boundary condition for ϕ; all of them tested differently from (2.3.4). Instead of (2.3.4), we will now focus on the following augmented formulation for the diffusion problem: find $(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\in\mathbf{H}_2 := \mathbf{L}^2(\Omega)\times\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$ such that

$$A_{\boldsymbol{\sigma}}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)) = G_{\boldsymbol{u}}(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\quad\forall\,(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\in\mathbf{H}_2, \tag{2.3.6}$$

where

$$\begin{aligned}
A_{\boldsymbol{\sigma}}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)) &:= \int_\Omega \vartheta(\boldsymbol{\sigma})\,\boldsymbol{t}\cdot\boldsymbol{s} - \int_\Omega \widetilde{\boldsymbol{\sigma}}\cdot\boldsymbol{s} + \int_\Omega \widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{t} + \int_\Omega \phi\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} - \int_\Omega \psi\,\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}\\
&+\;\kappa_1\int_\Omega \{\widetilde{\boldsymbol{\sigma}}-\vartheta(\boldsymbol{\sigma})\,\boldsymbol{t}\}\cdot\widetilde{\boldsymbol{\tau}} + \kappa_2\int_\Omega \mathrm{div}\,\widetilde{\boldsymbol{\sigma}}\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} + \kappa_3\int_\Omega \{\nabla\phi-\boldsymbol{t}\}\cdot\nabla\psi + \kappa_4\int_\Gamma \phi\,\psi,
\end{aligned} \tag{2.3.7}$$

and

$$G_{\boldsymbol{u}}(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi) := \langle\widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{\nu},\phi_{\mathrm{D}}\rangle_\Gamma + \int_\Omega \psi\,g(\boldsymbol{u}) - \kappa_2\int_\Omega g(\boldsymbol{u})\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}} + \kappa_4\int_\Gamma \phi_{\mathrm{D}}\,\psi. \tag{2.3.8}$$

Consequently, we arrive at the following augmented fully-mixed formulation for (2.2.1)-(2.2.3): find $\big((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\big)\in\mathbf{H}_1\times\mathbf{H}_2$, such that

$$\begin{aligned}
a(\boldsymbol{\sigma},\boldsymbol{\tau}) \;+\; b(\boldsymbol{\tau},(\boldsymbol{u},\boldsymbol{\rho})) &= G(\boldsymbol{\tau}) && \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\
b(\boldsymbol{\sigma},(\boldsymbol{v},\boldsymbol{\eta})) &= F_\phi(\boldsymbol{v},\boldsymbol{\eta}) && \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega),\\
A_{\boldsymbol{\sigma}}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)) &= G_{\boldsymbol{u}}(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi) && \forall\,(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\in\mathbf{H}_2.
\end{aligned} \tag{2.3.9}$$

2.3.2 A fixed-point approach

Here we utilise a fixed-point strategy to prove that problem (2.3.9) is well-posed. Let us first define the operator $\mathbf{S} : \mathrm{H}^1(\Omega) \to \mathbf{H}_1$ as

$$\mathbf{S}(\phi) := (\mathbf{S}_1(\phi), (\mathbf{S}_2(\phi), \mathbf{S}_3(\phi))) := (\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) \quad \forall \phi \in \mathrm{H}^1(\Omega),$$

where $(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho}))$ is the unique solution of (2.3.1) with the given ϕ. In turn, we define the operator $\widetilde{\mathbf{S}} : \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{L}^2(\Omega) \to \mathbf{H}_2$ as

$$\widetilde{\mathbf{S}}(\boldsymbol{\sigma}, \boldsymbol{u}) := (\widetilde{\mathbf{S}}_1(\boldsymbol{\sigma}, \boldsymbol{u}), \widetilde{\mathbf{S}}_2(\boldsymbol{\sigma}, \boldsymbol{u}), \widetilde{\mathbf{S}}_3(\boldsymbol{\sigma}, \boldsymbol{u})) := (\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) \quad \forall\, (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{L}^2(\Omega),$$

where $(\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi)$ is the unique solution of (2.3.6) with the given $(\boldsymbol{\sigma}, \boldsymbol{u})$. In this way, by introducing the operator $\mathbf{T} : \mathrm{H}^1(\Omega) \to \mathrm{H}^1(\Omega)$ as

$$\mathbf{T}(\phi) := \widetilde{\mathbf{S}}_3(\mathbf{S}_1(\phi), \mathbf{S}_2(\phi)) \quad \forall\, \phi \in \mathrm{H}^1(\Omega),$$

we realize that (2.3.9) can be rewritten as the fixed-point problem: find $\phi \in \mathrm{H}^1(\Omega)$ such that

$$\mathbf{T}(\phi) = \phi. \tag{2.3.10}$$

However, we remark in advance that the definition of $\mathbf{T}$ will be only in a closed ball of $\mathrm{H}^1(\Omega)$.

We also collect the following two technical lemmas, whose proofs can be found in [81, Lemma 2.3], and [77, Lemma 3.3], respectively.

Lemma 2.3.1. *There exists $c_1 > 0$ such that*

$$c_1 \, \|\boldsymbol{\tau}\|^2_{0,\Omega} \leq \|\boldsymbol{\tau}^{\mathrm{d}}\|^2_{0,\Omega} + \|\mathbf{div}\, \boldsymbol{\tau}\|^2_{0,\Omega} \quad \forall \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega).$$

Lemma 2.3.2. *There exists $c_2 > 0$ such that*

$$|\psi|^2_{1,\Omega} + \|\psi\|^2_{0,\Gamma} \geq c_2 \, \|\psi\|^2_{1,\Omega} \quad \forall\, \psi \in \mathrm{H}^1(\Omega).$$

In what follows we show that $\mathbf{T}$ has at least one fixed point. Firstly we will prove that the uncoupled problems defined by $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ are well-posed, where we emphasize that $\mathbf{S}$ is defined similarly as in [83], and therefore we omit parts of the proofs whenever necessary. Our analysis will focus on the uncoupled problem (2.3.6) and its repercussion on $\mathbf{T}$. Let us start by recalling the continuity of a and b. For a proof we refer to e.g. [81].

$$\begin{aligned} |a(\boldsymbol{\zeta}, \boldsymbol{\tau})| &\leq \frac{1}{\mu} \, \|\boldsymbol{\zeta}\|_{\mathbf{div};\Omega} \, \|\boldsymbol{\tau}\|_{\mathbf{div};\Omega} \quad \forall\, \boldsymbol{\zeta}, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \\ |b(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta}))| &\leq \|\boldsymbol{\tau}\|_{\mathbf{div};\Omega} \, \|(\boldsymbol{v}, \boldsymbol{\eta})\| \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \quad \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega). \end{aligned} \tag{2.3.11}$$

Furthermore, it is not difficult to see that a is strongly elliptic in the kernel of b. In fact, we denote the operator induced by the bilinear form b as $\mathbf{B}$, and note that

$$V := \mathrm{Ker}(\mathbf{B}) = \left\{ \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega) : \ \mathbf{div}\, \boldsymbol{\tau} = 0 \ \text{ in } \ \Omega, \quad \boldsymbol{\tau} = \boldsymbol{\tau}^{\mathrm{t}} \ \text{ in } \ \Omega \right\},$$

from which, we deduce that

$$a(\boldsymbol{\tau},\boldsymbol{\tau}) \geq \frac{1}{2\mu}\|\boldsymbol{\tau}^{\mathrm{d}}\|_{0,\Omega}^2 \geq \frac{c_1}{2\mu}\,\|\boldsymbol{\tau}\|_{0,\Omega}^2 = \alpha\,\|\boldsymbol{\tau}\|_{\mathbf{div};\Omega}^2 \quad \forall\,\boldsymbol{\tau}\in V, \tag{2.3.12}$$

where c_1 is the constant provided by Lemma 2.3.1. Additionally, as a slight modification of the proof of [81, Section 2.4.3], we find that $\mathbf{B}$ is surjective. Finally, we observe that G and F_ϕ are bounded with

$$\|G\| \leq \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \quad \text{and} \quad \|F_\phi\| \leq f_2|\Omega|^{1/2}. \tag{2.3.13}$$

This analysis confirms the well-posedness of (2.3.1), which is abridged in the following lemma.

Lemma 2.3.3. *For each $\phi\in \mathrm{H}^1(\Omega)$ the problem (2.3.1) has a unique solution $\mathbf{S}(\phi) := (\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\in\mathbf{H}_1$. Moreover, there exists $c_{\mathbf{S}}>0$, independent of ϕ, such that*

$$\|\mathbf{S}(\phi)\|_{\mathbf{H}_1} = \|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_{\mathbf{H}_1} \leq c_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\}. \tag{2.3.14}$$

Proof. It follows from estimates (2.3.11)–(2.3.13) and a direct application of the Babuška-Brezzi theory (see, e.g. [39] and [81, Thm. 2.3]). We refer to [83, Lemmma 2.2] for further details. □

In turn, we prove the well-posedness of problem (2.3.6) with the next result.

Lemma 2.3.4. *Assume that $\kappa_1\in\left(0,\frac{2\delta\vartheta_0}{\vartheta_2}\right)$ and $\kappa_3\in\left(0,2\widetilde{\delta}\left(\vartheta_0-\frac{\kappa_1\vartheta_2}{2\delta}\right)\right)$ with $\delta\in\left(0,\frac{2}{\vartheta_2}\right)$, $\widetilde{\delta}\in(0,2)$, and $\kappa_2,\kappa_4>0$. Then, for each $(\boldsymbol{\sigma},\boldsymbol{u})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega)$, problem (2.3.6) has a unique solution $\widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u}) = (\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\in\mathbf{H}_2$. Moreover, there exists $\widetilde{c}_{\mathbf{S}}>0$, independent of $(\boldsymbol{\sigma},\boldsymbol{u})$, such that*

$$\|\widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u})\|_{\mathbf{H}_2} = \|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\|_{\mathbf{H}_2} \leq \widetilde{c}_{\mathbf{S}}\left\{\|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2|\Omega|^{1/2}\right\}. \tag{2.3.15}$$

Proof. Firstly, we note from (2.3.7) that $A_{\boldsymbol{\sigma}}$ is a bilinear form. Next, applying Cauchy-Schwarz's inequality, the upper bound for ϑ (*cf.* (2.2.6)), and the trace theorem (with constant c_0), we find that

$$\begin{aligned}
|A_{\boldsymbol{\sigma}}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi))| &\leq \vartheta_2\|\boldsymbol{t}\|_{0,\Omega}\|\boldsymbol{s}\|_{0,\Omega} + \|\widetilde{\boldsymbol{\sigma}}\|_{0,\Omega}\,\|\boldsymbol{s}\|_{0,\Omega} + \|\widetilde{\boldsymbol{\tau}}\|_{0,\Omega}\,\|\boldsymbol{t}\|_{0,\Omega}\\
&\quad + \|\phi\|_{0,\Omega}\,\|\mathbf{div}\,\widetilde{\boldsymbol{\tau}}\|_{0,\Omega} + \|\psi\|_{0,\Omega}\,\|\mathbf{div}\,\widetilde{\boldsymbol{\sigma}}\|_{0,\Omega} + \kappa_1\|\widetilde{\boldsymbol{\sigma}}\|_{0,\Omega}\,\|\widetilde{\boldsymbol{\tau}}\|_{0,\Omega} + \kappa_1\vartheta_2\,\|\boldsymbol{t}\|_{0,\Omega}\,\|\widetilde{\boldsymbol{\tau}}\|_{0,\Omega}\\
&\quad + \kappa_2\|\mathbf{div}\,\widetilde{\boldsymbol{\sigma}}\|_{0,\Omega}\,\|\mathbf{div}\,\widetilde{\boldsymbol{\tau}}\|_{0,\Omega} + \kappa_3|\phi|_{1,\Omega}|\psi|_{1,\Omega} + \kappa_3\,\|\boldsymbol{t}\|_{0,\Omega}\,|\psi|_{1,\Omega} + c_0^2\,\kappa_4\,\|\phi\|_{1,\Omega}\,\|\psi\|_{1,\Omega}.
\end{aligned}$$

It follows that there exists a positive constant $\|A\|$, depending on $\vartheta_2, c_0, \kappa_1, \kappa_2, \kappa_3$ and κ_4, such that

$$|A_{\boldsymbol{\sigma}}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi))| \leq \|A\|\,\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\|_{\mathbf{H}_2}\,\|(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\|_{\mathbf{H}_2} \quad \forall\,(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),\,(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\in\mathbf{H}_2, \tag{2.3.16}$$

and hence $A_{\boldsymbol{\sigma}}$ is bounded independently of $(\boldsymbol{\sigma},\boldsymbol{u})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega)$. In turn, we now aim to show

that $A_{\boldsymbol{\sigma}}$ is $\mathbf{H}_2$-elliptic. To this end, given $(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\in\mathbf{H}_2$, we apply (2.2.7) and find that

$$\begin{aligned}
A_{\boldsymbol{\sigma}}((\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)) &\geq \int_\Omega \vartheta(\boldsymbol{\sigma})\,\boldsymbol{s}\cdot\boldsymbol{s}+\kappa_1\,\|\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}-\kappa_1\vartheta_2\,\|\boldsymbol{s}\|_{0,\Omega}\,\|\widetilde{\boldsymbol{\tau}}\|_{0,\Omega}+\kappa_2\,\|\mathbf{div}\,\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}\\
&\quad+\kappa_3|\psi|^2_{1,\Omega}-\kappa_3\,\|\boldsymbol{s}\|_{0,\Omega}\,|\psi|_{1,\Omega}+\kappa_4\,\|\psi\|^2_{0,\Gamma}\\
&\geq \vartheta_0\,\|\boldsymbol{s}\|^2_{0,\Omega}+\kappa_1\,\|\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}+\kappa_2\,\|\mathbf{div}\,\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}-\frac{\kappa_1\vartheta_2}{2\delta}\,\|\boldsymbol{s}\|^2_{0,\Omega}-\frac{\kappa_1\vartheta_2\delta}{2}\,\|\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}\\
&\quad+\kappa_3|\psi|^2_{1,\Omega}-\frac{\kappa_3}{2\widetilde{\delta}}\,\|\boldsymbol{s}\|^2_{0,\Omega}-\frac{\kappa_3\widetilde{\delta}}{2}|\psi|^2_{1,\Omega}+\kappa_4\,\|\psi\|^2_{0,\Gamma}\\
&=\left\{\left(\vartheta_0-\frac{\kappa_1\vartheta_2}{2\delta}\right)-\frac{\kappa_3}{2\widetilde{\delta}}\right\}\|\boldsymbol{s}\|^2_{0,\Omega}+\kappa_1\left(1-\frac{\vartheta_2\delta}{2}\right)\|\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}+\kappa_2\,\|\mathbf{div}\,\widetilde{\boldsymbol{\tau}}\|^2_{0,\Omega}\\
&\quad+\kappa_3\left(1-\frac{\widetilde{\delta}}{2}\right)|\psi|^2_{1,\Omega}+\kappa_4\,\|\psi\|^2_{0,\Gamma}\,.
\end{aligned}\tag{2.3.17}$$

Then, assuming the stipulated hypotheses on $\delta,\widetilde{\delta},\kappa_1,\kappa_2,\kappa_3,\kappa_4$ and applying Lemma 2.3.2, we can define

$$\widetilde{\alpha}_1:=\left\{\left(\vartheta_0-\frac{\kappa_1\vartheta_2}{2\delta}\right)-\frac{\kappa_3}{2\widetilde{\delta}}\right\},\quad \widetilde{\alpha}_2:=\min\left\{\kappa_1\left(1-\frac{\vartheta_2\delta}{2}\right),\kappa_2\right\},\quad \widetilde{\alpha}_3:=c_2\,\min\left\{\kappa_3\left(1-\frac{\widetilde{\delta}}{2}\right),\kappa_4\right\},$$

which allows us to deduce from (2.3.17) that

$$A_{\boldsymbol{\sigma}}((\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi),(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi))\geq\widetilde{\alpha}\,\|(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\|^2_{\mathbf{H}_2}\quad\forall\,(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\in\mathbf{H}_2,\tag{2.3.18}$$

where $\widetilde{\alpha}:=\min\{\widetilde{\alpha}_1,\widetilde{\alpha}_2,\widetilde{\alpha}_3\}$ is the ellipticity constant of $A_{\boldsymbol{\sigma}}$. Next, applying Cauchy-Schwarz's inequality, the trace estimates in $\mathbf{H}(\mathrm{div};\Omega)$ and $\mathrm{H}^1(\Omega)$, with constants 1 and c_0, respectively, the upper bound for g given in (2.2.5), and the fact that $\|\cdot\|_{0,\Gamma}\leq\|\cdot\|_{1/2,\Gamma}$, to (2.3.8), we find that

$$\begin{aligned}
\left|G_{\boldsymbol{u}}(\boldsymbol{s},\widetilde{\boldsymbol{\tau}},\psi)\right| &=\left|\langle\widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{\nu},\phi_{\mathrm{D}}\rangle_\Gamma+\int_\Omega\psi\,g(\boldsymbol{u})-\kappa_2\int_\Omega g(\boldsymbol{u})\,\mathrm{div}\,\widetilde{\boldsymbol{\tau}}+\kappa_4\int_\Gamma\phi_{\mathrm{D}}\,\psi\right|\\
&\leq\|\widetilde{\boldsymbol{\tau}}\cdot\boldsymbol{\nu}\|_{-1/2,\Gamma}\,\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}+g_2\,|\Omega|^{1/2}\,\|\psi\|_{0,\Omega}+\kappa_2\,g_2\,|\Omega|^{1/2}\,\|\mathrm{div}\,\widetilde{\boldsymbol{\tau}}\|_{0,\Omega}+\kappa_4\,\|\phi_{\mathrm{D}}\|_{0,\Gamma}\,\|\psi\|_{0,\Gamma}\\
&\leq\|\widetilde{\boldsymbol{\tau}}\|_{\mathbf{div};\Omega}\,\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}+g_2\,|\Omega|^{1/2}\,\|\psi\|_{1,\Omega}+\kappa_2\,g_2\,|\Omega|^{1/2}\,\|\mathrm{div}\,\widetilde{\boldsymbol{\tau}}\|_{0,\Omega}+\kappa_4\,c_0\,\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}\,\|\psi\|_{1,\Omega}\,,
\end{aligned}$$

which yields the existence of a positive constant $\|\widetilde{G}\|$, depending on κ_2,κ_4 and c_0, such that

$$\|G_{\boldsymbol{u}}\|_{\mathbf{H}_2}\leq\|\widetilde{G}\|\left\{\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}+g_2|\Omega|^{1/2}\right\}.\tag{2.3.19}$$

Finally, a direct application of the Lax-Milgram lemma proves that for each $(\boldsymbol{\sigma},\boldsymbol{u})\in\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{L}^2(\Omega)$, problem (2.3.6) has a unique solution $\widetilde{\mathbf{S}}(\boldsymbol{\sigma},\boldsymbol{u})=(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\in\mathbf{H}_2$. Moreover, a continuous dependence result is given by

$$\|\widetilde{S}(\boldsymbol{\sigma},\boldsymbol{u})\|_{\mathbf{H}_2}=\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\|_{\mathbf{H}_2}\leq\frac{1}{\widetilde{\alpha}}\,\|G_{\boldsymbol{u}}\|_{\mathbf{H}_2'}\leq\widetilde{c}_{\mathbf{S}}\left\{\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}+g_2|\Omega|^{1/2}\right\},$$

where $\widetilde{c}_{\mathbf{S}}:=\dfrac{\|\widetilde{G}\|}{\widetilde{\alpha}}$, completing the proof. □

Note that the constants $\widetilde{\alpha}_2$ and $\widetilde{\alpha}_3$, being each one defined by the minimum between two quantities, can be maximised separately by making the corresponding quantities equal, that is by choosing κ_2 and κ_4 such that

$$\kappa_2 = \kappa_1 \left(1 - \frac{\delta\vartheta_2}{2}\right) \quad \text{and} \quad \kappa_4 = \kappa_3 \left(1 - \frac{\widetilde{\delta}}{2}\right).$$

In turn, in order to guarantee that the rest of the constants involved are bounded away from zero, we take the parameters δ, $\widetilde{\delta}$, κ_1, and κ_3 as the middle points of their feasible ranges. According to the above, we adopt the following choices:

$$\begin{aligned} &\delta = \frac{1}{\vartheta_2}, \ \kappa_1 = \frac{\delta\vartheta_0}{\vartheta_2} = \frac{\vartheta_0}{\vartheta_2^2}, \ \widetilde{\delta} = 1, \ \kappa_3 = \widetilde{\delta}\left(\vartheta_0 - \frac{\kappa_1\vartheta_2}{2\delta}\right) = \frac{\vartheta_0}{2}, \\ &\kappa_2 = \kappa_1 \left(1 - \frac{\delta\vartheta_2}{2}\right) = \frac{\vartheta_0}{2\vartheta_2^2}, \ \kappa_4 = \kappa_3 \left(1 - \frac{\widetilde{\delta}}{2}\right) = \frac{\vartheta_0}{4}, \end{aligned} \tag{2.3.20}$$

which yield

$$\widetilde{\alpha}_1 = \frac{\vartheta_0}{4}, \quad \widetilde{\alpha}_2 = \frac{\vartheta_0}{2\vartheta_2^2}, \quad \widetilde{\alpha}_3 = c_2\,\frac{\vartheta_0}{4}, \quad \text{and} \quad \widetilde{\alpha} = \min\left\{\min\{c_2, 1\}\,\frac{\vartheta_0}{4}, \frac{\vartheta_0}{2\vartheta_2^2}\right\}.$$

We end this section by introducing suitable regularity estimates on $\mathbf{S}$ and $\widetilde{\mathbf{S}}$, exactly as in [83, Section 2.2]. In fact, we concentrate in the case where Ω is a convex polygonal domain and $n = 2$, recall that $\boldsymbol{f}(\psi) \in \mathbf{H}^1(\Omega)$ for each $\psi \in \mathrm{H}^1(\Omega)$, and assume from now on that $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{3/2+\gamma}(\Gamma)$, where γ is the positive constant whose existence is guaranteed in [27]. Then, applying precisely the estimate given in [27, eq. (3.9)] and recalling from the constitutive equation that the regularities of the unknowns are connected, we find that $\mathbf{S}(\psi) \in \mathbb{H}_0(\mathbf{div};\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega) \times \mathbf{H}^{2+\gamma}(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega)$.

In turn, for $\widetilde{\mathbf{S}}$ we note that, for a given pair $(\boldsymbol{\zeta}, \boldsymbol{w}) := (\mathbf{S}_1(\psi), \mathbf{S}_2(\psi)) \in \mathbb{H}_0(\mathbf{div};\Omega) \cap \mathbb{H}^{1+\gamma}(\Omega) \times \mathbf{H}^{2+\gamma}(\Omega)$ (which denote the first and second components of the unique solution produced by the operator $\mathbf{S}$), the hypothesis given by relation (2.2.8) implies in particular that $g(\boldsymbol{w}) \in \mathrm{H}^{\gamma}(\Omega)$. Additionally, we assume that the coefficients $\vartheta(\boldsymbol{\zeta})_{ij}$ are in $C^{1+\gamma}(\overline{\Omega})$ and $\phi_{\mathrm{D}} \in \mathrm{H}^{3/2+\gamma}(\Gamma)$, then elliptic regularity results (*cf.* [93], [121]) guarantee that $\phi := \widetilde{\mathbf{S}}_3(\boldsymbol{\zeta}, \boldsymbol{w}) \in \mathrm{H}^{2+\gamma}(\Omega)$, and therefore there exists $\widetilde{C}_1 > 0$ such that

$$\|\widetilde{\mathbf{S}}_1(\boldsymbol{\zeta}, \boldsymbol{w})\|_{1+\gamma,\Omega} = \|\boldsymbol{t}\|_{1+\gamma,\Omega} \leq \|\phi\|_{2+\gamma,\Omega} \leq \widetilde{C}_1 \left\{\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|g(\boldsymbol{w})\|_{\gamma,\Omega}\right\}. \tag{2.3.21}$$

On the other hand, the Sobolev embedding theorem (*cf.* [107, Thm. A.5]) establishes the continuous injection $i_\gamma : \mathbf{H}^{1+\gamma}(\Omega) \longrightarrow C^0(\overline{\Omega})$, with boundedness constant $\widetilde{C}_\gamma$. Then, applying (2.3.21) implies that

$$\|\widetilde{\mathbf{S}}_1(\boldsymbol{\zeta}, \boldsymbol{w})\|_{\infty,\Omega} = \|\boldsymbol{t}\|_{\infty,\Omega} \leq \widetilde{C}_\gamma \|\boldsymbol{t}\|_{1+\gamma,\Omega} \leq \widetilde{C}_\gamma \widetilde{C}_1 \left\{\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|g(\boldsymbol{w})\|_{\gamma,\Omega}\right\}. \tag{2.3.22}$$

Finally, replacing the estimates (2.2.8) and (2.3.14) into (2.3.22), we find that

$$\|\widetilde{\mathbf{S}}_1(\boldsymbol{\zeta}, \boldsymbol{w})\|_{\infty,\Omega} = \|\boldsymbol{t}\|_{\infty,\Omega} \leq C_\infty \left\{\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\}, \tag{2.3.23}$$

where C_∞ is a positive constant depending on C_γ, $c_{\mathbf{S}}$, $\widetilde{C}_\gamma$ and $\widetilde{C}_1$ (*cf.* (2.2.8), (2.3.14), (2.3.21), (2.3.22)).

2.3.3 Solvability analysis of the fixed-point equation

We now verify the hypotheses of the Schauder fixed-point theorem (see, e.g. [54, Theorem 9.12-1]). Before starting the result to be proved, we restrict $\mathbf{T}$ to a ball and show that this operator maps into itself.

Lemma 2.3.5. *Let W be the closed and convex subset of $\mathrm{H}^1(\Omega)$ defined by*

$$W := \left\{ \phi \in \mathrm{H}^1(\Omega) : \quad \|\phi\|_{1,\Omega} \leq \widetilde{c}_{\mathbf{S}} \left(\|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2 |\Omega|^{1/2} \right) \right\},$$

where $\widetilde{c}_{\mathbf{S}}$ is the constant given by (2.3.14). Then $\mathbf{T}(W) \subseteq W$.

Proof. It suffices to recall the definition of $\mathbf{T}$ and apply the estimate (2.3.15). □

The following estimate is key to derive Lipschitz continuity of $\mathbf{T}$. For a proof see [83, Lemma 2.6].

Lemma 2.3.6. *There exists a positive constant $C_{\mathbf{S}}$ depending on μ, L_f, α (cf.(2.2.4), (2.3.12)) and the inf-sup constant of b, such that*

$$\|\mathbf{S}(\phi) - \mathbf{S}(\varphi)\|_{\mathbf{H}_1} \leq C_{\mathbf{S}} \, \|\phi - \varphi\|_{0,\Omega} \quad \forall \, \phi, \varphi \in \mathrm{H}^1(\Omega). \tag{2.3.24}$$

We are in a position to establish the announced property of the operator $\mathbf{T}$.

Lemma 2.3.7. *Let $C_{\mathbf{S}}$ be the constant provided by* Lemma 2.3.6. *Then, for each $\phi, \varphi \in \mathrm{H}^1(\Omega)$, there holds*

$$\|\mathbf{T}(\phi) - \mathbf{T}(\varphi)\|_{1,\Omega} \leq \frac{C_{\mathbf{S}}}{\widetilde{\alpha}} \left\{ L_g (1 + \kappa_2^2)^{1/2} + L_\vartheta (1 + \kappa_1^2)^{1/2} \|\widetilde{\mathbf{S}}_1(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi))\|_{\infty,\Omega} \right\} \|\phi - \varphi\|_{0,\Omega} \, . \tag{2.3.25}$$

Proof. We begin by recalling that $\mathbf{T}(\phi) = \widetilde{\mathbf{S}}_3(\mathbf{S}_1(\phi), \mathbf{S}_2(\phi))$ and $\mathbf{T}(\varphi) = \widetilde{\mathbf{S}}_3(\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)) \quad \forall \, \phi, \varphi \in \mathrm{H}^1(\Omega)$. For notational purposes we rename

$$(\boldsymbol{\sigma}, \boldsymbol{u}) := (\mathbf{S}_1(\phi), \mathbf{S}_2(\phi)) \quad \text{and} \quad (\boldsymbol{\zeta}, \boldsymbol{w}) := (\mathbf{S}_1(\varphi), \mathbf{S}_2(\varphi)),$$

where $(\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\zeta}, \boldsymbol{w}) \in \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{L}^2(\Omega)$. Next, we consider $(\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) := \widetilde{\mathbf{S}}(\boldsymbol{\sigma}, \boldsymbol{u})$ and $(\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi) := \widetilde{\mathbf{S}}(\boldsymbol{\zeta}, \boldsymbol{w})$, that is, for each $(\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi) \in \mathbf{H}_2$, one has

$$A_{\boldsymbol{\sigma}}((\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi), (\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi)) = G_{\boldsymbol{u}}(\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi) \quad \text{and} \quad A_{\boldsymbol{\zeta}}((\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi), (\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi)) = G_{\boldsymbol{w}}(\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi).$$

Analogously to the proof of [83, Lemma 2.7], we apply the ellipticity of $A_{\boldsymbol{\sigma}}$ (*cf.* (2.3.18)) and then, by adding and subtracting appropiate terms, we find that

$$\begin{aligned} \widetilde{\alpha} \|(\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)\|^2_{\mathbf{H}_2} & \\ & \leq A_{\boldsymbol{\sigma}}((\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi), (\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)) - A_{\boldsymbol{\sigma}}((\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi), (\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)) \\ & = (G_{\boldsymbol{u}} - G_{\boldsymbol{w}})((\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)) + (A_{\boldsymbol{\zeta}} - A_{\boldsymbol{\sigma}})((\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi), (\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)). \end{aligned} \tag{2.3.26}$$

Using the definition of $A_{\boldsymbol{\sigma}}$,$G_{\boldsymbol{u}}$, Cauchy-Schwarz's inequality, and (2.2.5),(2.2.6), we can assert that

$$\begin{aligned}|(G_{\boldsymbol{u}}-G_{\boldsymbol{w}})((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi))| &= \left|\int_\Omega (g(\boldsymbol{u})-g(\boldsymbol{w}))\left\{(\phi-\varphi)-\kappa_2\,\mathrm{div}\,(\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\zeta}})\right\}\right| \\ &\leq L_g\,\|\boldsymbol{u}-\boldsymbol{w}\|_{0,\Omega}\left\{\|\phi-\varphi\|_{0,\Omega}+\kappa_2\|\mathrm{div}\,(\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\zeta}})\|_{0,\Omega}\right\} \\ &\leq L_g(1+\kappa_2^2)^{1/2}\,\|\boldsymbol{u}-\boldsymbol{w}\|_{0,\Omega}\,\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi)\|_{\mathbf{H}_2},\end{aligned} \tag{2.3.27}$$

and

$$\begin{aligned}|(A_{\boldsymbol{\zeta}}-A_{\boldsymbol{\sigma}})((\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi),(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi))| &= \left|\int_\Omega (\vartheta(\boldsymbol{\zeta})-\vartheta(\boldsymbol{\sigma}))\,\boldsymbol{r}\cdot\left\{(\boldsymbol{t}-\boldsymbol{r})-\kappa_1(\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\zeta}})\right\}\right| \\ &\leq L_\vartheta\,\|\boldsymbol{\sigma}-\boldsymbol{\zeta}\|_{0,\Omega}\,\|\boldsymbol{r}\|_{\infty,\Omega}\left\{\|\boldsymbol{t}-\boldsymbol{r}\|_{0,\Omega}+\kappa_1\|\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\zeta}}\|_{0,\Omega}\right\}. \\ &\leq L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\zeta}\|_{0,\Omega}\,\|\boldsymbol{r}\|_{\infty,\Omega}\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi)\|_{\mathbf{H}_2},\end{aligned} \tag{2.3.28}$$

whence the inequalities (2.3.26), (2.3.27) and (2.3.28) imply that

$$\begin{aligned}&\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{r},\widetilde{\boldsymbol{\zeta}},\varphi)\|_{\mathbf{H}_2} \\ &\quad\leq \frac{1}{\widetilde{\alpha}}\left\{L_g(1+\kappa_2^2)^{1/2}\,\|\boldsymbol{u}-\boldsymbol{w}\|_{0,\Omega}+L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\zeta}\|_{0,\Omega}\,\|\boldsymbol{r}\|_{\infty,\Omega}\right\}.\end{aligned} \tag{2.3.29}$$

Next, according to the definitions given when starting the proof, we can rewrite (2.3.29) as

$$\begin{aligned}\|\widetilde{\mathbf{S}}(\mathbf{S}_1(\phi),\mathbf{S}_2(\phi))-\widetilde{\mathbf{S}}(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi)\|_{\mathbf{H}_2} \leq \frac{1}{\widetilde{\alpha}}\Big\{&L_g(1+\kappa_2^2)^{1/2}\,\|\mathbf{S}_2(\phi)-\mathbf{S}_2(\varphi)\|_{0,\Omega} \\ &+L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\mathbf{S}_1(\phi)-\mathbf{S}_1(\varphi)\|_{0,\Omega}\,\|\widetilde{\mathbf{S}}_1(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi))\|_{\infty,\Omega}\Big\}.\end{aligned} \tag{2.3.30}$$

It is important to note here that, when needed, $\|\widetilde{\mathbf{S}}_1(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi))\|_{\infty,\Omega}$ can be bounded by (2.3.23), for each $\varphi\in\mathrm{H}^1(\Omega)$. Finally, applying estimates (2.3.24) and (2.3.30), we find that

$$\begin{aligned}\|\mathbf{T}(\phi)-\mathbf{T}(\varphi)\|_{1,\Omega} &= \|\widetilde{\mathbf{S}}_3(\mathbf{S}_1(\phi),\mathbf{S}_2(\phi))-\widetilde{\mathbf{S}}_3(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi))\|_{1,\Omega} \\ &\leq \frac{1}{\widetilde{\alpha}}\,C_{\mathbf{S}}\left\{L_g(1+\kappa_2^2)^{1/2}+L_\vartheta(1+\kappa_1^2)^{1/2}\|\widetilde{\mathbf{S}}_1(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi))\|_{\infty,\Omega}\right\}\|\phi-\varphi\|_{0,\Omega}.\end{aligned}$$

which gives (2.3.25), completing the proof. □

The next lemma establishes the continuity and compactness of $\mathbf{T}$.

Lemma 2.3.8. *Let W be as in Lemma 2.3.5. Then $\mathbf{T}:W\to W$ is continuous and $\overline{\mathbf{T}(W)}$ is compact.*

Proof. It follows straightforwardly from (2.3.25) and the continuity of $i_c:\mathrm{H}^1(\Omega)\to\mathrm{L}^2(\Omega)$ that

$$\|\mathbf{T}(\phi)-\mathbf{T}(\varphi)\|_{1,\Omega}\leq\frac{1}{\widetilde{\alpha}}\,C_{\mathbf{S}}\,\|i_c\|\left\{L_g(1+\kappa_2^2)^{1/2}+L_\vartheta(1+\kappa_1^2)^{1/2}\|\widetilde{\mathbf{S}}_1(\mathbf{S}_1(\varphi),\mathbf{S}_2(\varphi))\|_{\infty,\Omega}\right\}\|\phi-\varphi\|_{1,\Omega},$$

which proves continuity of $\mathbf{T}$. In turn, let $\{\phi_k\}_{k\in\mathbb{N}}$ be a sequence of W, which is clearly bounded. Then, there exists a subsequence $\{\phi_k^{(1)}\}_{k\in\mathbb{N}}\subseteq\{\phi_k\}_{k\in\mathbb{N}}$ and $\phi\in\mathrm{H}^1(\Omega)$ such that $\phi_k^{(1)}\xrightarrow{w}\phi\in\mathrm{H}^1(\Omega)$. In this way, thanks to the compactness of i_c, we deduce that $\phi_k^{(1)}\to\phi\in\mathrm{L}^2(\Omega)$, which, combined with (2.3.25), implies that $\mathbf{T}(\phi_k^{(1)})\to\mathbf{T}(\phi)\in\mathrm{H}^1(\Omega)$, and proves the compactness of $\overline{\mathbf{T}(W)}$. □

We are ready now to prove that (2.3.10) is well-posed. From Lemmas 2.3.5 and 2.3.8, the existence of solution is merely an application of Schauder's theorem. Futrhermore, assuming that the data is small enough, we can prove uniqueness of solution. This is indeed possible thanks to the regularity estimates established at the end of Section 2.3.2.

Details of the proof are similar to those available in [83, Thm. 2.9].

Theorem 2.3.9. *Let W be as in Lemma 2.3.5. Then, the augmented fully-mixed problem (2.3.9) has at least one solution $\big((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\big)\in\mathbf{H}_1\times\mathbf{H}_2$ with $\phi\in W$, satisfying the bounds*

$$\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)\|_{\mathbf{H}_2}\leq\widetilde{c}_{\mathbf{S}}\left\{\|\phi_{\mathrm{D}}\|_{1/2,\Gamma}+g_2|\Omega|^{1/2}\right\},$$
$$\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))\|_{\mathbf{H}_1}\leq c_{\mathbf{S}}\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right\}.$$

Moreover, if the data satisfy

$$\frac{1}{\widetilde{\alpha}}C_{\mathbf{S}}\left\{L_g(1+\kappa_2^2)^{1/2}+L_\vartheta(1+\kappa_1^2)^{1/2}C_\infty\left(\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma}+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\right\}<1,$$

then the solution ϕ is unique in W.

2.4 The Galerkin scheme and well-posedness of the discrete problem

In this section we introduce and analyse a Galerkin scheme for (2.3.9). We adopt the discrete analogue of the fixed-point strategy introduced in Section 2.3.2 and apply the Brouwer fixed-point theorem to prove existence of discrete solution. We start by considering generic finite dimensional subspaces

$$\mathbb{H}_h^{\boldsymbol{\sigma}}\subseteq\mathbb{H}_0(\mathbf{div};\Omega),\quad\mathbf{H}_h^{\boldsymbol{u}}\subseteq\mathbf{L}^2(\Omega),\quad\mathbb{H}_h^{\boldsymbol{\rho}}\subseteq\mathbb{L}^2_{\mathrm{skew}}(\Omega),\tag{2.4.1}$$
$$\mathbf{H}_h^{\boldsymbol{t}}\subseteq\mathbf{L}^2(\Omega),\quad\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\subseteq\mathbf{H}(\mathrm{div};\Omega),\quad\text{and}\quad\mathrm{H}_h^{\phi}\subseteq\mathrm{H}^1(\Omega),\tag{2.4.2}$$

which will be specified later on. Hereafter, h denotes the size of a regular partition $\mathcal{T}_h$ of $\overline{\Omega}$ into triangular (in 2D) or tetrahedral (in 3D) elements K of diameter h_K, i.e. $h:=\max\{h_K:\ K\in\mathcal{T}_h\}$. A Galerkin scheme for (2.3.9) reads: find $(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h),\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\times\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}$ such that

$$\begin{aligned}
a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)+b(\boldsymbol{\tau}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))&=G(\boldsymbol{\tau}_h)&&\forall\,\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}},\\
b(\boldsymbol{\sigma}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))&=F_{\phi_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h)&&\forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}},\\
A_{\boldsymbol{\sigma}_h}((\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h),(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h))&=G_{\boldsymbol{u}_h}(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)&&\forall\,(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}.
\end{aligned}\tag{2.4.3}$$

In order to address the well-posedness of (2.4.3), we proceed analogously as in Section 2.3.2 and apply a fixed-point strategy. In fact, we define the operator $\mathbf{S}_h:\mathrm{H}_h^{\phi}\to\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})$ as

$$\mathbf{S}_h(\phi_h):=(\mathbf{S}_{1,h}(\phi_h),(\mathbf{S}_{2,h}(\phi_h),\mathbf{S}_{3,h}(\phi_h))):=(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\quad\forall\phi_h\in\mathrm{H}_h^{\phi},$$

where $(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))$ is the unique solution of

$$\begin{aligned}
a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)+b(\boldsymbol{\tau}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))&=G(\boldsymbol{\tau}_h)&&\forall\,\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}},\\
b(\boldsymbol{\sigma}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))&=F_{\phi_h}(\boldsymbol{v}_h,\boldsymbol{\eta}_h)&&\forall\,(\boldsymbol{v}_h,\boldsymbol{\eta}_h)\in\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}},
\end{aligned}\tag{2.4.4}$$

with F_{ϕ_h} being defined by (2.3.2) with $\phi = \phi_h$. In turn, we introduce $\widetilde{\mathbf{S}}_h : \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \to \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^{\phi}$ as

$$\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) := (\widetilde{\mathbf{S}}_{1,h}(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), \widetilde{\mathbf{S}}_{2,h}(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), \widetilde{\mathbf{S}}_{3,h}(\boldsymbol{\sigma}_h, \boldsymbol{u}_h)) := (\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h) \quad \forall\, (\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}},$$

where $(\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h)$ is the unique solution of

$$A_{\boldsymbol{\sigma}_h}((\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h), (\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h)) = G_{\boldsymbol{u}_h}(\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h) \quad \forall\, (\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h) \in \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^{\phi}, \tag{2.4.5}$$

with $A_{\boldsymbol{\sigma}_h}$ and $G_{\boldsymbol{u}_h}$ being defined by (2.3.7) with $\boldsymbol{\sigma} = \boldsymbol{\sigma}_h$ and (2.3.8) with $\boldsymbol{u} = \boldsymbol{u}_h$, respectively. In this way, by introducing the operator $\mathbf{T}_h : \mathrm{H}_h^{\phi} \to \mathrm{H}_h^{\phi}$ as $\mathbf{T}_h(\phi_h) := \widetilde{\mathbf{S}}_{3,h}(\mathbf{S}_{1,h}(\phi_h), \mathbf{S}_{2,h}(\phi_h)) \quad \forall \phi_h \in \mathrm{H}_h^{\phi}$, we realize that (2.4.3) can be rewritten as the fixed-point problem: find $\phi_h \in \mathrm{H}_h^{\phi}$ such that

$$\mathbf{T}_h(\phi_h) = \phi_h. \tag{2.4.6}$$

Analogously to the continuous case, we first study the well-posedness of $\mathbf{S}_h$ and $\widetilde{\mathbf{S}}_h$, and hence the well-definiteness of $\mathbf{T}_h$. To this end we proceed as in [83, Section 3.2] and incorporate further hypotheses on the discrete spaces $\mathbb{H}_h^{\boldsymbol{\sigma}}, \mathbf{H}_h^{\boldsymbol{u}}$ and $\mathbb{H}_h^{\boldsymbol{\rho}}$. Let V_h be the discrete kernel of b given by

$$V_h := \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}} \ : \ b(\boldsymbol{\tau}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h)) = 0 \quad \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}} \right\},$$

and assume the following discrete inf-sup conditions:

(**H.0**) There exists a constant $\alpha_1 > 0$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h \in V_h \\ \boldsymbol{\tau}_h \neq 0}} \frac{a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)}{\|\boldsymbol{\tau}_h\|_{\mathbf{div};\Omega}} \geq \alpha_1 \|\boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega} \quad \forall\, \boldsymbol{\sigma}_h \in V_h. \tag{2.4.7}$$

(**H.1**) There exists a constant $\beta_1 > 0$, independent of h, such that

$$\sup_{\substack{\boldsymbol{\tau}_h \in \mathbf{H}_h^{\boldsymbol{\sigma}} \\ \boldsymbol{\tau}_h \neq 0}} \frac{b(\boldsymbol{\tau}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h))}{\|\boldsymbol{\tau}_h\|_{\mathbf{div};\Omega}} \geq \beta_1 \|(\boldsymbol{v}_h, \boldsymbol{\eta}_h)\|_{\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)} \quad \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}. \tag{2.4.8}$$

Deriving well-posedness of the discrete problem (2.4.4) results as a straightforward application of the discrete Babuška-Brezzi theory. Firstly, the operators related to a and b, and the functionals G and F_{ϕ_h} are all bounded on subspaces of the corresponding continuous spaces. Next, the inf-sup conditions are given by (**H.0**) and (**H.1**). The unique solvability of (2.4.4) is abridged in the following lemma.

Lemma 2.4.1. *For each $\phi_h \in \mathrm{H}_h^{\phi}$, problem (2.4.4) has a unique solution $\mathbf{S}_h(\phi_h) := (\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h)) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times (\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}})$. Moreover, there exists $\widetilde{C} > 0$, depending on μ, α_1 and β_1 (cf.(2.4.7), (2.4.8)), but independent of ϕ_h and h, such that*

$$\|\mathbf{S}_h(\phi_h)\|_{\mathbf{H}_1} = \|(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h))\|_{\mathbf{H}_1} \leq \widetilde{C} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}.$$

In regard to the problem defined by $\widetilde{\mathbf{S}}$ we state next the discrete analogue of Lemma 2.3.4.

Lemma 2.4.2. *Assume that* $\kappa_1 \in \left(0, \frac{2\delta\vartheta_0}{\vartheta_2}\right)$ *and* $\kappa_3 \in \left(0, 2\widetilde{\delta}\left(\vartheta_0 - \frac{\kappa_1\vartheta_2}{2\delta}\right)\right)$ *with* $\delta \in \left(0, \frac{2}{\vartheta_2}\right), \widetilde{\delta} \in (0,2)$, *and* $\kappa_2, \kappa_4 > 0$. *Then, for each* $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, *problem* (2.4.5) *has a unique solution* $\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) = (\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h) \in \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^{\phi}$. *Moreover, with the constant* $\widetilde{c}_{\mathbf{S}}$ *provided by* Lemma 2.3.4, *there holds*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h)\|_{\mathbf{H}_2} = \|(\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h)\|_{\mathbf{H}_2} \leq \widetilde{c}_{\mathbf{S}} \left\{ \|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2|\Omega|^{1/2} \right\}. \tag{2.4.9}$$

Proof. We first observe that for each $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, the operator $A_{\boldsymbol{\sigma}_h}$ is bounded and elliptic on $\mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^{\phi}$ with the same constants $\|A\|$ and $\widetilde{\alpha}$ from Lemma 2.3.4. In addition, $\widetilde{G}_{\boldsymbol{u}_h}$ restricted to $\mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^{\phi}$ is bounded as in (2.3.19) with $\boldsymbol{u}_h$ in place of $\boldsymbol{u}$. Therefore, the result is a direct application of the Lax-Milgram lemma. □

We notice in advance that, instead of the regularity estimates employed in the continuous case (not applicable in the present discrete case), we simply utilise properties of the discrete subspaces chosen. In what follows, we verify the hypotheses of the Brouwer fixed-point theorem (see, *e.g.* [54, Thm. 9.9-2]) to prove that $\mathbf{T}_h$ has at least one fixed point.

Lemma 2.4.3. *Let* $W_h := \left\{ \phi_h \in \mathrm{H}_h^{\phi} : \ \|\phi_h\|_{1,\Omega} \leq \widetilde{c}_{\mathbf{S}} \left(\|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2|\Omega|^{1/2} \right) \right\}$. *Then* $\mathbf{T}_h(W_h) \subseteq W_h$.

Proof. It is basically an application of the definition of $\mathbf{T}_h$ and the estimate (2.4.9). □

Lemma 2.4.4. *There exists* $C > 0$ *depending on* μ, L_f, α_1 *and* β_1 (*cf.*(2.2.4), (2.4.7), (2.4.8)) *such that*

$$\|\mathbf{S}_h(\phi_h) - \mathbf{S}_h(\varphi_h)\|_{\mathbf{H}_1} \leq C \, \|\phi_h - \varphi_h\|_{0,\Omega} \quad \forall \, \phi_h, \varphi_h \in \mathrm{H}_h^{\phi}.$$

Proof. See [83, Lemma 3.4]. □

Lemma 2.4.5. *For each* $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), (\boldsymbol{\zeta}_h, \boldsymbol{w}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, *there holds*

$$\begin{aligned} &\|\widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h) - \widetilde{\mathbf{S}}_h(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)\|_{\mathbf{H}_2} \\ &\leq \frac{1}{\widetilde{\alpha}} \left\{ L_g(1+\kappa_2^2)^{1/2} \, \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{0,\Omega} + L_\vartheta(1+\kappa_1^2)^{1/2} \|\widetilde{\mathbf{S}}_{1,h}(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)\|_{\infty,\Omega} \, \|\boldsymbol{\sigma}_h - \boldsymbol{\zeta}_h\|_{0,\Omega} \right\}. \end{aligned} \tag{2.4.10}$$

Proof. Proceeding as in [83, Lemma 3.5], given $(\boldsymbol{\sigma}_h, \boldsymbol{u}_h), (\boldsymbol{\zeta}_h, \boldsymbol{w}_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, we let $(\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h) = \widetilde{\mathbf{S}}_h(\boldsymbol{\sigma}_h, \boldsymbol{u}_h)$ and $(\boldsymbol{r}_h, \widetilde{\boldsymbol{\zeta}}_h, \varphi_h) = \widetilde{\mathbf{S}}(\boldsymbol{\zeta}_h, \boldsymbol{w}_h)$. Then, analogously to the proof of Lemma 2.3.7, we get

$$\begin{aligned} &\widetilde{\alpha}\|(\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{r}, \widetilde{\boldsymbol{\zeta}}, \varphi)\|_{\mathbf{H}_2} \\ &\quad \leq \left\{ L_g(1+\kappa_2^2)^{1/2} \, \|\boldsymbol{u}_h - \boldsymbol{w}_h\|_{0,\Omega} + L_\vartheta(1+\kappa_1^2)^{1/2} \|\boldsymbol{r}_h\|_{\infty,\Omega} \, \|\boldsymbol{\sigma}_h - \boldsymbol{\zeta}_h\|_{0,\Omega} \right\} \|\phi_h - \varphi_h\|_{0,\Omega}. \end{aligned}$$

Since the elements of $\mathbf{H}_h^{\boldsymbol{t}}$ are piecewise polynomials (to be specified later on) it follows that $\|\boldsymbol{r}_h\|_{\infty,\Omega} < +\infty$, and hence the foregoing equation yields (2.4.10). Futher details are omitted. □

As a consequence of the above Lemmas, we can state the Lipschitz continuity of $\mathbf{T}_h$.

Lemma 2.4.6. *Assume that* C *is as in* Lemma 2.4.4. *Then, for each* $\phi_h, \varphi_h \in \mathrm{H}_h^{\phi}$, *there holds*

$$\begin{aligned} &\|\mathbf{T}_h(\phi_h) - \mathbf{T}_h(\varphi_h)\|_{1,\Omega} \\ &\quad \leq \frac{C}{\widetilde{\alpha}} \left\{ L_g(1+\kappa_2^2)^{1/2} + L_\vartheta(1+\kappa_1^2)^{1/2} \|\widetilde{\mathbf{S}}_{1,h}(\mathbf{S}_{1,h}(\varphi), \mathbf{S}_{2,h}(\varphi))\|_{\infty,\Omega} \right\} \|\phi_h - \varphi_h\|_{0,\Omega}. \end{aligned}$$

Proof. It suffices to recall that $\mathbf{T}_h(\phi_h) = \widetilde{\mathbf{S}}_{3,h}(\mathbf{S}_{1,h}(\phi_h), \mathbf{S}_{2,h}(\phi_h))$ for $\phi_h \in \mathrm{H}_h^\phi$ and apply Lemmas 2.4.4 and 2.4.5. □

At this point, we are able to state the main result of this section.

Theorem 2.4.7. *Let* $W_h := \left\{ \phi_h \in \mathrm{H}_h^\phi : \ \|\phi_h\|_{1,\Omega} \leq \widetilde{c}_{\mathbf{S}} \big(\|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2 |\Omega|^{1/2} \big) \right\}$. *Then* (2.4.3) *has at least one solution* $(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h), \boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times (\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}) \times \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^\phi$ *with* $\phi_h \in W_h$, *and there holds*

$$\|(\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h)\|_{\mathbf{H}_2} \leq \widetilde{c}_{\mathbf{S}} \left\{ \|\phi_{\mathrm{D}}\|_{1/2,\Gamma} + g_2 |\Omega|^{1/2} \right\}, \tag{2.4.11}$$

$$\|(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h))\|_{\mathbf{H}_1} \leq \widetilde{C} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}. \tag{2.4.12}$$

Proof. After using Lemmas 2.4.3 and 2.4.6, the result is a straightforward consequence of Brouwer's fixed-point theorem. In turn, bounds (2.4.11) and (2.4.12) follow from Lemmas 2.4.2 and 2.4.1, respectively. □

2.5 Error analysis for the proposed Galerkin method

In this section we advocate the derivation of error estimates for (2.4.3). For this purpose, we consider in what follows $\big((\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})), (\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi)\big) \in \mathbf{H}_1 \times \mathbf{H}_2$, with $\phi \in W$, and $(\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h), \boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h) \in \mathbb{H}_h^{\boldsymbol{\sigma}} \times (\mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}) \times \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^\phi$, with $\phi_h \in W_h$, be the solutions of (2.3.9) and (2.4.3), respectively. We seek an upper bound for

$$\|(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho}), \boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h), \boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h)\| \,,$$

for which, we suggest to estimate $\|(\boldsymbol{\sigma}, (\boldsymbol{u}, \boldsymbol{\rho})) - (\boldsymbol{\sigma}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h))\|$ and $\|(\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi) - (\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h)\|$, separately. With this goal in mind, we first rearrange (2.3.9) and (2.4.3) as follows

$$\begin{aligned}
a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) &= G(\boldsymbol{\tau}) && \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \\
b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) &= F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) && \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \\
a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h)) &= G(\boldsymbol{\tau}_h) && \forall\, \boldsymbol{\tau}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}}, \\
b(\boldsymbol{\sigma}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h)) &= F_{\phi_h}(\boldsymbol{v}_h, \boldsymbol{\eta}_h) && \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\boldsymbol{\rho}}
\end{aligned} \tag{2.5.1}$$

and

$$\begin{aligned}
A_{\boldsymbol{\sigma}}((\boldsymbol{t}, \widetilde{\boldsymbol{\sigma}}, \phi), (\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi)) &= G_{\boldsymbol{u}}(\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi) \quad \forall\, (\boldsymbol{s}, \widetilde{\boldsymbol{\tau}}, \psi) \in \mathbf{H}_2, \\
A_{\boldsymbol{\sigma}_h}((\boldsymbol{t}_h, \widetilde{\boldsymbol{\sigma}}_h, \phi_h), (\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h)) &= G_{\boldsymbol{u}_h}(\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h) \quad \forall\, (\boldsymbol{s}_h, \widetilde{\boldsymbol{\tau}}_h, \psi_h) \in \mathbf{H}_h^{\boldsymbol{t}} \times \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} \times \mathrm{H}_h^\phi.
\end{aligned} \tag{2.5.2}$$

Next, we recall from [127, Thm. 11.2 and 11.1] two instrumental results. First, a Strang inequality for saddle point problems where continuous and discrete formulations differ only in the functional. This will be applied to (2.5.1). Second, the standard Strang Lemma for elliptic problems, which fits (2.5.2). We will not write them explicitly here but will refer to these lemmas as *Saddle-point Strang Lemma*, and *Elliptic Strang Lemma*, respectively.

From now on, we denote as usual

$$\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right):=\inf_{(\boldsymbol{\tau}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))\in\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})}\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\tau}_h,(\boldsymbol{v}_h,\boldsymbol{\eta}_h))\|_{\mathbf{H}_1},$$

and

$$\mathrm{dist}\left((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\right):=\inf_{(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}}\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\|_{\mathbf{H}_2}.$$

Next, a straightforward application of the Saddle-point Strang Lemma yields the following result concerning a priori estimates for $\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|$. Details of the proof can be found in [83, Lemma 3.10].

Lemma 2.5.1. *There exists a constant $C_{\mathrm{ST}}>0$, depending on μ,α_1 and β_1 (cf. (2.4.7), (2.4.8)), such that*

$$\begin{aligned}&\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_{\mathbf{H}_1}\\&\quad\leq C_{\mathrm{ST}}\left\{\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right)+L_f\,\|\phi-\phi_h\|_{0,\Omega}\right\}.\end{aligned}\tag{2.5.3}$$

In turn, an estimate for $\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|$ reads as follows.

Lemma 2.5.2. *Let $\widetilde{C}_{\mathrm{ST}}:=\widetilde{\alpha}^{-1}\max\{1,\|A\|\}$, where $\|A\|$ and $\widetilde{\alpha}$ are the boundedness and ellipticity constants, respectively, of the bilinear form $A_{\boldsymbol{\sigma}}$ (cf. (2.3.16), (2.3.18)). Then, there holds*

$$\begin{aligned}\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|_{\mathbf{H}_2}\leq\widetilde{C}_{\mathrm{ST}}\Big\{&(1+2\,\|A\|)\,\mathrm{dist}\left((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\right)\\&+L_g(1+\kappa_2^2)^{1/2}\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}+L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,\|\boldsymbol{t}\|_{\infty,\Omega}\Big\}.\end{aligned}\tag{2.5.4}$$

Proof. Applying the Elliptic Strang Lemma in the context of (2.5.2), gives

$$\begin{aligned}&\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|_{\mathbf{H}_2}\\&\leq\widetilde{C}_{\mathrm{ST}}\left\{\sup_{\substack{(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\\(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\neq0}}\frac{|G_{\boldsymbol{u}}(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)-G_{\boldsymbol{u}_h}(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)|}{\|(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\|}\right.\\&\quad+\inf_{\substack{(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\\(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\neq0}}\left(\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\|\right.\\&\quad\left.\left.+\sup_{\substack{(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\\(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\neq0}}\frac{|A_{\boldsymbol{\sigma}}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h))-A_{\boldsymbol{\sigma}_h}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h))|}{\|(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\|}\right)\right\}.\end{aligned}\tag{2.5.5}$$

Then, proceeding analogously as in the proof of Lemma 2.3.7, we deduce that

$$\sup_{\substack{(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\\(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\neq0}}\frac{|G_{\boldsymbol{u}}(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)-G_{\boldsymbol{u}_h}(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)|}{\|(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\|}\leq L_g(1+\kappa_2^2)^{1/2}\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}.\tag{2.5.6}$$

In turn, in much the same way as [15, Lemma 5.2], we add and subtract suitable terms to write

$$\begin{aligned}
&A_{\boldsymbol{\sigma}}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)) - A_{\boldsymbol{\sigma}_h}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h))\\
&\quad = A_{\boldsymbol{\sigma}}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)-(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)) + (A_{\boldsymbol{\sigma}}-A_{\boldsymbol{\sigma}_h})((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h))\\
&\qquad + A_{\boldsymbol{\sigma}_h}((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)),
\end{aligned}$$

thus, the estimates for the first and third terms follow by applying the boundedness (2.3.16), whereas for the second one, we find that

$$\begin{aligned}
(A_{\boldsymbol{\sigma}}-A_{\boldsymbol{\sigma}_h})((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)) &= \int_\Omega (\vartheta(\boldsymbol{\sigma})-\vartheta(\boldsymbol{\sigma}_h))\,\boldsymbol{t}\cdot(\boldsymbol{r}_h-\kappa_1\widetilde{\boldsymbol{\zeta}}_h)\\
&\leq L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,\|\boldsymbol{t}\|_{\infty,\Omega}\,\|(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\|,
\end{aligned}$$

whence, we deduce that

$$\begin{aligned}
&\sup_{\substack{(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\in\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\\(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\neq 0}} \frac{|A_{\boldsymbol{\sigma}}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)) - A_{\boldsymbol{\sigma}_h}((\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h),(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h))|}{\|(\boldsymbol{r}_h,\widetilde{\boldsymbol{\zeta}}_h,\varphi_h)\|}\\
&\qquad\leq 2\,\|A\|\,\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{s}_h,\widetilde{\boldsymbol{\tau}}_h,\psi_h)\| + L_\vartheta(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}\,\|\boldsymbol{t}\|_{\infty,\Omega}\,.
\end{aligned} \tag{2.5.7}$$

Finally, by replacing (2.5.6)-(2.5.7) into (2.5.5), we get (2.5.4), which ends the proof. □

Now, to derive the Céa estimate for the total error we combine Lemmas 2.5.1 and 2.5.2. To this end, and for notational convenience, we introduce the following constants

$$C_1 := C_{\mathrm{ST}}\,\widetilde{C}_{\mathrm{ST}}\,L_g\,(1+\kappa_2^2)^{1/2}, \qquad C_2 := C_{\mathrm{ST}}\,\widetilde{C}_{\mathrm{ST}}\,C_\infty\,L_\vartheta\,(1+\kappa_1^2)^{1/2}, \qquad C_3 := \widetilde{C}_{\mathrm{ST}}\,(1+2\,\|A\|).$$

Next we replace the bounds for $\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}$ and $\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,\Omega}$ into (2.5.4), and apply from (2.3.23) that

$$\|\boldsymbol{t}\|_{\infty,\Omega} \leq C_\infty\left\{\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right\}.$$

We then perform algebraic manipulations to find that

$$\begin{aligned}
&\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|_{\mathbf{H}_2}\\
&\quad\leq \left\{C_1 + C_2\left(\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right)\right\}\,\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}),\mathbf{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbf{H}_h^{\boldsymbol{\rho}})\right)\\
&\qquad + L_f\left\{C_1 + C_2\left(\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right)\right\}\|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|\\
&\qquad + C_3\,\mathrm{dist}\left((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\right)
\end{aligned} \tag{2.5.8}$$

Consequently, we can establish the following result which provides the complete Céa estimate.

Theorem 2.5.3. *Suppose that the data satisfy*

$$L_f\left\{C_1 + C_2\left(\|\phi_{\mathrm{D}}\|_{3/2+\gamma,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right)\right\} < \frac{1}{2}.$$

Then, there exist positive constants C_4 and C_5 independent of h, such that

$$\begin{aligned}
&\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho}))-(\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_{\mathbf{H}_1} + \|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi)-(\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|_{\mathbf{H}_2}\\
&\qquad\leq C_4\,\mathrm{dist}\left((\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})),\mathbb{H}_h^{\boldsymbol{\sigma}}\times(\mathbf{H}_h^{\boldsymbol{u}}\times\mathbb{H}_h^{\boldsymbol{\rho}})\right) + C_5\,\mathrm{dist}\left((\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi),\mathbf{H}_h^{\boldsymbol{t}}\times\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}\times\mathrm{H}_h^{\phi}\right).
\end{aligned} \tag{2.5.9}$$

Proof. It follows straightforwardly from (2.5.3) and (2.5.8). □

We now specify finite element subspaces satisfying (2.4.1)-(2.4.2) and the discrete inf-sup conditions (**H.0**)-(**H.1**). Given an integer $k \geq 0$, for each $K \in \mathcal{T}_h$ we let $\mathrm{P}_k(K)$ be the space of polynomial functions on K of degree $\leq k$ and define the local Raviart-Thomas space of order k as

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

where $\mathbf{P}_k(K) = [\mathrm{P}_k(K)]^2$, and $\boldsymbol{x}$ is the generic vector in $\mathbb{R}^2$. Let b_K be the element bubble function defined as the unique polynomial in $\mathrm{P}_{k+1}(K)$ vanishing on ∂K with $\int_K b_K = 1$. Then, for each $K \in \mathcal{T}_h$ we consider the bubble space of order k, defined by

$$\mathbf{B}_k(K) := \mathrm{P}_k(K)\left(\frac{\partial b_K}{\partial x_2}, -\frac{\partial b_K}{\partial x_1}\right).$$

One option to approximate stress, displacement and rotation is the classical PEERS elements [22]:

$$\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbf{RT}_k(K) \oplus \mathbf{B}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathbf{H}_h^{\boldsymbol{u}} &:= \{\boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathbb{H}_h^{\boldsymbol{\rho}} &:= \{\boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \quad \boldsymbol{\eta}_h \in \mathbf{C}(\Omega) \quad \text{and} \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\}\,.
\end{aligned} \tag{2.5.10}$$

We could also employ the Arnold-Falk-Winther (AFW, [24]) elements for the elasticity unknowns:

$$\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \{\boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbf{BDM}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathbf{H}_h^{\boldsymbol{u}} &:= \{\boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathbb{H}_h^{\boldsymbol{\rho}} &:= \{\boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,
\end{aligned} \tag{2.5.11}$$

and recall that both PEERS and AFW satisfy (**H.0**) and (**H.1**) (*cf.* [22, Lemma 4.4], [23, Thm. 11.9]).

In turn, we define the approximating spaces for the concentration gradient, diffusive flux and solute concentration as piecewise polynomials of degree $\leq k$, Raviart-Thomas elements of order k, and Lagrange finite elements up to degree $k+1$, respectively:

$$\begin{aligned}
\mathbf{H}_h^{\boldsymbol{t}} &:= \{\boldsymbol{t}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{t}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}} &:= \{\widetilde{\boldsymbol{\tau}}_h \in \mathbf{H}(\mathrm{div};\Omega) : \quad \widetilde{\boldsymbol{\tau}}_h|_K \in \mathbf{RT}_k(K) \quad \forall\, K \in \mathcal{T}_h\}\,,\\
\mathrm{H}_h^{\phi} &:= \{\psi_h \in \mathbf{C}(\Omega) \quad \boldsymbol{\psi}_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\}\,.
\end{aligned} \tag{2.5.12}$$

Approximation properties of the spaces in (2.5.10),(2.5.11), (2.5.12) can be found in e.g. [39, 81]. They can be combined with the Céa estimate (2.5.9) and the assumption of adequately small data, to produce the theoretical rates of convergence of (2.4.3), summarized in what follows.

Theorem 2.5.4. *In addition to the hypotheses of Theorems 2.3.9, 2.4.7 and 2.5.3, assume that there exists $s > 0$ such that $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega)$, $\mathbf{div}\,\boldsymbol{\sigma} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{u} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{\rho} \in \mathbb{H}^s(\Omega)$, $\boldsymbol{t} \in \mathbf{H}^s(\Omega)$, $\widetilde{\boldsymbol{\sigma}} \in \mathbf{H}^s(\Omega)$, $\mathrm{div}\,\widetilde{\boldsymbol{\sigma}} \in \mathrm{H}^s(\Omega)$ and $\phi \in \mathrm{H}^{1+s}(\Omega)$. Then, there exists $\widetilde{C} > 0$, independent of h, such that, with the finite element subspaces defined by either (2.5.10) or (2.5.11) and (2.5.12), there holds*

$$\begin{aligned}
&\|(\boldsymbol{\sigma},(\boldsymbol{u},\boldsymbol{\rho})) - (\boldsymbol{\sigma}_h,(\boldsymbol{u}_h,\boldsymbol{\rho}_h))\|_{\mathbf{H}_1} + \|(\boldsymbol{t},\widetilde{\boldsymbol{\sigma}},\phi) - (\boldsymbol{t}_h,\widetilde{\boldsymbol{\sigma}}_h,\phi_h)\|_{\mathbf{H}_2} \leq \widetilde{C} h^{\min\{s,k+1\}} \Big\{ \|\boldsymbol{\sigma}\|_{s,\Omega}\\
&\quad + \|\mathbf{div}\,\boldsymbol{\sigma}\|_{s,\Omega} + \|\boldsymbol{u}\|_{s,\Omega} + \|\boldsymbol{\rho}\|_{s,\Omega} + \|\boldsymbol{t}\|_{s,\Omega} + \|\widetilde{\boldsymbol{\sigma}}\|_{s,\Omega} + \|\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}\|_{s,\Omega} + \|\phi\|_{1+s,\Omega} \Big\}.
\end{aligned}$$

Augmented $\mathbf{BDM}_1 - \mathbf{P}_0 - \mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_0 - \mathrm{P}_1$ scheme							
N	h	$\mathrm{e}(\boldsymbol{\sigma})$	$r(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$r(\boldsymbol{u})$	$\mathrm{e}(\boldsymbol{\rho})$	$r(\boldsymbol{\rho})$
129	0.7071	1.36946	-	1.902e-02	-	6.669e-02	-
465	0.3536	0.71736	0.9328	9.897e-03	0.9422	3.412e-02	0.9668
1761	0.1768	0.36290	0.9831	4.989e-03	0.9884	1.716e-02	0.9918
6849	0.0883	0.18198	0.9957	2.499e-03	0.9976	8.590e-03	0.9982
27009	0.0441	0.09106	0.9989	1.250e-03	0.9994	4.296e-03	0.9996
107265	0.0221	0.04553	0.9997	6.249e-04	0.9999	2.148e-03	0.9999

$\mathrm{e}(\boldsymbol{t})$	$r(\boldsymbol{t})$	$\mathrm{e}(\widetilde{\boldsymbol{\sigma}})$	$r(\widetilde{\boldsymbol{\sigma}})$	$\mathrm{e}(\phi)$	$r(\phi)$	iter
3.767e-02	-	1.342e-01	-	4.632e-02	-	3
2.263e-02	0.7352	7.352e-02	0.8683	2.525e-02	0.8752	3
1.192e-02	0.9244	3.762e-02	0.9668	1.331e-02	0.9245	3
6.047e-03	0.9795	1.891e-02	0.9923	6.822e-03	0.9637	3
3.035e-03	0.9947	9.466e-03	0.9982	3.445e-03	0.9858	3
1.519e-03	0.9987	4.734e-03	0.9996	1.728e-03	0.9950	3

Augmented $\mathbf{BDM}_2 - \mathbf{P}_1 - \mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_1 - \mathrm{P}_2$ scheme							
N	h	$\mathrm{e}(\boldsymbol{\sigma})$	$r(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$r(\boldsymbol{u})$	$\mathrm{e}(\boldsymbol{\rho})$	$r(\boldsymbol{\rho})$
337	0.7071	0.40310	-	5.614e-03	-	1.770e-02	-
1265	0.3536	0.10681	1.9160	1.468e-03	1.9350	4.764e-03	1.8940
4897	0.1768	0.02705	1.9810	3.717e-04	1.9820	1.224e-03	1.9600
19265	0.0883	0.00678	1.9960	9.323e-05	1.9950	3.091e-04	1.9860
76417	0.0441	0.00169	2.0000	2.333e-05	1.9990	7.752e-05	1.9950
304385	0.0221	0.00042	2.0000	5.833e-06	2.0000	1.940e-05	1.9980

$\mathrm{e}(\boldsymbol{t})$	$r(\boldsymbol{t})$	$\mathrm{e}(\widetilde{\boldsymbol{\sigma}})$	$r(\widetilde{\boldsymbol{\sigma}})$	$\mathrm{e}(\phi)$	$r(\phi)$	iter
1.345e-02	-	4.492e-02	-	1.514e-02	-	4
3.993e-03	1.7520	1.284e-02	1.8070	4.342e-03	1.8020	3
1.054e-03	1.9220	3.321e-03	1.9510	1.156e-03	1.9100	3
2.685e-04	1.9730	8.378e-04	1.9870	2.987e-04	1.9520	3
6.763e-05	1.9890	2.100e-04	1.9960	7.599e-05	1.9750	3
1.696e-05	1.9950	5.254e-05	1.9990	1.917e-05	1.9870	3

Table 2.1: Example 1: Convergence history and Picard iteration count for the augmented $\mathbf{BDM}_{k+1} - \mathbf{P}_k - \mathbb{P}_k - \mathbf{RT}_k - \mathbf{P}_k - \mathrm{P}_{k+1}$ approximations with $k = 0, 1$. Here N stands for the number of degrees of freedom associated to the each triangulation $\mathcal{T}_h$ (table produced by the author).

2.6 Numerical results

In this section we present some examples illustrating the performance of our augmented fully-mixed scheme (2.4.3), and confirming the rates of convergence provided by Theorem 2.5.4. These numerical results also include examples in which some of the data do not necessarily satisfy all the hypotheses required, thus confirming the potentiality of the method proposed, and also evidencing that only technical limitations are preventing us from extending our theoretical analysis to more general cases.

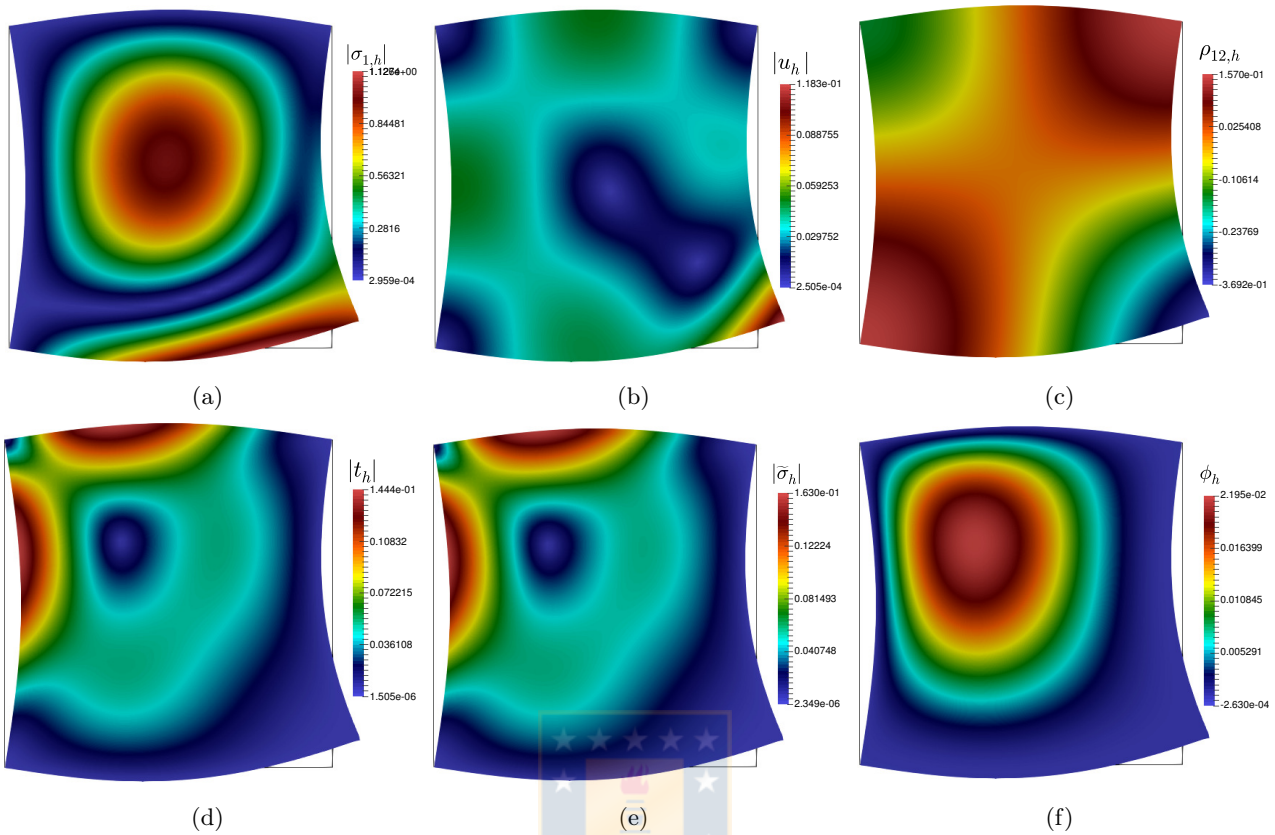


Figure 2.1: Example 1: Lowest-order approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), relevant component of the rotation $\boldsymbol{\rho}_h$ (c), gradient of concentration $|\boldsymbol{t}_h|$ (d), diffusive flux $|\widetilde{\boldsymbol{\sigma}}_h|$ (e), and solute concentration ϕ_h (f). All fields are plotted on the deformed domain (figure produced by the author).

