The Dissertation Committee for Federico Fuentes
certifies that this is the approved version of the following dissertation:

Various applications of discontinuous Petrov-Galerkin (DPG) finite element methods

Committee:

Leszek F. Demkowicz, Supervisor

Ivo M. Babuška

Luis A. Caffarelli

Thomas J. R. Hughes

J. Tinsley Oden

Aleta Wilder

Various applications of discontinuous Petrov-Galerkin (DPG) finite element methods

by

Federico Fuentes

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2018

Acknowledgments

First and foremost, I wish to thank my supervisor, Leszek Demkowicz. An almost paternal figure, he was always supportive of both my academic and personal goals. He provided a responsible amount of freedom when I needed it and gave me many opportunities to present my research worldwide as well as numerous chances to visit my family overseas. More importantly, his teaching style and seemingly unbounded knowledge of the fundamental mathematics underlying finite element methods were inspiring and incredibly useful. His appreciation and mastery of detail, which is unusual for a professor of his age, combined with his ability to not lose the big picture, will always be a guiding example for the years to come. Lastly, his academic advice, plagued with fascinating and memorable historical annotations, will surely be missed (although I hope it continues).

Next, I wish to thank the Institute for Computational Engineering and Sciences (ICES). The quality of all the faculty and its distinctive interdisciplinary research in computational engineering and mathematics is certainly top notch. I will always be honored to identify it as my alma mater. I appreciated the open research environment and the philosophy of unfiltered questioning and interruptions during seminars, which is so important in a university. During my time at ICES I was privileged to interact with numerous reputed professors and researchers, and was given the chance (by Professor Babuška) of being the co-host of the Babuška Forum, which I was glad to accept. I hope that the institute, pioneered by Professor Oden, continues along those same lines in the coming future. I wish to thank Gregory Rodin for pushing for me when I applied to ICES. Since I have always enjoyed teaching, I also thank Robert Moser, Todd Arbogast and my supervisor for allowing me to be their (at least part-time) teaching assistant. The University of Texas at Austin also provided many opportunities and support, in particular in improving my teaching skills and in giving a platform and facilities to play sports, and I wish to thank them for that.

To my dissertation committee, which looks like *the* Hall of Fame of finite element methods, I wish to express my admiration and to thank you for being my professors and mentors. Your advice throughout my doctoral experience was always appreciated and I was fortunate to receive it. The funding bodies that generously financed me and made the research possible also deserve acknowledgment. These include ICES, the Air Force Office of Scientific Research (AFOSR), the National Science Foundation (NSF), and the Office of Naval Research (ONR).

The summer I spent at the Woods Hole Oceanographic Institute (WHOI) as a Geophysical Fluid Dynamics (GFD) fellow was also instrumental in my growth as a researcher, and I wish to acknowledge them for that opportunity. In particular, I thank David Goluskin who supervised my work there. Moreover, I graciously thank my MSc supervisor, Sergei Chernyshenko, for his continued support and guidance after I graduated from Imperial College London.

My strong scientific background is the product of a liberal and complete education that started all the way back in The English School in Bogotá, Colombia, and continued at the Universidad de los Andes with bachelor's degrees in both mechanical engineering and mathematics. I wish to thank my (good) teachers for making that possible, most notably Óscar Moreno during high school and Monika Winklmeier and Álvaro Pinilla during my undergraduate years.

I am also glad to have had the fortune of working with great collaborators and to have been accompanied by good friends during my time in Austin: Sriram Nagaraj, Brendan Keith, Socratis Petrides, Jaime Mora, Ali Vaziri, René Hiemstra, Jane Garcés, Marisol Rodríguez, among others. In addition, I wish to thank all the different generations of soccer/football teams that I have ever played with. Of particular pride during my time in Austin was reaching the intramural final with the "ACES" team. I am happy to count many of my fellow players among my friends, and I think it would have been impossible without such a great sport to unite us. To my closest friends, I wish to thank you for your open sense of humor and appreciation for sarcasm and cynicism.

Last, but not least, I wish to thank my extended family and lifelong caretaker for their unwavering love and support throughout my entire life. Naturally, I wish to express special gratitude to my mother, my father and my brother, who have always been there for me. My parents provided a platform of free thought and inquiry since I was born, which is the hallmark of academic life and the starting point for the appreciation of the staunch pursuit of truth above all else. Without that, all this would have been impossible. Thank you.

Various applications of discontinuous Petrov-Galerkin (DPG) finite element methods

by

Federico Fuentes
The University of Texas at Austin, 2018

Supervisor: Leszek F. Demkowicz

Discontinuous Petrov-Galerkin (DPG) finite element methods have garnered significant attention since they were originally introduced. They discretize variational formulations with broken (discontinuous) test spaces and are crafted to be numerically stable by implicitly computing a near-optimal discrete test space as a function of a discrete trial space. Moreover, they are completely general in the sense that they can be applied to a variety of variational formulations, including non-conventional ones that involve non-symmetric functional settings, such as ultraweak variational formulations. In most cases, these properties have been harnessed to develop numerical methods that provide robust control of relevant equation parameters, like in convection-diffusion problems and other singularly perturbed problems.

In this work, other features of DPG methods are systematically exploited and applied to different problems. More specifically, the versatility of DPG methods is elucidated by utilizing the underlying methodology to discretize four distinct variational formulations of the equations of linear elasticity. By taking advantage of interface variables inherent to DPG discretizations, an approach to coupling different variational formulations within the same domain is described and used to solve interesting problems. Moreover, the convenient algebraic structure in DPG methods is harnessed to develop a new family of numerical methods called discrete least-squares (DLS) finite element methods. These involve solving, with improved conditioning properties, a discrete least-squares problem associated with an overdetermined rectangular system of equations, instead of directly solving the usual square systems. Their utility is demonstrated with illustrative examples. Additionally, high-order polygonal DPG (PolyDPG) methods are devised by using the intrinsic discontinuities present in ultraweak formulations. The resulting methods can handle heavily distorted non-convex polygonal elements and discontinuous material properties. A polygonal adaptive strategy was also proposed and compared with standard techniques. Lastly, the natural high-order residual-based a posteriori error estimator ingrained within DPG methods was further applied to problems of physical relevance, like the validation of dynamic mechanical analysis (DMA) calibration experiments of viscoelastic materials, and the modeling of form-wound medium-voltage stator coils sitting inside large electric machinery.

Table of Contents

Chapter 1

Introduction

1.1 Discontinuous Petrov-Galerkin (DPG) methods

The discontinuous Petrov-Galerkin (DPG) methodology is a finite element technique to solve differential equations which has many attractive properties [95, 55]. The most important one is its inherent capacity to craft as much numerical stability as allowed into the discretization of the underlying well-posed variational formulation associated to a given linear partial differential equation (PDE). Due to its solid mathematical structure, it has been used to solve many challenging problems, especially those involving numerical stability issues where the robust control of relevant equation parameters is crucial, like convection-diffusion problems [75, 102, 60, 184, 43, 44, 42] and other singularly perturbed problems [183, 147, 121, 118]. It is also very convenient to couple, within the same domain, with other numerical methods, such as boundary element methods and other finite element methods [113, 145, 117, 119, 146, 120]. Moreover, it has been applied to various physical problems such as wave propagation [237, 132, 96, 194], electromagnetism [55, 230], elasticity [158, 36, 113, 111, 56], fluid flow [202, 58, 109, 159] and optical fibers via Schrödinger's equation [97].

The DPG methodology was first devised by Demkowicz and Gopalakrishnan [92, 93, 98, 237]. It starts with a variational formulation of a linear PDE, which, as usual, has an infinite-dimensional trial space, where the solution of the PDE is sought, and an infinite-dimensional test space, where test functions lie. The formulation is assumed to be well-posed, so its bilinear form has a strictly positive inf-sup constant. Equivalently, when viewed in operator form (where the domain is the trial space), the associated operator is bounded below. Then, if the trial space is discretized, the idea is to find the element of the discrete trial space which minimizes the operator residual of the variational formulation. This can be shown to be equivalent to a Petrov-Galerkin method where the discrete test space is a so-called optimal test space, since it guarantees the discretized variational formulation has the best possible discrete inf-sup constant, thus ensuring numerical stability. Finding the optimal test space (or minimizing the residual) involves inverting the Riesz map of the test space, which is typically impossible due to its infinite-dimensional nature. Thus, in practice the test space is discretized separately into an enriched test space, which one can then use to compute a near-optimal test space instead. These methods are usually referred to as minimum residual finite element methods. In general, the (discrete) inversion of the Riesz map has to be performed at a global level, and this is extremely expensive, but if the Riesz map is somehow localized to each element, the associated minimum residual method becomes much cheaper and the overall method becomes useful from a practical standpoint. This is what happens when the test spaces are “broken” or discontinuous across elements. Variational formulations whose test spaces are broken are referred to as broken variational formulations, and they usually include extra “interface” variables in the trial space as well. The DPG methodology is the special case of applying minimum residual finite element methods to broken variational formulations.

Thus, the DPG methodology is able to craft numerical stability in such a way that it decouples the discretization process of the trial and test spaces: first the trial space is discretized, and then an independently tunable enriched test space is proposed (from which a near-optimal test space is computed). As the enriched test space is made larger, the resulting discrete inf-sup constant is designed to approach the best possible discrete inf-sup constant, and mathematically this can be analyzed through the introduction of a Fortin operator [133]. This philosophy of providing stability seems to be quite flexible and comes in contrast with the balancing act that arises in traditional mixed methods, where the discretization of distinct variables has to be considered simultaneously and carefully analyzed to establish the existence of numerical stability [38]. On a separate note, DPG methods are conforming finite element methods (the discrete trial and test spaces are subspaces of their infinite-dimensional counterparts). This, combined with their inherent numerical stability, means that DPG methods do not need to include ad hoc stabilization terms, as other methods do, and it also means that the mathematical analysis of the convergence of the methods is often simplified. Moreover, there is nothing that forbids the use of piecewise high-order polynomials as basis functions for the trial and test spaces. In fact, high-order DPG methods are the norm rather than the exception.

Besides the numerical stability and its consequences, DPG methods have several other assets. First, the DPG methodology is extremely general because it applies to any well-posed broken variational formulation. Moreover, the discretization possesses a rich algebraic structure. In fact, the associated stiffness matrices are always symmetric (or Hermitian) positive definite. Lastly, it carries a natural residual-based high-order a posteriori error estimator which is ideal to implement adaptivity. On the downside, it typically comes at a higher computational cost when compared to other methods.

These attractive properties have been harnessed to open new avenues of research in numerical methods and functional analysis. Indeed, the versatility with which DPG methods can be applied to different variational formulations has permitted the development of numerical methods that discretize non-conventional formulations involving non-symmetric functional settings, like ultraweak variational formulations. These formulations have many attractive properties which have been taken advantage of in several different contexts, including robust error control in singularly perturbed problems, superconvergence, and polygonal element methods [95, 96, 201, 46, 102, 43, 147, 60, 86, 116, 115, 229]. The interface variables that are often present in broken variational formulations have also facilitated the coupling of DPG methods with other numerical methods [113, 145, 117, 119, 146, 120]. In fact, the study of the broken variational formulations themselves has produced new interesting theoretical results [55, 97]. Moreover, the methodology's algebraic structure has been exploited theoretically and computationally, leading to novel numerical methods [160]. Lastly, the study of error estimation in DPG methods has generated fresh mathematical ideas, including the development of new numerical methods and goal-oriented a posteriori error estimators [54, 161, 157].

1.2 Goal

Most applications of DPG methods have been associated with providing robust control of relevant equation parameters. Examples include convection-diffusion problems and other singularly perturbed problems. The goal of this work is to further the study of DPG methods, from both the mathematical and engineering perspectives, by systematically exploiting other features of DPG methods and applying them to different problems. Some more specific goals are:

- To show the generality of the approach offered by the DPG methodology. In particular, to manifest the capacity to develop effective DPG methods for a variety of different variational formulations.
- To explain some of the ideas behind coupling DPG methods with other numerical methods and their potential applications.
- To apply DPG methods to real-world problems. This includes their use in calibration experiments of viscoelastic materials, and in the study of resins in form-wound medium-voltage stator coils sitting inside large electric machinery.
- To exploit the algebraic properties of DPG methods in order to design new discretization techniques and computational optimizations.
- To capitalize on the properties of ultraweak formulations to allow the development of polygonal DPG (PolyDPG) methods, which are useful for many engineering problems.

1.3 Outline

This dissertation is organized in three parts. The first part aims at describing DPG methods and serves as a prelude for the following chapters in terms of notation and main concepts. This is the content of Chapter 2. The reading of this chapter is aided by some knowledge in Sobolev spaces, and for this reason an auxiliary chapter introducing the basics of these spaces is provided as Appendix A.

The second part of the dissertation is about applications of DPG methods to linear elasticity, viscoelasticity and thermoviscoelasticity. Chapter 3 has the dual role of showing the flexibility of the DPG methodology in discretizing different variational formulations, and of explaining how to couple these different DPG formulations within the same domain. Its content is the product of two publications [158, 113]. The chapter is accompanied by Appendix B, which is a mathematical contribution proving the mutual well-posedness of the different variational formulations. Chapter 4 applies DPG methods to the problem of dynamic mechanical analysis (DMA) calibration experiments of viscoelastic polymers, which was part of one publication [111]. Meanwhile, Chapter 5 analyzes basic scenarios relevant to form-wound medium-voltage stator coils in electric machinery. These involve the linear viscoelasticity and thermoviscoelasticity equations, whose derivations from first principles are presented in Appendix C.

The third and final part of this work aims at studying other numerical methods that can be thought of as outgrowths of the DPG methodology. In Chapter 6 discrete least-squares (DLS) finite element methods are presented, and their utility is shown. This chapter was the subject of a publication of the same name [160]. Chapter 7 describes polygonal DPG (PolyDPG) methods and provides several relevant examples. The details of the proof of convergence are given in Appendix D. This chapter was the product of another publication [229].

Lastly, the overarching conclusions of the dissertation are presented in Chapter 8.

Chapter 2

Discontinuous Petrov-Galerkin (DPG) methods

2.1 Introduction

A rough overview of DPG methods was already given in Section 1.1. The aim of this chapter is to provide a more detailed account describing the DPG methodology from a mathematical and computational perspective. It serves as a preliminary to the rest of the dissertation. Note that with the exception of the variable names, the general notation will remain the same throughout this dissertation, so this chapter also serves the function of introducing such notation. As a driving example, Poisson's equation is analyzed. Some familiarity with functional analysis and finite element methods is recommended. Moreover, some knowledge of Sobolev spaces is also advised, including $H^1(K)$, $\boldsymbol{H}(\mathrm{div},K)$, $L^2(K)$, their relevant traces, and their mesh-broken (piecewise discontinuous) counterparts along with the respective interface spaces. For this reason Appendix A provides the basics of these Sobolev spaces and the interpolation properties of their discretizations.

2.2 Model problem and variational formulations

The goal of this section is to present a classical variational formulation of Poisson's equation, followed by a broken variational formulation of the same equation. It will be shown that if the classical formulation is well-posed in the sense of Hadamard (a unique solution satisfying a stability estimate exists), then the broken formulation is also well-posed. Broken variational formulations are suitable for discretization via the DPG methodology.

As a model problem, consider Poisson's equation in a Lipschitz domain $\Omega \subseteq \mathbb{R}^{n_d}$, with n_d being the number of spatial dimensions,

$$-\,\mathrm{div}(\nabla u) = r\,, \qquad \Leftrightarrow \qquad \begin{cases} \quad \mathrm{div}\,\boldsymbol{q} = r\,, \\ \boldsymbol{q} + \nabla u = 0\,, \end{cases} \tag{2.1}$$

where u is the temperature, $\boldsymbol{q}$ is the negative temperature gradient, and r is a source. Note that the equation can be written directly as a second order system (left) or as a first-order system (right). For simplicity, we assume temperature boundary conditions along all of $\partial\Omega$, so that $u = g$ at $\partial\Omega$, where g is a known function.

To solve the equation using finite element methods, a variational form is required, and in this respect, there are many possibilities. For now assume vanishing temperature boundary conditions so that $g = 0$. The classical approach stems directly from the second order equation by multiplying by a test function and integrating by parts once, leading to the primal formulation where the solution u is sought in the trial space $\mathcal{U}_0$ and must satisfy

$$\begin{aligned} b_0(u,v) &= \ell(v) \qquad && \forall v \in \mathcal{V}_0 = \mathcal{U}_0 = H_0^1(\Omega)\,, \\ b_0(u,v) &= (\nabla u, \nabla v)_\Omega\,, && \ell(v) = \langle r, v\rangle_{(H_0^1(\Omega))'\times H_0^1(\Omega)}\,. \end{aligned} \tag{2.2}$$

Here, $r \in (H_0^1(\Omega))'$, $\langle\,\cdot\,,\,\cdot\,\rangle_{(H_0^1(\Omega))'\times H_0^1(\Omega)}$ is a usual duality pairing, and the L^2 inner product in a domain $K \subseteq \Omega$ is defined as

$$(u,v)_K = \int_K \mathrm{tr}_{\mathbb{M}}(v^{\mathsf{T}}u)\,\mathrm{d}K\,, \tag{2.3}$$

where $\mathrm{tr}_{\mathbb{M}}$ is the usual algebraic trace of a matrix, so that depending on whether u and v take scalar, vector or matrix values, $\mathrm{tr}_{\mathbb{M}}(v^{\mathsf{T}}u)$ will be uv, $u\cdot v$ or $u : v$, respectively. Notice in this case the trial and test spaces are equal ($\mathcal{U}_0 = \mathcal{V}_0$), so both spaces can be discretized in the same way, leading to the Bubnov-Galerkin method (see next section), and the same is true for standard mixed formulations which stem from the first-order system. The primal formulation in (2.2) is known to be coercive and well-posed in view of the Lax-Milgram theorem and Poincaré's inequality.

The practicality of the DPG methodology relies on using broken (or discontinuous) test spaces, and this results in a slightly modified formulation called the *broken* primal formulation, which will be derived next. Consider a mesh (i.e. an open partition), $\mathcal{T}$, of Ω comprised of (disjoint) elements $K \in \mathcal{T}$, and recall the broken space $H^1(\mathcal{T}) = \{v \in L^2(\Omega) \mid v|_K \in H^1(K)\,, \forall K \in \mathcal{T}\}$, and the $L^2(\mathcal{T})$ piecewise integration,

$$(u,v)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} (u|_K, v|_K)_K\,. \tag{2.4}$$

Then, element-wise, multiply Poisson's equation by broken test functions $v = \mathfrak{v} \in \mathcal{V} = H^1(\mathcal{T})$, integrate by parts, and sum across all elements. The (informal) result is very similar to the primal formulation,

$$\sum_{K\in\mathcal{T}} (-\operatorname{div}(\nabla u), v)_K = \sum_{K\in\mathcal{T}} (\nabla u, \nabla v)_K - \sum_{K\in\mathcal{T}} \langle \nabla u|_{\partial K} \cdot \hat{\mathbf{n}}_K, v|_{\partial K}\rangle_{\partial K}\,, \tag{2.5}$$

but has new terms on the boundaries of the elements involving $\nabla u|_{\partial K} \cdot \hat{\mathbf{n}}_K$, where $\hat{\mathbf{n}}_K$ is the outward normal to the element K. These terms vanish if the test space is not broken (i.e. $v \in \mathcal{V}_0 = H^1(\Omega)$). Unfortunately, if we want $u \in H^1(\Omega)$, then $\nabla u \in \boldsymbol{L}^2(\Omega) = \big(L^2(\Omega)\big)^{n_d}$, so the traces $\nabla u|_{\partial K} \cdot \hat{\mathbf{n}}_K$ might not exist [170] and to incorporate them it is necessary to add a new interface variable.