Our implementation is based on the FEniCS library [11]. In turn, a Picard algorithm with tolerance of 10^{-6} on the ℓ^∞-norm of the residual has been employed for the fixed-point problem (2.4.6). The boundary conditions employed in Examples 2 and 3 were motivated by the specific application of stress-assisted diffusion problems in lithium batteries, and they correspond to mixed boundary conditions, which are currently not supported by our theoretical analysis. Nevertheless the obtained results still show stable and robust computations, which insinuates that only technical difficulties prevent us of extending our analysis to the case of mixed boundary conditions. For the diffusion sub-problem in Example 2 we have utilised the variational formulation (2.3.4) applying a fixed-point on $\boldsymbol{\sigma}$ and $\boldsymbol{t}$; and $\boldsymbol{\sigma}$ and ϕ, respectively.

Example 1. In our first numerical test we take the unit square as computational domain $\Omega = (0,1)^2$

and choose the following manufactured exact solutions and coupling terms to (2.3.9):

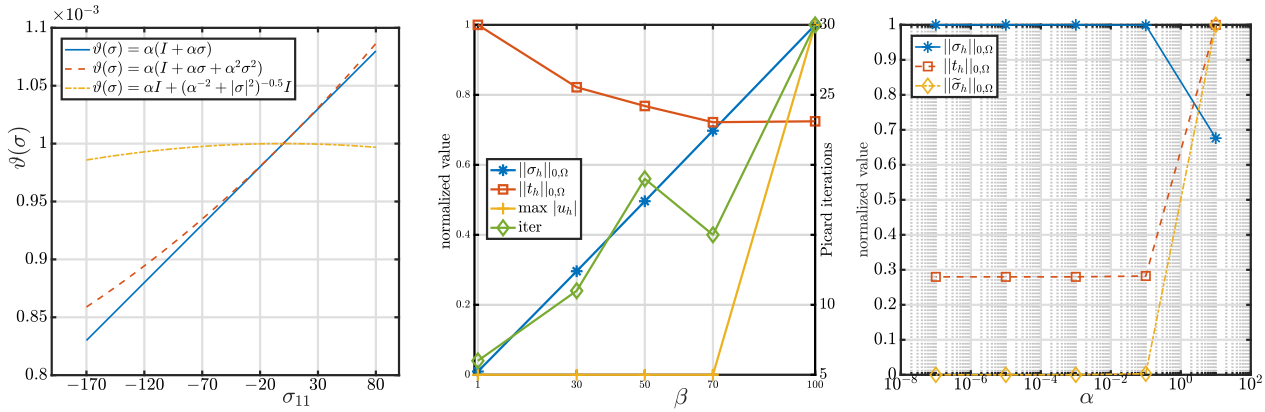
$$\boldsymbol{u} = \begin{pmatrix} d_1\cos(\pi x_1)\sin(\pi x_2) + \frac{x_1^2(1-x_2)^2}{2\lambda} \\ -d_1\sin(\pi x_1)\cos(\pi x_2) + \frac{x_1^3(1-x_2)^3}{2\lambda} \end{pmatrix}, \quad \boldsymbol{\sigma} = \lambda\,\mathrm{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbf{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}), \quad \boldsymbol{\rho} = \nabla u - \boldsymbol{\varepsilon}(\boldsymbol{u}),$$
$$\phi = (1-x_1)^2\,x_1\,(1-x_2)\,x_2^2, \quad \boldsymbol{t} = \nabla\phi, \quad \widetilde{\boldsymbol{\sigma}} = \vartheta(\boldsymbol{\sigma})\,\boldsymbol{t}, \tag{2.6.1}$$
$$\vartheta(\boldsymbol{\sigma}) = (D_0 + D_1(1+|\boldsymbol{\sigma}|^2)^{-0.5})\,\mathbf{I}, \quad \boldsymbol{f}(\phi) = d_2\begin{pmatrix}\cos(\phi)\\ -\sin(\phi)\end{pmatrix}, \quad g(\boldsymbol{u}) = 2 + \frac{1}{1+|\boldsymbol{u}|^2}.$$

We note that the tensorial diffusivity, body load and diffusive source terms satisfy (2.2.4)-(2.2.6) and (2.2.7). Moreover, the elasticity and diffusion equations are considered non-homogeneous and the extra source terms, as well as the non-homogeneous boundary data $\boldsymbol{u}_{\mathrm{D}}$ and ϕ_{D}, are chosen according to (2.6.1). This treatment does not compromise the continuous and discrete analysis, as the smoothness of the exact solution provides right-hand sides with terms in $\mathrm{L}^2(\Omega)$, thus only requiring a slight modification of the functionals in the variational formulation. The Lamé constants $\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}$ and $\mu = \frac{E}{2+2\nu}$ are computed using the values $E = 10$ and $\nu = 0.3$ [163]. The remaining model parameters are given by: $d_1 = 0.05, d_2 = 0.1$, $D_0 = 1.0, D_1 = 0.1$, and $\vartheta_0 = D_0$, $\vartheta_2 = \sqrt{2}(D_0 + D_1)$. According to (2.3.20), the stabilisation parameters are taken as $\kappa_1 = \vartheta_0/\vartheta_2^2$, $\kappa_2 = \vartheta_0/2\,\vartheta_2^2$, $\kappa_3 = \vartheta_0/2$ and $\kappa_4 = \vartheta_0/4$. The convergence of the approximate solutions is assessed by computing errors in the respective norms and experimental rates, that we define as usual

$$\mathrm{e}(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}, \ \mathrm{e}(\boldsymbol{u}) = \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}, \ \mathrm{e}(\boldsymbol{\rho}) = \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\Omega}, \ \mathrm{e}(\boldsymbol{t}) = \|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,\Omega},$$
$$\mathrm{e}(\widetilde{\boldsymbol{\sigma}}) = \|\widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h\|_{\mathbf{div};\Omega}, \ \mathrm{e}(\phi) = \|\phi - \phi_h\|_{1,\Omega}, \ r(\cdot) = \log(\mathrm{e}(\cdot)/\widehat{\mathrm{e}}(\cdot))[\log(h/\widehat{h})]^{-1},$$

where $\mathrm{e}, \widehat{\mathrm{e}}$ denote errors computed on two consecutive meshes of sizes $h, \widehat{h}$, respectively. We choose the finite element spaces (2.5.11) and (2.5.12), that is $\mathbf{BDM}_{k+1}-\mathbf{P}_k-\mathbb{P}_k-\mathbf{RT}_k-\mathbf{P}_k-\mathrm{P}_{k+1}$ approximations with $k = 0$ and $k = 1$. Errors and decay rates are summarised in Table 2.1, where we observe that optimal convergence $O(h^{k+1})$ is attained for all fields in their relevant norms. These findings are in agreement with the bounds given by Theorem 2.5.4. In all cases, three Picard steps were required to reach the desired tolerance. Sample solutions are displayed in Figure 2.1.

Example 2. Next we concentrate on the simulation of microscopic lithiation of an anode. Details on model derivation and physical considerations can be found, for instance, in [138, 110, 50], whereas the specific settings that motivate this example are summarised in [128]. The domain consists of a truncated sphere of radius 10 μm, representing the silicon core of a secondary particle (see Figure 2.3(a)), which we discretise using an unstructured mesh of 104913 tetrahedral elements. For this test we consider lowest order Raviart-Thomas elements for the flux and the concentration gradient, and piecewise linear and continuous polynomials for the concentration (see the method developed in [85]). We assume that the face of the truncated sphere which is closest to the plane $x_1 = 0$ (denoted Γ_{D}) is in contact with a region of electrolyte, that is, the zone between the sphere and the surrounding cube. On Γ_{D} we set zero-flux of lithium and also consider that the anode has an external layer that does not permit displacement of the body, so there we set $\boldsymbol{u} = \mathbf{0}$. On the remainder of the boundary, $\Gamma_{\mathrm{N}} = \Gamma \setminus \Gamma_{\mathrm{D}}$, we prescribe a maximum lithium concentration $\phi = \phi_{\max}$ with $\phi_{\max}$= 26390 mol m^{-3}, as well as $\boldsymbol{\sigma}\boldsymbol{\nu} = \beta\phi\mathbf{I}\boldsymbol{\nu}$, where β is a parameter to be specified later on. We assume that the source term is zero, and the diffusivity is specified as $\vartheta(\boldsymbol{\sigma}) = D_0\mathbf{I} + D_1\boldsymbol{\sigma}$ with D_0 = 1.2e-21 m^{2}s^{-1}, D_1 = 3.9e-14 m^{2}s^{-1},



(a) (b) (c)

Figure 2.2: Example 2. Approximation of different functions $\vartheta(\boldsymbol{\sigma})$ varying the $\boldsymbol{\sigma}_{11}$ component (a), normalised $\mathbf{L}^2$-norm for $\boldsymbol{\sigma}_h$ and $\boldsymbol{t}_h$, $\boldsymbol{\ell}^\infty$-norm for $\boldsymbol{u}_h$ and number of Picard iterations needed for different values of β with E=100 (b), and normalized $\mathbf{L}^2$-norm for $\boldsymbol{\sigma}_h, \boldsymbol{t}_h$ and $\widetilde{\boldsymbol{\sigma}}_h$ for five different values of α (c) (figure produced by the author).

and the elastic material properties of silicon are $E = 60$GPa and $\nu = 0.25$. Following the referenced models, here the total stress contains a contribution due to lithium concentration. More specifically, we consider $\boldsymbol{\sigma}^{\mathrm{tot}} = \boldsymbol{\sigma} - \beta\phi\mathbf{I}$, with $\beta = \widehat{\Omega}(3\lambda+2\mu)/3$, where $\widehat{\Omega} = 4.926\text{e-}6\ \mathrm{m}^3\,\mathrm{mol}^{-1}$ is the partial molar volume. The balance of momentum is then $-\mathbf{div}\,\boldsymbol{\sigma} = -\beta\nabla\phi$, or equivalently $-\mathbf{div}\,\boldsymbol{\sigma}^{\mathrm{tot}} = \mathbf{0}$ and the zero traction boundary condition can be recast as $\boldsymbol{\sigma}^{\mathrm{tot}}\boldsymbol{\nu} = \mathbf{0}$ on Σ.

In order to have a model with fewer chemical and physical parameters, and also to accommodate a model with adimensional units, we proceed to rescale the strong form of the governing equations and testing different deformation regimes to match the expected values found in the literature. We introduce the following parameters: the intrinsic size of the domain $L = 1.6\text{e-}5\,\mathrm{m}^2$, $\nabla^* = \nabla/L$, $\mathrm{div}^* = \mathrm{div}/L$, $\phi^* = \phi/\phi_{\max}$, $\boldsymbol{u}^* = \boldsymbol{u}/L$ and $\boldsymbol{\sigma}^* = L^2\boldsymbol{\sigma}$. Thus, taking D_0 =1.0e-2D_1L^2, we reduce the parameters $D_0, D_1, \beta, \widehat{\Omega}$ given above to only $\beta^* = \beta\phi_{\max}/L^2$ and $\alpha = 1.0\text{e-}2D_1L^2\phi_{\max}$. Making abuse of notation, we rename β^* by β, $\boldsymbol{u}^*$ by $\boldsymbol{u}$, and so on. The proper scaling of the parameters implies that the baseline case corresponds to $\beta = 5.0\text{e}1$ and $\alpha = 1.0\text{e-}3$.

Figure 2.3(i) illustrates the sharp transition between high and low concentrations as lithium diffuses from Γ_{N} into the secondary particle. In addition, Figure 2.3(c) shows more pronounced displacements near Γ_{N} (which is precisely the region where the silicon is fully lithiated), and the particle swelling is indeed influenced by the lithium gradient distribution. The stress-assisted diffusion mechanism together with the dilation-dependent source term, also contribute to maintain maximum lithium concentration near Γ_{N}. This two-way coupling effect implies in turn that the lithium concentration is less important in regions where the secondary particle is clamped.

In Figure 2.2 we show three different constitutive relations defining ϑ as a function of the first component of the Cauchy stress tensor. The first and third specifications correspond to the functions used in this test and in the accuracy example, respectively, whereas the second relationship has been used in [53] in the context of biological materials. Depending on the values attained by the stress,

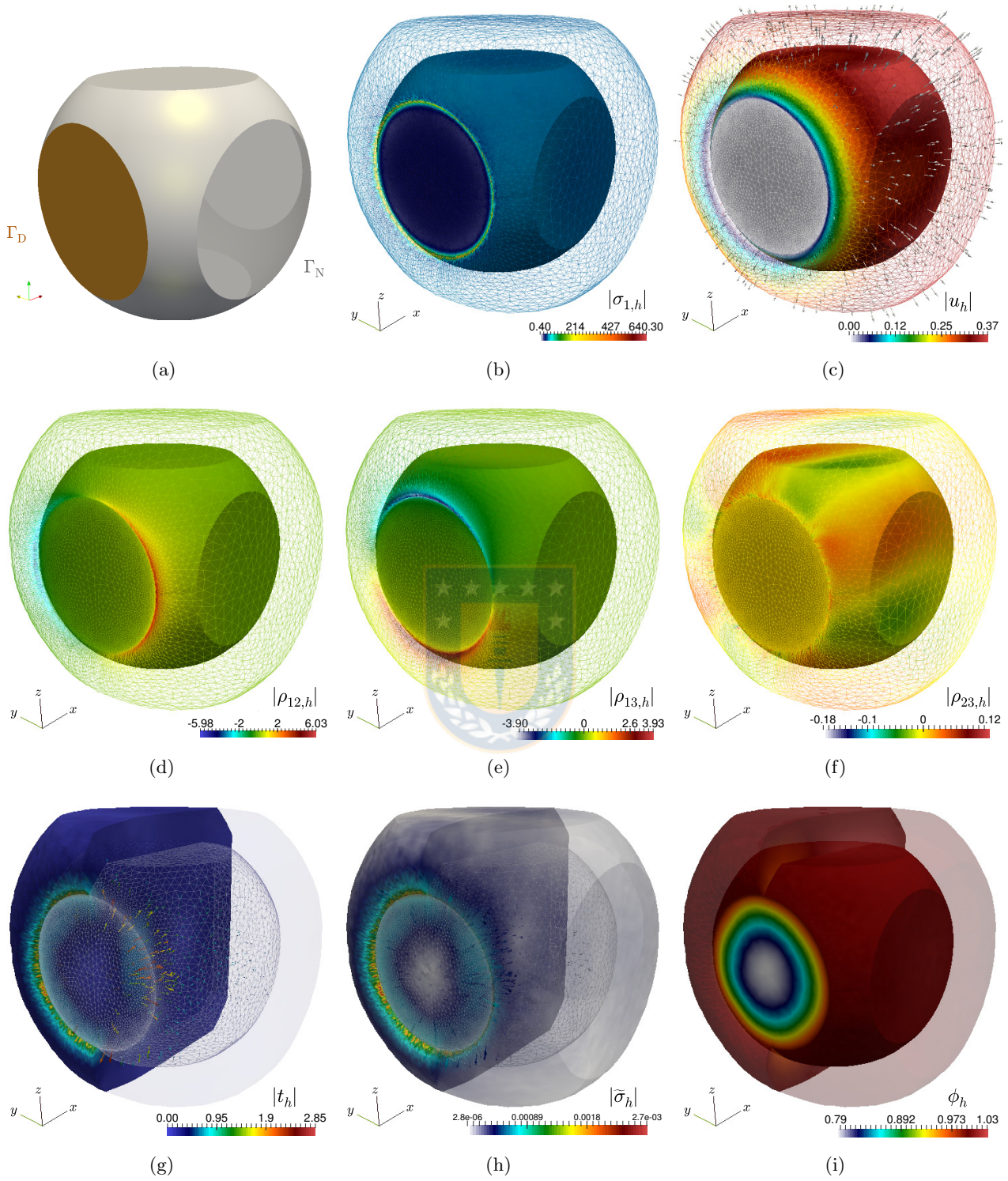


Figure 2.3: Example 2. Schematic representation of domain boundaries on a secondary particle silicone anode (a), lowest-order approximation of stress magnitude $|\boldsymbol{\sigma}_h|$ (b), displacement magnitude $|\boldsymbol{u}_h|$ (c), rotation components (d,e,f), concentration gradient $|\boldsymbol{t}_h|$ (g), diffusive flux $|\widetilde{\boldsymbol{\sigma}}_h|$ (h), and solute concentration ϕ_h (i) (figure produced by the author).

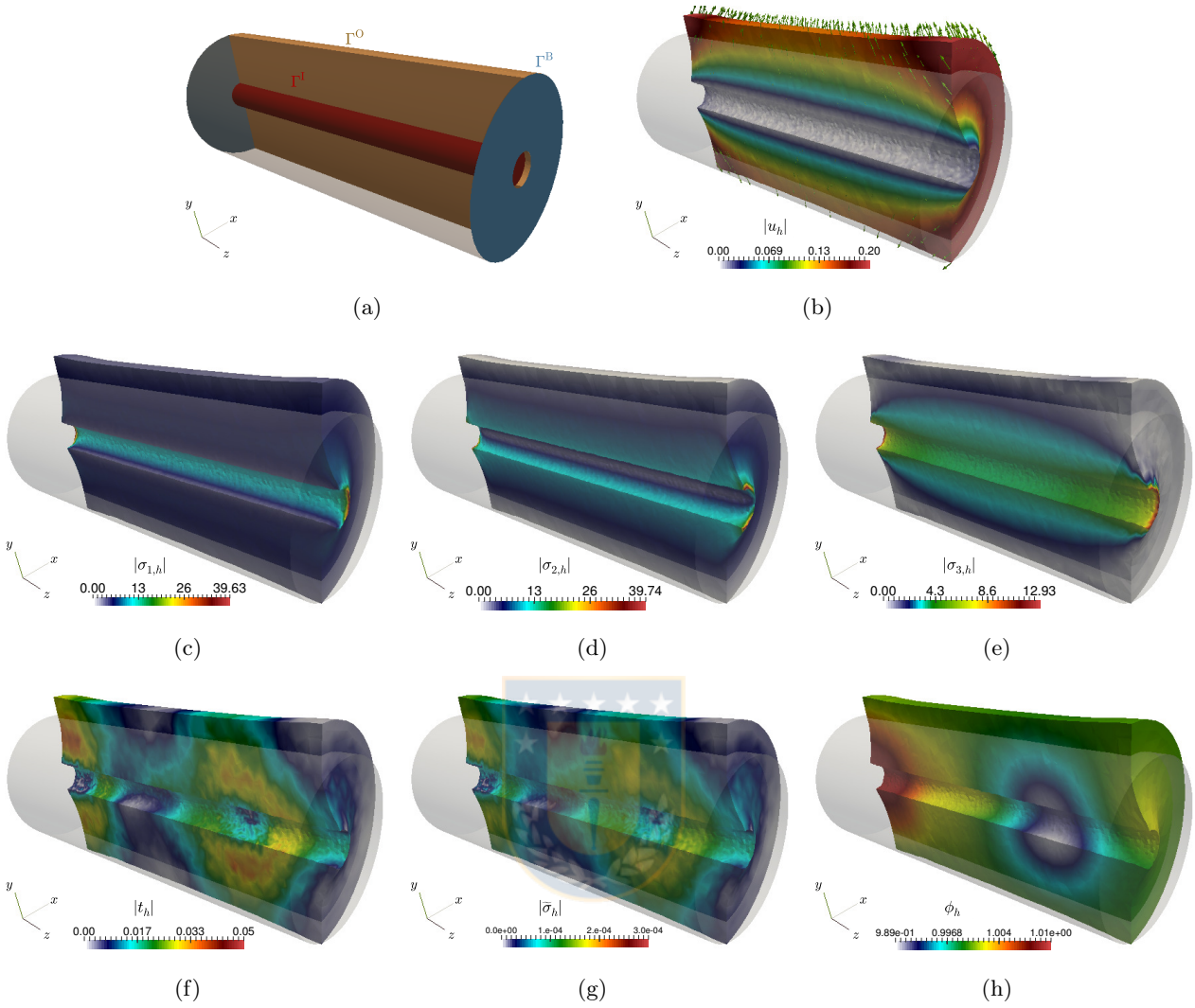


Figure 2.4: Example 3: Geometry for a perforated cylindrical particle (a), approximate displacement magnitude (b), magnitude of the rows of the approximate Cauchy stress (c,d,e), concentration gradient (f), diffusive flux (g), and concentration (h) shown on a clipped geometry (figure produced by the author).

one could then easily derive the values of the augmentation constants. On the other hand, in Figures 2.2(b) and 2.2(c) we report a study on the influence of different values of the coupling constants β and α into the norms of selected solution fields for the elasticity and diffusion problems. We remark that the $\boldsymbol{\ell}^\infty$-norm of $\boldsymbol{u}_h$ is practically invariant to moderate values of β, but it increases abruptly when this parameter approaches 70. Furthermore, the $\mathbf{L}^2$-norm of the stress increases linearly with β. As this constant drives the intensity of the deformation as well as the coupling strength, we also observe an increase in the Picard iteration count (where we stress that all fields are normalised). We also observe an increase of the $\mathbf{L}^2$-norm of the concentration gradient with respect to α, while for smaller values of α the method produces higher values of the $\mathbf{L}^2$-norm of $\widetilde{\boldsymbol{\sigma}}$.

Example 3. In our last example we test a similar model defined on a perforated cylindrical particle

(see a sketch in Figure 2.4(a)). The problem setup has been adapted from [138]. The outer and inner radii of the bases are 5 μm and 1 μm, respectively, and the height of the cylinder is 25 μm. We discretise the domain using an unstructured mesh of 101907 tetrahedral elements, and employ the method that uses the lowest-order spaces defined in (2.5.10)-(2.5.12). In this test we consider that the particle is clamped on the inner wall Γ^{I}, while zero lithium fluxes is prescribed on $\Gamma^{\mathrm{B}} \cup \Gamma^{\mathrm{I}}$. Also, we fix a maximum lithium concentration on Γ^{O}, whereas zero traction will be imposed on $\Gamma^{\mathrm{B}} \cup \Gamma^{\mathrm{O}}$. We let $E = 10\mathrm{GPa}$, $\nu = 0.3$ and $\widehat{\Omega} = 3.497\text{e-}6\ \mathrm{m}^3\,\mathrm{mol}^{-1}$. The diffusive source is zero and the diffusivity tensor and body load source are given by $\vartheta(\boldsymbol{\sigma}) = \alpha\mathbf{I} + \alpha^2\boldsymbol{\sigma} + \alpha^3\boldsymbol{\sigma}^2$ and $\boldsymbol{f}(\phi) = \beta\boldsymbol{r}\phi$, respectively, where $\boldsymbol{r}$ is the radial vector $\boldsymbol{r} = (x, y, 0)^{\mathrm{t}}$ and α, β are the adimensional parameters given in Example 2, assuming the values $\alpha = 5.0\text{e-}3$ and $\beta = 75$.

Figure 2.4 shows the approximate solutions, indicating that the cylindrical particle deforms on the faces and outer radius and having a more important displacement on the faces. Finally, as in Example 2, we observe that the lithium concentration induces the swelling of the cylindrical particle, however as on the faces Γ^{B} we now have zero-traction and zero concentration flux conditions coexisting, the lithium concentration is no longer maximal on the outer radius.

CHAPTER 3

A posteriori error analysis of mixed finite element methods for stress-assisted diffusion problems

3.1 Introduction

We have recently analysed in Chapters 1 and 2, mixed-primal and fully-mixed formulations for the stress-assisted diffusion problem. In these chapters we have used a mixed form for elasticity in terms of stress, rotation and displacement. For the diffusion problem we have studied primal and mixed approaches: the first one in terms of the solute concentration, whereas the second one has been formulated in terms of diffusive flux, solute concentration and its gradient. We have invoked regularity estimates that only hold for the specific case of convex domains and in two spatial dimensions (see details on the assumptions and their implications in [83, Section 2.2]). We have also formulated the nonlinear set of equations as a fixed-point problem, analysing it using Schauder fixed-point theory and classical tools for saddle-point equations. The associated methods use PEERS and Arnold-Falk-Winther elements for the elasticity, and either Lagrange finite elements, or a triplet of Raviart-Thomas elements and piecewise polynomials for the primal and mixed forms of the diffusion equation, respectively.

It is well known that in order to rectify the convergence of numerical schemes in pathological situations (such as in presence of singularities in the solutions, in the data, or in the domain geometry), one can introduce mesh adaptation guided by a posteriori error estimators. These indicators are essentially global quantities $\boldsymbol{\Theta}$ that are expressed in terms of local estimators $\boldsymbol{\Theta}_K$ (fully computable as a function of the discrete solution and of the data) defined on each element of a given mesh. Then, $\boldsymbol{\Theta}$ is said to be efficient (resp. reliable) if there exists a constant $C_{\rm eff} > 0$ (resp. $C_{\rm rel}$), independent of the meshsize, such that $C_{\rm eff}\boldsymbol{\Theta} + \text{h.o.t} \leq \|error\| \leq C_{\rm rel}\boldsymbol{\Theta} + \text{h.o.t}$, where h.o.t is a generic expression denoting one or several terms of higher order. Without knowing the exact solutions, these terms give an indication on which elements induce high errors (measured in a suitable norm) and should be considered for local refinement, thus guaranteeing that the discretisation error is controlled.

Diverse a posteriori error analyses for linear elasticity can be found in the literature, including for instance traditional primal schemes [48, 149], mixed finite element methods in stress-displacement-rotation form [36, 47, 51, 111], augmented mixed approaches [28], pseudostress-based mixed formulations [82], mixed schemes with pure traction boundary conditions [70], using discontinuous Galerkin methods [33, 52, 114, 156], virtual elements methods [119], or methods specifically tailored for incom-

pressible materials [102], among others. In turn, a posteriori error analyses for elliptic equations have been widely investigated by many authors (see, e.g. [7, 25, 26, 32, 147, 148] and the references therein). Although adaptive meshes are of key usefulness in stress-and-strain assisted diffusion of hydrogen in metals such as crack-capturing [142, 143] and fatigue crack growth [139, 140], a rigorous a posteriori error analysis specifically tailored for such coupled problems is still not available in the literature.

The lack of robustness of the two-way coupling between mechanical deformation and the chemical transport can affect the accuracy of the stress-assisted diffusion processes, especially under modelling peculiarities in either of the two problems. For instance, solutions with high gradients could lead to generating an excessive distortion of the finite element mesh. We therefore aim at developing robust and reliable a posteriori error estimators. Not many results are available for this particular type of problems, but we can draw inspiration from results where the elasticity and diffusion problems have been worked independently. Most of the a posteriori error estimators for elasticity in mixed form share similarities with those available for elliptic problems in divergence form, and therefore it is possible to establish an adequate analysis without the need of re-structuring the logical steps in the proofs of reliability and efficiency usually followed for classical approaches [12, 46, 47, 48, 111, 113], as well as some more recent references related with transport coupled with incompressible flow, as in e.g. [16, 17, 49, 60, 88]. For the latter, one needs to carefully handle the coupling terms, invoking properties of the nonlinear model functions (Lipschitz continuity, uniformly boundedness), as well as suitable regularity estimates.

The rest of this chapter is organised as follows. In Section 3.2 we introduce preliminary notation used throughout this chapter, and then we recall the model problem and establish some assumptions on data. The corresponding mixed-primal and fully-mixed variational formulations as well as their associated Galerkin schemes are presented in Section 3.3. Next, in Section 3.4, we derive the corresponding reliable and efficient residual-based a posteriori error indicators for our Galerkin schemes. Finally, in Section 3.5, our theoretical results are illustrated via some numerical examples, highlighting also the good performance of the associated adaptive schemes and properties of the proposed indicators.

3.2 The stress-assisted diffusion problem

3.2.1 Governing equations

Let us consider the following system of partial differential equations, governing the diffusion of a solute interacting with the motion of an elastic solid occupying the domain Ω:

$$\begin{array}{lllll} \boldsymbol{\sigma} = \lambda\,\mathrm{tr}\boldsymbol{\varepsilon}(\boldsymbol{u})\,\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) & \text{and} & -\,\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}(\phi) \text{ in } \Omega, & \boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} & \text{on } \Gamma, \\ \widetilde{\boldsymbol{\sigma}} = \vartheta(\boldsymbol{\sigma})\nabla\phi & \text{and} & -\,\mathrm{div}\,\widetilde{\boldsymbol{\sigma}} = g(\boldsymbol{u}) \text{ in } \Omega, & \phi = 0 & \text{on } \Gamma, \end{array} \tag{3.2.1}$$

where ϕ represents the local concentration of species; $\boldsymbol{\sigma}$ is the Cauchy solid stress; $\boldsymbol{u}$ is the displacement field; $\boldsymbol{\varepsilon}(\boldsymbol{u}) := \frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\mathrm{t}}\right)$ is the infinitesimal strain tensor; $\widetilde{\boldsymbol{\sigma}}$ is the diffusive flux; $\lambda, \mu > 0$ are the Lamé constants; $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{1/2}(\Gamma)$ is the corresponding Dirichlet condition for the displacement; $\vartheta : \mathrm{R}^{2\times 2} \to \mathrm{R}^{2\times 2}$ is a tensorial diffusivity; $\boldsymbol{f} : \mathrm{R} \to \mathrm{R}^2$ is a vector field of body loads (which will depend on the species concentration), and $g : \mathrm{R}^2 \to \mathrm{R}$ denotes an additional source term depending locally on the solid displacement. In what follows we will suppose that ϑ is of class C^1, uniformly positive

definite, uniformly bounded and Lipschitz continuous, meaning that there exist positive constants ϑ_0, ϑ_1, ϑ_2 and L_ϑ, such that

$$\begin{gathered}\vartheta(\boldsymbol{\tau})\boldsymbol{w}\cdot\boldsymbol{w} \geq \vartheta_0|\boldsymbol{w}|^2, \quad \vartheta_1 \leq |\vartheta(\boldsymbol{\tau})| \leq \vartheta_2 \quad \forall\, \boldsymbol{w}\in \mathrm{R}^2 \quad \forall\, \boldsymbol{\tau},\boldsymbol{\zeta}\in \mathrm{R}^{2\times 2},\\ |\vartheta(\boldsymbol{\tau})-\vartheta(\boldsymbol{\zeta})| \leq L_\vartheta|\boldsymbol{\tau}-\boldsymbol{\zeta}| \quad \forall\, \boldsymbol{\tau},\boldsymbol{\zeta}\in \mathrm{R}^{2\times 2}.\end{gathered} \tag{3.2.2}$$

Similar assumptions will be placed on the load and source functions $\boldsymbol{f}$ and g: we suppose that there exist positive constants f_1, f_2, L_f, g_1, g_2 and L_g, such that

$$f_1 \leq |\boldsymbol{f}(s)| \leq f_2, \quad |\boldsymbol{f}(s)-\boldsymbol{f}(t)| \leq L_f|s-t| \quad \forall\, s,t\in \mathrm{R}, \tag{3.2.3}$$

$$g_1 \leq g(\boldsymbol{w}) \leq g_2, \quad |g(\boldsymbol{v})-g(\boldsymbol{w})| \leq L_g|\boldsymbol{v}-\boldsymbol{w}| \quad \forall\, \boldsymbol{v},\boldsymbol{w}\in \mathrm{R}^2. \tag{3.2.4}$$

Moreover, for each $\gamma \in (0,1)$, there exists a constant $C_\gamma > 0$, such that $g(\boldsymbol{w}) \in \mathrm{H}^\gamma(\Omega)$ for each $\boldsymbol{w}\in \mathrm{H}^\gamma(\Omega)$, and

$$\|g(\boldsymbol{w})\|_{\gamma,\Omega} \leq C_\gamma\, \|\boldsymbol{w}\|_{\gamma,\Omega}\,.$$

Finally, we assume that for every $\phi \in \mathrm{H}^1(\Omega)$, we have $\boldsymbol{f}(\phi) \in \mathbf{H}^1(\Omega)$.

We point out that the reader may refer to [6, 53, 63, 83, 109] for further details concerning different variants of the model problem, as well as for specific examples of the nonlinear functions given above.

3.3 Continuous and discrete mixed formulations

In this section we recall the continuous and discrete mixed-primal and fully-mixed schemes for (3.2.1) derived in [83, Section 2] and [84, Section 3], respectively, and state their well-posedness.

3.3.1 Mixed-primal approach

The construction of a mixed formulation for the elasticity equation in (3.2.1) follows closely those in [81, 83]. Thus, from Hooke's law we have

$$\mathcal{C}^{-1}\boldsymbol{\sigma} = \boldsymbol{\varepsilon}(\boldsymbol{u}) = \nabla\boldsymbol{u} - \boldsymbol{\rho}, \quad \text{where} \quad \boldsymbol{\rho} := \frac{1}{2}(\nabla\boldsymbol{u} - \nabla\boldsymbol{u}^{\mathrm{t}}), \tag{3.3.1}$$

with $\boldsymbol{\rho} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) := \{\boldsymbol{\eta}\in \mathbb{L}^2(\Omega) : \quad \boldsymbol{\eta}+\boldsymbol{\eta}^{\mathrm{t}} = 0\}$. Moreover, an application of the orthogonal decomposition $\mathbf{H}(\mathrm{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega)\oplus \mathrm{R}\mathbb{I}$, where

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

allows us to only seek the $\mathbb{H}_0(\mathbf{div};\Omega)$-component of the stress, whereas the remaining unknows velocity and rotation are searched in $\mathbf{L}^2(\Omega)$ and $\mathbb{L}^2_{\mathrm{skew}}(\Omega)$, respectively. On the other hand, the boundary condition for ϕ indicates the appropriate trial and test space

$$\mathrm{H}^1_0(\Omega) := \left\{\psi\in \mathrm{H}^1(\Omega) : \quad \psi = 0 \ \text{ on } \ \Gamma\right\},$$

to deduce the corresponding primal formulation for the diffusion equation (second row of (3.2.1)). Therefore, denoting from now on $\vec{\boldsymbol{\sigma}} := (\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\rho}) \in \mathbf{H}_1 := \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)$, the mixed-primal variational formulation for our model problem (3.2.1) reads: Find $(\vec{\boldsymbol{\sigma}}, \phi) \in \mathbf{H}_1 \times \mathrm{H}^1_0(\Omega)$, such that

$$\begin{aligned} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) &= G(\boldsymbol{\tau}) && \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega), \\ b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) &= F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) && \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \\ A_{\boldsymbol{\sigma}}(\phi, \psi) &= G_{\boldsymbol{u}}(\psi) && \forall\, \psi \in \mathrm{H}^1_0(\Omega), \end{aligned} \tag{3.3.2}$$

where the bilinear forms $a : \mathbb{H}_0(\mathbf{div}; \Omega) \times \mathbb{H}_0(\mathbf{div}; \Omega) \to \mathbb{R}$, $b : \mathbb{H}_0(\mathbf{div}; \Omega) \times (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)) \to \mathbb{R}$ and $A_{\boldsymbol{\sigma}} : \mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega) \to \mathbb{R}$ are specified as

$$a(\boldsymbol{\zeta}, \boldsymbol{\tau}) := \int_\Omega \mathcal{C}^{-1}\boldsymbol{\sigma} : \boldsymbol{\tau} = \frac{1}{2\,\mu} \int_\Omega \boldsymbol{\zeta}^{\mathrm{d}} : \boldsymbol{\tau}^{\mathrm{d}} + \frac{1}{4(\lambda + \mu)} \int_\Omega \mathrm{tr}(\boldsymbol{\zeta})\, \mathrm{tr}(\boldsymbol{\tau}),$$

$$b(\boldsymbol{\tau}, (\boldsymbol{v}, \boldsymbol{\eta})) := \int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\tau} + \int_\Omega \boldsymbol{\eta} : \boldsymbol{\tau}, \qquad A_{\boldsymbol{\sigma}}(\varphi, \psi) := \int_\Omega \vartheta(\boldsymbol{\sigma}) \nabla\varphi \cdot \nabla\psi,$$

for $\boldsymbol{\zeta}, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega)$, $(\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega)$ and $\varphi, \psi \in \mathrm{H}^1_0(\Omega)$. In turn, the functionals $F_\phi \in (\mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega))'$, $G \in \mathbb{H}_0(\mathbf{div}; \Omega)'$ and $G_{\boldsymbol{u}} \in \mathrm{H}^1_0(\Omega)'$ are given by

$$G(\boldsymbol{\tau}) := \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma, \quad F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) := -\int_\Omega \boldsymbol{f}\,(\phi) \cdot \boldsymbol{v}, \qquad \text{and} \qquad G_{\boldsymbol{u}}(\psi) := \int_\Omega g(\boldsymbol{u})\, \psi\,,$$

for $\vec{\boldsymbol{\tau}} := (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{H}_1$ and $\psi \in \mathrm{H}^1_0(\Omega)$. Further details yielding the weak formulation (3.3.2) can be found in [83, Section 2.1], whereas its solvability follows from the fixed-point strategy developed in [83, Theorem 2.9]. We point out that for future purposes and according to the new meaning of $\boldsymbol{\sigma}$, the constitutive equation (3.3.1) now becomes

$$\mathcal{C}^{-1}\boldsymbol{\sigma} + \boldsymbol{\rho} + \frac{1}{2|\Omega|} \left(\int_\Gamma \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{\nu} \right) \mathbb{I} = \nabla \boldsymbol{u} \qquad \text{in} \quad \Omega. \tag{3.3.3}$$

In view of defining a Galerkin formulation, let us denote by $\mathcal{T}_h$ a regular partition of $\overline{\Omega}$ into triangles K of diameter h_K, where $h := \max\{h_K : K \in \mathcal{T}_h\}$ is the meshsize. Given an integer $k \geq 0$, for each $K \in \mathcal{T}_h$ we let $\mathrm{P}_k(K)$ be the space of polynomial functions on K of degree $\leq k$ and define the local Raviart-Thomas space of order k as

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

where $\mathbf{P}_k(K) = [\mathrm{P}_k(K)]^2$, and $\boldsymbol{x}$ is a generic vector in R^2. Furthermore, using the above notation, we define the Brezzi-Douglas-Marini space $\mathbb{BDM}_{k+1}(K) := [\mathrm{P}_{k+1}(K)]^{2\times 2}$. Now, let b_K be the element bubble function defined as the unique polynomial in $\mathrm{P}_3(K)$ vanishing on ∂K with $\int_K b_K = 1$. Then, for each $K \in \mathcal{T}_h$ we consider the bubble space of order k, defined by

$$\mathbf{B}_k(K) := \mathrm{P}_k(K) \left(\frac{\partial b_K}{\partial x_2}, -\frac{\partial b_K}{\partial x_1} \right).$$

Now, denoting by $\vec{\boldsymbol{\sigma}}_h := (\boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\rho}_h) \in \mathbf{H}_{1,h} := \mathbb{H}^{\boldsymbol{\sigma}}_h \times \mathbf{H}^{\boldsymbol{u}}_h \times \mathbb{H}^{\boldsymbol{\rho}}_h$, the Galerkin scheme for (3.3.2) is defined as: find $(\vec{\boldsymbol{\sigma}}_h, \phi_h) \in \mathbf{H}_{1,h} \times \mathrm{H}^\phi_h$ such that

$$\begin{aligned} a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h)) &= G(\boldsymbol{\tau}_h) && \forall\, \boldsymbol{\tau}_h \in \mathbb{H}^{\boldsymbol{\sigma}}_h, \\ b(\boldsymbol{\sigma}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h)) &= F_{\phi_h}(\boldsymbol{v}_h, \boldsymbol{\eta}_h) && \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}^{\boldsymbol{u}}_h \times \mathbb{H}^{\boldsymbol{\rho}}_h, \\ A_{\boldsymbol{\sigma}_h}(\phi_h, \psi_h) &= G_{\boldsymbol{u}_h}(\psi_h) && \forall\, \psi_h \in \mathrm{H}^\phi_h, \end{aligned} \tag{3.3.4}$$

where the involved finite element spaces are defined similar to [83, 84]. Thus, for the elasticity equation we consider the classical PEERS elements [22]:

$$\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}_h|_K \in [\mathbf{RT}_k(K)]^2 \oplus [\mathbf{B}_k(K)]^2 \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbf{H}_h^{\boldsymbol{u}} &:= \left\{ \boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbb{H}_h^{\boldsymbol{\rho}} &:= \left\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \quad \boldsymbol{\eta}_h \in \mathbf{C}(\overline{\Omega}) \quad \text{and} \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\},
\end{aligned} \tag{3.3.5}$$

and the Arnold-Falk-Winther elements [24]:

$$\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{\tau}_h|_K \in \mathbb{BDM}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \}, \\
\mathbf{H}_h^{\boldsymbol{u}} &:= \left\{ \boldsymbol{v}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbb{H}_h^{\boldsymbol{\rho}} &:= \left\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) : \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\},
\end{aligned} \tag{3.3.6}$$

whereas the approximating space for the concentration is defined as

$$\mathrm{H}_h^{\phi} := \left\{ \psi_h \in \mathrm{C}(\overline{\Omega}) \cap \mathrm{H}_0^1(\Omega) \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}.$$

The solvability and a priori error bounds for (3.3.2) and (3.3.4) are established in [83, Sections 2-3]. Denoting by c_p the Poincaré constant and defining the balls

$$W := \left\{ \phi \in \mathrm{H}_0^1(\Omega) : \quad \|\phi\|_{1,\Omega} \leq \frac{c_p^2}{\vartheta_0} g_2 |\Omega|^{1/2} \right\} \quad \text{and} \quad W_h := \left\{ \phi_h \in \mathrm{H}_h^{\phi} : \quad \|\phi_h\|_{1,\Omega} \leq \frac{c_p^2}{\vartheta_0} g_2 |\Omega|^{1/2} \right\},$$

we will assume through the rest of the paper that $(\vec{\boldsymbol{\sigma}}, \phi) \in \mathbf{H}_1 \times \mathrm{H}_0^1(\Omega)$ with $\phi \in W$, and $(\vec{\boldsymbol{\sigma}}_h, \phi_h) \in \mathbf{H}_{1,h} \times \mathrm{H}_h^{\phi}$ with $\phi_h \in W_h$, are the solutions of the continuous and discrete formulations (3.3.2) and (3.3.4), respectively. In addition, we recall from [83, Theorems 2.9 and 3.7] the following a priori estimates

$$\|\vec{\boldsymbol{\sigma}}\|_{\mathbf{H}_1} \leq c_{\mathbf{S}} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\}, \qquad \|\vec{\boldsymbol{\sigma}}_h\|_{\mathbf{H}_1} \leq \widetilde{C} \left\{ \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2} \right\},$$

where $c_{\mathbf{S}}$ and $\widetilde{C}$ are positive constants independent of ϕ and ϕ_h.

3.3.2 Fully-mixed approach

Having established in Section 3.3.1 the mixed formulation for the elasticity problem, it only remains to define a mixed formulation for the diffusion equation. Proceeding as in [84], we define $\boldsymbol{t} := \nabla\phi$, and consider the following Galerkin type terms:

$$\begin{aligned}
\kappa_1 \int_\Omega \{\widetilde{\boldsymbol{\sigma}} - \vartheta(\boldsymbol{\sigma})\, \boldsymbol{t}\} \cdot \widetilde{\boldsymbol{\tau}} = 0 \qquad & \forall\, \widetilde{\boldsymbol{\tau}} \in \mathbf{H}(\mathrm{div};\Omega), \\
\kappa_2 \int_\Omega \mathrm{div}\, \widetilde{\boldsymbol{\sigma}}\, \mathrm{div}\, \widetilde{\boldsymbol{\tau}} = -\kappa_2 \int_\Omega g(\boldsymbol{u})\, \mathrm{div}\, \widetilde{\boldsymbol{\tau}} \qquad & \forall\, \widetilde{\boldsymbol{\tau}} \in \mathbf{H}(\mathrm{div};\Omega), \\
\kappa_3 \int_\Omega \{\nabla\phi - \boldsymbol{t}\} \cdot \nabla\psi = 0 \qquad & \forall\, \psi \in \mathrm{H}_0^1(\Omega),
\end{aligned}$$

where κ_1, κ_2 and κ_3 are positive parameters to be suitably chosen. Let us group appropriately the unknowns and spaces of the diffusion problem as follows: $\widetilde{\underline{\boldsymbol{\sigma}}} := (\widetilde{\boldsymbol{\sigma}}, \boldsymbol{t}, \phi) \in \mathbf{H}_2 := \mathbf{H}(\mathrm{div};\Omega) \times \mathbf{L}^2(\Omega) \times \mathrm{H}^1_0(\Omega)$. We then have an augmented formulation for the diffusion problem: find $\widetilde{\underline{\boldsymbol{\sigma}}} \in \mathbf{H}_2$ such that

$$A_{\boldsymbol{\sigma}}(\widetilde{\underline{\boldsymbol{\sigma}}}, \widetilde{\underline{\boldsymbol{\tau}}}) = G_{\boldsymbol{u}}(\widetilde{\underline{\boldsymbol{\tau}}}) \quad \forall\, \widetilde{\underline{\boldsymbol{\tau}}} := (\widetilde{\boldsymbol{\tau}}, \boldsymbol{s}, \psi) \in \mathbf{H}_2,$$

where

$$\begin{aligned} A_{\boldsymbol{\sigma}}(\widetilde{\underline{\boldsymbol{\sigma}}}, \widetilde{\underline{\boldsymbol{\tau}}}) &:= \int_\Omega \vartheta(\boldsymbol{\sigma})\, \boldsymbol{t} \cdot \boldsymbol{s} - \int_\Omega \widetilde{\boldsymbol{\sigma}} \cdot \boldsymbol{s} + \int_\Omega \widetilde{\boldsymbol{\tau}} \cdot \boldsymbol{t} + \int_\Omega \phi\, \mathrm{div}\, \widetilde{\boldsymbol{\tau}} - \int_\Omega \psi\, \mathrm{div}\, \widetilde{\boldsymbol{\sigma}} \\ &+ \kappa_1 \int_\Omega \{\widetilde{\boldsymbol{\sigma}} - \vartheta(\boldsymbol{\sigma})\, \boldsymbol{t}\} \cdot \widetilde{\boldsymbol{\tau}} + \kappa_2 \int_\Omega \mathrm{div}\, \widetilde{\boldsymbol{\sigma}}\, \mathrm{div}\, \widetilde{\boldsymbol{\tau}} + \kappa_3 \int_\Omega \{\nabla \phi - \boldsymbol{t}\} \cdot \nabla \psi, \\ G_{\boldsymbol{u}}(\widetilde{\underline{\boldsymbol{\tau}}}) &:= \int_\Omega \psi\, g(\boldsymbol{u}) - \kappa_2 \int_\Omega g(\boldsymbol{u})\, \mathrm{div}\, \widetilde{\boldsymbol{\tau}}. \end{aligned}$$

Consequently, we arrive at the following augmented fully-mixed formulation to system (3.2.1): find $(\vec{\boldsymbol{\sigma}}, \widetilde{\underline{\boldsymbol{\sigma}}}) \in \mathbf{H}_1 \times \mathbf{H}_2$, such that

$$\begin{aligned} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, (\boldsymbol{u}, \boldsymbol{\rho})) &= G(\boldsymbol{\tau}) && \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div};\Omega), \\ b(\boldsymbol{\sigma}, (\boldsymbol{v}, \boldsymbol{\eta})) &= F_\phi(\boldsymbol{v}, \boldsymbol{\eta}) && \forall\, (\boldsymbol{v}, \boldsymbol{\eta}) \in \mathbf{L}^2(\Omega) \times \mathbb{L}^2_{\mathrm{skew}}(\Omega), \\ A_{\boldsymbol{\sigma}}(\widetilde{\underline{\boldsymbol{\sigma}}}, \widetilde{\underline{\boldsymbol{\tau}}}) &= G_{\boldsymbol{u}}(\widetilde{\underline{\boldsymbol{\tau}}}) && \forall\, \widetilde{\underline{\boldsymbol{\tau}}} \in \mathbf{H}_2. \end{aligned} \tag{3.3.7}$$

In turn, denoting from now on $\widetilde{\underline{\boldsymbol{\sigma}}}_h := (\widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{t}_h, \phi_h) \in \mathbf{H}_{2,h} := \mathbf{H}^{\widetilde{\boldsymbol{\sigma}}}_h \times \mathbf{H}^{\boldsymbol{t}}_h \times \mathrm{H}^{\phi}_h$, the associated Galerkin scheme reads: find $(\vec{\boldsymbol{\sigma}}_h, \widetilde{\underline{\boldsymbol{\sigma}}}_h) \in \mathbf{H}_{1,h} \times \mathbf{H}_{2,h}$ such that

$$\begin{aligned} a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, (\boldsymbol{u}_h, \boldsymbol{\rho}_h)) &= G(\boldsymbol{\tau}_h) && \forall\, \boldsymbol{\tau}_h \in \mathbb{H}^{\boldsymbol{\sigma}}_h, \\ b(\boldsymbol{\sigma}_h, (\boldsymbol{v}_h, \boldsymbol{\eta}_h)) &= F_{\phi_h}(\boldsymbol{v}_h, \boldsymbol{\eta}_h) && \forall\, (\boldsymbol{v}_h, \boldsymbol{\eta}_h) \in \mathbf{H}^{\boldsymbol{u}}_h \times \mathbb{H}^{\boldsymbol{\rho}}_h, \\ A_{\boldsymbol{\sigma}_h}(\widetilde{\underline{\boldsymbol{\sigma}}}_h, \widetilde{\underline{\boldsymbol{\tau}}}_h) &= G_{\boldsymbol{u}_h}(\widetilde{\underline{\boldsymbol{\tau}}}_h) && \forall\, \widetilde{\underline{\boldsymbol{\tau}}}_h := (\widetilde{\boldsymbol{\tau}}_h, \boldsymbol{s}_h, \psi_h) \in \mathbf{H}_{2,h}, \end{aligned} \tag{3.3.8}$$

where $\mathbf{H}_{1,h}$ is as in Section 3.3.1, and the remaining spaces are:

$$\begin{aligned} \mathbf{H}^{\widetilde{\boldsymbol{\sigma}}}_h &:= \{\widetilde{\boldsymbol{\tau}}_h \in \mathbf{H}(\mathrm{div};\Omega) : \quad \widetilde{\boldsymbol{\tau}}_h|_K \in \mathbf{RT}_k(K) \quad \forall\, K \in \mathcal{T}_h\}, \\ \mathbf{H}^{\boldsymbol{t}}_h &:= \{\boldsymbol{t}_h \in \mathbf{L}^2(\Omega) : \quad \boldsymbol{t}_h|_K \in \mathbf{P}_k(K) \quad \forall\, K \in \mathcal{T}_h\}, \\ \mathrm{H}^{\phi}_h &:= \{\psi_h \in \mathrm{C}(\overline{\Omega}) \cap \mathrm{H}^1_0(\Omega) : \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\}. \end{aligned} \tag{3.3.9}$$

Finally, similarly as in Section 3.3.1, and defining the balls

$$W := \left\{\phi \in \mathrm{H}^1_0(\Omega) : \ \|\phi\|_{1,\Omega} \leq \widetilde{c}_{\mathbf{S}} g_2 |\Omega|^{1/2}\right\} \quad \text{and} \quad W_h := \left\{\phi_h \in \mathrm{H}^{\phi}_h : \ \|\phi_h\|_{1,\Omega} \leq \widetilde{c}_{\mathbf{S}} g_2 |\Omega|^{1/2}\right\},$$

where $\widetilde{c}_{\mathbf{S}}$ is a constant depending only on data and other constants, we let $(\vec{\boldsymbol{\sigma}}, \widetilde{\underline{\boldsymbol{\sigma}}}) \in \mathbf{H}_1 \times \mathbf{H}_2$ with $\phi \in W$, and $(\vec{\boldsymbol{\sigma}}_h, \widetilde{\underline{\boldsymbol{\sigma}}}_h) \in \mathbf{H}_{1,h} \times \mathbf{H}_{2,h}$ with $\phi_h \in W_h$, be the solutions of the continuous and discrete formulations (3.3.7) and (3.3.8), respectively. Additionally, we recall from [84, Theorems 3.9 and 4.7] that the following a priori estimates hold

$$\begin{aligned} \|\widetilde{\underline{\boldsymbol{\sigma}}}\|_{\mathbf{H}_2} &\leq \widetilde{c}_{\mathbf{S}} g_2 |\Omega|^{1/2}, \qquad \|\vec{\boldsymbol{\sigma}}\|_{\mathbf{H}_1} \leq c_{\mathbf{S}} \left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2}\right\}, \\ \|\widetilde{\underline{\boldsymbol{\sigma}}}_h\|_{\mathbf{H}_2} &\leq \widetilde{c}_{\mathbf{S}} g_2 |\Omega|^{1/2}, \qquad \|\vec{\boldsymbol{\sigma}}_h\|_{\mathbf{H}_1} \leq \widetilde{C} \left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2 |\Omega|^{1/2}\right\}. \end{aligned}$$

3.4 Residual-based a posteriori error estimators

The main goal of this section is to derive reliable and efficient residual-based a posteriori error estimators for the Galerkin schemes (3.3.4) and (3.3.8).

3.4.1 Preliminaries

Further notation is needed for describing local information on elements and edges. Given $K \in \mathcal{T}_h$, we let $\mathcal{E}_h(K)$ be the set of its edges, and let $\mathcal{E}_h$ be the set of all edges of the triangulation $\mathcal{T}_h$, whose corresponding diameters are denoted by h_e. Then, we write $\mathcal{E}_h = \mathcal{E}_h(\Omega) \cup \mathcal{E}_h(\Gamma)$, where $\mathcal{E}_h(\Omega) := \{e \in \mathcal{E}_h : e \subseteq \Omega\}$ and $\mathcal{E}_h(\Gamma) := \{e \in \mathcal{E}_h : e \subseteq \Gamma\}$. Also, for each edge e of $\mathcal{E}_h$ we fix a unit normal and tangential vectors $\boldsymbol{\nu}$ and $\boldsymbol{s}$ to e. Thus, the usual jump operator $[\![\cdot]\!]$ across an internal edge $e \in \mathcal{E}_h(\Omega)$ is defined for piecewise continuous tensor, vector, or scalar-valued functions $\boldsymbol{\zeta}$ as

$$[\![\boldsymbol{\zeta}]\!] = \boldsymbol{\zeta}|_{K_+} - \boldsymbol{\zeta}|_{K_-} ,$$

where K_- and K_+ are the triangles of $\mathcal{T}_h$ sharing the edge e. Additionally, given scalar, vector and tensor fields φ, $\boldsymbol{\varphi} := (\varphi_1, \varphi_2)$ and $\boldsymbol{\tau} := (\tau_{ij})$, respectively, we set

$$\mathbf{rot}(\varphi) := \begin{pmatrix} \frac{\partial \varphi}{\partial x_2} \\ -\frac{\partial \varphi}{\partial x_1} \end{pmatrix}, \quad \mathrm{rot}(\boldsymbol{\varphi}) := \frac{\partial \varphi_2}{\partial x_1} - \frac{\partial \varphi_1}{\partial x_2},$$

$$\underline{\mathbf{curl}}(\boldsymbol{\varphi}) := \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_2} & -\frac{\partial \varphi_1}{\partial x_1} \\ \frac{\partial \varphi_2}{\partial x_2} & -\frac{\partial \varphi_2}{\partial x_1} \end{pmatrix}, \quad \text{and} \quad \mathbf{curl}(\boldsymbol{\tau}) := \begin{pmatrix} \frac{\partial \tau_{12}}{\partial x_1} - \frac{\partial \tau_{11}}{\partial x_2} \\ \frac{\partial \tau_{22}}{\partial x_1} - \frac{\partial \tau_{21}}{\partial x_2} \end{pmatrix}.$$

Next, we collect a few preliminary definitions and results that we need in what follows. We begin by recalling the usual Clément interpolation operator (cf. [56]) $\mathrm{I}_h : \mathrm{H}^1(\Omega) \to \mathrm{X}_h$, where

$$\mathrm{X}_h := \left\{ \varphi_h \in \mathrm{C}(\overline{\Omega}) : \quad \varphi_h|_K \in \mathrm{P}_1(K) \quad \forall\, K \in \mathcal{T}_h \right\}.$$

A vectorial version of I_h, say $\mathbf{I}_h : \mathbf{H}^1(\Omega) \to \mathbf{X}_h := \mathrm{X}_h \times \mathrm{X}_h$, which is defined component-wise by I_h, will be needed as well. Moreover, to satisfy homogeneous Dirichlet boundary conditions, we introduce the Clément-type interpolation operator $\widetilde{\mathrm{I}}_h : \mathrm{H}^1_0(\Omega) \to \widetilde{\mathrm{X}}_h$, where

$$\widetilde{\mathrm{X}}_h := \left\{ \varphi_h \in \mathrm{C}(\overline{\Omega}) \cap \mathrm{H}^1_0(\Omega) : \quad \varphi_h|_K \in \mathrm{P}_1(K) \quad \forall\, K \in \mathcal{T}_h \right\}.$$

The following lemma provides the local approximation properties of I_h (for a proof, see [56]). Analogue estimates hold for the operators $\mathbf{I}_h$ and $\widetilde{\mathrm{I}}_h$.

Lemma 3.4.1. *There exist $c_1, c_2 > 0$, independent of h, such that for each $\varphi \in \mathrm{H}^1(\Omega)$, there holds*

$$\|\varphi - \mathrm{I}_h(\varphi)\|_{0,K} \leq c_1 h_K \, |\varphi|_{1,\Delta(K)} \quad \forall\, K \in \mathcal{T}_h, \tag{3.4.1}$$

and

$$\|\varphi - \mathrm{I}_h(\varphi)\|_{0,e} \leq c_2 h_e^{1/2} \, |\varphi|_{1,\Delta(e)} \quad \forall\, e \in \mathcal{E}_h, \tag{3.4.2}$$

where $\Delta(K) := \cup\{K' \in \mathcal{T}_h : \quad K' \cap K \neq 0\}$ and $\Delta(e) := \cup\{K' \in \mathcal{T}_h : \quad K' \cap e \neq 0\}$.

Moreover, we also introduce the usual Raviart-Thomas interpolator $\Pi_h : \mathbf{H}^1(\Omega) \to \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}$ [81, Section 3.4.1], which, given $\boldsymbol{\tau} \in \mathbf{H}^1(\Omega)$, is characterised by

$$\int_e \Pi_h(\boldsymbol{\tau}) \cdot \boldsymbol{\nu} \Psi = \int_e \boldsymbol{\tau} \cdot \boldsymbol{\nu} \Psi, \qquad \forall \, \text{edge } e \in \mathcal{T}_h, \ \forall \, \Psi \in \mathrm{P}_k(e), \tag{3.4.3}$$

$$\int_K \Pi_h(\boldsymbol{\tau}) \cdot \boldsymbol{\xi} = \int_K \boldsymbol{\tau} \cdot \boldsymbol{\xi} \quad \forall \, K \in \mathcal{T}_h, \qquad \forall \, \boldsymbol{\xi} \in \mathbf{P}_{k-1}(K), \ \text{ when } k \geq 1. \tag{3.4.4}$$

Additionally, using (3.4.3) and (3.4.4), the commuting diagram property yields

$$\operatorname{div}(\Pi_h(\boldsymbol{\tau})) = \mathcal{P}_h(\operatorname{div} \boldsymbol{\tau}) \quad \forall \, \boldsymbol{\tau} \in \mathbf{H}^1(\Omega), \tag{3.4.5}$$

where $\mathcal{P}_h$ is the $\mathrm{L}^2(\Omega)$-orthogonal projector onto the space of piecewise scalar polynomials of degree $\leq k$. Further approximations properties of Π_h are summarised in the following lemma (see a proof in e.g [81, Lemmas 3.16 and 3.18]).

Lemma 3.4.2. *There exist $C_1, C_2 > 0$, independent of h, such that for all $\boldsymbol{\tau} \in \mathbf{H}^1(\Omega)$, there hold*

$$\|\boldsymbol{\tau} - \Pi_h(\boldsymbol{\tau})\|_{0,K} \leq C_1 h_K \, \|\boldsymbol{\tau}\|_{1,K} \quad \forall \, K \in \mathcal{T}_h, \tag{3.4.6}$$

$$\|(\boldsymbol{\tau} - \Pi_h(\boldsymbol{\tau}))\boldsymbol{\nu}\|_{0,e} \leq C_2 h_e^{1/2} \, \|\boldsymbol{\tau}\|_{1,K_e} \quad \forall \, e \in \mathcal{E}_h, \tag{3.4.7}$$

where K_e in (3.4.7) is a triangle of $\mathcal{T}_h$ containing the edge e on its boundary.

A tensor version of Π_h, say $\boldsymbol{\Pi}_h : \mathbb{H}^1(\Omega) \to \mathbb{RT}_k$, (where $\mathbb{RT}_k$ is the space of pure Raviart-Thomas tensors of order k), which is defined row-wise by Π_h, and a vector version of $\mathcal{P}_h$, say, $\boldsymbol{\mathcal{P}}_h$ which is the $\mathbf{L}^2(\Omega)$-orthogonal projector onto $\mathbf{H}_h^{\boldsymbol{u}}$ (cf. (3.3.5), (3.3.6)), that is the space of piecewise vector valued polynomials of degree $\leq k$, might also be required. For simplicity of the presentation we have focused on the Raviart-Thomas interpolator. However, if we would like to use the family (3.3.6), we might use the BDM interpolator, which also satisfies the approximation properties given above.

In addition, we recall a Helmholtz decomposition of $\mathbb{H}_0(\mathbf{div}; \Omega)$, which will be essential in the subsequent analysis. We refer to [89, Lemma 3.7] for further details.

Lemma 3.4.3. *For each $\boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega)$, there exist $\boldsymbol{z} \in \mathbf{H}^2(\Omega)$ and $\boldsymbol{\Phi} \in \mathbf{H}^1(\Omega)$, such that*

$$\boldsymbol{\tau} = \nabla \boldsymbol{z} + \underline{\mathbf{curl}} \, \boldsymbol{\Phi} \ \text{ in } \ \Omega, \quad \text{and} \quad \|\boldsymbol{z}\|_{2,\Omega} + \|\boldsymbol{\Phi}\|_{1,\Omega} \leq C \, \|\boldsymbol{\tau}\|_{\mathbf{div};\Omega} \, . \tag{3.4.8}$$

On the other hand, the main techniques involved below in the proof of efficiency include the localisation technique based on element-bubble and edge-bubble functions. In view of this, we let ψ_e and ψ_K be the usual edge-bubble and face-bubble functions (see [148]), respectively, which satisfy $\psi_e|_K \in \mathrm{P}_2(K)$, $\operatorname{supp} \psi_e \subseteq \omega_e := \cup \{K' \in \mathcal{T}_h : e \in \mathcal{E}_h(K')\}$, $\psi_e = 0$ on $\partial K \backslash e$ and $0 \leq \psi_e \leq 1$ in ω_e, and $\psi_K \in \mathrm{P}_3(K)$, $\operatorname{supp} \psi_K \subseteq K$, $\psi_K = 0$ on ∂K and $0 \leq \psi_K \leq 1$ in K, respectively. We also recall from [147] that, given $k \in \mathbb{N} \cup \{0\}$, there exists an extension operator $\mathrm{L} : C(e) \to C(K)$ satisfying $\mathrm{L}(p) \in \mathrm{P}_k(K)$ and $\mathrm{L}(p)|_e = p \ \ \forall \, p \in \mathrm{P}_k(e)$. A corresponding vector version of L, that is the component-wise application of L, is denote by $\mathbf{L}$. Additional properties of ψ_e, ψ_K and L are collected in the following lemma (see e.g. [148]).

Lemma 3.4.4. *Given $k \in \mathbb{N} \cup \{0\}$, there exist positive constants c_3, c_4 and c_5, depending only on k, and the shape regularity of the triangulations (minimum angle condition), such that for each $K \in \mathcal{T}_h$ and $e \in \mathcal{E}_h(K)$, there hold*

$$\|q\|_{0,K}^2 \leq c_3 \|\psi_K^{1/2} q\|_{0,K}^2 \quad \forall\, q \in \mathrm{P}_k(K), \tag{3.4.9}$$

$$\|p\|_{0,e}^2 \leq c_4 \|\psi_e^{1/2} p\|_{0,e}^2 \quad \forall\, p \in \mathrm{P}_k(e), \tag{3.4.10}$$

$$\|\psi_e\, \mathrm{L}(p)\|_{0,K}^2 \leq \|\psi_e^{1/2}\, \mathrm{L}(p)\|_{0,K}^2 \leq c_5\, h_e\, \|p\|_{0,e}^2 \quad \forall\, p \in \mathrm{P}_k(e). \tag{3.4.11}$$

Furthermore, we will also need the following inverse estimate (cf. [55, Theorem 3.2.6]) and discrete trace inequality (cf. [3, Theorem 3.10]), respectively.

Lemma 3.4.5. *Let k, l, $m \in \mathbb{N} \cup \{0\}$ such that $l \leq m$. Then, there exists $c_6 > 0$, depending only on k, l, m and the shape regularity of the triangulations, such that for each $K \in \mathcal{T}_h$, there holds*

$$|q|_{m,K} \leq c_6\, h_K^{l-m} |q|_{l,K} \quad \forall\, q \in \mathrm{P}_k(K).$$

Lemma 3.4.6. *There exists $c_7 > 0$, depending only on the shape regularity of the triangulations, such that for each $K \in \mathcal{T}_h$ and $e \in \mathcal{E}_h(K)$, there holds*

$$\|v\|_{0,e}^2 \leq c_7 \left\{ h_e^{-1} \|v\|_{0,K}^2 + h_e |v|_{1,K}^2 \right\} \quad \forall\, v \in \mathrm{H}^1(K). \tag{3.4.12}$$

Finally, the following lemma is applied next to the terms involving the **curl** and **rot** operators, and the tangential jumps across the edges of $\mathcal{T}_h$. Its proof, which makes use of Lemmas 3.4.4 and 3.4.6, can be found in [28, Lemmas 4.3 and 4.4].

Lemma 3.4.7. *Let $\boldsymbol{\xi}_h \in \mathbb{L}^2(\Omega)$ be a piecewise polynomial tensor of degree $k \geq 0$ on each $K \in \mathcal{T}_h$, and let $\boldsymbol{\xi} \in \mathbb{L}^2(\Omega)$ be such that $\mathbf{curl}(\boldsymbol{\xi}) = \mathbf{0}$ in Ω. Then, there exist $c_8, c_9 > 0$, independent of h, such that*

$$\|\mathbf{curl}(\boldsymbol{\xi}_h)\|_{0,K} \leq c_8\, h_K^{-1}\, \|\boldsymbol{\xi} - \boldsymbol{\xi}_h\|_{0,K} \quad \forall\, K \in \mathcal{T}_h,$$

$$\|[\![\boldsymbol{\xi}_h \boldsymbol{s}]\!]\|_{0,e} \leq c_9\, h_e^{-1/2}\, \|\boldsymbol{\xi} - \boldsymbol{\xi}_h\|_{0,\omega_e} \quad \forall\, e \in \mathcal{E}_h(\Omega).$$

3.4.2 A posteriori error analysis for the mixed-primal scheme

In this section we derive a reliable and efficient residual-based a posteriori error estimator for (3.3.4). We begin by defining for each $K \in \mathcal{T}_h$ the local error indicator $\boldsymbol{\Theta}_K := \boldsymbol{\Theta}_{E,K} + \boldsymbol{\Theta}_{D,K}$, where $\boldsymbol{\Theta}_{E,K}$ and $\boldsymbol{\Theta}_{D,K}$ are the corresponding quantities associated with the elasticity and diffusion equations, respectively, which are given by:

$$\begin{aligned}
\boldsymbol{\Theta}_{E,K}^2 := {} & \|\boldsymbol{f}(\phi_h) + \mathbf{div}\, \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,K}^2 + h_K^2 \left\|\nabla \boldsymbol{u}_h - \left(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}\right)\right\|_{0,K}^2 \\
& + h_K^2 \left\|\mathbf{curl}\left(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h\right)\right\|_{0,K}^2 + \sum_{e \in \mathcal{E}_h(\Omega) \cap \mathcal{E}_h(K)} h_e \left\|[\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!]\right\|_{0,e}^2 \\
& + \sum_{e \in \mathcal{E}_h(\Gamma) \cap \mathcal{E}_h(K)} h_e \left\{ \left\| \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s} \right\|_{0,e}^2 + \|\boldsymbol{u}_{\mathrm{D}} - \boldsymbol{u}_h\|_{0,e}^2 \right\},
\end{aligned} \tag{3.4.13}$$

and

$$\boldsymbol{\Theta}_{D,K}^2 := h_K^2 \left\| \mathrm{div}\left(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h\right) + g(\boldsymbol{u}_h)\right\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h(\Omega)\cap\mathcal{E}_h(K)} h_e \left\|[\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h\cdot\boldsymbol{\nu}]\!]\right\|_{0,e}^2, \tag{3.4.14}$$

where

$$c := \frac{1}{2|\Omega|}\int_\Gamma \boldsymbol{u}_{\mathrm{D}}\cdot\boldsymbol{\nu}. \tag{3.4.15}$$

We remark in advance that the above requires that $\frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} \in \mathbf{L}^2(e)$ for each $e\in\mathcal{E}_h(\Gamma)$. This is fixed below assuming that $\boldsymbol{u}_{\mathrm{D}}\in\mathbf{H}^1(\Gamma)$. Finally, we point out that the residual character of each term defining $\boldsymbol{\Theta}_{E,K}$, $\boldsymbol{\Theta}_{D,K}$, and hence $\boldsymbol{\Theta}_K$, is clear, and then, proceeding as usual, the global residual estimator can be defined as:

$$\boldsymbol{\Theta} := \left\{\sum_{K\in\mathcal{T}_h}\boldsymbol{\Theta}_K^2\right\}^{1/2}.$$

The remainder of this section advocates to show the existence of positive constants C_{eff} and C_{rel}, independent of the meshsizes and the continuous and discrete solutions, such that

$$C_{\mathrm{eff}}\boldsymbol{\Theta} \le \|(\vec{\boldsymbol{\sigma}},\phi)-(\vec{\boldsymbol{\sigma}}_h,\phi_h)\| \le C_{\mathrm{rel}}\boldsymbol{\Theta}. \tag{3.4.16}$$

The efficiency of the global a posteriori error estimator (lower bound in (3.4.16)) is proved below in Section 3.4.2, whereas the corresponding reliability (upper bound in (3.4.16)) is derived in Section 3.4.2.

In order to establish the reliability of the a posteriori error estimator $\boldsymbol{\Theta}$, we apply the global inf-sup condition and the uniform ellipticity of some bilinear forms, together with smallness-of-data assumptions.

We begin with a preliminary estimate for the partial elasticity error $\|\vec{\boldsymbol{\sigma}}-\vec{\boldsymbol{\sigma}}_h\|_{\mathbf{H}_1}$.