With this in mind, consider the mesh-trace of a variable in $\boldsymbol{H}(\operatorname{div}, \Omega)$, called $\hat{q}_{\mathbf{n}}$ (a heat flux), which is supposed to replace the $\mathcal{T}$-tuple $(-\nabla u)|_K\big|_{\partial K} \cdot \hat{\mathbf{n}}_K$, so that

$$\begin{aligned} \hat{q}_{\mathbf{n}} \in H^{-1/2}(\partial\mathcal{T}) = \operatorname{tr}^{\mathcal{T}}_{\text{div}}\big(\boldsymbol{H}(\operatorname{div}, \Omega)\big)\,, \\ \operatorname{tr}^{\mathcal{T}}_{\text{div}} \boldsymbol{q} = \textstyle\prod_{K\in\mathcal{T}} (\boldsymbol{q}|_K)\big|_{\partial K} \cdot \hat{\mathbf{n}}_K\,, \qquad \forall \boldsymbol{q} \in \boldsymbol{H}(\operatorname{div}, \mathcal{T})\,, \end{aligned} \tag{2.6}$$

where $\boldsymbol{H}(\operatorname{div}, \mathcal{T}) = \big\{\boldsymbol{q} \in \big(L^2(\Omega)\big)^{n_d} \mid \boldsymbol{q}|_K \in \boldsymbol{H}(\operatorname{div}, K)\,, \forall K \in \mathcal{T}\big\}$. Similarly,

$$\operatorname{tr}^{\mathcal{T}}_{\text{grad}} v = \textstyle\prod_{K\in\mathcal{T}} (v|_K)\big|_{\partial K}\,, \qquad \forall v \in H^1(\mathcal{T})\,. \tag{2.7}$$

Both $\operatorname{tr}^{\mathcal{T}}_{\text{div}} \boldsymbol{q}$ and $\operatorname{tr}^{\mathcal{T}}_{\text{grad}} v$ are $\mathcal{T}$-tuples indexed by $K \in \mathcal{T}$. The mesh inner product is

$$\langle \hat{u}, \hat{q}_{\mathbf{n}}\rangle_{\partial\mathcal{T}} = \sum_{K\in\mathcal{T}} \langle (\hat{u})_K, (\hat{q}_{\mathbf{n}})_K\rangle_{\partial K}\,, \tag{2.8}$$

where $\hat{u} \in \operatorname{tr}^{\mathcal{T}}_{\text{grad}}\big(H^1(\mathcal{T})\big)$, $\hat{q}_{\mathbf{n}} \in \operatorname{tr}^{\mathcal{T}}_{\text{div}}\big(H(\operatorname{div}, \mathcal{T})\big)$, and $\langle\,\cdot\,,\,\cdot\,\rangle_{\partial K}$ is the $H^{1/2}(\partial K) \times H^{-1/2}(\partial K)$ duality pairing. This duality pairing can be thought of as a boundary integral (for smooth enough inputs it is actually a boundary integral). Without loss of generality, we will use the same notation if the inputs are switched, i.e. $\langle \hat{q}_{\mathbf{n}}, \hat{u}\rangle_{\partial\mathcal{T}} = \langle \hat{u}, \hat{q}_{\mathbf{n}}\rangle_{\partial\mathcal{T}}$ (and $\langle\,\cdot\,,\,\cdot\,\rangle_{\partial K}$ will denote the $H^{-1/2}(\partial K) \times H^{1/2}(\partial K)$ duality pairing instead), so be aware of the context. For more information on these spaces consult Appendix A.

The resulting broken primal variational formulation seeks

$$\begin{aligned} (\mathfrak{u}_0, \hat{\mathfrak{u}}) = \mathfrak{u} \in \mathcal{U} = \mathcal{U}_0 \times \hat{\mathcal{U}}\,, \\ u = \mathfrak{u}_0 \in \mathcal{U}_0 = H^1_0(\Omega)\,, \qquad \hat{q}_{\mathbf{n}} = \hat{\mathfrak{u}} \in \hat{\mathcal{U}} = H^{-1/2}(\partial\mathcal{T})\,, \end{aligned} \tag{2.9}$$

such that

$$\begin{aligned}
&b(\mathfrak{u},\mathfrak{v}) = \ell(\mathfrak{v}) \qquad \forall v = \mathfrak{v} \in \mathcal{V} = H^1(\mathcal{T})\,,\\
&b\big((\mathfrak{u}_0,\hat{\mathfrak{u}}),\mathfrak{v}\big) = b_0(\mathfrak{u}_0,\mathfrak{v}) + \hat{b}(\hat{\mathfrak{u}},\mathfrak{v})\,, \qquad \ell(v) = \langle r, v\rangle_{(H^1(\mathcal{T}))'\times H^1(\mathcal{T})}\,,\\
&b_0(u,v) = (\nabla u, \nabla v)_{\mathcal{T}}\,,\\
&\hat{b}(\hat{q}_{\mathbf{n}}, v) = \langle \hat{q}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} v\rangle_{\partial\mathcal{T}}\,,
\end{aligned} \tag{2.10}$$

where $r \in (H^1(\mathcal{T}))'$. Note that $\mathcal{U} \neq \mathcal{V}$. To prove well-posedness of the broken primal formulation, a powerful theorem proved in [55, Theorem 3.1] is utilized.

Theorem 2.1. *Let $\mathcal{U}_0$, $\hat{\mathcal{U}}$ and $\mathcal{V}$ be Hilbert spaces over a fixed field $\mathbb{F} \in \{\mathbb{R},\mathbb{C}\}$. Let $\ell : \mathcal{V} \to \mathbb{F}$ be a continuous linear form, and let $b_0 : \mathcal{U}_0 \times \mathcal{V} \to \mathbb{F}$ and $\hat{b} : \hat{\mathcal{U}} \times \mathcal{V} \to \mathbb{F}$ be continuous bilinear forms if $\mathbb{F} = \mathbb{R}$ or sesquilinear forms if $\mathbb{F} = \mathbb{C}$. With $\mathcal{U} = \mathcal{U}_0 \times \hat{\mathcal{U}}$ and $\|\cdot\|^2_{\mathcal{U}} = \|\cdot\|^2_{\mathcal{U}_0} + \|\cdot\|^2_{\hat{\mathcal{U}}}$, define $b : \mathcal{U} \times \mathcal{V} \to \mathbb{F}$ for all $(\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathcal{U}$ and $\mathfrak{v} \in \mathcal{V}$ by*

$$b\big((\mathfrak{u}_0,\hat{\mathfrak{u}}),\mathfrak{v}\big) = b_0(\mathfrak{u}_0,\mathfrak{v}) + \hat{b}(\hat{\mathfrak{u}},\mathfrak{v})\,, \tag{2.11}$$

and let

$$\mathcal{V}_0 = \{\mathfrak{v} \in \mathcal{V} \mid \hat{b}(\hat{\mathfrak{u}},\mathfrak{v}) = 0,\ \forall \hat{\mathfrak{u}} \in \hat{\mathcal{U}}\}\,. \tag{2.12}$$

Assume:

(γ_0) *There exists $\gamma_0 > 0$ such that for all $\mathfrak{u}_0 \in \mathcal{U}_0$,*

$$\sup_{\mathfrak{v}_0 \in \mathcal{V}_0 \setminus \{0\}} \frac{|b_0(\mathfrak{u}_0,\mathfrak{v}_0)|}{\|\mathfrak{v}_0\|_{\mathcal{V}}} \geq \gamma_0 \|\mathfrak{u}_0\|_{\mathcal{U}_0}\,. \tag{2.13}$$

($\hat{\gamma}$) *There exists $\hat{\gamma} > 0$ such that for all $\hat{\mathfrak{u}} \in \hat{\mathcal{U}}$,*

$$\sup_{\mathfrak{v} \in \mathcal{V} \setminus \{0\}} \frac{|\hat{b}(\hat{\mathfrak{u}},\mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}}} \geq \hat{\gamma} \|\hat{\mathfrak{u}}\|_{\hat{\mathcal{U}}}\,. \tag{2.14}$$

Then:

(γ) *There exists $\gamma = \big(\frac{1}{\gamma_0^2} + \frac{1}{\hat{\gamma}^2}(\frac{M_0}{\gamma_0}+1)^2\big)^{-\frac{1}{2}} > 0$ such that for all $(\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathcal{U}$,*

$$\sup_{\mathfrak{v} \in \mathcal{V} \setminus \{0\}} \frac{|b\big((\mathfrak{u}_0,\hat{\mathfrak{u}}),\mathfrak{v}\big)|}{\|\mathfrak{v}\|_{\mathcal{V}}} \geq \gamma \|(\mathfrak{u}_0,\hat{\mathfrak{u}})\|_{\mathcal{U}}\,, \tag{2.15}$$

where $M_0 \geq \|b_0\| = \sup_{(\mathfrak{u}_0,\mathfrak{v}) \in \mathcal{U}_0 \times \mathcal{V} \setminus \{(0,0)\}} \frac{|b_0(\mathfrak{u}_0,\mathfrak{v})|}{\|\mathfrak{u}_0\|_{\mathcal{U}_0}\|\mathfrak{v}\|_{\mathcal{V}}}$.

Moreover, if ℓ satisfies the compatibility condition,

$$\ell(\mathfrak{v}) = 0 \qquad \forall \mathfrak{v} \in \mathcal{V}_{00}\,, \tag{2.16}$$

where

$$\mathcal{V}_{00} = \{\mathfrak{v}_0 \in \mathcal{V}_0 \mid b_0(\mathfrak{u}_0, \mathfrak{v}_0) = 0 \;\; \forall \mathfrak{u}_0 \in \mathcal{U}_0\}\,, \tag{2.17}$$

which is always true if $\mathcal{V}_{00} = \{0\}$, then the problem of finding $(\mathfrak{u}_0, \hat{\mathfrak{u}}) \in \mathcal{U}$ such that

$$b\big((\mathfrak{u}_0, \hat{\mathfrak{u}}), \mathfrak{v}\big) = \ell(\mathfrak{v}) \qquad \forall \mathfrak{v} \in \mathcal{V}\,, \tag{2.18}$$

has a unique solution $(\mathfrak{u}_0, \hat{\mathfrak{u}}) \in \mathcal{U}$ satisfying the estimate

$$\|(\mathfrak{u}_0, \hat{\mathfrak{u}})\|_{\mathcal{U}} \leq \frac{1}{\gamma}\|\ell\|_{\mathcal{V}'}\,. \tag{2.19}$$

Furthermore, the component $\mathfrak{u}_0$ from the unique solution is also the unique solution to the problem that seeks $\mathfrak{u}_0 \in \mathcal{U}_0$ such that

$$b_0(\mathfrak{u}_0, \mathfrak{v}_0) = \ell(\mathfrak{v}_0) \qquad \forall \mathfrak{v}_0 \in \mathcal{V}_0\,. \tag{2.20}$$

Therefore, it suffices to prove (γ_0) and $(\hat{\gamma})$ for (2.10). By Theorem A.1, it follows that

$$\mathcal{V}_0 = \{v \in H^1(\mathcal{T}) \mid \langle \hat{q}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} v\rangle_{\partial\mathcal{T}} = 0, \; \forall\, \hat{q}_{\mathbf{n}} \in H^{-1/2}(\partial\mathcal{T})\} = H^1_0(\Omega)\,. \tag{2.21}$$

Thus, (γ_0) is satisfied, because this is equivalent to the well-posedness of the original classical formulation (simply use Poincaré's inequality). Meanwhile, $(\hat{\gamma})$ follows directly from Theorem A.3 with $\hat{\gamma} = 1$, and

$$\mathcal{V}_{00} = \{v \in H^1_0(\Omega) \mid (\nabla u, \nabla v)_{\mathcal{T}} = 0, \; \forall\, u \in H^1_0(\Omega)\} = \{v \in H^1_0(\Omega) \mid \nabla v = 0\} = \{0\}\,. \tag{2.22}$$

Thus, by Theorem 2.1, the broken primal formulation is well-posed, and in fact its stability properties are independent of the choice of the mesh ($\hat{\gamma} = 1$ and $M_0 = 1$ do not depend on the mesh).

With nontrivial temperature boundary conditions, $g \neq 0$, simply consider the new right hand side $\ell(\mathfrak{v}) = \langle r, v\rangle_{(H^1(\mathcal{T}))' \times H^1(\mathcal{T})} - (\nabla\widetilde{g}, \nabla v)_{\mathcal{T}}$ instead, where $\widetilde{g} \in H^1(\Omega)$ is an extension of $g \in H^{1/2}(\partial\Omega)$, and add $\widetilde{g}$ to the solution u of (2.10) to obtain the final temperature. When other flux boundary conditions are present, simply modify the spaces accordingly and make the analogous changes to ℓ.

2.3 Minimum residual methods and DPG discretizations

In this section we present the procedure of discretizing broken variational formulations. The basics of minimum residual finite element methods and the DPG methodology are presented mostly from the point of view of linear algebra, but more optional details are given in the next section. Assume the field associated to the variational formulation is $\mathbb{F} = \mathbb{R}$ (as in Poisson's equation). The generalization to $\mathbb{F} = \mathbb{C}$ is straightforward, but does involve minor technicalities which will be left for the reader to ponder.

The Bubnov-Galerkin method is the widely used approach for classical formulations, since it employs the same test and trial spaces, leading to a square linear system of equations. Indeed, consider the primal formulation in (2.2), with $\{\mathfrak{u}_{0,j}\}_{j=1}^{N}$ being a basis for the discrete subspaces $\mathcal{U}_{0,h} = \mathcal{V}_{0,h} \subseteq \mathcal{U}_0 = \mathcal{V}_0$. Then, the discrete solution $u_h = \sum_{j=1}^{N} (\mathsf{u}_h)_j \mathfrak{u}_{0,j} \in \mathcal{U}_{0,h}$ for $\mathsf{u}_h \in \mathbb{R}^N$, satisfies

$$\mathsf{B}^{\mathrm{BG}} \mathsf{u}_h = \mathsf{l}^{\mathrm{BG}} , \tag{2.23}$$

where $\mathsf{B}_{ij}^{\mathrm{BG}} = b_0(\mathfrak{u}_{0,j}, \mathfrak{v}_{0,i})$ and $\mathsf{l}_i^{\mathrm{BG}} = \ell(\mathfrak{v}_{0,i})$ with $\mathfrak{v}_{0,i} = \mathfrak{u}_{0,i}$, so that $\mathsf{B}^{\mathrm{BG}} \in \mathbb{R}^{N \times N}$ and $\mathsf{l}^{\mathrm{BG}} \in \mathbb{R}^N$. The basis functions, $\mathfrak{u}_{0,j}$, are chosen with a very small support not exceeding a few neighboring elements, resulting in a computationally practical method due to the sparse structure of B^{BG}.

In general, when the trial and test spaces are different, $\mathcal{U} \neq \mathcal{V}$, this approach is still possible but requires finding bases $\{\mathfrak{u}_j\}_{j=1}^{N}$ and $\{\mathfrak{v}_i\}_{i=1}^{N}$ for $\mathcal{U}_h \subseteq \mathcal{U}$ and $\mathcal{V}_h \subseteq \mathcal{V}$ respectively. However, two issues immediately arise. First, the canonical polynomial-based discrete basis of $\mathcal{V}_h \subseteq \mathcal{V}$ typically is not of size N (the same size of the basis for $\mathcal{U}_h$). Second, even if a nonstandard basis for $\mathcal{V}_h$ of the right size is found, the resulting numerical method could very well be unstable, meaning that the inf-sup inequality,

$$\inf_{\delta\mathfrak{u}_h \in \mathcal{U}_h \setminus \{0\}} \sup_{\mathfrak{v}_h \in \mathcal{V}_h \setminus \{0\}} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_h)}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}} \|\mathfrak{v}_h\|_{\mathcal{V}}} = \gamma_h > 0 , \tag{2.24}$$

might *not* hold. In fact, depending on the equation and mesh size, even the Bubnov-Galerkin method can be unstable. Minimum residual finite element methods overcome these two difficulties by design.

Let $\mathcal{U}'$ and $\mathcal{V}'$ be the continuous dual spaces to the Hilbert spaces $\mathcal{U}$ and $\mathcal{V}$ respectively, and define the linear operator $\mathcal{B} : \mathcal{U} \to \mathcal{V}'$ and its continuous transpose $\mathcal{B}' : \mathcal{V} \to \mathcal{U}'$ (modulo the evaluation map since $\mathcal{V} = \mathcal{V}''$) through duality pairings as

$$\langle \mathcal{B}\mathfrak{u}, \mathfrak{v} \rangle_{\mathcal{V}' \times \mathcal{V}} = b(\mathfrak{u}, \mathfrak{v}) = \langle \mathcal{B}'\mathfrak{v}, \mathfrak{u} \rangle_{\mathcal{U}' \times \mathcal{U}} \qquad \forall \mathfrak{u} \in \mathcal{U},\ \forall \mathfrak{v} \in \mathcal{V}\,. \tag{2.25}$$

Recall the Riesz map, $\mathcal{R}_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$, which is an isometric isomorphism between $\mathcal{V}$ and $\mathcal{V}'$, defined by duality as

$$\langle \mathcal{R}_{\mathcal{V}}\mathfrak{v}, \delta\mathfrak{v} \rangle_{\mathcal{V}' \times \mathcal{V}} = (\mathfrak{v}, \delta\mathfrak{v})_{\mathcal{V}} \qquad \forall \mathfrak{v}, \delta\mathfrak{v} \in \mathcal{V}\,, \tag{2.26}$$

where $(\,\cdot\,,\,\cdot\,)_{\mathcal{V}}$ is the inner product of $\mathcal{V}$. Then, for a discrete trial space $\mathcal{U}_h \subseteq \mathcal{U}$, ideal minimum residual methods seek the minimizer of the residual,

$$\begin{aligned} \mathfrak{u}_h^{\text{opt}} = \underset{\delta\mathfrak{u}_h \in \mathcal{U}_h}{\arg\min}\, \|\mathcal{B}\delta\mathfrak{u}_h - \ell\|_{\mathcal{V}'}^2 \quad &\Leftrightarrow \quad (\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_h^{\text{opt}}, \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h)_{\mathcal{V}} = (\mathcal{R}_{\mathcal{V}}^{-1}\ell, \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h)_{\mathcal{V}} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h \\ &\Leftrightarrow \quad \langle \mathcal{B}\mathfrak{u}_h^{\text{opt}}, \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h \rangle_{\mathcal{V}' \times \mathcal{V}} = \langle \ell, \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h \rangle_{\mathcal{V}' \times \mathcal{V}} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h \\ &\Leftrightarrow \quad b(\mathfrak{u}_h^{\text{opt}}, \mathfrak{v}^{\text{opt}}) = \ell(\mathfrak{v}^{\text{opt}}) \quad \forall \mathfrak{v}^{\text{opt}} \in \mathcal{V}^{\text{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h\,. \end{aligned} \tag{2.27}$$

Here, $\mathcal{V}^{\text{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h$ is called the optimal test space, because this exact choice of discrete test space automatically results in the best inf-sup stable discrete method satisfying (2.24), as will be shown in the next section. For this reason, ideal minimum residual methods are sometimes referred to as *the* optimal Petrov-Galerkin methods. Given an element of the basis for $\mathcal{U}_h$, $\mathfrak{u}_i \in \{\mathfrak{u}_j\}_{j=1}^N$, the corresponding optimal test function is $\mathfrak{v}_i^{\text{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_i$. With these choices the resulting matrix $\mathsf{B}_{ij}^{\text{opt}} = b(\mathfrak{u}_j, \mathfrak{v}_i^{\text{opt}})$, called the optimal stiffness matrix, is always symmetric positive definite.