Lemma 3.4.8. *There exists $C_1>0$, independent of λ and h, such that*

$$\begin{aligned}\|\vec{\boldsymbol{\sigma}}-\vec{\boldsymbol{\sigma}}_h\|_{\mathbf{H}_1} \le C_1\Big\{&\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'} + L_f\,\|\phi-\phi_h\|_{1,\Omega}\\ &+\|f(\phi_h)+\mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega} + \|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega}\Big\},\end{aligned} \tag{3.4.17}$$

where the functional $\mathcal{R}^E$ is defined by

$$\mathcal{R}^E(\boldsymbol{\tau}) := G(\boldsymbol{\tau}) - a(\boldsymbol{\sigma}_h,\boldsymbol{\tau}) - b(\boldsymbol{\tau},(\boldsymbol{u}_h,\boldsymbol{\rho}_h)) \quad \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega). \tag{3.4.18}$$

Furthermore, there holds

$$\mathcal{R}^E(\boldsymbol{\tau}_h) = 0 \quad \forall\,\boldsymbol{\tau}_h\in\mathbb{H}_h^{\boldsymbol{\sigma}}. \tag{3.4.19}$$

Proof. We begin the derivation of (3.4.17) by recalling from [81, Section 2.4.3.1], that b satisfies the inf-sup condition and that a is elliptic in the kernel of b. Then, there exists $C>0$, independent of h, such that for each $\vec{\boldsymbol{\xi}} := (\boldsymbol{\xi},\boldsymbol{w},\boldsymbol{\zeta})\in\mathbf{H}_1$, the following global inf-sup condition holds (see [75, Proposition 2.36])

$$C\|\vec{\boldsymbol{\xi}}\|_{\mathbf{H}_1} \le \sup_{\substack{\vec{\boldsymbol{\tau}}\in\mathbf{H}_1\\ \vec{\boldsymbol{\tau}}\neq\mathbf{0}}} \frac{a(\boldsymbol{\xi},\boldsymbol{\tau})+b(\boldsymbol{\tau},(\boldsymbol{w},\boldsymbol{\zeta}))+b(\boldsymbol{\xi},(\boldsymbol{v},\boldsymbol{\eta}))}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}_1}}.$$

In particular, for the error $\vec{\boldsymbol{\xi}} := \vec{\boldsymbol{\sigma}} - \vec{\boldsymbol{\sigma}}_h$, using the notation introduced by (3.4.18), and applying some algebraic manipulations, we have

$$\begin{aligned} C\|\vec{\boldsymbol{\sigma}} - \vec{\boldsymbol{\sigma}}_h\|_{\mathbf{H}_1} &\leq \sup_{\substack{\vec{\boldsymbol{\tau}}\in\mathbf{H}_1 \\ \vec{\boldsymbol{\tau}}\neq\vec{\mathbf{0}}}} \frac{a(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,\boldsymbol{\tau}) + b(\boldsymbol{\tau},(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{\rho}-\boldsymbol{\rho}_h)) + b(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,(\boldsymbol{v},\boldsymbol{\eta}))}{\|\vec{\boldsymbol{\tau}}\|_{\mathbf{H}_1}} \\ &\leq \sup_{\substack{\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega) \\ \boldsymbol{\tau}\neq\mathbf{0}}} \frac{|\mathcal{R}^E(\boldsymbol{\tau})|}{\|\boldsymbol{\tau}\|} + \sup_{\substack{(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega) \\ (\boldsymbol{v},\boldsymbol{\eta})\neq\mathbf{0}}} \frac{|b(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,(\boldsymbol{v},\boldsymbol{\eta}))|}{\|(\boldsymbol{v},\boldsymbol{\eta})\|}. \end{aligned} \tag{3.4.20}$$

Now, according to the definition of the bilinear form b, adding and subtracting suitable terms, and then, applying the Lipschitz continuity of $\boldsymbol{f}$ (cf. (3.2.3)), the Cauchy-Schwarz inequality, and the fact that $\int_\Omega \boldsymbol{\sigma}_h : \boldsymbol{\eta} = \frac{1}{2}\int_\Omega (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}) : \boldsymbol{\eta}$, we get for all $(\boldsymbol{v},\boldsymbol{\eta}) \in \mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathrm{skew}}(\Omega)$

$$|b(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,(\boldsymbol{v},\boldsymbol{\eta}))| \leq \widetilde{C}\Big\{ L_f\,\|\phi-\phi_h\|_{1,\Omega} + \|f(\phi_h) + \mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega} + \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega}\Big\}\,\|(\boldsymbol{v},\boldsymbol{\eta})\|\,,$$

which, together with (3.4.20), yields (3.4.17). Finally, it is readily seen that (3.4.19) follows directly from the first row of (3.3.4) and (3.4.18). □

We now derive an analogous preliminary bound for the error associated with $\|\phi-\phi_h\|_{1,\Omega}$.

Lemma 3.4.9. *There exists $C_2 > 0$, independent of h, such that*

$$\begin{aligned} \|\phi-\phi_h\|_{1,\Omega} \leq C_2 \Big\{ &\|\mathcal{R}^D\|_{\mathrm{H}^1_0(\Omega)'} + L_\vartheta\, c_{\mathbf{S}} \Big(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\Big)\,\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega} \\ &+ \vartheta_2\,\|\phi-\phi_h\|_{1,\Omega} + L_g\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}\Big\}\,, \end{aligned} \tag{3.4.21}$$

where the functional $\mathcal{R}^D$ is defined by

$$\mathcal{R}^D(\psi) := G_{\boldsymbol{u}_h}(\psi) \;-\; A_{\boldsymbol{\sigma}_h}(\phi_h,\psi) \quad \forall\,\psi\in\mathrm{H}^1_0(\Omega). \tag{3.4.22}$$

Furthermore, there holds

$$\mathcal{R}^D(\psi_h) = 0 \quad \forall\,\psi_h\in\mathrm{H}^\phi_h. \tag{3.4.23}$$

Proof. Similarly to the proof of Lemma 3.4.8, we first observe from the $\mathrm{H}^1_0(\Omega)$-ellipticity of $A_{\boldsymbol{\sigma}}$ (cf. [83, Lemma 2.3]) that the global inf-sup condition holds

$$\alpha\,\|\varphi\|_{1,\Omega} \leq \sup_{\substack{\psi\in\mathrm{H}^1_0(\Omega) \\ \psi\neq 0}} \frac{A_{\boldsymbol{\sigma}}(\varphi,\psi)}{\|\psi\|_{1,\Omega}} \quad \forall\,\varphi\in\mathrm{H}^1_0(\Omega), \tag{3.4.24}$$

where α is the ellipticity constant of $A_{\boldsymbol{\sigma}}$ [83, eq. (2.18)]. Then, applying (3.4.24) to the error $\varphi := \phi-\phi_h$, bearing in mind the definition (3.4.22), and adding and subtracting suitable terms, we find that

$$\alpha\,\|\phi-\phi_h\|_{1,\Omega} \leq \sup_{\substack{\psi\in\mathrm{H}^1_0(\Omega) \\ \psi\neq 0}} \frac{\mathcal{R}^D(\psi) + A_{\boldsymbol{\sigma}_h}(\phi_h,\psi) - A_{\boldsymbol{\sigma}}(\phi_h,\psi) + G_{\boldsymbol{u}}(\psi) - G_{\boldsymbol{u}_h}(\psi)}{\|\psi\|_{1,\Omega}}. \tag{3.4.25}$$

Now, recalling from [83, Section 2] that there exists a constant $C_\infty > 0$, such that the following estimate for the solution of the diffusion problem $\phi\in\mathrm{H}^1_0(\Omega)$ holds

$$\|\phi\|_{1,\infty,\Omega} \leq C_\infty\, c_{\mathbf{S}} \Big(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\Big)\,,$$

we can deduce the following result

$$\begin{aligned}
&|A_{\boldsymbol{\sigma}_h}(\phi_h,\psi)-A_{\boldsymbol{\sigma}}(\phi_h,\psi)| \\
&\quad\leq \left\{\|\phi\|_{1,\infty,\Omega}\, L_\vartheta\, \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}+2\vartheta_2\,\|\phi-\phi_h\|_{1,\Omega}\right\}\|\psi\|_{1,\Omega}\,, \qquad (3.4.26)\\
&\quad\leq \left\{C_\infty\, L_\vartheta c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}+2\vartheta_2\,\|\phi-\phi_h\|_{1,\Omega}\right\}\|\psi\|_{1,\Omega}\,.
\end{aligned}$$

Moreover, applying the Lipschitz continuity of g (cf. (3.2.4)), we get

$$|G_{\boldsymbol{u}}(\psi)-G_{\boldsymbol{u}_h}(\psi)|\leq L_g\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}\,\|\psi\|_{0,\Omega}\,. \qquad (3.4.27)$$

Thus, replacing (3.4.26) and (3.4.27) back into (3.4.25) we obtain (3.4.21). Finally, using the fact that $G_{\boldsymbol{u}_h}(\psi_h)-A_{\boldsymbol{\sigma}_h}(\phi_h,\psi_h)=0\quad\forall\,\psi_h\in\mathrm{H}^{\phi}_h$, we get (3.4.23) and the proof concludes. □

Consequently, we can establish the following preliminary upper bound for the total error.

Theorem 3.4.10. *Assume that*

$$C_1L_f+C_2\left\{\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}+\vartheta_2+L_g\right\}<\frac{1}{2}. \qquad (3.4.28)$$

Then, there exists $C_3>0$, independent of λ and h, such that the total error satisfies

$$\begin{aligned}
&\|(\vec{\boldsymbol{\sigma}},\phi)-(\vec{\boldsymbol{\sigma}}_h,\phi_h)\| \\
&\quad\leq C_3\left\{\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'}+\|f(\phi_h)+\mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega}+\|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega}+\|\mathcal{R}^D\|_{\mathrm{H}^1_0(\Omega)'}\right\}. \qquad (3.4.29)
\end{aligned}$$

Proof. It follows as a straightforward application of (3.4.28) and Lemmas 3.4.8 and 3.4.9. □

It is clear from (3.4.29) that, in order to obtain an explicit estimate for the total error, it only remains to derive suitable upper bounds for $\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'}$ and $\|\mathcal{R}^D\|_{\mathrm{H}^1_0(\Omega)'}$. This is precisely the purpose of the next subsection.

Reliability

With the aim of estimating $\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'}$ we now take an arbitrary $\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega)$ and consider the Helmholtz decomposition provided by (3.4.8) (cf. Lemma 3.4.3). Then, we denote $\boldsymbol{\Phi}_h:=\mathbf{I}_h(\boldsymbol{\Phi})$ and define $\boldsymbol{\tau}_h:=\boldsymbol{\Pi}_h(\nabla\boldsymbol{z})+\underline{\mathbf{curl}}(\boldsymbol{\Phi}_h)-d_h\,\mathbb{I}\in\mathbb{RT}_k$, with $\boldsymbol{\Pi}_h$ the interpolator operator defined in Section 3.4.1, and where according to [82, Section 4.1], the constant d_h, which is defined by

$$d_h:=-\frac{1}{2|\Omega|}\int_\Omega \mathrm{tr}(\nabla\boldsymbol{z}-\boldsymbol{\Pi}_h(\nabla\boldsymbol{z})+\underline{\mathbf{curl}}(\boldsymbol{\Phi}-\boldsymbol{\Phi}_h)), \qquad (3.4.30)$$

is chosen so that $\boldsymbol{\tau}_h$ belongs to $\mathbb{H}^{\boldsymbol{\sigma}}_h$ (cf. (3.3.5)). It follows that $\boldsymbol{\tau}-\boldsymbol{\tau}_h=\nabla\boldsymbol{z}-\boldsymbol{\Pi}_h(\nabla\boldsymbol{z})+\underline{\mathbf{curl}}(\boldsymbol{\Phi}-\boldsymbol{\Phi}_h)+d_h\,\mathbb{I}$, and then, applying the tensor version of (3.4.5), we get

$$\mathbf{div}(\boldsymbol{\tau}-\boldsymbol{\tau}_h)=\mathbf{div}(\nabla\boldsymbol{z}-\boldsymbol{\Pi}_h(\nabla\boldsymbol{z}))=(\mathbb{I}-\boldsymbol{\mathcal{P}}_h)(\mathbf{div}\,\nabla\boldsymbol{z})=(\mathbb{I}-\boldsymbol{\mathcal{P}}_h)(\mathbf{div}\,\boldsymbol{\tau}),$$

which is $\mathbf{L}^2(\Omega)$-orthogonal to $\mathbf{H}^{\boldsymbol{u}}_h$, and hence,

$$\int_\Omega \boldsymbol{u}_h\cdot\mathbf{div}(\boldsymbol{\tau}-\boldsymbol{\tau}_h)=\int_\Omega \boldsymbol{u}_h\cdot(\mathbb{I}-\boldsymbol{\mathcal{P}}_h)(\mathbf{div}\,\boldsymbol{\tau})=0. \qquad (3.4.31)$$

Furthermore, taking into account that $\boldsymbol{\sigma}_h \in \mathbb{H}_h^{\boldsymbol{\sigma}}$ and $\boldsymbol{\rho}_h \in \mathbb{H}_h^{\boldsymbol{\rho}}$, and recalling that c and d_h are given by (3.4.15) and (3.4.30), respectively, we deduce from the definition of $\mathcal{R}^E$ (cf. (3.4.18)) that

$$\begin{aligned}\mathcal{R}^E(d_h\mathbb{I}) &= d_h \int_\Gamma \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{\nu} = -c \int_\Omega \mathrm{tr}(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z}) + \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)) \\ &= -\int_\Omega c\mathbb{I} \colon (\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z}) + \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)),\end{aligned} \tag{3.4.32}$$

where for the second row in (3.4.32) we have applied the equality $\mathrm{tr}(\boldsymbol{\xi}) = \boldsymbol{\xi} \colon \mathbb{I}$. Thus, applying the null property (3.4.19), we find that

$$\mathcal{R}^E(\boldsymbol{\tau}) = \mathcal{R}^E(\boldsymbol{\tau} - \boldsymbol{\tau}_h) = \mathcal{R}^E\big(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z})\big) + \mathcal{R}^E\big(\underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)\big) + \mathcal{R}^E(d_h\mathbb{I}),$$

from which, replacing the last adding by (3.4.32), recalling the definition of $\mathcal{R}^E$ (cf. (3.4.18)), and employing the identity (3.4.31), we deduce that $\mathcal{R}^E(\boldsymbol{\tau})$ can be decomposed as

$$\mathcal{R}^E(\boldsymbol{\tau}) = \mathcal{R}^E(\boldsymbol{\tau} - \boldsymbol{\tau}_h) = \mathcal{R}_1^E(\boldsymbol{z}) + \mathcal{R}_2^E(\boldsymbol{\Phi}) \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}_0(\mathbf{div}; \Omega)\,, \tag{3.4.33}$$

where

$$\begin{aligned}\mathcal{R}_1^E(\boldsymbol{z}) &:= \mathcal{R}^E\big(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z})\big) - \int_\Omega c\mathbb{I} \colon \big(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z})\big) \\ &= \langle \big(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z})\big)\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma - \int_\Omega (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) \colon \big(\nabla \boldsymbol{z} - \boldsymbol{\Pi}_h(\nabla \boldsymbol{z})\big),\end{aligned} \tag{3.4.34}$$

and

$$\begin{aligned}\mathcal{R}_2^E(\boldsymbol{\Phi}) &:= \mathcal{R}^E(\underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)) - \int_\Omega c\mathbb{I} \colon \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h) \\ &= \langle \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma - \int_\Omega (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) \colon \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h).\end{aligned} \tag{3.4.35}$$

The following two lemmas provide upper bounds for (3.4.34) and (3.4.35).

Lemma 3.4.11. *There exists $C_4 > 0$, independent of λ and h, such that*

$$|\mathcal{R}_1^E(\boldsymbol{z})| \leq C_4 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \left\| \nabla \boldsymbol{u}_h - \big(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}\big) \right\|_{0,K}^2 + \sum_{e \in \mathcal{E}_h(\Gamma)} h_e \left\| \boldsymbol{u}_{\mathrm{D}} - \boldsymbol{u}_h \right\|_{0,e}^2 \right\}^{1/2} \|\boldsymbol{\tau}\|_{\mathbf{div};\Omega}\,.$$

Proof. It follows from an application of the tensor version of properties (3.4.3) and (3.4.4) to $\boldsymbol{u}_h|_e \in \mathbf{P}_k(e)$ for each $e \in \mathcal{E}_h$ and $\nabla \boldsymbol{u}_h|_K \in \mathbf{P}_{k-1}(K)$ for each $K \in \mathcal{T}_h$, respectively, and approximation results (3.4.6) and (3.4.7), in conjunction with the continuous dependence given by the Helmholtz decomposition (cf. (3.4.8)). We omit further details and refer to [82, Lemma 4.4]. □

Lemma 3.4.12. *If $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^1(\Gamma)$, then there exists $C_5 > 0$, independent of λ and h, such that*

$$|\mathcal{R}_2^E(\boldsymbol{\Phi})| \leq C_5 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \left\| \mathbf{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h) \right\|_{0,K}^2 + \sum_{e \in \mathcal{E}_h(\Omega)} h_e \left\| [\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!] \right\|_{0,e}^2 \right.$$
$$\left. + \sum_{e \in \mathcal{E}_h(\Gamma)} h_e \left\| \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s} \right\|_{0,e}^2 \right\}^{1/2} \|\boldsymbol{\tau}\|_{\mathbf{div};\Omega} . \tag{3.4.36}$$

Proof. We begin by applying the result given by [89, Lemma 3.8], to obtain

$$\langle \underline{\mathbf{curl}}\,(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma = - \left\langle \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}}, \boldsymbol{\Phi} - \boldsymbol{\Phi}_h \right\rangle_\Gamma = - \sum_{e \in \mathcal{E}_h(\Gamma)} \int_e (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h) \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} . \tag{3.4.37}$$

In turn, integrating by parts the second term on the right-hand side of (3.4.35), we get

$$\int_\Omega (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) : \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h) = \sum_{K \in \mathcal{T}_h} \int_K \mathbf{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)$$
$$- \sum_{e \in \mathcal{E}_h(\Omega)} \int_e [\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!] \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h) - \sum_{e \in \mathcal{E}_h(\Gamma)} \int_e (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s} \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h),$$

which together with (3.4.37) yields

$$\langle \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma - \int_\Omega (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) : \underline{\mathbf{curl}}(\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)$$
$$= - \sum_{K \in \mathcal{T}_h} \int_K \mathbf{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}) \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h) + \sum_{e \in \mathcal{E}_h(\Omega)} \int_e [\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!] \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h)$$
$$- \sum_{e \in \mathcal{E}_h(\Gamma)} \int_e \left\{ \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s} \right\} \cdot (\boldsymbol{\Phi} - \boldsymbol{\Phi}_h).$$

Finally, employing the Cauchy-Schwarz inequality, the vector version of estimates (3.4.1) and (3.4.2), the fact that $\Delta(K)$ and $\Delta(e)$ are bounded, and the continuous dependence (3.4.8), we obtain (3.4.36). □

With the above two results, and bearing in mind the decomposition (3.4.33), we are in a position to complete an upper bound for $\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'}$.

Lemma 3.4.13. *Assume that $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^1(\Gamma)$. Then, there exists $\widehat{C}_1 > 0$, independent of λ and h, such that*

$$\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'} \leq \widehat{C}_1 \left\{ \sum_{K \in \mathcal{T}_h} h_K^2 \left\| \nabla \boldsymbol{u}_h - \left(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}\right) \right\|_{0,K}^2 + \sum_{K \in \mathcal{T}_h} h_K^2 \left\| \mathbf{curl}(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h) \right\|_{0,K}^2 \right.$$
$$+ \sum_{e \in \mathcal{E}_h(\Gamma)} h_e \left(\left\| \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s} \right\|_{0,e}^2 + \|\boldsymbol{u}_{\mathrm{D}} - \boldsymbol{u}_h\|_{0,e}^2 \right)$$
$$\left. + \sum_{e \in \mathcal{E}_h(\Omega)} h_e \left\| [\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!] \right\|_{0,e}^2 \right\}^{1/2} .$$

In turn, we now provide an upper bound for $\|\mathcal{R}^D\|_{\mathrm{H}_0^1(\Omega)'}$.

Lemma 3.4.14. *There exists a constant $\widehat{C}_2 > 0$, independent of h, such that*

$$\|\mathcal{R}^D\|_{\mathrm{H}_0^1(\Omega)'} \leq \widehat{C}_2 \left\{ \sum_{K\in\mathcal{T}_h} h_K^2 \, \|\mathrm{div}\,(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h) + g(\boldsymbol{u}_h)\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h(\Omega)} h_e \, \|[\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!]\|_{0,e}^2 \right\}^{1/2} . \tag{3.4.38}$$

Proof. Given $\psi \in \mathrm{H}_0^1(\Omega)$, we let $\Psi_h := \widetilde{\mathrm{I}}_h(\psi) \in \mathrm{H}_h^\phi$. Thus, recalling the null property (3.4.23), the definition of the involved residual (cf. (3.4.22)), and integrating by parts, we obtain

$$\begin{aligned}\mathcal{R}^D(\psi - \Psi_h) &= \int_\Omega g(\boldsymbol{u}_h)(\psi - \Psi_h) - \int_\Omega \vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \nabla(\psi - \Psi_h) \\ &= \sum_{K\in\mathcal{T}_h} \int_K \Big\{ g(\boldsymbol{u}_h) + \mathrm{div}\,(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h) \Big\}(\psi - \Psi_h) - \sum_{e\in\mathcal{E}_h(\Omega)} \int_e [\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!](\psi - \Psi_h).\end{aligned}$$

Finally, applying the Cauchy-Schwarz inequality and the estimates given by Lemma 3.4.1, we deduce the estimate

$$|\mathcal{R}^D(\psi - \Psi_h)| \leq \widehat{C}_2 \left\{ \sum_{K\in\mathcal{T}_h} h_K^2 \, \|g(\boldsymbol{u}_h) + \mathrm{div}\,(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h)\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h(\Omega)} h_e \, \|[\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!]\|_{0,e}^2 \right\}^{1/2} \|\psi\|_{1,\Omega} ,$$

which yields (3.4.38), concluding the proof. □

Finally, we point out that the reliability of the operator $\boldsymbol{\Theta}$ (cf. upper bound in (3.4.16)) essentially follows from Theorem 3.4.10, and Lemmas 3.4.13 and 3.4.14.

Efficiency

The goal of this section is to show the efficiency of our a posteriori error estimator $\boldsymbol{\Theta}$. In other words, we now provide upper bounds depending on the actual errors for the nine terms defining the local indicator $\boldsymbol{\Theta}_K$. We begin by establishing the main result of this section.

Theorem 3.4.15. *There exists $C_{\mathrm{eff}} > 0$, independent of λ and h, such that*

$$C_{\mathrm{eff}}\boldsymbol{\Theta} \leq \|(\vec{\boldsymbol{\sigma}}, \phi) - (\vec{\boldsymbol{\sigma}}_h, \phi_h)\| . \tag{3.4.39}$$

Throughout this section, as well as Section 3.4.3, we assume for simplicity that the nonlinear functions $\boldsymbol{f}$, g and ϑ are such that $\boldsymbol{f}(\phi_h)$, $g(\boldsymbol{u}_h)$ and $\vartheta(\boldsymbol{\sigma}_h)$, are all piecewise polynomials. The same is assumed for the data $\boldsymbol{u}_{\mathrm{D}}$. If this is not the case, but $\boldsymbol{f}$, g, ϑ and $\boldsymbol{u}_{\mathrm{D}}$ are sufficiently smooth, higher

order terms given by the errors arising from suitable polynomial approximations would appear in the right-hand side of (3.4.39), (3.4.63) and (3.4.67).

In order to prove (3.4.39), in the rest of this section we derive suitable upper bounds for the terms defining the local error indicator $\boldsymbol{\Theta}_K$ (cf. (3.4.13) - (3.4.14)). We begin by observing, thanks to the fact that $-\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}(\phi)$ in Ω, that there hold

$$\begin{aligned} \|\boldsymbol{f}(\phi_h) + \mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,K}^2 &\leq 2\,\|\boldsymbol{f}(\phi) - \boldsymbol{f}(\phi_h)\|_{0,K}^2 + 2\,\|\mathbf{div}\,(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{0,K}^2 \\ &\leq 2L_f^2\,\|\phi - \phi_h\|_{0,K}^2 + 2\,\|\mathbf{div}\,(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{0,K}^2\,, \end{aligned} \tag{3.4.40}$$

and

$$\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,K}^2 \leq 4\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2\,. \tag{3.4.41}$$

The following lemmas provide the corresponding upper bounds for the remaining estimates required to obtain the efficiency of $\boldsymbol{\Theta}$.

Lemma 3.4.16. *There exist $C_3, C_4 > 0$, independent of λ and h, such that*

$$h_K^2\,\left\|\mathbf{curl}\,(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h)\right\|_{0,K}^2 \leq C_3\left\{\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2\right\} \quad \forall\, K \in \mathcal{T}_h, \tag{3.4.42}$$

and

$$h_e\,\left\|[\![(\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}]\!]\right\|_{0,e}^2 \leq C_4\left\{\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,\omega_e}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\omega_e}^2\right\} \quad \forall\, e \in \mathcal{E}_h(\Omega).$$

Proof. It suffices to apply Lemma 3.4.7 with $\boldsymbol{\xi}_h := \mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I}$ and $\boldsymbol{\xi} := \mathcal{C}^{-1}\boldsymbol{\sigma} + \boldsymbol{\rho} + c\mathbb{I}$. □

Lemma 3.4.17. *There exists $C_5 > 0$, independent of λ and h, such that for each $K \in \mathcal{T}_h$, there holds*

$$h_K^2\,\left\|\nabla\boldsymbol{u}_h - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\right\|_{0,K}^2 \leq C_5\left\{\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K}^2 + h_K^2\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + h_K^2\,\|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K}^2\right\}.$$

Proof. It follows from an application of (3.4.9) with $\boldsymbol{q} := \nabla\boldsymbol{u}_h - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})$, the estimate (3.3.3) and then, the use of Lemma 3.4.5. We refer to [82, Lemma 4.12] and [47, Lemma 6.6] for further details. □

Lemma 3.4.18. *There exists $C_6 > 0$, independent of λ and h, such that for each $e \in \mathcal{E}_h(\Gamma)$, there holds*

$$h_e\left\|\frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - (\mathcal{C}^{-1}\boldsymbol{\sigma}_h + \boldsymbol{\rho}_h + c\mathbb{I})\boldsymbol{s}\right\|_{0,e}^2 \leq C_6\left\{\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K_e}^2 + \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,K_e}^2\right\},$$

where K_e is the triangle of $\mathcal{T}_h$ having e as an edge.

Proof. We begin by defining $\boldsymbol{\xi}$, $\boldsymbol{\xi}_h$ as in the proof of Lemma 3.4.16, and then, given $e \in \mathcal{E}_h(\Gamma)$, we denote $\boldsymbol{\chi}_e := \frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} - \boldsymbol{\xi}_h\boldsymbol{s}$ on e. Thus, applying the inequality (3.4.10) to $\boldsymbol{\chi}_e$, the extension operator $\mathbf{L} : \mathbf{C}(e) \to \mathbf{C}(K)$ and the fact that $\frac{d\boldsymbol{u}_{\mathrm{D}}}{d\boldsymbol{s}} = \nabla\boldsymbol{u}\boldsymbol{s}$, we obtain

$$\|\boldsymbol{\chi}_e\|_{0,e}^2 \leq c_4\|\psi_e^{1/2}\boldsymbol{\chi}_e\|_{0,e}^2 = c_4\int_e \psi_e\,\boldsymbol{\chi}_e\cdot\left\{\frac{d\boldsymbol{u}_{\mathrm{D}}}{ds} - \boldsymbol{\xi}_h\boldsymbol{s}\right\} = c_4\int_{\partial K_e}\psi_e\,\mathbf{L}(\boldsymbol{\chi}_e)\cdot\Big\{(\nabla\boldsymbol{u} - \boldsymbol{\xi}_h)\boldsymbol{s}\Big\}.$$

Then, we integrate by parts and use that $\boldsymbol{\xi} = \nabla\boldsymbol{u}$ in Ω (cf. (3.3.3)), to obtain

$$\int_{\partial K_e}\psi_e\,\mathbf{L}(\boldsymbol{\chi}_e)\cdot\Big\{(\nabla\boldsymbol{u} - \boldsymbol{\xi}_h)\boldsymbol{s}\Big\} = \int_{K_e}(\boldsymbol{\xi} - \boldsymbol{\xi}_h) : \underline{\mathbf{curl}}(\psi_e\,\mathbf{L}(\boldsymbol{\chi}_e)) + \int_{K_e}\mathbf{curl}(\boldsymbol{\xi}_h)\cdot\psi_e\,\mathbf{L}(\boldsymbol{\chi}_e).$$

Finally, by exploiting the Cauchy-Schwarz inequality, Lemmas 3.4.5 and 3.4.7, and then, invoking the estimates (3.4.11) and (3.4.42), we obtain the desired result. □

Lemma 3.4.19. *There exists $C_7 > 0$, independent of λ and h, such that for each $e \in \mathcal{E}_h(\Gamma)$, there holds*

$$h_e \, \|\boldsymbol{u}_{\mathrm{D}} - \boldsymbol{u}_h\|^2_{0,e} \leq C_7 \left\{ \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{0,K_e} + h^2_{K_e} \, \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|^2_{0,K_e} + h^2_{K_e} \, \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|^2_{0,K_e} \right\}.$$

Proof. It follows from an application of the discrete trace inequality (3.4.12), the estimate (3.3.3) and the fact that $\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}}$ on Γ. We refer to [82, Lemma 4.14] for further details. □

Lemma 3.4.20. *There exists $C_8 > 0$, independent of h, such that for each $K \in \mathcal{T}_h$, there holds*

$$h^2_K \, \|\mathrm{div}\,(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h) + g(\boldsymbol{u}_h)\|^2_{0,K} \leq C_8 \left\{ h^2_K \, \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{0,K} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|^2_{0,K} + \|\phi - \phi_h\|^2_{1,K} \right\}.$$

Proof. Proceeding as in [25, Lemma 4.4], given $K \in \mathcal{T}_h$, we define

$$\chi_K := \mathrm{div}\,(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h) + g(\boldsymbol{u}_h) \ \text{ on } \ K.$$

Thus, applying (3.4.9) with $q = \chi_K$, using that $\mathrm{div}\,(\vartheta(\boldsymbol{\sigma})\nabla\phi) = -g(\boldsymbol{u})$ in Ω, and integrating by parts, we find that

$$\|\chi_K\|^2_{0,K} \leq c_3 \|\psi_K^{1/2}\chi_K\|^2_{0,K} = c_3 \int_K (g(\boldsymbol{u}_h) - g(\boldsymbol{u}))\psi_K\chi_K + \int_K (\vartheta(\boldsymbol{\sigma})\nabla\phi - \vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h) \cdot \nabla(\psi_K\chi_K).$$

Now, applying the Cauchy-Schwarz inequality, the Lipschitz continuity of g (cf. (3.2.4)) and the estimate (3.4.26), we deduce that there exists $\widetilde{C}_8 > 0$, depending only on data and other constants, all of them independent of h, such that

$$\|\chi_K\|^2_{0,K} \leq \widetilde{C}_8 \left\{ \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K} \, \|\psi_K\chi_K\|_{0,K} + \left(\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K} + \|\phi - \phi_h\|_{1,K} \right) |\psi_K\chi_K|_{1,K} \right\}.$$

Next, using the inverse inequality provided by Lemma 3.4.5 and the fact that $0 \leq \psi_K \leq 1$ in K, we find that

$$\|\chi_K\|^2_{0,K} \leq \widetilde{C}_8 \left\{ \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K} + c_6 \, h_K^{-1} \left(\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K} + \|\phi - \phi_h\|_{1,K} \right) \right\} \|\chi_K\|_{0,K},$$

which gives

$$h^2_K \, \|\chi_K\|^2_{0,K} \leq C_8 \left\{ h^2_K \, \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{0,K} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|^2_{0,K} + \|\phi - \phi_h\|^2_{0,K} \right\},$$

completing the proof. □

Lemma 3.4.21. *There exists $C_9 > 0$, independent of h, such that for each $e \in \mathcal{E}_h(\Omega)$, there holds*

$$h_e \, \|[\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!]\|^2_{0,e} \leq C_9 \sum_{K \subseteq \omega_e} \left\{ h^2_K \, \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{0,K} + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|^2_{0,K} + \|\phi - \phi_h\|^2_{1,K} \right\}, \tag{3.4.43}$$

where ω_e is the union of the two triangles in $\mathcal{T}_h$ having e as an edge.

Proof. Proceeding analogously as in the proof of [25, Lemma 4.5], given $e \in \mathcal{E}_h(\Omega)$, we define

$$\chi_e := [\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!] \ \text{ on } \ e.$$

Thus, we apply (3.4.10) with $p = \chi_e$, and the integration by parts formula on each $K \in \omega_e$, to obtain

$$\|\chi_e\|_{0,e}^2 \leq c_4 \|\psi_e^{1/2}\chi_e\|_{0,e}^2 = c_4 \int_e [\![\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}]\!]\psi_e \,\mathrm{L}(\chi_e) = c_4 \sum_{K\subseteq\omega_e} \int_{\partial K} \vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \boldsymbol{\nu}\,\psi_e \,\mathrm{L}(\chi_e)$$

$$= c_4 \sum_{K\subseteq\omega_e} \left\{ \int_K \vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h \cdot \nabla(\psi_e \,\mathrm{L}(\chi_e)) + \int_K \mathrm{div}(\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h)\psi_e \,\mathrm{L}(\chi_e) \right\}.$$

Next, using that $\mathrm{div}\,(\vartheta(\boldsymbol{\sigma})\nabla\phi) = -g(\boldsymbol{u})$ in Ω and then, integrating by parts once more, we get

$$\|\chi_e\|_{0,e}^2 \leq c_4 \sum_{K\subseteq\omega_e} \left\{ \int_K (\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h - \vartheta(\boldsymbol{\sigma})\nabla\phi) \cdot \nabla(\psi_e \,\mathrm{L}(\chi_e)) \right.$$

$$\left. + \int_K (g(\boldsymbol{u}) - g(\boldsymbol{u}_h))\psi_e \,\mathrm{L}(\chi_e) + \int_K (\mathrm{div}(\vartheta(\boldsymbol{\sigma}_h)) + g(\boldsymbol{u}_h))\psi_e \,\mathrm{L}(\chi_e) \right\}.$$

Then, employing the Cauchy-Schwarz inequality, the Lipschitz continuity of g, the inverse inequality provided by Lemma 3.4.5, the fact that $0 \leq \psi_e \leq 1$ in ω_e, and the estimate (3.4.11), we see that

$$\|\chi_e\|_{0,e}^2 \leq \widetilde{C}_9 \sum_{K\subseteq\omega_e} \left\{ h_K^{-1} \|\vartheta(\boldsymbol{\sigma}_h)\nabla\phi_h - \vartheta(\boldsymbol{\sigma})\nabla\phi\|_{0,K} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K} \right.$$

$$\left. + \|\mathrm{div}(\vartheta(\boldsymbol{\sigma}_h)) + g(\boldsymbol{u}_h)\|_{0,K} \right\} h_e^{1/2} \|\chi_e\|_{0,K},$$

from which, noting that $h_e \leq h_K$, applying the estimate (3.4.26) and performing simple algebraic manipulations, we deduce that there exists $\widehat{C}_9 > 0$, depending only on data and other constants, all of them independent of h, such that

$$h_e \|\chi_e\|_{0,e}^2 \leq \widehat{C}_9 \sum_{K\subseteq\omega_e} \left\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{0,K}^2 + \|\phi - \phi_h\|_{0,K}^2 + h_K^2 \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,K}^2 \right.$$

$$\left. + h_K^2 \|\mathrm{div}(\vartheta(\boldsymbol{\sigma}_h)) + g(\boldsymbol{u}_h)\|_{0,K}^2 \right\}. \tag{3.4.44}$$

Finally, (3.4.44) and the efficiency estimate given by Lemma 3.4.20 imply (3.4.43), completing the proof. □

We end this section by observing that the efficiency of the a posteriori error indicator $\boldsymbol{\Theta}$ follows straightforwardly from the estimates (3.4.40) and (3.4.41), and Lemmas 3.4.16 - 3.4.21.

3.4.3 A posteriori error analysis for the fully-mixed scheme

In this section we derive two reliable and efficient residual-based a posteriori error estimators for the Galerkin scheme (3.3.8). We introduce the global a posteriori error estimators

$$\widetilde{\boldsymbol{\Theta}} := \left\{ \sum_{K\in\mathcal{T}_h} \widetilde{\boldsymbol{\Theta}}_K^2 \right\}^{1/2} \quad \text{and} \quad \widehat{\boldsymbol{\Theta}} := \left\{ \sum_{K\in\mathcal{T}_h} \widehat{\boldsymbol{\Theta}}_K^2 \right\}^{1/2},$$

where we define for each $K \in \mathcal{T}_h$

$$\widetilde{\boldsymbol{\Theta}}_K^2 := \boldsymbol{\Theta}_{E,K}^2 + \|\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2 + \|g(\boldsymbol{u}_h) + \operatorname{div}\widetilde{\boldsymbol{\sigma}}_h\|_{0,K}^2 + \|\nabla\phi_h - \boldsymbol{t}_h\|_{0,K}^2,$$

$$\widehat{\boldsymbol{\Theta}}_K^2 := \widetilde{\boldsymbol{\Theta}}_K^2 + h_K^2\|\operatorname{rot}(\boldsymbol{t}_h)\|_{0,K}^2 + \sum_{e\in\mathcal{E}_h(K)} h_e\,\|[\![\boldsymbol{t}_h\cdot\boldsymbol{s}]\!]\|_{0,e}^2\,, \tag{3.4.45}$$

with $\boldsymbol{\Theta}_{E,K}^2$ defined by (3.4.13).

The main goal of this section is to establish, under suitable assumptions, the existence of positive constants $C_{\rm rel}, C_{\rm eff}$, $c_{\rm rel}$, $c_{\rm eff}$, independent of the meshsizes and the continuous and discrete solutions, such that

$$C_{\rm eff}\widetilde{\boldsymbol{\Theta}} \le \|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}}) - (\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\| \le C_{\rm rel}\widetilde{\boldsymbol{\Theta}}, \quad \text{and} \quad c_{\rm eff}\widehat{\boldsymbol{\Theta}} \le \|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}}) - (\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\| \le c_{\rm rel}\widehat{\boldsymbol{\Theta}}, \tag{3.4.46}$$

A general a posteriori error estimate

We now focus here on the mixed diffusion equation. Applying the uniform ellipticity of the bilinear form $A_{\boldsymbol{\sigma}}$, we conclude a preliminary upper bound for the total error under smallness-of-data assumptions. More precisely, we begin with the following auxiliary result.

Lemma 3.4.22. *There exists $C_2 > 0$, independent of h, such that*

$$\begin{aligned}\|\underline{\widetilde{\boldsymbol{\sigma}}} - \underline{\widetilde{\boldsymbol{\sigma}}}_h\|_{\mathbf{H}_2} \le C_2\big\{&\|\mathcal{R}^D\|_{\mathbf{H}(\operatorname{div};\Omega)'} + \|g(\boldsymbol{u}_h) + \operatorname{div}\widetilde{\boldsymbol{\sigma}}_h\|_{0,\Omega} + \|\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,\Omega}\\ &+ \|\nabla\phi_h - \boldsymbol{t}_h\|_{0,\Omega} + L_\vartheta\, c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\rm D}\|_{1/2,\Gamma} + f_2|\Omega|^{1/2}\right)\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}\\ &+ \vartheta_2\,\|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,\Omega} + L_g\,\|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}\big\},\end{aligned} \tag{3.4.47}$$

where the functional $\mathcal{R}^D$ is defined by

$$\mathcal{R}^D(\widetilde{\boldsymbol{\tau}}) := -\kappa_2\int_\Omega (g(\boldsymbol{u}_h) + \operatorname{div}\widetilde{\boldsymbol{\sigma}}_h)\cdot\operatorname{div}\widetilde{\boldsymbol{\tau}} - \int_\Omega \boldsymbol{t}_h\cdot\widetilde{\boldsymbol{\tau}} - \int_\Omega \phi_h\operatorname{div}\widetilde{\boldsymbol{\tau}} - \kappa_1\int_\Omega(\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h)\cdot\widetilde{\boldsymbol{\tau}}, \tag{3.4.48}$$

for each $\widetilde{\boldsymbol{\tau}} \in \mathbf{H}(\operatorname{div};\Omega)$. Furthermore, there holds

$$\mathcal{R}^D(\widetilde{\boldsymbol{\tau}}_h) = 0 \quad \forall\,\widetilde{\boldsymbol{\tau}}_h \in \mathbf{H}_h^{\widetilde{\boldsymbol{\sigma}}}. \tag{3.4.49}$$

Proof. We proceed similar as in the proof of Lemma 3.4.9, applying the global inf-sup condition to the error between $\underline{\widetilde{\boldsymbol{\sigma}}}$ and $\underline{\widetilde{\boldsymbol{\sigma}}}_h$, to obtain

$$\alpha\,\|\underline{\widetilde{\boldsymbol{\sigma}}} - \underline{\widetilde{\boldsymbol{\sigma}}}_h\|_{\mathbf{H}_2} \le \sup_{\substack{\underline{\widetilde{\boldsymbol{\tau}}}\in\mathbf{H}_2\\ \underline{\widetilde{\boldsymbol{\tau}}}\neq\mathbf{0}}} \frac{G_{\boldsymbol{u}}(\underline{\widetilde{\boldsymbol{\tau}}}) - A_{\boldsymbol{\sigma}}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}})}{\|\underline{\widetilde{\boldsymbol{\tau}}}\|_{\mathbf{H}_2}},$$

where α is the ellipticity constant of $A_{\boldsymbol{\sigma}}$ given in [84, eq. (3.18)]. Now, adding and subtracting terms appropriately, we can write

$$G_{\boldsymbol{u}}(\underline{\widetilde{\boldsymbol{\tau}}}) - A_{\boldsymbol{\sigma}}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}}) = G_{\boldsymbol{u}_h}(\underline{\widetilde{\boldsymbol{\tau}}}) - A_{\boldsymbol{\sigma}_h}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}}) + A_{\boldsymbol{\sigma}_h}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}}) - A_{\boldsymbol{\sigma}}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}}) + G_{\boldsymbol{u}}(\underline{\widetilde{\boldsymbol{\tau}}}) - G_{\boldsymbol{u}_h}(\underline{\widetilde{\boldsymbol{\tau}}}), \tag{3.4.50}$$

In this way, by using the definitions of $A_{\boldsymbol{\sigma}}$, $A_{\boldsymbol{\sigma}_h}$, $G_{\boldsymbol{u}}$, and $G_{\boldsymbol{u}_h}$, we notice that

$$|(A_{\boldsymbol{\sigma}_h}-A_{\boldsymbol{\sigma}})(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}})| \leq \widetilde{C}_2\left\{L_\vartheta c_{\mathbf{S}}(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2})\,\|\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\sigma}}_h\|_{0,\Omega} + \vartheta_2\,\|\boldsymbol{t}-\boldsymbol{t}_h\|_{0,\Omega}\right\}\|\underline{\widetilde{\boldsymbol{\tau}}}\|_{\mathbf{H}_2}\,, \tag{3.4.51}$$

$$|(G_{\boldsymbol{u}}-G_{\boldsymbol{u}_h})(\underline{\widetilde{\boldsymbol{\tau}}})| \leq \widehat{C}_2 L_g\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}\,\|\underline{\widetilde{\boldsymbol{\tau}}}\|_{\mathbf{H}_2}, \tag{3.4.52}$$

and

$$|G_{\boldsymbol{u}_h}(\underline{\widetilde{\boldsymbol{\tau}}})-A_{\boldsymbol{\sigma}_h}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}})| \leq \overline{C}_2\left\{|\mathcal{R}^D(\widetilde{\boldsymbol{\tau}})|+\|g(\boldsymbol{u}_h)+\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h\|_{0,\Omega}+\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,\Omega} + \|\nabla\phi_h-\boldsymbol{t}_h\|_{0,\Omega}\right\}\|\underline{\widetilde{\boldsymbol{\tau}}}\|_{\mathbf{H}_2}, \tag{3.4.53}$$

and then, the estimate (3.4.47) follows by replacing (3.4.51), (3.4.52) and (3.4.53) back into (3.4.50). Finally, using the fact that $G_{\boldsymbol{u}_h}(\underline{\widetilde{\boldsymbol{\tau}}}_h)-A_{\boldsymbol{\sigma}_h}(\underline{\widetilde{\boldsymbol{\sigma}}}_h,\underline{\widetilde{\boldsymbol{\tau}}}_h)=0 \quad \forall\,\underline{\widetilde{\boldsymbol{\tau}}}_h\in\mathbf{H}_{2,h}$, and taking in particular $\boldsymbol{s}_h=0$ and $\psi_h=0$, we arrive at (3.4.49), which completes the proof. □

Consequently, we can establish the following preliminary upper bound for the total error.

Theorem 3.4.23. *Assume that*

$$C_1L_f+C_2\left\{L_\vartheta c_{\mathbf{S}}\left(\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+f_2|\Omega|^{1/2}\right)+\vartheta_2+L_g\right\}<\frac{1}{2}. \tag{3.4.54}$$

Then, there exists $C_3>0$, independent of λ and h, such that the total error satisfies

$$\|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}})-(\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\| \leq C_3\left\{\|\mathcal{R}^E\|_{\mathbb{H}_0(\mathbf{div};\Omega)'}+\|f(\phi_h)+\mathbf{div}\,\boldsymbol{\sigma}_h\|_{0,\Omega}+\|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}_h^{\mathrm{t}}\|_{0,\Omega}+\|\mathcal{R}^D\|_{\mathbf{H}(\mathrm{div};\Omega)'}\right.$$
$$\left.+\|g(\boldsymbol{u}_h)+\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h\|_{0,\Omega}+\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,\Omega}+\|\nabla\phi_h-\boldsymbol{t}_h\|_{0,\Omega}\right\}.$$

Proof. It follows as a straightforward application of (3.4.54) and Lemmas 3.4.8 and 3.4.22. □

We end this section by rewriting equivalently the residual $\mathcal{R}^D$. In fact, given $\widetilde{\boldsymbol{\tau}}\in\mathbf{H}(\mathrm{div};\Omega)$, we apply integration by parts to the third term on the right-hand side of (3.4.48), to obtain

$$\mathcal{R}^D(\widetilde{\boldsymbol{\tau}})=-\kappa_2\int_\Omega(g(\boldsymbol{u}_h)+\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h)\cdot\mathrm{div}\,\widetilde{\boldsymbol{\tau}}+\int_\Omega(\nabla\phi_h-\boldsymbol{t}_h)\cdot\widetilde{\boldsymbol{\tau}}-\kappa_1\int_\Omega(\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h)\cdot\widetilde{\boldsymbol{\tau}}. \tag{3.4.55}$$

Reliability of the a posteriori error estimators

The main goal of this section is to establish an upper bound for the residual $\mathcal{R}^D$ in its respective norm. This task is actually performed in two different ways, which leads to the reliability of the a posteriori error estimators $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$. We begin with the upper bound for the first inequality in (3.4.46).

Lemma 3.4.24. *There exists a constant $C_{\mathrm{rel}}>0$, independent of λ and h, such that*

$$\|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}})-(\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\| \leq C_{\mathrm{rel}}\widetilde{\boldsymbol{\Theta}}.$$

Proof. The proof follows straightforwardly from the application of the Cauchy-Schwarz inequality to the residual $\mathcal{R}^D$ (cf. (3.4.55)), Lemma 3.4.13, and the definition of $\widetilde{\boldsymbol{\Theta}}$. □

In turn, we now aim at establishing an upper bound for the second inequality in (3.4.46). For that, we will apply the vector form of the Helmholtz decomposition in Lemma 3.4.8, to bound $\|\mathcal{R}^D\|_{\mathbf{H}(\mathrm{div};\Omega)'}$. In fact, given $\widetilde{\boldsymbol{\tau}} \in \mathbf{H}(\mathrm{div};\Omega)$, there exist $z \in \mathrm{H}^2(\Omega)$ and $\Phi \in \mathrm{H}^1(\Omega)$ such that

$$\widetilde{\boldsymbol{\tau}} = \nabla z + \mathbf{rot}\,\Phi \quad \in \quad \Omega, \quad \text{and} \quad \|z\|_{2,\Omega} + \|\Phi\|_{1,\Omega} \leq C\,\|\widetilde{\boldsymbol{\tau}}\|_{\mathrm{div};\Omega}\,, \tag{3.4.56}$$

and then, denoting $\Phi_h := \mathrm{I}_h(\Phi)$, we define $\widetilde{\boldsymbol{\tau}}_h := \Pi_h(\nabla z) + \mathbf{rot}(\Phi_h) \in \mathbf{H}^{\widetilde{\boldsymbol{\sigma}}}_h$. In this way, noticing from (3.4.49) that $\mathcal{R}^D(\widetilde{\boldsymbol{\tau}}_h) = 0$, it readily follows that $\mathcal{R}^D(\widetilde{\boldsymbol{\tau}})$ can be decomposed as

$$\mathcal{R}^D(\widetilde{\boldsymbol{\tau}}) = \mathcal{R}^D(\widetilde{\boldsymbol{\tau}} - \widetilde{\boldsymbol{\tau}}_h) = \mathcal{R}^D(\nabla z - \Pi_h(\nabla z)) + \mathcal{R}^D(\mathbf{rot}(\Phi - \Phi_h)). \tag{3.4.57}$$

In the following two lemmas, we provide upper bounds for the terms on the right-hand side of (3.4.57).

Lemma 3.4.25. *There exists $C > 0$, independent of h, such that for each $z \in \mathrm{H}^2(\Omega)$, there holds*

$$\begin{aligned} |\mathcal{R}^D(\nabla z - \Pi_h(\nabla z))| \leq C \Bigg\{ \sum_{K \in \mathcal{T}_h} \Big(& \|g(\boldsymbol{u}_h) + \mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h\|^2_{0,K} + h_K^2 \|\nabla \phi_h - \boldsymbol{t}_h\|^2_{0,K} \\ & + \; h_K^2 \|\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|^2_{0,K} \Big) \Bigg\}^{1/2} \|z\|_{2,\Omega}\,. \end{aligned} \tag{3.4.58}$$

Proof. Given $z \in \mathrm{H}^2(\Omega)$, we first notice, from the definition of $\mathcal{R}^D$ (cf. (3.4.55)), that there holds

$$\begin{aligned} \mathcal{R}^D(\nabla z - \Pi_h(\nabla z)) = & -\kappa_2 \int_\Omega (g(\boldsymbol{u}_h) + \mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h) \cdot \mathrm{div}\,(\nabla z - \Pi_h(\nabla z)) \\ & + \int_\Omega (\nabla \phi_h - \boldsymbol{t}_h) \cdot (\nabla z - \Pi_h(\nabla z)) - \kappa_1 \int_\Omega (\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h) \cdot (\nabla z - \Pi_h(\nabla z)). \end{aligned} \tag{3.4.59}$$

For the first term on the right-hand side of (3.4.59) we proceed as in [16, Lemma 3.10], whereas for the remaining terms, we simply apply the Cauchy-Schwarz inequality, and subsequently use the approximation properties of Π_h provided by Lemma 3.4.2. □

Lemma 3.4.26. *There exists $C > 0$, independent of h, such that for each $\Phi \in \mathrm{H}^1(\Omega)$, there holds*

$$\begin{aligned} & |\mathcal{R}^D(\mathbf{rot}(\Phi - \Phi_h))| \\ & \leq C \left\{ \sum_{K \in \mathcal{T}_h} \left(\|\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|^2_{0,K} + h_K^2 \|\mathrm{rot}(\boldsymbol{t}_h)\|^2_{0,K} + \sum_{e \in \mathcal{E}_h(K)} h_e\, \|[\![\boldsymbol{t}_h \cdot \boldsymbol{s}]\!]\|^2_{0,e} \right) \right\}^{1/2} \|\Phi\|_{1,\Omega}\,. \end{aligned}$$

Proof. Given $\Phi \in \mathrm{H}^1(\Omega)$, we notice from the original definition of $\mathcal{R}^D$ (cf. (3.4.48)) that there holds

$$\mathcal{R}^D(\mathbf{rot}(\Phi - \Phi_h)) = -\kappa_1 \int_\Omega (\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h) \cdot \mathbf{rot}(\Phi - \Phi_h) - \int_\Omega \boldsymbol{t}_h \cdot \mathbf{rot}(\Phi - \Phi_h). \tag{3.4.60}$$

For the first term, we proceed as in the proof of [89, Lemma 3.9], applying the boundedness of $\mathrm{I}_h : \mathrm{H}^1(\Omega) \to \mathrm{H}^1(\Omega)$ (cf. [75, Lemma 1.127]), and the Cauchy-Schwarz and triangle inequalities, to deduce that

$$\left| \kappa_1 \int_\Omega (\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h) \cdot \mathbf{rot}(\Phi - \Phi_h) \right| \leq C_1 \|\widetilde{\boldsymbol{\sigma}}_h - \vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,\Omega}\, \|\Phi\|_{1,\Omega}\,. \tag{3.4.61}$$

Now, for the second term, we proceed as in the proof of [49, Lemma 3.9], to obtain

$$\left|\int_\Omega \boldsymbol{t}_h\cdot\mathbf{rot}(\Phi-\Phi_h)\right| \leq C_2\left\{\sum_{K\in\mathcal{T}_h}\left(h_K^2\|\mathrm{rot}(\boldsymbol{t}_h)\|_{0,K}^2+\sum_{e\in\mathcal{E}_h(K)}h_e\,\|[\![\boldsymbol{t}_h\cdot\boldsymbol{s}]\!]\|_{0,e}^2\right)\right\}^{1/2}\|\Phi\|_{1,\Omega}\,. \quad (3.4.62)$$

Finally, the desired result follows by replacing (3.4.61) and (3.4.62) back into (3.4.60). □

As a consequence of Lemmas 3.4.25 and 3.4.26, the identity (3.4.57), and the stability result given by (3.4.56), we can deduce the required upper bound for $\|\mathcal{R}^D\|_{\mathbf{H}(\mathrm{div};\Omega)'}$, that is

$$\|\mathcal{R}^D\|_{\mathbf{H}(\mathrm{div};\Omega)'}\leq C\left\{\sum_{K\in\mathcal{T}_h}\Big(\|g(\boldsymbol{u}_h)+\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h\|_{0,K}^2+h_K^2\|\nabla\phi_h-\boldsymbol{t}_h\|_{0,K}^2+h_K^2\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2\right.$$
$$\left.+\,\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2+h_K^2\|\mathrm{rot}(\boldsymbol{t}_h)\|_{0,K}^2+\sum_{e\in\mathcal{E}_h(K)}h_e\,\|[\![\boldsymbol{t}_h\cdot\boldsymbol{s}]\!]\|_{0,e}^2\Big)\right\}^{1/2},$$

where C is a positive constant independent of h.

Finally, we point out that the existence of a constant $c_{\mathrm{rel}}>0$, such that

$$\|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}})-(\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\|\leq c_{\mathrm{rel}}\widehat{\boldsymbol{\Theta}},$$

follows from Theorem 3.4.23, and Lemmas 3.4.13, 3.4.25 and 3.4.26, after observing that, for sufficiently small elements, the terms $h_K^2\|\nabla\phi_h-\boldsymbol{t}_h\|_{0,K}^2$ and $h_K^2\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2$ in (3.4.58), are dominated by $\|\nabla\phi_h-\boldsymbol{t}_h\|_{0,K}^2$ and $\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2$, respectively.

Efficiency of the a posteriori error estimators

Let us begin with the efficiency estimate for $\widetilde{\boldsymbol{\Theta}}$.

Lemma 3.4.27. *There exists $C_{\mathrm{eff}}>0$, independent of λ and h, such that*

$$C_{\mathrm{eff}}\widetilde{\boldsymbol{\Theta}}\leq\|(\vec{\boldsymbol{\sigma}},\underline{\widetilde{\boldsymbol{\sigma}}})-(\vec{\boldsymbol{\sigma}}_h,\underline{\widetilde{\boldsymbol{\sigma}}}_h)\|. \quad (3.4.63)$$

Proof. We recall that $g(\boldsymbol{u})=-\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}$ in Ω. In this way, it is clear that

$$\|g(\boldsymbol{u}_h)+\mathrm{div}\,\widetilde{\boldsymbol{\sigma}}_h\|_{0,K}^2\leq 2\,\|\mathrm{div}(\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\sigma}}_h)\|_{0,K}^2+2L_g^2\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,K}^2\,. \quad (3.4.64)$$

Moreover, since $\widetilde{\boldsymbol{\sigma}}=\vartheta(\boldsymbol{\sigma})\boldsymbol{t}$ in Ω, applying the Lipschitz continuity of ϑ, the regularity estimate [83, eq. (2.23)], and the Cauchy-Schwarz inequality, we deduce

$$\|\widetilde{\boldsymbol{\sigma}}_h-\vartheta(\boldsymbol{\sigma}_h)\boldsymbol{t}_h\|_{0,K}^2\leq C\Big\{\|\widetilde{\boldsymbol{\sigma}}-\widetilde{\boldsymbol{\sigma}}_h\|_{0,K}^2+\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{0,K}^2+\|\boldsymbol{t}-\boldsymbol{t}_h\|_{0,K}^2\Big\}. \quad (3.4.65)$$

Additionally, since $\boldsymbol{t}=\nabla\phi$ in Ω, we get

$$\|\nabla\phi_h-\boldsymbol{t}_h\|_{0,K}^2\leq C_1\Big\{\|\phi-\phi_h\|_{1,K}^2+\|\boldsymbol{t}-\boldsymbol{t}_h\|_{0,K}^2\Big\}. \quad (3.4.66)$$

Finally, the result follows from the definition of $\widetilde{\boldsymbol{\Theta}}$, estimates (3.4.40), (3.4.41), (3.4.64), (3.4.65) and (3.4.66), and Lemmas 3.4.16 - 3.4.19. □

On the other hand, we derive the efficiency of the estimator $\widehat{\boldsymbol{\Theta}}$. The following lemma provides the required upper bounds for the second and third terms on the right-hand side of (3.4.45).

Lemma 3.4.28. *There exist $c_1, c_2 > 0$, independent of h, such that*

$$h_K^2 \, \|\mathrm{rot}(\boldsymbol{t}_h)\|_{0,K}^2 \leq c_1 \, \|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,K}^2 \quad \forall \, K \in \mathcal{T}_h,$$

$$h_e \, \|[\![\boldsymbol{t}_h \cdot \boldsymbol{s}]\!]\|_{0,e}^2 \leq c_2 \, \|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,\omega_e}^2 \quad \forall \, e \in \mathcal{E}_h.$$

Proof. For the first inequality, we simply apply the vector version of the first inequality in Lemma 3.4.7 with $\boldsymbol{\xi}_h = \boldsymbol{t}_h$ and $\boldsymbol{\xi} = \boldsymbol{t} = \nabla\phi$, whereas for the second one, we can follow the proof given by [28, Lemma 4.4]. We omit further details. □

Finally, as a consequence of the estimates (3.4.40), (3.4.41), (3.4.64), (3.4.65) and (3.4.66), and Lemmas 3.4.16 - 3.4.19, and 3.4.28, we are now in position to state the efficiency of $\widehat{\boldsymbol{\Theta}}$.

Lemma 3.4.29. *There exists a $c_{\mathrm{eff}} > 0$, independent of λ and h, such that*

$$c_{\mathrm{eff}} \widehat{\boldsymbol{\Theta}} \leq \|(\vec{\boldsymbol{\sigma}}, \underline{\widetilde{\boldsymbol{\sigma}}}) - (\vec{\boldsymbol{\sigma}}_h, \underline{\widetilde{\boldsymbol{\sigma}}}_h)\|. \tag{3.4.67}$$

3.5 Numerical results

In this section we present some numerical results illustrating the properties of the estimator introduced in Section 3.4 and showing the behaviour of the associated adaptive algorithm. The individual errors and rates of convergence of the unknowns will be computed as usual

$$\mathrm{e}(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}\,, \ \mathrm{e}(\boldsymbol{u}) = \|\boldsymbol{u} - \boldsymbol{u}_h\|_{0,\Omega}\,, \ \mathrm{e}(\boldsymbol{\rho}) = \|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{0,\Omega}\,, \ \mathrm{e}(\widetilde{\boldsymbol{\sigma}}) = \|\widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h\|_{\mathrm{div};\Omega}\,,$$

$$\mathrm{e}(\boldsymbol{t}) = \|\boldsymbol{t} - \boldsymbol{t}_h\|_{0,\Omega}\,, \ \mathrm{e}(\phi) = \|\phi - \phi_h\|_{1,\Omega}\,, \ r(\cdot) = \frac{\log(\mathrm{e}(\cdot)/\widehat{\mathrm{e}}(\cdot))}{\log(h/\widehat{h})},$$

where e and $\widehat{\mathrm{e}}$ denote errors computed on two consecutive meshes of sizes h and $\widehat{h}$. When the adaptive algorithm is applied, the expression $\log(h/\widehat{h})$ appearing in the computation of the above rates is replaced by $-0.5\log(N/\widehat{N})$, where N and $\widehat{N}$, denote the corresponding degrees of freedom of each triangulation. In addition, given the total errors

$$\mathrm{e}_1 = \big\{[\mathrm{e}(\boldsymbol{\sigma})]^2 + [\mathrm{e}(\boldsymbol{u})]^2 + [\mathrm{e}(\boldsymbol{\rho})]^2 + [\mathrm{e}(\phi)]^2\big\}\,, \quad \text{and} \quad \mathrm{e}_2 = \{e_1 + \mathrm{e}[(\boldsymbol{t})]^2 + \mathrm{e}[(\widetilde{\boldsymbol{\sigma}})]^2\},$$

the effectivity indexes associated with $\boldsymbol{\Theta}$, $\widetilde{\boldsymbol{\Theta}}$, and $\widehat{\boldsymbol{\Theta}}$ are defined, respectively, as

$$\mathrm{eff}(\boldsymbol{\Theta}) = \frac{\mathrm{e}_1}{\boldsymbol{\Theta}}, \quad \mathrm{eff}(\widetilde{\boldsymbol{\Theta}}) = \frac{\mathrm{e}_2}{\widetilde{\boldsymbol{\Theta}}}, \quad \text{and} \quad \mathrm{eff}(\widehat{\boldsymbol{\Theta}}) = \frac{\mathrm{e}_2}{\widehat{\boldsymbol{\Theta}}}.$$

The linearisation of the systems associated with the assembled forms of (3.3.4) and (3.3.8) is carried out by Newton's method. In turn, the solution of the resulting linear systems at each Newton step are conducted using the Multifrontal Massively Parallel Sparse direct Solver (MUMPS). In addition, the examples use a classical adaptive mesh refinement procedure based on the equi-distribution of the error indicators, where the diameter of each element in the new adapted mesh (contained in a generic

element K on the initial coarse mesh) is proportional to the diameter of the initial element times the ratio $\widehat{\boldsymbol{\Theta}}_h/\boldsymbol{\Theta}_K$, where $\widehat{\boldsymbol{\Theta}}_h$ is the mean value of a given indicator $\boldsymbol{\Theta}$ over the initial mesh (cf. [147]).

On the other hand, we recall that given the Young modulus E and the Poisson ratio ν of an isotropic linear elastic solid, the corresponding Lamé parameters are defined as $\lambda = E\nu(1+\nu)^{-1}(1-2\nu)^{-1}$ and $\mu = E/(2+2\nu)$. Thus, in the following examples, we will consider $E = 1.0e3$ and $\nu = 0.4$.

Moreover, we point out that given $D_0 = D_1 = 0.1$, the nonlinear functions

$$\vartheta(\boldsymbol{\sigma}) = (D_0 + D_1(1+|\boldsymbol{\sigma}|^2)^{-0.5})\,\mathbb{I}, \quad \boldsymbol{f}(\phi) = \begin{pmatrix} -\sin(\phi) \\ \cos(\phi) \end{pmatrix}, \quad \text{and} \quad g(\boldsymbol{u}) = 2 + \frac{1}{1+|\boldsymbol{u}|^2},$$

satisfying (3.2.2)-(3.2.4), will be used in the following computational tests, and remark that for the examples described below, the elasticity and diffusion equations are considered non-homogeneous and the extra source terms are chosen according to the given exact solutions. This treatment does not compromise the analysis, as the regularity of the exact solution provides sufficiently smooth right-hand sides, thus only requiring a slight modification of the functionals in the variational formulation.

Finally, for the nonlinear diffusivity, the parameters appearing in (3.2.2) are given by: $\vartheta_0 = D_0$, $\vartheta_2 = \sqrt{2}(D_0 + D_1)$, and then, according to [84, eq. (3.20)], the stabilisation parameters for the fully-mixed scheme (3.3.8) can be taken as $\kappa_1 = \vartheta_0/\vartheta_2$, $\kappa_2 = \vartheta_0/2\vartheta_2$ and $\kappa_3 = \vartheta_0/2$.

Example 1. In the first example, we consider the following exact solutions to (3.2.1):

$$\boldsymbol{u} = \frac{1}{\lambda}\begin{pmatrix} d_1\cos(\pi x_1)\sin(2\pi x_2) \\ -d_1\sin(\pi x_1)\cos(\pi x_2) \end{pmatrix}, \quad \phi = 1.0 - e^{-x_1(x_1-1)x_2(x_2-1)}, \tag{3.5.1}$$

defined on the unit square $\Omega = (0,1)^2$, satisfying the boundary conditions $\boldsymbol{u}_{\mathrm{D}} = \boldsymbol{u}$ on Γ and $\phi = 0$ on Γ. The involved coefficient in (3.5.1) is taken as $d_1 = 0.05$.

The manufactured solutions on the considered domain are smooth, and the a posteriori error indicators show effectivity indexes close to one. The results reported in Tables 3.1 and 3.2 indicate optimal convergence rates for the two lowest-order methods. Approximate solutions obtained after seven steps of uniform refinement are depicted in Figure 3.1.

Example 2. In our second example we design a mesh convergence test using a closed-form solution, and performing uniform and adaptive mesh refinements. Thus, we consider the same computational domain as the one given in Example 1, and propose the following exact solutions

$$\boldsymbol{u} = \frac{1}{\lambda}\begin{pmatrix} \frac{-d_1\sin(x_1)\cos(x_2)}{(x_2+0.02)^2+(x_1+0.02)^2} \\ -d_1\cos(x_1)\sin(2x_2) \end{pmatrix}, \quad \phi = \frac{x_1(x_1-1)x_2(x_2-1)}{(10x_1+0.1)^2}, \tag{3.5.2}$$

where the manufactured displacement is used as Dirichlet datum on Γ, and the involved coefficient in (3.5.2) is taken as in Example 1. Notice that the first component of the displacement, and the concentration, in (3.5.2), exhibit singularities just outside the domain, at $(0,0)$ and the line $x_1 = -0.01$, respectively, therefore, high gradients are also expected in the approximation of these fields, and optimal convergence is no longer evidenced under uniform mesh refinement (see second row of Tables 3.3 and 3.4). In Tables 3.3 and 3.4, we show the individual errors, the effectivity indexes and experimental

N	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e(ϕ)	$r(\phi)$	eff($\boldsymbol{\Theta}$)
Lowest-order mixed-primal method									
346	0.372	-	8.4e-6	-	3.3e-5	-	5.8e-2	-	1.00
1298	0.191	1.00	4.2e-6	1.03	1.6e-5	1.05	2.8e-2	1.10	0.99
5026	0.096	1.01	2.1e-6	1.01	8.4e-6	1.02	1.4e-2	1.01	0.99
19778	0.048	1.00	1.0e-6	1.01	4.2e-6	1.01	7.0e-3	1.01	0.99
78466	0.024	1.00	5.3e-7	1.00	2.1e-6	1.00	3.5e-3	1.00	0.99
312578	0.012	1.00	2.6e-7	1.00	1.0e-6	1.00	1.7e-3	1.00	0.99
Second-order mixed-primal method									
898	0.0825	-	1.7e-6	-	7.6e-6	-	8.2e-3	-	0.99
3458	0.0213	2.00	4.4e-7	2.01	1.9e-6	2.03	2.1e-3	2.02	0.98
13570	0.0053	2.01	1.1e-7	2.01	4.8e-7	2.02	5.2e-4	2.00	0.98
53762	0.0013	2.01	2.8e-8	2.01	1.2e-7	2.01	1.3e-4	2.01	0.98
214018	0.0003	2.00	7.0e-9	2.00	3.0e-8	2.00	3.2e-5	2.00	0.98

Table 3.1: Example 1: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity index for the first- and second-order mixed-primal finite element methods (table produced by the author).

Lowest-order augmented fully-mixed scheme														
N	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e($\boldsymbol{t}$)	$r(\boldsymbol{t})$	e($\widetilde{\boldsymbol{\sigma}}$)	$r(\widetilde{\boldsymbol{\sigma}})$	e(ϕ)	$r(\phi)$	eff($\widetilde{\boldsymbol{\Theta}}$)	eff($\widehat{\boldsymbol{\Theta}}$)
466	3.728	-	8.4e-5	-	3.3e-4	-	1.9e-2	-	4.5e-2	-	5.6e-2	-	1.001	1.00
1762	1.918	0.99	4.2e-5	1.02	1.6e-4	1.05	9.6e-3	1.02	2.3e-2	0.99	2.7e-2	1.08	0.999	0.999
6850	0.965	1.01	2.1e-5	1.01	8.4e-5	1.02	4.7e-3	1.03	1.1e-2	1.00	1.4e-2	0.99	0.999	0.998
27010	0.483	1.00	1.0e-5	1.00	4.2e-5	1.01	2.3e-3	1.01	5.8e-3	1.00	7.0e-3	0.99	0.999	0.998
107266	0.242	1.00	5.3e-6	1.00	2.1e-5	1.00	1.1e-3	1.00	2.9e-3	1.00	3.5e-3	1.00	0.999	0.998
427522	0.121	1.00	2.6e-6	1.00	1.0e-5	1.00	5.9e-4	1.00	1.4e-3	1.00	1.7e-3	1.00	0.999	0.998
Second-order augmented fully-mixed scheme														
1266	0.825	-	1.7e-5	-	7.6e-5	-	2.4e-3	-	6.6e-3	-	7.4e-3	-	1.000	0.997
4898	0.213	1.99	4.4e-6	2.00	1.9e-5	2.02	6.2e-4	2.04	1.7e-3	1.99	1.9e-3	1.96	0.999	0.996
19266	0.053	2.01	1.1e-6	2.01	4.8e-6	2.02	1.5e-4	2.02	4.3e-4	2.01	5.0e-4	1.98	0.999	0.996
76418	0.013	2.00	2.8e-7	2.00	1.2e-6	2.01	3.9e-5	2.01	1.0e-4	2.00	1.2e-4	1.98	0.999	0.996
304386	0.003	2.00	7.0e-8	2.00	3.0e-7	2.00	9.8e-6	2.00	2.7e-5	2.00	3.2e-5	1.99	0.999	0.996

Table 3.2: Example 1: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity indexes for the first- and second-order augmented fully-mixed finite element methods (table produced by the author).

rates of convergence for the uniform and adaptive refinements of the mixed-primal and augmented fully-mixed schemes. As expected, we observe that the errors decrease faster through the adaptive procedure, and that in each case, the effectivity indexes remain bounded, which confirms the reliability and efficiency of $\boldsymbol{\Theta}$, $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$, in the cases of non-smooth solutions. Moreover, although super-convergence of the concentration can be seen when the adaptive scheme is applied for the augmented fully-mixed system (see the last two blocks of Table 3.4), we notice from Figure 3.3(a) that the global rate of convergence remains optimal. Furthermore, it is important to remark that when the adaptive algorithms are applied, optimal convergence can be restored, as shown in the last block of Table 3.3 and

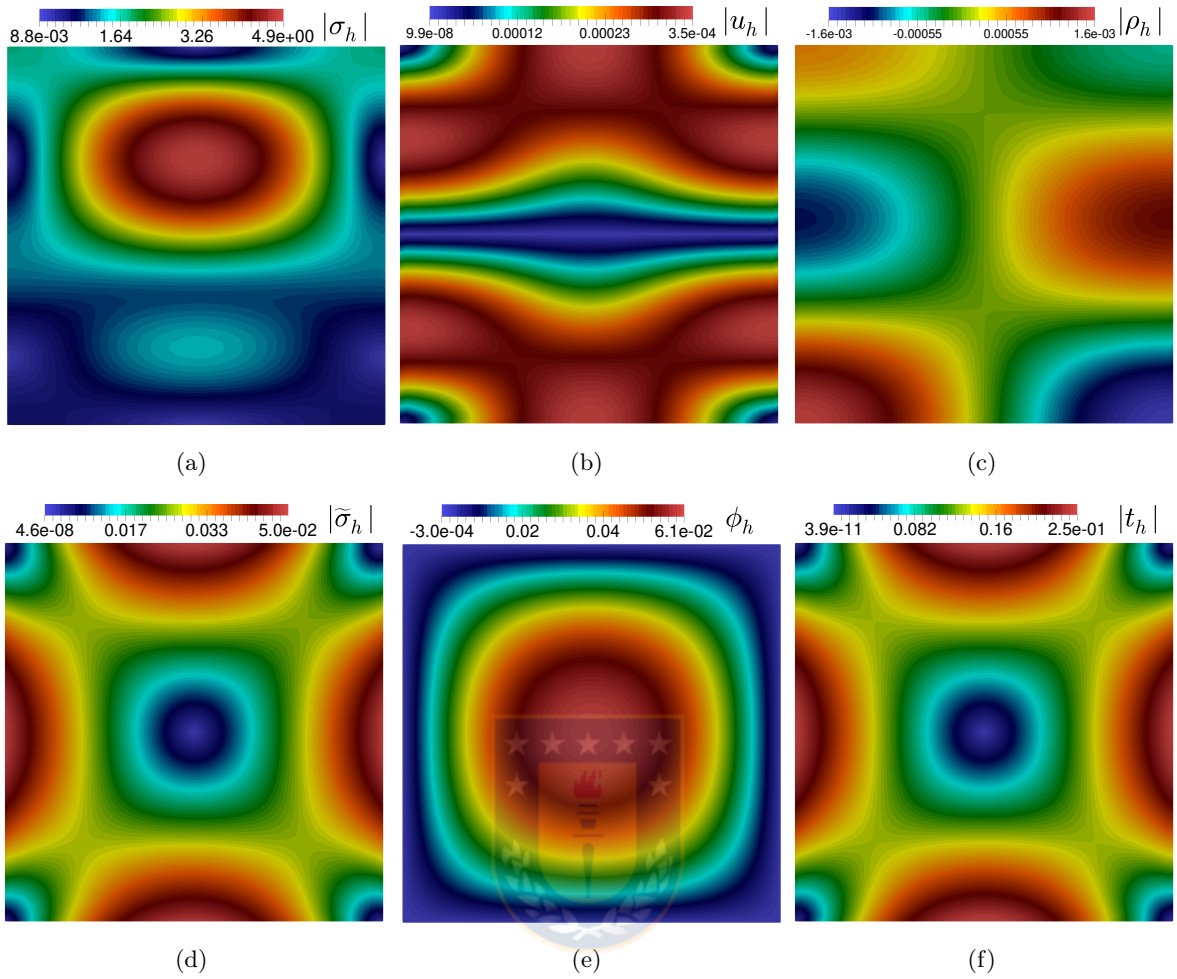


Figure 3.1: Example 1: Approximation of the stress magnitude $|\boldsymbol{\sigma}_h|$ (a), displacement magnitude $|\boldsymbol{u}_h|$ (b), rotation magnitude $|\boldsymbol{\rho}_h|$ (c), diffusive flux magnitude $|\widetilde{\boldsymbol{\sigma}}_h|$ (d), concentration of the diffusive substance ϕ_h (e), and concentration gradient magnitude $|\boldsymbol{t}_h|$ (f), by using the lowest-order augmented fully-mixed scheme with adaptive refinement according to $\widetilde{\boldsymbol{\Theta}}$ (figure produced by the author).

the last two blocks of Table 3.4. Additionally, we display in Figure 3.2 some adapted meshes obtained during the adaptive refinements according to $\boldsymbol{\Theta}$, $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$, and observe that they are concentrated around $(0,0)$ and the line $x_1 = -0.01$, which shows how the method is able to identify the regions in which the accuracy of the numerical approximation is deteriorated. Finally, approximation solutions are shown in Figure 3.3(b-e) after eight steps of adaptive refinement according to the indicator $\boldsymbol{\Theta}$.