Unfortunately, computing $\mathcal{R}_{\mathcal{V}}^{-1}$ is usually impossible since $\mathcal{V}$ is infinite-dimensional. Thus, minimum residual methods simply make a choice of an *enriched* test space $\mathcal{V}_r \subseteq \mathcal{V}$ over which the operator is inverted. The enriched test space must be chosen to be large enough, and the minimum requirement is that $M = \dim(\mathcal{V}_r) \geq \dim(\mathcal{U}_h) = N$. The advantage is that this enriched space may be discretized with a standard canonical polynomial-based basis, $\{\mathfrak{v}_i\}_{i=1}^M$, and ultimately the resulting *near*-optimal space is $\mathcal{V}_h = \mathcal{V}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h$ and its corresponding *near*-optimal basis is $\mathfrak{v}_i^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathfrak{u}_i$ for every $\mathfrak{u}_i \in \{\mathfrak{u}_j\}_{j=1}^N$. The resulting discrete method can be shown to be equivalent to the linear system,

$$\mathsf{B}^{\text{n-opt}}\mathsf{u}_h = \mathsf{B}^{\mathsf{T}}\mathsf{G}^{-1}\mathsf{B}\mathsf{u}_h = \mathsf{B}^{\mathsf{T}}\mathsf{G}^{-1}\mathsf{l} = \mathsf{l}^{\text{n-opt}}\,, \tag{2.28}$$

where $\mathfrak{u}_h = \sum_{j=1}^{N} (\mathsf{u}_h)_j \mathfrak{u}_j \in \mathcal{U}_h$ is the discrete solution; the Gram matrix $\mathsf{G}_{ij} = (\mathfrak{v}_i, \mathfrak{v}_j)_{\mathcal{V}}$ is a discretization of $\mathcal{R}_{\mathcal{V}_r}$; $\mathsf{B}_{ij} = b(\mathfrak{u}_j, \mathfrak{v}_i)$ and $\mathsf{l}_i = \ell(\mathfrak{v}_i)$ are called the enriched stiffness matrix and load; and $\mathsf{B}_{ij}^{\text{n-opt}} = b(\mathfrak{u}_j, \mathfrak{v}_i^{\text{n-opt}})$ and $\mathsf{l}_i^{\text{n-opt}} = \ell(\mathfrak{v}_i^{\text{n-opt}})$ are the near-optimal stiffness matrix and load. Clearly the enriched stiffness matrix is rectangular and tall, $\mathsf{B} \in \mathbb{R}^{M \times N}$ with $M \geq N$, while the near-optimal stiffness matrix is square and symmetric positive definite, $\mathsf{B}^{\text{n-opt}} \in \mathbb{R}^{N \times N}$. To implement, one has to form the Gram matrix ($\mathsf{G} \in \mathbb{R}^{M \times M}$), enriched stiffness matrix ($\mathsf{B} \in \mathbb{R}^{M \times N}$) and enriched load vector ($\mathsf{l} \in \mathbb{R}^{M}$) first; then proceed to calculate the near-optimal stiffness matrix ($\mathsf{B}^{\text{n-opt}} = \mathsf{B}^{\mathsf{T}} \mathsf{G}^{-1} \mathsf{B} \in \mathbb{R}^{N \times N}$) and near-optimal load vector ($\mathsf{l}^{\text{n-opt}} = \mathsf{B}^{\mathsf{T}} \mathsf{G}^{-1} \mathsf{l} \in \mathbb{R}^{N}$); and finally solve for the basis coefficients of the discrete solution ($\mathsf{u}_h \in \mathbb{R}^{N}$).

These methods are referred to as minimum residual finite element methods. All this derivation holds for any arbitrary linear variational formulation including the primal formulations in (2.2) and (2.10). The method is near-optimal in that it is designed to approximate the optimal method (with B^{opt}), so in principle it is not *known* to be stable, but in practice it typically is or can be made stable (if it is not stable simply enrich $\mathcal{V}_r$ even more so that $M \gg N$). In fact, the stability of the near-optimal method can rigorously be analyzed by constructing a Fortin operator, $\Pi_F : \mathcal{V} \to \mathcal{V}_r$ (see Section 2.8.1).

However, there are major differences between applying this method to the primal formulation in (2.2) and the broken primal formulation in (2.10). Namely, for the classical primal formulation the enriched (sparse) stiffness matrix, B, and the Gram matrix, G, are assembled globally first and then the near-optimal stiffness matrix, $\mathsf{B}^{\text{n-opt}}$, is computed using (2.28). This is very expensive, especially due to the inversion of G. Thus, despite many advantages, the method is not very practical. On the other hand, when using broken test spaces, as in the broken primal formulation, the matrix G has a disjoint diagonal block structure, where each block corresponds to one element. Hence, the Gram matrix can be inverted locally, allowing the local near-optimal stiffness matrices $\mathsf{B}_K^{\text{n-opt}}$ to be computed directly for each element $K \in \mathcal{T}$. This in turn allows $\mathsf{B}^{\text{n-opt}}$ to be assembled as in any other finite element method. Thus, using formulations with broken test spaces localizes the computations and parallelizes the assembly, and makes it a viable method from a practical standpoint. However, when compared to traditional finite element methods, the local computations are in general more expensive due to the extra operations involving enriched test functions. Note that the broken primal formulation in (2.10) has an enriched stiffness matrix with the structure,

$$\mathsf{B}=\left.\begin{bmatrix} \overbrace{\begin{matrix} \qquad | \qquad \\ \mathsf{B}_0 \\ | \end{matrix}}^{\{(\mathfrak{u}_0)_j\}_{j=1}^{N_0}} & \overbrace{\begin{matrix} \qquad | \qquad \\ \hat{\mathsf{B}} \\ | \end{matrix}}^{\{\hat{\mathfrak{u}}_j\}_{j=1}^{\hat{N}}} \end{bmatrix}\right\}\{\mathfrak{v}_i\}_{i=1}^{M} \tag{2.29}$$

where $(\mathsf{B}_0)_{ij} = b_0((\mathfrak{u}_0)_j, \mathfrak{v}_i)$ and $\hat{\mathsf{B}}_{ij} = \hat{b}(\hat{\mathfrak{u}}_j, \mathfrak{v}_i)$, with $\{\mathfrak{u}_j\}_{j=1}^{N} = \{((\mathfrak{u}_0)_j, 0)\}_{j=1}^{N_0} \cup \{(0, \hat{\mathfrak{u}}_j)\}_{j=1}^{\hat{N}}$ being the $\mathcal{U}_h$-basis, so that $N = N_0 + \hat{N}$.

In the literature, the application of minimum residual methods to broken variational formulations is referred to as the DPG methodology. The methodology is quite general as it can be applied to variational formulations other than the broken primal such as broken ultraweak or broken mixed formulations, as will be seen in Chapter 3. Each application case results in a different DPG method similar to how the Bubnov-Galerkin methodology can be applied to primal and mixed formulations (where $\mathcal{U}_h = \mathcal{V}_h$). They are conforming finite element methods, where the trial spaces usually have some form of continuity across the mesh, while the test spaces are broken.

Remark 2.1. Unless otherwise specified, the norm of $\mathcal{V}$, $\|\cdot\|_{\mathcal{V}}$, will be its natural Hilbert norm ($\|\cdot\|_{H^1(\mathcal{T})}$ if $\mathcal{V} = H^1(\mathcal{T})$, etc.). However, it should be noted that the choice of norm of $\mathcal{V}$, which enters the method through G, can be very important in DPG methods, as it can affect the convergence behavior (the residual being minimized is measured in the $\|\cdot\|_{\mathcal{V}'}$ norm). Indeed, when dealing with broken ultraweak formulations, another common choice has been to use the adjoint graph norm, which has some important properties. Having said that, different choices have been exploited in the literature to produce robust control of the error in some perturbation parameter [132, 102, 60, 43, 147, 117].

2.4 Equivalent reformulations

This section extends the previous section as it relates to minimum residual methods. Other equivalences are explored and the "optimal" labels are justified more rigorously. However, this section is not necessary to understand the rest of this work, so the reader may skip it if so desired.

As before, consider a variational formulation defined by a bilinear form $b : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ and linear form $\ell : \mathcal{V} \to \mathbb{R}$, where $\mathcal{U}$ and $\mathcal{V}$ are Hilbert trial and test spaces respectively. Using the operator $\mathcal{B}$ defined in (2.25), the formulation can be equivalently rewritten in operator form: it seeks a solution $\mathfrak{u} \in \mathcal{U}$ such that

$$b(\mathfrak{u}, \mathfrak{v}) = \ell(\mathfrak{v}) \qquad \forall \mathfrak{v} \in \mathcal{V} \qquad \Leftrightarrow \qquad \mathcal{B}\mathfrak{u} = \ell\,. \tag{2.30}$$

Now consider a discrete trial space $\mathcal{U}_h \subseteq \mathcal{U}$. By using the definitions of $\mathcal{B}$, $\mathcal{B}'$ and the Riesz operator $\mathcal{R}_{\mathcal{V}}$ given in (2.25) and (2.26), as in (2.27), it can be deduced that the following statements are equivalent:

$$\mathfrak{u}_h^{\mathrm{opt}} = \underset{\delta\mathfrak{u}_h \in \mathcal{U}_h}{\arg\min} \, \|\mathcal{B}\delta\mathfrak{u}_h - \ell\|_{\mathcal{V}'}^2\,, \tag{2.31}$$

$$b(\mathfrak{u}_h^{\mathrm{opt}}, \mathfrak{v}^{\mathrm{opt}}) = \ell(\mathfrak{v}^{\mathrm{opt}}) \quad \forall \mathfrak{v}^{\mathrm{opt}} \in \mathcal{V}^{\mathrm{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h\,, \tag{2.32}$$

$$\langle \mathcal{B}'\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_h^{\mathrm{opt}}, \delta\mathfrak{u}_h \rangle_{\mathcal{U}' \times \mathcal{U}} = \langle \mathcal{B}'\mathcal{R}_{\mathcal{V}}^{-1}\ell, \delta\mathfrak{u}_h \rangle_{\mathcal{U}' \times \mathcal{U}} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h\,. \tag{2.33}$$

Next, notice the last expression looks like a Schur complement, and proceed to write the equivalent (mixed) system both in operator form and variational form:

$$\begin{pmatrix} \mathcal{R}_{\mathcal{V}} & \mathcal{B} \\ \mathcal{B}' & 0 \end{pmatrix} \begin{pmatrix} \psi^{\mathrm{opt}} \\ \mathfrak{u}_h^{\mathrm{opt}} \end{pmatrix} = \begin{pmatrix} \ell \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{cases} (\psi^{\mathrm{opt}}, \mathfrak{v})_{\mathcal{V}} + b(\mathfrak{u}_h^{\mathrm{opt}}, \mathfrak{v}) = \ell(\mathfrak{v}) & \forall \mathfrak{v} \in \mathcal{V}\,, \\ b(\delta\mathfrak{u}_h, \psi^{\mathrm{opt}}) \qquad\qquad = 0 & \forall \delta\mathfrak{u}_h \in \mathcal{U}_h\,, \end{cases} \tag{2.34}$$

where $\psi^{\mathrm{opt}} \in \mathcal{V}$ is a new solution variable called the error representation function, since it represents the residual in $\mathcal{V}$, i.e. $\psi^{\mathrm{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}(\ell - \mathcal{B}\mathfrak{u}_h^{\mathrm{opt}})$. Thus, the equivalent mixed variational formulation has the same (infinite-dimensional) trial and test space, namely $\mathcal{U}^{\mathrm{mix}} = \mathcal{V}^{\mathrm{mix}} = \mathcal{V} \times \mathcal{U}_h$. To summarize: (2.31), (2.32), (2.33) and (2.34) are all equivalent statements. In particular, an ideal minimum residual method can be viewed as a Petrov-Galerkin method with an optimal test space, or like a mixed variational formulation with the same infinite-dimensional trial and test spaces.

Regarding the Petrov-Galerkin method in (2.32), its test space $\mathcal{V}^{\mathrm{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h$, as mentioned previously, is called the optimal test space. To justify this label, first fix the discrete trial space $\mathcal{U}_h$ and notice the discrete inf-sup constant has a fixed upper bound no matter which discrete test space $\mathcal{V}_h \subseteq \mathcal{V}$ is considered,

$$\gamma_h = \inf_{\delta\mathfrak{u}_h \in \mathcal{U}_h} \sup_{\mathfrak{v}_h \in \mathcal{V}_h} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_h)}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}} \|\mathfrak{v}_h\|_{\mathcal{V}}} \leq \inf_{\delta\mathfrak{u}_h \in \mathcal{U}_h} \sup_{\mathfrak{v} \in \mathcal{V}} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v})}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}} \|\mathfrak{v}\|_{\mathcal{V}}} = \gamma^{\mathrm{opt}}\,. \tag{2.35}$$

Note that throughout this document, in order to lighten the notation, the zero element is sometimes tacitly omitted from infima and suprema taken over vector spaces. With the aim of attaining that optimal bound, fix a $\delta\mathfrak{u}_h \in \mathcal{U}_h$ and notice that

$$\begin{aligned}\sup_{\mathfrak{v}\in\mathcal{V}} \frac{b(\delta\mathfrak{u}_h,\mathfrak{v})}{\|\mathfrak{v}\|_{\mathcal{V}}} &= \|\mathcal{B}\delta\mathfrak{u}_h\|_{\mathcal{V}'} = \|\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h\|_{\mathcal{V}} = \frac{(\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h,\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h)_{\mathcal{V}}}{\|\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h\|_{\mathcal{V}}}\\ &= \frac{b(\delta\mathfrak{u}_h,\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h)}{\|\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h\|_{\mathcal{V}}} \leq \sup_{\mathfrak{v}^{\mathrm{opt}}\in\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h} \frac{b(\delta\mathfrak{u}_h,\mathfrak{v}^{\mathrm{opt}})}{\|\mathfrak{v}^{\mathrm{opt}}\|_{\mathcal{V}}} \leq \sup_{\mathfrak{v}\in\mathcal{V}} \frac{b(\delta\mathfrak{u}_h,\mathfrak{v})}{\|\mathfrak{v}\|_{\mathcal{V}}}\,,\end{aligned}\tag{2.36}$$

so the supremum is attained by choosing the optimal test function $\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h$. Clearly, all inequalities above are in fact equalities. The result also holds for all $\delta\mathfrak{u}_h \in \mathcal{U}_h$, so taking the infimum at both sides of the last two expressions yields that $\gamma_h = \gamma^{\mathrm{opt}}$ if $\mathcal{V}_h = \mathcal{V}^{\mathrm{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h$. Moreover,

$$\gamma^{\mathrm{opt}} = \inf_{\delta\mathfrak{u}_h\in\mathcal{U}_h} \sup_{\mathfrak{v}^{\mathrm{opt}}\in\mathcal{V}^{\mathrm{opt}}} \frac{b(\delta\mathfrak{u}_h,\mathfrak{v}^{\mathrm{opt}})}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}}\|\mathfrak{v}^{\mathrm{opt}}\|_{\mathcal{V}}} = \inf_{\delta\mathfrak{u}_h\in\mathcal{U}_h} \sup_{\mathfrak{v}\in\mathcal{V}} \frac{b(\delta\mathfrak{u}_h,\mathfrak{v})}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}}\|\mathfrak{v}\|_{\mathcal{V}}} \geq \inf_{\mathfrak{u}\in\mathcal{U}} \sup_{\mathfrak{v}\in\mathcal{V}} \frac{b(\mathfrak{u},\mathfrak{v})}{\|\mathfrak{u}\|_{\mathcal{U}}\|\mathfrak{v}\|_{\mathcal{V}}} = \gamma\,, \tag{2.37}$$

so the discrete method is more stable than the original infinite-dimensional variational formulation. This is why $\mathcal{V}^{\mathrm{opt}}$ is called the optimal test space and why the underlying method is referred to as *the* optimal Petrov-Galerkin method. Lastly, note that given a basis for $\mathcal{U}_h$, $\mathfrak{u}_i \in \{\mathfrak{u}_j\}_{j=1}^N$, the corresponding optimal test function associated to a basis element $\mathfrak{u}_i$ is $\mathfrak{v}_i^{\mathrm{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_i$, as shown above. As mentioned in the previous section, these choices result in the matrix $\mathsf{B}_{ij}^{\mathrm{opt}} = b(\mathfrak{u}_j,\mathfrak{v}_i^{\mathrm{opt}})$, called the optimal stiffness matrix, which was claimed to be symmetric positive definite. This is easily shown, since for any nonzero $\mathsf{a} \in \mathbb{R}^N$,

$$\mathsf{a}^{\mathsf{T}}\mathsf{B}^{\mathrm{opt}}\mathsf{a} = \mathsf{a}_i\mathsf{B}_{ij}^{\mathrm{opt}}\mathsf{a}_j = \mathsf{a}_i\mathsf{a}_j b(\mathfrak{u}_j,\mathfrak{v}_i^{\mathrm{opt}}) = \mathsf{a}_i\mathsf{a}_j(\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_j,\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathfrak{u}_i)_{\mathcal{V}} = \|\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathsf{a}_i\mathfrak{u}_i\|_{\mathcal{V}}^2 > 0\,, \tag{2.38}$$

where the Einstein summation convention is adopted. The symmetry of the matrix is proved similarly.

Next, recall that due to the infinite-dimensional nature of $\mathcal{V}$, inverting the Riesz map exactly is not usually possible, so the Riesz map is inverted over a finite-dimensional enriched test space $\mathcal{V}_r \subseteq \mathcal{V}$ instead. To define those maps, first consider the orthogonal projections $\mathsf{P}_{\mathcal{V}_r} : \mathcal{V} \to \mathcal{V}_r$ and $\mathsf{P}_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r} : \mathcal{V}' \to \mathcal{R}_{\mathcal{V}}\mathcal{V}_r$ (note Riesz map's properties imply $\mathcal{R}_{\mathcal{V}}\mathcal{V}_r^{\perp} = (\mathcal{R}_{\mathcal{V}}\mathcal{V}_r)^{\perp}$); and the canonical embeddings $\iota_{\mathcal{V}_r} : \mathcal{V}_r \to \mathcal{V}$ and $\iota_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r} : \mathcal{R}_{\mathcal{V}}\mathcal{V}_r \to \mathcal{V}'$. Then, the $\mathcal{V}_r$ Riesz maps are,

$$\mathcal{R}_{\mathcal{V}_r} = \mathcal{R}_{\mathcal{V}}\iota_{\mathcal{V}_r}\mathsf{P}_{\mathcal{V}_r} : \mathcal{V} \to \mathcal{V}'\,, \qquad\qquad \mathcal{R}_{\mathcal{V}_r}^{-1} = \mathcal{R}_{\mathcal{V}}^{-1}\iota_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r}\mathsf{P}_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r} : \mathcal{V}' \to \mathcal{V}\,. \tag{2.39}$$

This is an abuse of notation, because neither $\mathcal{R}_{\mathcal{V}_r}$ nor $\mathcal{R}_{\mathcal{V}_r}^{-1}$ are invertible, but this is partly justified since $\mathcal{R}_{\mathcal{V}_r}^{-1}\big(\mathcal{R}_{\mathcal{V}_r}|_{\mathcal{V}_r}\big) = \mathrm{id}_{\mathcal{V}_r}$ and $\mathcal{R}_{\mathcal{V}_r}\big(\mathcal{R}_{\mathcal{V}_r}^{-1}|_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r}\big) = \mathrm{id}_{\mathcal{R}_{\mathcal{V}}\mathcal{V}_r}$. With these definitions, all the equivalences still hold, but the variables previously designated as optimal are now near-optimal:

$$\mathfrak{u}_h = \underset{\delta\mathfrak{u}_h\in\mathcal{U}_h}{\arg\min}\, \|\mathcal{R}_{\mathcal{V}_r}^{-1}(\mathcal{B}\delta\mathfrak{u}_h - \ell)\|_{\mathcal{V}}^2 \,, \tag{2.40}$$

$$b(\mathfrak{u}_h, \mathfrak{v}^{\text{n-opt}}) = \ell(\mathfrak{v}^{\text{n-opt}}) \quad \forall \mathfrak{v}^{\text{n-opt}} \in \mathcal{V}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h \,, \tag{2.41}$$

$$\langle \mathcal{B}'\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathfrak{u}_h, \delta\mathfrak{u}_h\rangle_{\mathcal{U}'\times\mathcal{U}} = \langle \mathcal{B}'\mathcal{R}_{\mathcal{V}_r}^{-1}\ell, \delta\mathfrak{u}_h\rangle_{\mathcal{U}'\times\mathcal{U}} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h \,, \tag{2.42}$$

$$\begin{pmatrix} \mathcal{R}_{\mathcal{V}_r} & \mathcal{B} \\ \mathcal{B}' & 0 \end{pmatrix} \begin{pmatrix} \psi_r \\ \mathfrak{u}_h \end{pmatrix} = \begin{pmatrix} \ell \\ 0 \end{pmatrix} \quad \Leftrightarrow \quad \begin{cases} (\psi_r, \mathfrak{v}_r)_{\mathcal{V}} + b(\mathfrak{u}_h, \mathfrak{v}_r) = \ell(\mathfrak{v}_r) & \forall \mathfrak{v}_r \in \mathcal{V}_r \,, \\ b(\delta\mathfrak{u}_h, \psi_r) \qquad\qquad\quad = 0 & \forall \delta\mathfrak{u}_h \in \mathcal{U}_h \,, \end{cases} \tag{2.43}$$

where ψ_r is now sought in $\mathcal{V}_r$ instead of $\mathcal{V}$. Rigorously speaking, this implies that (2.37) no longer holds (intuitively the larger $\mathcal{V}_r$ is, the closer it is to $\mathcal{V}$ and the closer γ_h will be to γ^{opt}) and this will be analyzed through a Fortin operator in Section 2.8.1.

The mixed method in (2.43), now has the same *finite*-dimensional trial and test spaces $\mathcal{U}_h^{\text{mix}} = \mathcal{V}_h^{\text{mix}} = \mathcal{V}_r \times \mathcal{U}_h$. This means it can be discretized and solved using the Bubnov-Galerkin method. Indeed, consider the bases $\{\mathfrak{v}_i\}_{i=1}^M$ and $\{\mathfrak{u}_j\}_{j=1}^N$ of $\mathcal{V}_r$ and $\mathcal{U}_h$ respectively, so their union produces a basis for $\mathcal{U}_h^{\text{mix}} = \mathcal{V}_h^{\text{mix}}$ of size $N + M$. The resulting discretization yields the mixed system,

$$\begin{bmatrix} \mathsf{G} & \mathsf{B} \\ \mathsf{B}^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} \psi_r \\ \mathsf{u}_h \end{bmatrix} = \begin{bmatrix} \mathsf{l} \\ 0 \end{bmatrix} \,, \tag{2.44}$$

where the final solution is given by $\psi_r = \sum_{i=1}^M (\psi_r)_i \mathfrak{v}_i \in \mathcal{V}_r$ and $\mathfrak{u}_h = \sum_{j=1}^N (\mathsf{u}_h)_j \mathfrak{u}_j \in \mathcal{U}_h$. The positive definite Gram matrix, $\mathsf{G}_{ij} = (\mathfrak{v}_i, \mathfrak{v}_j)_{\mathcal{V}}$, and enriched stiffness matrix and load, $\mathsf{B}_{ij} = b(\mathfrak{u}_j, \mathfrak{v}_i)$ and $\mathsf{l}_i = \ell(\mathfrak{v}_i)$, are defined as in the previous section. The Schur complement of (2.44) precisely recovers (2.28), but (2.28) can also be derived directly from (2.41) or (2.42). Lastly, the resulting near-optimal stiffness matrix which governs the system of equations, $\mathsf{B}^{\text{n-opt}} = \mathsf{B}^{\mathsf{T}}\mathsf{G}^{-1}\mathsf{B}$, is obviously symmetric positive definite because $\mathsf{B}^{\text{n-opt}} = (\mathsf{G}^{-1/2}\mathsf{B})^{\mathsf{T}}(\mathsf{G}^{-1/2}\mathsf{B})$, and full rank because $M \geq N$. This concludes the analysis of the different equivalent characterizations of minimum residual methods, both at the semi-discrete level and fully discrete level.

2.5 Least-squares methods and $\boldsymbol{L}^2$ test spaces

This section is not fundamental to understand the vast majority of this dissertation, so there is no harm if the reader wishes to jump ahead. It explores the connection of minimum residual methods and least-squares finite element methods, and also analyzes the more complicated case where only part of the test space is in L^2. Throughout this section, any spaces comprised of copies of $L^2(\Omega)$ (e.g. $\boldsymbol{L}^2(\Omega)$, $\mathsf{L}^2(\Omega;\mathbb{S})$, etc.) will be liberally referred to as simply L^2. It will be shown that when some of the test variables are in L^2, it is possible to exploit that $(L^2)' \cong L^2$ to avoid, at least to some degree, the discrete inversion of the Riesz map.

The most salient case occurs when $\mathcal{V} = L^2$, so $\mathcal{B} : \mathcal{U} \to (L^2)'$ and it must take the form $b(\mathfrak{u}, \mathfrak{v}) = \langle \mathcal{B}\mathfrak{u}, \mathfrak{v}\rangle_{(L^2)'\times L^2} = (\mathcal{L}\mathfrak{u}, \mathfrak{v})_{L^2}$ for all $\mathfrak{v} \in L^2$, meaning $\mathcal{L} = \mathcal{R}_{L^2}^{-1}\mathcal{B} : \mathcal{U} \to L^2$ is an easily identified operator. Similarly, the load is easily identified as a function $f \in L^2$, so that $\ell(\mathfrak{v}) = (f, \mathfrak{v})_{L^2}$ and $f = \mathcal{R}_{L^2}^{-1}\ell$. Then, simply rewrite the first variation in (2.27), which seeks $\mathfrak{u}_h^{\mathrm{opt}} \in \mathcal{U}_h$ such that

$$(\mathcal{L}\mathfrak{u}_h^{\mathrm{opt}}, \mathcal{L}\delta\mathfrak{u}_h)_{L^2} = (f, \mathcal{L}\delta\mathfrak{u}_h)_{L^2} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h\,. \tag{2.45}$$

The trivial identification of the $(L^2)'$ functions $\mathcal{B}\mathfrak{u}_h^{\mathrm{opt}}$, $\mathcal{B}\delta\mathfrak{u}_h$ and ℓ with the L^2 functions $\mathcal{L}\mathfrak{u}_h^{\mathrm{opt}}$, $\mathcal{L}\delta\mathfrak{u}_h$ and f corresponds precisely the exact inverse of the Riesz map, which is otherwise difficult to compute. In general, when $\mathcal{V} \neq L^2$, this identification is not accessible, so $\mathcal{L}\mathfrak{u}_h^{\mathrm{opt}}$, $\mathcal{L}\delta\mathfrak{u}_h$ and f would be unknown (i.e. they have to be computed inexactly). In this case, (2.45) can easily be discretized by considering a basis $\{\mathfrak{u}_j\}_{j=1}^N$ of $\mathcal{U}_h$, so that

$$\mathsf{B}^{\mathrm{opt}}\mathsf{u}_h^{\mathrm{opt}} = \mathsf{l}^{\mathrm{opt}}\,, \qquad \mathsf{B}_{ij}^{\mathrm{opt}} = (\mathcal{L}\mathfrak{u}_j, \mathcal{L}\mathfrak{u}_i)_{L^2}\,, \quad \mathsf{l}_i^{\mathrm{opt}} = (f, \mathcal{L}\mathfrak{u}_i)_{L^2}\,, \tag{2.46}$$

where $\mathfrak{u}_h^{\mathrm{opt}} = \sum_{j=1}^N (\mathsf{u}_h^{\mathrm{opt}})_j \mathfrak{u}_j \in \mathcal{U}_h$ is the optimal discrete solution. Note the notation reflects that $\mathsf{B}^{\mathrm{opt}}$ and $\mathsf{l}^{\mathrm{opt}}$ are the optimal stiffness matrix and load, not the near-optimal ones. This is because the Riesz map is being inverted exactly, so (assuming exact integration) the optimal stability of the original formulation is reproduced exactly, and the numerical error that arises when discretizing $\mathcal{V}_r$ is completely avoided. Moreover, the computational cost is significantly lowered when compared to the discretization in (2.28). When applied to PDEs written as a first-order system, these methods are known as first-order system least-squares (FOSLS) finite element methods [33].

When part of the test space is in L^2, similar optimizations are also possible, but only in the L^2 part of the test space, where the Riesz map is trivial. This both lowers computational cost and helps to better approach optimal stability. A derivation for those cases in a general setting is now presented.

Let $\mathcal{W}$ be a Hilbert space and assume the test space has the form $\mathcal{V} = \mathcal{W} \times L^2$ with the Hilbert norm $\|(\mathfrak{v}_{\mathcal{W}}, \mathfrak{v}_{L^2})\|_{\mathcal{V}}^2 = \|\mathfrak{v}_{\mathcal{W}}\|_{\mathcal{W}}^2 + \|\mathfrak{v}_{L^2}\|_{L^2}^2$. The operators associated to the variational formulation can then be easily decomposed as $\mathcal{B} = \mathcal{B}_{\mathcal{W}} \times \mathcal{B}_{L^2}$ and $\ell = \ell_{\mathcal{W}} \times \ell_{L^2}$, so that (2.33) can be rewritten as,

$$\langle \mathcal{B}_{\mathcal{W}}' \mathcal{R}_{\mathcal{W}}^{-1} \mathcal{B}_{\mathcal{W}} \mathfrak{u}_h^{\mathrm{opt}}, \delta\mathfrak{u}_h \rangle_{\mathcal{U}' \times \mathcal{U}} + (\mathcal{L}\mathfrak{u}_h^{\mathrm{opt}}, \mathcal{L}\delta\mathfrak{u}_h)_{L^2} = \langle \mathcal{B}_{\mathcal{W}}' \mathcal{R}_{\mathcal{W}}^{-1} \ell_{\mathcal{W}}, \delta\mathfrak{u}_h \rangle_{\mathcal{U}' \times \mathcal{U}} + (f, \mathcal{L}\delta\mathfrak{u}_h)_{L^2} \quad \forall \delta\mathfrak{u}_h \in \mathcal{U}_h \,, \tag{2.47}$$

where $(\mathcal{L}\mathfrak{u}, \mathfrak{v}_{L^2})_{L^2} = \langle \mathcal{B}_{L^2}\mathfrak{u}, \mathfrak{v}_{L^2} \rangle_{(L^2)' \times L^2}$ and $(f, \mathfrak{v}_{L^2})_{L^2} = \ell_{L^2}(\mathfrak{v}_{L^2})$. The inversion of $\mathcal{R}_{\mathcal{W}}^{-1}$ cannot be done exactly, so consider $\mathcal{R}_{\mathcal{W}_r}^{-1}$ instead to get a near-optimal method, where $\mathcal{W}_r \subseteq \mathcal{W}$ is an enriched test space. Choosing bases $\{\mathfrak{u}_j\}_{j=1}^N$ and $\{\mathfrak{w}_i\}_{i=1}^{M_{\mathcal{W}}}$ of $\mathcal{U}_h$ and $\mathcal{W}_r$ respectively yields the following linear system of equations,

$$(\mathsf{B}_{\mathcal{W}}^{\text{n-opt}} + \mathsf{B}_{L^2}^{\text{opt}})\mathsf{u}_h = (\mathsf{l}_{\mathcal{W}}^{\text{n-opt}} + \mathsf{l}_{L^2}^{\text{opt}}) \,, \qquad \mathsf{B}_{\mathcal{W}}^{\text{n-opt}} = \mathsf{B}_{\mathcal{W}}^{\mathsf{T}} \mathsf{G}_{\mathcal{W}}^{-1} \mathsf{B}_{\mathcal{W}} \,, \quad \mathsf{l}_{\mathcal{W}}^{\text{n-opt}} = \mathsf{B}_{\mathcal{W}}^{\mathsf{T}} \mathsf{G}_{\mathcal{W}}^{-1} \mathsf{l}_{\mathcal{W}} \,, \tag{2.48}$$

where $\mathfrak{u}_h = \sum_{j=1}^N (\mathsf{u}_h)_j \mathfrak{u}_j \in \mathcal{U}_h$, $\big(\mathsf{B}_{L^2}^{\text{opt}}\big)_{ij} = (\mathcal{L}\mathfrak{u}_j, \mathcal{L}\mathfrak{u}_i)_{L^2}$, $\big(\mathsf{l}_{L^2}^{\text{opt}}\big)_i = (f, \mathcal{L}\mathfrak{u}_i)_{L^2}$, $\big(\mathsf{G}_{\mathcal{W}}\big)_{ij} = (\mathfrak{w}_i, \mathfrak{w}_j)_{\mathcal{W}}$, $\big(\mathsf{B}_{\mathcal{W}}\big)_{ij} = \langle \mathcal{B}_{\mathcal{W}}\mathfrak{u}_j, \mathfrak{w}_i \rangle_{\mathcal{W}' \times \mathcal{W}}$ and $\big(\mathsf{l}_{\mathcal{W}}\big)_i = \ell_{\mathcal{W}}(\mathfrak{w}_i)$.

2.6 Adaptivity

This section discusses measuring the residual norm in minimum residual methods and a posteriori error estimators in DPG methods. These allow for adaptivity to be implemented.

One big advantage of minimum residual methods is that they have a built-in way to measure the residual norm, which is the same as the error measured in the special norm $\|\cdot\|_E = \|\mathcal{B}(\cdot)\|_{\mathcal{V}'}$, since $\|\mathfrak{u}_h - \mathfrak{u}\|_E = \|\mathcal{B}\mathfrak{u}_h - \ell\|_{\mathcal{V}'}$, where $\mathfrak{u}$ is the exact solution (so $\mathcal{B}\mathfrak{u} = \ell$) and $\mathfrak{u}_h$ is the computed solution of the minimum residual method. Ideally, minimum residual methods minimize the residual $\|\mathcal{B}\mathfrak{u}_h - \ell\|_{\mathcal{V}'}$ as in (2.31), so as long as $\mathcal{V}$ remains the same, the exact residual will decrease as the trial space is refined in a consistent nested manner (i.e. a series of refinements $\mathcal{U}_{h_1} \subseteq \mathcal{U}_{h_2} \subseteq \ldots \subseteq \mathcal{U}_{h_n}$).

In practice, however, usually the residual norm cannot be computed exactly due to the infinite-dimensional nature of $\mathcal{V}$, so the enriched test space $\mathcal{V}_r$ is considered again, and an expression for the approximate global residual norm can be deduced to be

$$\eta_h^2 = \|\mathcal{R}_{\mathcal{V}_r}^{-1}(\mathcal{B}\mathsf{u}_h - \ell)\|_{\mathcal{V}}^2 = (\mathsf{B}\mathsf{u}_h - \mathsf{l})^{\mathsf{T}}\mathsf{G}^{-1}(\mathsf{B}\mathsf{u}_h - \mathsf{l})\,, \tag{2.49}$$

where B and l are the enriched stiffness matrix and load defined previously, and u_h is the solution vector of coefficients computed from the minimum residual method in (2.28). Unfortunately, this expression is not of much use to develop adaptivity, since it is only a global value, so it is not an element-wise a posteriori error estimator.

This changes when the test spaces are broken, like in broken variational formulations discretized via the DPG methodology. In these cases, the computations are localized to each element. Thus, enriched stiffness matrices, Gram matrices and loads are computed for each element, B_K, G_K and l_K for all $K \in \mathcal{T}$. In DPG methods, as mentioned in Section 2.3, the local near-optimal stiffness matrices $\mathsf{B}_K^{\text{n-opt}} = \mathsf{B}_K^{\mathsf{T}}\mathsf{G}_K^{-1}\mathsf{B}_K$ and loads $\mathsf{l}_K^{\text{n-opt}} = \mathsf{B}_K^{\mathsf{T}}\mathsf{G}_K^{-1}\mathsf{l}_K$ are computed and assembled into $\mathsf{B}^{\text{n-opt}}$ and $\mathsf{l}^{\text{n-opt}}$ as in any other finite element method, and then the solution vector u_h in (2.28) is solved for. A posteriori, one can arrange a vector $\mathsf{u}_{h,K}$ for each $K \in \mathcal{T}$ comprised of the components associated to the trial basis functions with support in K. The residual then becomes,

$$\eta_h^2 = \|\mathcal{R}_{\mathcal{V}_r}^{-1}(\mathcal{B}\mathsf{u}_h - \ell)\|_{\mathcal{V}}^2 = \sum_{K\in\mathcal{T}} \eta_K^2\,, \qquad \eta_K^2 = (\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K)^{\mathsf{T}}\mathsf{G}_K^{-1}(\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K)\,, \tag{2.50}$$

where the η_K are element-wise a posteriori error estimators. These then allow to develop adaptive strategies by marking the desired elements for refinement under some proposed criterion. Technically speaking, the global residual in DPG methods need not decrease as the mesh is refined (except in 1D), since the mesh-dependent test space changes (and along with it $\|\cdot\|_{\mathcal{V}'}$ and $\mathcal{V}_r$) and so does the trial space (in a way that can violate the nesting properties due to the presence of interface variables). Nevertheless, in practice the residual does typically decrease with successive refinements. Note that the residual-based estimator is natural since the residual itself is a foundation that drives DPG methods, and note that expressions for η_K need not be modified if high-order discretizations are being considered. Therefore, all DPG methods have a very convenient natural residual-based high-order a posteriori error estimator that can be used to implement adaptivity.