N	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e(ϕ)	$r(\phi)$	eff($\boldsymbol{\Theta}$)
Lowest-order mixed-primal scheme upon uniform refinement									
346	136.2	-	3.4e-5	-	2.2e-4	-	1.54	-	1.13
1298	77.45	0.85	2.4e-5	0.50	2.6e-4	-0.23	1.10	0.50	1.06
5026	56.68	0.46	1.1e-5	1.19	1.8e-4	0.48	0.81	0.45	1.00
19778	42.11	0.43	4.3e-6	1.37	9.9e-5	0.93	0.66	0.30	0.99
78466	26.29	0.68	1.8e-6	1.21	4.3e-5	1.19	0.56	0.24	0.99
312578	14.25	0.88	8.8e-7	1.08	1.8e-5	1.24	0.41	0.42	0.99
Lowest-order mixed-primal scheme with adaptive refinement according to $\boldsymbol{\Theta}$									
346	136.2	-	3.4e-5	-	2.2e-4	-	1.54	-	1.13
898	77.45	1.18	2.3e-5	0.81	2.1e-4	0.46	1.10	0.71	1.06
2239	56.68	0.68	1.0e-5	1.76	1.5e-4	0.70	0.80	0.67	1.00
4985	42.06	0.74	4.5e-6	2.08	8.5e-5	1.56	0.65	0.52	0.99
10968	25.83	1.23	2.8e-6	1.15	4.0e-5	1.88	0.54	0.44	0.99
27366	13.32	1.44	1.7e-6	1.07	1.9e-5	1.62	0.40	0.66	0.99
77382	6.484	1.38	1.1e-6	0.75	9.4e-6	1.36	0.24	0.94	0.99
244093	3.190	1.23	6.5e-7	1.05	4.7e-6	1.21	0.13	1.08	0.99

Table 3.3: Example 2: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity index for the lowest-order mixed-primal finite element method (table produced by the author).

N	e($\boldsymbol{\sigma}$)	$r(\boldsymbol{\sigma})$	e($\boldsymbol{u}$)	$r(\boldsymbol{u})$	e($\boldsymbol{\rho}$)	$r(\boldsymbol{\rho})$	e($\widetilde{\boldsymbol{\sigma}}$)	$r(\widetilde{\boldsymbol{\sigma}})$	e($\boldsymbol{t}$)	$r(\boldsymbol{t})$	e(ϕ)	$r(\phi)$	eff($\widetilde{\boldsymbol{\Theta}}$)	eff($\widehat{\boldsymbol{\Theta}}$)
Lowest order augmented fully-mixed scheme upon uniform refinement														
466	136.2	-	3.2e-5	-	2.2e-4	-	130.1	-	17.2	-	48.1	-	1.14	1.11
1762	77.46	0.84	2.1e-5	0.64	2.5e-4	-0.22	91.73	0.52	6.85	1.39	17.1	1.55	1.09	1.08
6850	56.68	0.46	1.0e-5	1.08	1.8e-4	0.46	65.41	0.49	2.84	1.29	6.64	1.39	1.05	1.04
27010	42.11	0.43	4.1e-6	1.28	9.8e-5	0.92	48.19	0.44	1.53	0.89	3.16	1.08	1.02	1.02
107266	26.29	0.68	1.8e-6	1.16	4.3e-5	1.18	37.02	0.38	0.95	0.69	1.66	0.93	1.00	1.00
427522	14.25	0.88	8.9e-7	1.07	1.8e-5	1.24	26.99	0.45	0.56	0.74	0.79	1.07	1.00	0.99
Lowest order augmented fully-mixed scheme with adaptive refinement according to $\widetilde{\boldsymbol{\Theta}}$														
466	136.2	-	3.2e-5	-	2.2e-4	-	130.1	-	17.2	-	48.1	-	1.14	
1762	77.45	0.84	2.1e-5	0.58	2.1e-4	0.10	91.72	0.52	0.73	1.28	18.0	1.47	1.09	
6014	56.68	0.50	9.8e-6	1.29	1.5e-4	0.52	65.41	0.55	3.02	1.44	7.04	1.53	1.05	
13206	42.06	0.75	4.1e-6	2.20	8.5e-5	1.58	48.19	0.77	1.61	1.59	3.35	1.88	1.02	
24044	25.83	1.62	2.2e-6	2.10	4.0e-5	2.47	37.02	0.88	0.98	1.64	1.74	2.17	1.00	
48542	13.32	1.88	1.4e-6	1.25	1.9e-5	2.12	26.99	0.89	0.57	1.51	0.81	2.17	1.00	
127678	6.484	1.49	1.0e-6	0.68	9.4e-6	1.47	16.97	0.95	0.31	1.25	0.33	1.81	1.00	
423282	3.190	1.18	6.3e-7	0.78	4.6e-6	1.16	9.344	0.99	0.16	1.09	0.14	1.41	1.00	
Lowest order augmented fully-mixed scheme with adaptive refinement according to $\widehat{\boldsymbol{\Theta}}$														
466	136.2	-	3.2e-5	-	2.2e-4	-	130.1	-	17.2	-	48.1	-		1.11
1762	77.45	0.84	2.1e-5	0.58	2.1e-4	0.11	91.72	0.52	7.34	1.28	18.0	1.47		1.08
6708	56.68	0.46	9.8e-6	1.19	1.5e-4	0.48	65.41	0.50	3.02	1.32	7.04	1.40		1.05
14956	42.06	0.74	4.1e-6	2.18	8.5e-5	1.55	48.19	0.76	1.61	1.56	3.35	1.85		1.02
26122	25.83	1.74	2.1e-6	2.34	4.0e-5	2.66	37.02	0.94	0.98	1.76	1.74	2.34		1.00
52180	13.32	1.91	1.3e-6	1.37	1.9e-5	2.15	26.99	0.91	0.57	1.53	0.81	2.20		0.99
128846	6.484	1.59	9.1e-7	0.83	9.3e-6	1.58	16.97	1.02	0.31	1.33	0.33	1.94		0.99
431240	3.190	1.17	6.3e-7	0.60	4.6e-6	1.14	9.344	0.98	0.16	1.09	0.14	1.40		0.99

Table 3.4: Example 2: Degrees of freedom, individual absolute errors, rates of convergence, and effectivity indexes for the lowest-order augmented fully-mixed finite element method (table produced by the author).

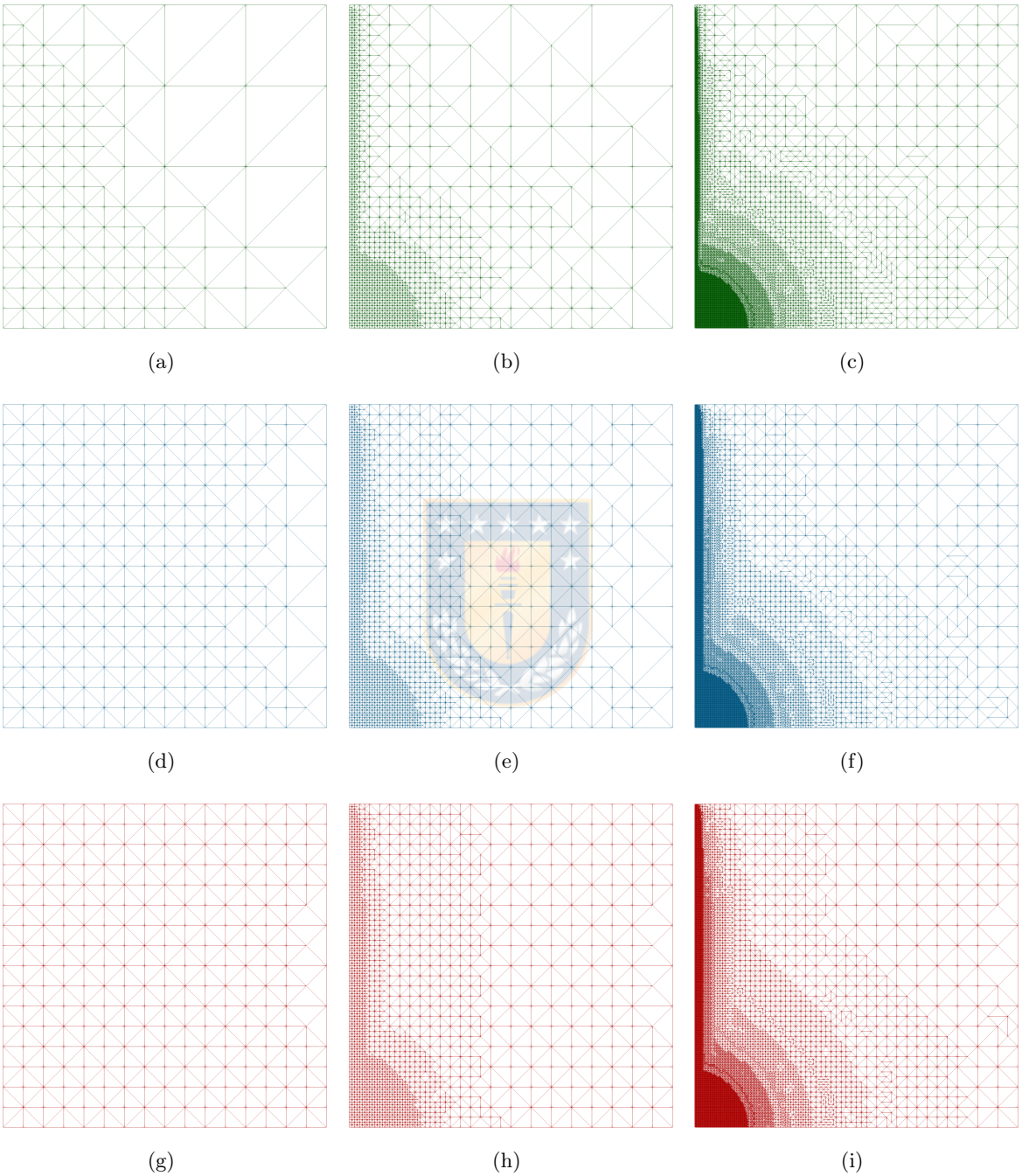


Figure 3.2: Example 2: From left to right, three snapshots of successively refined meshes according to the indicators $\boldsymbol{\Theta}$ (a,b,c), $\widetilde{\boldsymbol{\Theta}}$ (d,e,f), and $\widehat{\boldsymbol{\Theta}}$ (g,h,i) (figure produced by the author).

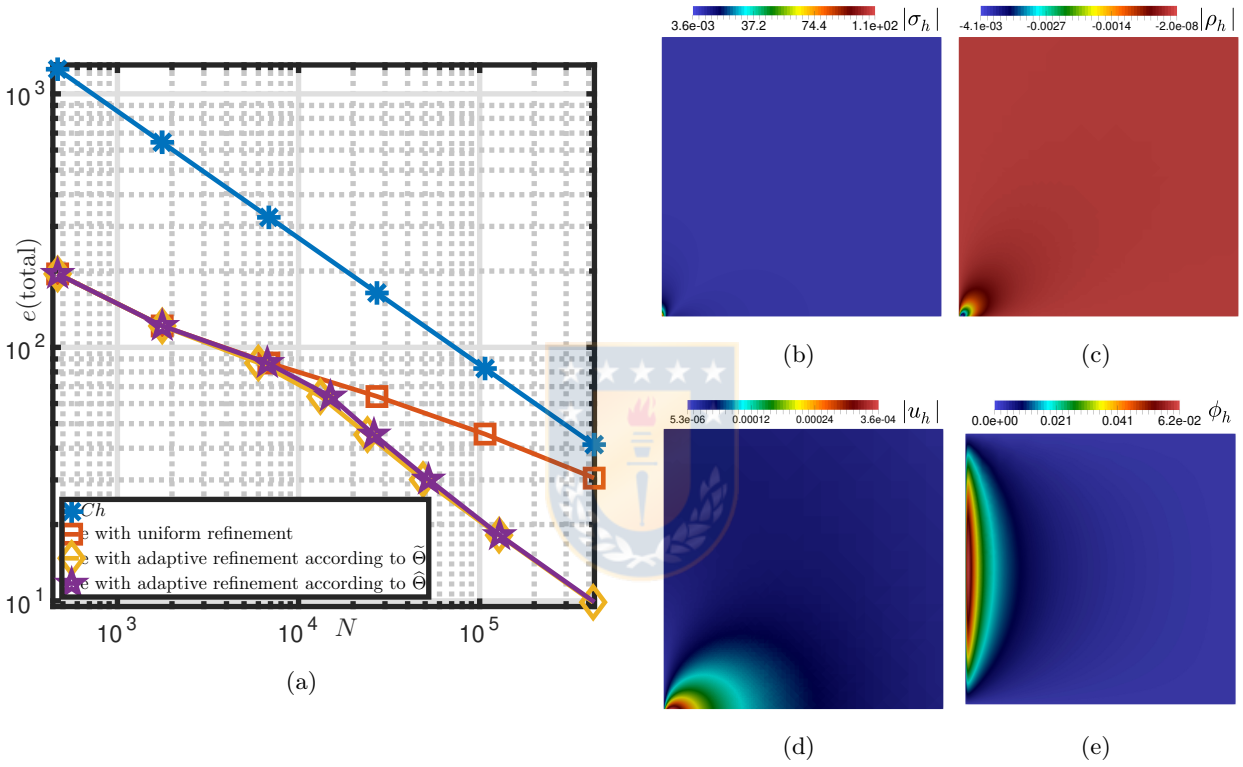


Figure 3.3: Example 2: Plot of the total error *versus* the number of degrees of freedom N associated with the uniform mesh refinement and adaptive algorithms according to $\widetilde{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Theta}}$ (a); and approximate stress magnitude (b), rotation magnitude (c), displacement magnitude (d), and solute concentration (e) computed using the lowest-order scheme where mesh adaptation is done via the estimator $\boldsymbol{\Theta}$ after eight steps of refinement (figure produced by the author).

CHAPTER 4

Stability and finite element approximation of phase change models for natural convection in porous media

4.1 Introduction

The phenomenon of natural convection driven by variations in temperature distribution has been extensively studied from the viewpoint of physical properties and also using computational methods. Common applications include ocean and atmosphere dynamics, design of double glass windows and ventilation devices. If the density of the fluid is approximately constant and the buoyancy contribution depends on temperature, the model equations consist of the so-called Boussinesq approximation [35]. Phenomena that involve phase change in addition to these elements have also a great relevance in many industrial and natural processes, as in e.g. the melting and solidification in the refining of metals.

As the nature of the physical scenario abruptly changes, modelling and computing formalisms usually have difficulty in reproducing the behaviour of the system especially near the liquid-solid interface. Phase changes have been incorporated into the Boussinesq approximation mainly using two different approaches. One is based on enthalpy-porosity models (as in e.g. [134]), where a jump function arising from the so-called Carman-Kozeny equations (see e.g. [45, 105]) enforces a large drag force in the solid regime. In other approaches, phase change has been modelled by embedding a jump function into the viscosity, as in e.g. [64]. One objective in the present work is to give a numerical comparison between these two models. Difficulties in incorporating phase change models are related to the choice of regularisation and jump size parameters. We address this issue with a new viscosity-based model that highlights an appropriate choice of parameters. This model considers the presence of microscopic particles in the solid, which resembles porosity-based models. We choose a transition from fluid to solid having a large gradient, which creates additional numerical challenges.

Recent numerical methods dedicated for phase change Boussinesq models include a class of stabilised discontinuous Galerkin and finite volume methods proposed for porosity-based models in [134] and [155], respectively; and the primal finite element scheme for viscosity-based models, introduced in [64]. However these contributions do not address the stability of the discrete or continuous problems. Theoretical studies are available for other (typically simpler) related stationary models, as natural convection [164, 115] including stabilisation analysis and errors estimates, and also time-dependent Boussinesq-type problems under different contexts in [4, 8, 29, 67, 135, 137]. The discretisation of these

problems has been associated with Taylor-Hood finite elements for mass and momentum equations and piecewise quadratic approximations for the temperature (as in [164, 115, 67, 123]), or Taylor-Hood and piecewise linear elements as in [64], the MINI-element and Lagrange elements [8] and also piecewise quadratic, piecewise linear and piecewise quadratic for velocity-pressure-temperature as in [4]. Exactly divergence-free methods are available for the stationary Boussinesq equations [122], and other related mixed formulations including a posteriori error estimates can be found in [14, 58, 61, 76, 31]. Finite volume, finite difference and Lagrangian schemes have also been used to simulate solidification problems [104, 95, 124, 18]. In summary, a large variety of methods could be employed to solve numerically the equations we look at here. However the stability of the numerical methods applied to this specific case has not yet been addressed, and this is precisely another objective of this chapter.

The remainder of this chapter is structured as follows. Section 4.2 recalls the model problems of flow and temperature with and without phase change. In Section 4.3 we derive a weak formulation of the governing equations and outline a solvability and stability analysis. In Section 4.4 we introduce two finite element methods based on the primal velocity-pressure-temperature formulation and on the mixed-primal stress-velocity-temperature formulation of the generalised Boussinesq equations. We specify the fully discrete implicit scheme and write down the corresponding Newton linearisation. Next, in Section 4.5 we present several tests serving as numerical validation for the enthalpy-free case, and we then present a set of comparisons and concluding insights drawn from the simulations of the melting case, collected in Section 4.6. These tests also include a qualitative analysis on the micro-structure and its relationship with our modelling assumptions.

4.2 Phase-change Boussinesq models

4.2.1 Main assumptions and model equations

The model problem arises from the description of flow (which has kinematic viscosity ν, thermal expansion coefficient α, and nondimensional specific heat C) using Navier-Stokes and Stefan problems. Applying the so-called Oberbeck-Boussinesq approximation, one ends up with the following set of governing equations written in terms of the velocity $\boldsymbol{u}(t) : \Omega \to \mathbb{R}^d$, the pressure $p(t) : \Omega \to \mathbb{R}$, and the temperature $\theta(t) : \Omega \to \mathbb{R}$:

$$\partial_t \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} - \frac{1}{\mathrm{Re}} \mathbf{div}\big[2\mu(\theta)\boldsymbol{\varepsilon}(\boldsymbol{u})\big] + \nabla p + \eta(\theta)\boldsymbol{u} = f(\theta)\boldsymbol{k}, \tag{4.2.1}$$

$$\mathrm{div}\, \boldsymbol{u} = 0, \qquad \text{in } \Omega \times (0, t_f], \tag{4.2.2}$$

$$\partial_t \theta + \boldsymbol{u} \cdot \nabla \theta - \frac{1}{C\mathrm{Pr}} \mathrm{div}(\kappa \nabla \theta) + \partial_t s + \boldsymbol{u} \cdot \nabla s = 0, \tag{4.2.3}$$

and which state the conservation of momentum, mass, and energy with enthalpy, respectively. In (4.2.1)-(4.2.3), $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)$ is the strain rate tensor; the function s is the enthalpy; the symbol $\boldsymbol{k}$ stands for the unit vector pointing in the opposite direction to gravity; η, μ are nonlinear functions of temperature that encode the permeability of the porous material and the viscosity of the fluid, respectively. These functions will assume different specifications depending on the phase change model, to be discussed later on. We also represent by $\mathrm{Re} = \rho_{\mathrm{ref}} V_{\mathrm{ref}} L_{\mathrm{ref}} \mu^{-1}$ the Reynolds number, the adimensional buoyancy force $f(\theta) = \mathrm{Ra}\theta(\mathrm{Pr}\,\mathrm{Re}^2)^{-1}$ (depending linearly on the temperature dis-

tribution as in the classical Boussinesq approximation [35]), κ is the adimensional heat conductivity tensor (here assumed isotropic), $\mathrm{Ra} = g\beta L_{\mathrm{ref}}(\theta_h - \theta_c)[\nu\alpha]^{-1}$ is the Rayleigh number, g is the gravity magnitude, $L_{\mathrm{ref}}, \rho_{\mathrm{ref}}, V_{\mathrm{ref}}$ are the reference length, density, and velocity defining the flow, θ_h, θ_c are maximum and minimum temperatures, and $\mathrm{Pr} = \nu\alpha^{-1}$ is the Prandtl number. Here $s(\theta)$ denotes the regularised enthalpy function and it accounts for the latent heat of fusion, i.e. the energy needed to change the phase of a material.

In order to analyse the coupled system (4.2.1)-(4.2.3), we will suppose that the functions μ, η are uniformly bounded and Lipschitz continuous: there exist positive constants $\mu_1, \mu_2, \eta_1, \eta_2, L_\mu$ and L_η, such that

$$\eta_1 \leq \eta(\psi) \leq \eta_2, \quad |\eta(\psi) - \eta(\varphi)| \leq L_\eta|\psi - \varphi| \quad \forall\, \psi, \varphi \in \mathbb{R}, \tag{4.2.4}$$

$$\mu_1 \leq \mu(\psi) \leq \mu_2, \quad |\mu(\psi) - \mu(\varphi)| \leq L_\mu|\psi - \varphi| \quad \forall\, \psi, \varphi \in \mathbb{R}. \tag{4.2.5}$$

Similar assumptions will be placed on the source function f: we suppose that there exists positive constants C_f and L_f such that

$$|f(\psi)| \leq C_f|\psi|, \quad |f(\psi) - f(\varphi)| \leq L_f|\psi - \varphi| \quad \forall\, \psi, \varphi \in \mathbb{R}. \tag{4.2.6}$$

On the other hand, we will suppose that for every $\psi \in \mathrm{H}^1(\Omega)$, we have $s(\psi) \in \mathrm{H}^1(\Omega)$, and that there exist positive constants s_1, s_2, L_{s_1} and L_{s_2} such that

$$|s(\psi)| \leq s_1, \quad |s(\psi) - s(\varphi)| \leq L_{s_1}|\psi - \varphi| \quad |\nabla s(\psi)| \leq s_2|\nabla\psi|, \quad |\nabla s(\psi) - \nabla s(\varphi)| \leq L_{s_2}|\psi - \varphi|, \tag{4.2.7}$$

for all $\psi, \varphi \in \mathbb{R}$. Finally, we suppose that κ is a uniform bounded and uniformly positive definite tensor, meaning that there exist positive constants κ_0 and κ_1 such that

$$|\kappa| \leq \kappa_1, \quad \kappa\boldsymbol{v} \cdot \boldsymbol{v} \geq \kappa_0|\boldsymbol{v}|^2 \quad \forall\, \boldsymbol{v} \in \mathbb{R}^d. \tag{4.2.8}$$

Boundary and initial data. Equations (4.2.1)-(4.2.3) are supplemented with boundary conditions as follows. No-slip boundary conditions are prescribed on the velocity over the whole $\partial\Omega$, and therefore an additional condition is required for pressure uniqueness; as usual we impose a zero-mean property. Regarding the energy equation, we assume that the domain boundary admits a splitting between two disjoint sets Γ_D^θ and Γ_N^θ, where temperature and normal heat fluxes are prescribed, respectively. The system is supposed to be initially at rest and isothermal, and so we set $\boldsymbol{u}(0) = \boldsymbol{0}$, $p(0) = 0$ and $\theta(0) = \theta_0$ with θ_0 constant.

4.2.2 Enthalpy-porosity models for phase change

The permeability function η appearing in the drag term is usually defined in such a way that (4.2.1) behaves as the well-known Carman-Kozeny equations (see e.g. the review [126]). That is, one uses a phase change (or liquid fraction) field ϕ with

$$\eta(\phi) = \xi\frac{(1-\phi)^2}{\phi^3 + m}, \quad \text{with} \quad \phi = \frac{1}{2}\Big[\tanh\left(\frac{5}{\delta\theta}(\theta - \theta_f)\right) + 1\Big], \tag{4.2.9}$$

where $m > 0$ is a small parameter that prevents division by zero. The term $\delta\theta$ represents the temperature range corresponding to the width of the mushy region, and θ_f is a constant the jump function is regularised about and it corresponds to the melting point subject to appropriate scaling (in the sense that in the fluid one has $\phi = 1$ by setting $\eta = 0$, and in the solid $\phi = 0$ corresponds to $\eta = \xi[m]^{-1}$). This also implies that in the solid region one imposes a low permeability field that generates a large drag force. The parameter ξ is a large constant that represents the morphology of the melt front. It is related to the imposed permeability through $\xi = \frac{180}{\rho_{\text{ref}} d_m^2}$ where d_m is the particle diameter, and the constant 180 depends on the material under consideration [38]. Alternatively to (4.2.9), in our examples we will include a regularised jump function that takes the form

$$\eta = \frac{\eta_s}{2}[\tanh(M_\eta(\theta_f - \theta)) + 1], \tag{4.2.10}$$

where η_s corresponds to the relative size of the imposed force and M_η is the size of the mushy region. These constants determine the degree of regularisation of the jump. As above, in the liquid phase we have $\eta = 0$, and in the solid $\eta = \eta_s$, with η_s a large constant accounting for the morphology of the melt front.

In many models from the literature, the phase change is often described by combining the permeability (or porosity) regularised jumps with an enthalpy formulation [38]. The non-dimensional enthalpy function s should ideally be a Heaviside function assuming the values s_s in the solid, and s_l in the liquid. After regularisation using the phase change field ϕ we employ

$$s(\theta) = s_s + (s_l - s_s)\phi(\theta). \tag{4.2.11}$$

We will adopt this form in all of our models, so that the mushy region for the enthalpy will be predetermined by the temperature range, $\delta\theta$. The corresponding constants are $s_l = 0$ and $s_s = [\text{Ste}]^{-1}$, where the Stephan number is inversely proportional to the latent heat of fusion L_a, and proportional to the temperature range $\delta\theta$, and specific heat α. More succinctly $\text{Ste} = \frac{\alpha\delta\theta}{L_a}$. For example a material with a larger latent heat of fusion, will have a larger Stephan number embed a larger jump in the enthalpy jump function. Another observation regarding this jump, is that its height depends on the mushy region size. For instance, if $\delta\theta$ decreases then Ste decreases and s_s increases. In Section 4.5 we refer to mushy regions but always in relation with the jump functions imposed on η, μ. In order to avoid cross-effects associated with having effectively two mushy regions (one for viscosity/drag terms and another for enthalpy), we fix in advance a relation that specifies a fixed latent heat.

4.2.3 Enthalpy-viscosity models for phase change

The incorporation of phase change can be alternatively embedded in the form of a temperature-dependent viscosity combined with an enthalpy formulation (as in e.g. [64]). The information about phase variation from solid to fluid is then carried by two different scaled viscosities. This family of models can also be linked to the principle of packing spheres (more naturally associated with the porosity model above, as a dense packing of spheres in the solid translates into a drag force that slows down the flow), and its influence on viscosity variations.

Defining Φ as the ratio of volume occupied by solid particles, we notice that it is related to the phase function ϕ by $\Phi = \Phi_m(1 - \phi)$, where Φ_m is a constant depending on the maximum packing of

the imposed particles, and so the latter corresponds to the ratio of volume not corresponding to solid particles. As a given solid melts, then there are less solid particles suspended in the liquid, implying an adequate decrease in the viscosity $\mu = \phi^{-B\Phi_m}$. Defining $n = B\Phi_m$ and using a security constant m (mimicking the regularisation in the Carman-Kozeny equation) we obtain

$$\mu = \frac{1}{\phi^n + m}. \tag{4.2.12}$$

In analogy to (4.2.10) we will also employ

$$\mu(\theta) = \mu_l + \frac{(\mu_s - \mu_l)}{2}[\tanh(M_\mu(\theta_f - \theta)) + 1], \tag{4.2.13}$$

so that in the solid we have $\mu = \mu_s$, and in the liquid $\mu = \mu_l$. Here the constant M_μ encodes the width of the mushy region.

4.2.4 Relationship with the rheology of suspended particles

Historic models for the rheology of suspensions go back to [72], where the relation $\mu = 1 + B\Phi$, with $B = 2.5$ is postulated. This model can be derived by the consideration of slow flow past a sphere, and it corresponds to linear Newton rheology and is only valid for very dilute suspensions ($\Phi \lesssim 0.01$), so it would not be suitable for phase change models. Extensions to the case of particles with varying size were derived from first principles in [40] and [129]. These models also capture the property that the viscosity tends to infinity as the ratio of solid volume to liquid volume tends to 1, and they deal much better with non Newtonian viscosities observed at higher values of Φ. This form of the viscosity model is still used for nano particles [74], or for the sedimentation-consolidation in macroscopic models [133], and it relates to our model of phase change. Traditionally, these models incorporate differences in density and enthalpy by considering nano-particles made of different materials. See for instance [68], where thermal and density properties of copper nano-particles are taken into account. In this context, the models in (4.2.12) and (4.2.13) consider nano-particles with the same density and thermal properties as the liquid, another important distinction is that the concentration of nano-particles is dependent on temperature in such a way that the particles are not present in the liquid phase, and they exhibit maximum concentration in the solid phase. This is linked to the concept of critical fraction introduced in [129]. There, a packing density becomes high enough that the fluid behaves as a solid, $\mu = (1 - \frac{\Phi}{\Phi_m})^{-2.5}$ where Φ_m is a constant depending on the maximum particle packing, and the model was later extended to $\mu = (1 - \frac{\Phi}{\Phi_m})^{-B\Phi_m}$, see [106].

4.3 Analysis of Boussinesq phase change models

4.3.1 Weak formulation

Firstly let us recall some recurrent notation. For instance, we will write $\mathrm{L}^2(\Omega)$ to denote the space of square integrable functions, and will use $\mathrm{H}^1(\Omega), \mathbf{H}^1(\Omega)$ to refer to the scalar and vector-valued Sobolev spaces $W^{1,2}(\Omega)$ and $\boldsymbol{W}^{1,2}(\Omega)$, respectively; whose norms will be denoted as $\|\cdot\|_{1,\Omega}$. The inner product in $\mathrm{L}^2(\Omega)$ (or in its vectorial and tensorial counterparts) will be simply denoted as $(\cdot,\cdot)$ and its associated

norm as $\|\cdot\|$. In addition, the space $\mathrm{L}^2_0(\Omega)$ denotes the restriction of $\mathrm{L}^2(\Omega)$ to functions with zero mean value over Ω. In view of incorporating the boundary conditions for velocity and temperature, we also introduce the space $\mathbf{H}^1_0(\Omega)$ of vector functions in $\mathbf{H}^1(\Omega)$ whose trace vanishes on $\partial\Omega$, and the space $\mathrm{H}^1_D(\Omega)$ of scalar functions in $\mathrm{H}^1(\Omega)$ whose trace vanishes on the sub-boundary Γ^θ_D.

Associated with the spaces introduced above, the following nonlinear, bilinear and trilinear forms are defined for all $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbf{H}^1(\Omega)$, $p,q\in\mathrm{L}^2(\Omega)$, and $\theta,\psi\in\mathrm{H}^1(\Omega)$

$$\begin{aligned} a_1^\theta(\boldsymbol{u},\boldsymbol{v}) &:= \frac{2}{\mathrm{Re}}\int_\Omega \mu(\theta)\boldsymbol{\varepsilon}(\boldsymbol{u}):\boldsymbol{\varepsilon}(\boldsymbol{v}),\ b(\boldsymbol{v},q) := -\int_\Omega q\ \mathrm{div}\,\boldsymbol{v},\ c_1(\boldsymbol{w};\boldsymbol{u},\boldsymbol{v}) := \int_\Omega [(\boldsymbol{w}\cdot\nabla)\boldsymbol{u}]\cdot\boldsymbol{v},\\ a_3(\theta,\psi) &:= \frac{1}{C\,\mathrm{Pr}}\int_\Omega \kappa\nabla\theta\cdot\nabla\psi,\quad c_3(\boldsymbol{w};\theta,\psi) := \int_\Omega [\boldsymbol{w}\cdot\nabla\theta]\psi. \end{aligned} \tag{4.3.1}$$

On account of these definitions, we proceed to test (4.2.1)-(4.2.2) against adequate functions and integrate by parts conveniently in order to arrive at the following problem in weak form. For all $t\in(0,t_f]$, find $(\boldsymbol{u},p,\theta)\in\mathbf{H}^1_0(\Omega)\times\mathrm{L}^2_0(\Omega)\times\mathrm{H}^1_D(\Omega)$ such that

$$\begin{aligned} (\partial_t\boldsymbol{u},\boldsymbol{v})+c_1(\boldsymbol{u};\boldsymbol{u},\boldsymbol{v})+a_1^\theta(\boldsymbol{u},\boldsymbol{v})+(\eta(\theta)\boldsymbol{u},\boldsymbol{v})+b(\boldsymbol{v},p) &= (f(\theta)\boldsymbol{k},\boldsymbol{v}) \qquad \forall\boldsymbol{v}\in\mathbf{H}^1_0(\Omega),\\ b(\boldsymbol{u},q) &= 0 \qquad \forall q\in\mathrm{L}^2_0(\Omega),\\ (\partial_t[\theta+s],\psi)+c_3(\boldsymbol{u};\theta+s,\psi)+a_3(\theta,\psi) &= 0 \qquad \forall\psi\in\mathrm{H}^1_D(\Omega). \end{aligned} \tag{4.3.2}$$

The forms defined in (4.3.1) enjoy the following properties, established in e.g. [34]

$$\begin{aligned} |a_1^\theta(\boldsymbol{u},\boldsymbol{v})|&\le C\|\boldsymbol{u}\|_{1,\Omega}\|\boldsymbol{v}\|_{1,\Omega},\quad |a_1^\theta(\boldsymbol{v},\boldsymbol{v})|\ge C\|\boldsymbol{v}\|^2_{1,\Omega},\\ |a_3(\theta,\psi)|&\le C\|\theta\|_{1,\Omega}\|\psi\|_{1,\Omega},\quad |a_3(\psi,\psi)|\ge C\|\psi\|^2_{1,\Omega},\quad b(\boldsymbol{v},q)\le\|\boldsymbol{v}\|_{1,\Omega}\|q\|,\\ |c_1(\boldsymbol{w};\boldsymbol{u},\boldsymbol{v})|&\le C\|\boldsymbol{w}\|_{1,\Omega}\|\boldsymbol{u}\|_{1,\Omega}\|\boldsymbol{v}\|_{1,\Omega},\quad |c_3(\boldsymbol{w};\theta,\psi)|\le C\|\boldsymbol{w}\|_{1,\Omega}\|\theta\|_{1,\Omega}\|\psi\|_{1,\Omega}, \end{aligned}$$

for all $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\mathbf{H}^1_0(\Omega)$, $p,q\in\mathrm{L}^2(\Omega)$, and $\theta,\psi\in\mathrm{H}^1_D(\Omega)$. Also, there exists $C>0$ depending only on the domain, such that

$$\sup_{\boldsymbol{v}\in\mathbf{H}^1(\Omega)\setminus\mathbf{0}}\frac{b(\boldsymbol{v},q)}{\|\boldsymbol{v}\|_{1,\Omega}}\ge C\|q\|_{0,\Omega}\qquad\forall\, q\in\mathrm{L}^2_0(\Omega). \tag{4.3.3}$$

We can now define the kernel of the bilinear form $b(\cdot,\cdot)$, characterised by the space of divergence-free velocities $\mathbf{V}$, as

$$\mathbf{V}=\{\boldsymbol{v}\in\mathbf{H}^1_0(\Omega);\ \ \mathrm{div}\,\boldsymbol{v}=0\ \text{ on }\ \Omega\}.$$

Thus, based on (4.3.3) and the definition of $\mathbf{V}$, the incompressibility condition is included in the functional space and the pressure can be removed from the formulation. That is, problem (4.3.2) is equivalent to the following problem: For all $t\in(0,t_f]$, find $(\boldsymbol{u},\theta)\in\mathbf{V}\times\mathrm{H}^1_D$ such that

$$\begin{aligned} (\partial_t\boldsymbol{u},\boldsymbol{v})+c_1(\boldsymbol{u};\boldsymbol{u},\boldsymbol{v})+a_1^\theta(\boldsymbol{u},\boldsymbol{v})+(\eta(\theta)\boldsymbol{u},\boldsymbol{v}) &= (f(\theta)\boldsymbol{k},\boldsymbol{v}) \qquad \forall\boldsymbol{v}\in\mathbf{V},\\ (\partial_t[\theta+s],\psi)+c_3(\boldsymbol{u};\theta+s,\psi)+a_3(\theta,\psi) &= 0 \qquad \forall\psi\in\mathrm{H}^1_D(\Omega), \end{aligned} \tag{4.3.4}$$

(a proof can be carried out following e.g. [67]). Moreover, one can see that for all $\boldsymbol{u}\in\mathbf{V}$, $\boldsymbol{v}\in\mathbf{H}^1(\Omega)$, and $\vartheta,\psi\in\mathrm{H}^1(\Omega)$, we have

$$c_1(\boldsymbol{u};\boldsymbol{v},\boldsymbol{v})=0,\quad c_3(\boldsymbol{u},\psi,\psi)=0,\quad c_3(\boldsymbol{u};\psi,\vartheta)=-c_3(\boldsymbol{u};\vartheta,\psi).$$

4.3.2 Stability analysis

Note also that the convective and advective terms can be rewritten using skew-symmetric forms as follows

$$c_1(\boldsymbol{w};\boldsymbol{u},\boldsymbol{v}) = \frac{1}{2}\int_\Omega [(\boldsymbol{w}\cdot\nabla)\boldsymbol{u}]\cdot\boldsymbol{v} - \frac{1}{2}\int_\Omega [(\boldsymbol{w}\cdot\nabla)\boldsymbol{v}]\cdot\boldsymbol{u}, \quad c_3(\boldsymbol{w};\theta,\psi) = \frac{1}{2}\int_\Omega [\boldsymbol{w}\cdot\nabla\theta]\psi - \frac{1}{2}\int_\Omega [\boldsymbol{w}\cdot\nabla\psi]\theta. \tag{4.3.5}$$

Single phase flows. Let us consider an enthalpy-free counterpart of (4.3.2), and proceed to derive energy estimates. Testing the energy equation against the temperature solution, and using (4.3.5) we obtain

$$\frac{1}{2}\partial_t\|\theta\|^2_{0,\Omega} + \frac{\kappa}{\mathrm{CPr}}\|\nabla\theta\|^2_{0,\Omega} = 0,$$

and then one can use Gronwall's Lemma to assert that

$$\|\theta\|^2_{0,\Omega} + \int_0^t \frac{2\kappa}{\mathrm{CPr}}\|\nabla\theta\|^2_{0,\Omega}ds \leq \|\theta_0\|^2_{0,\Omega}.$$

Testing now the momentum equation against the velocity solution, exploiting again (4.3.5), and applying Young's inequality we obtain

$$\partial_t\|\boldsymbol{u}\|^2_{0,\Omega} + \big\|\frac{2\mu(\theta)^{0.5}}{\mathrm{Re}^{0.5}}\boldsymbol{\varepsilon}(\boldsymbol{u})\big\|^2_{0,\Omega} + 2A\|\boldsymbol{u}\|^2_{0,\Omega} \leq A_1^2\|\theta_0\|^2_{0,\Omega} + \|\boldsymbol{u}\|^2_{0,\Omega},$$

where we have also used that $|\boldsymbol{k}| = 1$. We can then we invoke Gronwall's Lemma once again to get

$$\|\boldsymbol{u}\|^2_{0,\Omega} + \int_0^t \big\|\frac{2\mu(\theta)^{0.5}}{\mathrm{Re}^{0.5}}\boldsymbol{\varepsilon}(\boldsymbol{u})\big\|^2_{0,\Omega}ds \leq \exp((1-2A)t)\big(\|\boldsymbol{u}_0\|^2_{0,\Omega} + \int_0^t A_1^2\|\theta_0\|^2_{0,\Omega}\big).$$

Enthalpy-based flows. The following two lemmas will be employed to show the stability and uniqueness analysis of (4.3.2).

Lemma 4.3.1. *For $d = 2$, there holds*

$$\|\boldsymbol{v}\|^2_{4,\Omega} \leq 2^{1/2}\,\|\boldsymbol{v}\|_{0,\Omega}\,|\boldsymbol{v}|_{1,\Omega} \quad \forall\boldsymbol{v}\in\mathbf{H}^1(\Omega).$$

Lemma 4.3.2. *There holds*

$$\|\boldsymbol{\varepsilon}(\boldsymbol{v})\|^2_{0,\Omega} \geq \frac{1}{2}|\boldsymbol{v}|^2_{1,\Omega} \quad \forall\boldsymbol{v}\in\mathbf{H}^1_0(\Omega).$$

We now establish the stability analysis of problem (4.3.4).

Theorem 4.3.3. *Assume that $\kappa_0 > \kappa_1 s_2$. Then, for any solution of (4.3.4) and for any $t\in(0,t_f]$, there exists a constant $\widetilde{C}$ depending on $\mu_1, \mathrm{Re}, \eta_1, C_f, \kappa_0, \kappa_1, s_2, C, \mathrm{Pr}, \Omega$, t_f and c_p (positive constant provided by Poincaré's inequality), such that*

$$\|\boldsymbol{u}\|_{\mathbf{L}^2(0,t;\mathbf{H}^1_0(\Omega))} + \|\theta\|_{\mathrm{L}^2(0,t;\mathrm{H}^1_D(\Omega))} \leq \widetilde{C}|\boldsymbol{k}|\left\{\|s_0\|_{0,\Omega} + \|\theta_0\|_{0,\Omega}\right\}. \tag{4.3.6}$$

Proof. Let $(\boldsymbol{u},\theta)$ be the solution of (4.3.4). Taking $\boldsymbol{v}=\boldsymbol{u}$ in the first equation of (4.3.4) and applying (4.2.4), (4.2.5) and (4.2.6), we obtain that

$$\frac{1}{2}\partial_t\|\boldsymbol{u}\|_{0,\Omega}^2+\frac{2\mu_1}{\mathrm{Re}}\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{0,\Omega}^2+\eta_1\|\boldsymbol{u}\|_{0,\Omega}^2\leq C_f|\boldsymbol{k}|\,\|\theta\|_{0,\Omega}\|\boldsymbol{u}\|_{0,\Omega}\,.$$

Next, applying Young's inequality with constant $\varepsilon=\frac{\eta_1}{C_f|\boldsymbol{k}|}$, we find that

$$\frac{1}{2}\partial_t\|\boldsymbol{u}\|_{0,\Omega}^2+\frac{2\mu_1}{\mathrm{Re}}\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{0,\Omega}^2+\frac{\eta_1}{2}\|\boldsymbol{u}\|_{0,\Omega}^2\leq\frac{C_f^2|\boldsymbol{k}|^2}{2\eta_1}\|\theta\|_{0,\Omega}^2\,.$$

By using Lemma 4.3.2, we deduce that

$$\frac{1}{2}\partial_t\|\boldsymbol{u}\|_{0,\Omega}^2+\alpha_1\|\boldsymbol{u}\|_{1,\Omega}^2\leq\frac{C_f^2|\boldsymbol{k}|^2}{2\eta_1}\|\theta\|_{0,\Omega}^2\,,$$

where $\alpha_1:=\min\left\{\frac{\mu_1}{\mathrm{Re}},\frac{\eta_1}{2}\right\}$. Now, integrating this equation between 0 and t yields

$$\|\boldsymbol{u}\|_{0,\Omega}^2+\|\boldsymbol{u}\|_{\mathbf{L}^2(0,t;\mathbf{H}_0^1(\Omega))}^2\leq C_1|\boldsymbol{k}|^2\,\|\theta\|_{\mathrm{L}^2(0,t;\mathrm{H}_D^1(\Omega))}^2\,,\tag{4.3.7}$$

where C_1 is a constant depending on $C_f,\mu_1,\eta_1,\mathrm{Re},\Omega$ and t_f. Similarly, we take $\psi=\theta+s$ in the third row of (4.3.2), and apply (4.2.7) and (4.2.8) to obtain

$$\frac{1}{2}\partial_t\|\theta+s\|_{0,\Omega}^2+\frac{\kappa_0}{C\mathrm{Pr}}|\theta|_{1,\Omega}^2\leq\frac{\kappa_1 s_2}{C\mathrm{Pr}}|\theta|_{1,\Omega}^2.$$

Now, we integrate between 0 and t to obtain

$$\|\theta+s\|_{0,\Omega}^2+\|\theta\|_{\mathrm{L}^2(0,t;\mathrm{H}_D^1(\Omega))}^2\leq C_2\left\{\|s_0\|_{0,\Omega}^2+\|\theta_0\|_{0,\Omega}^2\right\},\tag{4.3.8}$$

where $s_0:=s(\theta_0)$ and C_2 is a constant depending on $c_p,C,\mathrm{Pr},\kappa_0,\kappa_1,s_2,\Omega$ and t_f. Finally, we derive the result (4.3.6) from (4.3.7) and (4.3.8). □

Theorem 4.3.4. *Assume that the data $\theta_0\in\mathrm{L}^2(\Omega)$. Then, problem (4.3.2) has a solution $(\boldsymbol{u},p,\theta)\in\mathbf{L}^2(0,t_f;\mathbf{H}_0^1(\Omega))\times\mathrm{L}^2(0,t_f;\mathrm{L}_0^2(\Omega))\times\mathrm{L}^2(0,t_f;\mathrm{H}_D^1(\Omega))$.*

Proof. It follows from an argument based on Galerkin's method and assumptions (4.2.4)-(4.2.7). For more details see [4, Theorem 2.3]. □

The following result establishes the uniqueness of problem (4.3.2).

Theorem 4.3.5. *Let $d=2$. If the problem (4.3.2) admits a solution $(\boldsymbol{u},\theta,p)\in\mathbf{L}^p(0,t_f;\boldsymbol{W}^{1,r}(\Omega))\times\mathrm{L}^2(0,t_f;\mathrm{H}_0^1(\Omega))\times\mathrm{L}^2(0,t_f;\mathrm{L}_0^2(\Omega))$, with $p\geq 4$ and $r\geq 4$, and $L_{s_1}<1/2$, then this solution is unique.*

Proof. Let $(\boldsymbol{u}_1,p_1,\theta_1)$ and $(\boldsymbol{u}_2,p_2,\theta_2)$ be two solutions of (4.3.2). With the aim to prove uniqueness, we denote $\overline{\boldsymbol{u}}=\boldsymbol{u}_1-\boldsymbol{u}_2$, $\overline{p}=p_1-p_2$ and $\overline{\theta}=\theta_1-\theta_2$. Now, from the third equation in (4.3.2), by adding and subtracting $c_3(\boldsymbol{u}_2,\theta_1+s(\theta_1),\psi)$, with $\psi=\overline{\theta}+s(\theta_1)-s(\theta_2)$ and applying Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}&\frac{1}{2}\partial_t\|\overline{\theta}+s(\theta_1)-s(\theta_2)\|_{0,\Omega}^2+\frac{1}{C\mathrm{Pr}}\kappa_0|\overline{\theta}|_{1,\Omega}^2\\&\qquad\leq\frac{\kappa_1 L_{s_2}}{C\mathrm{Pr}}|\overline{\theta}|_{1,\Omega}\|\overline{\theta}\|_{0,\Omega}+|\theta_1+s(\theta_1)|_{1,\Omega}\|\overline{\boldsymbol{u}}\|_{4,\Omega}\|\overline{\theta}+s(\theta_1)-s(\theta_2)\|_{4,\Omega}.\end{aligned}$$

Next, using Lemma 4.3.1 and Young's inequality with constant ε_3, we deduce that

$$\begin{aligned}
&\frac{1}{2}\partial_t\|\overline{\theta}+s(\theta_1)-s(\theta_2)\|^2_{0,\Omega}+\frac{1}{C\mathrm{Pr}}\kappa_0|\overline{\theta}|^2_{1,\Omega}\\
&\qquad\leq\frac{\kappa_1 L_{s_2}}{C\mathrm{Pr}}\left\{\frac{1}{2\varepsilon_3}|\overline{\theta}|^2_{1,\Omega}+\frac{\varepsilon_3}{2}\|\overline{\theta}\|^2_{0,\Omega}\right\}+\frac{1}{\sqrt{2}}|\theta_1+s(\theta_1)|_{1,\Omega}\|\overline{\boldsymbol{u}}\|_{0,\Omega}|\overline{\boldsymbol{u}}|_{1,\Omega}\\
&\qquad\quad+\frac{1}{\sqrt{2}}|\overline{\theta}+s(\theta_1)-s(\theta_2)|_{1,\Omega}\|\overline{\theta}+s(\theta_1)-s(\theta_2)\|_{0,\Omega}|\theta_1+s(\theta_1)|_{1,\Omega}.
\end{aligned}$$

Finally, by applying again Young's inequality with constants ε_4 and ε_5, we get

$$\begin{aligned}
&\frac{1}{2}\partial_t\|\overline{\theta}+s(\theta_1)-s(\theta_2)\|^2_{0,\Omega}+\frac{1}{C\mathrm{Pr}}\kappa_0|\overline{\theta}|^2_{1,\Omega}\\
&\leq\frac{\kappa_1 L_{s_2}}{C\mathrm{Pr}}\left\{\frac{1}{2\varepsilon_3}|\overline{\theta}|^2_{1,\Omega}+\frac{\varepsilon_3}{2}\|\overline{\theta}\|^2_{0,\Omega}\right\}+\frac{1}{\sqrt{2}}\left\{\frac{1}{2}\varepsilon_4|\overline{\boldsymbol{u}}|^2_{1,\Omega}+\frac{1}{2\varepsilon_4}|\theta_1+s(\theta_1)|^2_{1,\Omega}\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}\right\}\\
&+\frac{1}{\sqrt{2}}\left\{\varepsilon_5\left(|\overline{\theta}|^2_{1,\Omega}+L^2_{s_2}\|\overline{\theta}\|^2_{0,\Omega}\right)+\frac{(L^2_{s_1}+1)}{\varepsilon_5}\|\overline{\theta}\|^2_{0,\Omega}|\theta_1+s(\theta_1)|^2_{1,\Omega}\right\}.
\end{aligned}\tag{4.3.9}$$

Analogously, from the first equation in (4.3.2), adding and subtracting the terms $c_1(\boldsymbol{u}_2,\boldsymbol{u}_1,\boldsymbol{v})$, $(\eta(\theta_2)\boldsymbol{u}_1,\boldsymbol{v})$ and $\frac{2}{\mathrm{Re}}(\mu(\theta_2)\boldsymbol{\varepsilon}(\boldsymbol{u}_1),\boldsymbol{\varepsilon}(\boldsymbol{v}))$, with $\boldsymbol{v}=\overline{\boldsymbol{u}}$, and applying Cauchy-Schwarz inequality, we find that

$$\begin{aligned}
&\frac{1}{2}\partial_t\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}+\frac{2\mu_1}{\mathrm{Re}}\|\boldsymbol{\varepsilon}(\overline{\boldsymbol{u}})\|^2_{0,\Omega}+\eta_1\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}\\
&\quad\leq\|\overline{\boldsymbol{u}}\|^2_{4,\Omega}|\overline{\boldsymbol{u}}|_{1,\Omega}+\frac{2}{\mathrm{Re}}|((\mu(\theta_1)-\mu(\theta_2))\boldsymbol{\varepsilon}(\boldsymbol{u}_1),\boldsymbol{\varepsilon}(\overline{\boldsymbol{u}}))|+L_\eta\|\boldsymbol{u}_1\|_{0,\Omega}\|\overline{\theta}\|_{4,\Omega}\|\overline{\boldsymbol{u}}\|_{4,\Omega}+L_f|\boldsymbol{k}|\|\overline{\theta}\|_{0,\Omega}\|\overline{\boldsymbol{u}}\|_{0,\Omega}.
\end{aligned}\tag{4.3.10}$$

Thus, on the left hand side of (4.3.10), we apply Korn inequality, while on the right hand side, for the first, third and fourth term, we apply Young's inequality with constants $\varepsilon_6,\varepsilon_7$ and ε_8 to obtain

$$\begin{aligned}
&\frac{1}{2}\partial_t\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}+\alpha_3\|\overline{\boldsymbol{u}}\|^2_{1,\Omega}\\
&\leq\frac{1}{2}\varepsilon_6|\overline{\boldsymbol{u}}|^2_{1,\Omega}+\frac{1}{2\varepsilon_6}\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}|\boldsymbol{u}_1|^2_{1,\Omega}+\frac{2}{\mathrm{Re}}|((\mu(\theta_1)-\mu(\theta_2))\boldsymbol{\varepsilon}(\boldsymbol{u}_1),\boldsymbol{\varepsilon}(\overline{\boldsymbol{u}}))|\\
&+\frac{1}{\sqrt{2}}L_\eta\|\boldsymbol{u}_1\|_{0,\Omega}\left\{\frac{1}{2}\varepsilon_7\|\overline{\theta}\|^2_{0,\Omega}+\frac{1}{2\varepsilon_7}|\overline{\theta}|^2_{1,\Omega}\right\}+\frac{1}{\sqrt{2}}L_\eta\|\boldsymbol{u}_1\|_{0,\Omega}\left\{\frac{1}{2\varepsilon_8}|\overline{\boldsymbol{u}}|^2_{1,\Omega}+\frac{1}{2}\varepsilon_8\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}\right\}\\
&+\frac{1}{2}L_f|\boldsymbol{k}|\left\{\|\overline{\theta}\|^2_{0,\Omega}+\|\overline{\boldsymbol{u}}\|^2_{0,\Omega}\right\},
\end{aligned}\tag{4.3.11}$$

where $\alpha_3:=\min\left\{\frac{\mu_1}{\mathrm{Re}},\eta_1\right\}$. Now, since the exact solution $\boldsymbol{u}_1\in\mathbf{L}^p(0,t_f,\mathbf{W}^{1,r}(\Omega))$, $r\geq 4$, $p\geq 4$, using Hölder's and Young's inequalities with constants r, $r^*:=\frac{2r}{r-2}$ and ε_9, ε_{10}, respectively, to the last term of (4.3.11), we deduce that

$$\begin{aligned}
\frac{2}{\mathrm{Re}}|((\mu(\theta_1)-\mu(\theta_2))\boldsymbol{\varepsilon}(\boldsymbol{u}_1):\boldsymbol{\varepsilon}(\overline{\boldsymbol{u}}))|&\leq\frac{2L_\mu}{\mathrm{Re}}\left\{\frac{\varepsilon_9}{2}\|\boldsymbol{u}_1\|_{1,r,\Omega}\|\overline{\boldsymbol{u}}\|^2_{1,\Omega}+\frac{1}{2\varepsilon_9}\|\boldsymbol{u}_1\|_{1,r,\Omega}\|\overline{\theta}\|^2_{r^*,\Omega}\right\}\\
&\leq\frac{L_\mu\varepsilon_9}{\mathrm{Re}}\|\boldsymbol{u}_1\|_{1,r,\Omega}\|\overline{\boldsymbol{u}}\|^2_{1,\Omega}+\frac{L_\mu}{\varepsilon_9\mathrm{Re}}\|\boldsymbol{u}_1\|_{1,r,\Omega}\left\{\frac{1}{\sqrt{2}\varepsilon_{10}}\|\overline{\theta}\|^2_{0,\Omega}+\frac{\varepsilon_{10}}{\sqrt{2}}|\overline{\theta}|^2_{1,\Omega}\right\}.
\end{aligned}\tag{4.3.12}$$

Now, choosing the parameters as

$$\varepsilon_3=\frac{2\kappa_1 L_{s_2}}{\kappa_0},\ \varepsilon_4=\frac{\alpha_3}{\sqrt{2}},\ \varepsilon_5=\frac{\kappa_0}{2\sqrt{2}C\mathrm{Pr}},\ \varepsilon_6=\frac{\alpha_3}{2},\ \varepsilon_7=\frac{\sqrt{2}C\mathrm{Pr}L_\eta\|\boldsymbol{u}_1\|_{0,\Omega}}{\kappa_0},$$

$$\varepsilon_8 = \frac{\sqrt{2}L_\eta \|\boldsymbol{u}_1\|_{0,\Omega}}{\alpha_3},\ \varepsilon_9 = \frac{\alpha_3 \mathrm{Re}}{4L_\mu \|\boldsymbol{u}_1\|_{1,r,\Omega}}, \varepsilon_{10} = \frac{\kappa_0 \varepsilon_9 \mathrm{Re}}{2\sqrt{2}C\mathrm{Pr}L\mu \|\boldsymbol{u}_1\|_{1,r,\Omega}},$$

we obtain from (4.3.9), (4.3.11) and (4.3.12) that

$$\begin{aligned}&\frac{1}{2}\partial_t \left(\|\overline{\theta} + s(\theta_1) - s(\theta_2)\|^2_{0,\Omega} + \|\overline{\boldsymbol{u}}\|^2_{0,\Omega}\right)\\ &\qquad \leq \overline{C}\left(|\theta_1 + s(\theta_1)|^2_{1,\Omega} + \|\boldsymbol{u}_1\|^4_{1,r,\Omega} + \|\boldsymbol{u}_1\|^2_{0,\Omega} + 1\right)\left(\|\overline{\boldsymbol{u}}\|^2_{0,\Omega} + \|\overline{\theta}\|^2_{0,\Omega}\right).\end{aligned}$$

Integrating between 0 and t on the last inequality, and then, adding on both sides the term $(|\theta|, |s(\theta_1) - s(\theta_2)|)$, and using the assumption given in the theorem statement, we get

$$\|\overline{\theta}\|^2_{0,\Omega} + \|\overline{\boldsymbol{u}}\|^2_{0,\Omega} \leq \widehat{C}\int_0^t \left(|\theta_1 + s(\theta_1)|^2_{1,\Omega} + \|\boldsymbol{u}_1\|^4_{1,r,\Omega} + \|\boldsymbol{u}_1\|^2_{0,\Omega} + 1\right)\left(\|\overline{\boldsymbol{u}}\|^2_{0,\Omega} + \|\overline{\theta}\|^2_{0,\Omega}\right).$$

Finally, by applying Gronwall's Lemma, we obtain $\overline{\boldsymbol{u}} = \boldsymbol{0}$ and $\overline{\theta} = 0$. Moreover, from the relation

$$(\mathrm{div}\,\boldsymbol{v}, \overline{p}) = 0 \quad \forall\, \boldsymbol{v} \in \mathrm{H}^1_0(\Omega),$$

we deduce $\overline{p} = 0$, concluding the proof. □

4.4 Two families of finite element schemes

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of partitions of the region $\bar{\Omega}$, by triangles (or tetrahedrons in 3D) K of diameter h_K, with overall meshsize $h := \max\{h_K : K \in \mathcal{T}_h\}$. In what follows, given an integer $k \geq 1$ and a subset S of $\mathbb{R}^d$, $\mathbb{P}_k(S)$ will denote the space of polynomial functions defined locally in S and being of total degree $\leq k$.

4.4.1 A conforming method in primal formulation

The spatial discretisation will be based on the finite element method. Accordingly, we define the following finite-dimensional spaces for the approximation of velocity, pressure, and temperature respectively:

$$\begin{aligned}\mathbf{V}_h &:= \{\boldsymbol{v}_h \in \mathbf{C}(\overline{\Omega}) : \boldsymbol{v}_h|_K \in [\mathbb{P}_{k+1}(K)]^d\ \forall K \in \mathcal{T}_h,\ \text{and } \boldsymbol{v}_h = \boldsymbol{0} \text{ on } \partial\Omega\},\\ \mathrm{Q}_h &:= \{q_h \in \mathrm{C}(\overline{\Omega}) : q_h|_K \in \mathbb{P}_k(K)\ \forall K \in \mathcal{T}_h,\ \text{ and } \int_\Omega q_h = 0\},\\ \mathrm{Z}_h &:= \{\psi_h \in \mathrm{C}(\overline{\Omega}) : \psi_h|_K \in \mathbb{P}_{k+1}(K)\ \forall K \in \mathcal{T}_h,\ \text{and } \psi_h = 0 \text{ on } \Gamma_D^\theta\},\end{aligned} \tag{4.4.1}$$

for $k \geq 1$, which satisfy the discrete inf-sup condition: There exists a constant $C^* \geq 0$ independent of h such that

$$\sup_{v_h \in \mathbf{V}_h \setminus \mathbf{0}} \frac{b(v_h, q_h)}{\|v_h\|_{1,\Omega}} \geq C^* \|q_h\|_{0,\Omega} \quad \forall\, q_h \in \mathrm{Q}_h. \tag{4.4.2}$$

Then the semi-discrete Galerkin method associated with (4.3.2) reads: For all $t \in (0, t_f]$, find $(\boldsymbol{u}_h, p_h, \theta_h) \in \mathbf{V}_h \times \mathrm{Q}_h \times \mathrm{Z}_h$ such that

$$\begin{aligned}(\partial_t \boldsymbol{u}_h, \boldsymbol{v}_h) + c_1(\boldsymbol{u}_h; \boldsymbol{u}_h, \boldsymbol{v}_h) + a_1^{\theta_h}(\boldsymbol{u}_h, \boldsymbol{v}_h) + (\eta(\theta_h)\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) &= (f(\theta_h)\boldsymbol{k}, \boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in \mathbf{V}_h,\\ b(\boldsymbol{u}_h, q_h) &= 0 \qquad \forall q_h \in \mathrm{Q}_h,\\ (\partial_t[\theta_h + s_h], \psi_h) + c_3(\boldsymbol{u}_h; \theta_h + s_h, \psi_h) + a_3(\theta_h, \psi_h) &= 0 \qquad \forall \psi_h \in \mathrm{Z}_h\end{aligned} \tag{4.4.3}$$

A fully discrete method will be obtained after applying the method of lines. Regarding the time discretisation of (4.4.3), and in view of the overall second order space discretisation expected when choosing $k = 1$, we here employ a fully implicit second-order backward differentiation formula (BDF2). This choice provides unconditional stability and permits to take sufficiently large timesteps to reach approximate steady state solutions, should they exist. Let $0 = t^0 < t^1 < \cdots t^N = t_f$ be a uniform partition of the time interval into equi-spaced subintervals of size Δt, then the method reads: starting from the initial values $\boldsymbol{u}_h^0, \theta_h^0, \boldsymbol{u}_h^1, \theta_h^1$ taken as interpolates of the initial data onto $\mathbf{V}_h$ and Z_h, solve for $n = 1, \ldots$ the nonlinear system

$$\begin{aligned}&\frac{3}{2\Delta t}(\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + c_1(\boldsymbol{u}_h^{n+1}; \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h^{n+1}\, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + a_1^{\theta_h^{n+1}}(\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h)\\ &\quad + (\eta(\theta_h^{n+1})\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h^{n+1}) - (f(\theta_h^{n+1})\boldsymbol{k}, \boldsymbol{v}_h) = \frac{1}{\Delta t}(2\boldsymbol{u}_h^n - \frac{1}{2}\boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) \qquad \forall \boldsymbol{v}_h \in \mathbf{V}_h,\\ &\qquad b(\boldsymbol{u}_h^{n+1}, q_h) = 0 \qquad \forall q_h \in \mathrm{Q}_h,\\ &\frac{3}{2\Delta t}(\theta_h^{n+1} + s_h^{n+1}, \psi_h) + c_3(\boldsymbol{u}_h^{n+1}; \theta_h^{n+1} + s_h^{n+1}, \psi_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h^{n+1}\, (\theta_h^{n+1} + s_h^{n+1}), \psi_h)\\ &\qquad + a_3(\theta_h^{n+1}, \psi_h) = \frac{1}{\Delta t}(2[\theta_h^n + s_h^n] - \frac{1}{2}[\theta_h^{n-1} + s_h^{n-1}], \psi_h) \qquad \forall \psi_h \in \mathrm{Z}_h,\end{aligned} \tag{4.4.4}$$

and for $n = 0$ one applies a first order backward Euler method. An advantage of introducing these additional terms is that we have the following relations (for a proof, see e.g. [1])

$$c_1(\boldsymbol{u}_h; \boldsymbol{v}_h, \boldsymbol{v}_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h\, \boldsymbol{v}_h, \boldsymbol{v}_h) = c_3(\boldsymbol{u}_h; \psi_h, \psi_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h\, \psi_h, \psi_h) = 0 \quad \forall\, \boldsymbol{u}_h, \boldsymbol{v}_h \in \mathbf{V}_h, \quad \forall\, \psi_h \in \mathrm{Z}_h, \tag{4.4.5}$$

which will be useful in order to establish the stability of the non-linear problem (4.4.4). More specifically, in Theorem 4.4.1 below, when we take $\boldsymbol{v}_h = 4\boldsymbol{u}_h^{n+1}$ and $\psi_h = 4(\theta_h^{n+1} + s_h^{n+1})$, the properties (4.4.5) immediately hold and the corresponding analysis is substantially simpler.

In much the same way as the continuous case, we define $\mathbf{V}_h^* = \{\boldsymbol{v}_h \in \mathbf{V}_h;\ \ b(\boldsymbol{v}_h, q_h) = 0 \quad \forall\, q_h \in \mathrm{Q}_h\}$ and thanks to the inf-sup condition (4.4.2), consider the following equivalent problem (see [123], Lemma 3.1): Find $(\boldsymbol{u}_h^{n+1}, \theta_h^{n+1}) \in \mathbf{V}_h^* \times \mathrm{Z}_h$ such that

$$\begin{aligned}&\frac{3}{2\Delta t}(\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + c_1(\boldsymbol{u}_h^{n+1}; \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h^{n+1}\, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) + a_1^{\theta_h^{n+1}}(\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h)\\ &\quad + (\eta(\theta_h^{n+1})\boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) - (f(\theta_h^{n+1})\boldsymbol{k}, \boldsymbol{v}_h) = \frac{1}{\Delta t}(2\boldsymbol{u}_h^n - \frac{1}{2}\boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) \qquad \forall \boldsymbol{v}_h \in \mathbf{V}_h^*,\\ &\frac{3}{2\Delta t}(\theta_h^{n+1} + s_h^{n+1}, \psi_h) + c_3(\boldsymbol{u}_h^{n+1}; \theta_h^{n+1} + s_h^{n+1}, \psi_h) + \frac{1}{2}(\operatorname{div} \boldsymbol{u}_h^{n+1}\, (\theta_h^{n+1} + s_h^{n+1}), \psi_h)\\ &\qquad + a_3(\theta_h^{n+1}, \psi_h) = \frac{1}{\Delta t}(2[\theta_h^n + s_h^n] - \frac{1}{2}[\theta_h^{n-1} + s_h^{n-1}], \psi_h) \qquad \forall \psi_h \in \mathrm{Z}_h.\end{aligned} \tag{4.4.6}$$

In what follows, we establish the stability result for (4.4.4), for which the following algebraic identity will be essential: for any real numbers a^{n+1}, a^n and a^{n-1}, we have

$$2(3a^{n+1}-4a^n+a^{n-1},a^{n+1}) = |a^{n+1}|^2+|2a^{n+1}-a^n|^2+|\Lambda a^n|^2-|a^n|^2-|2a^n-a^{n-1}|^2, \tag{4.4.7}$$

where $\Lambda a^n = a^{n+1}-2a^n+a^{n-1}$.

Theorem 4.4.1. *Let $(\boldsymbol{u}_h^{n+1},\theta_h^{n+1})\in \mathbf{V}_h^*\times \mathrm{Z}_h$ be a solution of (4.4.4). Assume that $\kappa_0>2\kappa_1 s_2$. Then,*

$$\begin{aligned}&\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\|_{0,\Omega}^2+\sum_{m=1}^n\|\Lambda\boldsymbol{u}_h^m\|_{0,\Omega}^2+\sum_{m=1}^n\Delta t\,\|\boldsymbol{u}_h^{m+1}\|_{1,\Omega}^2\\ &\quad\le\quad C_1\left(\|\theta_h^1+s_h^1\|_{0,\Omega}^2+\|2(\theta_h^1+s_h^1)-(\theta_h^0+s_h^0)\|_{0,\Omega}^2+\|\boldsymbol{u}_h^1\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^1-\boldsymbol{u}_h^0\|_{0,\Omega}^2\right)\end{aligned} \tag{4.4.8}$$

and

$$\begin{aligned}&\|\theta_h^{n+1}+s_h^{n+1}\|_{0,\Omega}^2+\|2(\theta_h^{n+1}+s_h^{n+1})-(\theta_h^n+s_h^n)\|_{0,\Omega}^2+\sum_{m=1}^n\|\Lambda(\theta_h^m+s_h^m)\|_{0,\Omega}^2+\sum_{m=1}^n\Delta t\,|\theta_h^{m+1}|_{1,\Omega}^2\\ &\quad\le\quad C_2\left(\|\theta_h^1+s_h^1\|_{0,\Omega}^2+\|2(\theta_h^1+s_h^1)-(\theta_h^0+s_h^0)\|_{0,\Omega}^2\right),\end{aligned} \tag{4.4.9}$$

where C_1 and C_2 are constants independent on h and Δt.

Proof. For (4.4.9), we take $\psi_h = 4(\theta_h^{n+1}+s_h^{n+1})$ in the third equation of (4.4.4) and use the relation (4.4.7) to deduce

$$\begin{aligned}&\|\theta_h^{n+1}+s_h^{n+1}\|_{0,\Omega}^2+\|2(\theta_h^{n+1}+s_h^{n+1})-(\theta_h^n+s_h^n)\|_{0,\Omega}^2+\|\Lambda(\theta_h^n+s_h^n)\|_{0,\Omega}^2+\frac{4\kappa_0}{C\mathrm{Pr}}\Delta t\,|\theta_h^{n+1}|_{1,\Omega}^2\\ &\le\ \frac{4s_2\kappa_1}{C\mathrm{Pr}}\Delta t\,|\theta_h^{n+1}|_{1,\Omega}^2+\|\theta_h^n+s_h^n\|_{0,\Omega}^2+\|2(\theta_h^n+s_h^n)-(\theta_h^{n-1}+s_h^{n-1})\|_{0,\Omega}^2.\end{aligned}$$

Thus, summing this inequality over n, we obtain

$$\begin{aligned}&\|\theta_h^{n+1}+s_h^{n+1}\|_{0,\Omega}^2+\|2(\theta_h^{n+1}+s_h^{n+1})-(\theta_h^n+s_h^n)\|_{0,\Omega}^2+\sum_{m=1}^n\|\Lambda(\theta_h^m+s_h^m)\|_{0,\Omega}^2\\ &\quad+\ \frac{4(\kappa_0-2\kappa_1 s_2)}{C\mathrm{Pr}}\sum_{m=1}^n\Delta t\,|\theta_h^{m+1}|_{1,\Omega}^2\le\|\theta_h^1+s_h^1\|_{0,\Omega}^2+\|2(\theta_h^1+s_h^1)-(\theta_h^0-s_h^0)\|_{0,\Omega}^2.\end{aligned} \tag{4.4.10}$$

Similarly, taking $\boldsymbol{v}_h = 4\boldsymbol{u}_h^{n+1}$ in the first equation of (4.4.4) and applying (4.4.7), we find that

$$\begin{aligned}&\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\|_{0,\Omega}^2+\|\Lambda\boldsymbol{u}_h^n\|_{0,\Omega}^2+\frac{8\mu_1}{\mathrm{Re}}\Delta t\,\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h^{n+1})\|_{0,\Omega}^2+4\,\eta_1\Delta t\,\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}^2\\ &\quad\le\quad\|\boldsymbol{u}_h^n\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}\|_{0,\Omega}^2+4C_f|\boldsymbol{k}|\Delta t\|\theta_h^{n+1}\|_{0,\Omega}\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}.\end{aligned}$$

Applying Young's inequality with constant $\varepsilon=\frac{\eta_1}{C_f|\boldsymbol{k}|}$ and then, Korn's Lemma, we obtain that

$$\begin{aligned}&\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\|_{0,\Omega}^2+\|\Lambda\boldsymbol{u}_h^n\|_{0,\Omega}^2+\min\left\{\frac{4\mu_1}{\mathrm{Re}},2\eta_1\right\}\Delta t\,\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2\\ &\quad\le\quad\|\boldsymbol{u}_h^n\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}\|_{0,\Omega}^2+\frac{2C_f^2|\boldsymbol{k}|^2}{\eta_1}\Delta t|\theta_h^{n+1}|_{1,\Omega}^2.\end{aligned}$$

Summing over n and using the estimate (4.4.10), we finally deduce that

$$\begin{aligned}\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\|_{0,\Omega}^2+\sum_{m=1}^n\|\Lambda\boldsymbol{u}_h^m\|_{0,\Omega}^2+\min\left\{\frac{4\mu_1}{\mathrm{Re}},2\eta_1\right\}\sum_{m=1}^n\Delta t\,\|\boldsymbol{u}_h^{m+1}\|_{1,\Omega}^2\\ \leq\quad\frac{2C_f^2|\boldsymbol{k}|^2}{\eta_1}\sum_{m=1}^n\Delta t|\theta_h^{m+1}|_{1,\Omega}^2+\|\boldsymbol{u}_h^1\|_{0,\Omega}^2+\|2\boldsymbol{u}_h^1-\boldsymbol{u}_h^0\|_{0,\Omega}^2.\end{aligned}\tag{4.4.11}$$

Finally, from bounds (4.4.11) and (4.4.10) we obtain the results (4.4.8) and (4.4.9). □

We now establish the existence of the solution of (4.4.4).