Remark 2.2. The criterion to refine elements usually follows a basic greedy algorithm where the element $K \in \mathcal{T}$ is marked for refinement if,

$$\eta_K \geq \alpha_\eta \eta_{\max} , \qquad \eta_{\max} = \max_{K\in\mathcal{T}} \eta_K , \tag{2.51}$$

where $\alpha_\eta \in [0,1]$. This means that the elements marked for refinement are those whose residual norm lies within α_η·100% of the maximum element residual norm (among all elements in the mesh). A common choice is $\alpha_\eta = 0.5$.

Remark 2.3. In the cases considered in Section 2.5 the residual can also be computed. In fact, it can be computed exactly (assuming exact integration) when $\mathcal{V} = L^2$ as

$$\eta^2 = \|\mathcal{B}\mathfrak{u}_h^{\mathrm{opt}} - \ell\|_{\mathcal{V}'}^2 = \sum_{K\in\mathcal{T}} \eta_K^2 , \qquad \eta_K^2 = (\mathcal{L}\mathfrak{u}_h^{\mathrm{opt}} - f, \mathcal{L}\mathfrak{u}_h^{\mathrm{opt}} - f)_{L_K^2} , \tag{2.52}$$

where $\mathfrak{u}_h^{\mathrm{opt}} = \sum_{j=1}^N (\mathsf{u}_h^{\mathrm{opt}})_j \mathfrak{u}_j$, and L_K^2 is a restriction to $K \in \mathcal{T}$ of the domain associated to L^2 (e.g. $L_K^2 = L^2(K)$ if $L^2 = L^2(\Omega)$). Similarly, when only part of the test space is in L^2, like when $\mathcal{V} = \mathcal{W} \times L^2$, the local residual can also be approximated as

$$\eta_K^2 = (\mathsf{B}_{\mathcal{W},K}\mathsf{u}_{h,K} - \mathsf{l}_{\mathcal{W},K})^{\mathsf{T}} \mathsf{G}_{\mathcal{W},K}^{-1} (\mathsf{B}_{\mathcal{W},K}\mathsf{u}_{h,K} - \mathsf{l}_{\mathcal{W},K}) + (\mathcal{L}\mathfrak{u}_h - f, \mathcal{L}\mathfrak{u}_h - f)_{L_K^2} . \tag{2.53}$$

Remark 2.4. Note that the enriched test space used to compute the residual can actually be different from that used to obtain the discrete solution. However, if that is the case, then B_K, G_K and l_K must be recomputed as a function of the new $\mathcal{V}_r$ (but u_h stays the same). This can sometimes facilitate comparisons among residuals. For example, if u_{h_1} and u_{h_2} represent discrete solutions from solving the problems with different polynomial orders (using $\mathcal{V}_{r_1}$ and $\mathcal{V}_{r_2}$), then the residuals from both solutions can be more accurately compared if the residuals are computed using a fixed $\mathcal{V}_r$.

2.7 Choice of trial and test spaces

For simplicity assume the number of spatial dimensions is $n_d = 3$ and consider a (simply connected) polyhedral element $K \in \mathcal{T}$. Assume there exists a family of high-order finite-dimensional discretizations of the spaces $H^1(K)$, $\boldsymbol{H}(\mathrm{curl}, K)$, $\boldsymbol{H}(\mathrm{div}, K)$ and $L^2(K)$ forming a differential de Rham exact sequence (or complex) as follows,

$$\begin{tikzcd}
H^1(K) \arrow[d, phantom, sloped, "\supseteq"] & \boldsymbol{H}(\mathrm{curl},K) \arrow[d, phantom, sloped, "\supseteq"] & \boldsymbol{H}(\mathrm{div},K) \arrow[d, phantom, sloped, "\supseteq"] & L^2(K) \arrow[d, phantom, sloped, "\supseteq"] \\
W^p(K) \arrow[r, "\nabla"] \arrow[d, phantom, sloped, "\supseteq"] & \boldsymbol{Q}^p(K) \arrow[r, "\mathrm{curl}"] \arrow[d, phantom, sloped, "\supseteq"] & \boldsymbol{V}^p(K) \arrow[r, "\mathrm{div}"] \arrow[d, phantom, sloped, "\supseteq"] & Y^p(K) \arrow[d, phantom, sloped, "\supseteq"] \\
\mathcal{P}^p & (\mathcal{P}^{p-1})^3 & (\mathcal{P}^{p-1})^3 & \mathcal{P}^{p-1}
\end{tikzcd} \tag{2.54}$$

where $\mathcal{P}^p$ are the high-order polynomials in $\boldsymbol{x} = (x_1, x_2, x_3)$ of total order at most p. Note the parameter p represents the order of the discrete *sequence* composed of $W^p(K)$, $\boldsymbol{Q}^p(K)$, $\boldsymbol{V}^p(K)$ and $Y^p(K)$, but does not necessarily coincide with the polynomial order of a specific discrete space in the sequence (e.g. sometimes $Y^p(K) = \mathcal{P}^{p-1}$ even though it comes from the sequence of order p). Instead, the parameter p is intended to eventually coincide with the order of convergence of the numerical method (i.e. $\|\mathfrak{u}-\mathfrak{u}_h\|_{\mathfrak{U}} \leq Ch^p$). These discretizations are referred to as Sobolev-de Rham discretizations, or simply SdR discretizations, and we refer to Appendix A (Section (A.5)) for more subtleties and requirements about this definition. Discretization of the local traces is also possible simply by using the appropriate definitions of traces (see Appendix A),

$$\begin{aligned}
W^p(\partial K) &= \big\{\hat{\phi}_K = \phi|_{\partial K} \mid \phi \in W^p(K)\big\} \subseteq H^{1/2}(\partial K)\,, \\
\boldsymbol{Q}^p_{\top}(\partial K) &= \big\{\hat{\boldsymbol{E}}_{\top_K} = \big(\hat{\mathbf{n}}_K \times \boldsymbol{E}|_{\partial K}\big) \times \hat{\mathbf{n}}_K \mid \boldsymbol{E} \in \boldsymbol{Q}^p(K)\big\} \subseteq \boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)\,, \\
\boldsymbol{Q}^p_{\dashv}(\partial K) &= \big\{\hat{\boldsymbol{F}}_{\dashv_K} = \hat{\mathbf{n}}_K \times \boldsymbol{F}|_{\partial K} \mid \boldsymbol{F} \in \boldsymbol{Q}^p(K)\big\} \subseteq \boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)\,, \\
V^p(\partial K) &= \big\{\hat{v}_{\mathbf{n}_K} = \boldsymbol{v}|_{\partial K} \cdot \hat{\mathbf{n}}_K \mid \boldsymbol{v} \in \boldsymbol{V}^p(K)\big\} \subseteq H^{-1/2}(\partial K)\,,
\end{aligned} \tag{2.55}$$

where $\hat{\mathbf{n}}_K$ is the outward boundary normal vector of the element $K \in \mathcal{T}$.

The sequence of spaces depends on the type of element $K \in \mathcal{T}$, and for the conventional elements, like hexahedra, tetrahedra, triangular prisms and pyramids, several different families that satisfy these properties have been proposed in the literature [114, 74, 103, 34, 205, 236, 57]. Naturally, to have globally conforming discretizations of $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl},\Omega)$ and $\boldsymbol{H}(\mathrm{div},\Omega)$ across the whole mesh, then some interelement compatibility of the spaces is required. In other words, if F is a common face between elements K_1 and K_2, then the face trace restrictions of discretizations of both elements should coincide, $W^p(\partial K_1)|_F = W^p(\partial K_2)|_F$, $V^p(\partial K_1)|_F = V^p(\partial K_2)|_F$, etc. This criterion is relatively easy to satisfy when all elements of the mesh are of the same type (all hexahedra or all tetrahedra), but becomes more complicated when combining different types of elements in the same mesh. This aspect is recognized and well covered in the literature.

Turning back to DPG methods, it will be assumed that the trial and test spaces, $\mathcal{U}$ and $\mathcal{V}$, only involve copies of the usual Sobolev spaces, $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl},\Omega)$, $\boldsymbol{H}(\mathrm{div},\Omega)$ and $L^2(\Omega)$, their broken counterparts and their relevant traces. This is a reasonable assumption as most variational formulations involve these spaces or can be modified to solely involve these spaces. If it is not possible, then appropriate discretizations satisfying the right approximation properties have to be developed. In any case, with these assumptions it is clear that when the functions $\delta\mathfrak{u} = (\delta\mathfrak{u}_0, \delta\hat{\mathfrak{u}}) \in \mathcal{U}_0 \times \hat{\mathcal{U}} = \mathcal{U}$ have their domain restricted to K, their exists a trial space with the same canonical structure (involving no boundary conditions) which contains the restricted function, $\delta\mathfrak{u}|_K = (\delta\mathfrak{u}_0|_K, \delta\hat{\mathfrak{u}}_K) \in \mathcal{U}_0(K) \times \hat{\mathcal{U}}(K) = \mathcal{U}(K)$, where $\delta\hat{\mathfrak{u}}_K$ is the K-th component of a $\mathcal{T}$-tuple. The space $\mathcal{U}_0(K)$ will be composed of copies of $H^1(K)$, $\boldsymbol{H}(\mathrm{curl},K)$, $\boldsymbol{H}(\mathrm{div},K)$ and $L^2(K)$, while $\hat{\mathcal{U}}(K)$ will be composed of copies of $H^{1/2}(\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)$ and $H^{-1/2}(\partial K)$. Thus, the discretization of $\mathcal{U}(K)$ is

$$\mathcal{U}_h^p(K) = \mathcal{U}_{0,h}^p(K) \times \hat{\mathcal{U}}_h^p(K) \subseteq \mathcal{U}(K)\,, \qquad \mathcal{U}_{0,h}^p(K) \subseteq \mathcal{U}_0(K)\,, \quad \hat{\mathcal{U}}_h^p(K) \subseteq \hat{\mathcal{U}}(K)\,, \tag{2.56}$$

where $\mathcal{U}_{0,h}^p(K)$ is composed to the corresponding copies of $W^p(K)$, $\boldsymbol{Q}^p(K)$, $\boldsymbol{V}^p(K)$ and $Y^p(K)$, and $\hat{\mathcal{U}}_h^p(K)$ is composed to the corresponding copies of $W^p(\partial K)$, $\boldsymbol{Q}_\top^p(\partial K)$, $\boldsymbol{Q}_\dashv^p(\partial K)$ and $V^p(\partial K)$.

The same can be said of the test space $\mathcal{V}$, which will be composed of copies of the broken spaces $H^1(\mathcal{T})$, $\boldsymbol{H}(\mathrm{curl},\mathcal{T})$, $\boldsymbol{H}(\mathrm{div},\mathcal{T})$ and $L^2(\mathcal{T})$, so its K-restricted version, $\mathcal{V}(K)$, will also be composed of the analogous copies of $H^1(K)$, $\boldsymbol{H}(\mathrm{curl},K)$, $\boldsymbol{H}(\mathrm{div},K)$ and $L^2(K)$. However, the DPG discretization of $\mathcal{V}$ is not directly $\mathcal{V}_h = \mathcal{V}^{\text{n-opt}}$, but the enriched test space $\mathcal{V}_r$ (used to implicitly compute $\mathcal{V}^{\text{n-opt}}$). As mentioned previously, at least it is required that $M = \dim(\mathcal{V}_r) \geq \dim(\mathcal{U}_h) = N$ for minimum residual methods to make sense (so that the matrix $\mathsf{B}^{\text{n-opt}}$ is invertible). This is obviously satisfied by the conservative criterion that $M_K = \dim\big(\mathcal{V}_r(K)\big) \geq \dim\big(\mathcal{U}_h^p(K)\big) = N_K$ for every $K \in \mathcal{T}$, where $\mathcal{V}_r(K)$ is some discretization of $\mathcal{V}(K)$. Thus, satisfying this criterion is usually the approach taken. With this in mind, there are many ways to find an appropriate discretization $\mathcal{V}_r(K) \subseteq \mathcal{V}(K)$, but the modus operandi has typically been that of choosing $\mathcal{V}_r(K)$ to come from a sequence of enriched order $p+\Delta p$,

$$\mathcal{V}_r(K)=\mathcal{V}_r^{p+\Delta p}(K)\subseteq\mathcal{V}(K)\,, \tag{2.57}$$

where $\mathcal{V}_r^{p+\Delta p}(K)$ is composed to the corresponding copies of $W^p(K)$, $\boldsymbol{Q}^p(K)$, $\boldsymbol{V}^p(K)$ and $Y^p(K)$, and where p is the order of the local trial space discretization $\mathcal{U}_h^p(K)$. The value of the enrichment parameter, Δp, can then be chosen locally (a Δp_K for each $K\in\mathcal{T}$) or globally (same Δp everywhere) to at least satisfy $M_K=\dim\big(\mathcal{V}_r^{p+\Delta p}(K)\big)\geq\dim\big(\mathcal{U}_h^p(K)\big)=N_K$ at each $K\in\mathcal{T}$.

Finally, the global high-order discrete trial and enriched test spaces are,

$$\begin{aligned}\mathcal{U}_h&=\mathcal{U}_h^p=\big\{\delta\mathfrak{u}_h\in\mathcal{U}\mid\delta\mathfrak{u}_h|_K\in\mathcal{U}_h^p(K)\big\}\subseteq\mathcal{U}\,,\\ \mathcal{V}_r&=\mathcal{V}_r^{p+\Delta p}=\big\{\mathfrak{v}_r\mid\mathfrak{v}_r|_K\in\mathcal{V}_r^{p+\Delta p}(K)\big\}\subseteq\mathcal{V}\,.\end{aligned} \tag{2.58}$$

Note that the requirement that $\delta\mathfrak{u}_h\in\mathcal{U}$ ensures the compatibility across elements. Otherwise it could occur that $\delta\mathfrak{u}_h\in\prod_{K\in\mathcal{T}}\mathcal{U}_h^p(K)\backslash\mathcal{U}$. No such requirement exists for the enriched test functions, because these spaces are assumed to be naturally broken. In the terminology of Appendix A (see Definition A.2), $\mathcal{U}_h$ is said to be a compatible SdR discretization of order p, while $\mathcal{V}_r$ is an SdR discretization of order $p+\Delta p$. Lastly, notice these discretizations imply that DPG methods are conforming finite element methods (even though the test spaces are broken).

It is always possible to increase the parameter Δp (or to grow $\mathcal{V}_r(K)$ some other way) so that the accuracy of the method is improved, since the approximation of $\mathcal{V}^{\text{n-opt}}=\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h$ to $\mathcal{V}^{\text{opt}}=\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h$ will be better. This means that once $\mathcal{U}_h$ has been chosen, Δp acts as a convenient tunable parameter that, if sufficiently high, will ensure the stability of the numerical method. This is in contrast with the more complicated balancing act that arises in traditional high-order mixed methods, where both the simultaneous discretizations of all variables must be considered and carefully analyzed to determine numerical stability [38].

As an example of this discretization process, consider the variational formulation of Poisson's equation in Section 2.2 given in (2.10). The trial and test space discretizations would then be,

$$\begin{aligned}&\mathcal{U}_h=\big\{(\phi,\hat{\tau}_{\mathbf{n}})\in\mathcal{U}\mid\phi|_K\in W^p(K),\ (\hat{\tau}_{\mathbf{n}})_K\in V^p(\partial K)\big\}\subseteq\mathcal{U}=H_0^1(\Omega)\times H^{-1/2}(\partial\mathcal{T})\,,\\ &\mathcal{V}_r=\big\{w\mid w|_K\in W^{p+\Delta p}(K)\big\}\subseteq\mathcal{V}=H^1(\mathcal{T})\,.\end{aligned} \tag{2.59}$$

The requirement that $(\phi, \hat{\tau}_{\mathbf{n}}) \in \mathcal{U}$ implies the interelement compatibilities $\phi|_{K_1}\big|_F = \phi|_{K_2}\big|_F$ and $(\hat{\tau}_{\mathbf{n}})_{K_1}|_F = -(\hat{\tau}_{\mathbf{n}})_{K_2}|_F$ for every common face F between mesh elements K_1 and K_2. To be even more explicit, consider a mesh of only tetrahedra, and use the spaces coming from the classical Nédélec sequence of the first type [181], so that $W^p(K) = \mathcal{P}^p$ and $V^p(\partial K)$ is defined as in (2.55) with $\boldsymbol{V}^p(K) = \mathcal{RT}^p = (\mathcal{P}^{p-1})^3 + \boldsymbol{x}\mathcal{P}^{p-1}$ being the high-order Raviart-Thomas space.

2.8 Convergence

To prove high-order convergence of DPG methods, first the issue of stability in the fully discrete case will be analyzed (Section 2.8.1), and then using interpolation inequalities, the final convergence estimates will be produced (Section 2.8.2). As shown in Section 2.4, minimum residual methods are optimally stable provided the trial space has been discretized, and that the Riesz map can be inverted exactly (at least over a relevant part of its range). In two special cases outlined in remarks below one can show this is indeed the case. In the remaining cases, this inversion is assumed not to be done exactly, so a cost is incurred in the stability of the method due to the use of the enriched test space. To get a conservative estimate of that cost, one can posit the existence of a Fortin operator, which can then be deduced to provide a bound for the discrete inf-sup constant. This allows to make a fully rigorous analysis of the convergence in particular situations where such a Fortin operator can be explicitly constructed. Otherwise, the existence of such Fortin operators is assumed, and an a priori bound for the best approximation error can be established. Later, interpolation estimates provide the bound resulting in asymptotic high-order convergence.

Remark 2.5. When the test space is $\mathcal{V} = L^2$ as in Section 2.5, then the Riesz map is inverted exactly (because it is trivial to do so). Thus, the resulting discrete solution, $\mathfrak{u}_h^{\text{opt}}$, is indeed optimal and so are the stiffness matrix B^{opt} and inf-sup constant γ^{opt} as in (2.31) and (2.37). In other words, the method is in fact an ideal minimum residual method, and $\mathfrak{u}_h^{\text{opt}}$ attains such minimum exact residual.