Theorem 4.4.2. *Assume that the data satisfy*

$$2\kappa_1 s_2<\kappa_0\quad\text{and}\quad 4C_f^2|\boldsymbol{k}|^2+9s_1^2<\frac{2}{C\mathrm{Pr}}\kappa_0 c_p\min\left\{\frac{\mu_1}{\mathrm{Re}},\eta_1\right\}.\tag{4.4.12}$$

Then, problem (4.4.6) admits at least a solution $(\boldsymbol{u}_h^{n+1},\theta_h^{n+1},p_h^{n+1})\in\mathbf{V}_h\times\mathrm{Z}_h\times\mathrm{Q}_h$.

Proof. We proceed analogously as in [67, Thm. 3.2]. For notational convenience, we introduce the following constants

$$\begin{aligned}C_{\boldsymbol{u}}&=C_1\left(\|\theta_h^1+s_h^1\|_{0,\Omega}+\|2(\theta_h^1+s_h^1)-(\theta_h^0+s_h^0)\|_{0,\Omega}+\|\boldsymbol{u}_h^1\|_{0,\Omega}+\|2\boldsymbol{u}_h^1-\boldsymbol{u}_h^0\|_{0,\Omega}\right),\\ C_\theta&=C_2\left(\|\theta_h^1+s_h^1\|_{0,\Omega}+\|2(\theta_h^1+s_h^1)-(\theta_h^0+s_h^0)\|_{0,\Omega}\right).\end{aligned}$$

Now, proceeding by induction on $n\geq 2$, we define a mapping Φ from $\mathbf{V}_h^*\times\mathrm{Z}_h$ into itself by

$$\begin{aligned}&\Phi(\boldsymbol{u}_h^{n+1},\theta_h^{n+1}),(\boldsymbol{v}_h,\psi_h))=\tfrac{1}{2\Delta t}(3\boldsymbol{u}_h^{n+1}-4\boldsymbol{u}_h^n+\boldsymbol{u}_h^{n-1},\boldsymbol{v}_h)+c_1(\boldsymbol{u}_h^{n+1};\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h)+\tfrac{1}{2}(\mathrm{div}\,\boldsymbol{u}_h^{n+1}\,\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h)\\ &+a_1^{\theta_h^{n+1}}(\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h)+(\eta(\theta_h^{n+1})\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h)-(f(\theta_h^{n+1})\boldsymbol{k},\boldsymbol{v}_h)+c_3(\boldsymbol{u}_h^{n+1};\theta_h^{n+1}+s_h^{n+1},\psi_h)+a_3(\theta_h^{n+1},\psi_h)\\ &+\frac{1}{2\Delta t}(3(\theta_h^{n+1}+s_h^{n+1})-4(\theta_h^n+s_h^n)+(\theta_h^{n-1}+s_h^{n-1}),\psi_h)+\frac{1}{2}(\mathrm{div}\,\boldsymbol{u}_h^{n+1}\,(\theta_h^{n+1}+s_h^{n+1}),\psi_h).\end{aligned}\tag{4.4.13}$$

We can note that this mapping is well defined and continuous on $\mathbf{V}_h^*\times\mathrm{Z}_h$. In order to use Brouwer's fixed-point theorem, we take $(\boldsymbol{v}_h,\psi_h)=(\boldsymbol{u}_h^{n+1},\theta_h^{n+1})$ in (4.4.13), apply (4.4.8)-(4.4.9) and denote $C_3:=\min\{\frac{\mu_1}{\mathrm{Re}},\eta_1\}$ to get

$$\begin{aligned}&(\Phi(\boldsymbol{u}_h^{n+1},\theta_h^{n+1}),(\boldsymbol{u}_h^{n+1},\theta_h^{n+1}))\geq C_3\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2-\frac{1}{2\Delta t}\|4\boldsymbol{u}_h^n-\boldsymbol{u}_h^{n-1}\|_{0,\Omega}\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}-C_f|\boldsymbol{k}|\|\theta_h^{n+1}\|_{0,\Omega}\times\\ &\quad\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}+\frac{\kappa_0}{C\mathrm{Pr}}|\theta_h^{n+1}|_{1,\Omega}^2-s_1\|\boldsymbol{u}_h^{n+1}\|_{0,\Omega}|\theta_h^{n+1}|_{1,\Omega}-\frac{1}{2}s_1\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}\|\theta_h^{n+1}\|_{0,\Omega}\\ &\quad-3s_1|\Omega|^{1/2}\|\theta_h^{n+1}\|_{0,\Omega}-\frac{1}{2\Delta t}\|4(\theta_h^n+s_h^n)-(\theta_h^{n-1}+s_h^{n-1})\|_{0,\Omega}\|\theta_h^{n+1}\|_{0,\Omega}\\ &\geq\ C_3\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2+\frac{\kappa_0 c_p}{C\mathrm{Pr}}\|\theta_h^{n+1}\|_{1,\Omega}-\frac{1}{2\Delta t}\widehat{C}_1C_{\boldsymbol{u}}\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}-\left(\frac{1}{2\Delta t}\widehat{C}_2C_\theta-3s_1|\Omega|^{1/2}\right)\|\theta_h^{n+1}\|_{1,\Omega}\\ &\quad-\frac{C_f|\boldsymbol{k}|\varepsilon_1}{2}\|\theta_h^{n+1}\|_{1,\Omega}^2-\frac{C_f|\boldsymbol{k}|}{2\varepsilon_1}\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2-\frac{3s_1\varepsilon_2}{4}\|\theta_h^{n+1}\|_{1,\Omega}^2-\frac{3s_1}{4\varepsilon_2}\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2.\end{aligned}$$

Now, taking $\varepsilon_1 = \frac{2C_f|\boldsymbol{k}|}{C_3}$, $\varepsilon_2 = \frac{3s_1}{C_3}$, applying the second inequality given in (4.4.12), and recalling that $(\alpha+\beta) \leq \sqrt{2}(\alpha^2+\beta^2)^{1/2} \ \ \forall\, \alpha, \beta \in \mathrm{R}$, we obtain

$$(\Phi(\boldsymbol{u}_h^{n+1}, \theta_h^{n+1}), (\boldsymbol{u}_h^{n+1}, \theta_h^{n+1})) \geq \widetilde{C}_1 \left(\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2 + \|\theta_h^{n+1}\|_{1,\Omega}^2 \right) - \widetilde{C}_2 \left(\|\boldsymbol{u}_h^{n+1}\|_{1,\Omega}^2 + \|\theta_h^{n+1}\|_{1,\Omega}^2 \right)^{1/2},$$

where $\widetilde{C}_1 = \frac{1}{2} \min \left\{ C_3, \frac{\kappa_0 c_p}{C\mathrm{Pr}} \right\}$ and $\widetilde{C}_2 = \max \left\{ \frac{1}{2\Delta t} \widehat{C}_1 C_{\boldsymbol{u}}, \frac{1}{2\Delta t} \widehat{C}_2 C_\theta - 3s_1 |\Omega|^{1/2} \right\}$. So, the right hand side is nonnegative on the sphere of radius $r := \frac{\widetilde{C}_2}{\widetilde{C}_1}$. Consequently, applying Brouwer's fixed-point theorem, we get the existence of a solution $(\boldsymbol{u}_h^{n+1}, \theta^{n+1_h})$ of $\Phi(\boldsymbol{u}_h^{n+1}, \theta_h^{n+1}) = 0$, concluding the proof. □

It is important to remark here that the existence of solution of system (4.4.4) could be obtained directly from an application of Brouwer fixed-point theorem. In fact, now we define the auxiliary problem $(\widetilde{\Phi}(\boldsymbol{u}_h^{n+1}, \theta_h^{n+1}, p_h^{n+1}), (\boldsymbol{v}_h, \psi_h, q_h)) := (\Phi(\boldsymbol{u}_h^{n+1}, \theta_h^{n+1}), (\boldsymbol{u}_h^{n+1}, \theta_h^{n+1})) + b(\boldsymbol{v}_h, p_h^{n+1}) - b(\boldsymbol{u}_h^{n+1}, q_h)$, which is continuous from $\mathbf{V}_h \times \mathrm{Z}_h \times \mathrm{Q}_h$ into itself. Thus, when one take $\boldsymbol{v}_h = \boldsymbol{u}_h^{n+1}$, $\psi_h = \theta_h^{n+1}$ and $q_h = p_h^{n+1}$, the additional term disappears and the proof would be exactly the same. The following result asserts the unique solvability of (4.4.4).

Theorem 4.4.3. *Assume that Δt sufficiently small and that $L_{s_1} < 1/2$. Then the scheme defined in (4.4.4) has a unique solution.*

Proof. It follows analogously to the proof of Theorem 4.3.5. For more details see [8, Thm. 5.3]. □

4.4.2 A mixed-primal finite element method

Numerical methods based on mixed-primal variational formulations have the advantage that important physical variables (as pseudostress and vorticity) can be approximated directly. Before stating the corresponding Galerkin scheme, we proceed to derive a mixed-primal weak formulation for the model problem. In order to simplify the exposition of this section we will restrict to the stationary counterpart of (4.2.1)-(4.2.3), in this case written as follows

$$\begin{aligned} \boldsymbol{u} \cdot \nabla \boldsymbol{u} \ - \ \mathrm{Re}^{-1} \, \mathbf{div}[\mu(\theta)\boldsymbol{\varepsilon}(\boldsymbol{u})] \ + \ \nabla p + \eta(\theta)\boldsymbol{u} &= f(\theta)\boldsymbol{k}, \\ \mathrm{div} \ \boldsymbol{u} &= 0, \\ -\mathrm{CPr}^{-1} \ \mathrm{div}(\kappa \nabla \theta) \ + \ \boldsymbol{u} \cdot \nabla \theta \ + \ \boldsymbol{u} \cdot \nabla s(\theta) \ &= \ 0, \end{aligned} \tag{4.4.14}$$

and associated with Dirichlet boundary conditions $\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}}$ and $\theta = 0$ on Γ. Here the velocity datum $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{1/2}(\Gamma)$ verifies the compatibility condition $\int_\Gamma \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{\nu} = 0$, where $\boldsymbol{\nu}$ denotes the unit outward normal on Γ. Proceeding as in [10], we introduce the strain rate tensor as an auxiliary unknown

$$\mathbf{t} := \boldsymbol{\varepsilon}(\boldsymbol{u}) = \nabla \boldsymbol{u} - \boldsymbol{\gamma} \in \mathbb{L}^2_{\mathtt{tr}}(\Omega),$$

where $\boldsymbol{\gamma} = \boldsymbol{\omega}(\boldsymbol{u}) \in \mathbb{L}^2_{\mathtt{skew}}(\Omega)$ is the skew-symmetric part of the velocity gradient $\nabla \boldsymbol{u}$ and the involved functional spaces are

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

Also the total stress (or pseudostress tensor, including diffusive, convective, and pressure contributions) is regarded as a new unknown

$$\boldsymbol{\sigma} := \mathrm{Re}^{-1}\,\mu(\theta)\mathbf{t} - (\boldsymbol{u}\otimes\boldsymbol{u}) - p\,\mathbb{I}, \tag{4.4.15}$$

which implies that the second equation in (4.4.14) together with (4.4.15) are equivalent to the relations

$$\mathrm{Re}^{-1}\,\mu(\theta)\,\mathbf{t} - (\boldsymbol{u}\otimes\boldsymbol{u})^{\mathtt{d}} = \boldsymbol{\sigma}^{\mathtt{d}}, \quad \text{and} \quad p = -\frac{1}{n}\mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{u}\otimes\boldsymbol{u}) \quad \text{in} \quad \Omega.$$

Consequently, we arrive at the following coupled system without pressure

$$\begin{aligned}
\mathrm{Re}^{-1}\,\mu(\theta)\,\mathbf{t} - (\boldsymbol{u}\otimes\boldsymbol{u})^{\mathtt{d}} &= \boldsymbol{\sigma}^{\mathtt{d}} && \text{in} \quad \Omega, && (4.4.16)\\
\mathbf{t} + \boldsymbol{\gamma} &= \nabla\boldsymbol{u} && \text{in} \quad \Omega, && (4.4.17)\\
\eta(\theta)\,\boldsymbol{u} - \mathbf{div}\,\boldsymbol{\sigma} &= f(\theta)k && \text{in} \quad \Omega, && (4.4.18)\\
-\mathrm{CPr}^{-1}\,\mathrm{div}(\kappa\nabla\theta) + \boldsymbol{u}\cdot\nabla\theta + \boldsymbol{u}\cdot\nabla s(\theta) &= 0 && \text{in} \quad \Omega, && (4.4.19)\\
\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} \quad \text{and} \quad \theta &= 0 && \text{on} \quad \Gamma, && (4.4.20)\\
\int_\Omega \mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{u}\otimes\boldsymbol{u}) &= 0. && && (4.4.21)
\end{aligned}$$

Note that the incompressibility constraint is implicitly present in (4.4.16), and the pressure being in $\mathrm{L}^2_0(\Omega)$ is guaranteed by the equivalent statement in (4.4.21).

A weak form for this problem is obtained after testing (4.4.16)-(4.4.17) against suitable functions and imposing the symmetry of $\boldsymbol{\sigma}$; multiplying (4.4.18) by $\boldsymbol{\tau}\in\mathbb{H}(\mathbf{div};\Omega)$, integrating by parts and using the velocity boundary condition in (4.4.20); and taking a similar weak formulation for the energy equation as in (4.3.2), appropriately modified to incorporate the temperature boundary data in (4.4.20). That is, the temperature trial and test space will be

$$\mathrm{H}^1_0(\Omega) := \left\{\psi\in\mathrm{H}^1(\Omega): \quad \psi = 0 \ \text{ on } \ \Gamma\right\}.$$

Moreover, the specific structure of the problem and the orthogonal decomposition $\mathbb{H}(\mathbf{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega)\oplus\mathbb{RI}$, allows us to look for stresses in the space

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

The weak formulation then reads: Find $(\mathbf{t},\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\gamma},\theta)\in\mathbb{L}^2_{\mathtt{tr}}(\Omega)\times\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathtt{skew}}(\Omega)\times\mathrm{H}^1_0(\Omega)$ such that

$$\begin{aligned}
\mathrm{Re}^{-1}\int_\Omega \mu(\theta)\mathbf{t}:\mathbf{s} - \int_\Omega (\boldsymbol{u}\otimes\boldsymbol{u})^{\mathtt{d}}:\mathbf{s} - \int_\Omega \boldsymbol{\sigma}^{\mathrm{d}}:\mathbf{s} &= 0 \quad \forall\,\mathbf{s}\in\mathbb{L}^2_{\mathtt{tr}}(\Omega),\\
\int_\Omega \mathbf{t}:\boldsymbol{\tau}^{\mathrm{d}} + \int_\Omega \boldsymbol{\gamma}:\boldsymbol{\tau} + \int_\Omega \boldsymbol{u}\cdot\mathbf{div}\,\boldsymbol{\tau} &= \langle\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{u}_{\mathrm{D}}\rangle_\Gamma \quad \forall\,\boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega),\\
-\int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\sigma} - \int_\Omega \boldsymbol{\sigma}:\boldsymbol{\eta} + \int_\Omega \eta(\theta)\boldsymbol{u}\cdot\boldsymbol{v} &= \int_\Omega f(\theta)\boldsymbol{k}\cdot\boldsymbol{v} \quad \forall\,(\boldsymbol{v},\boldsymbol{\eta})\in\mathbf{L}^2(\Omega)\times\mathbb{L}^2_{\mathtt{skew}}(\Omega),\\
\mathrm{CPr}^{-1}\int\kappa\nabla\theta\cdot\nabla\psi &= -\int_\Omega \psi\,\boldsymbol{u}\cdot\nabla(\theta+s(\theta)) \quad \forall\ \psi\ \in\mathrm{H}^1_0(\Omega).
\end{aligned}$$

A further inspection of this formulation eventually reveals the lack of sufficient regularity (in particular for velocity and stress), which suggests the use of augmentation techniques in the spirit of e.g. [9]. We therefore incorporate residual, Galerkin-type terms in weak form

$$
\begin{aligned}
\kappa_1 \int_\Omega \left\{ \boldsymbol{\sigma}^{\mathtt{d}} + (\boldsymbol{u}\otimes\boldsymbol{u})^{\mathtt{d}} + \mathrm{Re}^{-1}\mu(\theta)\mathbf{t}\right\} : \boldsymbol{\tau}^{\mathrm{d}} = 0 \qquad & \forall\, \boldsymbol{\tau}\in \mathbb{H}_0(\mathbf{div};\Omega)\,,\\
\kappa_2 \int_\Omega \{\mathbf{div}\,\boldsymbol{\sigma} - \eta(\theta)\boldsymbol{u}\}\cdot \mathbf{div}\boldsymbol{\tau} = -\,\kappa_2 \int_\Omega f(\theta)\boldsymbol{k}\cdot\mathbf{div}\boldsymbol{\tau} \qquad & \forall\, \boldsymbol{\tau}\in \mathbb{H}_0(\mathbf{div};\Omega)\,,\\
\kappa_3 \int_\Omega \{\boldsymbol{\varepsilon}(\boldsymbol{u}) - \mathbf{t}\}\cdot\boldsymbol{\varepsilon}(\boldsymbol{v}) = 0 \qquad & \forall\, \boldsymbol{v}\in \mathbf{H}^1(\Omega)\,,\\
\kappa_4 \int_\Omega \{\boldsymbol{\gamma} - \boldsymbol{\omega}(\boldsymbol{u})\} : \boldsymbol{\eta} = 0 \qquad & \forall\, \boldsymbol{\eta}\in \mathbb{L}^2_{\mathtt{skew}}(\Omega),
\end{aligned}
$$

where κ_i, $i\in\{1,2,3,4\}$ are positive parameters. Denoting $H := \mathbb{L}^2_{\mathtt{tr}}(\Omega)\times\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)\times\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, $\vec{\mathbf{t}} := (\mathbf{t},\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\gamma})$, and $\vec{\mathbf{s}} := (\mathbf{s},\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{\eta})$, we arrive at the following augmented mixed-primal formulation of the initial coupled problem (4.4.14): Find $(\vec{\mathbf{t}},\theta)\in H\times \mathrm{H}^1_0(\Omega)$ such that

$$
\begin{aligned}
\mathbf{A}_\theta(\vec{\mathbf{t}},\vec{\mathbf{s}}) + \mathbf{B}_{\boldsymbol{u}}(\vec{\mathbf{t}},\vec{\mathbf{s}}) &= \mathcal{F}_\theta(\vec{\mathbf{s}}) + \mathcal{F}_{\mathrm{D}}(\vec{\mathbf{s}}) \qquad \forall\, \vec{\mathbf{s}}\in H,\\
\mathbf{a}(\theta,\psi) &= \mathcal{G}_{\boldsymbol{u},\theta}(\psi) \qquad \forall\, \psi\in \mathrm{H}^1_0(\Omega),
\end{aligned}
\tag{4.4.22}
$$

where, given an arbitrary $(\boldsymbol{w},\phi)\in\mathbf{H}^1(\Omega)\times\mathrm{H}^1_0(\Omega)$, the forms $\mathbf{A}_\phi$, $\mathbf{B}_{\boldsymbol{w}}$, $\mathbf{a}$, and the fuctionals $\mathcal{F}_\phi$, $\mathcal{F}_{\mathrm{D}}$, and $\mathcal{G}_{\boldsymbol{w},\phi}$ are defined as

$$
\begin{aligned}
\mathbf{A}_\phi(\vec{\mathbf{t}},\vec{\mathbf{s}}) &:= \mathrm{Re}^{-1}\int \mu(\phi)\mathbf{t} : \left\{\mathbf{s}-\kappa_1\boldsymbol{\tau}^{\mathrm{d}}\right\} + \int_\Omega \mathbf{t} : \left\{\boldsymbol{\tau}^{\mathrm{d}} - \kappa_3\boldsymbol{\varepsilon}(\boldsymbol{u})\right\} - \int_\Omega \boldsymbol{\sigma}^{\mathtt{d}} : \left\{\mathbf{s}-\kappa_1\boldsymbol{\tau}^{\mathrm{d}}\right\}\\
&\quad + \int_\Omega \boldsymbol{u}\cdot\mathbf{div}\,\boldsymbol{\tau} - \int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\sigma} + \int_\Omega \boldsymbol{\gamma}:\boldsymbol{\tau} - \int_\Omega \boldsymbol{\sigma}:\boldsymbol{\eta} + \int_\Omega \eta(\phi)\boldsymbol{u}\cdot\left\{\boldsymbol{v}-\kappa_2\,\mathbf{div}\,\boldsymbol{\tau}\right\}\\
&\quad - \kappa_4\int_\Omega \boldsymbol{\omega}(\boldsymbol{u}):\boldsymbol{\eta} + \kappa_2\int_\Omega \mathbf{div}\,\boldsymbol{\sigma}\cdot\mathbf{div}\,\boldsymbol{\tau} + \kappa_3\int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}):\boldsymbol{\varepsilon}(\boldsymbol{v}) + \kappa_4\int_\Omega \boldsymbol{\gamma}:\boldsymbol{\eta},\\
\mathbf{B}_{\boldsymbol{w}}(\vec{\mathbf{t}},\vec{\mathbf{s}}) &:= \int_\Omega (\boldsymbol{u}\otimes\boldsymbol{w})^{\mathtt{d}} : \left\{\kappa_1\boldsymbol{\tau}^{\mathrm{d}} - \mathbf{s}\right\}, \qquad \mathbf{a}(\theta,\psi) := \mathrm{CPr}^{-1}\int_\Omega \kappa\nabla\theta\cdot\nabla\psi,\\
\mathcal{F}_\phi(\vec{\mathbf{s}}) &:= \int_\Gamma f(\phi)\boldsymbol{k}\cdot\left\{\boldsymbol{v}-\kappa_2\mathbf{div}\,\boldsymbol{\tau}\right\}, \quad \mathcal{F}_{\mathrm{D}}(\vec{\mathbf{s}}) := \langle\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{u}_{\mathrm{D}}\rangle_\Gamma, \quad \mathcal{G}_{\boldsymbol{w},\phi}(\psi) := -\int_\Omega \psi\,\boldsymbol{w}\cdot\nabla(\phi+s(\phi)),
\end{aligned}
$$

for all $\vec{\mathbf{t}},\vec{\mathbf{s}}\in H$, $\theta,\psi\in\mathrm{H}^1_0(\Omega)$.

The solvability of (4.4.22) can be established by invoking a fixed-point approach between a mixed formulation for the momentum and mass equations and a primal formulation for the energy equation; which can be carried out following [10, 9, 43]. In contrast with those works, the solvability analysis of the Navier-Stokes equations can be done in our case by applying a further fixed-point iteration that relies on existence results for Brinkman equations, and we require four residual terms and the use of Korn's inequality. Since θ lives in $\mathrm{H}^1_0(\Omega)$, we can then apply a similar treatment as in [14, 83]. These steps go beyond the scope of this chapter and will be addressed in the next chapter of this thesis.

The Galerkin scheme and specific finite element subspaces. For each $K\in\mathcal{T}_h$ let us recall the definition of the local Raviart-Thomas space of order k as $\mathbf{RT}_k(K) := \mathbb{P}_k(K)^d \oplus \mathbb{P}_k(K)\,\boldsymbol{x}$, where $\boldsymbol{x}$ is a generic vector in $\mathbb{R}$. Then, we consider finite-dimensional subspaces $\mathbb{H}^{\boldsymbol{t}}_h\subset\mathbb{L}^2_{\mathtt{tr}}(\Omega)$, $\mathbb{H}^{\boldsymbol{\sigma}}_h\subset\mathbb{H}_0(\mathbf{div};\Omega)$, $\mathbf{H}^{\boldsymbol{u}}_h\subset\mathbf{H}^1(\Omega)$, $\mathbb{H}^{\boldsymbol{\gamma}}_h\subset\mathbb{L}^2_{\mathtt{skew}}(\Omega)$, $\mathrm{H}^\theta_h\subset\mathrm{H}^1_0(\Omega)$ defined as

$$
\mathbb{H}^{\boldsymbol{t}}_h := \left\{\mathbf{s}_h\in\mathbb{L}^2_{\mathtt{tr}}(\Omega) : \quad \mathbf{s}_h\Big|_K\in\mathbb{P}_k(K) \quad \forall\, K\in\mathcal{T}_h\right\},
$$

$$
\begin{aligned}
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \Big\{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \mathbf{c}^{\mathtt{t}}\,\boldsymbol{\tau}_h|_K \in \mathbf{RT}_k(K) \quad \forall\, \mathbf{c} \in \mathbb{R}^d\,, \quad \forall\, K \in \mathcal{T}_h \Big\}\,,\\
\mathbf{H}_h^{\boldsymbol{u}} &:= \Big\{ \boldsymbol{v}_h \in \mathbf{C}(\overline{\Omega}) : \quad \boldsymbol{v}_h|_K \in \mathbb{P}_{k+1}(K)^d \quad \forall\, K \in \mathcal{T}_h \Big\}\,,\\
\mathbb{H}_h^{\gamma} &:= \Big\{ \boldsymbol{\eta}_h \in \mathbb{L}^2_{\mathtt{skew}}(\Omega) : \quad \boldsymbol{\eta}_h|_K \in \mathbb{P}_k(K)^{d\times d} \quad \forall\, K \in \mathcal{T}_h \Big\},\\
\mathrm{H}_h^{\theta} &:= \Big\{ \psi_h \in C(\overline{\Omega}) \cap \mathrm{H}^1_0(\Omega) : \quad \psi_h|_K \in \mathbb{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \Big\}\,.
\end{aligned}
$$

Denoting $H_h := \mathbb{H}_h^{\boldsymbol{t}} \times \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}} \times \mathbb{H}_h^{\gamma}$, $\vec{\mathbf{t}}_h := (\mathbf{t}_h, \boldsymbol{\sigma}_h, \boldsymbol{u}_h, \boldsymbol{\gamma}_h)$, and $\vec{\mathbf{s}}_h := (\mathbf{s}_h, \boldsymbol{\tau}_h, \boldsymbol{v}_h, \boldsymbol{\eta}_h)$, a Galerkin scheme reads: Find $(\vec{\mathbf{t}}_h, \theta_h) \in H_h \times \mathrm{H}_h^{\theta}$ such that

$$
\begin{aligned}
\mathbf{A}_{\theta_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) \,+\, \mathbf{B}_{\boldsymbol{u}_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) &= \mathcal{F}_{\theta_h}(\vec{\mathbf{s}}_h) \,+\, \mathcal{F}_{\mathrm{D}}(\vec{\mathbf{s}}_h) \qquad \forall\, \vec{\mathbf{s}}_h \in H_h,\\
\mathbf{a}(\theta_h, \psi_h) &= \mathcal{G}_{\boldsymbol{u}_h,\theta_h}(\psi_h) \qquad \forall\, \psi_h \in \mathrm{H}_h^{\theta}.
\end{aligned}
\tag{4.4.23}
$$

The well-posedness of (4.4.23) will be also based on a suitable adaptation of the continuous analysis of (4.4.22) to the present context (see, e.g. [10, 43]).

4.4.3 Consistent linearisation

Problem (4.4.4) entails solving a set of nonlinear equations at each timestep. For this we will employ Newton Raphson's method, which features quadratic convergence provided the initial guess is sufficiently close to the zone of attraction. For a generic nonlinear problem $\mathbf{F}(\mathbf{w}) = \mathbf{0}$, one produces a sequence $\{\mathbf{w}^k\}_k$ converging quadratically to $\mathbf{w}$, through the iterates

$$
\mathcal{D}\mathbf{F}(\mathbf{w}^k)[\delta\mathbf{w}] = -\mathbf{F}(\mathbf{w}^k) \quad \text{where} \quad \delta\mathbf{w} = \mathbf{w}^{k+1} - \mathbf{w}^k, \quad \mathbf{w}^0 = (\boldsymbol{u}_h^n, p_h^n, \theta_h^n),
$$

where $\mathcal{D}\mathbf{F}(v)[\delta v]$ denotes the Gâteaux derivative of the functional $\mathbf{F}$ along the direction δv. Then at each Newton step k we solve the linear problem

$$
\begin{aligned}
&\frac{3}{2\Delta t}(\delta\boldsymbol{u}_h, \boldsymbol{v}_h) + c_1(\delta\boldsymbol{u}_h; \boldsymbol{u}_h^k, \boldsymbol{v}_h) + c_1(\boldsymbol{u}_h^k; \delta\boldsymbol{u}_h, \boldsymbol{v}_h) + \frac{2}{\mathrm{Re}}\big(\mu'(\theta_h^k)\delta\theta_h\boldsymbol{\varepsilon}(\boldsymbol{u}_h^k), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big)\\
&\quad + \frac{2}{\mathrm{Re}}\big(\mu(\theta_h^k)\boldsymbol{\varepsilon}(\delta\boldsymbol{u}_h), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big) + (\eta'(\theta_h^k)\delta\theta_h\boldsymbol{u}_h^k, \boldsymbol{v}_h) + (\eta(\theta_h)\delta\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, \delta p_h) - (f'(\theta_h^k)\delta\theta_h\boldsymbol{k}, \boldsymbol{v}_h)\\
&\qquad = \frac{1}{\Delta t}(2\boldsymbol{u}_h^n - \frac{1}{2}\boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h) + \big(\mathcal{R}_h^1(\boldsymbol{u}_h^k, p_h^k, \theta_h^k), \boldsymbol{v}_h\big) \qquad \forall \boldsymbol{v}_h \in \mathbf{V}_h,\\
&b(\delta\boldsymbol{u}_h, q_h) = (\mathcal{R}_h^2(\boldsymbol{u}_h^k, p_h^k, \theta_h^k), q_h) \qquad \forall q_h \in \mathrm{Q}_h,\\
&\frac{3}{2\Delta t}(\delta\theta_h + s'(\theta_h)\delta\theta_h, \psi_h) + c_3(\delta\boldsymbol{u}_h; \theta_h^k + s(\theta_h^k), \psi_h) + c_3(\boldsymbol{u}_h^k; \delta\theta_h + s'(\theta_h^k)\delta\theta_h, \psi_h) + a_3(\delta\theta_h, \psi_h)\\
&\qquad = \frac{1}{\Delta t}(2[\theta_h^n + s_h^n] - \frac{1}{2}[\theta_h^{n-1} + s_h^{n-1}], \psi_h) + (\mathcal{R}_h^3(\boldsymbol{u}_h^k, p_h^k, \theta_h^k), \psi_h) \qquad \forall \psi_h \in \mathrm{Z}_h.
\end{aligned}
$$

where the terms $\mathcal{R}_h^i$ stand for the Newton residuals associated to the momentum, mass, and energy-enthalpy equations.

One readily notes that as we increase the Rayleigh number, the coupling between the Navier-Stokes and the temperature equation becomes stronger. This makes the radius of convergence for the Newton method smaller (see e.g. [73]). As the new initial guess in the time-dependent method is the solution at the previous timestep, this condition reflects on a restriction on the timestep, independently of the CFL condition.

4.5 Numerical verification

We stress that the zero-mean condition enforcing the uniqueness of the pressure (for the schemes based on the primal finite element formulation from Section 4.4.1) is implemented using a pressure penalisation approach. All nonlinear systems undergo a Newton linearisation with fixed residual tolerance of 1E-6. In addition, the resulting linear solves are performed with the direct method SuperLU.

4.5.1 Experimental convergence for the semidiscrete and fully discrete methods

For our first example we produce the error history associated with the finite element approximation. Let us consider the following closed-form solutions to the stationary Boussinesq equations with enthalpy, defined on the unit square domain $\Omega = (0,1)^2$:

$$\boldsymbol{u}(x,y) = \begin{pmatrix} \sin(\pi x)^2 \sin(\pi y)^2 \cos(\pi y) \\ -\frac{1}{3}\sin(2\pi x)\sin(\pi y)^3 \end{pmatrix}, \quad p(x,y) = 10(x^4 - y^4), \quad \theta(x,y) = 1 + \sin(\pi x)\cos(\pi y). \tag{4.5.1}$$

These functions are smooth and they are used to generate non-homogeneous forcing and source terms. The vertical walls constitute Γ_D^θ and the bottom and top lids of the domain conform Γ_N^θ. The temperature-dependent viscosity, porosity, buoyancy, and enthalpy functions are taken as

$$\eta(\theta) := 2 + \tanh(\frac{1}{2} - \theta), \quad \mu(\theta) := \exp(-\theta), \quad f(\theta) := \frac{\mathrm{Ra}}{\mathrm{Pr}\,\mathrm{Re}^2}\theta, \quad s(\theta) := 1 + \tanh(1-\theta), \tag{4.5.2}$$

respectively, and the remaining parameters specifying this steady-state version of (4.2.1)-(4.2.3) are $\mathrm{Re} = 10$, $\mathrm{Ra} = 100$, $\mathrm{Pr} = 0.71$, γ =1E-6. We then construct a sequence of successively refined meshes for Ω and proceed to compute errors between approximate and exact solutions, together with local convergence rates. The outcome of this error study is depicted in Table 4.1, which shows optimal convergence $O(h^{k+1})$ for velocity, pressure, and temperature in their natural norms.

The error associated to the time discretisation is assessed by considering the original transient problem, with enthalpy, and employing the following exact solutions (proposed in [67] for the study of the Boussinesq approximation without enthalpy), defined on the three-dimensional domain $\Omega = (0,1)^3$:

$$\boldsymbol{u}(x,y,z,t) = \begin{pmatrix} x^2 + xy - z^2 + yz \\ -2xy - \frac{1}{2}y^2 + 2yz - 2xz \\ z^2 + y^2 - x^2 + 3xy \end{pmatrix} \sin(t), \quad p(x,y,z,t) = [x - y + 3z - 3/2]\sin(t),$$
$$\theta(x,y,z,t) = 2 + [x^2 + y^2 + z^2 + 1]\sin(t).$$

The regularity of these solutions implies that the spatial finite element discretisation using a method with $k = 1$ will be machine-precision accurate, and so the total error will practically coincide with the time discretisation error. The temperature-dependent functions are taken as in (4.5.2). We proceed to discretise the time interval into a sequence of successively refined grids and compute accumulative errors, defined for a generic field scalar or vector field v as $E(v) := [\Delta t \sum_{n=1}^{N} |v_h^n - v(t^n)|^2]^{1/2}$, up to the adimensional final time $t_f = 1$. The error history is displayed in Table 4.2, showing a second order convergence, consistent with the BDF2 algorithm employed.

DoF	h	$\|\boldsymbol{u}-\boldsymbol{u}_h\|_1$	rate	$\|p-p_h\|$	rate	$\|\theta-\theta_h\|_1$	rate	it
				$k=1$				
84	0.7071	4.4051	–	0.6275	–	0.3791	–	5
268	0.3536	0.8306	2.407	0.1468	2.096	0.1027	1.884	4
948	0.1768	0.1334	2.638	0.0357	2.038	0.02654	1.953	4
3556	0.0884	0.0239	2.478	0.0088	2.013	0.0067	1.983	4
13764	0.0442	0.0051	2.226	0.0022	2.004	0.0017	1.993	4
54148	0.0221	0.0012	2.074	0.0006	2.001	0.0004	1.997	5
214788	0.0110	0.0003	2.015	0.0001	1.999	0.0001	1.999	4
				$k=2$				
172	0.7071	0.8695	–	0.102	–	0.0994	–	4
588	0.3536	0.1741	2.320	0.0145	2.813	0.0130	2.931	5
2164	0.1768	0.0253	2.779	0.0019	2.873	0.0016	2.991	5
8292	0.0884	0.0034	2.907	0.0003	2.928	0.0002	3.000	5
32452	0.0442	0.0004	2.954	3.38e-5	2.949	2.56e-5	3.001	5
128388	0.0221	6.16e-5	2.824	6.36e-6	2.961	3.21e-6	2.997	5
510724	0.0110	1.43e-5	2.983	1.51e-6	2.984	8.01e-7	2.997	4

Table 4.1: Error history (errors on a sequence of successively refined grids, experimental convergence rates, and Newton iteration count at each refinement level) associated to the spatial discretisation using the finite element spaces (4.4.1) with $k=1$ and $k=2$ (table produced by the author).

We also generate the error history associated with the mixed-primal discretisation discussed in Section 4.4.2. The closed-form solutions are those in (4.5.1) and the model parameters and temperature-dependent functions are as in the first example above; while the stabilisation constants needed for the augmented formulation take the values $\kappa_1 = \kappa_2 = \mu_1\mu_2^{-2}$, $\kappa_3 = \mu_1/2$, $\kappa_4 = \mu_1/4$ (and where the constants μ_1, μ_2 are the bounds for the viscosity introduced in (4.2.5)). The problem in this case is solved in terms of strain rate, total stress (including viscous and convective contributions), velocity, vorticity tensor, and temperature; and each individual error is measured in its natural norm. The error decay and the corresponding convergence rates are reported in Table 4.3, which confirms an optimal convergence rate. The Newton iteration count is similar to the one observed in Table 4.1.

4.5.2 Benchmark test: natural convection of air

We further validate the numerical method against a well-documented benchmark, the natural convection of air in a differentially heated square cavity $\Omega = (0,1)^2$. In this problem we do not have the enthalpy terms, we do not consider the temperature-dependent drag contribution, and the viscosity and conduction coefficients are constant. Therefore the scaling of the equations is as follows

$$\begin{aligned} \partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} - \mathrm{Pr}\,\Delta\boldsymbol{u} + \nabla p + \eta(\theta)\boldsymbol{u} &= \mathrm{Ra}\,\mathrm{Pr}\,\theta\boldsymbol{k}, \\ \operatorname{div}\boldsymbol{u} &= 0 \qquad \text{in } \Omega\times(0,t_f], \\ \partial_t\theta + \boldsymbol{u}\cdot\nabla\theta - \Delta\theta &= 0, \end{aligned}$$

Δt	$E(\boldsymbol{u})$	rate	$E(p)$	rate	$E(\theta)$	rate	avg(it)
1	0.1355	–	1.2943	–	0.2627	–	5
0.25	0.0176	1.938	0.1640	1.925	0.0329	1.890	4.2
0.0613	0.0023	1.990	0.0251	1.934	0.0041	1.958	4.2
0.0151	0.0003	1.983	0.0032	1.962	0.0005	1.947	4.2

Table 4.2: Time discretisation errors produced with a BDF2 method on different timestep resolutions, convergence rates, and average number of Newton iterations (table produced by the author).

DoF	h	$\|\mathbf{t}-\mathbf{t}_h\|$	rate	$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div}}$	rate	$\|\boldsymbol{u}-\boldsymbol{u}_h\|_1$	rate	$\|\gamma-\gamma_h\|$	rate	$\|\theta-\theta_h\|_1$	rate	it
83	0.7071	1.2330	–	1.2681	–	1.5671	–	1.2547	–	1.5651	–	3
283	0.3536	0.6253	0.826	0.7302	0.796	1.0193	0.633	0.7187	0.803	0.8465	0.886	4
1043	0.1768	0.3402	0.904	0.4056	0.847	0.5788	0.803	0.3641	0.895	0.4328	0.967	4
4003	0.0884	0.1764	0.927	0.2003	0.934	0.3077	0.911	0.1842	0.954	0.2177	0.991	4
15683	0.0442	0.0831	0.954	0.1401	0.929	0.1623	0.925	0.0917	1.022	0.1090	0.997	4
62083	0.0221	0.0462	0.960	0.0892	0.972	0.0958	0.946	0.0483	0.972	0.0548	0.998	4
247043	0.0111	0.0247	0.985	0.0426	0.977	0.0521	0.951	0.0258	0.931	0.0274	0.994	4
985603	0.0055	0.0124	0.992	0.0225	0.985	0.0311	0.954	0.0123	1.003	0.0139	0.978	4

Table 4.3: Error history (errors on a sequence of successively refined grids, experimental convergence rates, and Newton iteration count at each refinement level) associated to the spatial discretisation using the mixed-primal formulation from Section 4.4.2, using a lowest-order scheme with $k = 0$ (table produced by the author).

where $\boldsymbol{k} = (0, 1)^T$, and the only non-dimensional parameters and their values are Ra =1E5 and Pr = 0.71. We use a constant timestep of $\Delta t = 0.001$ and employ a rather coarse mesh with meshsize $h = \sqrt{2}/64$. The initial conditions on the domain interior are $\boldsymbol{u}(0) = \mathbf{0}$ and $\theta(0) = 0.5$, and we prescribe $\theta = 1$ on the left and $\theta = 0$ on the right walls of the cavity (also for the initial datum). The upper and lower plates constitute the boundary Γ_N^θ, where we set zero-flux boundary conditions for temperature (representing insulated walls); and on all four sides of the container we impose no-slip velocities $\boldsymbol{u} = \mathbf{0}$.

The simulation is run until the final, dimensionless time $t_f = 0.5$, using a Taylor-Hood finite element family for the approximation of velocity and pressure (i.e., $k = 1$), and the pressure penalty parameter takes the value γ =1E-7. The flow is driven by the difference of temperature and examples of velocity, pressure and temperature distribution at the final time are depicted in Figure 4.1(a-c). We observe well-defined temperature profiles and the expected recirculation velocity patterns. A more quantitative study is done by extracting the approximate solutions for temperature and vertical velocity on the horizontal midlines at $y = 0.5$ and plotting them against properly rotated published benchmark values from [154] (which were generated using the method of discrete singular convolution). A reasonably close match is confirmed by looking at Figure 4.1(d,e), where we emphasise that our results come from coarse mesh computations.

We carry out further comparisons based on the average Nusselt number on the hot wall of the cavity,

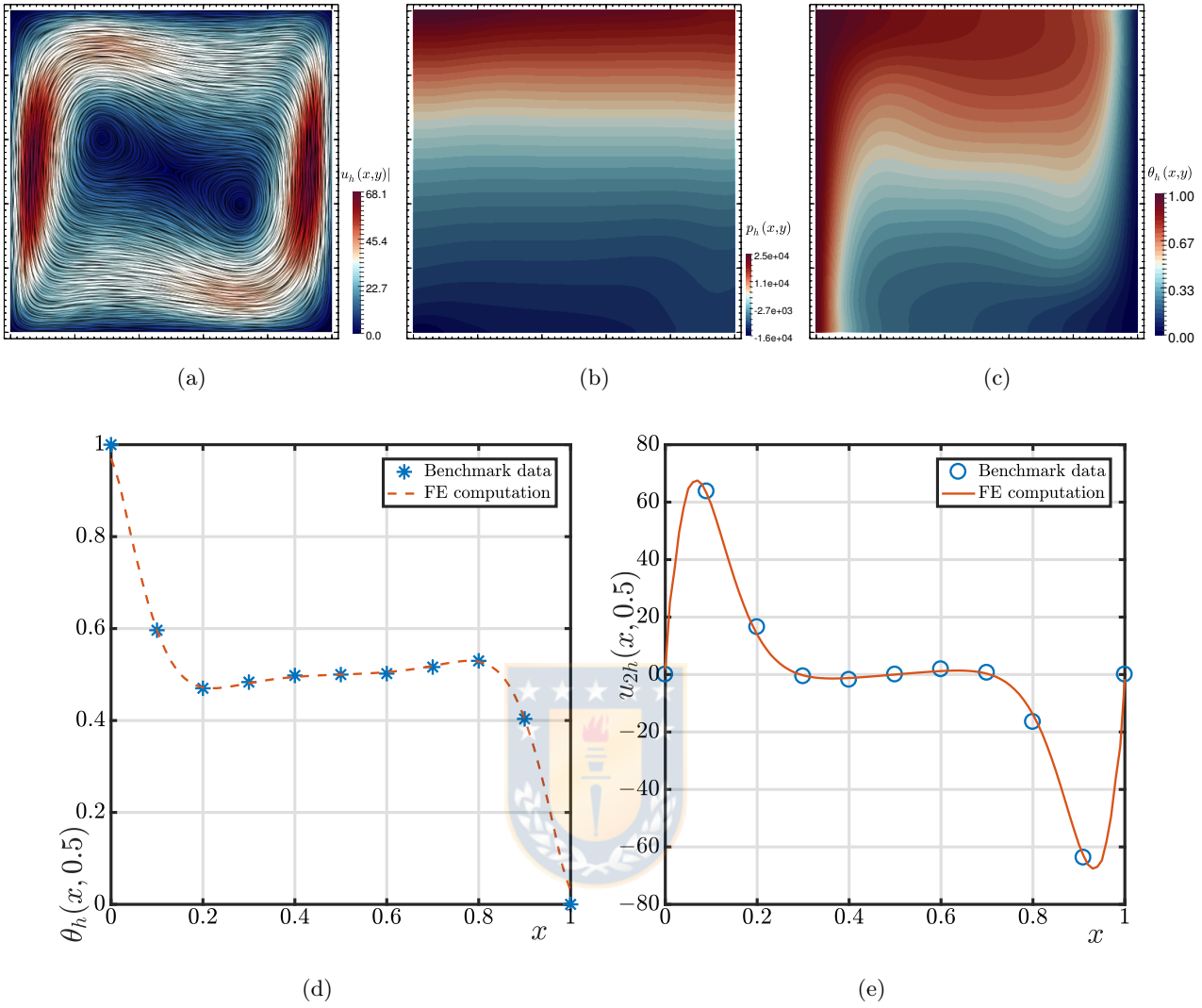


Figure 4.1: Velocity, pressure, and temperature profiles for the 2D differentially heated cavity (a,b,c respectively) and comparisons to the benchmark data in [154] (d and e) (figure produced by the author).

that is, at $x = 0$. The value is here defined as

$$\overline{\mathrm{Nu}} = |\int_M \mathrm{Pr}\,\mathrm{Re}\,u_1\theta - \partial_x\theta|, \tag{4.5.3}$$

where M denotes the hot wall, and it encodes the rate of heat transfer along M (including the total flux, even the part coming from advection). We also record the maximum and minimum velocities and temperatures attained on the symmetry lines $x = 0.5$ and $y = 0.5$. The computed values are collected in Table 4.4, where we also include reference values from the literature (see also [18, 108]).

	Ra	$\overline{\mathrm{Nu}}$	$\max(\lvert\hat{u}_{1,h}\rvert)$	$\max(\lvert\hat{u}_{2,h}\rvert)$	x_∞	y_∞
Computed	10^3	1.105	0.133	0.137	0.177	0.815
Reference value	10^3	1.117	0.136	0.138	0.178	0.813
Computed	10^4	2.002	0.188	0.239	0.117	0.820
Reference value	10^4	2.054	0.192	0.234	0.119	0.823
Computed	10^5	4.430	0.161	0.258	0.068	0.851
Reference value	10^5	4.337	0.153	0.261	0.066	0.855

Table 4.4: Average Nusselt number (4.5.3) and maximum velocities on the midplanes attained at $(0.5, y_\infty)$ and $(x_\infty, 0.5)$, computed for different values of the Rayleigh number and compared with respect to reference values from [66] (table produced by the author).

4.6 Examples using phase-change models

4.6.1 Simulating the melting of N-octadecane

We now consider the melting of a solid phase change material (N-octadecane) contained in a square box and subject to heating on the left side of the domain. The problem set up is taken from [64], where the boundary conditions are as above (no-slip velocities on the whole boundary, the temperature has zero flux on the top and bottom walls, and high and low temperatures are maintained on the left and right walls, respectively). The low temperature imposed on the right wall $\theta_C = -0.01$ is lower than the phase change temperature $\theta = 0$, in order to allow the phase change to occur. We here employ a structured mesh of 100000 elements.

We first consider an enthalpy-viscosity model. The model parameters are as follows Ra = 3.27E5, Pr= 56.2, Re= 1, $\kappa = 1$, Ste= 0.045, high temperature on the left wall $\theta_H = 1$, and size of the mushy region $\delta\theta = 0.1$. We then use (4.2.11) with $s_l = 0$, $s_s = [\mathrm{Ste}]^{-1}$, and we also employ the regularised viscosity specification (4.2.13) with the constants $\mu_l = 1$, $\mu_s = 10^8$, $M_\mu = 50$, and $\theta_f = 0$. These values give a constitutive relationship for the phase change that can also be recovered from (4.2.9) using the phase change function ϕ together with the viscosity $\mu(\phi) = \mu_s + (\mu_l - s_s)\phi$. In that case, the jump is sufficiently large so that the mushy region acts as a solid near the melting point. Notice that in [64] the convection of enthalpy (the last term in the LHS of (4.2.3)) is neglected. This term is zero almost everywhere, except within the mushy region. So if this region is relatively large, then the term will have an important effect in the generated flow patterns.

Secondly, we use an enthalpy-porosity model having a very similar jump behaviour. We focus on the regularised permeability field (4.2.10) and setting the parameter values to $\eta_s = 10^8$, $M_\eta = 50$, $\theta_f = 0$. Again, this can be regularised alternatively using the phase change function ϕ and putting $\eta(\phi) = \eta_s(1 - \phi)$ (see also [134]).

In Figure 4.2 we present snapshots of the numerical solutions obtained using an enthalpy-viscosity model (and setting constant drag), and an enthalpy-porosity model (using constant viscosity). After an initial stage where the dynamics of the system are dictated predominantly by heat conduction, the convective effects in the mixture start to dominate and the layer that determines the phase change

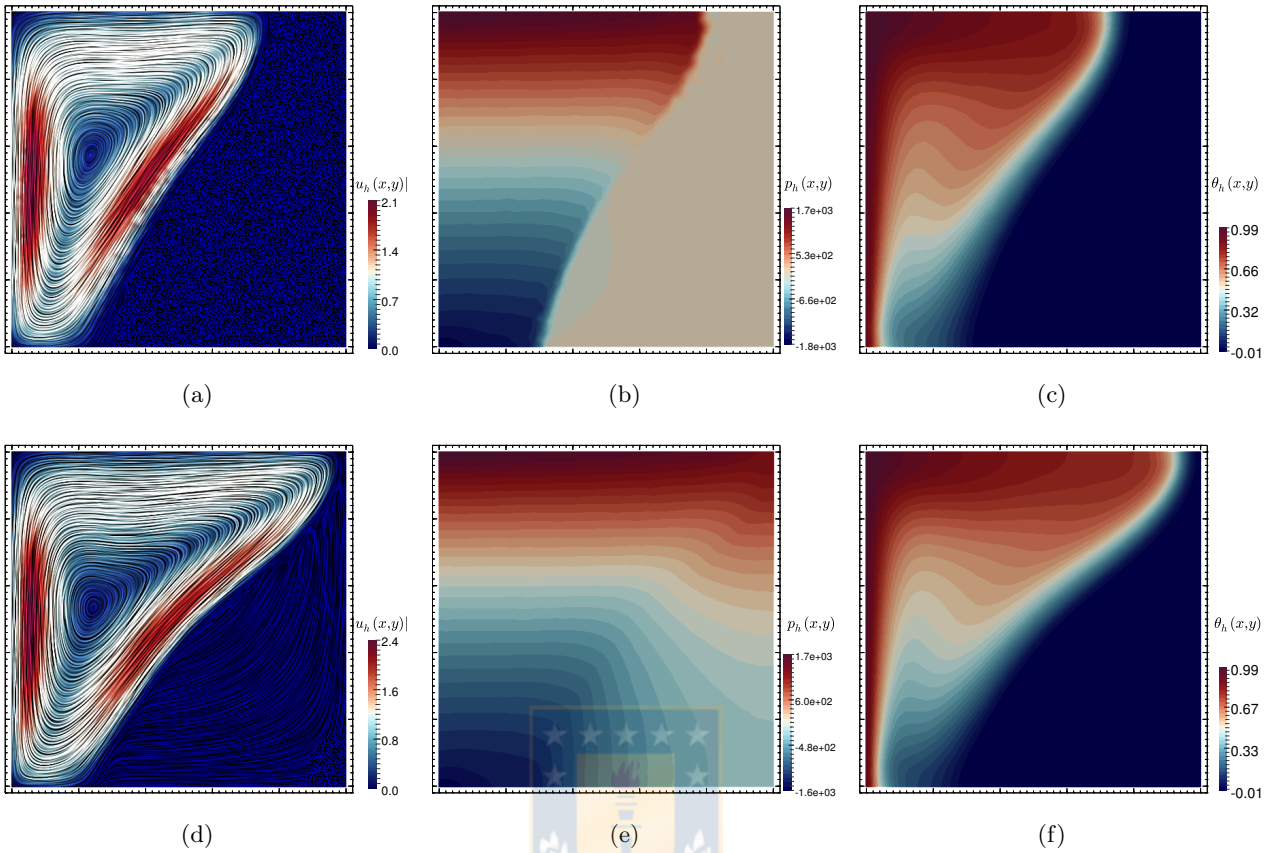
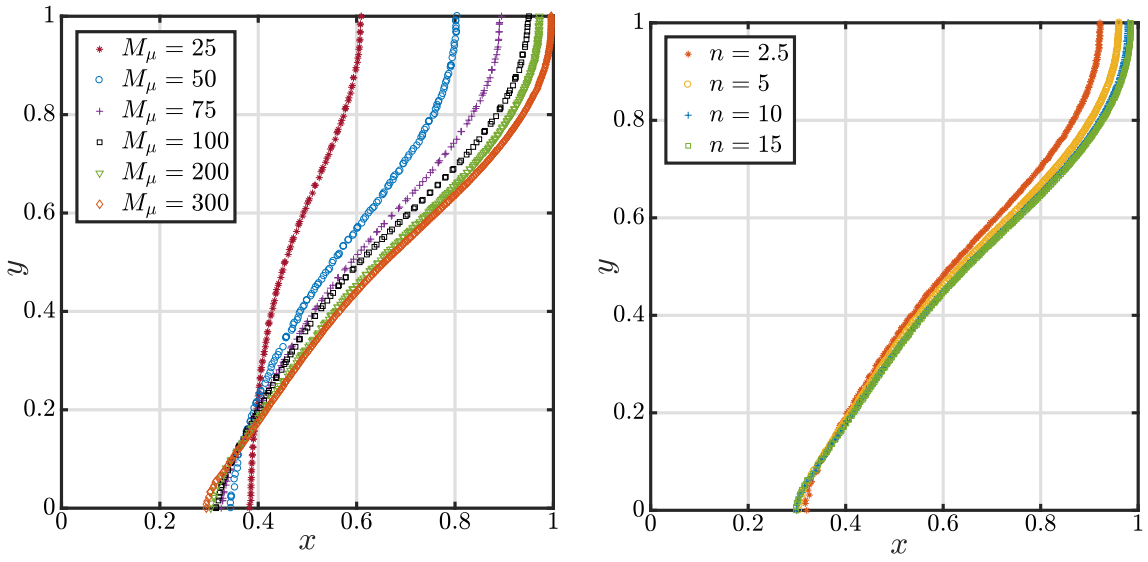


Figure 4.2: Example 3. Comparison between the melting of N-octadecane using enthalpy-viscosity (a,b,c) and enthalpy-porosity models (d,e,f) (figure produced by the author).

moves on to the right on the top of the cavity. We can observe that even if the jump definitions are qualitatively similar, substantial differences appear in terms of the pressure profiles in the solid region. The velocity patterns are also different (the porosity-based model permits recirculation even in the solid), but this can be straightforwardly remediated by taking a larger value for η_s. Secondly, we also observe that the the phase change boundary is more advanced using the enthalpy-porosity model, and this is more pronounced at the top of the container. In the next set of tests we will address these differences in further detail.

4.6.2 Changing the size of the mushy region and the jump nonlinearity

In view to investigate the differences between porosity and viscosity based models, we recall that enthalpy-viscosity models can be very sensitive to dynamic viscosity effects, thermal conductivity variations, and the size of the mushy region (see [21]). On the other hand, the enthalpy-porosity model for the melting of gallium, and studied in [134] is sufficiently accurate for a specific mushy region size. We then proceed to vary the size of the mushy region in an adequate manner to conciliate enthalpy-viscosity and enthalpy-porosity models. It is of note, we do not vary the enthalpy from that of the last experiment. This is because the enthalpy-porosity formulation allows the regularisation of the jump



(a) (b)

Figure 4.3: Example 4: Contour plots of the phase change depending on the size of the mushy region on enthalpy-viscosity models (a), or on the nonlinearity of the regularisation in enthalpy-porosity models (b) (figure produced by the author).

for the drag force $\eta(\theta)$ independent to the regularisation of the jump function for enthalpy, i.e. by changing the value for n. So for consistency we vary M_μ without varying $s(\theta)$.

We start with different values for M_μ in the enthalpy-viscosity model, and we observe that with large mushy regions one can mimic the results obtained using enthalpy-porosity models. However, smaller mushy regions in enthalpy-viscosity models do not necessarily imply a higher irregularity in the model coefficients, as their counterpart in enthalpy-porosity models need to impose a larger jump in order to prevent unwanted flow in the solid region. The results of mushy region variations in enthalpy-viscosity models are collected in Figure 4.3(a).

A second investigation is done by modifying the degree of nonlinearity in the Brinkman term. Setting the parameters $\xi = 10^{7.4}$, $m = 10^{-2.6}$ in the specification of the enthalpy-porosity model (4.2.9) imply that the permeability η is regularised over the desired mushy region, with a constant larger than 10^8 reducing the flow in the solid region. We recall that ξ is related to the material morphology, and in the context of the study of melting N-octadecane, one could choose a more appropriate value. Also, the phase change function ϕ could be regularised around some artificial melting temperature with a different choice of mushy region $\delta\theta$. A closer inspection of the model coefficients reveals that the present form is equivalent to (4.2.10) with a mushy region of size $M_\eta = 150$ (which is a large value, especially considering that enthalpy-porosity models require less regularisation than enthalpy-viscosity models).

Making the link with variable viscosity we recall that in the model by Brinkman one has (4.2.12). We use the value $n = 2.5$, and choose $m = 10^{-8}$ to produce a jump of order 10^8. Plotting then η it is clear that the phase change function ϕ requires a different parametrisation in order to get the jump

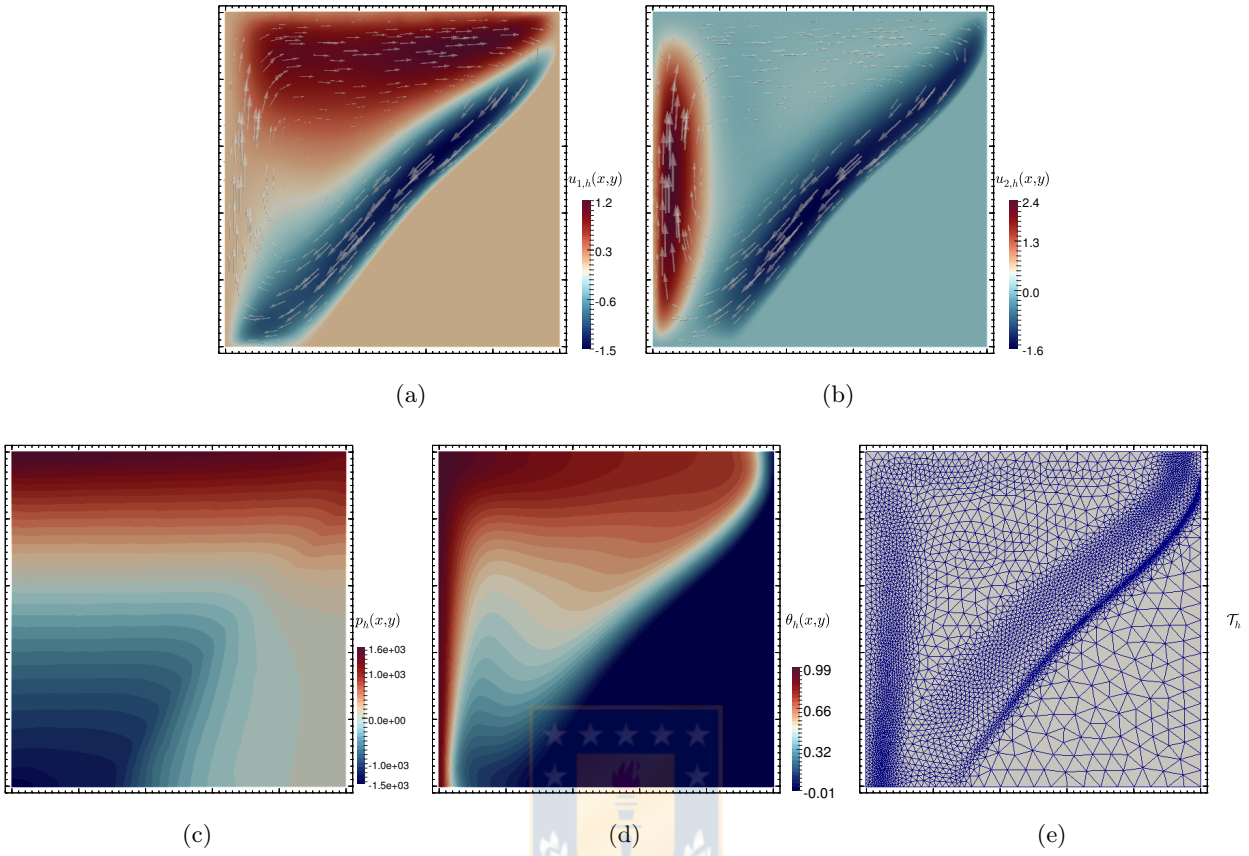


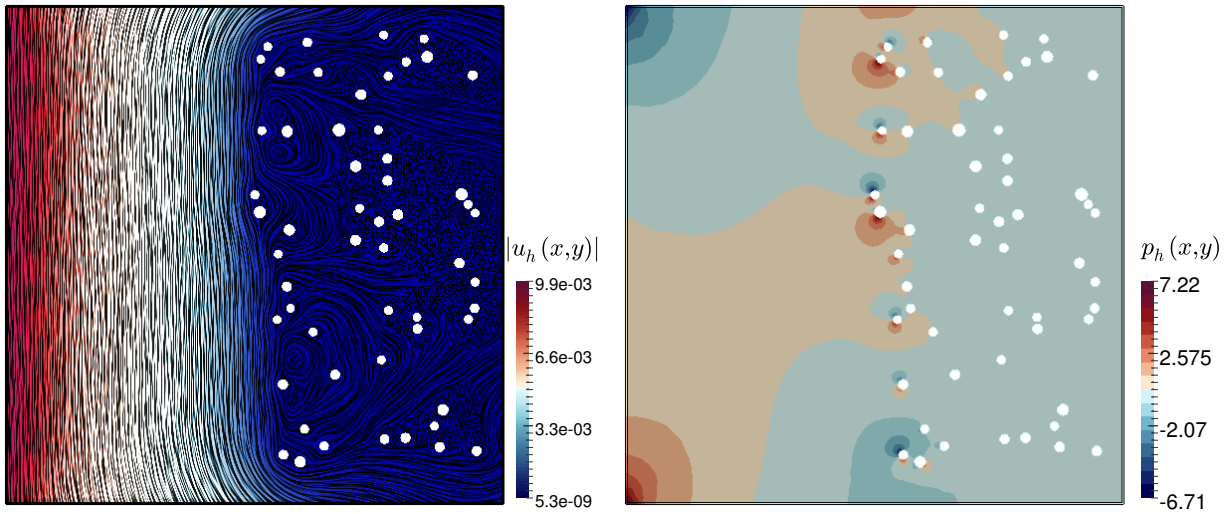
Figure 4.4: Example 4. Velocity components, pressure, temperature, and adapted mesh at $t = 160$, using an enthalpy-porosity model with $n = 5$ (figure produced by the author).

over the required mushy region. We then choose $\phi = \frac{1}{2}[\tanh(50(x - 0.074)) + 1]$, and stress that this model is based on the theory of suspended particles and how the packing density affects viscosity.

Again, using (4.2.12) and varying n, one can mimic the effects of (4.2.9). Comparisons of the melting fronts produced with these two last family of models are displayed in Figure 4.3(b), whereas a sample of the numerical solutions choosing $n = 5$ (and having the following regularisation of the phase change function $\phi = \frac{1}{2}[\tanh(50(\theta - 0.0366)) + 1]$), and a locally refined mesh are portrayed in Figure 4.4.

4.6.3 Flow patterns in a local element

We next turn to the simulation of buoyancy-driven flow within a mesoscopic subdomain (or local element) lying on the boundary layer between liquid and solid. In line with the observations made above, the solid phase can be regarded as a porous medium filled with solid particles. It is expected that these obstacles will generate a drag as the one encoded in the macroscopic form (4.2.9). In a local element of diameter 0.01, we create a so-called blockage arrangement of randomly distributed particles of varying sizes with mean $d_m = 0.0002$ and a 1% variance. The boundary conditions for the flow are different than in the macroscopic case. We impose a slip velocity $\boldsymbol{u} = (0, 0.01)^T$ on the left wall, no-slip velocities on the particles, and Neumann conditions on the remaining walls. We observe a drag force

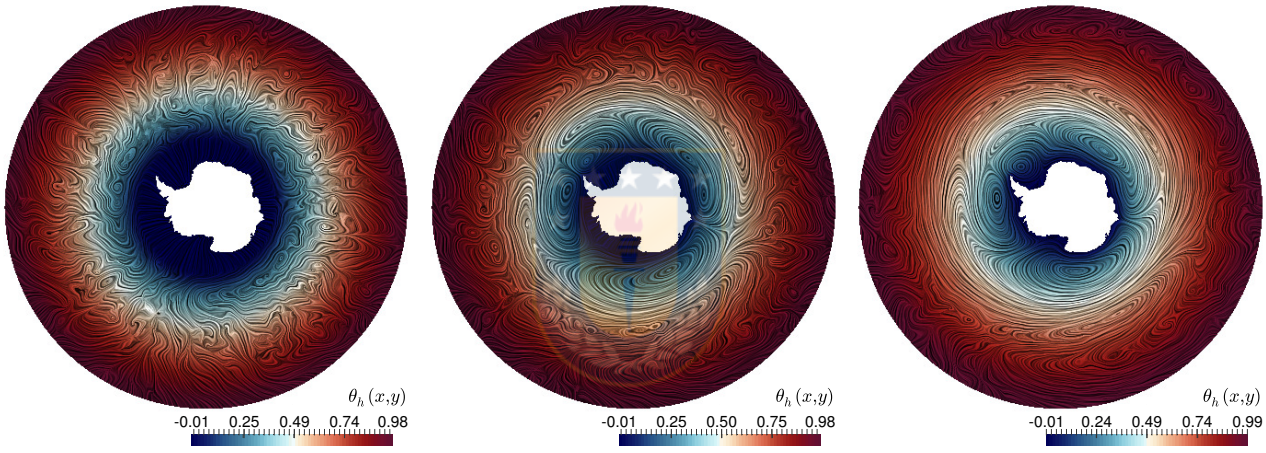


(a) (b)

Figure 4.5: Example 5. Velocity and pressure profiles on a local element (figure produced by the author).

exerted on the fluid by the array of particles. The velocity is reduced on a very small interval, but flow is allowed through the permeable field. These observations suggest why a viscosity-based formulation would require a much smaller mushy region in order to produce the same melting front as in porosity-based models, and also why slow flow is allowed in the solid regime in the porosity formulation. For instance, for the form (4.2.12) one should consider nano particles that melt at the melting temperature of the material, and have the same properties as the liquid, therefore their only effect on the fluid would be an increased viscosity. This behaviour is achieved by simply by not being fluid and by colliding with each other. This phenomenon amounts to impose a moving permeable field, eventually leading to the formation of a larger mushy region. In summary, in enthalpy-porosity models the particles are fixed and the drag force hinders the flow, whereas for enthalpy-viscosity models, the holes are allowed to move and the amount of fluid present is reduced. An example is given in Figure 4.5, and a somewhat similar study can be found in [100].

Finally, with the purpose of extend the applicability of the methods proposed here, in the study of large scale models including the thermal evolution of magma/ocean interfaces [145], or ice-shelf melting [153], we propose a preliminary test given in Figure 4.6 where we have implemented an enthalpy-viscosity model to simulate the ice-shelf melting around Antarctica, using an unstructured mesh consisting of 20500 elements and the transient version of the mixed-primal finite element scheme described in Section 4.4.2. For a relatively large mushy region we can observe a stabilisation of the recirculation patterns and a concentration of the temperature gradients towards the phase change layer.



(a) (b) (c)

Figure 4.6: Temperature profiles at three different times of the phase change dynamics using (4.2.10) (figure produced by the author).

CHAPTER 5

New mixed finite element methods for natural convection with phase-change in porous media

5.1 Introduction

We are interested in the mathematical and numerical investigation of phase change models for natural convection in porous media. Natural convection is a largely studied phenomenon due to its presence in different applications: melting and solidification processes [64, 120, 150, 155], design of latent heat based energy storage devices [71], ocean and atmosphere dynamics [69, 99], crystalization in magma chambers [37, 145], etc. Differently from other works where the phase change is incorporated into the Boussinesq approximation by means of enthalpy-porosity methods [134] or enthalpy-viscosity models [64], in this chapter the problem is modeled either as a viscous Newtonian fluid where the change of phase is encoded in the viscosity itself, or using a Brinkman-Boussinesq approximation where the solidification process influences the drag directly.

A variety of numerical methods dealing with phase change Boussinesq models have been proposed in recent years, including bioconvective flows [44, 101], porosity-based models [134, 155], and viscosity based-models [64, 158, 165]. Mathematical analysis of other related models as natural convection [115, 164], and Boussinesq-type, such as time-dependent problems under different contexts [4, 8, 67], primal and mixed formulations [123, 57, 61, 76], with viscosity of the fluid depending on the temperature [10, 9], and exactly divergence free [123] are available in the literature. However, up to our knowledge, a rigorous mixed analysis for phase change models for natural convection is something that has not had great attention until now. Therefore, in the present chapter, we focus on the mathematical and numerical analysis of that problem, which has been proposed in [158, Section 4.2], and where the authors studied a fully-primal formulation for a non-stationary phase-change model. Here, and similarly to [61], we propose mixed-primal and fully-mixed approaches.

The rest of this chapter is organised as follows. In the remainder of this section, we recall some preliminary notations. The nonlinear model of interest, and the definitive unknowns to be considered in the variational formulation are presented in Section 5.2. For the Navier-Stokes-Brinkman equations, the main unknowns are the velocity, a pseudostress tensor relating the strain tensor with the convective term and the strain rate tensor. The pressure is eliminated using the fluid incompressibility and can be recovered as a post-process of the pseudostress. Moreover, because of the convective term,

the velocity is sought in $\mathbf{H}^1(\Omega)$, which requires augmentation via Galerkin terms arising from the constitutive and equilibrium equations, and therefore, imposing in an ultra-weak sense the symmetry of the pseudostress, we do not need to introduce the vorticity as unknown in our variational formulation. In turn, for the energy equation, and in addition to the temperature, we introduce the normal heat flux through the boundary as a Lagrange multiplier for the primal formulation and a further unknown for the mixed approach. We remark that including these Galerkin terms allows us to circumvent the necessity of proving inf-sup conditions for both problems, and as a result, to relax the hypotheses on the corresponding discrete spaces. In this way, the classical Banach fixed-point theorem, the Lax-Milgram lemma, the Babuška-Brezzi theory, suitable regularity and smallness-of-data assumptions, can be applied to prove well-posedness of the continuous problem. In Section 5.3, we also define the Galerkin scheme considering arbitrary finite dimensional subspaces and provide its unique solvability (this time, by means of Brouwer fixed-point theorem), together with the corresponding Céa estimate. Then, we make precise the definition of the involved discrete spaces. In Section 5.4 we establish the corresponding fully-mixed variational formulation and its associated Galerkin scheme, and show that both systems are well-posed. Then, considering specific finite element spaces for the unknowns together with its approximation properties, we deduce the corresponding rates of convergence. We close in Section 5.5 with several numerical examples illustrating the performance of the augmented mixed-primal and fully-mixed finite element methods, as well as confirming the theoretical rates of convergence.

5.2 The model problem

Let us consider the following PDE system, describing phase change mechanisms involving viscous fluids within porous media. The governing equations in this case correspond to the Navier-Stokes-Brinkman equations coupled with a generalised energy equation (related to the well-known Stefan problem)

$$
\begin{aligned}
(\nabla \boldsymbol{u})\boldsymbol{u} - \alpha\, \mathbf{div}\big[\mu(\theta)\mathbf{e}(\boldsymbol{u})\big] + \nabla p + \eta(\theta)\boldsymbol{u} &= f(\theta)\boldsymbol{k} && \text{in} \quad \Omega\,, && \text{(5.2.1a)}\\
\operatorname{div} \boldsymbol{u} &= 0 && \text{in} \quad \Omega\,, && \text{(5.2.1b)}\\
-\rho \operatorname{div}(\kappa \nabla\theta) + \boldsymbol{u}\cdot\nabla\theta + \boldsymbol{u}\cdot\nabla s(\theta) &= 0 && \text{in} \quad \Omega\,, && \text{(5.2.1c)}\\
\boldsymbol{u} \;=\; \boldsymbol{u}_{\mathrm{D}} \quad \text{and} \quad \theta &= \theta_{\mathrm{D}} && \text{on} \quad \Gamma\,, && \text{(5.2.1d)}
\end{aligned}
$$

with $\alpha := \frac{1}{\mathrm{Re}}$, $\rho := \frac{1}{C\mathrm{Pr}}$, where Re and Pr are the Reynolds and Prandtl numbers, respectively; κ and C are the non-dimensional heat conductivity tensor (here assumed isotropic) and specific heat, respectively; $\boldsymbol{k}$ stands for the unit vector pointing oppositely to gravity; $\mathbf{e}(\boldsymbol{u}) = \dfrac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^{\mathtt{t}})$ is the strain rate tensor, and $\boldsymbol{u} : \Omega \to \mathbb{R}^n$, $p : \Omega \to \mathbb{R}$ and $\theta : \Omega \to \mathbb{R}$, correspond to the velocity, pressure, and the temperature of the fluid flow, respectively. Finally, μ, η, s and f are the nonlinear viscosity, porosity, enthalpy and buoyancy terms, respectively, which depend on the temperature. Notice that here $s(\theta)$ denotes the regularised enthalpy function and it accounts for the latent heat of fusion, i.e. the energy needed to change the phase of a material [158, 151, 152].