Remark 2.6. When $\mathcal{B}\mathcal{U}_h \subseteq \mathcal{R}_{\mathcal{V}}\mathcal{V}_r$, then the resulting method will function like an ideal minimum residual method. To see this, define $\mathcal{R}_{\mathcal{V}_r^\perp} = \mathcal{V} \to \mathcal{V}'$ and $\mathcal{R}_{\mathcal{V}_r^\perp}^{-1} = \mathcal{V} \to \mathcal{V}'$ analogously to $\mathcal{R}_{\mathcal{V}_r}$ and

$\mathcal{R}_{\mathcal{V}_r}^{-1}$ in (2.39) (as noted there, there is an abuse of notation), so it follows that

$$\begin{aligned} &\mathcal{R}_{\mathcal{V}} = \mathcal{R}_{\mathcal{V}_r} + \mathcal{R}_{\mathcal{V}_r^\perp}\,, && \mathcal{R}_{\mathcal{V}}\mathfrak{v} = \mathcal{R}_{\mathcal{V}_r}\mathfrak{v} + \mathcal{R}_{\mathcal{V}_r^\perp}\mathfrak{v} \in \mathcal{V}'\,, && \mathcal{R}_{\mathcal{V}}\mathcal{V}_r \ni \mathcal{R}_{\mathcal{V}_r}\mathfrak{v} \perp \mathcal{R}_{\mathcal{V}_r^\perp}\mathfrak{v} \in \mathcal{R}_{\mathcal{V}}\mathcal{V}_r^\perp\,,\\ &\mathcal{R}_{\mathcal{V}}^{-1} = \mathcal{R}_{\mathcal{V}_r}^{-1} + \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}\,, && \mathcal{R}_{\mathcal{V}}^{-1}\mathfrak{v}' = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathfrak{v}' + \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}\mathfrak{v}' \in \mathcal{V}\,, && \mathcal{V}_r \ni \mathcal{R}_{\mathcal{V}_r}^{-1}\mathfrak{v}' \perp \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}\mathfrak{v}' \in \mathcal{V}_r^\perp\,. \end{aligned} \tag{2.60}$$

Hence, if $\mathcal{B}\mathcal{U}_h \subseteq \mathcal{R}_{\mathcal{V}}\mathcal{V}_r$, then $\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h \subseteq \mathcal{V}_r$ and more importantly,

$$\mathcal{V}^{\text{opt}} = \mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\mathcal{U}_h = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h + \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}\mathcal{B}\mathcal{U}_h = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h + \{0\} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h = \mathcal{V}^{\text{n-opt}}\,. \tag{2.61}$$

Therefore, using (2.37), and comparing (2.32) with (2.41), it is clear that the solution is optimal ($\mathfrak{u}_h = \mathfrak{u}_h^{\text{opt}}$) and so are the inf-sup constant ($\gamma_h = \gamma^{\text{opt}}$), and stiffness matrix and load ($\mathsf{B}^{\text{n-opt}} = \mathsf{B}^{\text{opt}}$ and $\mathfrak{l}^{\text{n-opt}} = \mathfrak{l}^{\text{opt}}$).

2.8.1 Fortin operator

When DPG methods are used, in general $\mathcal{V}^{\text{n-opt}} \neq \mathcal{V}^{\text{opt}}$, so the inf-sup constant is not exactly optimal, $\gamma_h \neq \gamma^{\text{opt}}$. Thus a lower positive bound for γ_h must be deduced. This can be done via a Fortin operator, as shown in the next theorem, first proved in [133].

Theorem 2.2. *Let $b : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be the bilinear form associated to a well-posed linear variational formulation, and let $\mathcal{U}_h \subseteq \mathcal{U}$ and $\mathcal{V}_r \subseteq \mathcal{V}$ be a discrete trial and enriched test space respectively. Suppose there exists a continuous linear operator $\Pi_F : \mathcal{V} \to \mathcal{V}_r$, such that*

$$\begin{aligned} &\|\Pi_F \mathfrak{v}\|_{\mathcal{V}} \leq C_\Pi \|\mathfrak{v}\|_{\mathcal{V}} && \forall \mathfrak{v} \in \mathcal{V}\,,\\ &b(\delta\mathfrak{u}_h, \mathfrak{v} - \Pi_F\mathfrak{v}) = 0 && \forall \delta\mathfrak{u}_h \in \mathcal{U}_h,\ \forall \mathfrak{v} \in \mathcal{V}\,. \end{aligned} \tag{2.62}$$

Then,

$$\inf_{\delta\mathfrak{u}_h \in \mathcal{U}_h} \sup_{\mathfrak{v}_h \in \mathcal{V}^{\text{n-opt}}} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_h)}{\|\delta\mathfrak{u}_h\|_{\mathcal{U}} \|\mathfrak{v}_h\|_{\mathcal{V}}} \geq \frac{\gamma}{C_\Pi}\,, \tag{2.63}$$

where $\gamma = \inf_{\delta\mathfrak{u} \in \mathcal{U}} \sup_{\mathfrak{v} \in \mathcal{V}} \frac{|b(\delta\mathfrak{u}, \mathfrak{v})|}{\|\delta\mathfrak{u}\|_{\mathcal{U}} \|\mathfrak{v}\|_{\mathcal{V}}}$ and $\mathcal{V}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h$, with $\mathcal{R}_{\mathcal{V}_r}^{-1}$ defined in (2.39) *and $\mathcal{B}$ defined in* (2.25)*.*

Proof. Let $\delta\mathfrak{u}_h \in \mathcal{U}_h$. Then,

$$\gamma \|\delta\mathfrak{u}_h\|_{\mathcal{U}} \leq \sup_{\mathfrak{v} \in \mathcal{V}} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v})}{\|\mathfrak{v}\|_{\mathcal{V}}} = \sup_{\mathfrak{v} \in \mathcal{V}} \frac{b(\delta\mathfrak{u}_h, \Pi_F\mathfrak{v})}{\|\mathfrak{v}\|_{\mathcal{V}}} \leq C_\Pi \sup_{\mathfrak{v} \in \mathcal{V}} \frac{b(\delta\mathfrak{u}_h, \Pi_F\mathfrak{v})}{\|\Pi_F\mathfrak{v}\|_{\mathcal{V}}} \leq C_\Pi \sup_{\mathfrak{v}_r \in \mathcal{V}_r} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_r)}{\|\mathfrak{v}_r\|_{\mathcal{V}}}\,. \tag{2.64}$$

For the final step, note the orthogonal decomposition of $\mathcal{R}_{\mathcal{V}}^{-1} = \mathcal{R}_{\mathcal{V}_r}^{-1} + \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}$ provided in (2.60) in Remark 2.6, which implies

$$b(\delta\mathfrak{u}_h, \mathfrak{v}_r) = (\mathcal{R}_{\mathcal{V}}^{-1}\mathcal{B}\delta\mathfrak{u}_h, \mathfrak{v}_r)_{\mathcal{V}} = (\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h + \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}\mathcal{B}\delta\mathfrak{u}_h, \mathfrak{v}_r)_{\mathcal{V}} = (\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h, \mathfrak{v}_r)_{\mathcal{V}} \quad \forall \mathfrak{v}_r \in \mathcal{V}_r \,, \tag{2.65}$$

so that

$$\sup_{\mathfrak{v}_r \in \mathcal{V}_r} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_r)}{\|\mathfrak{v}_r\|_{\mathcal{V}}} = \sup_{\mathfrak{v}_r \in \mathcal{V}_r} \frac{(\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h, \mathfrak{v}_r)_{\mathcal{V}}}{\|\mathfrak{v}_r\|_{\mathcal{V}}} = \frac{b(\delta\mathfrak{u}_h, \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h)}{\|\mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h\|_{\mathcal{V}}} = \sup_{\mathfrak{v}_h \in \mathcal{V}^{\text{n-opt}}} \frac{b(\delta\mathfrak{u}_h, \mathfrak{v}_h)}{\|\mathfrak{v}_h\|_{\mathcal{V}}} \,, \tag{2.66}$$

because the supremum is clearly attained by $\mathfrak{v}_r = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\delta\mathfrak{u}_h \in \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h = \mathcal{V}^{\text{n-opt}}$. □

The seminal work by Babuška then establishes the following result [12, 235].

Corollary 2.1. *If the assumptions of Theorem 2.2 are satisfied, then the solution error is bounded by the best approximation error,*

$$\|\mathfrak{u} - \mathfrak{u}_h\|_{\mathcal{U}} \leq \frac{C_\Pi M}{\gamma} \inf_{\delta\mathfrak{u}_h \in \mathcal{U}_h} \|\delta\mathfrak{u}_h - \mathfrak{u}_h\|_{\mathcal{U}} \,, \tag{2.67}$$

where $|b(\delta\mathfrak{u}, \mathfrak{v})| \leq M\|\delta\mathfrak{u}\|_{\mathcal{U}}\|\mathfrak{v}\|_{\mathcal{V}}$ for all $\delta\mathfrak{u} \in \mathcal{U}$ and $\mathfrak{v} \in \mathcal{V}$, $\mathfrak{u}$ is the exact solution of the original variational formulation, and $\mathfrak{u}_h$ is the solution of the discretized variational formulation with $\mathcal{U}_h$ and $\mathcal{V}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_r}^{-1}\mathcal{B}\mathcal{U}_h$ as trial and test spaces.

Remark 2.7. Babuška's theory and even DPG methods have been generalized to the Banach space setting, along with the Fortin operator [213, 177]. In fact, an analogous estimate just like (2.67) is available in the literature [177].

Remark 2.8. Obviously, when $\mathcal{V}^{\text{n-opt}} = \mathcal{V}^{\text{opt}}$, application of Babuška's theorem yields the same result as in (2.67) but without the constant C_Π coming from the Fortin operator.

The issue is now to construct such a Fortin operator. In general, this is a complicated task. Fortunately, due to the broken test spaces in DPG methods it becomes a little easier, since the problem decouples and it suffices to find local Fortin operators for each specific element. The first construction in the context of DPG methods was done in [133], but since then, more elaborate and general constructions have been made [179, 55, 56]. All constructions are in the context of simplices, so it is an open problem to develop these operators for other types of elements. The following theorem represents the most general construction for tetrahedra at this moment [55].

Theorem 2.3. *Let $K \in \mathcal{T}$ be a tetrahedron. Then, there exist a sequence of commuting linear and continuous Fortin operators with domains $H^1(K)$, $\boldsymbol{H}(\mathrm{curl}, K)$, $\boldsymbol{H}(\mathrm{div}, K)$ and $L^2(K)$ as follows,*

$$\begin{tikzcd}
H^1(K) \arrow[r, "\nabla"] \arrow[d, "\Pi^{p,\Delta p}_{F,\mathrm{grad},K}"] & \boldsymbol{H}(\mathrm{curl},K) \arrow[r, "\mathrm{curl}"] \arrow[d, "\Pi^{p,\Delta p}_{F,\mathrm{curl},K}"] & \boldsymbol{H}(\mathrm{div},K) \arrow[r, "\mathrm{div}"] \arrow[d, "\Pi^{p,\Delta p}_{F,\mathrm{div},K}"] & L^2(K) \arrow[d, "\Pi^{p,\Delta p}_{F,\int,K}"] \\
W^{p+\Delta p}(K) \arrow[r, "\nabla"] & \boldsymbol{Q}^{p+\Delta p}(K) \arrow[r, "\mathrm{curl}"] & \boldsymbol{V}^{p+\Delta p}(K) \arrow[r, "\mathrm{div}"] & Y^{p+\Delta p}(K)
\end{tikzcd} \tag{2.68}$$

where $p + \Delta p \geq 4$, $p \geq 1$, and the spaces coming from the Nédélec sequence of the first type,

$$\begin{aligned}
W^p(K) &= \mathcal{P}^p\,, \qquad \boldsymbol{Q}^p(K) = \mathcal{N}^p = \big(\mathcal{P}^{p-1}\big)^3 + \boldsymbol{x} \times \big(\mathcal{P}^{p-1}\big)^3\,,\\
\boldsymbol{V}^p(K) &= \mathcal{RT}^p = (\mathcal{P}^{p-1})^3 + \boldsymbol{x}\mathcal{P}^{p-1}\,, \qquad Y^p(K) = \mathcal{P}^{p-1}\,.
\end{aligned} \tag{2.69}$$

with $\mathcal{P}^p$ being the polynomials in $\boldsymbol{x} = (x_1, x_2, x_3)$ of total order at most p. The operators satisfy the following identities,

$$\big(\phi, v - \Pi^{p,\Delta p}_{F,\mathrm{grad},K} v\big)_K = 0 \qquad \forall \phi \in \mathcal{P}^{p+\Delta p-4},\ \forall v \in H^1(K)\,, \tag{2.70}$$

$$\big(\boldsymbol{\phi}, \boldsymbol{E} - \Pi^{p,\Delta p}_{F,\mathrm{curl},K} \boldsymbol{E}\big)_K = 0 \qquad \forall \boldsymbol{\phi} \in \big(\mathcal{P}^{p+\Delta p-3}\big)^3,\ \forall \boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, K)\,, \tag{2.71}$$

$$\big(\boldsymbol{\phi}, \nabla(v - \Pi^{p,\Delta p}_{F,\mathrm{grad},K} v)\big)_K = 0 \qquad \forall \boldsymbol{\phi} \in \big(\mathcal{P}^{p+\Delta p-3}\big)^3,\ \forall v \in H^1(K)\,, \tag{2.72}$$

$$\big(\boldsymbol{\phi}, \boldsymbol{v} - \Pi^{p,\Delta p}_{F,\mathrm{div},K} \boldsymbol{v}\big)_K = 0 \qquad \forall \boldsymbol{\phi} \in \big(\mathcal{P}^{p+\Delta p-2}\big)^3,\ \forall \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, K)\,, \tag{2.73}$$

$$\big(\boldsymbol{\phi}, \mathrm{curl}(\boldsymbol{E} - \Pi^{p,\Delta p}_{F,\mathrm{curl},K} \boldsymbol{E})\big)_K = 0 \qquad \forall \boldsymbol{\phi} \in \big(\mathcal{P}^{p+\Delta p-2}\big)^3,\ \forall \boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, K)\,, \tag{2.74}$$

$$\big(\phi, w - \Pi^{p,\Delta p}_{F,\int,K} w\big)_K = 0 \qquad \forall \phi \in \mathcal{P}^{p+\Delta p-1},\ \forall w \in L^2(K)\,, \tag{2.75}$$

$$\big(\phi, \mathrm{div}(\boldsymbol{v} - \Pi^{p,\Delta p}_{F,\mathrm{div},K} \boldsymbol{v})\big)_K = 0 \qquad \forall \phi \in \mathcal{P}^{p+\Delta p-1},\ \forall \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, K)\,, \tag{2.76}$$

$$\big\langle \mathrm{tr}^K_{\mathrm{div}} \boldsymbol{\phi}, \mathrm{tr}^K_{\mathrm{grad}}\big(v - \Pi^{p,\Delta p}_{F,\mathrm{grad},K} v\big)\big\rangle_{\partial K} = 0 \qquad \forall \boldsymbol{\phi} \in \boldsymbol{V}^{p+\Delta p-2}(K),\ \forall v \in H^1(K)\,, \tag{2.77}$$

$$\big\langle \mathbf{tr}^K_{\mathrm{curl},\top} \boldsymbol{\phi}, \mathbf{tr}^K_{\mathrm{curl},\dashv}\big(\boldsymbol{E} - \Pi^{p,\Delta p}_{F,\mathrm{curl},K} \boldsymbol{E}\big)\big\rangle_{\partial K} = 0 \qquad \forall \boldsymbol{\phi} \in (\mathcal{P}^{p+\Delta p-2})^3,\ \forall \boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, K)\,, \tag{2.78}$$

$$\big\langle \mathbf{tr}^K_{\mathrm{curl},\dashv} \boldsymbol{\phi}, \mathbf{tr}^K_{\mathrm{curl},\top}\big(\boldsymbol{F} - \Pi^{p,\Delta p}_{F,\mathrm{curl},K} \boldsymbol{F}\big)\big\rangle_{\partial K} = 0 \qquad \forall \boldsymbol{\phi} \in (\mathcal{P}^{p+\Delta p-2})^3,\ \forall \boldsymbol{F} \in \boldsymbol{H}(\mathrm{curl}, K)\,, \tag{2.79}$$

$$\big\langle \mathrm{tr}^K_{\mathrm{grad}} \phi, \mathrm{tr}^K_{\mathrm{div}}\big(\boldsymbol{v} - \Pi^{p,\Delta p}_{F,\mathrm{div},K} \boldsymbol{v}\big)\big\rangle_{\partial K} = 0 \qquad \forall \phi \in \mathcal{P}^{p+\Delta p-1},\ \forall \boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}, K)\,, \tag{2.80}$$

where the notation $\langle\,\cdot\,,\,\cdot\,\rangle_{\partial K}$ is expanded to include the two duality pairings between $\boldsymbol{H}^{-1/2}(\mathrm{curl}, \partial K)$ and $\boldsymbol{H}^{-1/2}(\mathrm{div}, \partial K)$, and where the outward element normal, $\hat{\mathbf{n}}_K$, defines the relevant traces as shown in Appendix A.

Remark 2.9. The continuity constants of the Fortin operators in Theorem 2.3, and in all other current high-order constructions [133, 179, 55], depend on the shape-regularity of the element and on $p+\Delta p$. The open question is whether the constants are independent of $p+\Delta p$ for sufficiently high $p+\Delta p$, meaning that for a fixed large enough Δp, the Fortin constant, C_Π, would be independent of p. Eventually, this could allow to demonstrate algebraic hp-convergence estimates, instead of merely h-convergence estimates. Numerical evidence suggests this is the case [179], but as it stands at the moment, no such claim has been mathematically proved. However, the reader should be reminded that Fortin operators are simply a mathematical tool that establishes a conservative bound for the discrete stability. In practice, the real discrete stability is observed to lie much closer to the ideal one and hp-estimates are often numerically detected. Moreover, computations with lower values of Δp than those advocated by the Fortin operator also result in the desired convergence behavior.

As an example, consider the variational formulation of Poisson's equation in Section 2.2 given in (2.10) on a shape-regular tetrahedral mesh. The discrete trial and test spaces were given in (2.59). Define a Fortin operator $\Pi_F : H^1(\mathcal{T}) \to \mathcal{V}_r$ by $(\Pi_F v)|_K = \Pi^{p,\Delta p}_{F,\mathrm{grad},K} v|_K$ for $K \in \mathcal{T}$ with the local Fortin operator coming from Theorem 2.3. Take $(\phi, \hat{\tau}_{\mathbf{n}}) \in \mathcal{U}_h$, and notice that for every tetrahedral element $K \in \mathcal{T}$,

$$\big(\nabla(\phi|_K), (\hat{\tau}_{\mathbf{n}})_K\big) \in \big(\mathcal{P}^{p-1}\big)^3 \times \mathrm{tr}^K_{\mathrm{div}}\big(\boldsymbol{V}^p(K)\big) \subseteq \big(\mathcal{P}^{p+\Delta p-3}\big)^3 \times \mathrm{tr}^K_{\mathrm{div}}\big(\boldsymbol{V}^{p+\Delta p-2}(K)\big)\,, \tag{2.81}$$

for $\Delta p \geq 2$, where $\boldsymbol{V}^p(K) = \mathcal{RT}^p$, so that using (2.72) and (2.78) it follows

$$b\big((\phi,\hat{\tau}_{\mathbf{n}}), v - \Pi_F v\big) = \big(\nabla\phi, \nabla(v-\Pi_F v)\big)_{\mathcal{T}} + \big\langle \hat{\tau}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}}(v-\Pi_F v)\big\rangle_{\partial\mathcal{T}} = 0 \quad \forall v \in H^1(\mathcal{T})\,. \tag{2.82}$$

This means that as long as $\Delta p \geq 2$ and $p+\Delta p \geq 4$ an existence of a Fortin operator is guaranteed.

2.8.2 Final estimates

This section simply gives the final convergence estimates. The main result comes from Theorem A.4 in Appendix A which provides h-convergence interpolation estimates for variables in the usual Sobolev spaces, $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl},\Omega)$, $\boldsymbol{H}(\mathrm{div},\Omega)$ and $L^2(\Omega)$, and their traces $H^{1/2}(\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)$ and $H^{-1/2}(\partial K)$. Then, combining this with Corollary 2.1 yields the desired high-order convergence result. The reader should consult Section A.5 and Section A.6 for subtleties on the definitions and terminology used in the theorem, although the main result is that the only dependence of the final estimates on the mesh is through the element size.