For the subsequent analysis we assume that the functions μ, η, s are uniformly bounded and Lipschitz

continuous: there exist positive constants $\mu_1, \mu_2, \eta_1, \eta_2, s_1, s_2, L_\mu, L_\eta, L_s$ such that

$$\begin{array}{lll} \mu_1 \leq \mu(\psi) \leq \mu_2, & |\mu(\psi) - \mu(\phi)| \leq L_\mu |\psi - \phi| & \forall\, \psi, \phi \in \mathrm{R}, \\ \eta_1 \leq \eta(\psi) \leq \eta_2, & |\eta(\psi) - \eta(\phi)| \leq L_\eta |\psi - \phi| & \forall\, \psi, \phi \in \mathrm{R}, \\ s_1 \leq s(\psi) \leq s_2, & |s(\psi) - s(\phi)| \leq L_s |\psi - \phi| & \forall\, \psi, \phi \in \mathrm{R}. \end{array} \tag{5.2.2}$$

Similar assumptions are placed on the buoyancy function f: there exist positive constants C_f and L_f such that

$$|f(\psi)| \leq C_f |\psi|, \quad |f(\psi) - f(\phi)| \leq L_f |\psi - \phi| \quad \forall\, \psi, \phi \in \mathrm{R}. \tag{5.2.3}$$

On the other hand, we will suppose that for every $\psi \in \mathrm{H}^1(\Omega)$, we have $s(\psi) \in \mathrm{H}^1(\Omega)$, and that there exist positive constants s_3 and $L_{\tilde{s}}$ such that

$$|\nabla s(\psi)| \leq s_3 |\nabla \psi|, \quad |\nabla s(\psi) - \nabla s(\phi)| \leq L_{\tilde{s}} |\psi - \phi|, \quad \forall\, \psi, \phi \in \mathrm{H}^1(\Omega)\,. \tag{5.2.4}$$

Finally, we suppose that κ and κ^{-1} are uniform bounded and uniformly positive definite tensors, meaning that there exist positive constants $K_0, K_1, \widetilde{K}_0$ and $\widetilde{K}_1$ such that

$$|\kappa| \leq K_1, \quad \kappa \boldsymbol{v} \cdot \boldsymbol{v} \geq K_0 |\boldsymbol{v}|^2, \quad |\kappa^{-1}| \leq \widetilde{K}_1, \quad \kappa^{-1} \boldsymbol{v} \cdot \boldsymbol{v} \geq \widetilde{K}_0 |\boldsymbol{v}|^2 \quad \forall\, \boldsymbol{v} \in \mathrm{R}^n. \tag{5.2.5}$$

With respect to the boundary conditions in (5.2.1d), we assume that $\boldsymbol{u}_{\mathrm{D}} \in \mathbf{H}^{1/2}(\Gamma)$, $\theta_{\mathrm{D}} \in \mathrm{H}^{1/2}(\Gamma)$, and that $\boldsymbol{u}_{\mathrm{D}}$ verifies the compatibility condition

$$\int_\Gamma \boldsymbol{u}_{\mathrm{D}} \cdot \boldsymbol{\nu} \,=\, 0. \tag{5.2.6}$$

In addition, it is well-known (see, e.g [122]) that uniqueness of pressure is ensured in the space

$$\mathrm{L}^2_0(\Omega) \,:=\, \Big\{ q \in \mathrm{L}^2(\Omega) : \, \int_\Omega q \,=\, 0 \Big\}\,.$$

We end this section by remarking that, due to the laminar regime of the fluid in each one of the numerical tests reported in Section 5.5, the module of the velocity field is small, and hence it might not be necessary to compute the Reynolds number, besides the fact that there seems to be no formula available in the literature for the case (as the present) of a non-constant viscosity. Nevertheless, if an estimation of this number is in fact needed, we would suggest to take a characteristic viscosity μ_c defined as the mean value of it, that is $\mu_c := \dfrac{1}{|\Omega|} \displaystyle\int_\Omega \mu(\theta)$, which can be controlled by μ_1 and μ_2 (cf. (5.2.2)), and then compute the model parameters (Reynolds and Prandtl numbers) and rewritte the coupled-system based on this choice.

5.3 The mixed-primal approach

In this section we proceed similarly as in [10, 42, 57] to propose a mixed-primal approach for (5.2.1). Then, we establish the corresponding continuous and discrete formulations, analyse their solvability by using a fixed-point approach, and derive the corresponding a priori error estimates.

5.3.1 The continuous formulation

We first proceed as in [10] and set the strain rate tensor as an auxiliary unknown:

$$\mathbf{t} := \mathbf{e}(\boldsymbol{u}) = \nabla \boldsymbol{u} - \boldsymbol{\gamma}(\boldsymbol{u}) \in \mathbb{L}^2_{\mathtt{tr}}(\Omega),$$

where, for each $\boldsymbol{v} \in \mathbf{H}^1(\Omega)$, $\boldsymbol{\gamma}(\boldsymbol{v}) = \frac{1}{2}(\nabla \boldsymbol{v} - (\nabla \boldsymbol{v})^{\mathrm{t}})$ is the skew-symmetric part of the velocity gradient tensor $\nabla \boldsymbol{v}$, and

$$\mathbb{L}^2_{\mathtt{tr}}(\Omega) := \Big\{ \mathbf{s} \in \mathbb{L}^2(\Omega) : \quad \mathbf{s} = \mathbf{s}^{\mathtt{t}} \quad \text{and} \quad \mathrm{tr}(\mathbf{s}) = 0 \Big\}.$$

Then, introducing also the pseudostress tensor as a new unknown:

$$\boldsymbol{\sigma} := \alpha\, \mu(\theta) \mathbf{t} - (\boldsymbol{u} \otimes \boldsymbol{u}) - p\, \mathbb{I}, \tag{5.3.1}$$

we deduce that (5.2.1b) together with (5.3.1) are equivalent to the pair of equations

$$\alpha\, \mu(\theta)\, \mathbf{t} - (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathtt{d}} = \boldsymbol{\sigma}^{\mathtt{d}} \quad \text{in} \quad \Omega,$$
$$p = -\frac{1}{n} \mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{u} \otimes \boldsymbol{u}) \quad \text{in} \quad \Omega.$$

Consequently, we arrive at the following coupled system without pressure:

$$\mathbf{t} + \boldsymbol{\gamma}(\boldsymbol{u}) = \nabla \boldsymbol{u} \quad \text{in} \quad \Omega, \tag{5.3.2a}$$
$$\alpha\, \mu(\theta)\, \mathbf{t} - (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathtt{d}} = \boldsymbol{\sigma}^{\mathtt{d}} \quad \text{in} \quad \Omega, \tag{5.3.2b}$$
$$\eta(\theta)\, \boldsymbol{u} - \mathbf{div}\, \boldsymbol{\sigma} = f(\theta) \boldsymbol{k} \quad \text{in} \quad \Omega, \tag{5.3.2c}$$
$$-\rho\, \mathrm{div}(\kappa \nabla \theta) + \boldsymbol{u} \cdot \nabla \theta + \boldsymbol{u} \cdot \nabla s(\theta) = 0 \quad \text{in} \quad \Omega, \tag{5.3.2d}$$
$$\boldsymbol{u} = \boldsymbol{u}_{\mathrm{D}} \quad \text{on} \quad \Gamma, \tag{5.3.2e}$$
$$\theta = \theta_{\mathrm{D}} \quad \text{on} \quad \Gamma, \tag{5.3.2f}$$
$$\int_\Omega \mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{u} \otimes \boldsymbol{u}) = 0. \tag{5.3.2g}$$

Note that the incompressibility constraint is implicitly present in (5.3.2b), relating $\boldsymbol{\sigma}$ and $\boldsymbol{u}$. In turn, the fact that the pressure p must belong to $\mathrm{L}^2_0(\Omega)$ (for uniqueness reasons) is guaranteed by the equivalent statement given by (5.3.2g).

Thus, in order to derive a primal formulation for the energy equation, we proceed to multiply (5.3.2d) by $\psi \in \mathrm{H}^1(\Omega)$, integrate by parts, and introduce, as a new unknown, the normal heat flux on Γ, $\lambda := -\rho\, \kappa \nabla \theta \cdot \boldsymbol{\nu} \in \mathrm{H}^{-1/2}(\Gamma)$, so that we arrive at

$$\rho \int \kappa \nabla \theta \cdot \nabla \psi + \langle \lambda, \psi \rangle_\Gamma + \int_\Omega \psi \boldsymbol{u} \cdot \nabla(\theta + s(\theta)) = 0 \quad \forall\, \psi \in \mathrm{H}^1(\Omega)\,, \tag{5.3.3}$$

where $\langle \cdot, \cdot \rangle_\Gamma$ denotes from now on the duality pairing between $\mathrm{H}^{-1/2}(\Gamma)$ and $\mathrm{H}^{1/2}(\Gamma)$. In turn, the Dirichlet condition $\theta = \theta_{\mathrm{D}}$ on Γ is imposed weakly as

$$\langle \xi, \theta \rangle_\Gamma = \langle \xi, \theta_{\mathrm{D}} \rangle_\Gamma \quad \forall\, \xi \in \mathrm{H}^{-1/2}(\Gamma).$$

On the other hand, multiplying (5.3.2b) by a suitable test function, we obtain

$$\alpha \int_\Omega \mu(\theta) \mathbf{t} : \mathbf{s} - \int_\Omega \boldsymbol{\sigma}^{\mathrm{d}} : \mathbf{s} - \int_\Omega (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathtt{d}} : \mathbf{s} = 0 \quad \forall \mathbf{s} \in \mathbb{L}^2_{\mathtt{tr}}(\Omega). \tag{5.3.4}$$

Here we readily note that in order to bound the third terms on the LHS of (5.3.3) and (5.3.4), and thanks to the continuous injection of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^4(\Omega)$, we require the unknown $\boldsymbol{u}$ to live in $\mathbf{H}^1(\Omega)$ (see e.g. [10, 9, 14]). Such regularity can be exploited to cast the Navier-Stokes-Brinkman equations uniquely in terms of the pseudostress and the velocity. Indeed, testing (5.3.2a) against $\boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega)$ and employing (5.3.2e), we readily obtain

$$\int_\Omega \mathbf{t} : \boldsymbol{\tau}^{\mathrm{d}} + \int_\Omega \boldsymbol{\gamma}(\boldsymbol{u}) : \boldsymbol{\tau} + \int_\Omega \boldsymbol{u} \cdot \mathbf{div}\, \boldsymbol{\tau} = \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma \qquad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega)\,.$$

Afterwards, testing (5.3.2c) against $\boldsymbol{v} \in \mathbf{H}^1(\Omega)$, we deduce that

$$-\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\sigma} + \int_\Omega \eta(\theta)\boldsymbol{u} \cdot \boldsymbol{v} = \int_\Omega f(\theta)\boldsymbol{k} \cdot \boldsymbol{v} \qquad \forall\, \boldsymbol{v} \in \mathbf{H}^1(\Omega),$$

and finally, defining $\mathcal{A} := \Big\{ \boldsymbol{\gamma}(\boldsymbol{v}) : \ \boldsymbol{v} \in \mathbf{H}^1(\Omega) \Big\}$, we impose the symmetry of $\boldsymbol{\sigma}$ in an ultra-weak sense, as follows:

$$\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\gamma}(\boldsymbol{v}) = 0 \qquad \forall\, \boldsymbol{v} \in \mathbf{H}^1(\Omega). \tag{5.3.5}$$

We stress here that the usual way of imposing this property of $\boldsymbol{\sigma}$ is in the form: $\displaystyle\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\eta} = 0 \quad \forall\, \boldsymbol{\eta} \in \mathbb{L}^2_{\mathrm{skew}}(\Omega) := \Big\{ \boldsymbol{\omega} \in \mathbb{L}^2(\Omega) : \quad \boldsymbol{\omega} + \boldsymbol{\omega}^{\mathrm{t}} = \mathbf{0} \Big\}$, which is known as the weak sense. However, in the present approach we propose to take advantage of the further regularity of $\boldsymbol{u}$ and its corresponding test functions, which are all now in $\mathbf{H}^1(\Omega)$, and simply test $\boldsymbol{\sigma}$ against tensors in $\mathcal{A}$. In this way, the fact that $\mathcal{A}$ is a proper subspace of $\mathbb{L}^2_{\mathrm{skew}}(\Omega)$ constitutes the reason why this alternative imposition of the symmetry of the pseudostress is called *ultra-weak*.

In this way, a preliminary weak formulation for the coupled problem (5.2.1) reads: Find $(\mathbf{t}, \boldsymbol{\sigma}, \boldsymbol{u}, \theta, \lambda) \in \mathbb{L}^2_{\mathbf{tr}}(\Omega) \times \mathbb{H}(\mathbf{div};\Omega) \times \mathbf{H}^1(\Omega) \times \mathrm{H}^1(\Omega) \times \mathrm{H}^{-1/2}(\Gamma)$ such that

$$\begin{aligned}
\alpha \int_\Omega \mu(\theta)\mathbf{t} : \mathbf{s} - \int_\Omega (\boldsymbol{u} \otimes \boldsymbol{u})^{\mathtt{d}} : \mathbf{s} - \int_\Omega \boldsymbol{\sigma}^{\mathrm{d}} : \mathbf{s} &= 0 \quad \forall\, \mathbf{s} \in \mathbb{L}^2_{\mathbf{tr}}(\Omega), \\
\int_\Omega \mathbf{t} : \boldsymbol{\tau}^{\mathrm{d}} + \int_\Omega \boldsymbol{\gamma}(\boldsymbol{u}) : \boldsymbol{\tau} + \int_\Omega \boldsymbol{u} \cdot \mathbf{div}\, \boldsymbol{\tau} &= \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{u}_{\mathrm{D}} \rangle_\Gamma \quad \forall\, \boldsymbol{\tau} \in \mathbb{H}(\mathbf{div};\Omega)\,, \\
-\int_\Omega \boldsymbol{v} \cdot \mathbf{div}\, \boldsymbol{\sigma} - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\gamma}(\boldsymbol{v}) + \int_\Omega \eta(\theta)\boldsymbol{u} \cdot \boldsymbol{v} &= \int_\Omega f(\theta)\boldsymbol{k} \cdot \boldsymbol{v} \quad \forall\, \boldsymbol{v} \in \mathbf{H}^1(\Omega), \\
\rho \int \kappa \nabla\theta \cdot \nabla\psi + \langle \lambda, \psi \rangle_\Gamma &= -\int_\Omega \psi \boldsymbol{u} \cdot \nabla(\theta + s(\theta)) \quad \forall\, \psi \in \mathrm{H}^1(\Omega), \\
\langle \xi, \theta \rangle_\Gamma &= \langle \xi, \theta_{\mathrm{D}} \rangle_\Gamma \quad \forall\, \xi \in \mathrm{H}^{-1/2}(\Gamma).
\end{aligned} \tag{5.3.6}$$

On the other hand, by virtue of the orthogonal decomposition $\mathbb{H}(\mathbf{div};\Omega) = \mathbb{H}_0(\mathbf{div};\Omega) \oplus \mathbb{R}\mathbb{I}$, where

$$\mathbb{H}_0(\mathbf{div};\Omega) := \Big\{ \boldsymbol{\zeta} \in \mathbb{H}(\mathbf{div};\Omega) : \quad \int_\Omega \mathrm{tr}(\boldsymbol{\zeta}) = 0 \Big\},$$

and (5.3.2g), we can write $\boldsymbol{\sigma} = \boldsymbol{\sigma}_0 + c\mathrm{I}$, with $\boldsymbol{\sigma}_0$ in $\mathbb{H}_0(\mathbf{div};\Omega)$, and c given explicitly in terms of $\boldsymbol{u}$ by

$$c = -\frac{1}{n|\Omega|} \int_\Omega \mathrm{tr}(\boldsymbol{u} \otimes \boldsymbol{u}).$$

Then, denoting from now on the unknown $\boldsymbol{\sigma}_0$ simply by $\boldsymbol{\sigma}$, the variational formulation (5.3.6) can be reformulated in terms of the $\mathbb{H}_0(\mathbf{div};\Omega)$-component of the pseudostress (see [9, Lemma 3.1]). Accordingly, in order to analyse (5.3.6) we augment using residual Galerkin-type terms arising from (5.3.2), but all them tested differently from (5.3.6), namely:

$$\begin{aligned}
\kappa_1 \int_\Omega \{\boldsymbol{\sigma}^{\mathtt{d}} + (\boldsymbol{u}\otimes\boldsymbol{u})^{\mathtt{d}} - \alpha\mu(\theta)\mathbf{t}\} : \boldsymbol{\tau}^{\mathrm{d}} &= 0 && \forall\, \boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega)\,,\\
\kappa_2 \int_\Omega \{\mathbf{div}\,\boldsymbol{\sigma} - \eta(\theta)\boldsymbol{u}\}\cdot\mathbf{div}\boldsymbol{\tau} &= -\kappa_2\int_\Omega f(\theta)\boldsymbol{k}\cdot\mathbf{div}\boldsymbol{\tau} && \forall\, \boldsymbol{\tau}\in\mathbb{H}_0(\mathbf{div};\Omega)\,,\\
\kappa_3 \int_\Omega \{\mathbf{e}(\boldsymbol{u}) - \mathbf{t}\}\cdot\mathbf{e}(\boldsymbol{v}) &= 0 && \forall\, \boldsymbol{v}\in\mathbf{H}^1(\Omega),
\end{aligned}$$

where κ_1, κ_2 and κ_3 are positive parameters to be specified later on. In this way, denoting $H := \mathbb{L}^2_{\mathtt{tr}}(\Omega)\times\mathbb{H}_0(\mathbf{div};\Omega)\times\mathbf{H}^1(\Omega)$, $\vec{\mathbf{t}} := (\mathbf{t},\boldsymbol{\sigma},\boldsymbol{u})$, and $\vec{\mathbf{s}} := (\mathbf{s},\boldsymbol{\tau},\boldsymbol{v})$, we arrive at the following augmented mixed-primal formulation for (5.2.1): Find $(\vec{\mathbf{t}},(\theta,\lambda))\in H\times\mathrm{H}^1(\Omega)\times\mathrm{H}^{-1/2}(\Gamma)$ such that

$$\mathbf{A}_\theta(\vec{\mathbf{t}},\vec{\mathbf{s}}) + \mathbf{B}_{\boldsymbol{u}}(\vec{\mathbf{t}},\vec{\mathbf{s}}) = F_\theta(\vec{\mathbf{s}}) + F_{\mathrm{D}}(\vec{\mathbf{s}}) \qquad \forall\, \vec{\mathbf{s}}\in H, \tag{5.3.7a}$$

$$\mathbf{a}(\theta,\psi) + \mathbf{b}(\psi,\lambda) = H_{\boldsymbol{u},\theta}(\psi) \qquad \forall\,\psi\in\mathrm{H}^1(\Omega)\,, \tag{5.3.7b}$$

$$\mathbf{b}(\theta,\xi) = G(\xi) \qquad \forall\,\xi\in\mathrm{H}^{-1/2}(\Gamma), \tag{5.3.7c}$$

where, given an arbitrary $(\boldsymbol{w},\phi)\in\mathbf{H}^1(\Omega)\times\mathrm{H}^1(\Omega)$, the forms $\mathbf{A}_\phi$, $\mathbf{B}_{\boldsymbol{w}}$, $\mathbf{a}$, $\mathbf{b}$, and the functionals F_ϕ, F_{D}, $H_{\boldsymbol{w},\phi}$, and G are defined as

$$\begin{aligned}
\mathbf{A}_\phi(\vec{\mathbf{t}},\vec{\mathbf{s}}) := {}& \alpha\int\mu(\phi)\mathbf{t}:\Big\{\mathbf{s}-\kappa_1\boldsymbol{\tau}^{\mathrm{d}}\Big\} + \int_\Omega \mathbf{t}:\Big\{\boldsymbol{\tau}^{\mathrm{d}}-\kappa_3\,\mathbf{e}(\boldsymbol{v})\Big\} - \int_\Omega \boldsymbol{\sigma}^{\mathtt{d}}:\Big\{\mathbf{s}-\kappa_1\boldsymbol{\tau}^{\mathrm{d}}\Big\}\\
&+ \int_\Omega \boldsymbol{u}\cdot\mathbf{div}\,\boldsymbol{\tau} - \int_\Omega \boldsymbol{v}\cdot\mathbf{div}\,\boldsymbol{\sigma} + \int_\Omega \boldsymbol{\gamma}(\boldsymbol{u}):\boldsymbol{\tau} - \int_\Omega \boldsymbol{\sigma}:\boldsymbol{\gamma}(\boldsymbol{v})\\
&+ \int_\Omega \eta(\phi)\boldsymbol{u}\cdot\Big\{\boldsymbol{v}-\kappa_2\,\mathbf{div}\boldsymbol{\tau}\Big\} + \kappa_2\int_\Omega \mathbf{div}\,\boldsymbol{\sigma}\cdot\mathbf{div}\boldsymbol{\tau} + \kappa_3\int_\Omega \mathbf{e}(\boldsymbol{u}):\mathbf{e}(\boldsymbol{v}),
\end{aligned} \tag{5.3.8a}$$

$$\mathbf{B}_{\boldsymbol{w}}(\vec{\mathbf{t}},\vec{\mathbf{s}}) := \int_\Omega (\boldsymbol{u}\otimes\boldsymbol{w})^{\mathtt{d}} : \Big\{\kappa_1\boldsymbol{\tau}^{\mathrm{d}}-\mathbf{s}\Big\}, \tag{5.3.8b}$$

for all $\vec{\mathbf{t}},\vec{\mathbf{s}}\in H$, and

$$\mathbf{a}(\theta,\psi) := \rho\int_\Omega \kappa\nabla\theta\cdot\nabla\psi \qquad \forall\,\theta,\psi\in\mathrm{H}^1(\Omega), \tag{5.3.9a}$$

$$\mathbf{b}(\psi,\xi) := \langle\xi,\psi\rangle_\Gamma \qquad \forall\,(\psi,\xi)\in\mathrm{H}^1(\Omega)\times\mathrm{H}^{-1/2}(\Gamma), \tag{5.3.9b}$$

$$F_\phi(\vec{\mathbf{s}}) := \int_\Omega f(\phi)\boldsymbol{k}\cdot\Big\{\boldsymbol{v}-\kappa_2\,\mathbf{div}\,\boldsymbol{\tau}\Big\} \qquad \forall\,\vec{\mathbf{s}}\in H, \tag{5.3.9c}$$

$$F_{\mathrm{D}}(\vec{\mathbf{s}}) := \langle\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{u}_{\mathrm{D}}\rangle_\Gamma \qquad \forall\,\vec{\mathbf{s}}\in H, \tag{5.3.9d}$$

$$H_{\boldsymbol{w},\phi}(\psi) := -\int_\Omega \psi\boldsymbol{w}\cdot\nabla(\phi+s(\phi)) \qquad \forall\,\psi\in\mathrm{H}^1(\Omega), \tag{5.3.9e}$$

$$G(\xi) := \langle\xi,\theta_{\mathrm{D}}\rangle_\Gamma \qquad \forall\,\xi\in\mathrm{H}^{-1/2}(\Gamma). \tag{5.3.9f}$$

We notice in advance that the forms $\mathbf{B}_{\boldsymbol{w}}$, $\boldsymbol{a}$ and $\mathbf{b}$ are exactly defined as in [59, Section 3.1] and therefore we omit parts of the proofs whenever necessary. Finally, we remark that, in contrast with other recent strain-based formulations [10, 43, 79, 83], here we do not introduce vorticity as additional unknown. Also, the presence of the drag term in the momentum equation allows us to complete the $\mathbf{H}^1(\Omega)-$norm of the velocity without the need of a fourth residual term in the augmentation procedure.

5.3.2 Solvability analysis

We proceed similarly as in [10, 83] and utilise a fixed-point scheme to prove the well-posedness of the continuous formulation (5.3.7). Let us write $\mathbf{H} := \mathbf{H}^1(\Omega) \times \mathrm{H}^1(\Omega)$ and define $\mathbf{S} : \mathbf{H} \to H$ as

$$\mathbf{S}(\boldsymbol{w}, \phi) = (\mathbf{S}_1(\boldsymbol{w}, \phi), \mathbf{S}_2(\boldsymbol{w}, \phi), \mathbf{S}_3(\boldsymbol{w}, \phi)) := \vec{\mathbf{t}} \qquad \forall\, (\boldsymbol{w}, \phi) \in \mathbf{H}\,, \tag{5.3.10}$$

where $\vec{\mathbf{t}} \in H$ is the unique solution of the problem defined by (5.3.7a) with $(\boldsymbol{w}, \phi)$ instead of $(\boldsymbol{u}, \theta)$, that is

$$\mathbf{A}_\phi(\vec{\mathbf{t}}, \vec{\mathbf{s}}) \,+\, \mathbf{B}_{\boldsymbol{w}}(\vec{\mathbf{t}}, \vec{\mathbf{s}}) = F_\phi(\vec{\mathbf{s}}) \,+\, F_{\mathrm{D}}(\vec{\mathbf{s}}) \quad \forall\, \vec{\mathbf{s}} \in H. \tag{5.3.11}$$

In turn, let $\widetilde{\mathbf{S}} : \mathbf{H} \to \mathrm{H}^1(\Omega)$ be the operator defined by

$$\widetilde{\mathbf{S}}(\boldsymbol{w}, \phi) := \theta \qquad \forall\, (\boldsymbol{w}, \phi) \in \mathbf{H}\,, \tag{5.3.12}$$

where θ is the first component of the unique solution $(\theta, \lambda) \in \mathrm{H}^1(\Omega) \times \mathrm{H}^{-1/2}(\Gamma)$ of the problem defined by (5.3.7b)-(5.3.7c) with $(\boldsymbol{w}, \phi)$ instead of $(\boldsymbol{u}, \theta)$, that is

$$\begin{aligned} \mathbf{a}(\theta, \psi) \,+\, \mathbf{b}(\psi, \lambda) &= H_{\boldsymbol{w},\phi}(\psi) \quad \forall\, \psi \in \mathrm{H}^1(\Omega)\,, \\ \mathbf{b}(\theta, \xi) &= G(\xi) \qquad \forall\, \xi \in \mathrm{H}^{-1/2}(\Gamma). \end{aligned} \tag{5.3.13}$$

Then, we define the operator $\mathbf{T} : \mathbf{H} \to \mathbf{H}$ by

$$\mathbf{T}(\boldsymbol{w}, \phi) = \Big(\mathbf{S}_3(\boldsymbol{w}, \phi), \widetilde{\mathbf{S}}(\mathbf{S}_3(\boldsymbol{w}, \phi), \phi)\Big) \quad \forall\, (\boldsymbol{w}, \phi) \in \mathbf{H}, \tag{5.3.14}$$

and one readily realises that solving (5.3.7) is equivalent to seeking a fixed point of $\mathbf{T}$, that is: Find $(\boldsymbol{u}, \theta) \in \mathbf{H}$ such that

$$\mathbf{T}(\boldsymbol{u}, \theta) = (\boldsymbol{u}, \theta).$$

We now provide sufficient conditions under which the uncoupled problems (5.3.11) and (5.3.13) are indeed uniquely solvable. In what follows, for each $\vec{\mathbf{s}} \in H$, $\|\vec{\mathbf{s}}\|$ denotes the corresponding product norm.

Lemma 5.3.1. *Assume that* $\kappa_1 \in \left(0, \frac{2\mu_1\delta_1}{\mu_2}\right)$, $\kappa_2 \in \left(0, \frac{2\eta_1\delta_3}{\eta_2}\right)$ *and* $\kappa_3 \in \left(0, 2\alpha\delta_2\left(\mu_1 - \frac{\kappa_1\mu_2}{2\delta_1}\right)\right)$ *with* $\delta_1 \in \left(0, \frac{2}{\alpha\mu_2}\right)$, $\delta_2 \in (0, 2)$, $\delta_3 \in \left(0, \frac{2}{\eta_2}\right)$. *Then, there exists* $r_0 > 0$ *such that for each* $r \in (0, r_0)$, *problem* (5.3.11) *has a unique solution* $\mathbf{S}(\boldsymbol{w}, \phi) := \vec{\mathbf{t}} \in H$, *for each* $(\boldsymbol{w}, \phi) \in \mathbf{H}$ *with* $\|\boldsymbol{w}\|_{1,\Omega} \leq r$. *Moreover, there exists* $c_{\mathbf{S}} > 0$, *independent of* $(\boldsymbol{w}, \phi)$, *such that*

$$\|\mathbf{S}(\boldsymbol{w}, \phi)\| = \|\vec{\mathbf{t}}\| \leq c_{\mathbf{S}} \left\{ C_f \, \|\phi\|_{0,\Omega} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \right\} \quad \forall\, (\boldsymbol{w}, \phi) \in \mathbf{H}. \tag{5.3.15}$$

Proof. Let us start the discussion by deriving the continuity of the forms involved. First, employing the assumptions (5.2.2), we deduce that

$$|\mathbf{A}_\phi(\vec{\mathbf{t}}, \vec{\mathbf{s}})| \leq C_{\mathbf{A}} \|\vec{\mathbf{t}}\| \|\vec{\mathbf{s}}\| \quad \forall\, \vec{\mathbf{t}}, \vec{\mathbf{s}} \in H, \tag{5.3.16}$$

where $C_{\mathbf{A}}$ is a constant depending on $\alpha, \kappa_1, \kappa_2, \kappa_3, \mu_2$, and η_2. In turn, by applying the continuous injection $\boldsymbol{i}_c : \mathbf{H}^1(\Omega) \to \mathbf{L}^4(\Omega)$, we obtain that

$$|\mathbf{B}_{\boldsymbol{w}}(\vec{\mathbf{t}}, \vec{\mathbf{s}})| \leq \|\boldsymbol{i}_c\|^2 \, (1 + \kappa_1^2)^{1/2} \|\boldsymbol{u}\| \|\boldsymbol{w}\| \|\vec{\mathbf{s}}\| \quad \forall\, \vec{\mathbf{t}}, \vec{\mathbf{s}} \in H. \tag{5.3.17}$$

Hence, from (5.3.16) and (5.3.17), there exists a positive constant denoted by $\|\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}\|$, such that

$$|(\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}})(\vec{\mathbf{t}}, \vec{\mathbf{s}})| \leq \|\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}\| \|\vec{\mathbf{t}}\| \|\vec{\mathbf{s}}\| \quad \forall\, \vec{\mathbf{t}}, \vec{\mathbf{s}} \in H.$$

On the other hand, in order to show that $(\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}})$ is elliptic, we first prove that $\mathbf{A}_\phi$ satisfies this property. In fact, by using Cauchy-Schwarz and Young inequalities, and the results provided in [39, Prop. 3.1] and [54, Thm. 6.15-1] with constants $c_3(\Omega)$ and $\kappa_0(\Omega)$, respectively, it is possible to find a constant $\widetilde{\alpha}(\Omega) := \min\{\alpha_1, \alpha_3\kappa_0(\Omega), \alpha_4\}$, independent of $(\boldsymbol{w}, \phi)$, such that

$$\mathbf{A}_\phi(\vec{\mathbf{s}}, \vec{\mathbf{s}}) \geq \widetilde{\alpha}(\Omega) \|\vec{\mathbf{s}}\|^2 \quad \forall\, \vec{\mathbf{s}} \in H, \tag{5.3.18}$$

where

$$\alpha_1 := \alpha\mu_1 - \frac{\kappa_1\alpha\mu_2}{2\delta_1} - \frac{\kappa_3}{2\delta_2}, \quad \alpha_2 := \min\left\{\kappa_1\left(1 - \frac{\alpha\mu_2\delta_1}{2}\right), \frac{\kappa_2}{2}\left(1 - \frac{\eta_2\delta_3}{2}\right)\right\},$$
$$\alpha_3 := \min\left\{\kappa_3\left(1 - \frac{\delta_2}{2}\right), \eta_1 - \frac{\kappa_2\eta_2}{2\delta_3}\right\}, \quad \alpha_4 := \min\left\{\alpha_2 c_3(\Omega), \frac{\kappa_2}{2}\left(1 - \frac{\eta_2\delta_3}{2}\right)\right\},$$

where $\kappa_1, \kappa_2, \kappa_3, \delta_1, \delta_2$ and δ_3 are defined as in the statement of the present lemma. Moreover, by combining (5.3.17) and (5.3.18), we obtain that

$$(\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}})(\vec{\mathbf{t}}, \vec{\mathbf{s}}) \geq \frac{\widetilde{\alpha}(\Omega)}{2} \|\vec{\mathbf{s}}\|^2 \quad \forall\, \vec{\mathbf{s}} \in H, \tag{5.3.19}$$

provided $\|\boldsymbol{w}\|_{1,\Omega} \leq r_0$, with

$$r_0 := \frac{\widetilde{\alpha}(\Omega)}{2\,\|\boldsymbol{i}_c\|^2\,(1 + \kappa_1^2)^{1/2}}, \tag{5.3.20}$$

which confirms the ellipticity of the nonlinear operator $\mathbf{A}_\phi + \mathbf{B}_{\boldsymbol{w}}$. On the other hand, by applying Cauchy-Schwarz inequality and the trace theorem in $\mathbb{H}(\mathbf{div}; \Omega)$, we deduce that $F_\phi, F_{\mathrm{D}} \in H'$ with

$$\|F_\phi\| \leq C_f(1 + \kappa_2^2)^{1/2}\,\|\phi\|_{0,\Omega} \quad \text{and} \quad \|F_{\mathrm{D}}\| \leq \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}. \tag{5.3.21}$$

Consequently, a straightforward application of the Lax-Milgram lemma implies that there exists a unique solution $\vec{\mathbf{t}} \in H$ of (5.3.11). Finally, using (5.3.19) and (5.3.21), and performing simple algebraic manipulations, we derive (5.3.15) with $c_{\mathbf{S}} := \frac{2(1+\kappa_2^2)^{1/2}}{\widetilde{\alpha}(\Omega)} > 0$, independent of $(\boldsymbol{w}, \phi)$. □

Lemma 5.3.2. *For each $(\boldsymbol{w}, \phi) \in \mathbf{H}$, problem (5.3.13) has a unique solution $(\theta, \lambda) = \big(\widetilde{\mathbf{S}}(\boldsymbol{w}, \phi), \lambda\big) \in \mathrm{H}^1(\Omega) \times \mathrm{H}^{-1/2}(\Gamma)$. Moreover, there exists a constant $\widetilde{c}_{\mathrm{S}} > 0$ independent of $(\boldsymbol{w}, \phi)$, such that*

$$\|\widetilde{\mathbf{S}}(\boldsymbol{w}, \phi)\| \leq \|(\theta, \lambda)\| \leq \widetilde{c}_{\mathbf{S}}\Big\{\|\boldsymbol{w}\|_{1,\Omega}|\phi|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\Big\}. \tag{5.3.22}$$

Proof. From [57, Lemma 3.4] we know that $\boldsymbol{a}$ and $\mathbf{b}$ are bounded independently of $(\boldsymbol{w}, \phi)$, and that the bilinear form $\mathbf{b}$ satisfies the inf-sup condition. Furthermore, recalling that ϑ (cf. (5.2.5)) is a uniformly positive definite tensor, and using the Friedrichs-Poincaré inequality, we also deduce that $\boldsymbol{a}$ is V-elliptic with constant $\alpha_{\boldsymbol{a}}(\Omega)$, where V is the kernel of the operator induced by $\mathbf{b}$. Now, using (5.3.9e), (5.3.9f) and applying the continuous injection $i_c : \mathrm{H}^1(\Omega) \to \mathrm{L}^4(\Omega)$, we find that

$$\|H_{\boldsymbol{w},\phi}\| \leq \|i_c\|^2\,\big\{1 + s_3\big\}\,\|\boldsymbol{w}\|_{1,\Omega}|\phi|_{1,\Omega} \quad \text{and} \quad \|G\| \leq \|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\,,$$

which implies that $H_{\boldsymbol{w},\phi}$ and G are bounded functionals. Thus, a straightforward application of the Babŭska-Brezzi theory (see, e.g. [81, Thm. 2.3]) proves that for each $(\boldsymbol{w},\phi)\in\mathbf{H}$, problem (5.3.13) has a unique solution $(\theta,\lambda)\in\mathrm{H}^1(\Omega)\times\mathrm{H}^{-1/2}(\Gamma)$. Moreover, there exists a positive constant $\widetilde{c}_{\mathrm{S}}$ depending on ρ, K_1, $\alpha_{\boldsymbol{a}}(\Omega)$, $\|i_c\|$, s_3 and the inf-sup constant of $\mathbf{b}$, such that the estimate (5.3.22) holds. □

At this point, we remark that for computational purposes, the constants α_1, α_2 and α_3 defining $\widetilde{\alpha}(\Omega)$ in Lemma 5.3.1, can be maximised by choosing the parameters δ_1, δ_2, δ_3, κ_1 κ_2, and κ_3 as the middle points of their feasible ranges. Adequate choices for these parameters are then

$$\delta_1=\frac{1}{\alpha\mu_2},\quad \kappa_1=\frac{\mu_1}{\alpha\mu_2^2},\quad \delta_2=1,\quad \kappa_3=\frac{\alpha\mu_1}{2},\quad \delta_3=\frac{1}{\eta_2},\quad \kappa_2=\frac{\eta_1}{\eta_2^2}. \tag{5.3.23}$$

Continuing with the analysis, we assume further regularity on the problem defining $\mathbf{S}$. More precisely, we assume that $\boldsymbol{u}_{\mathrm{D}}\in\mathbf{H}^{1/2+\varepsilon}(\Gamma)$ for some $\varepsilon\in(0,1)$ (when $n=2$) or $\varepsilon\in\left[\frac{1}{2},1\right)$ (when $n=3$), and that for each $(\boldsymbol{w},\phi)\in\mathbf{H}$ with $\|\boldsymbol{w}\|_{1,\Omega}+\|\phi\|_{1,\Omega}\leq r$, $r>0$ given, there holds $(\mathbf{r},\zeta,\boldsymbol{z}):=\mathbf{S}(\boldsymbol{w},\phi)\in\mathbb{L}^2_{\mathtt{tr}}(\Omega)\cap\mathbb{H}^{\varepsilon}(\Omega)\times\mathbb{H}_0(\mathbf{div};\Omega)\cap\mathbb{H}^{\varepsilon}(\Omega)\times\mathbf{H}^{1+\varepsilon}(\Omega)$, with

$$\|\mathbf{r}\|_{\varepsilon,\Omega}+\|\zeta\|_{\varepsilon,\Omega}+\|\boldsymbol{z}\|_{1+\varepsilon,\Omega}\leq\widehat{C}(r)\left\{C_f\,\|\phi\|_{0,\Omega}+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma}\right\}, \tag{5.3.24}$$

where C_f is given by (5.2.3) and $\widehat{C}(r)$ is a positive constant independent of $(\boldsymbol{w},\phi)$ but depending on the upper bound r of its norm. The reason of the stipulated ranges for ε will be clarified in the forthcoming analysis (specifically in the proofs of Lemmas 5.3.4 and 5.3.6, below). Also, we pay attention to the fact that while the estimate (5.3.24) will be employed only to bound $\|\mathbf{r}\|_{\varepsilon,\Omega}$, we have stated it including the terms $\|\zeta\|_{\varepsilon,\Omega}$ and $\|\boldsymbol{z}\|_{1+\varepsilon,\Omega}$ as well, since due to the first and second equations of (5.3.2), the regularities of $\mathbf{r}$, ζ and $\boldsymbol{z}$ will most likely be connected.

On the other hand, we emphasize that the well-posedness of the uncoupled problems (5.3.11) and (5.3.13) ensure that the operators $\mathbf{S}$, $\widetilde{\mathbf{S}}$ and $\mathbf{T}$ are well-defined. Hence, the existence of a unique fixed-point of $\mathbf{T}$ follows after verifying the hypotheses of the Banach fixed-point theorem.

Lemma 5.3.3. *Given $r\in(0,r_0)$, with r_0 given by (5.3.20), we let $W:=\left\{(\boldsymbol{w},\phi)\in\mathbf{H}:\|(\boldsymbol{w},\phi)\|\leq r\right\}$, and assume that*

$$c(r)\left\{C_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}\right\}+\widetilde{c}_{\mathbf{S}}\,\|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\leq r, \tag{5.3.25}$$

where $c(r):=(1+\widetilde{c}_{\mathbf{S}})\,c_{\mathbf{S}}\max\{1,r\}$, and C_f, $c_{\mathbf{S}}$ and $\widetilde{c}_{\mathbf{S}}$ are the constants specified in (5.2.3), and Lemmas 5.3.1 and 5.3.2, respectively. Then $\mathbf{T}(W)\subseteq W$.

Proof. It follows exactly as in [57, Lemma 3.5]. □

Next, the Lipschitz continuity of $\mathbf{T}$ will essentially be a direct consequence of the following two lemmas providing the same property for $\mathbf{S}$ and $\widetilde{\mathbf{S}}$, respectively.

Lemma 5.3.4. *Let $r\in(0,r_0)$ with r_0 given by (5.3.20). Then, there exists a constant $\widetilde{C}_{\mathbf{S}}>0$, independent of r, such that for all $(\boldsymbol{w}_1,\phi_1)$, $(\boldsymbol{w}_2,\phi_2)\in\mathbf{H}$, with $\|\boldsymbol{w}_1\|_{1,\Omega}$, $\|\boldsymbol{w}_2\|_{1,\Omega}\leq r$, there holds*

$$\begin{aligned}\|\mathbf{S}(\boldsymbol{w}_1,\phi_1)-\mathbf{S}(\boldsymbol{w}_2,\phi_2)\|\leq\widetilde{C}_{\mathbf{S}}\Big\{&\|\mathbf{S}_3(\boldsymbol{w}_2,\phi_2)\|_{1,\Omega}\Big(\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{1,\Omega}+\|\phi_1-\phi_2\|_{1,\Omega}\Big)\\&+\|\phi_1-\phi_2\|_{\mathrm{L}^{n/\varepsilon}(\Omega)}\,\|\mathbf{S}_1(\boldsymbol{w}_2,\phi_2)\|_{\varepsilon,\Omega}+L_f\,\|\phi_1-\phi_2\|_{0,\Omega}\Big\}.\end{aligned} \tag{5.3.26}$$

Proof. Given $(\boldsymbol{w}_1,\phi_1)$, $(\boldsymbol{w}_2,\phi_2)$ as stated, we let $\vec{\mathbf{t}}_j := (\mathbf{t}_j,\boldsymbol{\sigma}_j,\boldsymbol{u}_j) = \mathbf{S}(\boldsymbol{w}_j,\phi_j) \in H$, $j\in\{1,2\}$, which, according to (5.3.11), means that for all $\vec{\mathbf{s}}\in H$ there hold:

$$\mathbf{A}_{\phi_1}(\vec{\mathbf{t}}_1,\vec{\mathbf{s}}) + \mathbf{B}_{\boldsymbol{w}_1}(\vec{\mathbf{t}}_1,\vec{\mathbf{s}}) = F_{\phi_1}(\vec{\mathbf{s}}) + F_{\mathrm{D}}(\vec{\mathbf{s}}) \quad \text{and} \quad \mathbf{A}_{\phi_2}(\vec{\mathbf{t}}_2,\vec{\mathbf{s}}) + \mathbf{B}_{\boldsymbol{w}_2}(\vec{\mathbf{t}}_2,\vec{\mathbf{s}}) = F_{\phi_2}(\vec{\mathbf{s}}) + F_{\mathrm{D}}(\vec{\mathbf{s}}).$$

Now, applying the ellipticity of $\mathbf{A}_{\phi_1}+\mathbf{B}_{\boldsymbol{w}_1}$ (cf. (5.3.19)), and then adding and subtracting the equality $\mathbf{A}_{\phi_2}(\vec{\mathbf{t}}_2,\vec{\mathbf{s}}) + \mathbf{B}_{\boldsymbol{w}_2}(\vec{\mathbf{t}}_2,\vec{\mathbf{s}}) = F_{\phi_2}(\vec{\mathbf{s}}) + F_{\mathrm{D}}(\vec{\mathbf{s}})$, we find that

$$\begin{aligned}\frac{\widetilde{\alpha}(\Omega)}{2}\|\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2\|^2 &\leq (\mathbf{A}_{\phi_1}+\mathbf{B}_{\boldsymbol{w}_1})(\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2,\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2)\\ &= (F_{\phi_1}-F_{\phi_2})(\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2) + (\mathbf{A}_{\phi_2}-\mathbf{A}_{\phi_1})(\vec{\mathbf{t}}_2,\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2) + (\mathbf{B}_{\boldsymbol{w}_2-\boldsymbol{w}_1})(\vec{\mathbf{t}}_2,\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2).\end{aligned} \tag{5.3.27}$$

Next, for the first and third terms on the right hand side of (5.3.27), we exploit the assumption (5.2.3) and the estimate given in [57, Lemma 3.6], respectively, to obtain

$$\begin{aligned}&\left|\int_\Omega \Big(f(\phi_1)-f(\phi_2)\Big)\boldsymbol{k}\cdot\Big\{(\boldsymbol{u}_1-\boldsymbol{u}_2)-\kappa_2\,\mathbf{div}\,(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)\Big\}\right|\\ &\qquad\leq L_f(1+\kappa_2^2)^{1/2}\,\|\phi_1-\phi_2\|_{0,\Omega}\,\|\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2\|,\end{aligned} \tag{5.3.28}$$

and

$$\begin{aligned}&\left|\int_\Omega \Big(\boldsymbol{u}_2\otimes(\boldsymbol{w}_2-\boldsymbol{w}_1)\Big)^{\mathrm{d}}:\Big\{\kappa_1(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)^{\mathrm{d}}-(\mathbf{t}_1-\mathbf{t}_2)\Big\}\right|\\ &\qquad\leq \|\boldsymbol{i}_c\|^2\,(1+\kappa_1^2)^{1/2}\,\|\boldsymbol{u}_2\|_{1,\Omega}\,\|\boldsymbol{w}_2-\boldsymbol{w}_1\|_{1,\Omega}\,\|\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2\|.\end{aligned} \tag{5.3.29}$$

On the other hand, for the second term of (5.3.27), we apply the assumptions (5.2.2), and the Cauchy-Schwarz and Hölder inequalities, to deduce that

$$\begin{aligned}\Big|(\mathbf{A}_{\phi_2}-\mathbf{A}_{\phi_1})(\vec{\mathbf{t}}_2,\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2)\Big| &= \left|\alpha\int_\Omega\Big(\mu(\phi_2)-\mu(\phi_1)\Big)\mathbf{t}_2:\Big\{(\mathbf{t}_1-\mathbf{t}_2)-\kappa_1(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)^{\mathrm{d}}\Big\}\right.\\ &\quad\left.+\int_\Omega\Big(\eta(\phi_2)-\eta(\phi_1)\Big)\boldsymbol{u}_2\cdot\Big\{(\boldsymbol{u}_1-\boldsymbol{u}_2)-\kappa_2\,\mathbf{div}\,(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)\Big\}\right|\\ &\leq\Big(\alpha L_\mu(1+\kappa_1^2)^{1/2}\,\|\phi_2-\phi_1\|_{\mathrm{L}^{2q}(\Omega)}\,\|\mathbf{t}_2\|_{\mathbb{L}^{2p}(\Omega)}\\ &\quad+L_\eta\,\|i_c\|^2\,(1+\kappa_2^2)^{1/2}\,\|\phi_2-\phi_1\|_{1,\Omega}\,\|\boldsymbol{u}_2\|_{1,\Omega}\Big)\|\vec{\mathbf{t}}_1-\vec{\mathbf{t}}_2\|\,,\end{aligned} \tag{5.3.30}$$

with $p,q\in(1,+\infty)$ such that $\frac{1}{p}+\frac{1}{q}=1$. At this point, we proceed as in [14, Lemma 3.9]. In fact, given the further regularity ε assumed in (5.3.24), we recall that the Sobolev embedding theorem (see, e.g [2, Thm. 4.12]) establishes the continuous injection $i_\varepsilon:\mathrm{H}^\varepsilon(\Omega)\to\mathrm{L}^{2p}(\Omega)$ with boundedness constant C_ε, where

$$2p=\begin{cases}\dfrac{2}{1-\varepsilon} & \text{if } n=2,\\[2mm] \dfrac{6}{3-2\varepsilon} & \text{if } n=3,\end{cases}$$

and $2q=\frac{n}{\varepsilon}$, and therefore, there holds

$$\|\mathbf{t}_2\|_{\mathbb{L}^{2p}(\Omega)}\leq C_\varepsilon\,\|\mathbf{t}_2\|_{\varepsilon,\Omega}\quad\forall\,\mathbf{t}_2\in\mathbb{H}^\varepsilon(\Omega). \tag{5.3.31}$$

Then, (5.3.31) could be bounded by (5.3.24), yielding for each $(\boldsymbol{w}_2,\phi_2)\in\mathbf{H}$ with $\|\boldsymbol{w}_2\|_{1,\Omega}+\|\phi_2\|_{1,\Omega}\leq r$, the estimate

$$\|\mathbf{t}_2\|_{\mathbb{L}^{2p}(\Omega)}\leq C_\varepsilon\widehat{C}(r)\left\{C_f\,\|\phi_2\|_{0,\Omega}+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma}\right\}.$$

Finally, denoting

$$\widetilde{C}_{\mathbf{S}}:=\frac{2}{\widetilde{\alpha}(\Omega)}\max\left\{(1+\kappa_2^2)^{1/2},\|\boldsymbol{i}_c\|^2\,(1+\kappa_1^2)^{1/2},\alpha C_\varepsilon L_\mu(1+\kappa_1^2)^{1/2},L_\eta\,\|i_c\|^2\,(1+\kappa_2^2)^{1/2}\right\},$$

inequalities (5.3.27), (5.3.28), (5.3.29), (5.3.30) and (5.3.31), imply (5.3.26) and complete the proof. □

Lemma 5.3.5. *There exists $\widetilde{C}_{\widetilde{\mathbf{S}}}>0$, such that for all $(\boldsymbol{w}_1,\phi_1)$, $(\boldsymbol{w}_2,\phi_2)\in\mathbf{H}$ there holds*

$$\begin{aligned}&\|\widetilde{\mathbf{S}}(\boldsymbol{w}_1,\phi_1)-\widetilde{\mathbf{S}}(\boldsymbol{w}_2,\phi_2)\|\\&\quad\leq\widetilde{C}_{\widetilde{\mathbf{S}}}\Big\{\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{1,\Omega}\,|\phi_1|_{1,\Omega}+\|\boldsymbol{w}_2\|_{1,\Omega}\,|\phi_1-\phi_2|_{1,\Omega}+\|\boldsymbol{w}_2\|_{1,\Omega}\,\|\phi_1-\phi_2\|_{0,\Omega}\Big\}.\end{aligned}\tag{5.3.32}$$

Proof. Given $(\boldsymbol{w}_1,\phi_1),(\boldsymbol{w}_2,\phi_2)\in\mathbf{H}$, we let (θ_1,λ_1), $(\theta_2,\lambda_2)\in\mathrm{H}^1(\Omega)\times\mathrm{H}^{-1/2}(\Gamma)$ be solutions to (5.3.13) corresponding to $(\boldsymbol{w}_1,\phi_1)$ and $(\boldsymbol{w}_2,\phi_2)$, respectively, that is $\theta_j=\widetilde{\mathbf{S}}(\boldsymbol{w}_j,\phi_j)$, $j\in\{1,2\}$. Then invoking the linearity of the forms $\boldsymbol{a}$ and $\mathbf{b}$, and performing algebraic manipulations, we deduce (using both formulations arising from (5.3.13)) that

$$\begin{aligned}\boldsymbol{a}(\theta_1-\theta_2,\psi)+(\psi,\lambda_1-\lambda_2)&=H_{\boldsymbol{w}_1-\boldsymbol{w}_2,\phi_1}(\psi)+H_{\boldsymbol{w}_2,\phi_1}(\psi)-H_{\boldsymbol{w}_2,\phi_2}(\psi)\quad\forall\,\psi\in\mathrm{H}^1(\Omega),\\(\theta_1-\theta_2,\xi)&=0\qquad\qquad\qquad\qquad\qquad\qquad\qquad\forall\,\xi\in\mathrm{H}^{-1/2}(\Gamma).\end{aligned}\tag{5.3.33}$$

Next, noting from the second equation of (5.3.33) that $\theta_1-\theta_2$ belongs to the kernel V of $\mathbf{b}$, taking $\psi=\theta_1-\theta_2$ and $\xi=\lambda_1-\lambda_2$ in (5.3.33), applying the ellipticity of $\boldsymbol{a}$ in V, and using the assumption (5.2.4), we readily deduce from the first equation of (5.3.33) that

$$\begin{aligned}\alpha_{\boldsymbol{a}}(\Omega)\,\|\theta_1-\theta_2\|_{1,\Omega}^2&\leq\boldsymbol{a}(\theta_1-\theta_2,\theta_1-\theta_2)\\&=H_{\boldsymbol{w}_1-\boldsymbol{w}_2,\phi_1}(\theta_1-\theta_2)+H_{\boldsymbol{w}_2,\phi_1}(\theta_1-\theta_2)-H_{\boldsymbol{w}_2,\phi_2}(\theta_1-\theta_2)\\&\leq\|i_c\|^2\Big\{(1+s_3)\,\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_{1,\Omega}\,|\phi_1|_{1,\Omega}+\|\boldsymbol{w}_2\|_{1,\Omega}\,|\phi_1-\phi_2|_{1,\Omega}\\&\quad+L_{\tilde{s}}\,\|\boldsymbol{w}_2\|_{1,\Omega}\,\|\phi_1-\phi_2\|_{0,\Omega}\Big\}\,\|\theta_1-\theta_2\|_{1,\Omega}\,,\end{aligned}$$

which gives (5.3.32) with $\widetilde{C}_{\widetilde{\mathbf{S}}}:=\frac{\|i_c\|^2}{\alpha_{\boldsymbol{a}}}\max\{1+s_3,L_{\tilde{s}}\}$. □

The announced property of $\mathbf{T}$ is proved now.

Lemma 5.3.6. *Let r and W as in Lemma 5.3.3. Then, there exists a positive constant $C_{\mathbf{T}}$ such that for all $(\boldsymbol{w}_1,\phi_1),(\boldsymbol{w}_2,\phi_2)\in W$ there holds*

$$\|\mathbf{T}(\boldsymbol{w}_1,\phi_1)-\mathbf{T}(\boldsymbol{w}_2,\phi_2)\|\leq C_{\mathbf{T}}\Big\{C_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma}+L_f\Big\}\,\|(\boldsymbol{w}_1,\phi_1)-(\boldsymbol{w}_2,\phi_2)\|\,.$$

Proof. It follows directly from the definition of $\mathbf{T}$ (cf. (5.3.14)) and the estimates (5.3.26) and (5.3.32). We remit to [57, Lemma 3.8] for similar further details. □

Finally, the main result of this section is given as follows.

Theorem 5.3.7. *Suppose that the parameters κ_1, κ_2 and κ_3 satisfy the conditions required by Lemma 5.3.1. Let r and W as in Lemma 5.3.3, and assume that the data satisfy (5.3.25) and*

$$C_{\mathbf{T}}\Big\{C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma} + L_f\Big\} < 1. \tag{5.3.34}$$

Then, problem (5.3.7) has a unique solution $(\vec{\mathbf{t}}, (\theta, \lambda)) \in H \times \mathrm{H}^1(\Omega) \times \mathrm{H}^{-1/2}(\Gamma)$, with $(\boldsymbol{u}, \theta) \in W$, and there holds

$$\|\vec{\mathbf{t}}\| \leq c_{\mathbf{S}}\left\{C_f\, r + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}\right\},$$

and

$$\|(\theta, \lambda)\| \leq \widetilde{c}_{\mathrm{S}}\{r\|\boldsymbol{u}\|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\}.$$

Proof. It follows as a combination of Lemmas 5.3.3 and 5.3.6, the assumption (5.3.34), the Banach fixed-point theorem, and the a priori estimates (5.3.15) and (5.3.22). We omit further details. □

5.3.3 The Galerkin scheme

In this section we analyse a Galerkin scheme associated with (5.3.7). We remark in advance that most of the details are omitted since they follow straightforwardly by adapting the fixed-point strategy from Section 5.3.2. We start by considering generic finite dimensional subspaces

$$\mathbb{H}_h^{\mathbf{t}} \subseteq \mathbb{L}^2_{\mathbf{tr}}(\Omega), \quad \mathbb{H}_h^{\boldsymbol{\sigma}} \subseteq \mathbb{H}_0(\mathbf{div};\Omega), \quad \mathbf{H}_h^{\boldsymbol{u}} \subseteq \mathbf{H}^1(\Omega), \quad \mathrm{H}_h^{\theta} \subseteq \mathrm{H}^1(\Omega), \quad \text{and} \quad \mathrm{H}_h^{\lambda} \subseteq \mathrm{H}^{-1/2}(\Gamma),$$

which will be specified later on. Hereafter, h denotes the size of a regular triangulation $\mathcal{T}_h$ of $\overline{\Omega}$ made up of triangles K (in $\mathbb{R}^2$) or tetrahedra K (in $\mathbb{R}^3$) of diameter h_K, i.e. $h := \max\{h_K : K \in \mathcal{T}_h\}$. Defining $H_h := \mathbb{H}_h^{\mathbf{t}} \times \mathbb{H}_h^{\boldsymbol{\sigma}} \times \mathbf{H}_h^{\boldsymbol{u}}$, and denoting $\vec{\mathbf{t}}_h := (\mathbf{t}_h, \boldsymbol{\sigma}_h, \boldsymbol{u}_h)$ and $\vec{\mathbf{s}}_h := (\mathbf{s}_h, \boldsymbol{\tau}_h, \boldsymbol{v}_h)$, the Galerkin scheme for (5.3.7) reads: Find $(\vec{\mathbf{t}}_h, (\theta_h, \lambda_h)) \in H_h \times \mathrm{H}_h^{\theta} \times \mathrm{H}_h^{\lambda}$ such that

$$\begin{array}{rcll} \mathbf{A}_{\theta_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) + \mathbf{B}_{\boldsymbol{u}_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) & = & F_{\theta_h}(\vec{\mathbf{s}}_h) + F_{\mathrm{D}}(\vec{\mathbf{s}}_h) & \forall\, \vec{\mathbf{s}}_h \in H_h, \\ \mathbf{a}(\theta_h, \psi_h) + \mathbf{b}(\psi_h, \lambda_h) & = & H_{\boldsymbol{u}_h,\theta_h}(\psi_h) & \forall\, \psi_h \in \mathrm{H}_h^{\theta}, \\ \mathbf{b}(\theta_h, \xi_h) & = & G(\xi_h) & \forall\, \xi_h \in \mathrm{H}_h^{\lambda}. \end{array} \tag{5.3.35}$$

Next, we set $\mathbf{H}_h := \mathbf{H}_h^{\boldsymbol{u}} \times \mathrm{H}_h^{\theta}$ and let $\mathbf{S}_h : \mathbf{H}_h \to H_h$ be the operator defined as

$$\mathbf{S}_h(\boldsymbol{w}_h, \phi_h) = (\mathbf{S}_{1,h}(\boldsymbol{w}_h, \phi_h), \mathbf{S}_{2,h}(\boldsymbol{w}_h, \phi_h), \mathbf{S}_{3,h}(\boldsymbol{w}_h, \phi_h)) := \vec{\mathbf{t}}_h \qquad \forall\, (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h, \tag{5.3.36}$$

where $\vec{\mathbf{t}}_h \in H_h$ is the unique solution of the problem given by the first equation of (5.3.35) with $(\boldsymbol{w}_h, \phi_h)$ instead of $(\boldsymbol{u}_h, \theta_h)$, that is

$$\mathbf{A}_{\phi_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) + \mathbf{B}_{\boldsymbol{w}_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) = F_{\phi_h}(\vec{\mathbf{s}}_h) + F_{\mathrm{D}}(\vec{\mathbf{s}}_h) \quad \forall\, \vec{\mathbf{s}}_h \in H_h. \tag{5.3.37}$$

Just for sake of completeness, we recall here that the functional F_{D} is defined in (5.3.9d). In turn, for a given pair $(\boldsymbol{w}_h, \phi_h)$, the bilinear forms $\mathbf{A}_{\phi_h}$ and $\mathbf{B}_{\boldsymbol{w}_h}$, and the functional F_{ϕ_h} are those corresponding to (5.3.8a), (5.3.8b) and (5.3.9c), respectively, with $\boldsymbol{w} = \boldsymbol{w}_h$ and $\phi = \phi_h$.

Furthermore, we define $\widetilde{\mathbf{S}}_h : \mathbf{H}_h \to \mathrm{H}_h^\theta$ as

$$\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h) := \theta_h \qquad \forall\, (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h\,, \tag{5.3.38}$$

where θ_h is the first component of the unique solution $(\theta_h, \lambda_h) \in \mathrm{H}_h^\theta \times \mathrm{H}_h^\lambda$ of the problem given by the second and third equations of (5.3.35) with $(\boldsymbol{w}_h, \phi_h)$ instead of $(\boldsymbol{u}_h, \theta_h)$, that is

$$\begin{aligned} \mathbf{a}(\theta_h, \psi_h) \,+\, \mathbf{b}(\psi_h, \lambda_h) &= H_{\boldsymbol{w}_h, \phi_h}(\psi_h) \quad \forall\, \psi_h \in \mathrm{H}_h^\theta(\Omega)\,, \\ \mathbf{b}(\theta_h, \xi_h) &= G(\xi_h) \qquad\quad \forall\, \xi_h \in \mathrm{H}_h^\lambda. \end{aligned} \tag{5.3.39}$$

The forms $\boldsymbol{a}$ and $\mathbf{b}$ and the functional G are defined in (5.3.9a), (5.3.9b) and (5.3.9f), respectively, whereas $H_{\boldsymbol{w}_h, \phi_h}$ is defined as in (5.3.9e) with $\boldsymbol{w} = \boldsymbol{w}_h$ and $\phi = \phi_h$.

Finally, by introducing the operator $\mathbf{T}_h : \mathbf{H}_h \to \mathbf{H}_h$ as

$$\mathbf{T}_h(\boldsymbol{w}_h, \phi_h) = \Big(\mathbf{S}_{3,h}(\boldsymbol{w}_h, \phi_h), \widetilde{\mathbf{S}}_h(\mathbf{S}_{3,h}(\boldsymbol{w}_h, \phi_h), \phi_h)\Big) \qquad \forall\, (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h,$$

we see that solving (5.3.35) is equivalent to seeking $(\boldsymbol{u}_h, \theta_h) \in \mathbf{H}_h$ such that

$$\mathbf{T}_h(\boldsymbol{u}_h, \theta_h) = (\boldsymbol{u}_h, \theta_h). \tag{5.3.40}$$

Certainly, all the above makes sense if we guarantee that the uncoupled discrete problems (5.3.37) and (5.3.39) are well-posed, which is addressed in what follows. We begin with the corresponding result for $\mathbf{S}_h$, which actually follows almost verbatim to that of its continuous counterpart $\mathbf{S}$, and proof can be omitted.

Lemma 5.3.8. *Assume that* $\kappa_1 \in \Big(0, \frac{2\mu_1\delta_1}{\mu_2}\Big)$, $\kappa_2 \in \Big(0, \frac{2\eta_1\delta_3}{\eta_2}\Big)$ *and* $\kappa_3 \in \Big(0, 2\alpha\delta_2\Big(\mu_1 - \frac{\kappa_1\mu_2}{2\delta_1}\Big)\Big)$, *with* $\delta_1 \in \Big(0, \frac{2}{\alpha\mu_2}\Big)$, $\delta_2 \in (0,2)$ *and* $\delta_3 \in \Big(0, \frac{2}{\eta_2}\Big)$. *Then, there exists* $r_0 > 0$ *such that for each* $r \in (0, r_0)$, *problem* (5.3.37) *has a unique solution* $\mathbf{S}_h(\boldsymbol{w}_h, \phi_h) := \vec{\mathbf{t}}_h \in H_h$ *for each* $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$ *with* $\|\boldsymbol{w}_h\|_{1,\Omega} \leq r$. *Moreover, there exists* $c_{\mathbf{S}} > 0$, *independent of* $(\boldsymbol{w}_h, \phi_h)$, *such that*

$$\|\mathbf{S}_h(\boldsymbol{w}_h, \phi_h)\| = \|\vec{\mathbf{t}}_h\| \leq c_{\mathbf{S}} \Big\{ C_f \, \|\phi_h\|_{0,\Omega} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \Big\} \quad \forall\, (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h. \tag{5.3.41}$$

In turn, in order to analyse the problem (5.3.39), we need to incorporate further hypotheses on the discrete spaces H_h^θ and H_h^λ. For this purpose, we now let

$$V_h := \Big\{ \psi_h \in \mathrm{H}_h^\theta : \quad (\psi_h, \xi_h) = 0 \quad \forall\, \xi_h \in \mathrm{H}_h^\lambda \Big\}\,,$$

be the discrete kernel of $\mathbf{b}$. Then, assuming the following discrete inf-sup conditions (which do hold for some finite element spaces, as those listed at the end of this section):

(**H.0**) There exists a constant $\alpha_1 > 0$, independent of h, such that

$$\sup_{\substack{\psi_h \in V_h \\ \psi_h \neq 0}} \frac{\boldsymbol{a}(\psi_h, \varphi_h)}{\|\psi_h\|_{1,\Omega}} \geq \alpha_1 \, \|\varphi_h\|_{1,\Omega} \quad \forall\, \varphi_h \in V_h. \tag{5.3.42}$$

(**H.1**) There exists a constant $\alpha_2 > 0$, independent of h, such that

$$\sup_{\substack{\psi_h \in \mathrm{H}_h^\theta \\ \psi_h \neq 0}} \frac{(\psi_h, \xi_h)}{\|\psi_h\|_{1,\Omega}} \geq \alpha_2 \|\xi_h\|_{-1/2,\Gamma} \quad \forall\, \xi_h \in \mathrm{H}_h^\lambda,$$

we can prove that the operator $\widetilde{\mathbf{S}}_h$ is well-posed, which is abridged in the following lemma. We refer to [57, Lemma 4.2] for further details.

Lemma 5.3.9. *For each $(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h$, problem (5.3.39) has a unique solution $(\theta_h, \lambda_h) = (\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h), \lambda_h) \in \mathrm{H}_h^\theta \times \mathrm{H}_h^\lambda$. Moreover, there exists a constant $\widetilde{C} > 0$ independent of $(\boldsymbol{w}_h, \phi_h)$, such that*

$$\|\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h)\| \leq \|(\theta_h, \lambda_h)\| \leq \widetilde{C} \left\{ \|\boldsymbol{w}_h\|_{1,\Omega} |\phi_h|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \right\}.$$

The solvability of the fixed-point problem (5.3.40) is now proved by means of the Brouwer fixed-point theorem (see, e.g. [54, Thm. 9.9-2]). We begin with the discrete version of Lemma 5.3.3.

Lemma 5.3.10. *Given $r \in (0, r_0)$, with r_0 as in (5.3.20), we let $W_h := \left\{ (\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h : \|(\boldsymbol{w}_h, \phi_h)\| \leq r \right\}$, and assume that*

$$\widetilde{c}(r) \left\{ C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \right\} + \widetilde{C} \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \leq r, \tag{5.3.43}$$

where $\widetilde{c}(r) := (1 + \widetilde{C})\, c_{\mathbf{S}} \max\{1, r\}$, and $c_{\mathbf{S}}$ and $\widetilde{C}$ are the constants specified in Lemmas 5.3.1 and 5.3.9, respectively. Then $\mathbf{T}_h(W_h) \subseteq W_h$.

The discrete analogue of Lemma 5.3.4 is provided next. We notice in advance that, instead of the regularity assumptions employed in the continuous case (not applicable in the present discrete case), we simply utilise an $\mathrm{L}^4 - \mathrm{L}^4 - \mathrm{L}^2$ argument.

Lemma 5.3.11. *Let $r \in (0, r_0)$ with r_0 given by (5.3.20). Then, there exists a constant $\widetilde{C}_{\mathbf{S}} > 0$, independent of r, such that for all $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in \mathbf{H}_h$, with $\|\boldsymbol{w}_h\|_{1,\Omega}$, $\|\widetilde{\boldsymbol{w}}_h\|_{1,\Omega} \leq r$, there holds*

$$\begin{aligned} \|\mathbf{S}_h(\boldsymbol{w}_h, \phi_h) - \mathbf{S}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \leq \widetilde{C}_{\mathbf{S}} \Big\{ & \|\mathbf{S}_{3,h}(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\|_{1,\Omega} \left(\|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega} + \|\phi_h - \widetilde{\phi}_h\|_{1,\Omega} \right) \\ & + \|\phi_h - \widetilde{\phi}_h\|_{4,\Omega} \|\mathbf{S}_{1,h}(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\|_{4,\Omega} + L_f \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} \Big\}. \end{aligned}$$

Proof. It proceeds exactly as in the proof of Lemma 5.3.4, except for the derivation of the discrete analogue of (5.3.30), where, instead of choosing the values of p, q determined by the regularity parameter ε, it suffices to take $p = q = 2$, thus obtaining

$$\begin{aligned} |(\mathbf{A}_{\widetilde{\phi}_h} - \mathbf{A}_{\phi_h})(\vec{\boldsymbol{r}}_h, \vec{\mathbf{t}}_h - \vec{\boldsymbol{r}}_h)| \leq & \left(\alpha L_\mu (1 + \kappa_1^2)^{1/2} \|\widetilde{\phi}_h - \phi_h\|_{4,\Omega} \|\boldsymbol{r}_h\|_{4,\Omega} \right. \\ & \left. + L_\eta c_4(\Omega) (1 + \kappa_2^2)^{1/2} \|\widetilde{\phi}_h - \phi_h\|_{1,\Omega} \|\boldsymbol{z}_h\|_{1,\Omega} \right) \|\vec{\mathbf{t}}_h - \vec{\boldsymbol{r}}_h\|, \end{aligned}$$

for all $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)$, with $\vec{\mathbf{t}}_h = (\mathbf{t}_h, \boldsymbol{\sigma}_h, \boldsymbol{u}_h) := \mathbf{S}_h(\boldsymbol{w}_h, \phi_h) \in H_h$ and $\vec{\boldsymbol{r}}_h := (\boldsymbol{r}_h, \boldsymbol{\zeta}_h, \boldsymbol{z}_h) = \mathbf{S}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in H_h$. Thus, since the elements of $\mathbb{H}_h^{\mathbf{t}}$ are piecewise polynomials, we know that $\|\boldsymbol{r}_h\|_{4,\Omega} < +\infty$ for each $\boldsymbol{r}_h \in \mathbb{H}_h^{\mathbf{t}}$. □

The discrete version of Lemma 5.3.5 is given as follows.

Lemma 5.3.12. *There exists a constant $\widehat{C}_{\widetilde{\mathbf{S}}} > 0$, such that for all $(\boldsymbol{w}_h, \phi_h)$, $(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in \mathbf{H}_h$, there holds*

$$\begin{aligned}&\|\widetilde{\mathbf{S}}_h(\boldsymbol{w}_h, \phi_h) - \widetilde{\mathbf{S}}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \\ &\qquad \leq \widehat{C}_{\widetilde{\mathbf{S}}}\Big\{ \|\boldsymbol{w}_h - \widetilde{\boldsymbol{w}}_h\|_{1,\Omega}\, \|\phi_h\|_{1,\Omega} + \|\widetilde{\boldsymbol{w}}_h\|_{1,\Omega} |\phi_h - \widetilde{\phi}_h|_{1,\Omega} + \|\widetilde{\boldsymbol{w}}_h\|_{1,\Omega} \|\phi_h - \widetilde{\phi}_h\|_{0,\Omega} \Big\}.\end{aligned}$$

Proof. It follows the same arguments from Lemma 5.3.5, but now using the inf-sup condition (5.3.42) rather than the V-ellipticity of $\boldsymbol{a}$. □

Next, utilizing Lemmas 5.3.11 and 5.3.12, we can prove the discrete version of Lemma 5.3.6.

Lemma 5.3.13. *Let r and W_h as in Lemma 5.3.10. Then, there exists a constant $\widetilde{C}_{\mathbf{T}} > 0$, such that for all $(\boldsymbol{w}_h, \phi_h), (\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h) \in \mathbf{H}_h$, there holds*

$$\begin{aligned}&\|\mathbf{T}_h(\boldsymbol{w}_h, \phi_h) - \mathbf{T}_h(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\| \\ &\qquad \leq \widetilde{C}_{\mathbf{T}}\Big\{ \|\mathbf{S}_{3,h}(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\|_{1,\Omega} + \|\mathbf{S}_{1,h}(\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\|_{4,\Omega} + L_f \Big\} \|(\boldsymbol{w}_h, \phi_h) - (\widetilde{\boldsymbol{w}}_h, \widetilde{\phi}_h)\|.\end{aligned}$$

Notice that the previous lemma provides the continuity required by the Brouwer fixed-point theorem, in the convex and compact set $W_h \subseteq \mathbf{H}_h$. Therefore, we have the following result.