Theorem 2.4. *Let Ω be a polytopal domain and $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ be family of meshes of Ω comprised of simply connected polytopal elements $K\in\mathcal{T}_{\mathfrak{h}}$ with simply connected faces. Consider linear well-posed variational formulations associated to $b_{\mathfrak{h}} : \mathcal{U}_{\mathfrak{h}} \times \mathcal{V}_{\mathfrak{h}} \to \mathbb{R}$ and $\ell_{\mathfrak{h}} : \mathcal{V}_{\mathfrak{h}} \to \mathbb{R}$, where $\mathcal{U}_{\mathfrak{h}}$ and $\mathcal{V}_{\mathfrak{h}}$ are SdR spaces. Let $\{\mathfrak{u}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ be the exact solutions to the corresponding formulations, and assume they are attached to some $\mathfrak{u}_\Omega \in \mathcal{U}_\Omega$ through $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$, where $\mathcal{U}_\Omega$ is a compatible SdR space. Let $p \in \mathbb{N}$, $\mathcal{U}_{h,\mathfrak{h}} \subseteq \mathcal{U}_{\mathfrak{h}}$ be compatible SdR discretizations of order p, and $\mathcal{V}_{r,\mathfrak{h}} \subseteq \mathcal{V}_{\mathfrak{h}}$ be SdR discretizations. Suppose there exists a continuous linear Fortin operator $\Pi_{F,\mathfrak{h}} : \mathcal{V}_{\mathfrak{h}} \to \mathcal{V}_{r,\mathfrak{h}}$ such that for all $\mathfrak{v} \in \mathcal{V}_{\mathfrak{h}}$ and $\delta\mathfrak{u}_h \in \mathcal{U}_{h,\mathfrak{h}}$, $\|\Pi_{F,\mathfrak{h}}\mathfrak{v}\|_{\mathcal{V}_{\mathfrak{h}}} \leq C_\Pi \|\mathfrak{v}\|_{\mathcal{V}_{\mathfrak{h}}}$ and $b_{\mathfrak{h}}(\delta\mathfrak{u}_h, \mathfrak{v} - \Pi_{F,\mathfrak{h}}\mathfrak{v}) = 0$, for some $C_\Pi = C_\Pi(p) > 0$ that does not depend on the family of meshes $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$. Then, there exists a unique solution, $\mathfrak{u}_{h,\mathfrak{h}} \in \mathcal{U}_{h,\mathfrak{h}}$, solving the discrete variational formulation,*

$$b_{\mathfrak{h}}(\mathfrak{u}_{h,\mathfrak{h}}, \mathfrak{v}_{h,\mathfrak{h}}) = \ell_{\mathfrak{h}}(\mathfrak{v}_{h,\mathfrak{h}}) \qquad \forall \mathfrak{v}_{h,\mathfrak{h}} \in \mathcal{V}_{\mathfrak{h}}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_{r,\mathfrak{h}}}^{-1} \mathcal{B}_{\mathfrak{h}} \mathcal{U}_{h,\mathfrak{h}} \,, \tag{2.83}$$

where $\mathcal{R}_{\mathcal{V}_{r,\mathfrak{h}}}^{-1}$ is defined in (2.39) *and $\mathcal{B}_{\mathfrak{h}}$ is defined in* (2.25)*. Assume the attached exact solution $\mathfrak{u}_\Omega \in \mathcal{U}_\Omega^s$ for some $s > \frac{1}{2}$, where $\mathcal{U}_\Omega^s$ is the fractional counterpart to $\mathcal{U}_\Omega$. Then, provided all elements $K \in \mathcal{T}_{\mathfrak{h}}$ among all $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ are shape-regular, the following h-convergence estimate holds,*

$$\|\mathfrak{u}_{\mathfrak{h}} - \mathfrak{u}_{h,\mathfrak{h}}\|_{\mathcal{U}_{\mathfrak{h}}} \leq C h_{\mathfrak{h}}^{\min\{s,p\}} \|\mathfrak{u}_\Omega\|_{\mathcal{U}_\Omega^s} \,, \tag{2.84}$$

where $h_{\mathfrak{h}} = \sup_{K\in\mathcal{T}_{\mathfrak{h}}} \operatorname{diam}(K)$ and $C = C(s,p) > 0$. Moreover, if C_Π is p-independent as well and if all elements are either tetrahedra or hexahedra, then in the p-asymptotic limit an algebraic hp-estimate holds with $C = C_s(\ln p)^2 p^{-s}$ where $C_s = C(s)$ is independent of p.

Remark 2.10. When $\mathcal{V}^{\text{n-opt}} = \mathcal{V}^{\text{opt}}$, as in Remark 2.5 and Remark 2.6, then the assumption of the Fortin operator can be dropped altogether. Regarding the p-convergence estimate, the requirement that all the elements be either tetrahedra or hexahedra is due to the fact that these are the only elements for which there is a proof establishing a family of polynomial preserving extension operators with a bound independent of p [90, 99, 100, 101, 84]. If such a result can be proved for other elements (e.g. triangular prisms), then those elements can be included in the list too. Lastly, there is evidence that the logarithms in the p-convergence estimate may be removed [11, 172], but no rigorous proof in the context of projection-based interpolation has been given.

Finally, return to the variational formulation of Poisson's equation in Section 2.2 given in (2.10) and consider a family of shape-regular tetrahedral meshes. Clearly, the assumption of the Fortin operator in Theorem 2.4 holds provided $\Delta p \geq 2$ and $p + \Delta p \geq 4$ (see (2.82) for the existence of the Fortin operator). From Theorem 2.1, it is known that the underlying exact solution $\mathfrak{u}_0 = u \in H_0^1(\Omega)$ is the same, regardless of the mesh being considered, and it is easily observed that $\hat{\mathfrak{u}} = \hat{q}_{\mathbf{n}} = \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}}(-\nabla u)$ for every mesh $\mathcal{T}$, so that all the exact solutions $(\mathfrak{u}_0, \hat{\mathfrak{u}})$ are attached to $\mathfrak{u}_\Omega = (u, -\nabla u) \in H_0^1(\Omega) \times \boldsymbol{H}(\mathrm{div}, \Omega)$ through the corresponding family of meshes. Thus, the h-convergence estimate holds, and (2.84) can be explicitly rewritten as,

$$\|u - u_h\|_{H^1(\Omega)}^2 + \|\hat{q}_{\mathbf{n}} - \hat{q}_{\mathbf{n},h}\|_{H^{-1/2}(\partial\mathcal{T})}^2 \leq C^2 h^{2\min\{s,p\}} \left(\|u\|_{H^{1+s}(\Omega)}^2 + \|\nabla u\|_{\boldsymbol{H}^s(\mathrm{div},\Omega)}^2 \right), \tag{2.85}$$

where $u \in H^{1+s}(\Omega)$ (for some $s > \frac{1}{2}$) is the exact solution, $(u_h, \hat{q}_{\mathbf{n},h})$ are the discrete solutions after solving at $\mathcal{T}$, and h is the maximum element size in $\mathcal{T}$. The p-convergence estimate is only true if for large enough Δp, the bound of the Fortin operator is proved to be independent of p.

2.9 Numerical implementation

Based on Section 2.3, it should be clear how to computationally implement DPG methods. However, some of the more explicit details might be missing. This section aims to at least partially cover some of those minor gaps, show some optimizations, and to review the current software with support for DPG methods.

As is conventional in finite element methods, the stiffness matrix and load, $\mathsf{B}^{\text{n-opt}}$ and $\mathsf{l}^{\text{n-opt}}$, are assembled not by evaluating each component (i.e. computing $\mathsf{B}_{11}^{\text{n-opt}}$, then $\mathsf{B}_{12}^{\text{n-opt}}$, etc.), but by looking at the contribution of each element $K \in \mathcal{T}$ to $\mathsf{B}^{\text{n-opt}}$ and $\mathsf{l}^{\text{n-opt}}$ and looping over the elements. The local contributions are made by only considering the basis functions that have support intersecting that particular element. As previously mentioned, this is not possible for general minimum residual methods, but the broken test spaces discretized in DPG methods provide a natural and compatible decoupling of G, B and l, which allows the global expressions $\mathsf{B}^{\mathsf{T}}\mathsf{G}^{-1}\mathsf{B}$ and $\mathsf{B}^{\mathsf{T}}\mathsf{G}^{-1}\mathsf{l}$ to be written as a sum of properly assembled local contributions of the same type,

$$\mathsf{B}_K^{\text{n-opt}} = \mathsf{B}_K^{\mathsf{T}}\mathsf{G}_K^{-1}\mathsf{B}_K\,, \qquad \mathsf{l}_K^{\text{n-opt}} = \mathsf{B}_K^{\mathsf{T}}\mathsf{G}_K^{-1}\mathsf{l}_K\,, \tag{2.86}$$

where for every $K \in \mathcal{T}$, G_K, B_K and l_K are local Gram matrices, enriched (rectangular) stiffness matrices, and loads. The connectivities used in the assembly of the $\mathsf{B}_K^{\text{n-opt}}$ and $\mathsf{l}_K^{\text{n-opt}}$ are the ones traditionally used by finite element codes. The structure of these local variables is very important, because it can be exploited computationally. Indeed, notice that if L_K is the Cholesky factorization of $\mathsf{G}_K = \mathsf{L}_K\mathsf{L}_K^{\mathsf{T}}$, then $\mathsf{G}_K^{-1} = \mathsf{L}_K^{-\mathsf{T}}\mathsf{L}_K^{-1}$, and the local near-optimal stiffness matrix and load become

$$\mathsf{B}_K^{\text{n-opt}} = (\mathsf{L}_K^{-1}\mathsf{B}_K)^{\mathsf{T}}(\mathsf{L}_K^{-1}\mathsf{B}_K)\,, \qquad \mathsf{l}_K^{\text{n-opt}} = (\mathsf{L}_K^{-1}\mathsf{B}_K)^{\mathsf{T}}\mathsf{L}_K^{-1}\mathsf{l}_K\,. \tag{2.87}$$

This is in contrast with computing $\mathsf{G}_K^{-1}\mathsf{B}_K$ first (which typically involves computing the Cholesky factorization anyway), and then $\mathsf{B}_K^{\mathsf{T}}\mathsf{G}_K^{-1}\mathsf{B}_K$. Indeed, exploiting this structure provides a slight computational optimization if correctly implemented (by using the right `BLAS` and `LAPACK` routines). Other factorizations producing a similar result as above, like $\mathsf{G}_K = (\mathsf{G}_K^{1/2})^{\mathsf{T}}\mathsf{G}_K^{1/2}$, are possible, but much more complicated and expensive to compute. The a posteriori error estimator is subject to the same optimization, since (2.50) can be rewritten as

$$\eta_K^2 = (\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K)^{\mathsf{T}}\mathsf{G}_K^{-1}(\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K) = \big(\mathsf{L}_K^{-1}(\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K)\big)^{\mathsf{T}}\big(\mathsf{L}_K^{-1}(\mathsf{B}_K\mathsf{u}_{h,K} - \mathsf{l}_K)\big)\,. \tag{2.88}$$

This type of optimization was first reported in [113, 111], and since then it has been exploited at a global level to develop a new family of finite element methods (the subject of Chapter 6) [160].

To compute the individual components in G_K, B_K and l_K, usually shape functions local to each element are considered. These are basically restrictions of the relevant global basis functions $\{\mathfrak{u}_j\}_{j=1}^N$ and $\{\mathfrak{v}_i\}_{i=1}^M$ to $K \in \mathcal{T}$. The shape functions are essentially bases for the local discrete spaces $\mathcal{U}_h^p(K)$ and $\mathcal{V}_r^{p+\Delta p}(K)$ described in Section 2.7, which are in turn composed of products of the discrete spaces $W^p(K)$, $\boldsymbol{Q}^p(K)$, $\boldsymbol{V}^p(K)$, $Y^p(K)$, $W^p(\partial K)$, $\boldsymbol{Q}_{\dashv}^p(\partial K)$, $\boldsymbol{Q}_{\top}^p(\partial K)$ and $V^p(\partial K)$ proposed in (2.54) and (2.55). As usual, one must be careful in satisfying interelement compatibilities, and this implies that the shape functions must be chosen such that they have the correct orientation at each face in order to match the shape functions from the adjacent element (thus producing a properly defined globally conforming basis element). As mentioned previously, these spaces and compatibilities depend on the type of element $K \in \mathcal{T}$ and can become more elaborate to satisfy when dealing with high-order methods, especially if local p-refinements are a desired feature.

Fortunately, all those features are supported and covered by the hierarchical orientation-embedded high-order shape functions proposed in [114]. This work gives a unified coverage of all the conventional elements in 1D, 2D (quadrilaterals and triangles), and 3D (hexahedra, tetrahedra, triangular prisms, and pyramids) and provides discretizations of arbitrary high-order for $H^1(K)$, $\boldsymbol{H}(\mathrm{curl}, K)$, $\boldsymbol{H}(\mathrm{div}, K)$ and $L^2(K)$ of the type described (2.54). These discretizations have also been referred in this document as compatible SdR discretizations (see Section A.5 in Appendix A). For each element, the hierarchical shape functions are classified by association with topological entities: vertices, edges, faces, and interior. This, along with the concept of orientation embedding, facilitates the computation of compatible shape functions across different elements, while also allowing for the possibility of local p-refinements. In particular, the sequences of discrete spaces for tetrahedra and hexahedra in [114] are the classical Nédélec sequences of the first type [181, 103, 34]. Other explicit shape functions with some of these features can be found in [24, 25, 57, 205].

The classification of interior shape functions, also referred to as bubbles, is important in high-order shape functions, because these correspond to basis functions that only have support in a given element $K \in \mathcal{T}$. The number of bubbles obviously grows with p. Their local support implies that they are at least partially decoupled algebraically from the rest of the degrees of freedom, so the Schur complement associated to these bubbles can be computed. This is referred to as static condensation. It is done at the local level (so can be computed in parallel), and it results in a considerably smaller global matrix, $\mathsf{B}^{\text{n-opt,c}}$, which makes the implementation of high-order methods much more efficient. Later, the degrees of freedom associated to the bubbles can be recovered. This will be covered in more detail in Chapter 6. Previous to the static condensation, however, some modifications are recommended to account for the degrees of freedom associated to Dirichlet boundary conditions and to account for constrained approximations associated to hanging nodes resulting from adaptive mesh refinements. For these, we refer to [89, 103].

2.9.1 DPG software

There are a few software packages that were explicitly designed to support DPG methods. The main ones are `Camellia` [199, 200], `DUNE-DPG` [136], `hp3d`, `hp2d`, and `PolyDPG` [229]. The codes `hp3d` and `hp2d` are based on the same architecture, and the ideas are similar to those outlined in [89, 103]. Among the 3D codes, the ones with the most features are `Camellia` and `hp3d`.

The package used throughout most of this dissertation is the in-house code `hp3d` written in `Fortran 90`. The shape functions were taken from [114], as described previously, so `hp3d` supports compatible discretizations of all the conventional element shapes (hexahedra, tetrahedra, triangular prisms and pyramids) and for all the spaces $H^1(\Omega)$, $\boldsymbol{H}(\text{curl},\Omega)$, $\boldsymbol{H}(\text{div},\Omega)$ and $L^2(\Omega)$ and their relevant traces. Thus, it can discretize hybrid meshes containing elements of different types in the same mesh. As its name suggests, `hp3d` supports both h and p local anisotropic refinements via constrained approximations and hanging nodes. Moreover, it has sophisticated multi-physics support (facilitating the identification of variables to couple between distinct subdomains), projection-based interpolation of all the spaces [103, 90] (to enforce non-homogeneous boundary conditions), the ability to handle isoparametric geometries with inherited curvilinear refinements through local transfinite interpolation [103, 123], and support for complex-valued variables. The assembly and computation of the residuals was done using multiple threads via `OpenMP`. The direct solvers used for the global solution of the problems in this dissertation were `MUMPS 5.0.1` [168, 3] and `Intel MKL PARDISO` (and `qr_mumps 1.2` [48] in Chapter 6).

The package `PolyDPG` [229, 228] was used in Chapter 7 to implement (2D) high-order polygonal DPG (PolyDPG) methods, as described there. It is based in `MATLAB`$^{®}$, but it also uses the shape functions in [114, 112]. Lastly, as its name suggests, it has support for polygonal elements and polygonal mesh refinement strategies.

Chapter 3

DPG methods for linear elasticity

This chapter is the combination of two published research papers by the author ([158]* and [113]†). It is about the application of the DPG methodology to the equations of linear elasticity. The main purpose for being included in this dissertation is twofold. First, it shows the versatility of the DPG methodology in being able to discretize distinct variational formulations. Second, it shows the framework and motivation on coupling different formulations within the same domain using DPG methods. The contributions of the author to the multi-authored articles ranged from the computational implementation of the numerical methods to the development of the theory, the mathematical proofs, and the writing of the manuscripts.

3.1 Introduction

In this chapter we demonstrate the fitness of the discontinuous Petrov-Galerkin (DPG) methodology by applying it to various variational formulations of the equations of linear elasticity. These equations are often solved by applying the Bubnov-Galerkin finite element methodology to the classical displacement-based variational formulation, which is derived from the second-order equations of linear elasticity. Alternatively, it is the formulation resulting from the minimization of the potential energy functional [68, 154]. In a similar fashion, the minimization of the Hellinger-Reissner energy functional leads to the other well-known mixed variational formulation that can also be solved using the Bubnov-Galerkin methodology [38]. At the infinite-dimensional level, under suitable regularity assumptions, they can be shown to be equivalent, since they result in exactly the same solution [70, 108], namely, the unique solution of the equations of linear elasticity. However, at

* Keith, B., Fuentes, F., and Demkowicz, L. (2016). The DPG methodology applied to different variational formulations of linear elasticity. *Comput. Methods Appl. Mech. Engrg.*, 309:579–609.

† Fuentes, F., Keith, B., Demkowicz, L., and Le Tallec, P. (2017b). Coupled variational formulations of linear elasticity and the DPG methodology. *J. Comput. Phys.*, 348:715–731.

the discrete finite-dimensional level, which is the setting involving finite element methods, there are differences that might be important. For example, the mixed formulation results in a discretization which is not as efficient computationally, but remains robustly well-posed for nearly incompressible materials (it avoids volumetric locking) and also guarantees a locally conservative stress tensor [110, 38, 8]. Likewise, other energy principles leading to distinct variational formulations are possible. In fact, just for this single problem, a total of fourteen complementary-dual energy principles are presented in [190], each leading to a different variational formulation. Some may not be easily amenable to computation, but perspective should be given that there is little to regard as sacred or more physical about one formulation over another.

In this chapter we present eight different variational formulations that solve the equations of linear elasticity, mostly derived naturally by formal integration by parts of the equations written in their first-order form. We include a proof of what we believe is the contemporary observation that all of the variational formulations which we have considered are mutually well- or ill-posed (a similar assertion has been proved in the context of Maxwell equations in [55]). This is important because it avoids having to present an independent proof of well-posedness for each different formulation.