Theorem 5.3.14. *Suppose that the parameters κ_1, κ_2 and κ_3 satisfy the conditions required by Lemma 5.3.8. Let r and W_h as in Lemma 5.3.10, and assume that the data satisfy (5.3.43). Then, the problem (5.3.35) has at least one solution $(\vec{\mathbf{t}}_h, (\theta_h, \lambda_h)) \in H_h \times \mathrm{H}^{\theta}_h \times \mathrm{H}^{\lambda}_h$, with $(\boldsymbol{u}_h, \theta_h) \in W_h$, and there holds*

$$\|\vec{\mathbf{t}}_h\| \leq c_{\mathbf{S}} \left\{ C_f\, r + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \right\},$$

and

$$\|(\theta_h, \lambda_h)\| \leq \widetilde{C} \left\{ r \|\boldsymbol{u}_h\|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \right\}.$$

5.3.4 A priori error analysis

Our next goal is to derive an a priori error estimate for our Galerkin scheme (5.3.35). More precisely, given $(\mathbf{t}, \boldsymbol{\sigma}, \boldsymbol{u}, (\theta, \lambda)) := (\vec{\mathbf{t}}, (\theta, \lambda)) \in H \times \mathrm{H}^1(\Omega) \times \mathrm{H}^{-1/2}(\Gamma)$, with $(\boldsymbol{u}, \theta) \in W$, and $(\mathbf{t}_h, \boldsymbol{\sigma}_h, \boldsymbol{u}_h, (\theta_h, \lambda_h))$ $:= (\vec{\mathbf{t}}_h, (\theta_h, \lambda_h)) \in H_h \times \mathrm{H}^{\theta}_h \times \mathrm{H}^{\lambda}_h$, with $(\boldsymbol{u}_h, \theta_h) \in W_h$, solutions of the problems (5.3.7) and (5.3.35), respectively, we are interested in obtaining an upper bound for

$$\|(\vec{\mathbf{t}}, (\theta, \lambda)) - (\vec{\mathbf{t}}_h, (\theta_h, \lambda_h))\|.$$

To this end, we apply two instrumental results from [127, Thm. 11.1 and 11.2] concerning Strang-type estimates for elliptic and saddle point problems, respectively, where continuous and discrete formulations differ only in the functionals involved. We begin with the following preliminary estimate.

Lemma 5.3.15. *There exists a constant $C_{\mathbf{ST}} > 0$, independent of h, such that*

$$\begin{aligned}\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| \leq C_{\mathbf{ST}} \Big\{ & \operatorname{dist}(\vec{\mathbf{t}}, H_h) + L_f \|\theta - \theta_h\|_{1,\Omega} + \|\theta - \theta_h\| \|\mathbf{t}\|_{\varepsilon,\Omega} \\ & + \|\boldsymbol{u}\|_{1,\Omega} \|\theta - \theta_h\|_{1,\Omega} + \|\boldsymbol{u}\|_{1,\Omega} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} \Big\}.\end{aligned} \tag{5.3.44}$$

Proof. From Lemma (5.3.1) we observe that $\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}}$ and $\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h}$ are bounded and uniformly elliptic bilinear forms with ellipticity constant $\frac{\widetilde{\alpha}(\Omega)}{2}$. Also, $F_\theta + F_{\mathrm{D}}$ and $F_{\theta_h} + F_{\mathrm{D}}$ are linear bounded functionals in H and H_h, respectively. Thus, a straightforward application of [127, Thm. 11.1] to the context given by the first equations of (5.3.7) and (5.3.35), yields

$$\begin{aligned}\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| \leq \overline{C}_1 \Bigg\{ & \sup_{\substack{\vec{\mathbf{s}}_h \in H_h \\ \vec{\mathbf{s}}_h \neq 0}} \frac{|F_\theta(\vec{\mathbf{s}}_h) - F_{\theta_h}(\vec{\mathbf{s}}_h)|}{\|\vec{\mathbf{s}}_h\|} \\ & + \inf_{\substack{\vec{\mathbf{q}}_h \in H_h \\ \vec{\mathbf{q}}_h \neq 0}} \left(\|\vec{\mathbf{t}} - \vec{\mathbf{q}}_h\| + \sup_{\substack{\vec{\mathbf{s}}_h \in H_h \\ \vec{\mathbf{s}}_h \neq 0}} \frac{|(\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h) - (\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h)|}{\|\vec{\mathbf{s}}_h\|} \right) \Bigg\},\end{aligned} \tag{5.3.45}$$

where $\overline{C}_1 := \frac{2}{\widetilde{\alpha}(\Omega)} \max\{1, \|\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}}\|\}$. Hence, in order to estimate the last supremum in (5.3.45), we add and subtract suitable terms to obtain

$$\begin{aligned}(\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h) - (\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h) &= (\mathbf{A}_\theta - \mathbf{A}_{\theta_h})(\vec{\mathbf{t}}, \vec{\mathbf{s}}_h) + (\mathbf{B}_{\boldsymbol{u}} - \mathbf{B}_{\boldsymbol{u}_h})(\vec{\mathbf{t}}, \vec{\mathbf{s}}_h) \\ &\quad + (\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h})(\vec{\mathbf{q}}_h - \vec{\mathbf{t}}, \vec{\mathbf{s}}_h) + (\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}})(\vec{\mathbf{q}}_h - \vec{\mathbf{t}}, \vec{\mathbf{s}}_h),\end{aligned}$$

and then, using the boundedness of the bilinear forms $\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}}$ and $\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h}$, the estimate (5.3.31), and the continuous embedding $\mathrm{H}^1(\Omega) \to \mathrm{L}^{n/\varepsilon}(\Omega)$ with constant $\widetilde{C}_\varepsilon$, we obtain

$$\begin{aligned}&|(\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h) - (\mathbf{A}_{\theta_h} + \mathbf{B}_{\boldsymbol{u}_h})(\vec{\mathbf{q}}_h, \vec{\mathbf{s}}_h)| \\ &\leq \Big\{ \alpha L_\mu C_\varepsilon \widetilde{C}_\varepsilon (1 + \kappa_1^2)^{1/2} \|\mathbf{t}\|_{\varepsilon,\Omega} \|\theta - \theta_h\|_{1,\Omega} + L_\eta (1 + \kappa_2^2)^{1/2} \|\theta - \theta_h\|_{1,\Omega} \|\boldsymbol{u}\|_{1,\Omega} \\ &\quad + \|\boldsymbol{i}_c\|^2 (1 + \kappa_1^2)^{1/2} \|\boldsymbol{u}\|_{1,\Omega} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} + 2\|\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}}\| \|\vec{\mathbf{q}}_h - \vec{\mathbf{t}}\| \Big\} \|\vec{\mathbf{s}}_h\|.\end{aligned} \tag{5.3.46}$$

In turn, similarly as in (5.3.28), we note that

$$|(F_{\theta_h} - F_\theta)(\vec{\mathbf{s}}_h)| \leq L_f (1 + \kappa_2^2)^{1/2} \|\theta - \theta_h\|_{0,\Omega} \|\vec{\mathbf{s}}_h\|. \tag{5.3.47}$$

Finally, by replacing (5.3.46) and (5.3.47) back into (5.3.45), one obtains (5.3.44) with constant $C_{\mathbf{ST}}$ depending on $\widetilde{\alpha}(\Omega), L_\mu, C_\varepsilon, \widetilde{C}_\varepsilon, L_\eta, \|\boldsymbol{i}_c\|$ and $\|\mathbf{A}_\theta + \mathbf{B}_{\boldsymbol{u}}\|$. □

Next, we have the following complementary result.

Lemma 5.3.16. *There exists a constant $\widetilde{C}_{\mathbf{ST}} > 0$ independent of h, such that*

$$\begin{aligned}\|(\theta, \lambda) - (\theta_h, \lambda_h)\| \leq \widetilde{C}_{\mathbf{ST}} \Big\{ & \operatorname{dist}\Big((\theta, \lambda), \mathrm{H}_h^\theta \times \mathrm{H}_h^\lambda\Big) + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} \, |\theta|_{1,\Omega} \\ & + \|\boldsymbol{u}_h\|_{1,\Omega} \, |\theta - \theta_h|_{1,\Omega} + \|\boldsymbol{u}_h\|_{1,\Omega} \|\theta - \theta_h\|_{0,\Omega} \Big\}.\end{aligned} \tag{5.3.48}$$

Proof. We first observe that (**H.0**) and (**H.1**) guarantee the main hypothesis in [127, Thm. 11.2]. Hence, by applying this lemma to the context given by the second and third equations of (5.3.7) and (5.3.35), we arrive at

$$\|(\theta,\lambda)-(\theta_h,\lambda_h)\| \leq \overline{C}_2 \left\{ \|(H_{\boldsymbol{u},\theta}-H_{\boldsymbol{u}_h,\theta_h})|_{\mathrm{H}_h^\theta}\| + \mathrm{dist}\left((\theta,\lambda),\mathrm{H}_h^\theta\times\mathrm{H}_h^\lambda\right)\right\}, \quad (5.3.49)$$

where $\overline{C}_2$ is a constant depending on $\alpha_1,\alpha_2,\|\boldsymbol{a}\|,\|\mathbf{b}\|$. Next, analogously to the proof of Lemma 5.3.5, we can assert that

$$\begin{aligned}\|(H_{\boldsymbol{u},\theta}-H_{\boldsymbol{u}_h,\theta_h})|_{\mathrm{H}_h^\theta}\| &= \|(H_{\boldsymbol{u}-\boldsymbol{u}_h,\theta}+H_{\boldsymbol{u}_h,\theta}-H_{\boldsymbol{u}_h,\theta_h})|_{\mathrm{H}_h^\theta}\| \\ &\leq \|i_c\|^2\left\{(1+s_3)\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\,|\theta|_{1,\Omega}+\|\boldsymbol{u}_h\|_{1,\Omega}\,|\theta-\theta_h|_{1,\Omega}+L_{\tilde{s}}\,\|\boldsymbol{u}_h\|_{1,\Omega}\,\|\theta-\theta_h\|_{0,\Omega}\right\}. \end{aligned} \quad (5.3.50)$$

Finally, the required estimate (5.3.48) follows by replacing (5.3.50) back into (5.3.49), with constant $\widetilde{C}_{\mathbf{ST}}$ depending on $\alpha_1,\alpha_2,\|\boldsymbol{a}\|,\|\mathbf{b}\|,\|i_c\|,s_3$ and $L_{\tilde{s}}$. □

We remark that an alternative way to prove the previous results follows similarly as in [83, Lemma 3.11] and [81, Thm. 2.6], respectively.

Having established bounds for $\|\vec{\mathbf{t}}-\vec{\mathbf{t}}_h\|$ and $\|(\theta,\lambda)-(\theta_h,\lambda_h)\|$, we are now able to derive the Céa estimate for the global error. In fact, by adding the estimates (5.3.44) and (5.3.48), and applying the continuous injection $\mathrm{H}^1(\Omega)\to\mathrm{L}^2(\Omega)$, we obtain

$$\begin{aligned}\|\vec{\mathbf{t}}-\vec{\mathbf{t}}_h\|+\|(\theta,\lambda)-(\theta_h,\lambda_h)\| &\leq C_{\mathbf{ST}}\,\mathrm{dist}(\vec{\mathbf{t}},H_h)+\widetilde{C}_{\mathbf{ST}}\,\mathrm{dist}\left((\theta,\lambda),\mathrm{H}_h^\theta\times\mathrm{H}_h^\lambda\right) \\ &+\left\{C_{\mathbf{ST}}\left(L_f+\|\mathbf{t}\|_{\varepsilon,\Omega}+\|\boldsymbol{u}\|_{1,\Omega}\right)+2\widetilde{C}_{\mathbf{ST}}\,\|\boldsymbol{u}_h\|_{1,\Omega}\right\}\|\theta-\theta_h\|_{1,\Omega} \\ &+\left\{C_{\mathbf{ST}}\,\|\boldsymbol{u}\|_{1,\Omega}+\widetilde{C}_{\mathbf{ST}}|\theta|_{1,\Omega}\right\}\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}.\end{aligned}$$

Now, we note that the terms $\|\boldsymbol{u}\|_{1,\Omega}$, $|\theta|_{1,\Omega}$, $\|\boldsymbol{u}_h\|_{1,\Omega}$ and $\|\mathbf{t}\|_{\varepsilon,\Omega}$ can be bounded by data using the estimates (5.3.15), (5.3.22), (5.3.41) and (5.3.24), respectively. Therefore, performing some algebraic manipulations, and introducing the constants:

$$C_5 := C_{\mathbf{ST}}C_\varepsilon\widehat{C}(r), \quad C_6 := 2C_{\mathbf{ST}}c_{\mathbf{S}}+\widetilde{C}_{\mathbf{ST}}\widetilde{c}_{\mathbf{S}}c_{\mathbf{S}}r+2\widetilde{C}_{\mathbf{ST}}c_{\mathbf{S}}, \quad (5.3.51)$$

$$C_7 := \max\{C_{\mathbf{ST}},C_5,(C_5+C_6)r,C_6,\widetilde{C}_{\mathbf{ST}}\widetilde{c}_{\mathbf{S}}\},$$

it can be show that

$$\begin{aligned}\|\vec{\mathbf{t}}-\vec{\mathbf{t}}_h\|+\|(\theta,\lambda)-(\theta_h,\lambda_h)\| &\leq C_{\mathbf{ST}}\,\mathrm{dist}(\vec{\mathbf{t}},H_h)+\widetilde{C}_{\mathbf{ST}}\,\mathrm{dist}\left((\theta,\lambda),\mathrm{H}_h^\theta\times\mathrm{H}_h^\lambda\right) \\ &+C_7\left(L_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Omega}+C_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Omega}+\|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\right)\left\{\|\vec{\mathbf{t}}-\vec{\mathbf{t}}_h\|+\|(\theta,\lambda)-(\theta_h,\lambda_h)\|\right\}.\end{aligned} \quad (5.3.52)$$

Consequently, we can establish the following main result.

Theorem 5.3.17. *Assume that the data satisfy*

$$C_7\left\{L_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Omega}+C_f+\|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Omega}+\|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\right\}<\frac{1}{2}. \quad (5.3.53)$$

Then, there exists a positive constant C_8 independent of h, such that

$$\|\vec{\mathbf{t}}-\vec{\mathbf{t}}_h\|+\|(\theta,\lambda)-(\theta_h,\lambda_h)\| \leq C_8\left\{\mathrm{dist}(\vec{\mathbf{t}},H_h)+\mathrm{dist}\left((\theta,\lambda),\mathrm{H}_h^\theta\times\mathrm{H}_h^\lambda\right)\right\}. \quad (5.3.54)$$

Proof. It follows directly from (5.3.52) and (5.3.53). □

As a first remark of the previous theorem, we stress that the ultra-weak sense in which the symmetry of $\boldsymbol{\sigma}$ was imposed (cf. (5.3.5)) does not affect the expected asymptotic symmetry of the discrete tensor $\boldsymbol{\sigma}_h$. In fact, adding and subtracting the symmetric unknown $\boldsymbol{\sigma}$ in the below estimate, we obtain

$$\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\| = \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma} + \boldsymbol{\sigma}^{\mathrm{t}} - \boldsymbol{\sigma}_h^{\mathrm{t}}\| \leq C_8 \Big\{ \mathrm{dist}(\vec{\mathbf{t}}, H_h) + \mathrm{dist}\left((\theta,\lambda), \mathrm{H}_h^{\theta} \times \mathrm{H}_h^{\lambda}\right) \Big\}, \tag{5.3.55}$$

which yields $\lim_{h\to 0} \|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{t}}\| = 0$, and then, we have actually proved that $\boldsymbol{\sigma}_h$ tends to a symmetric tensor. In second place, exactly as in [43, Section 4] we obtain the error for the postprocessed pressure: there exists a positive constant $\widehat{C}$, independent of h, such that

$$\|p - p_h\|_{0,\Omega} \leq \widehat{C} \Big\{ \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega} + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} \Big\}.$$

5.3.5 Specific finite element subspaces

In this section we specify concrete discrete subspaces and make precise the convergence rate for (5.3.35). Given an integer $k \geq 0$, for each $K \in \mathcal{T}_h$ we let $\mathrm{P}_k(K)$ be the space of polynomial functions on K of degree $\leq k$ and define the local Raviart-Thomas space of order k as

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

where $\mathbf{P}_k(K) = [\mathrm{P}_k(K)]^n$, and $\boldsymbol{x}$ is the generic vector in R^n. Then, we consider piecewise polynomials of degree $\leq k$ for approximating entries of the strain rate $\mathbf{t}$, the global Raviart-Thomas space of order k to approximate rows of the pseudostress $\boldsymbol{\sigma}$, and the Lagrange space given by the continuous piecewise polynomial vectors of degree $\leq k+1$ for the velocity $\boldsymbol{u}$, respectively, that is

$$\begin{aligned}
\mathbb{H}_h^{\mathbf{t}} &:= \left\{ \mathbf{s}_h \in \mathbb{L}^2_{\mathrm{tr}}(\Omega) : \quad \mathbf{s}_h|_K \in \mathbb{P}_k(K) \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbb{H}_h^{\boldsymbol{\sigma}} &:= \left\{ \boldsymbol{\tau}_h \in \mathbb{H}_0(\mathbf{div};\Omega) : \quad \boldsymbol{c}^{\mathrm{t}}\boldsymbol{\tau}_h|_K \in \mathbf{RT}_k(K), \quad \forall\, \boldsymbol{c} \in \mathrm{R}^n \quad \forall\, K \in \mathcal{T}_h \right\}, \\
\mathbf{H}_h^{\boldsymbol{u}} &:= \left\{ \boldsymbol{v}_h \in \mathbf{C}(\overline{\Omega}) : \quad \boldsymbol{v}_h|_K \in \mathbf{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}.
\end{aligned} \tag{5.3.56}$$

The approximating space for temperature will consist of continuous piecewise polynomials of degree $\leq k+1$

$$\mathrm{H}_h^{\theta} := \left\{ \psi_h \in \mathrm{C}(\overline{\Omega}) : \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h \right\}. \tag{5.3.57}$$

For the normal heat flux, we let $\{\widetilde{\Gamma}_1, \widetilde{\Gamma}_2, ..., \widetilde{\Gamma}_m\}$ be an independent triangulation of Γ (made of straight segments in R^2, or triangles in R^3), and define $\widetilde{h} := \max_{j\in\{1,...,m\}} |\widetilde{\Gamma}_j|$. Then, with the same integer $k \geq 0$ used in definitions (5.3.56) and (5.3.57), we approximate λ by piecewise polynomials of degree $\leq k$ over this new mesh, that is

$$\mathrm{H}_{\widetilde{h}}^{\lambda} := \left\{ \xi_{\widetilde{h}} \in \mathrm{L}^2(\Gamma) : \quad \xi_{\widetilde{h}}|_{\widetilde{\Gamma}_j} \in \mathrm{P}_k(\widetilde{\Gamma}_j) \quad \forall\, j \in \{1, ..., m\} \right\}. \tag{5.3.58}$$

We remark that the spaces H_h^{θ} and $\mathrm{H}_{\widetilde{h}}^{\lambda}$ satisfy the inf-sup conditions **H.0** and **H.1**. We remit to (cf. [57, Lemma 4.10], [81, Lemma 4.7]) for further details.

Finally, approximation properties of the spaces in (5.3.56), (5.3.57) and (5.3.58) can be found in e.g [10, 39, 81], which combined with the Céa estimate (5.3.54) produce the theoretical rate of convergence of (5.3.35), summarised in what follows.

Theorem 5.3.18. *In addition to the hypotheses of Theorems 5.3.7, 5.3.14 and 5.3.17, assume that there exists $s > 0$ such that $\mathbf{t} \in \mathbb{H}^s(\Omega)$, $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega)$, $\mathbf{div}\,\boldsymbol{\sigma} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{u} \in \mathbf{H}^{1+s}(\Omega)$, $\theta \in \mathrm{H}^{1+s}(\Omega)$ and $\lambda \in \mathrm{H}^{-1/2+s}(\Gamma)$. Then, there exist positive constants C_0, $C > 0$, independent of h and $\widetilde{h}$, such that for all $h \leq C_0\widetilde{h}$, with the finite element subspaces defined by (5.3.56), (5.3.57) and (5.3.58), there holds*

$$\begin{aligned}\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| + \big\|(\theta,\lambda) - (\theta_h,\lambda_{\widetilde{h}})\big\| \;\leq\; & C\widetilde{h}^{\min\{s,k+1\}}\,\|\lambda\|_{-1/2+s,\Gamma} \\ & + Ch^{\min\{s,k+1\}}\Big\{\,\|\mathbf{t}\|_{s,\Omega} + \;\|\boldsymbol{\sigma}\|_{s,\Omega} + \|\mathbf{div}\,\boldsymbol{\sigma}\|_{s,\Omega} \;+\; \|\boldsymbol{u}\|_{1+s,\Omega} + \|\theta\|_{1+s,\Omega}\,\Big\}.\end{aligned}$$

Finally, we point out that (5.3.55) and the previous theorem imply that, under the same foregoing regularity assumptions, the approximating unknown $\boldsymbol{\sigma}_h$ converges to a symmetric tensor with the same rate of convergence of all the unknowns involved.

5.4 The fully-mixed approach

In this section we proceed similarly as in [59] to put forward a fully-mixed approach for (5.2.1). Then, we establish the corresponding continuous and discrete formulations, analyse their solvability by using fixed-point strategies, and derive the corresponding a priori error estimates.

5.4.1 The continuous formulation

Having established in Section 5.3 the mixed formulation for the Navier-Stokes-Brinkman problem, it only remains to define a mixed formulation for the energy equation. Let us introduce the unknown

$$\boldsymbol{\Theta} := \rho\kappa\nabla\theta - \theta\boldsymbol{u} - s(\theta)\boldsymbol{u} \quad \text{in } \Omega,$$

and then, denoting from now on the tensor $\rho^{-1}\kappa^{-1}$ simply as κ^{-1}, applying (5.2.1b) and performing some algebraic computations, we obtain

$$\kappa^{-1}\boldsymbol{\Theta} + \kappa^{-1}\theta\boldsymbol{u} + \kappa^{-1}s(\theta)\boldsymbol{u} = \nabla\theta \quad \text{in } \Omega, \quad \mathrm{div}\,\boldsymbol{\Theta} = 0 \quad \text{in } \Omega, \quad \theta = \theta_{\mathrm{D}} \quad \text{on } \Gamma. \tag{5.4.1}$$

In this way, testing the first equation in (5.4.1) against functions $\boldsymbol{\Phi} \in \mathbf{H}(\mathrm{div};\Omega)$, integrating by parts, and using the Dirichlet boundary condition for θ, we obtain

$$\int_\Omega \kappa^{-1}\boldsymbol{\Theta}\cdot\Phi + \int_\Omega \theta\,\mathrm{div}\,\Phi + \int_\Omega \kappa^{-1}\theta\boldsymbol{u}\cdot\boldsymbol{\Phi} \;=\; -\int_\Omega \kappa^{-1}s(\theta)\boldsymbol{u}\cdot\boldsymbol{\Phi} + \langle\boldsymbol{\Phi}\cdot\boldsymbol{\nu},\theta_{\mathrm{D}}\rangle_\Gamma\,. \tag{5.4.2}$$

In turn, testing the equilibrium equation in (5.4.1) against a suitable function ψ, we get

$$-\int_\Omega \psi\,\mathrm{div}\,\boldsymbol{\Theta} \;=\; 0.$$

Similarly as in Section 5.3, we note from the last term on the left-hand side of (5.4.2), that we require to seek the temperature θ in $\mathrm{H}^1(\Omega)$. Thus we are left with the preliminary weak formulation: Find $(\boldsymbol{\Theta},\theta) \in \mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$, such that

$$\begin{aligned}\int_\Omega \kappa^{-1}\boldsymbol{\Theta}\cdot\Phi + \int_\Omega \theta\,\mathrm{div}\,\Phi + \int_\Omega \kappa^{-1}\theta\boldsymbol{u}\cdot\boldsymbol{\Phi} \;&=\; -\int_\Omega \kappa^{-1}s(\theta)\boldsymbol{u}\cdot\boldsymbol{\Phi} + \langle\boldsymbol{\Phi}\cdot\boldsymbol{\nu},\theta_{\mathrm{D}}\rangle_\Gamma\,, \\ -\int_\Omega \psi\,\mathrm{div}\,\boldsymbol{\Theta} \;&=\; 0\,,\end{aligned} \tag{5.4.3}$$

for all $(\boldsymbol{\Phi},\psi) \in \mathbf{H}(\mathrm{div};\Omega)\times \mathrm{H}^1(\Omega)$. Again, the analysis will require to incorporate the following redundant terms:

$$\begin{aligned}\kappa_4\int_\Omega \big(\nabla\theta-\kappa^{-1}\theta\boldsymbol{u}-\kappa^{-1}s(\theta)\boldsymbol{u}-\kappa^{-1}\boldsymbol{\Theta}\big)\cdot\nabla\psi &= 0 \quad \forall\,\psi\in\mathrm{H}^1(\Omega),\\ \kappa_5\int_\Omega \mathrm{div}\,\boldsymbol{\Theta}\,\mathrm{div}\,\boldsymbol{\Phi} &= 0 \quad \forall\,\boldsymbol{\Phi}\in\mathbf{H}(\mathrm{div};\Omega),\\ \kappa_6\int_\Gamma \theta\,\psi &= \kappa_6\int_\Gamma \theta_{\mathrm{D}}\,\psi \quad \forall\,\psi\in\mathrm{H}^1(\Omega),\end{aligned}$$

where κ_4,κ_5 and κ_6 are positive parameters to be specified later on. Then, now we may consider the following mixed formulation for the energy equation: Find $(\boldsymbol{\Theta},\theta)\in\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$, such that

$$\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi))+\widetilde{\mathbf{b}}_{\boldsymbol{u}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi))=\widetilde{F}_{\boldsymbol{u},\theta}(\boldsymbol{\Phi},\psi)+\widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi},\psi), \tag{5.4.4}$$

for all $(\boldsymbol{\Phi},\psi)\in\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$, where, given an arbitrary $(\boldsymbol{w},\phi)\in\mathbf{H}$, the forms $\widetilde{\boldsymbol{a}}$, $\widetilde{\mathbf{b}}_{\boldsymbol{w}}$ and the functionals $\widetilde{F}_{\boldsymbol{w},\phi}$ and $\widetilde{F}_{\mathrm{D}}$ are defined, respectively, as

$$\begin{aligned}\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)) &:= \int_\Omega \kappa^{-1}\boldsymbol{\Theta}\cdot(\Phi-\kappa_4\nabla\psi)+\int_\Omega \theta\,\mathrm{div}\,\Phi-\int_\Omega\psi\,\mathrm{div}\,\boldsymbol{\Theta}\\ &\quad+\kappa_4\int_\Omega\nabla\theta\cdot\nabla\psi+\kappa_5\int_\Omega\mathrm{div}\,\boldsymbol{\Theta}\,\mathrm{div}\,\boldsymbol{\Phi}+\kappa_6\int_\Gamma\theta\,\psi\,, \qquad (5.4.5a)\\ \widetilde{\mathbf{b}}_{\boldsymbol{w}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)) &:= \int_\Omega \kappa^{-1}\theta\boldsymbol{w}\cdot(\boldsymbol{\Phi}-\kappa_4\nabla\psi)\,, \qquad (5.4.5b)\end{aligned}$$

for all $(\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)\in\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$, and

$$\widetilde{F}_{\boldsymbol{w},\phi}(\boldsymbol{\Phi},\psi) := \int_\Omega \kappa^{-1}s(\phi)\boldsymbol{w}\cdot(\kappa_4\nabla\psi-\boldsymbol{\Phi}), \tag{5.4.6a}$$

$$\widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi},\psi) := \langle\boldsymbol{\Phi}\cdot\boldsymbol{\nu},\theta_{\mathrm{D}}\rangle_\Gamma+\kappa_6\int_\Gamma\theta_{\mathrm{D}}\,\psi, \tag{5.4.6b}$$

for all $(\boldsymbol{\Phi},\psi)\in\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$. The fully-mixed variational formulation for (5.2.1) reduces therefore to the first equation of (5.3.7) and (5.4.4), i.e.: Find $(\vec{\mathbf{t}},(\boldsymbol{\Theta},\theta))\in H\times\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$ such that

$$\begin{aligned}\mathbf{A}_\theta(\vec{\mathbf{t}},\vec{\mathbf{s}})+\mathbf{B}_{\boldsymbol{u}}(\vec{\mathbf{t}},\vec{\mathbf{s}}) &= F_\theta(\vec{\mathbf{s}})+F_{\mathrm{D}}(\vec{\mathbf{s}}),\\ \widetilde{\boldsymbol{a}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi))+\widetilde{\mathbf{b}}_{\boldsymbol{u}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)) &= \widetilde{F}_{\boldsymbol{u},\theta}(\boldsymbol{\Phi},\psi)+\widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi},\psi),\end{aligned} \tag{5.4.7}$$

for all $(\vec{\mathbf{s}},(\boldsymbol{\Phi},\psi))\in H\times\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$.

We end this section by noticing that the present use of a mixed approach for the heat equation avoids the introduction of the unknown given by the normal boundary heat flux λ, as it was required in the primal formulation from Section 5.3.

5.4.2 Solvability analysis

The forms $\widetilde{\boldsymbol{a}}$ and $\widetilde{\mathbf{b}}_{\boldsymbol{u}}$ are defined exactly as in [59, Section 3.1] and therefore we omit parts of the proofs whenever necessary. On the other hand, for the solvability of (5.4.7), we propose a fixed-point approach as in Section 5.3.2. More precisely, in addition to using the operator $\mathbf{S}$ (cf. (5.3.10)

- (5.3.11)), and instead of (5.3.12) and (5.3.14), we define the operators $\widehat{\mathbf{S}} : \mathbf{H} \to \mathrm{H}^1(\Omega)$ and $\widehat{\mathbf{T}} := \mathbf{H} \to \mathbf{H}$ as $\widehat{\mathbf{S}}(\boldsymbol{w},\phi) := \theta \quad \forall\, (\boldsymbol{w},\phi) \in \mathbf{H}$, where θ is the second component of the unique solution $(\boldsymbol{\Theta},\theta) \in \mathbf{H}(\mathrm{div};\Omega) \times \mathrm{H}^1(\Omega)$ of the problem given by the second equation of (5.4.7) with $(\boldsymbol{w},\phi)$ instead of $(\boldsymbol{u},\theta)$, that is

$$\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)) + \widetilde{\mathbf{b}}_{\boldsymbol{w}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi)) = \widetilde{F}_{\boldsymbol{w},\phi}(\boldsymbol{\Phi},\psi) + \widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi},\psi), \tag{5.4.8}$$

for all $(\boldsymbol{\Phi},\psi) \in \mathbf{H}(\mathrm{div};\Omega) \times \mathrm{H}^1(\Omega)$, and

$$\widehat{\mathbf{T}}(\boldsymbol{w},\phi) = \Big(\mathbf{S}_3(\boldsymbol{w},\phi), \widehat{\mathbf{S}}(\mathbf{S}_3(\boldsymbol{w},\phi),\phi)\Big) \quad \forall\, (\boldsymbol{w},\phi) \in \mathbf{H},$$

respectively. A first result concerning the solvability of the mixed formulation (5.4.8) is provided next.

Lemma 5.4.1. *Assume that* $\kappa_4 \in \left(0, \frac{2\widetilde{K}_0\delta_4}{\widetilde{K}_1}\right)$, *with* $\delta_4 \in \left(0, \frac{2}{\widetilde{K}_1}\right)$, *and* $\kappa_5, \kappa_6 > 0$. *Then, there exists* $\widetilde{r}_0 > 0$ *such that for each* $\widetilde{r} \in (0,\widetilde{r}_0)$, *problem* (5.4.8) *has a unique solution* $(\boldsymbol{\Theta}, \widehat{\mathbf{S}}(\boldsymbol{w},\phi)) := (\boldsymbol{\Theta},\theta) \in \mathbf{H}(\mathrm{div};\Omega) \times \mathrm{H}^1(\Omega)$ *for each* $(\boldsymbol{w},\phi) \in \mathbf{H}$ *with* $\|\boldsymbol{w}\|_{1,\Omega} \leq \widetilde{r}$. *Moreover, there exists* $k_{\mathbf{S}} > 0$, *independent of* $(\boldsymbol{w},\phi)$, *such that*

$$\|\widehat{\mathbf{S}}(\boldsymbol{w},\phi)\| \;=\; \|\theta\|_{1,\Omega} \;\leq\; \|(\boldsymbol{\Theta},\theta)\| \leq k_{\mathbf{S}} \left\{ \|\boldsymbol{w}\|_{0,\Omega} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \right\} \quad \forall\, (\boldsymbol{w},\phi) \in \mathbf{H}. \tag{5.4.9}$$

Proof. From [59, Lemma 3.3] we recall that the bilinear form $\widetilde{\boldsymbol{a}} + \widetilde{\mathbf{b}}_{\boldsymbol{w}}$ (cf. (5.4.5a), (5.4.5b)) is elliptic with constant $\frac{\widetilde{\alpha}_1(\Omega)}{2}$, provided $\|\boldsymbol{w}\|_{1,\Omega} \leq \widetilde{r}_0$, with

$$\widetilde{r}_0 := \frac{\widetilde{\alpha}_1(\Omega)}{2\,\|i_c\|^2\,(\Omega)(1+\kappa_4^2)^{1/2}\widetilde{K}_1}. \tag{5.4.10}$$

Now, from (5.4.6a) and (5.4.6b) we note that the functionals $\widetilde{F}_{\boldsymbol{w},\phi}$ and $\widetilde{F}_{\mathrm{D}}$ are bounded with

$$\|\widetilde{F}_{\boldsymbol{w},\phi}\| \leq \widetilde{K}_1 s_2 (1+\kappa_4^2)^{1/2} \|\boldsymbol{w}\|_{0,\Omega} \quad \text{and} \quad \|\widetilde{F}_{\mathrm{D}}\| \leq \kappa_6 c_0(\Omega) \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma},$$

where $c_0(\Omega)$ is the norm of the trace operator in $\mathrm{H}^1(\Omega)$. Finally, a direct application of the Lax-Milgram lemma proves that for each $(\boldsymbol{w},\phi) \in \mathbf{H}$, problem (5.4.8) has a unique solution $(\boldsymbol{\Theta},\theta) \in \mathbf{H}(\mathrm{div};\Omega) \times \mathrm{H}^1(\Omega)$. Moreover, the continuous dependence result establishes that

$$\|\widehat{\mathbf{S}}(\boldsymbol{w},\phi)\| \leq \|(\boldsymbol{\Theta},\theta)\| \leq \frac{2}{\widetilde{\alpha}_1}\|\widetilde{F}_{\boldsymbol{w},\phi} + \widetilde{F}_{\mathrm{D}}\| \leq k_{\mathbf{S}} \left\{ \|\boldsymbol{w}\|_{0,\Omega} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \right\},$$

where $k_{\mathbf{S}} := \frac{2}{\widetilde{\alpha}_1} \max\{\widetilde{K}_1 s_2 (1+\kappa_4^2)^{1/2}, \kappa_6 c_0(\Omega), 1\}$, which ends the proof. □

The analogue of Lemma 5.3.3 is stated next.

Lemma 5.4.2. *Given* $r \in (0, \min\{r_0, \widetilde{r}_0\})$, *with* r_0 *and* $\widetilde{r}_0$ *given by* (5.3.20) *and* (5.4.10), *respectively, we let* $\widehat{W} := \{(\boldsymbol{w},\phi) \in \mathbf{H} : \|(\boldsymbol{w},\phi)\| \leq r\}$, *and assume that*

$$c(r) \left\{ C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} \right\} + k_{\mathbf{S}} \left\{ \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} \right\} \leq r, \tag{5.4.11}$$

where $c(r) := (1 + k_{\mathbf{S}})\, c_{\mathbf{S}} \max\{1, r\}$, *and* $c_{\mathbf{S}}$ *and* $k_{\mathbf{S}}$ *are the constants specified in Lemmas* 5.3.1 *and* 5.4.1, *respectively. Then* $\widehat{\mathbf{T}}(\widehat{W}) \subseteq \widehat{W}$.

Proof. It follows exactly as in [59, Lemma 3.5]. □

Next, we aim to prove the continuity of $\widehat{\mathbf{T}}$, which basically will be direct consequence of Lemma 5.3.4 and the following result providing the continuity of $\mathbf{S}$ and $\widehat{\mathbf{S}}$, respectively.

Lemma 5.4.3. *There exists $\widetilde{K}_{\widetilde{\mathbf{S}}} > 0$, such that for all $(\boldsymbol{w}_1, \phi_1)$, $(\boldsymbol{w}_2, \phi_2) \in \mathbf{H}$, there holds*

$$\begin{aligned} &\|\widehat{\mathbf{S}}(\boldsymbol{w}_1, \phi_1) - \widehat{\mathbf{S}}(\boldsymbol{w}_2, \phi_2)\| \\ &\leq \widetilde{K}_{\widetilde{\mathbf{S}}} \Big\{ \|\widehat{\mathbf{S}}(\boldsymbol{w}_2, \phi_2)\|_{1,\Omega} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{1,\Omega} + \|\boldsymbol{w}_2\|_{1,\Omega} \|\phi_1 - \phi_2\|_{0,\Omega} + \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{1,\Omega} \Big\}. \end{aligned} \tag{5.4.12}$$

Proof. Given $r \in (0, \widetilde{r}_0)$, and $(\boldsymbol{w}_1, \phi_1)$, $(\boldsymbol{w}_2, \phi_2) \in \mathbf{H}$ with $\|\boldsymbol{w}_1\|_{1,\Omega}, \|\boldsymbol{w}_2\|_{1,\Omega} \leq r$, we let $(\boldsymbol{\Theta}_1, \theta_1)$, $(\boldsymbol{\Theta}_2, \theta_2) \in \mathbf{H}(\mathrm{div}; \Omega) \times \mathrm{H}^1(\Omega)$ be solutions to (5.4.8) corresponding to $(\boldsymbol{w}_1, \phi_1)$ and $(\boldsymbol{w}_2, \phi_2)$, respectively, that is

$$\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta}_1, \theta_1), (\boldsymbol{\Phi}, \psi)) + \widetilde{\mathbf{b}}_{\boldsymbol{w}_1}((\boldsymbol{\Theta}_1, \theta_1), (\boldsymbol{\Phi}, \psi)) = \widetilde{F}_{\boldsymbol{w}_1, \phi_1}(\boldsymbol{\Phi}, \psi) + \widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi}, \psi),$$

and

$$\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta}_2, \theta_2), (\boldsymbol{\Phi}, \psi)) + \widetilde{\mathbf{b}}_{\boldsymbol{w}_2}((\boldsymbol{\Theta}_2, \theta_2), (\boldsymbol{\Phi}, \psi)) = \widetilde{F}_{\boldsymbol{w}_2, \phi_2}(\boldsymbol{\Phi}, \psi) + \widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi}, \psi),$$

for all $(\boldsymbol{\Phi}, \psi) \in \mathbf{H}(\mathrm{div}; \Omega) \times \mathrm{H}^1(\Omega)$. Then, similarly to Lemma 5.3.4, we add and subtract suitable terms to get

$$\begin{aligned} &(\widetilde{\boldsymbol{a}} + \widetilde{\mathbf{b}}_{\boldsymbol{w}_2})((\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2), (\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)) \\ &\qquad = -\widetilde{\mathbf{b}}_{\boldsymbol{w}_1 - \boldsymbol{w}_2}((\boldsymbol{\Theta}_1, \theta_1), (\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)) + (\widetilde{F}_{\boldsymbol{w}_1, \phi_1} - \widetilde{F}_{\boldsymbol{w}_2, \phi_2})((\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)), \end{aligned}$$

from which, applying the ellipticity of $\widetilde{\boldsymbol{a}} + \widetilde{\mathbf{b}}_{\boldsymbol{w}_2}$, we deduce that

$$\begin{aligned} &\frac{\widetilde{\alpha}_1}{2} \|(\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)\|^2 \\ &\qquad \leq -\widetilde{\mathbf{b}}_{\boldsymbol{w}_1 - \boldsymbol{w}_2}((\boldsymbol{\Theta}_1, \theta_1), (\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)) + (\widetilde{F}_{\boldsymbol{w}_1, \phi_1} - \widetilde{F}_{\boldsymbol{w}_2, \phi_2})((\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)) \\ &\qquad \leq \widetilde{K}_1 \Big\{ (1 + \kappa_4^2)^{1/2} \|i_c\|^2 \|\theta_1\|_{1,\Omega} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + L_s (1 + \kappa_4^2)^{1/2} \|\boldsymbol{w}_2\|_{1,\Omega} \|\phi_1 - \phi_2\|_{1,\Omega} \\ &\qquad + s_2 \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_{0,\Omega} \Big\} \|(\boldsymbol{\Theta}_1, \theta_1) - (\boldsymbol{\Theta}_2, \theta_2)\|. \end{aligned}$$

The foregoing inequality yields (5.4.12) with $\widetilde{K}_{\widetilde{\mathbf{S}}} := \frac{2\widetilde{K}_1}{\widetilde{\alpha}_1} \max\{(1+\kappa_4^2)^{1/2} \|i_c\|^2, L_s(1+\kappa_4^2)^{1/2}, s_2\}$, which finishes the proof. □

We are now in a position to establish the announced property of the operator $\widehat{\mathbf{T}}$. We omit the corresponding proof and refer to [59, Lemma 3.8] for details.

Lemma 5.4.4. *Given $r \in (0, \min\{r_0, \widetilde{r}_0\})$, with r_0 and $\widetilde{r}_0$ given by (5.3.20) and (5.4.10), respectively, we let $\widehat{W}$ as in Lemma 5.4.2. Then, there exists a constant $K_{\mathbf{T}} > 0$ such that for all $(\boldsymbol{w}_1, \phi_1), (\boldsymbol{w}_2, \phi_2) \in \widehat{W}$, there holds*

$$\begin{aligned} &\|\widehat{\mathbf{T}}(\boldsymbol{w}_1, \phi_1) - \widehat{\mathbf{T}}(\boldsymbol{w}_2, \phi_2)\| \\ &\qquad \leq K_{\mathbf{T}} \Big\{ C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma} + L_f \Big\} \|(\boldsymbol{w}_1, \phi_1) - (\boldsymbol{w}_2, \phi_2)\|. \end{aligned} \tag{5.4.13}$$

The existence and uniqueness of a fixed point of $\widehat{\mathbf{T}}$ (and therefore well-posedness of (5.4.7)), is stated as follows.

Theorem 5.4.5. *Suppose that the parameters κ_4, κ_5 and κ_6 satisfy the conditions required by Lemma (5.4.1). In addition, let r and $\widehat{W}$ as in Lemma 5.4.2, and assume that the data verify (5.4.11) and*

$$K_{\mathbf{T}}\Big\{C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma} + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Gamma} + L_f\Big\} < 1. \tag{5.4.14}$$

Then (5.4.7) has a unique solution $(\vec{\mathbf{t}}, (\boldsymbol{\Theta}, \theta)) \in H \times \mathbf{H}(\mathrm{div};\Omega) \times \mathrm{H}^1(\Omega)$ with $(\boldsymbol{u}, \theta) \in \widehat{W}$. Moreover

$$\|\vec{\mathbf{t}}\| \leq c_{\mathbf{S}}\Big\{C_f\, r + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}\Big\},$$

and

$$\|(\boldsymbol{\Theta}, \theta)\| \leq k_{\mathbf{S}}\{\|\boldsymbol{u}\|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\}.$$

Proof. It suffices to apply the Banach fixed-point Theorem (bearing in mind (5.4.13) - (5.4.14)), and then employ the a priori estimates (5.3.15) and (5.4.9). We omit further details. □

5.4.3 The Galerkin scheme

Similarly to Section 5.3.3, we begin by considering the arbitrary finite dimensional subspaces

$$\mathbb{H}_h^{\mathbf{t}} \subseteq \mathbb{L}^2_{\mathbf{tr}}(\Omega), \quad \mathbb{H}_h^{\boldsymbol{\sigma}} \subseteq \mathbb{H}_0(\mathbf{div};\Omega), \quad \mathbf{H}_h^{\boldsymbol{u}} \subseteq \mathbf{H}^1(\Omega), \quad \mathbf{H}_h^{\boldsymbol{\Theta}} \subseteq \mathbf{H}(\mathrm{div};\Omega), \quad \text{and} \quad \mathrm{H}_h^{\theta} \subseteq \mathrm{H}^1(\Omega). \tag{5.4.15}$$

A Galerkin scheme for (5.4.7) then reads: Find $(\vec{\mathbf{t}}_h, (\boldsymbol{\Theta}_h, \theta_h)) \in H_h \times \mathbf{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}$ such that

$$\begin{aligned} \mathbf{A}_{\theta_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) + \mathbf{B}_{\boldsymbol{u}_h}(\vec{\mathbf{t}}_h, \vec{\mathbf{s}}_h) &= F_{\theta_h}(\vec{\mathbf{s}}_h) + F_{\mathrm{D}}(\vec{\mathbf{s}}_h), \\ \widetilde{\boldsymbol{a}}((\boldsymbol{\Theta}_h, \theta_h), (\boldsymbol{\Phi}_h, \psi_h)) + \widetilde{\mathbf{b}}_{\boldsymbol{u}_h}((\boldsymbol{\Theta}_h, \theta_h), (\boldsymbol{\Phi}_h, \psi_h)) &= \widetilde{F}_{\boldsymbol{u}_h,\theta_h}(\boldsymbol{\Phi}_h, \psi_h) + \widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi}_h, \psi_h), \end{aligned} \tag{5.4.16}$$

for all $(\vec{\mathbf{s}}_h, (\boldsymbol{\Phi}_h, \psi_h)) \in H_h \times \mathbf{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}$. We emphasize that the analysis of (5.4.16) uses the discrete version of the fixed-point strategy from Section 5.4.2. Results and the used arguments are almost verbatim to those in that section, and we omit them here simply stating the main result.

Theorem 5.4.6. *Suppose that the parameters κ_4, κ_5 and κ_6 satisfy the conditions required by Lemma 5.4.1. In addition, let $\widehat{W}_h := \big\{(\boldsymbol{w}_h, \phi_h) \in \mathbf{H}_h^{\boldsymbol{u}} \times \mathrm{H}_h^{\theta} : \|(\boldsymbol{w}_h, \phi_h)\| \leq r\big\}$, with r defined as in Lemma 5.4.2, and assume that the data satisfy (5.4.11). Then, the problem (5.4.16) has at least one solution $(\vec{\mathbf{t}}_h, (\boldsymbol{\Theta}_h, \theta_h)) \in H_h \times \mathbf{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}$, with $(\boldsymbol{u}_h, \theta_h) \in \widehat{W}_h$, and there holds*

$$\|\vec{\mathbf{t}}_h\| \leq c_{\mathbf{S}}\Big\{C_f\, r + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Gamma}\Big\},$$

and

$$\|(\boldsymbol{\Theta}_h, \theta_h)\| \leq k_{\mathbf{S}}\{\|\boldsymbol{u}_h\|_{1,\Omega} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma}\}.$$

5.4.4 A priori error analysis

Let $(\boldsymbol{\Theta},\theta)$ and $(\boldsymbol{\Theta}_h,\theta_h)$ be solutions to the problems

$$\begin{aligned}\widetilde{\boldsymbol{a}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi))+\widetilde{\mathbf{b}}_{\boldsymbol{u}}((\boldsymbol{\Theta},\theta),(\boldsymbol{\Phi},\psi))&=\widetilde{F}_{\boldsymbol{u},\theta}(\boldsymbol{\Phi},\psi)+\widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi},\psi)\quad\text{and}\\ \widetilde{\boldsymbol{a}}((\boldsymbol{\Theta}_h,\theta_h),(\boldsymbol{\Phi}_h,\psi_h))+\widetilde{\mathbf{b}}_{\boldsymbol{u}_h}((\boldsymbol{\Theta}_h,\theta_h),(\boldsymbol{\Phi}_h,\psi_h))&=\widetilde{F}_{\boldsymbol{u}_h,\theta_h}(\boldsymbol{\Phi}_h,\psi_h)+\widetilde{F}_{\mathrm{D}}(\boldsymbol{\Phi}_h,\psi_h)\,,\end{aligned}\tag{5.4.17}$$

for all $(\boldsymbol{\Phi},\psi)\in\mathbf{H}(\mathrm{div};\Omega)\times\mathrm{H}^1(\Omega)$, and for all $(\boldsymbol{\Phi}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}$, respectively. A preliminary error estimate is provided by the following lemma.

Lemma 5.4.7. *There exists a positive constant $K_{\mathbf{ST}}$, independent of h, such that*

$$\begin{aligned}\|(\boldsymbol{\Theta},\theta)-(\boldsymbol{\Theta}_h,\theta_h)\|\;\leq\;K_{\mathbf{ST}}\Big\{&\Big(1+\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\Big)\,\mathrm{dist}\,\Big((\boldsymbol{\Theta},\theta),\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\Big)\\&+\;\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\,\|\theta\|_{1,\Omega}\;+\;\|\boldsymbol{u}_h\|_{1,\Omega}\,\|\theta-\theta_h\|_{1,\Omega}+s_2\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{0,\Omega}\Big\}\,.\end{aligned}\tag{5.4.18}$$

Proof. Proceeding as in the proof of Lemma 5.3.15, a straightforward application of the Strang lemma provided in [127, Thm. 11.1] to the context (5.4.17), yields

$$\begin{aligned}\|(\boldsymbol{\Theta},\theta)-(\boldsymbol{\Theta}_h,\theta_h)_h\|\leq\overline{K}_1\Bigg\{&\sup_{\substack{(\boldsymbol{\Phi}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\\(\boldsymbol{\Phi}_h,\psi_h)\neq0}}\frac{|\widetilde{F}_{\boldsymbol{u},\theta}(\boldsymbol{\Phi}_h,\psi_h)-\widetilde{F}_{\boldsymbol{u}_h,\theta_h}(\boldsymbol{\Phi}_h,\psi_h)|}{\|(\boldsymbol{\Phi}_h,\psi_h)\|}\\&+\inf_{\substack{(\boldsymbol{\Psi}_h,\phi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\\(\boldsymbol{\Psi}_h,\phi_h)\neq0}}\left(\|(\boldsymbol{\Theta},\theta)-(\boldsymbol{\Psi}_h,\phi_h)\|+\sup_{\substack{(\boldsymbol{\Phi}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\\(\boldsymbol{\Phi}_h,\psi_h)\neq0}}\frac{|\widetilde{\mathbf{b}}_{\boldsymbol{u}-\boldsymbol{u}_h}((\boldsymbol{\Psi}_h,\phi_h),(\boldsymbol{\Phi}_h,\psi_h))|}{\|(\boldsymbol{\Phi}_h,\psi_h)\|}\right)\Bigg\}\,,\end{aligned}\tag{5.4.19}$$

where $\overline{K}_1:=\frac{2}{\widetilde{\alpha}_1(\Omega)}\max\{1,\|\widetilde{\boldsymbol{a}}+\widetilde{\mathbf{b}}_{\boldsymbol{u}}\|\}$. Thus, employing [59, Lemma 5.3], we have

$$\begin{aligned}\sup_{\substack{(\boldsymbol{\Phi}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\\(\boldsymbol{\Phi}_h,\psi_h)\neq0}}\frac{|\widetilde{\mathbf{b}}_{\boldsymbol{u}-\boldsymbol{u}_h}((\boldsymbol{\Psi}_h,\phi_h),(\boldsymbol{\Phi}_h,\psi_h))|}{\|(\boldsymbol{\Phi}_h,\psi_h)\|}\;\leq\;&\|i_c\|^2\,(1+\kappa_4^2)^{1/2}\widetilde{K}_1\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\,\|\theta\|_{1,\Omega}\\&+\;\|i_c\|^2\,(1+\kappa_4^2)^{1/2}\widetilde{K}_1\,\|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\,\|(\boldsymbol{\Theta},\theta)-(\boldsymbol{\Psi}_h,\phi_h)\|\,,\end{aligned}\tag{5.4.20}$$

and similarly as in Lemma 5.4.3 we get

$$\begin{aligned}&\sup_{\substack{(\boldsymbol{\Phi}_h,\psi_h)\in\mathbf{H}_h^{\boldsymbol{\Theta}}\times\mathrm{H}_h^{\theta}\\(\boldsymbol{\Phi}_h,\psi_h)\neq0}}\frac{|\widetilde{F}_{\boldsymbol{u},\theta}(\boldsymbol{\Phi}_h,\psi_h)-\widetilde{F}_{\boldsymbol{u}_h,\theta_h}(\boldsymbol{\Phi}_h,\psi_h)|}{\|(\boldsymbol{\Phi}_h,\psi_h)\|}\\&\qquad\leq\widetilde{K}_1(1+\kappa_4^2)^{1/2}L_s\,\|\boldsymbol{u}_h\|_{1,\Omega}\,\|\theta_h-\theta\|_{1,\Omega}+\widetilde{K}_1 s_2\,\|\boldsymbol{u}_h-\boldsymbol{u}\|_{0,\Omega}\,.\end{aligned}\tag{5.4.21}$$

Therefore, (5.4.18) follows by replacing (5.4.20) and (5.4.21) back into (5.4.19), with a constant $K_{\mathbf{ST}}$ depending on $\widetilde{\alpha}_1$, $\|\widetilde{\boldsymbol{a}}+\widetilde{\mathbf{b}}_{\boldsymbol{u}}\|$, $\widetilde{K}_1$, $\|i_c\|$, κ_4, and L_s. □

In much the same way as in Section 5.3.3, denoting

$$C_9 = K_{\mathbf{ST}} k_{\mathbf{S}} c_{\mathbf{S}}, \quad C_{10} := C_{\mathbf{ST}} c_{\mathbf{S}} + K_{\mathbf{ST}} c_{\mathbf{S}} + C_{\mathbf{ST}} c_{\mathbf{S}},$$

$$C_{11} := \max\{C_{\mathbf{ST}}, C_5, (C_5 + C_{10} + C_9 r)r, C_9 + C_{10}, K_{\mathbf{ST}} k_{\mathbf{S}}, K_{\mathbf{ST}}\},$$

where C_5 is the constant defined in (5.3.51), and applying the estimates given in Lemmas 5.3.15 and 5.4.7, we can prove that

$$\begin{aligned}\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| + \|(\boldsymbol{\Theta}, \theta) - (\boldsymbol{\Theta}_h, \theta_h)\| &\leq C_{\mathbf{ST}} \operatorname{dist}(\vec{\mathbf{t}}, H_h) + K_{\mathbf{ST}} \left(1 + \|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}\right) \operatorname{dist}\left((\boldsymbol{\Theta}, \theta), \mathrm{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}\right) \\ &\quad + C_{11}\Big(L_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Omega} + C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + s_2\Big) \\ &\quad \times \Big\{\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| + \|(\boldsymbol{\Theta}, \theta) - (\boldsymbol{\Theta}_h, \theta_h)\|\Big\}.\end{aligned}$$

We stress here that the constants multiplying $\operatorname{dist}(\vec{\mathbf{t}}, H_h)$ and $\operatorname{dist}\left((\boldsymbol{\Theta}, \theta), \mathrm{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}\right)$ are both controlled by constants, parameters, and data only since $\|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega}$ can be controlled by (5.3.15) and (5.3.41). Consequently, we can establish the following main result.

Theorem 5.4.8. *Assume that the data satisfy*

$$C_7\Big\{L_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2+\varepsilon,\Omega} + C_f + \|\boldsymbol{u}_{\mathrm{D}}\|_{1/2,\Omega} + \|\theta_{\mathrm{D}}\|_{1/2,\Gamma} + \|\theta_{\mathrm{D}}\|_{0,\Gamma} + s_2\Big\} < \frac{1}{2}.$$

Then, there exists a positive constant C_{12}, independent of h, such that

$$\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| + \|(\boldsymbol{\Theta}, \theta) - (\boldsymbol{\Theta}_h, \theta_h)\| \leq C_{12}\Big\{\operatorname{dist}(\vec{\mathbf{t}}, H_h) + \operatorname{dist}\left((\boldsymbol{\Theta}, \theta), \mathrm{H}_h^{\boldsymbol{\Theta}} \times \mathrm{H}_h^{\theta}\right)\Big\}. \tag{5.4.22}$$

5.4.5 Specific finite element subspaces

Here we consider the global Raviart-Thomas space of order k to approximate $\boldsymbol{\Theta}$, and the Lagrange space of degree $\leq k+1$ for the temperature θ, that is

$$\begin{aligned}\mathbf{H}_h^{\boldsymbol{\Theta}} &:= \left\{\boldsymbol{\Phi}_h \in \mathbf{H}(\operatorname{div}; \Omega) : \quad \boldsymbol{c}^{\mathrm{t}} \boldsymbol{\Phi}_h|_K \in \mathbf{RT}_k(K), \quad \forall\, \boldsymbol{c} \in \mathrm{R}^n \quad \forall\, K \in \mathcal{T}_h\right\}, \\ \mathrm{H}_h^{\theta} &:= \left\{\psi_h \in \mathrm{C}(\overline{\Omega}) : \quad \psi_h|_K \in \mathrm{P}_{k+1}(K) \quad \forall\, K \in \mathcal{T}_h\right\}.\end{aligned} \tag{5.4.23}$$

The approximation properties of the spaces in (5.3.56) and (5.4.23) (that can be found in e.g [10, 39, 81]) are then combined with the Céa estimate (5.4.22) to produce the theoretical rate of convergence of (5.4.16), summarised as follows.

Theorem 5.4.9. *Appart from the hypotheses of Theorems 5.4.5, 5.4.6 and 5.4.8, assume that there exists $s > 0$ such that $\mathbf{t} \in \mathbb{H}^s(\Omega)$, $\boldsymbol{\sigma} \in \mathbb{H}^s(\Omega)$, $\mathbf{div}\, \boldsymbol{\sigma} \in \mathbf{H}^s(\Omega)$, $\boldsymbol{u} \in \mathbf{H}^{1+s}(\Omega)$, $\boldsymbol{\Theta} \in \mathbf{H}^s(\Omega)$, $\operatorname{div} \boldsymbol{\Theta} \in \mathrm{H}^s(\Omega)$, and $\theta \in \mathrm{H}^{1+s}(\Omega)$. Then there exists $C > 0$ independent of h, such that with (5.3.56) and (5.4.23), one has*

$$\begin{aligned}\|\vec{\mathbf{t}} - \vec{\mathbf{t}}_h\| + \|(\boldsymbol{\Theta}, \theta) - (\boldsymbol{\Theta}_h, \theta_h)\| \leq\ & C h^{\min\{s, k+1\}} \Big\{\|\mathbf{t}\|_{s,\Omega} + \|\boldsymbol{\sigma}\|_{s,\Omega} + \|\mathbf{div}\, \boldsymbol{\sigma}\|_{s,\Omega} \\ &+ \|\boldsymbol{u}\|_{1+s,\Omega} + \|\boldsymbol{\Theta}\|_{s,\Omega} + \|\operatorname{div} \boldsymbol{\Theta}\|_{s,\Omega} + \|\theta\|_{1+s,\Omega}\Big\}.\end{aligned} \tag{5.4.24}$$

5.5 Numerical tests

We now present a set of computational tests. For the mixed-primal scheme (5.3.35) we consider an example that shows the convergence rates anticipated by Theorem 5.3.18, and a second test that addresses the application of our method to the three-dimensional modelling of gallium melting in a cuboid cavity. We will also present two examples that illustrate the performance of the fully-mixed scheme (5.4.16), and that will serve as conformation for the rates of convergence provided by Theorem 5.4.9.

5.5.1 Preliminary notations

A Picard algorithm with tolerance of $1E-6$ on the ℓ^2-norm of the residual has been employed for our fixed-point problems. The convergence of the approximate solutions is assessed by computing errors in the respective norms and experimental rates, that we define as usual

$$\begin{aligned}
&\mathrm{e}(\mathbf{t}) = \|\mathbf{t}-\mathbf{t}_h\|_{0,\Omega}\,, \quad \mathrm{e}(\boldsymbol{u}) = \|\boldsymbol{u}-\boldsymbol{u}_h\|_{1,\Omega}\,, \quad \mathrm{e}(p) = \|p-p_h\|_{0,\Omega}\,, \quad \mathrm{e}(\theta) = \|\theta-\theta_h\|_{1,\Omega}\,,\\
&\mathrm{e}(\lambda) = \big\|\lambda-\lambda_{\widetilde{h}}\big\|_{0,\Gamma}\,, \quad \mathrm{e}(\boldsymbol{\sigma}) = \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathbf{div};\Omega}\,, \quad \mathrm{e}(\boldsymbol{\Theta}) = \|\boldsymbol{\Theta}-\boldsymbol{\Theta}_h\|_{\mathbf{div};\Omega} \quad \widehat{\mathrm{e}}(\boldsymbol{\sigma}) = \big\|\boldsymbol{\sigma}_h-\boldsymbol{\sigma}_h^{\mathrm{t}}\big\|_{0,\Omega}\,,\\
&r(\lambda) = \frac{\log(\mathrm{e}(\lambda)/\mathrm{e}'(\lambda))}{\log(\widetilde{h}/\widetilde{h}')}, \quad r(\%) = \frac{\log(\mathrm{e}(\%)/\mathrm{e}'(\%))}{\log(h/h')},
\end{aligned}$$

with $\% \in \{\mathbf{t},\boldsymbol{\sigma},\boldsymbol{u},p,\boldsymbol{\Theta},\theta\}$, and where e,e' denote errors computed on two consecutive meshes of sizes h,h' ($\widetilde{h}$ and $\widetilde{h}'$ for λ), respectively. The trace condition on the stress is enforced through a penalisation strategy. Furthermore, for the Examples 5.2.1, 5.3.1 and 5.3.2 described below, we remark that the Navier-Stokes-Brinkman and heat equations are considered non-homogeneous and the extra source terms are chosen according to the given exact solutions. This treatment does not compromise the continuous and discrete analysis, as the regularity of the exact solution provides sufficiently smooth right-hand sides, thus only requiring a slight modification of the functionals in the variational formulation.

5.5.2 Tests for the mixed-primal scheme

Example 5.2.1. In our first numerical test, we consider problem (5.2.1) defined in the unit square $\Omega = (0,1)^2$ and choose the following manufactured exact solutions, viscosity, porosity, enthalpy, buoyancy and thermal conductivity:

$$\begin{aligned}
&\boldsymbol{u} = \begin{pmatrix} \sin(\pi x)\cos(\pi y) \\ -\sin(\pi y)\cos(\pi x) \end{pmatrix}, \quad \theta = 1+\sin(\pi x)\cos(\pi y), \quad p = x^2-y^2, \quad \mathbf{t} = \mathbf{e}(\boldsymbol{u}),\\
&\boldsymbol{\sigma} = \alpha\,\mu(\theta)\mathbf{t} - (\boldsymbol{u}\otimes\boldsymbol{u}) - p\,\mathbb{I}, \quad \lambda = -\rho\kappa\nabla\theta\cdot\boldsymbol{\nu}, \quad \mu(\theta) = \exp(-0.25\,\theta),\\
&\eta(\theta) = 2-\tanh(0.5-\theta), \quad s(\theta) = 1+\tanh(1-\theta), \quad f(\theta) = 0.01\frac{\mathrm{Ra}}{\mathrm{PrRe}^2}\theta, \quad \kappa = \mathbb{I}.
\end{aligned} \tag{5.5.1}$$

These closed-form solutions feature a divergence-free velocity that satisfies the compatibility condition (5.2.6) and it is used as a non-homogeneous Dirichlet datum on Γ. In turn, the exact temperature is uniformly bounded and it is also exploited as Dirichlet datum. Moreover, the nonlinear functions

Mixed-primal $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathrm{P}_1 - \mathrm{P}_0$ scheme							
DoFs	h	$\mathrm{e}(\mathbf{t})$	$\mathrm{e}(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$\mathrm{e}(p)$	$\mathrm{e}(\theta)$	$\mathrm{e}(\lambda)$
1114	0.1900	0.27795	0.81133	0.46689	0.08997	0.32250	1.04403
4138	0.0950	0.14164	0.39563	0.23877	0.04233	0.16783	0.50832
16088	0.0490	0.07030	0.19703	0.11721	0.02047	0.08227	0.25242
63531	0.0244	0.03513	0.09902	0.05920	0.01045	0.04157	0.12559
255319	0.0139	0.01751	0.04928	0.02931	0.00514	0.02057	0.06272
1010150	0.0077	0.00878	0.02435	0.01450	0.00249	0.01020	0.03135
$\widehat{\mathrm{e}}(\boldsymbol{\sigma})$	$\widetilde{h}$	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\theta)$	$r(\lambda)$
0.28394	0.5000	-	-	-	-	-	-
0.14441	0.2500	0.97260	1.03613	0.96744	1.08761	0.94224	1.03826
0.07225	0.1250	1.05788	1.05290	1.07467	1.09690	1.07679	1.00987
0.03542	0.0625	0.99623	0.98802	0.98078	0.96534	0.98014	1.00705
0.01767	0.0312	1.24455	1.24779	1.25722	1.26803	1.25808	1.00167
0.00901	0.0156	1.17723	1.20254	1.19973	1.23193	1.19550	1.00041
Mixed-primal $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathrm{P}_2 - \mathrm{P}_1$ scheme							
DoFs	h	$\mathrm{e}(\mathbf{t})$	$\mathrm{e}(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$\mathrm{e}(p)$	$\mathrm{e}(\theta)$	$\mathrm{e}(\lambda)$
3610	0.1900	0.02055	0.06020	0.03517	0.01120	0.02617	0.08327
13690	0.1025	0.00494	0.01494	0.00824	0.00324	0.00607	0.01984
53826	0.0492	0.00120	0.00365	0.00200	0.00078	0.00145	0.00476
213782	0.0256	0.00030	0.00092	0.00051	0.00020	0.00036	0.00116
861670	0.0139	0.00008	0.00023	0.00013	0.00006	0.00008	0.00028
$\widehat{\mathrm{e}}(\boldsymbol{\sigma})$	$\widetilde{h}$	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\theta)$	$r(\lambda)$
0.01935	0.5000	-	-	-	-	-	-
0.00427	0.0250	2.30696	2.25787	2.35075	2.01078	2.36693	2.06950
0.00108	0.1250	1.91309	1.90661	1.91476	1.91739	1.93990	2.05693
0.00027	0.0625	2.10199	2.12297	2.10057	2.09550	2.14255	2.03644
0.00006	0.0312	2.15117	2.26398	2.23191	1.95851	2.29949	2.01195

Table 5.1: Example 5.2.1. Convergence history for $k = 0, 1$ (table produced by the author).

satisfy (5.2.2)-(5.2.4). We consider $\boldsymbol{k} = (0, 1)^{\mathrm{t}}$ and the parameters given by: $\mathrm{Re} = 1, \mathrm{Pr} = 0.71, C = 1$ and $\mathrm{Ra} = 100$, where Ra is the Rayleigh number. The stabilisation parameters κ_1, κ_2 and κ_3 are taken as in (5.3.23), where the viscosity and porosity bounds are estimated as $\mu_1 = 0.6$, $\mu_2 = 1$ and $\eta_1 = 1$, $\eta_2 = 3$, respectively, thus resulting in $\kappa_1 = 0.6$, $\kappa_2 = 0.33$ and $\kappa_3 = 0.3$. An average of six Picard steps were required to reach the desired tolerance. Errors and corresponding rates associated with first and second order approximations are summarised in Table 5.1. The results show optimal asymptotic convergence rates for all fields, which are the expected ones according to Theorem 5.3.18. Also, Figure 5.2 shows that the rates of convergence for $\widehat{e}(\boldsymbol{\sigma})$ are the expected ones. Finally, samples of augmented mixed-primal approximations obtained with 1M DoFs are depicted in Figure 5.1.

Example 5.2.2. We continue with a simulation involving phase change in a cuboid cavity. The problem corresponds to the steady thermal convective flow occurring in the melting of gallium. Numerical results for the transient version of this problem, as well as detailed experimental considerations, can be found in e.g. [30, 157, 165]. We have adapted the model to comply with (5.2.1), using a

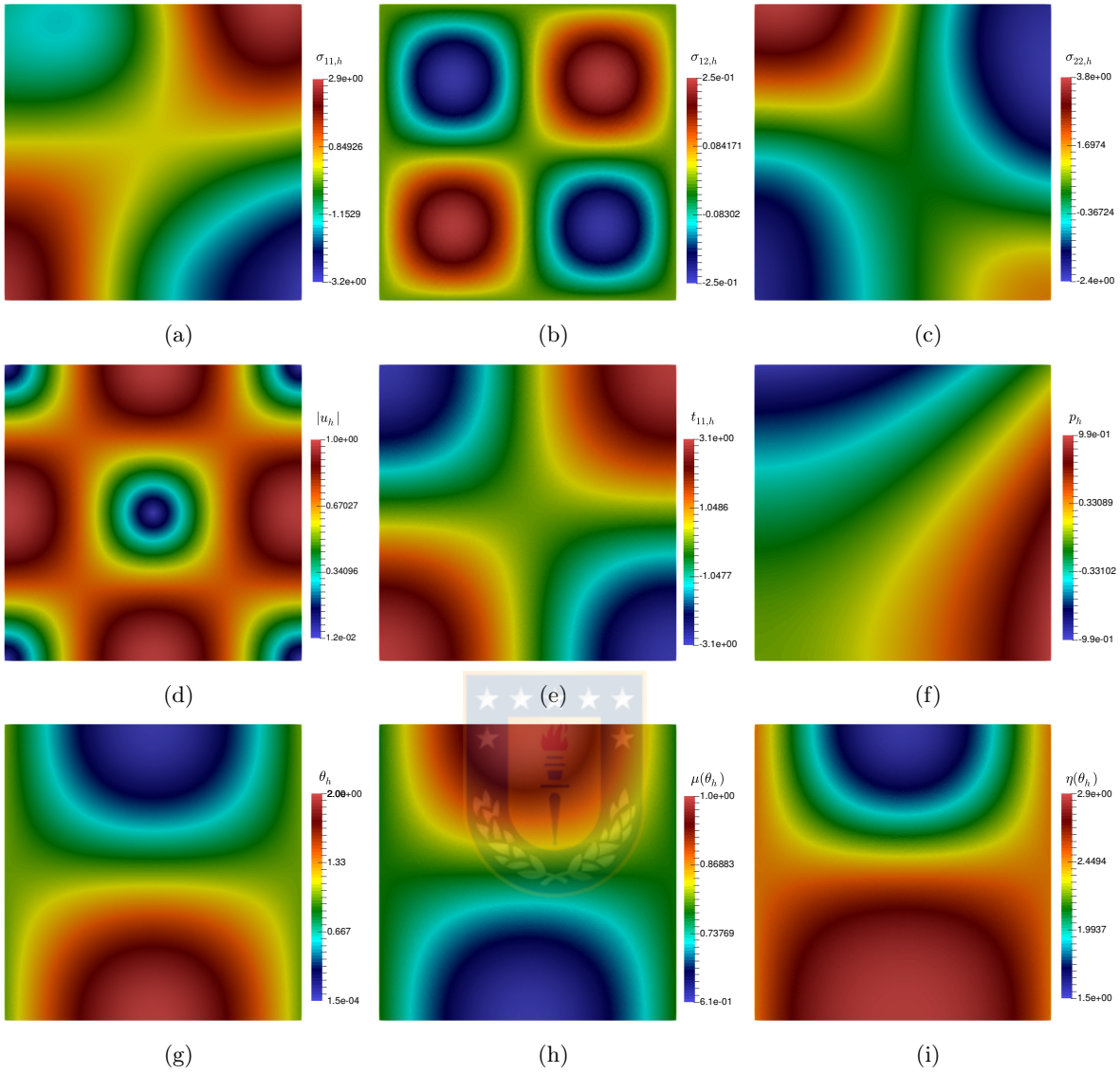


Figure 5.1: Example 5.2.1. Lowest-order approximate solutions: (a)-(c) pseudostress entries, (d) displacement magnitude, (e) strain rate, (f) postprocessed pressure, (g) temperature, (h) effective viscosity, and (i) effective porosity fields (figure produced by the author).

porosity-enthalpy framework (i.e., setting a constant viscosity), but employing mixed boundary conditions as prescribed below. The physical properties of the problem are defined by the model constants $\mathrm{Ra} = 2E5$, $\mathrm{Re} = 10$, $\mathrm{Pr} = 0.71$, $\mu = 1$, $\eta_1 = 1E-3$, $\eta_2 = 1E5$, $\theta_r = 0.01$, $r = 0.05$, $\boldsymbol{g} = (0,0,1)^{\mathtt{t}}$. The temperature-dependent enthalpy and porosity functions adopt the forms

$$s(\theta) = \frac{1}{2}\Big\{1 + \tanh(\frac{\theta_r - \theta}{r})\Big\}, \quad \eta(\theta) = \eta_1 + \eta_2\Big\{1 + \tanh(\frac{\theta_r - \theta}{r})\Big\}.$$

The computational domain is the box $\Omega = (0,2) \times (0,2) \times (0,1)$ and we generate a structured mesh composed by 255K tetrahedral elements and about 46K vertices. Considering the lowest-order mixed-primal finite element method (5.3.35), the assembled linear systems appearing at each Picard iteration consist of about 3M DoFs for the Navier-Stokes-Brinkman block and near 46K DoFs for the energy

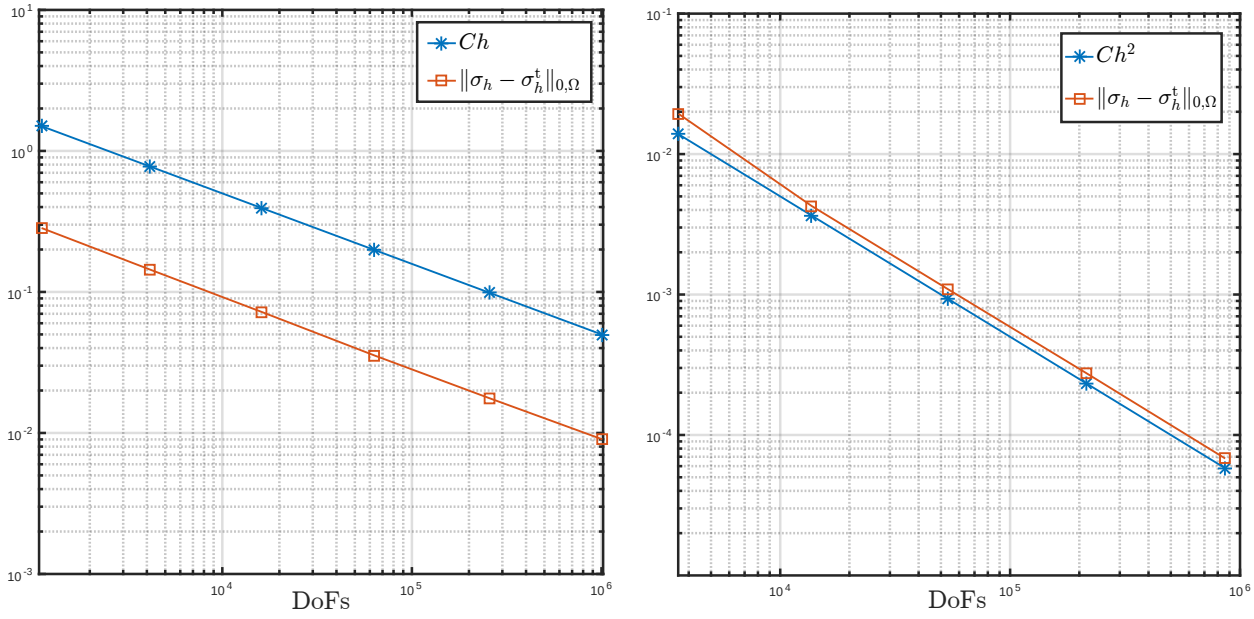


Figure 5.2: Example 5.2.1. Errors associated with the mixed-primal approximation *versus* DoFs for $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathrm{P}_1 - \mathrm{P}_0$ and $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathrm{P}_2 - \mathrm{P}_1$ finite elements (left and right, respectively) (figure produced by the author).

conservation equation. No-slip conditions were imposed for the velocity on the whole boundary. Moreover, the walls defined by $x = 0$ and $x = 2$ are maintained at fixed temperatures of $\theta = 1$ and $\theta = -0.01$, respectively; whereas on the remaining walls we impose zero-flux boundary conditions for the temperature. Such a setting implies in particular, that the Lagrange multiplier λ is not required in the formulation. Primary features of the flow can be observed in Figure 5.3. We do not expect to produce the flow separation vortices as those seen in [157] because our test focuses on the steady regime and we employ a enthalpy-porosity model. Nevertheless, we do see streamlines avoiding the solid region (on the right side of the gray wall), as well as a qualitative match with the temperature profiles observed in [30, 157], where thermal convection occurs mainly on the xy plane. Under the considered flow regime, 15 fixed-point iterations were needed to reach the desired residual tolerance of 1E-6.

5.5.3 Tests for the fully-mixed scheme

Example 5.3.1. In this example we consider the domain, exact solution, nonlinear functions, parameters and stabilisation parameters for the Navier-Stokes-Brinkman equation exactly as in Example 5.2.1 of Section 5.5.2 (cf. (5.5.1)). We recall that $\boldsymbol{\Theta} := \rho\kappa\nabla\theta - \theta\boldsymbol{u} - s(\theta)$ and for the values κ_4, κ_5 and κ_6, we follow [59, Section 6] to obtain $\kappa_4 = 0.99$, $\kappa_5 = 0.5$ and κ_6=0.49. Values and plots of errors and corresponding rates associated with first and second order approximations are summarised in Table 5.2 and Fig. 5.4. The results show optimal asymptotic convergence rates for all fields, which are the expected ones according to Theorem 5.4.9. We remark here that the errors reported in Tables 5.1 and 5.2 for the unknowns $\mathbf{t}$, $\boldsymbol{\sigma}$, and $\boldsymbol{u}$, are basically the same for the two methods considered

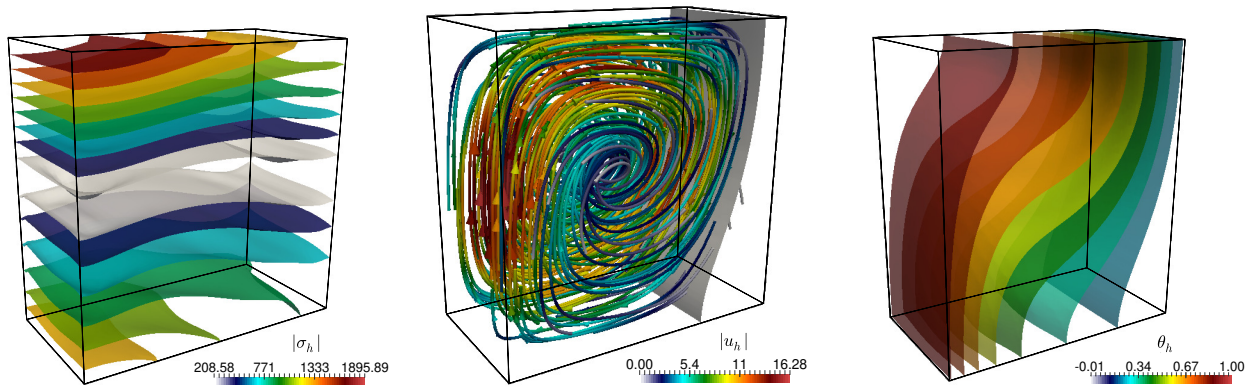


Figure 5.3: Example 5.2.2. Computed solutions with the lowest-order mixed-primal scheme. (a) pseudo-stress magnitude, (b) velocity magnitude, (c) temperature (figure produced by the author).

in this chapter, which is due to the fact that both formulations consider a mixed approach for the Navier-Stokes-Brinkman equation. However, since for the heat equation primal and mixed approaches are employed, which yields the two different coupled schemes that are proposed and analysed in this chapter, some very slight changes (even only after two or three decimals) can be observed in those tables for the rates of convergences of $\mathbf{t}$, $\boldsymbol{\sigma}$, $\boldsymbol{u}$, and p.

Example 5.3.2 In our second example, we produce the error and rate history associated with the finite element approximation for the three-dimensional case. Let us consider the following closed-form solutions to the model problem, defined on the unit cube domain $\Omega = (0,1)^3$:

$$\boldsymbol{u} = \begin{pmatrix} \cos(x)\sin(y)\sin(z) \\ \sin(x)\cos(y)\sin(z) \\ -2\sin(x)\sin(y)\cos(z) \end{pmatrix}, \quad \theta = 1 + \sin(\pi x)\cos(\pi y)\sin(\pi z), \quad p = x^2 - 2y^2 - z^2.$$

These functions are smooth and they are used to generate non-homogeneous forcing and source terms. Also, the manufactured velocity and temperature are used as Dirichlet datum on Γ. The porosity, enthalpy and thermal conductivity are taken as in Example 5.2.1, and the remaining nonlinear functions are defined as: $\mu(\theta) = \exp(-\theta)$, $f(\theta) = \theta$. All model constants assume the adimensional value 1. The stabilisation parameters are taken again as in Example 5.3.1. Part of the solution is shown in Figure 5.5, and a convergence history for a set of quasi-uniform refinements is shown in Table 5.3, confirming that this fully-mixed finite element method converges optimally with order $\mathcal{O}(h^{k+1})$.

Fully-mixed $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ scheme							
DoFs	h	$\mathrm{e}(\mathbf{t})$	$\mathrm{e}(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$\mathrm{e}(p)$	$\mathrm{e}(\boldsymbol{\Theta})$	$\mathrm{e}(\theta)$
1332	0.1900	0.27796	0.81134	0.46690	0.08977	1.75126	0.32186
5012	0.0950	0.14164	0.39564	0.23877	0.04228	0.86291	0.16783
19636	0.0490	0.07030	0.19703	0.11721	0.02047	0.43528	0.08226
77860	0.0244	0.03513	0.09902	0.05920	0.01045	0.21694	0.04158
313572	0.0139	0.01751	0.04928	0.02931	0.0051	0.10825	0.02057
1241924	0.0077	0.00878	0.02435	0.01450	0.00249	0.05320	0.01020

iter	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\boldsymbol{\Theta})$	$r(\theta)$
6	-	-	-	-	-	-
6	0.97261	1.03610	0.96744	1.08623	1.02110	0.93942
6	1.05793	1.05293	1.07468	1.09557	1.03359	1.07691
6	0.99624	0.98803	0.98078	0.96496	1.00003	0.97988
6	1.24456	1.24779	1.25722	1.26790	1.24230	1.25816
6	1.17723	1.20254	1.19973	1.23190	1.21162	1.19554

Fully-mixed $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2$ scheme							
DoFs	h	$\mathrm{e}(\mathbf{t})$	$\mathrm{e}(\boldsymbol{\sigma})$	$\mathrm{e}(\boldsymbol{u})$	$\mathrm{e}(p)$	$\mathrm{e}(\boldsymbol{\Theta})$	$\mathrm{e}(\theta)$
4354	0.1900	0.02055	0.06020	0.03517	0.01120	0.10995	0.02402
16642	0.1025	0.00494	0.01494	0.00824	0.00324	0.02824	0.00571
65734	0.0492	0.00120	0.00365	0.00200	0.00078	0.00694	0.00139
261712	0.0256	0.00030	0.00092	0.00051	0.00020	0.00174	0.00035
1056184	0.0139	0.00008	0.00023	0.00013	0.00006	0.00042	0.00008

iter	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\boldsymbol{\Theta})$	$r(\theta)$
6	-	-	-	-	-	-
6	2.30685	2.25790	2.35075	2.01040	2.20238	2.32777
6	1.91308	1.90661	1.91475	1.91735	1.90140	1.90685
6	2.10194	2.12295	2.10055	2.09544	2.12817	2.11582
6	2.15690	2.26665	2.23441	1.97252	2.31585	2.28408

Table 5.2: Example 5.3.1. Convergence history and Picard iteration count for $k = 0, 1$ (table produced by the author).

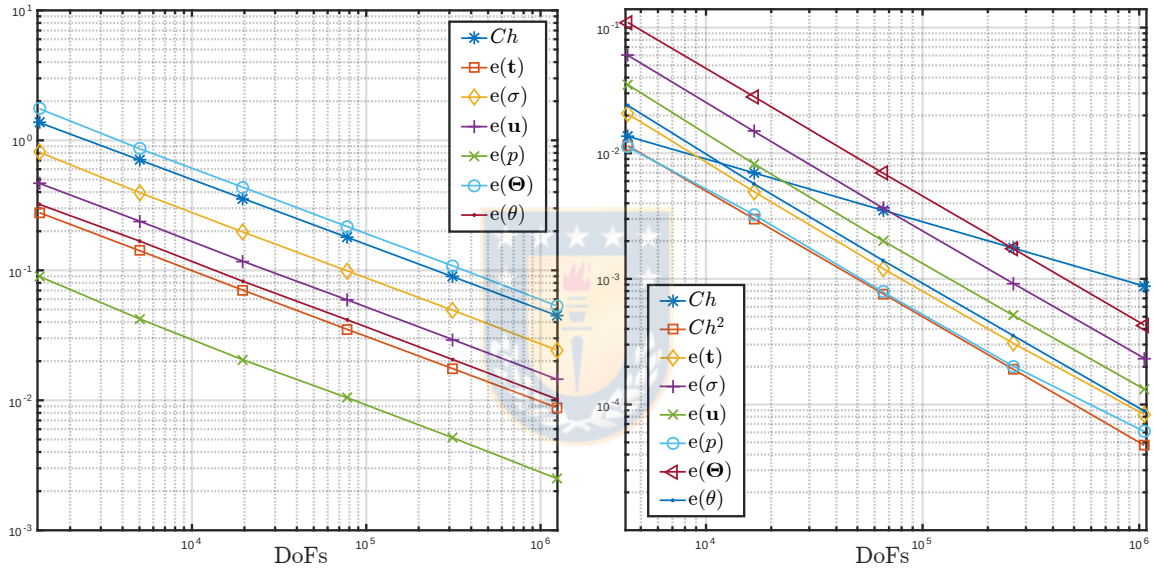


Figure 5.4: Example 5.3.1. Errors associated with the fully-mixed approximation *versus* DoFs for $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ and $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2$ finite elements (left and right, respectively) (figure produced by the author).

Fully-mixed $\mathbb{P}_0 - \mathbf{RT}_0 - \mathbf{P}_1 - \mathbf{RT}_0 - \mathrm{P}_1$ scheme							
DoFs	h	e($\mathbf{t}$)	e($\boldsymbol{\sigma}$)	e($\boldsymbol{u}$)	e(p)	e($\boldsymbol{\Theta}$)	e(θ)
828	0.7071	0.38526	0.79824	0.54203	0.23637	5.37964	1.85821
5876	0.3535	0.24931	0.39517	0.27068	0.13072	2.88208	0.95736
44388	0.1767	0.13775	0.19102	0.12628	0.06088	1.46682	0.48293
345284	0.0883	0.07088	0.09400	0.06013	0.02935	0.73668	0.24289
2724228	0.0441	0.03572	0.04676	0.02947	0.01449	0.36875	0.12175

iter	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\boldsymbol{\Theta})$	$r(\theta)$
7	-	-	-	-	-	-
6	0.62787	1.01432	1.00177	1.02698	0.90040	0.95678
6	0.85581	1.04871	1.09991	1.10229	0.97441	0.98723
6	0.95859	1.02296	1.07049	1.05256	0.99357	0.99151
6	0.98869	1.00727	1.02845	1.01791	0.99839	0.99630

Fully-mixed $\mathbb{P}_1 - \mathbf{RT}_1 - \mathbf{P}_2 - \mathbf{RT}_1 - \mathrm{P}_2$ scheme							
DoFs	h	e($\mathbf{t}$)	e($\boldsymbol{\sigma}$)	e($\boldsymbol{u}$)	e(p)	e($\boldsymbol{\Theta}$)	e(θ)
3476	0.7071	0.05677	0.11816	0.07065	0.04174	1.90822	0.59405
25572	0.3535	0.01657	0.03067	0.01769	0.01044	0.51742	0.15859
196292	0.1767	0.00441	0.00784	0.00437	0.00260	0.13213	0.04294
1538436	0.0883	0.00113	0.00199	0.00108	0.00065	0.03354	0.01128

iter	$r(\mathbf{t})$	$r(\boldsymbol{\sigma})$	$r(\boldsymbol{u})$	$r(p)$	$r(\boldsymbol{\Theta})$	$r(\theta)$
6	-	-	-	-	-	-
6	1.77670	1.94583	1.99761	1.99851	1.88281	1.90524
6	1.90946	1.96778	2.01746	2.00408	1.96929	1.88463
6	1.96175	1.97533	2.00942	1.99138	1.97773	1.92771

Table 5.3: Example 5.3.2. Convergence history and Picard iteration count for $k = 0, 1$ (table produced by the author).

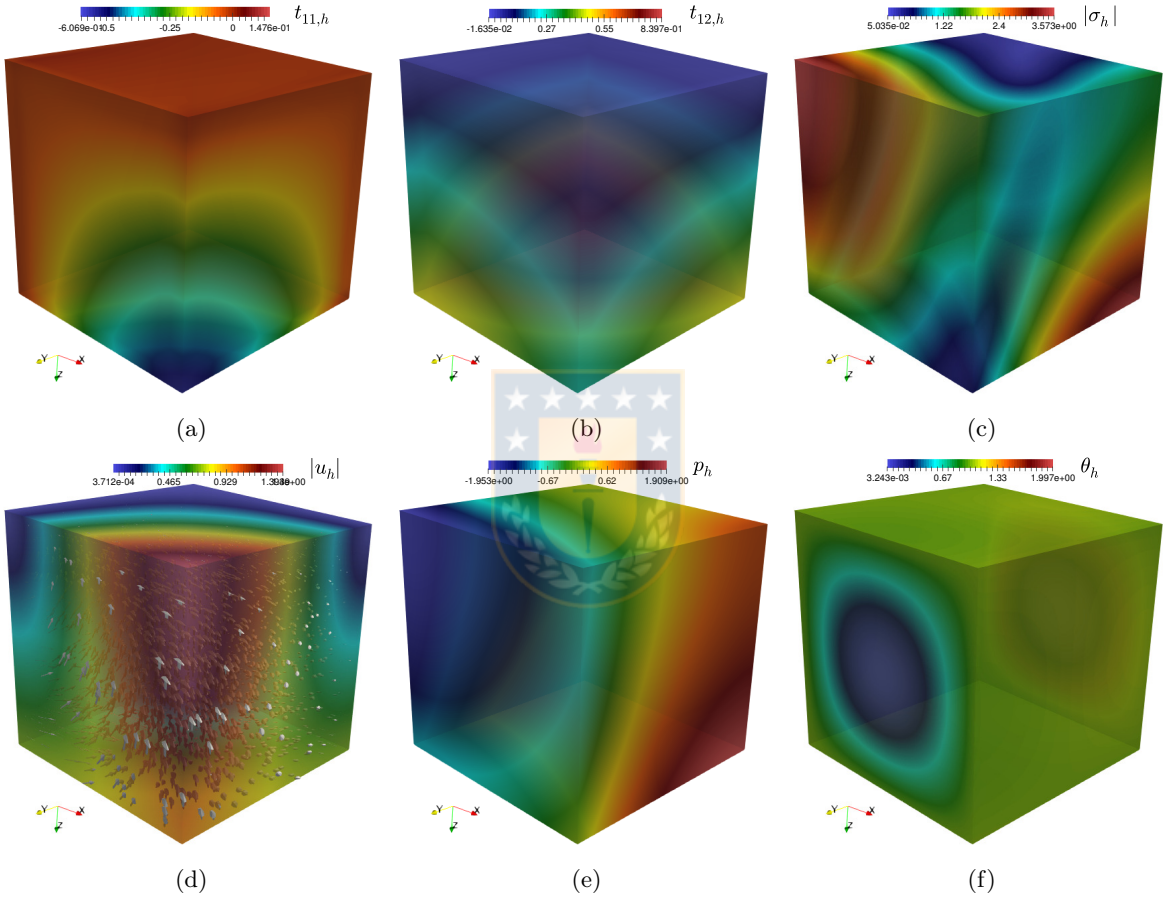


Figure 5.5: Example 5.3.2. Lowest-order approximate solutions: (a)-(b) relevant components of the strain rate, (c) pseudostress magnitude, (d) displacement magnitude, (e) postprocessed pressure, and (f) temperature (figure produced by the author).

Conclusions and summary of the thesis

In this thesis we have developed primal and mixed finite element methods for a set of partial differential equations arising from solid and fluid mechanics problems, more precisely, stress-assisted diffusion, and a phase change problem. We have proved the solvability of the continuous and discrete problems as well as their convergence results, and we have also provided the corresponding numerical tests and simulations. The main conclusions for each of the models are:

Stress-assisted diffusion: We begin by remarking few novelties of our work. We have provided a rigorous mathematical and numerical analysis for these type of problems, which is significant, since in the context of stress-assisted diffusion processes, the available literature has been focused mainly on a modelling point of view, being its mathematical counterpart really scarce (see for example [97, 63, 109]). We have also applied our numerical approach to the modelling of problems of interest, such as the simulation of microscopic lithiation of an anode, obtaining results comparable to those published by [128, 138]. With respect to the main difficulties encountered, we point out the treatment of the trilinear form when proving the Lipschitz continuity of the fixed point operators (cf. (1.3.26) and (2.3.25)). We had to appeal to regularity estimates, which have restricted our analysis to convex domains and the two-dimensional case. However, our numerical results seem to indicate that the restriction is technical and that the scheme can work without it. The above encourages us to think that our analysis could be improved in the future. Regarding extensions of this model, a number of generalisations are envisaged. First, the physical context of Example 3 in Section 1.6, was motivated by the study of stress-assisted diffusion in actively deforming hyperelastic media (see e.g. [53, 112]). The analysis of this class of problems constitutes one of the forthcoming extensions of the present thesis, where the regime of nonlinear elasticity and the difficulties associated to nearly incompressible and incompressible material poses a great challenge. Secondly, it is left to investigate other constitutive relations for the tensor diffusivity ϑ, possibly depending also on the concentration and entailing the study of non-monotone operators and embedded fixed-point schemes [130]. Finally, we remark tha the regularity assumptions and the structure of the mixed formulations will also need to be rewritten once we incorporate concentration gradient modulations of the body loads $\boldsymbol{f}(\phi) = O(\nabla\phi)$, as in [132].

Phase change in porous media: As a new contribution, we have introduced a new viscosity-based model for phase change problems, which does not require the choice of regularisation and jump size parameters. The analysis of this new model can be conducted after adapting classical techniques used to analyse Boussinesq-type problems. For the mixed approach, the novelty of the treatment consists in the ultra-weak imposition of the symmetry of the pseudostress. This fact avoid us the introduction

of one additional unknown, and hence, two additional terms in our scheme, allowing a decrease in computational costs. Furthermore, we have presented several tests serving as numerical validation for the enthalpy-free case, and a set of comparisons and concluding insights drawn from the simulations of the melting case. Regarding the difficulties encountered, we remark the need to assume additional regularity for the Navier-Stokes-Brinkman problem. However, as we can see in Section 5.3.2, that assumption is indeed quite reasonable for $n = 2$ since just $\varepsilon \in (0, 1)$ is required in this case. Moreover, taking into account that we are working with Dirichlet boundary conditions, we conjecture that the assumption for $n = 3$ can be allowed as well. Finally, future extensions of our work include the derivation of error estimates for the time-dependent problem presented in Chapter 4, the generalisation of the enthalpy-viscosity and enthalpy-porosity models to include nonlocal contributions, and performing further comparisons with alternative models, such as those based on the error function and reviewed in [62, 116].

In summary, the thesis has included the following stages:

1. We have focused on a coupled system consisting of the three-field equations of linear elastostatics imposing weakly the symmetry of the Cauchy stress, and a generalised diffusion problem where the diffusion tensor depends nonlinearly on the stress. We have analysed the mathematical properties of this system (existence, uniqueness, and regularity of weak solutions) by means of fixed-point theory and the classical theory for elliptic and saddle-point PDEs. We have also introduced two main families of finite element schemes for the discretisation of the model problem: one that adopts the mixed-primal character of the set of governing equations, and another one based on augmentation and penalisation. The properties of the resulting discrete problems were also established, and we have rigorously proved convergence estimates under suitable assumptions. Finally, we have presented some 2D and 3D tests that exemplify the accuracy of the methods under different regimes.

2. We have introduced a fully-mixed finite element method for the stress-assisted diffusion of a solute into an elastic material. The equations of elastostatics were written in mixed form using stress, rotation and displacements. Then, with the aim of applying regularity estimates, the diffusion equation was reformulated through an augmented variational approach, solving for the solute concentration, for its gradient, and for the diffusive flux. For the existence and uniqueness of the continuous scheme, we have applied the classical Schauder fixed-point theorem in combination with Babuška-Brezzi and Lax-Milgram theories. Concerning the numerical approach, we have proposed two families of finite element methods, based on either PEERS or Arnold-Falk-Winther elements for elasticity, and a Raviart-Thomas and piecewise polynomial triplet approximating the mixed diffusion equation. Then, we have proved the well-posedness of the discrete problems by means of the Brouwer fixed-point theorem together again with Babuška-Brezzi and Lax-Milgram theories, and derived optimal error bounds by using a Strang-type inequality. Finally, the theorical rates of convergence of our method have been confirmed by means of numerical examples, and the applicability of our scheme has been shown through the simulation of 3D microscopic lithiation processes.

3. The numerical analyses of the aforementioned mixed-primal and fully-mixed schemes were complemented by carrying out residual based a posteriori error estimations in two dimensions. We

have proposed estimators mainly based on the use of suitable ellipticity and inf-sup conditions together with a Helmholtz decomposition, and the local approximation properties of the Cleément interpolant and Raviart-Thomas operator. A global efficiency properties with respect to the natural norms were further proved via usual localisation techiques of bubble functions and/or well-known results from previous a posteriori analyses of related mixed elasticity schemes, and elliptic equations.

4. We have studied phase change in Boussinesq models within porous media, and established stability, existence and uniqueness of the continuous and discrete approaches. Furthermore, a finite element method has been proposed in order to obtain numerical approximations. After that, we have proposed a new fully-mixed finite element method for the stationary case. Although we have not theoretically derived the error bounds for any of these methods, we have examined numerically their rates of convergence. Finally, we have tested the performance of the method using a classical benchmark for air convection, and simulated the melting of a solid material.

5. We have introduced a new mixed finite element method for the phase change problem. We formulated the system in terms of pseudostress, strain rate and velocity for the Navier-Stokes-Brinkman equation, whereas temperature, normal heat flux on the boundary, and an auxiliary unknown were introduced for the energy conservation equation. Moreover, we have imposed the symmetry of the pseudostress in an ultra-weak sense, which is one of the novelties of our work, and at the same time it has avoided the introduction of the vorticity as an additional unknown. Two variational formulations were proposed, namely mixed-primal and fully-mixed approaches. The corresponding solvability analysis was established by combining fixed-point arguments, Sobolev embedding theorems and certain regularity assumptions. Based on adequate finite element spaces, we derived two Galerkin schemes, and then we showed its well-posedness. Afterwards, suitable Strang-type inequalities were utilised to rigorously derive a priori error estimates in their natural norms. Finally, several numerical results were provided in order to validate the good performance of the method and confirm the corresponding rate of convergence.

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 for the phase change problem**
   We are interested in developing a posteriori error analysis for the method studied in Chapter 5 in order to improve its robustness in the context of problems involving complex geometries or solutions with high gradients. To do this, we will apply the techniques used in Chapter 3, adapting them to our context of phase change problems.

2. **Development of new mixed finite element methods for poroelasticity**
   We are interested in extending our mixed approximation for the elasticity equations in Chapters 1 and 2 to the poroelasticity problem. In contrast to other works [160, 161] where the author studied a two-dimensional Biot's model depending on the Lamé constants, we propose here a

two or three-dimensional locking-free finite element method for the poroelasticity system. For that, we suggest introducing $\widetilde{p} := \alpha p - \lambda \mathrm{div}\, \boldsymbol{u}$ in Ω, as auxiliary unknown (which will disappear later in our formulation), and then rewriting the system as a two-fold saddle-point scheme. The main advantage of using this new approach lies in the fact that all the resulting equations are independent of λ, which is particularly important to prevent volumetric locking. Furthermore, since the stress is directly approximate in our method, another advantage is the applicability of our approach to problems of interest, such as interface problems, where the interaction between a poromechanic system with other mechanical system requires transmission conditions across the interface, some of which involve the stress tensor.

3. **Development of new mixed finite element approximation for an elasticity - poroelasticity interface problem**
   Taking into account the advantages of applying the method outlined in Point 2, we are interested in extending the analysis of [20] to the fully-mixed finite element case, namely mixed for the elasticity system and mixed for the poromechanic system. In this way, the transmission conditions for this problem (see e.g. [91]) can be imposed naturally and then it is not necessary to rewrite the interface conditions as it was done in [20, eq. (5.4)].

4. **Finite element method for the coupling of the elastodinamyc equation and transport**
   We are interested in extending the results and techniques of Chapters 1 and 2 to the non-stationary case. For this model, we consider the regime of infinitesimal deformations, and assuming that species ϕ can be transported within Ω, then a Lagrangian set of equations representing conservation of mass and momentum gives the following system of linear elasticity equations coupled with a diffusion equation through external loads only:

$$\begin{aligned} \partial_{tt}\boldsymbol{u} - \nabla \cdot \boldsymbol{\sigma} &= \boldsymbol{f}(\phi) \quad \text{in } \Omega, \\ \boldsymbol{\sigma} &= \lambda \,\mathrm{tr}\, \boldsymbol{\varepsilon}(\boldsymbol{u})\mathbb{I} + 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{in } \Omega, \\ \partial_t \phi - \nabla \cdot (\vartheta(\boldsymbol{\sigma})\nabla\phi) &= g(\boldsymbol{u}) \quad \text{in } \Omega. \end{aligned} \tag{5}$$

   In the case where ϕ represents electric potential, then $(5)_3$ states conservation of current and the system models a basic electromechanical phenomenon. The two-way coupling is carried out through a modification of the diffusion properties of species ϕ and by the source term $g(\boldsymbol{u})$. Problem (5) can be regarded as a very rough linearisation of active-stress models for soft tissue, but it also has many other applications. A similar model can be found in [146]; however, it introduces viscous damping forces into the stress, and writes the reaction-diffusion equation in the current configuration (and therefore in terms of velocity transport).

Conclusiones, sumario y trabajo futuro

Conclusiones y sumario de la tesis

En esta tesis hemos desarrollado métodos de elementos finitos primales y mixtos para sistemas de ecuaciones diferenciales parciales relacionados con la mecánica de sólidos y fluidos, más precisamente, difusión asistida por esfuerzo y un problema de cambio de fase. Hemos demostrado solubilidad de los problemas continuo y discreto, así como sus resultados de convergencia, para luego proporcionar sus correspondientes resultados numéricos y simulaciones. Las principales conclusiones de este trabajo son:

Difusión asistida por esfuerzo: Empezamos señalando algunas novedades de nuestro trabajo. Así, hemos proporcionado un análisis matemático y numérico riguroso para este tipo de problemas, lo cual es significativo ya que la literatura para este tipo de problemas, se basa principalmente en la parte de modelamiento, siendo su contraparte matemática realmente escasa (ver por ejemplo [97, 63, 109]). Hemos aplicado nuestra aproximación numérica al modelamiento de problema de interés, tales como la simulación de la litiación en un anodo microscópico, obteniendo resultados comparables a los presentados por ejemplo en [128, 138]. Con respecto a las principales dificultades encontradas, apuntamos al tratamiento de la forma trilineal resultante cuando se está probando la Lipschitz continuity de los operadores de punto fijo (cf. (1.3.26) and (2.3.25)). Así, con la idea de acotar estas formas trilineales, hemos apelado a estimaciones de regularidad, las cuales restringen nuestro análisis al caso de dominios convexos y bidimensionales. Sin embargo, se puede ver de nuestros resultados numéricos que estas restricciones son técnicas, y nuestro esquema numérico funciona sin este tipo de restricciones. Lo anterior nos permite pensar en el hecho de que posiblemente nuestro análisis puede ser mejorado a futuro. Respecto a extensiones de este modelo, un número de generalizaciones son previstas. Primero, el contexto físico del Ejemplo 3 en la Sección 1.6, fue motivado por el estudio de difusión asistida por esfuerzo en deformación activa en medios hiperelásticos (ver [53, 112]). El análisis de esta clase de problemas constituye uno de las futuras extensiones de la presente tesis, donde el régimen de elasticidad no lineal y las dificultades asociadas a materiales incompresibles y casi incompresibles constituye un gran desafío. En segundo lugar, se busca investigar otras relaciones constitutivas para el tensor de difusividad ϑ, posiblemente dependiendo además de la concentration y posiblemente dependiendo además de la concentración e implicando el estudio de operadores no monótonos y esquemas de punto fijo [130]. Finalmente, remarcamos que los supuestos de regularidad y la estructura de las formulaciones mixtas, serán necesarias para ser reescritas una vez que incorporemos el módulo del gradiente de concentración en el término fuente $\boldsymbol{f}(\phi) = O(\nabla\phi)$, como en [132].

Cambio de fase en medios porosos: Como una nueva contribución, hemos propuesto un nuevo

modelo basado en viscosidad para problemas de cambio de fase, el cual no requiere de la escogencia de parámetros de regularización y tamaño de salto. El análisis de este modelo, puede ser establecido después de adaptar técnicas clásicas bien conocidas utilizadas para analizar problemas del tipo Boussinesq. Para la aproximación mixta, la novedad en el tratamiento consiste de la imposición ultra-débil de la simetría del pseudostress. Este hecho nos evita la introducción de una incógnita adicional y por ende, de dos términos adicionales en nuestro esquema, permitiendo una disminución de los costos computacionales. Además, hemos presentado varias pruebas que permiten validar la teoría para el caso de entalpía libre, y un conjunto de comparaciones y conclusiones finales extraídas de las simulaciones del caso de derretimiento. Respecto a las dificultades encontradas, señalamos la necesidad de asumir regularidad adicional para el problema de Navier-Stokes-Brinkman problem. Sin embargo, como podemos ver en la Sección 5.3.2, este supuesto es en efecto razonable para el caso $n = 2$ ya que únicamente necesitamos que $\varepsilon \in (0, 1)$ en este caso. Más aún, aprovechando el hecho de que estamos trabajando con condiciones de borde del tipo Dirichlet, conjeturamos que el supuesto para $n = 3$ puede es razonable también. Finalmente, futuras extensiones de nuestro trabajo incluyen la derivación de estimaciones de error para el problema transiente estudiado en el Capítulo 4, la generalización de los modelos de entalpía-viscosidad y entalpía-viscosidad para incluir contribuciones no locales y desarrollar además comparaciones con modelos alternativos, tales como aquellos basados en la función de error y revisados en [62, 116].

Un sumario de la tesis incluye las siguientes etapas:

1. Nos hemos enfocado en un sistema acoplado que involucra tres incógnitas para las ecuaciones de elastodinámica, en donde se impone débilmente la simetría del esfuerzo de Cauchy, y un problema de difusión generalizado, donde el tensor de difusividad depende no linealmente del esfuerzo. Hemos analizado las propiedades matemáticas de este sistema (existencia, unicidad y regularidad de las soluciones débiles) por medio de la teoría de punto fijo y la teoría clásica para EDPs elípticas y de punto silla. También, hemos introducido dos familias de elementos finitos para la discretización del problema modelo: una del tipo mixto-primal, y otra basada en la aumentación y penalización. Las propiedades del problema discreto resultante, fueron también establecidas, y hemos demostrado rigurosamente las estimaciones de convergencia bajo supuestos adecuados. Finalmente, hemos presentado algunas pruebas en 2D y 3D que ejemplifican la precisión de los métodos bajo diferentes regímenes.

2. Hemos introducido un método de elementos finitos completamente mixto para la difusión asistida por esfuerzo de un soluto en un material elástico. Las ecuaciones de elasticidad se escribieron en forma mixta utilizando, esfuerzo, rotación y desplazamientos. Luego, con el objetivo de aplicar algunas estimaciones de regularidad, la ecuación de difusión fue reformulada para obtener una aproximación variacional aumentada, resolviendo para la concentración de soluto, su gradiente y el flujo difusivo. Para la existencia y unicidad del esquema continuo, hemos aplicado el clásico teorema de punto fijo de Schauder en combinación con las teorías de Babuška-Brezzi y Lax-Milgram. Respecto a la aproximación numérica, hemos propuesto dos familias de elementos finitos, basados ya sea en elementos PEERS o Arnold-Falk-Winther para la elasticidad, y un triplete que involucra Raviart-Thomas y elementos polinomiales a trozos para aproximar la ecuación de difusión mixta. Luego, hemos demostrado que el problema discreto está bien definido mediante el teorema de punto fijo de Brouwer, y las teorías de Babuška-Brezzi y Lax-Milgram, y hemos derivado

cotas de error óptimas mediante el uso de la desigualdad de Strang. Finalmente, los órdenes de convergencia teóricos de nuestro método han sido confirmados mediante ejemplos numéricos, y la aplicabilidad de nuestro esquema ha sido utilizada para estudiar la simulación 3D de procesos de litiación microscópica.

3. Los análisis numéricos de los anteriores esquemas mixto-primal y completamente mixto, fueron complementados desarrollando estimaciones de error a posteriori en bidimensionales del tipo residual. Hemos propuesto estimadores principalmente basados en el uso de condiciones de elipticidad e inf-sup, junto con descomposiciones de Helmholtz, y las propiedades de aproximación locales de los operadores interpolantes de Clément y Raviart-Thomas. Propiedades de eficiencia global respecto a las normas naturales fueron probadas a través de técnicas de localización de funciones burbuja y/o resultados bien conocidos de análisis previos relacionados con esquemas mixtos de elasticidad y ecuaciones elípticas.

4. Hemos estudiado el modelado de problemas de cambio de fase en problemas del tipo Boussinesq, dentro de medios porosos, y hemos establecido la estabilidad, existencia y unicidad, tanto de los problemas continuo como discreto. De esta manera, hemos propuesto un método de elementos finitos para obtener aproximaciones numéricas. Después de eso, hemos propuesto un nuevo método de elementos finitos completamente mixto para el caso estacionario. A pesar de que teóricamente no hemos derivado las cotas de error para ninguno de los métodos, hemos examinado numéricamente sus órdenes de convergencia. Finalmente, hemos probado el rendimiento del método utilizando pruebas clásicas de referencia para la convección de aire, y hemos simulado el derretimiento de un material sólido.

5. Hemos introducido un nuevo método de elementos finitos mixtos para el problema de cambio de fase. Por conveniencia del análisis, formulamos el sistema en términos del pseudo-esfuerzo, tensión y velocidad para la ecuación Navier-Stokes-Brinkman, mientras que temperatura, flujo de calor normal en el límite y una incógnita auxiliar han sido introducidas para la ecuación de conservación de energía. Así, con la intención de proponer la correspondiente formulación variacional mixta para nuestro modelo, hemos impuesto la simetría del pseudo-esfuerzo de manera ultra-débil, lo cual ha sido una de las novedades de nuestro trabajo, y al mismo tiempo, nos ha permitido evitar el uso del tensor de vorticidad como una incógnita adicional. Luego, para el análisis matemático se propusieron dos formulaciones variacionales, las cuales llamamos: aproximaciones mixta-primal y completamente mixta. El correspondiente análisis de solubilidad se estableció combinando, argumentos de punto fijo, teoremas de inclusión de Sobolev y ciertos supuestos de regularidad. En consecuencia, utilizando espacios de elementos finitos adecuados, hemos derivado los correspondientes dos esquemas de Galerkin, y luego demostrado que ambos esquemas están bien puestos. Posteriormente, se utilizaron desigualdades de tipo Strang para derivar rigurosamente estimaciones de error a priori en sus normas naturales. Finalmente, se proporcionaron varios resultados numéricos para validar el buen desempeño del método y confirmar los órdenes de convergencia correspondientes.

Trabajos futuros

Los métodos desarrollados y los resultados obtenidos en esta tesis han motivado varios proyectos en curso y futuros. Algunos de ellos se describen a continuación:

1. **Análisis de error a posteriori para el problema de cambio de fase**
   Estamos interesados en desarrollar análisis de error a posteriori para el método estudiado en el Capítulo 5, y de esta forma, mejorar su solidez en el contexto de problemas que involucran geometrías complejas o soluciones con altos gradientes. Para desarrollar esto, aplicaremos las técnicas utilizadas en el Capítulo 3, adaptándolas a nuestro contexto de cambio de fase

2. **Desarrollo de nuevos métodos de elementos finitos mixtos para poroelasticidad**
   Estamos interesados en extender nuestra aproximación mixta para el problema de elasticidad, la cual se muestra en los Capítulos 1 y 2, al problema de la poroelasticidad. A diferencia de otros trabajos [160, 161] donde el autor estudió un modelo bidimensional completamente mixto para Biot, el cual depende de las constantes de Lamé, nosotros proponemos aquí, un método de elementos finitos libre de bloqueo, ya sea en dos o tres dimensiones, para la poroelasticidad. Para eso, sugerimos introducir $\widetilde{p} := \alpha p - \lambda \mathrm{div}\, \boldsymbol{u}$ in Ω, como incógnita auxiliar (la cual desaparecerá más adelante en nuestra formulación), y luego reescribir el sistema como un esquema de doble punto silla. Por lo tanto, la principal ventaja de utilizar este nuevo enfoque, radica en el hecho de que todas las ecuaciones resultantes son independientes de λ, lo cual es particularmente importante para evitar el bloqueo volumétrico. Además, dado que el esfuerzo es aproximado directamente en nuestro método, otra ventaja es la aplicabilidad de nuestro enfoque a problemas de interés. En particular, podemos aplicar nuestro método a problemas de interfaz, donde la interacción entre un sistema poromecánico con otro sistema mecánico necesita condiciones de transmisión a través de la interfaz, algunas de las cuales involucran el tensor de esfuerzo.

3. **Desarrollo de una nueva aproximación de elementos finitos mixtos para un problema de interfaz elasticidad-poroelasticidad**
   Teniendo en cuenta las ventajas de aplicar nuestro método descrito en el Punto 2, estamos interesados en extender el análisis de [20] al caso de elementos finitos completamente mixtos, es decir, mixto para elasticidad y mixto para el sistema poromecánico. De esta forma, las condiciones de transmisión para este problema (véase, por ejemplo, [91]) se pueden imponer de forma natural, sin la necesidad por ejemplo de reescribirlas, tal y como fue necesario hacerlo en [20, eq. (5.4)].

4. **Método de elementos finitos para el acoplamiento de la ecuación de elastodinámica y transporte**
   Estamos interesados en extender los contenidos de los Capítulos 1 y 2 al caso no estacionario. Para este modelo, consideramos el régimen de deformaciones infinitesimales y admitimos que las especies ϕ pueden transportarse dentro de Ω. Luego, un conjunto de ecuaciones lagrangianas que representan la conservación de masa y momento, proporcionan el siguiente sistema de elasticidad

lineal junto con una ecuación de difusión que toma en cuenta solo cargas externas:

$$\begin{aligned} \partial_{tt}\boldsymbol{u} - \nabla \cdot \boldsymbol{\sigma} &= \boldsymbol{f}(\phi) \quad \text{in } \Omega, \\ \boldsymbol{\sigma} &= \lambda \operatorname{tr} \boldsymbol{\varepsilon}(\boldsymbol{u}) \mathbb{I} + 2\mu \boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{in } \Omega, \\ \partial_t \phi - \nabla \cdot (\vartheta(\boldsymbol{\sigma}) \nabla \phi) &= g(\boldsymbol{u}) \quad \text{in } \Omega. \end{aligned} \tag{5}$$

En caso de que ϕ represente el potencial eléctrico, entonces (5) establece la conservación de la corriente, y el sistema modela un fenómeno electromecánico básico. El acoplamiento bidireccional se lleva a cabo mediante una modificación de las propiedades de difusión de las especies ϕ y por el término fuente $g(\boldsymbol{u})$. El problema (5) puede ser considerado como una linealización muy aproximada de modelos de esfuerzo activo para tejidos blandos, pero también tiene muchas otras aplicaciones. Un modelo similar, pero que introduce fuerzas de amortiguamiento viscosas en la tensión, y escribe la ecuación de reacción-difusión en la configuración actual (y, por lo tanto, en términos de transporte de velocidad), se puede encontrar en [146].

References

[1] H. Abboud, V. Girault, and T. Sayah, *A second order accuracy for a full discretized time-dependent Navier-Stokes equations by a two-grid scheme*, Numerische Mathematik, 114 (2009), pp. 189–231.

[2] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, Academic Press, Elsevier Ltd, 2003.

[3] S. Agmon, *Lectures on elliptic boundary value problems*, D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London, 1965.

[4] R. Agroum, C. Bernardi, and J. Satouri, *Spectral discretization of the time-dependent Navier-Stokes problem coupled with the heat equation*, Applied Mathematics and Computation, 268 (2015), pp. 59–82.

[5] J. Ahrens, B. Geveci, and C. Law, *ParaView: An End-User Tool for Large Data Visualization*, Visualization Handbook, Elsevier, 2005.

[6] E. Aifantis, *On the problem of diffusion in solids*, Acta Mechanica, 37 (1980), pp. 265–296.

[7] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Computer Methods in Applied Mechanics and Engineering, 142 (1997), pp. 1–88.

[8] R. Aldbaissy, F. Hecht, G. Mansour, and T. Sayah, *A full discretisation of the time-dependent Boussinesq (buoyancy) model with nonlinear viscosity*, Calcolo, 55 (2018), pp. Art. 44, 49.

[9] J. A. Almonacid, G. N. Gatica, and R. Oyarzúa, *A mixed–primal finite element method for the Boussinesq problem with temperature-dependent viscosity*, Calcolo, 55 (2018), pp. Art. 36, 42.

[10] ——, *A new mixed finite element method for the n-dimensional boussinesq problem with temperature-dependent viscosity*, Preprint 2018-18, Centro de Investigación en Ingeniería Matemática (CI^2MA), Universidad de Concepción, Chile, (2018).

[11] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes, and G. Wells, *The fenics project version 1.5*, Archive of Numerical Software, 100 (2015), pp. 9–23.

[12] A. Alonso, *Error estimators for a mixed method*, Numerische Mathematik, 74 (1996), pp. 385–395.

[13] M. Álvarez, G. N. Gatica, B. Gomez-Vargas, and R. Ruiz-Baier, *New mixed finite element methods for natural convection with phase-change in porous media*, Journal of Scientific Computing, 80 (2019), pp. 141–174.

[14] M. Álvarez, 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.

[15] ——, *A mixed-primal finite element approximation of a sedimentation-consolidation system*, Mathematical Models and Methods in Applied Sciences, 49 (2016), pp. 867–900.

[16] M. Alvarez, G. N. Gatica, and R. Ruiz-Baier, *A posteriori error analysis for a viscous flow-transport problem*, ESAIM: Mathematical Modelling and Numerical Analysis, 50 (2016), pp. 1789–1816.

[17] M. Alvarez, G. N. Gatica, and R. Ruiz-Baier, *A posteriori error estimation for an augmented mixed-primal method applied to sedimentation-consolidation systems*, Journal of Computational Physics, 367 (2018), pp. 322–346.

[18] F. Ampofo and T. G. Karayiannis, *Experimental benchmark data for turbulent natural convection in an air filled square cavity*, International Journal of Heat and Mass Transfer, 46 (2003), pp. 3551–3572.

[19] Y. An and H. Jiang, *A finite element simulation on transient large deformation and mass diffusion in electrodes for lithium ion batteries*, Modelling and Simulation in Materials Science and Engineering, 21 (2013), p. 074007.

[20] V. Anaya, Z. D. Wijn, B. Gómez-Vargas, D. Mora, and R. Ruiz-Baier, *Rotation-based mixed formulations for an elasticity-poroelasticity interface problem*, SIAM Journal on Scientific Computing, in press (2020).

[21] S. Arena, E. Casti, J. Gasia, L. Cabeza, and G. Cau, *Numerical simulation of a finned-tube lhtes system: influence of the mushy zone constant on the phase change behaviour*, Energy Procedia, 126 (2017), pp. 517–524.

[22] 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.

[23] D. N. Arnold, R. S. Falk, and R. Winther, *Finite element exterior calculus, homological techniques, and applications*, Acta Numerica, 15 (2006), pp. 1–155.

[24] ——, *Mixed finite element methods for linear elasticity with weakly imposed symmetry*, Mathematics of Computation, 76 (2007), pp. 1699–1723.

[25] I. Babuška and G. N. Gatica, *A residual-based a posteriori error estimator for the Stokes-Darcy coupled problem*, SIAM Journal on Numerical Analysis, 48 (2010), pp. 498–523.

[26] I. Babuška and W. Rheinboldt, *A posteriori error estimates for the finite element method*, International Journal for Numerical Methods in Engineering, 12 (1978), pp. 1597–1615.

[27] C. Bacuta and J. Bramble, *Regularity estimates for solutions of the equations of linear elasticity in convex plane polygonal domains*, Zeitschrift für angewandte Mathematik und Physik, 54 (2003), pp. 874–878.

[28] T. P. Barrios, G. N. Gatica, M. González, and N. Heuer, *A residual based a posteriori error estimator for an augmented mixed finite element method in linear elasticity*, M2AN Mathematical Modelling and Numerical Analysis, 40 (2006), pp. 843–869.

[29] C. Beckermann and R. Viskanta, *Natural convection solid/liquid phase change in porous media*, International Journal of Heat and Mass Transfer, 31 (1988), pp. 35–46.

[30] O. Ben-David, A. Levy, B. Mikhailovich, and A. Azulay, *3D numerical and experimental study of gallium melting in a rectangular container*, International Journal of Heat and Mass Transfer, 67 (2013), pp. 260–271.

[31] C. Bernardi, L. El Alaoui, and Z. Mghazli, *A posteriori analysis of a space and time discretization of a nonlinear model for the flow in partially saturated porous media*, IMA Journal of Numerical Analysis, 34 (2014), pp. 1002–1036.

[32] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numerische Mathematik, 85 (2000), pp. 579–608.

[33] R. E. Bird, W. M. Coombs, and S. Giani, *A posteriori discontinuous Galerkin error estimator for linear elasticity*, Applied Mathematics and Computation, 344/345 (2019), pp. 78–96.

[34] J. Boland and W. Layton, *An analysis of the finite element method for natural convection problems*, Numerical Methods for Partial Differential Equations, 6 (1990), pp. 115–126.

[35] J. Boussinesq, *Théorie de l'écoulement tourbillonnant et tumultueux des liquides dans les lits rectilignes a grande section*, Gauthier-Villars et fils, 1897.

[36] D. Braess, O. Klaas, R. Niekamp, E. Stein, and F. Wobschal, *Error indicators for mixed finite elements in 2-dimensional linear elasticity*, Computer Methods in Applied Mechanics and Engineering, 127 (1995), pp. 345–356.

[37] G. Brandeis and B. Marsh, *The convective liquidus in solidifying magma chamber: a fluid dynamic investigation*, Nature, 339 (1989), pp. 613–616.

[38] A. Brent, V. Voller, and K. Reid, *Enthalpy-porosity technique for modeling convection-diffusion phase change: application to the melting of a pure metal*, Numerical Heat Transfer, 13 (1988), pp. 297–318.

[39] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Series in Computational Mathematics, 15. Springer–Verlag, New York, 1991.

[40] H. Brinkman, *The viscosity of concentrated suspensions and solutions*, The Journal of Chemical Physics, 20 (1952), p. 571.

[41] R. Bürger, C. Liu, and W. L. Wendland, *Existence and stability for mathematical models of sedimentation–consolidation processes in several space dimensions*, Journal of Computational Physics, 264 (2001), pp. 288–310.

[42] J. Camaño, R. Oyarzúa, R. Ruiz-Baier, and G. Tierra, *Error analysis of an augmented mixed method for the Navier-Stokes problem with mixed boundary conditions*, IMA Journal of Numerical Analysis, 38 (2018), pp. 1452–1484.

[43] J. Camaño, G. N. Gatica, R. Oyarzúa, and R. Ruiz-Baier, *An augmented stress-based mixed finite element method for the steady state Navier–Stokes equations with nonlinear viscosity*, Numerical Methods for Partial Differential Equations, 33 (2017), pp. 1692–1725.

[44] Y. Cao and S. Chen, *Analysis and finite element approximation of bioconvection flows with concentration dependent viscosity*, International Journal of Numerical Analysis and Modelling, 11 (2014), pp. 86–101.

[45] P. Carman, *Fluid flow through granular beds*, Transactions of the Institution of Chemical Engineers, 15 (1937), pp. 150–166.

[46] C. Carstensen, *A posteriori error estimate for the mixed finite element method*, Mathematics of Computation, 66 (1997), pp. 465–476.

[47] C. Carstensen and G. Dolzmann, *A posteriori error estimates for mixed FEM in elasticity*, Numerische Mathematik, 81 (1998), pp. 187–209.

[48] C. Carstensen, O. Scherf, and P. Wriggers, *Adaptive finite elements for elastic bodies in contact*, SIAM Journal on Scientific Computing, 20 (1999), pp. 1605–1626.

[49] S. Caucao, G. N. Gatica, and R. Oyarzúa, *A posteriori analysis of an augmented fully mixed formulation for the nonisothermal Oldroyd-Stokes problem*, Numerical Methods for Partial Differential Equations, 35 (2019), pp. 295–324.

[50] S. Chang, J. Moon, and M. Cho, *Stress-diffusion coupled multiscale analysis of si anode for li-ion battery*, Journal of Mechanical Science and Technology, 29 (2015), pp. 4807–4816.

[51] L. Chen, J. Hu, X. Huang, and H. Man, *Residual-based a posteriori error estimates for symmetric conforming mixed finite elements for linear elasticity problems*, Science China. Mathematics, 61 (2018), pp. 973–992.

[52] Y. Chen, J. Huang, X. Huang, and Y. Xu, *On the local discontinuous Galerkin method for linear elasticity*, Mathematical Problems in Engineering, (2010), pp. Art. ID 759547, 20.

[53] C. Cherubini, S. Filippi, A. Gizzi, and R. Ruiz-Baier, *A note on stress-driven anisotropic diffusion and its role in active deformable media*, Journal of Theoretical Biology, 430 (2017), pp. 221–228.

[54] P. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[55] P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[56] P. Clément, *Approximation by finite element functions using local regularisation*, RAIRO Modélisation Mathématique et Analyse Numérique, 9 (1975), pp. 77–84.

[57] 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.

[58] ——, *Fixed point strategies for mixed variational formulations of the stationary Boussinesq problem*, Comptes Rendus Mathématique. Académie des Sciences. Paris, 354 (2016), pp. 57–62.

[59] ——, *An augmented fully-mixed finite element method for the stationary Boussinesq problem*, Calcolo, 54 (2017), pp. 167–205.

[60] ——, *A posteriori error analysis of an augmented mixed-primal formulation for the stationary Boussinesq model*, Calcolo, 54 (2017), pp. 1055–1095.

[61] E. Colmenares and M. Neilan, *Dual-mixed finite element methods for the stationary Boussinesq problem*, Computers and Mathematics with Applications, 72 (2016), pp. 1828–1850.

[62] A. Costa, *Viscosity of high crystal content melts: Dependence on solid fraction*, Geophysical Research Letters, 32 (2005), p. L22308.

[63] R. Cox, *Stress-assisted diffusion: A free boundary problem*, SIAM Journal on Applied Mathematics, 51 (1991), pp. 1522–1537.

[64] I. Danaila, R. Moglan, F. Hecht, and S. L. Masson, *A newton method with adaptive finite elements for solving phase-change problems with natural convection*, Journal of Computational Physics, 274 (2014), pp. 826–840.

[65] B. V. de Fliert and R. V. der Hout, *Stress-driven diffusion in a drying liquid paint layer*, European Journal of Applied Mathematics, 9 (1988), pp. 447–461.

[66] G. de Vahl Davis, *Natural convection of air in a square cavity: A benchmark numerical solution*, International Journal for Numerical Methods in Fluids, 3 (1983), pp. 249–264.

[67] J. Deteix, A. Jendoubi, and D. Yakoubi, *A coupled prediction scheme for solving the Navier-Stokes and convection-diffusion equations*, SIAM Journal on Numerical Analysis, 52 (2014), pp. 2415–2439.

[68] N. S. Dhaidan, J. Khodadadi, T. A. Al-Hattab, and S. M. Al-Mashat, *Experimental and numerical study of constrained melting of n-octadecane with cuo nanoparticle dispersions in a horizontal cylindrical capsule subjected to a constant heat flux*, International Journal of Heat and Mass Transfer, 67 (2013), pp. 523–534.

[69] M. S. Dinniman, X. S. Asay-Davis, B. K. Galton-Fenzi, P. R. Holland, A. Jenkins, and R. Timmermann, *Modeling ice shelf/ocean interaction in antarctica: A review*, Oceanography, 29 (2016), pp. 144–153.

[70] C. 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.

[71] Y. Dutil, D. R. Rousse, N. B. Salah, S. Lassue, and L. Zalewski, *A review on phase-change materials: Mathematical modeling and simulations*, Renewable and Sustainable Energy Reviews, 15 (2011), pp. 112–130.

[72] A. Einstein, *Eine neue bestimmung der molekuldimensionen*, Annalen der Physik, 19 (1906), pp. 286–306.

[73] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, 2nd ed., 2014.

[74] K. Ensslin, *Quantum physics in quantum dots. nanophysics: Coherence and transport*, École d'été de Physique des Houches, Session LXXXI Les Houches, (2005), pp. 585–586.

[75] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Applied Mathematical Sciences, 159. Springer–Verlag, New York, 2004.

[76] M. Farhloul, S. Nicaise, and L. Paquet, *A mixed formulation of Boussinesq equations: analysis of nonsingular solutions*, Mathematics of Computation, 69 (2000), pp. 965–986.

[77] L. E. Figueroa, G. N. Gatica, and A. Márquez, *Augmented mixed finite element methods for the stationary Stokes equations*, SIAM Journal on Scientific Computing, 31 (2008/09), pp. 1082–1119.

[78] J. M. Foster, S. J. Chapman, G. Richardson, and B. Protas, *A mathematical model for mechanically-induced deterioration of the binder in lithium-ion electrodes*, SIAM Journal on Applied Mathematics, 77 (2016), pp. 2172–2198.

[79] G. Gatica, *Analysis of a new augmented mixed finite element method for linear elasticity allowing* $\mathrm{RT}_0 - \mathrm{P}_1 - \mathrm{P}_0$ *approximations*, M2AN Mathematical Modelling and Numerical Analysis, 40 (2006), pp. 1–28.

[80] G. N. Gatica, *An augmented mixed finite element method for linear elasticity with non-homogeneous dirichlet conditions*, Electronic Transactions on Numerical Analysis, 26 (2007), pp. 421–438.

[81] ——, *A Simple Introduction to the Mixed Finite Element Method: Theory and Applications*, SpringerBriefs in Mathematics. Springer, Cham, 2014.

[82] G. N. Gatica, L. F. Gatica, and F. Sequeira, *A priori and a posteriori error analyses of a pseudostress-based mixed formulation for linear elasticity*, Computers and Mathematics with Applications, 71 (2016), pp. 585–614.

[83] G. N. Gatica, B. Gomez-Vargas, and R. Ruiz-Baier, *Analysis and mixed-primal finite element discretisations for stress-assisted diffusion problems*, Computer Methods in Applied Mechanics and Engineering, 337 (2018), pp. 411–438.

[84] ——, *Formulation and analysis of fully-mixed methods for stress-assisted diffusion problems*, Computers and Mathematics with Applications, 77 (2019), pp. 1312–1330.

[85] G. N. Gatica and N. Heuer, *An expanded mixed finite element approach via a dual–dual formulation and the minimum residual method*, Journal of Computational and Applied Mathematics, 132 (2001), pp. 371–385.

[86] G. N. Gatica, A. Márquez, and S. Meddahi, *An augmented mixed finite element method for 3d linear elasticity problems*, Journal of Computational and Applied Mathematics, 231 (2009), pp. 526–540.

[87] G. N. Gatica, A. Márquez, R. Oyarzúa, and R. Rebolledo, *Analysis of an augmented fully-mixed approach for the coupling of quasi-newtonian fluids and porous media*, Computer Methods in Applied Mechanics and Engineering, 270 (2014), pp. 76–112.

[88] G. N. Gatica, A. Márquez, and M. A. Sánchez, *Analysis of a velocity-pressure-pseudostress formulation for the stationary Stokes equations*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 1064–1079.

[89] G. N. Gatica, R. Ruiz-Baier, and G. Tierra, *A posteriori error analysis of an augmented mixed method for the Navier–Stokes equations with nonlinear viscosity*, Computers and Mathematics with Applications, 72 (2016), pp. 2289–2310.

[90] C. Geuzaine and J. F. Remacle, *Gmsh Reference Manual*, 1.12 ed., Aug. 2003.

[91] V. Girault, G. Pencheva, M. F. Wheeler, and T. Wildey, *Domain decomposition for poroelasticity and elasticity with DG jumps and mortars*, Mathematical Models and Methods in Applied Sciences, 21 (2011), pp. 169–213.

[92] P. Grisvard, *Smoothness of the solution of a monotonic boundary value problem for a second order elliptic equation in a general convex domain*, Lecture Notes in Mathematics, 564 (1976), pp. 135–151.

[93] ——, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics, 24. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[94] C. Gritton, J. Guilkey, J. Hopper, D. Bedrov, R. Kirby, and M. Berzins, *Using the material point method to model chemical/mechanical coupling in the deformation of a silicon anode*, Modelling and Simulation in Materials Science and Engineering, 25 (2017), pp. 1–22.

[95] S. Gupta, *A moving grid numerical scheme for multi-dimensional solidification with transition temperature range*, Computer Methods in Applied Mechanics and Engineering, 189 (2000), pp. 525–544.

[96] F. Hecht, *New development in freefem++*, Journal of Numerical Mathematics, 20 (2012), pp. 251–265.

[97] J. Hill, *Plane steady solutions for stress-assisted diffusion*, Mechanics Research Communications, 6 (1979), pp. 147–150.

[98] W. Hong, X. Zhao, J. Zhou, and Z. Suo, *A theory of coupled diffusion and large deformation in polymeric gels*, Journal of the Mechanics and Physics of Solids, 56 (2008), pp. 1779–1793.

[99] H. Hsiang-Wen, F. Postberg, Y. Sekine, T. Shibuya, S. Kempf, M. Horányi, A. Juhász, N. Altobelli, K. Suzuki, Y. Masaki, T. Kuwatani, S. Tachibana, S. Sin-iti, G. Moragas-Klostermeyer, and R. Srama, *Ongoing hydrothermal activities within enceladus*, Nature, 7549 (2015), pp. 207–210.

[100] D. Iyi and R. Hasan, *Natural convection flow and heat transfer in an enclosure containing staggered arrangement of blockages*, Procedia Engineering, 105 (2015), pp. 176–183.

[101] 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.

[102] A. Khan, C. E. Powell, and D. J. Silvester, *Robust a posteriori error estimators for mixed approximation of nearly incompressible elasticity*, International Journal for Numerical Methods in Engineering, 119 (2019), pp. 18–37.

[103] D. Klepach and T. Zohdi, *Strain assisted diffusion: Modeling and simulation of deformation-dependent diffusion in composite media*, Computers and Mathematics with Applications, 56 (2014), pp. 1728–1738.

[104] T. A. Kowalewski and M. Rebow, *Freezing of water in a differentially heated cubic cavity*, International Journal of Computational Fluid Dynamics, 11 (1999), pp. 193–210.

[105] J. Kozeny, *Ueber kapillare leitung des wassers im boden*, Sitzungsber. Akad. Wiss. Wien Math.-Naturwiss, 136 (1927), pp. 271–306.

[106] I. M. Krieger and T. J. Dougherty, *A mechanism for non-newtonian flow in suspension of rigid spheres*, Transactions of the Society of Rheology, 3 (1959), pp. 137–152.

[107] S. Larsson and V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer-Verlag, Berlin, 2003.

[108] P. Lenarda, M. Paggi, and R. Ruiz Baier, *Partitioned coupling of advection-diffusion-reaction systems and Brinkman flows*, Journal of Computational Physics, 344 (2017), pp. 281–302.

[109] M. Lewicka and P. Mucha, *A local and global well-posedness results for the general stress-assisted diffusion systems*, Journal of Elasticity, 123 (2016), pp. 19–41.

[110] J. Li, N. Lotfi, R. Landers, and J. Park, *A single particle model for lithium-ion batteries with electrolyte and stress-enhanced diffusion physics*, Journal of Electrochemical Society, 164 (2017), pp. A874–A883.

[111] M. Lonsing and R. Verfürth, *A posteriori error estimators for mixed finite element methods in linear elasticity*, Numerische Mathematik, 97 (2004), pp. 757–778.

[112] A. Loppini, A. Gizzi, R. Ruiz-Baier, C. Cherubini, F. H. Fenton, and S. Filippi, *Competing mechanisms of stress-assisted diffusivity and stretch-activated currents in cardiac electromechanics*, Frontiers in Physiology, 9 (2018), p. 1714.

[113] C. Lovadina and R. Stenberg, *Energy norm a posteriori error estimates for mixed finite element methods*, Mathematics of Computation, 75 (2006), pp. 1659–1674.

[114] T. H. C. Luong and C. Daveau, *A posteriori estimates for discontinuous Galerkin method to the elasticity problem*, Numerical Methods for Partial Differential Equations, 34 (2018), pp. 1348–1369.

[115] X. Ma, Z. Tao, and T. Zhang, *A variational multiscale method for steady natural convection problem based on two-grid discretization*, Advances in Difference Equations, (2016), pp. Paper No. 85, 20pp.

[116] H. Mader, E. Llewellin, and S. Mueller, *The rheology of two-phase magmas: A review and analysis*, Journal of Volcanology and Geothermal Research, 257 (2013), pp. 135–158.

[117] M. L. Manda, R. Shepard, B. Fair, and H. Massoud, *Stress-assisted diffusion of boron and arsenic in silicon*, Materials Research Society Symposium Proceedings, 36 (1985), pp. 71–76.

[118] L. Meisel, *Stress-assisted diffusion to dislocations and its role in strain aging*, Journal of Applied Physics, 38 (1967), pp. 4780–4784.

[119] D. Mora and G. Rivera, *A priori and a posteriori error estimates for a virtual element spectral analysis for the elasticity equations*, IMA Journal of Numerical Analysis, (2018).

[120] K. Morgan, *A numerical analysis of freezing and melting with convection*, Computer Methods in Applied Mechanics and Engineering, 28 (1981), pp. 275–284.

[121] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague, 1967.

[122] 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.

[123] R. Oyarzúa and P. Zúñiga, *Analysis of a conforming finite element method for the Boussinesq problem with temperature-dependent parameters*, Journal of Computational and Applied Mathematics, 323 (2017), pp. 71–94.

[124] Y. L. Pentrec and G. Lauriat, *Effects of the heat transfer at the side walls on natural convection in cavities*, Journal of Heat Transfer, 112 (1990), pp. 370–378.

[125] Y. Podstrigach and V. Pavlina, *Differential equations of thermodynamic processes in n−component solid solutions*, Soviet Materials Science, 1 (1966), pp. 259–264.

[126] H. Quinn, *A reconciliation of packed column permeability data: Column permeability as a function of particle porosity*, Journal of Materials Research, 323 (2014), p. 636507.

[127] J. E. Roberts and J.-M. Thomas, *Mixed and Hybrid Methods*, Handbook of Numerical Analysis, Vol. II, 523–639, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991.

[128] I. Roper, *Stress distributions in silicon electrodes during lithiation*, InFoMM CDT miniproject, University of Oxford, (2017).

[129] R. Roscoe, *The viscosity of suspensions of rigid spheres*, British Journal of Applied Physics, 3 (1952), pp. 267–269.

[130] T. Roubiček, *Nonlinear Partial Differential Equations with Applications*, Int. Ser. Numer. Math. Vol. 153. Birkhäuser, Basel, 2005.

[131] S. Roy, K. Vengadassalam, Y. Wang, S. Park, and K. Liechti, *Characterization and modeling of strain assisted diffusion in an epoxy adhesive layer*, Transport in Porous Media, 43 (2006), pp. 27–52.

[132] R. Ruiz-Baier, *Primal-mixed formulations for reaction-diffusion systems on deforming domains*, Journal of Computational Physics, 299 (2015), pp. 320–338.

[133] R. Ruiz-Baier and H. Torres, *Numerical solution of a multidimensional sedimentation problem using finite volume-element methods*, Applied Numerical Mathematics, 95 (2015), pp. 280–291.

[134] P. W. Schroeder and G. Lube, *Stabilised dG-FEM for incompressible natural convection flows with boundary and moving interior layers on non-adapted meshes*, Journal of Computational Physics, 335 (2017), pp. 760–779.

[135] M. A. Sheremet, T. Grosan, and I. Pop, *Natural convection and entropy generation in a square cavity with variable temperature side walls filled with a nanofluid: Buongiorno's mathematical model*, Entropy, 19 (2017), p. 337.

[136] Y. Song, X. Shao, Z. Guo, and J. Zhang, *Role of material properties and mechanical constraint on stress-assisted diffusion in plate electrodes of lithium ion batteries*, Journal of Physics D: Applied Physics, 46 (2013), p. 105307.

[137] M. Tabata and D. Tagami, *Error estimates of finite element methods for nonstationary thermal convection problems with temperature-dependent coefficients*, Numerische Mathematik, 100 (2005), pp. 351–372.

[138] V. Taralova, O. Iliev, and Y. Efendiev, *Derivation and numerical validation of a homogenized isothermal li-ion battery model*, Journal of Engineering Mathematics, 101 (2016), pp. 1–27.

[139] J. Toribio and V. Kharin, *High-resolution numerical modelling of stress-strain fields in the vicinity of a crack tip subjected to fatigue*, Fracture from Defects, EMAS, (1998), pp. 1059–1064.

[140] ——, *Role of fatigue crack closure stresses in hydrogen assisted cracking*, Advances in Fatigue Crack Closure Measurement and Analysis, ASTM STP 1343, R.C. McClung, J.C. Newman, Eds., ASTM International, West Gonshohocken, (1999), pp. 440–458.

[141] ——, *A hydrogen diffusion model for applications in fusion nuclear technology*, Fusion Engineering and Design, 299 (2000), pp. 213–218.

[142] ——, *Role of cyclic pre-loading in hydrogen assisted cracking*, Environmentally Assisted Cracking: Predictive Methods for Risk Assessment and Evaluation of Materials, Equipment, and Structures. ASTM STP 1401. ASTM, West Conshohocken (PA), (2000), pp. 329–342.

[143] J. Toribio, V. Kharin, D. Vergara, and M. Lorenzo, *Two-dimensional numerical modelling of hydrogen diffusion in metals assisted by both stress and strain*, in Light Weight Metal Corrosion and Modeling, vol. 138 of Advanced Materials Research, Trans Tech Publications Ltd, 2010, pp. 117–126.

[144] C. Truesdell, *Mechanical basis of diffusion*, The Journal of Chemical Physics, 37 (1962), pp. 2336–2344.

[145] M. Ulvrová, S. Labrosse, N. Coltice, P. Røaback, and P. Tackley, *Numerical modelling of convection interacting with a melting and solidification front: Application to the thermal evolution of the basal magma ocean*, Physics of the Earth and Planetary Interiors, 206-207 (2012), pp. 51–66.

[146] B. L. Vaughan, Jr., R. E. Baker, D. Kay, and P. K. Maini, *A modified Oster-Murray-Harris mechanical model of morphogenesis*, SIAM Journal on Applied Mathematics, 73 (2013), pp. 2124–2142.

[147] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, Journal of Computational and Applied Mathematics, 50 (1994), pp. 67–83.

[148] R. Verfürth, *A Review of A-Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley Teubner, Chichester, 1996.

[149] R. Verfürth, *A review of a posteriori error estimation techniques for elasticity problems*, Computer Methods in Applied Mechanics and Engineering, 176 (1999), pp. 419–440.

[150] G. Vidalain, L. Gosselin, and M. Lacroix, *An enhanced thermal conduction model for the prediction of convection dominated solid-liquid phase change*, International Journal of Heat and Mass Transfer, 52 (2009), pp. 1753–1760.

[151] V. R. Voller, M. Cross, and N. C. Markatos, *An enthalpy method for convection/diffusion phase change*, International Journal for Numerical Methods in Engineering, 24 (1987), pp. 271–284.

[152] V. R. Voller and C. Prakash, *A fixed grid numerical modelling methodology for convection and phase transition efficiently*, Journal of Computational Physics, 30 (1987), pp. 1709–1719.

[153] R. T. Walker and D. M. Holland, *A two-dimensional coupled model for ice shelf–ocean interaction*, Ocean Modelling, 17 (2007), pp. 123–139.

[154] D. C. Wan, B. S. V. Patnaik, and G. W. Wei, *A new benchmark quality solution for the buoyancy-driven cavity by discrete singular convolution*, Numerical Heat Transfer, 40 (2001), pp. 199–228.

[155] S. Wang, A. Faghri, and T. L. Bergman, *A comprehensive numerical model for melting with natural convection*, International Journal of Heat and Mass Transfer, 53 (2010), pp. 1986–2000.

[156] T. P. Wihler, *Locking-free adaptive discontinuous Galerkin FEM for linear elasticity problems*, Mathematics of Computation, 75 (2006), pp. 1087–1102.

[157] K. Wittig and P. Nikrityuk, *Three-dimensionality of fluid flow in the benchmark experiment for a pure metal melting on a vertical wall*, IOP Conference Series Materials Science and Engineering, 27 (2012), p. 012054.

[158] J. Woodfield, M. Alvarez, B. Gómez-Vargas, and R. Ruiz-Baier, *Stability and finite element approximation of phase change models for natural convection in porous media*, Journal of Computational and Applied Mathematics, 360 (2019), pp. 117–137.

[159] F. Xuan, S. S. Shao, Z. Wang, and S. T. Tu, *Coupling effects of chemical stresses and external mechanical stresses on diffusion*, Journal of Physics D: Applied Physics, 42 (2009), p. 015401.

[160] S.-Y. Yi, *A coupling of nonconforming and mixed finite element methods for Biot's consolidation model*, Numerical Methods for Partial Differential Equations, 29 (2013), pp. 1749–1777.

[161] ——, *Convergence analysis of a new mixed finite element method for Biot's consolidation model*, Numerical Methods for Partial Differential Equations, 30 (2014), pp. 1189–1210.

[162] F. G. Yost, D. E. Amos, and A. D. Roming, *Stress-driven diffusive voiding of aluminum conductor lines*, Proceedings International Reliability Physics Symposium, (1989), pp. 193–201.

[163] X. Zhang, A. M. Sastry, and W. Shyy, *Intercalation-induced stress and heat generation within single lithium-ion battery cathode particles*, Journal of Electrochemical Society, 155 (2008), pp. A542–A552.

[164] Y. Zhang, Y. Hou, and H. Jia, *Subgrid stabilized defect-correction method for a steady-state natural convection problem*, Computers and Mathematics with Applications, 67 (2014), pp. 497–514.

[165] A. G. Zimmerman and J. Kowalski, *Monolithic simulation of convection-coupled phase-change - verification and reproducibility*, Schäfer M., Behr M., Mehl M., Wohlmuth B. (eds) Recent Advances in Computational Engineering. ICCE 2017. Lecture Notes in Computational Science and Engineering, vol 124. Springer, Cham, (2018).