Often finite element methodologies cannot be applied to particular variational formulations, which might make them seem like they are not accommodated for computation. Indeed, some formulations have different trial and test spaces, in which case they are said to have a non-symmetric functional setting. What is powerful about minimum residual finite element methodologies, is that as long as there is a discretization of the trial and test spaces, they can be used to solve these non-conventional formulations, while at the same time being crafted to produce numerical stability. The DPG methodology is a minimum residual methodology where the test spaces are broken (discontinuous) along the mesh. In this chapter, we derive broken variational formulations for four out of the eight formulations proposed initially, and show they are also well-posed. Moreover, the DPG methodology, which is compatible with these formulations, is used to develop four distinct high-order DPG methods to numerically solve the equations in 3D. This is done to show the versatility of DPG methods in solving different formulations. Examples involving smooth and singular solutions are considered in order to corroborate the theory. Moreover, the different behavior of the residual-based a posteriori error estimators coming from the four DPG methods is analyzed. Other work involving DPG methods and 2D linear elasticity include [36, 133, 56, 54].

Additionally, this chapter studies the scenario where distinct variational formulations are implemented in different subdomains of the same physical domain. This can be useful in situations where a certain behavior of the equations to be solved is known (or expected) in particular parts of the domain. Hence, in each region one can choose a variational formulation which is well-suited to the expected behavior. For example, consider a material with heterogeneous material properties varying within the domain. The properties can vary continuously, as in cloaking applications or biological materials, or discontinuously, as in multi-material problems. Then, in the parts of the domain where it can be an issue (e.g. a nearly incompressible material in linear elasticity), one can choose a variational formulation that is robustly well-posed with respect to the material properties. In the remaining regions, where such robustness is not fundamental, one can choose a more computationally efficient formulation. Another example occurs when a near singularity is expected in a particular area, so that one would hope to use a variational formulation (with possibly an associated adaptivity scheme) which is desirable in that subdomain, but not necessarily in the entire physical domain [204].

The main issue with such an implementation arises at the interfaces between the two subdomains having distinct variational formulations. At this interface, information must pass between the two subdomains to enable communication. This imposes a coupling with both theoretical and practical compatibility issues which can be difficult to resolve and analyze. Moreover, the coupling must be constructed properly so that the entire problem is well-posed. This is not immediate, even if each of the interacting variational formulations is well-posed when considered independently across the whole domain.

At the theoretical and infinite-dimensional level, an attractive possibility that naturally unburdens the compatibility and well-posedness requirements is the use of broken variational formulations. These mesh-dependent formulations are extensions of the usual variational formulations to the case involving broken (or discontinuous) test spaces, which are precisely the formulations acquiescent to the DPG methodology. The family comprised of the four well-posed broken formulations mentioned before, is analyzed in this setting. Those formulations will be observed to inescapably possess interface variable unknowns which are a desirable means of communicating the necessary solution variable information across subdomains. This is what allows introducing a proper definition of *coupled* variational formulations, which we will later prove to be globally well-posed as well. Thus, another goal of this chapter is to demonstrate the use of the DPG methodology in solving the equations of linear elasticity via coupled variational formulations. The derived DPG method will be used to discretize and solve the coupled formulation in 3D and with high-order methods while retaining numerical stability. The use of the DPG methodology will corroborate the expected theoretical convergence results in this heterogeneous functional setting. Examples showing the viability of the approach at a practical level will be illustrated, including a case where the demanding scenario of a fully incompressible material is considered. This last case has physical applications in modeling steel braided rubber hoses and even stents.

Regarding the coupling of formulations, similarity exists between the approach in this work and that taken in [145] (used for elliptic transmission problems). There, a variational formulation similar to those considered here is coupled with a variational formulation composed of boundary integral operators. Afterward, the coupled formulation is discretized with the DPG methodology throughout the entire computational domain. A remark is also warranted for the contributions in [117, 119] where the ideas in [145] are extended to couple the DPG methodology with more standard boundary element methods (BEMs), so that different discretization methods are considered across the domain. More recently, there have also been further developments [146, 120].

This chapter is organized as follows. In Section 3.2 first the non-conventional Hilbert spaces involved in linear elasticity are defined (for the conventional spaces see Appendix A). Then, the equations of linear elasticity are introduced and eight different variational formulations are derived. The proof that all these formulations are actually mutually well-posed is relegated to Appendix B. Next, four broken variational formulations are deduced and it is argued that they are well-posed too. In Section 3.2.6 coupled variational formulations are described. Their proof of well-posedness is also left to Appendix B. In Section 3.3, numerical results are presented for the subset of the four broken formulations and their coupled counterparts. Smooth solutions are used to corroborate the expected numerical results for all the formulations being considered. Then, a singular solution involving an L-shaped domain is solved for using the four DPG methods derived, and their adaptive behavior (of their built-in a posteriori error estimators) is compared. Finally, an illustrative, physically-relevant, and challenging example of a sheathed hose is examined and solved using coupled variational formulations.

3.2 Variational formulations in linear elasticity

3.2.1 Hilbert spaces in linear elasticity

Before introducing the equations of linear elasticity, it is useful to define the functional spaces relevant to the problem. The conventional spaces are defined in Appendix A. Fortunately, most formulations will only involve copies of those spaces, including all of the formulations which will be numerically solved. The notation for those spaces is only slightly different, and we introduce it for the sake of compactness and brevity. Additionally, there are other non-conventional spaces important in linear elasticity, which will be defined below.

The vector- and matrix-valued Sobolev spaces on a domain $K \subseteq \mathbb{R}^3$ are,

$$\begin{aligned}
\boldsymbol{L}^2(K) &= \big\{\boldsymbol{u} : K \to \mathbb{R}^3 \mid u_i \in L^2(K), i = 1,2,3\big\} = \big(L^2(K)\big)^3\,,\\
\mathsf{L}^2(K) &= \big\{\boldsymbol{\sigma} : K \to \mathbb{M} \mid \sigma_{ij} \in L^2(K), i,j = 1,2,3\big\}\,,\\
\mathsf{L}^2(K;\mathbb{S}) &= \big\{\boldsymbol{\sigma} : K \to \mathbb{S} \mid \boldsymbol{\sigma} \in \mathsf{L}^2(K)\big\} \subseteq \mathsf{L}^2(K)\,,\\
\mathsf{L}^2(K;\mathbb{A}) &= \big\{\boldsymbol{\sigma} : K \to \mathbb{A} \mid \boldsymbol{\sigma} \in \mathsf{L}^2(K)\big\} \subseteq \mathsf{L}^2(K)\,,\\
\boldsymbol{H}^1(K) &= \big\{\boldsymbol{u} : K \to \mathbb{R}^3 \mid u_i \in H^1(K), i = 1,2,3\big\} = \big(H^1(K)\big)^3\,,\\
\mathsf{H}(\mathbf{div},K) &= \big\{\boldsymbol{\sigma} : K \to \mathbb{M} \mid (\sigma_{i1},\sigma_{i2},\sigma_{i3}) \in \boldsymbol{H}(\mathrm{div},K), i = 1,2,3\big\} = \big(\boldsymbol{H}(\mathrm{div},K)\big)^3\,,\\
\mathsf{H}(\mathbf{div},K;\mathbb{S}) &= \big\{\boldsymbol{\sigma} : K \to \mathbb{S} \mid \boldsymbol{\sigma} \in \mathsf{H}(\mathbf{div},K)\big\} \subseteq \mathsf{H}(\mathbf{div},K)\,,
\end{aligned} \tag{3.1}$$

where $\mathbb{M} = \mathbb{R}^{3\times 3}$ are the real-valued 3×3 matrices, $\mathbb{S} \subseteq \mathbb{M}$ are the subset of symmetric matrices and $\mathbb{A} \subseteq \mathbb{M}$ are the subset of antisymmetric (or skew-symmetric) matrices. Note the divergence in $\mathsf{H}(\mathbf{div},K)$ and $\mathsf{H}(\mathbf{div},K;\mathbb{S})$ is taken row-wise. The norms are the Hilbert norms which are naturally induced. The trace spaces are,

$$\boldsymbol{H}^{1/2}(\partial K) = \big(H^{1/2}(\partial K)\big)^3\,, \qquad \boldsymbol{H}^{-1/2}(\partial K) = \big(H^{-1/2}(\partial K)\big)^3\,, \tag{3.2}$$

with their induced Hilbert norms. Meanwhile, for $\boldsymbol{u} \in \boldsymbol{H}^1(K)$ and $\boldsymbol{\sigma} \in \mathsf{H}(\mathbf{div}, K)$ the traces are,

$$\big(\mathbf{tr}^K_{\mathrm{grad}}\boldsymbol{u}\big)_i = \mathrm{tr}^K_{\mathrm{grad}} u_i \,, \qquad \big(\mathbf{tr}^K_{\mathrm{div}}\boldsymbol{\sigma}\big)_i = \mathrm{tr}^K_{\mathrm{div}}(\sigma_{i1}, \sigma_{i2}, \sigma_{i3}) \,, \qquad i = 1, 2, 3 \,. \tag{3.3}$$

Next, if $\Omega \subseteq \mathbb{R}^3$ is a domain, the spaces with boundary conditions, $\boldsymbol{H}^1_{\Gamma_u}(\Omega)$, $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$, $\boldsymbol{H}^{1/2}_{\Gamma_u}(\partial\Omega)$ and $\boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\Omega)$, are defined analogously to their simpler counterparts, where $\Gamma_u \subseteq \partial\Omega$ and $\Gamma_\sigma \subseteq \partial\Omega$ are relatively open in $\partial\Omega$. Meanwhile, $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S}) = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \cap \mathsf{H}(\mathbf{div}, \Omega; \mathbb{S})$.

Lastly, if $\Omega \subseteq \mathbb{R}^3$ is partitioned into a mesh, $\mathcal{T}$, then it is easy to analogously define $\boldsymbol{H}^1(\mathcal{T})$, $\mathsf{H}(\mathbf{div}, \mathcal{T})$, $\boldsymbol{H}^{1/2}(\partial\mathcal{T})$ and $\boldsymbol{H}^{-1/2}(\partial\mathcal{T})$ as the corresponding copies of the broken spaces $H^1(\mathcal{T})$, $\boldsymbol{H}(\mathrm{div}, \mathcal{T})$, $H^{1/2}(\partial\mathcal{T})$ and $H^{-1/2}(\partial\mathcal{T})$. The same is true for the vector trace operators, $\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}$ and $\mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}$. For $\boldsymbol{L}^2(\mathcal{T})$, $\mathsf{L}^2(\mathcal{T})$, $\mathsf{L}^2(\mathcal{T}; \mathbb{S})$ and $\mathsf{L}^2(\mathcal{T}; \mathbb{A})$, they are simply equal to $\boldsymbol{L}^2(\Omega)$, $\mathsf{L}^2(\Omega)$, $\mathsf{L}^2(\Omega; \mathbb{S})$ and $\mathsf{L}^2(\Omega; \mathbb{A})$ respectively.

Note that with the exception of $\mathsf{H}(\mathbf{div}, K; \mathbb{S})$, it is very easy to discretize these spaces, since $\boldsymbol{L}^2(K)$ is three copies of $L^2(K)$, $\mathsf{L}^2(K)$ is nine copies of $L^2(K)$, $\mathsf{L}^2(K; \mathbb{S})$ is six copies of $L^2(K)$ (but the norm measured in $\mathsf{L}^2(K)$), $\mathsf{L}^2(K; \mathbb{A})$ is three copies of $L^2(K)$ (but the norm measured in $\mathsf{L}^2(K)$), $\boldsymbol{H}^1(K)$ is three copies of $H^1(K)$, $\mathsf{H}(\mathbf{div}, K)$ is three copies of $\boldsymbol{H}(\mathrm{div}, K)$, $\boldsymbol{H}^{1/2}(\partial K)$ is three copies of $H^{1/2}(\partial K)$, and $\boldsymbol{H}^{-1/2}(\partial K)$ is three copies of $H^{-1/2}(\partial K)$. Thus, simply take copies of the discretizations of the underlying space. Regarding the space $\mathsf{H}(\mathbf{div}, K; \mathbb{S})$, it is notoriously difficult to discretize as a high-order space with mathematically desirable properties [7, 10, 153], and for that reason it is often avoided when it comes to computations.

3.2.2 Linear elasticity equations

In this work, the classical equations of static linear elasticity will be solved [67]. These are simply the linearization in the reference configuration about a stress-free state of the general constitutive model for solids and the conservation of momentum in the static case. The equations of linear elasticity in a domain $\Omega \subseteq \mathbb{R}^3$ can be written as follows,

$$-\,\mathbf{div}(\mathsf{C}\!:\!\boldsymbol{\varepsilon}(\boldsymbol{u})) = \boldsymbol{f} \qquad \Leftrightarrow \qquad \begin{cases} \boldsymbol{\sigma} - \mathsf{C}\!:\!\boldsymbol{\varepsilon}(\boldsymbol{u}) = 0 \,, \\ \qquad -\,\mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f} \,, \end{cases} \tag{3.4}$$

where $\boldsymbol{u}$ is the displacement, $\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^{\mathsf{T}})$ is its associated strain, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathsf{T}}$ is the stress (which must be symmetric in order to satisfy conservation of angular momentum), and $\boldsymbol{f}$ is the known body force. Lastly, $\mathsf{C} : \mathbb{S} \to \mathbb{S}$ is the stiffness tensor. For isotropic materials,

$$\mathsf{C}_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\,, \quad \mathsf{S}_{ijkl} = \frac{1}{4\mu}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) - \frac{\lambda}{2\mu(3\lambda+2\mu)}\delta_{ij}\delta_{kl}\,, \tag{3.5}$$

where $\mathsf{C}^{-1} = \mathsf{S} : \mathbb{S} \to \mathbb{S}$ is the compliance tensor, and λ and μ are the Lamé parameters. All variables are assumed to be appropriately nondimensionalized.

The constitutive equation can be rewritten as

$$\mathsf{S}:\boldsymbol{\sigma} - \boldsymbol{\varepsilon}(\boldsymbol{u}) = 0\,. \tag{3.6}$$

This form is preferred when dealing with nearly incompressible materials (as $\lambda \to \infty$) because the norm of S remains finite, while that of C diverges. This is the underlying reason why using this form in a variational setting prevents volumetric locking phenomena.

Note the equations are written in their first- and second-order forms in (3.4). This single second-order equation with the corresponding boundary conditions is the starting point for the more traditional variational formulations. However, as it will be seen, using the first-order system gives more versatility to construct variational formulations.

The goal is to solve the equations in (3.4) for the unknown displacement and stress, provided the forcing and the dynamic stiffness tensor of the material are known throughout the domain $\Omega \subseteq \mathbb{R}^3$. For this to be possible, boundary conditions need to be specified, so it will be assumed that the boundary is partitioned into relatively open subsets $\Gamma_u \subseteq \partial\Omega$ and $\Gamma_\sigma \subseteq \partial\Omega$ satisfying $\overline{\Gamma_u \cup \Gamma_\sigma} = \partial\Omega$ and $\Gamma_u \cap \Gamma_\sigma = \varnothing$, where displacement and traction boundary conditions are set by the known functions $\boldsymbol{u} = \boldsymbol{u}^{\Gamma_u}$ and $(\mathsf{C}^* : \boldsymbol{\nabla u})\cdot\hat{\mathbf{n}} = \boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$ on Γ_u and Γ_σ respectively, with $\hat{\mathbf{n}}$ being the outward normal at $\partial\Omega$. From now on it will be assumed that $\Gamma_u \neq \varnothing$ and Ω is bounded and Lipschitz.

3.2.3 Basic variational testing

For simplicity, assume for now that there are no traction boundary conditions, so that $\Gamma_\sigma = \varnothing$, and $\Gamma_u = \partial\Omega$. If we assume that $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$, the conservation law is equivalent to the variational equation,

$$-\,(\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_\Omega = (\boldsymbol{f},\boldsymbol{v})_\Omega \qquad \forall \boldsymbol{v}\in \boldsymbol{L}^2(\Omega)\,. \tag{3.7}$$

Due to the symmetry of the stress tensor, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\mathsf{T}}$, it is natural to consider $\boldsymbol{\sigma} \in \mathbf{H}(\mathbf{div},\Omega;\mathbb{S})$. However, as mentioned previously, the space $\mathbf{H}(\mathbf{div},\Omega;\mathbb{S})$ is difficult to discretize [153, 7, 10, 195, 192, 193], so it is typically impractical to consider this space. Instead it is often assumed $\boldsymbol{\sigma} \in \mathbf{H}(\mathbf{div},\Omega)$, with the symmetry of $\boldsymbol{\sigma}$ being imposed weakly through the extra equation,

$$(\boldsymbol{\sigma},\boldsymbol{w})_\Omega = 0 \qquad \forall \boldsymbol{w}\in \mathbf{L}^2(\Omega;\mathbb{A})\,. \tag{3.8}$$

Formally integrating (3.7) by parts gives an equation closely related to the principle of virtual work,

$$(\boldsymbol{\sigma},\boldsymbol{\nabla}\boldsymbol{v})_\Omega = (\boldsymbol{f},\boldsymbol{v})_\Omega \qquad \forall \boldsymbol{v}\in \boldsymbol{H}_0^1(\Omega)\,. \tag{3.9}$$

Here, to enforce the symmetry it makes sense to take $\boldsymbol{\sigma}\in\mathbf{L}^2(\Omega;\mathbb{S})$ which is easy to discretize. Note that $\boldsymbol{v}\in\boldsymbol{H}_0^1(\Omega)$ in (3.9), while $\boldsymbol{v}\in\boldsymbol{L}^2(\Omega)$ in (3.7).

Likewise, after testing with $\boldsymbol{\tau}$, the constitutive law in (3.4) may be written as,

$$(\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega - (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\tau})_\Omega = 0 \qquad \forall \boldsymbol{\tau}\in\mathbf{L}^2(\Omega;\mathbb{S})\,, \tag{3.10}$$

where it was used $\mathsf{C}:\boldsymbol{\varepsilon}(\boldsymbol{u}) = \mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u}$, with the domain of $\mathsf{C}:\mathbb{S}\to\mathbb{S}$ being extended naturally to $\mathsf{C}:\mathbb{M}\to\mathbb{S}$ (i.e., $\mathsf{C}|_{\mathbb{A}}=0$). Here, due to the presence of $\boldsymbol{\nabla}\boldsymbol{u}$, it makes sense to have $\boldsymbol{u}\in\widetilde{\boldsymbol{u}}^{\partial\Omega}+\boldsymbol{H}_0^1(\Omega)$, where $\widetilde{\boldsymbol{u}}^{\partial\Omega}\in\boldsymbol{H}^1(\Omega)$ is an extension of the prescribed boundary displacement $\boldsymbol{u}^{\partial\Omega}$ from $\partial\Omega$ to Ω.

To get an alternate variational form of the constitutive equation it is more convenient to consider the characterization provided in (3.6). This equation is easier to integrate by parts and avoids volumetric locking in the limit of incompressible materials, as alluded previously. Testing with $\boldsymbol{\tau}$ yields the expression $(\boldsymbol{\varepsilon}(\boldsymbol{u}):\boldsymbol{\tau})_\Omega$, which cannot be integrated by parts unless $\boldsymbol{\tau}=\boldsymbol{\tau}^{\mathsf{T}}$, in which case $\boldsymbol{\varepsilon}(\boldsymbol{u}):\boldsymbol{\tau} = \boldsymbol{\nabla}\boldsymbol{u}:\boldsymbol{\tau}$ and $(\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\tau})_\Omega = -(\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega$. Thus, it makes sense to test using $\boldsymbol{\tau}\in\mathbf{H}(\mathbf{div},\Omega;\mathbb{S})$, and integration of (3.6) by parts yields,

$$(\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega = \langle \boldsymbol{u}^{\partial\Omega},\mathbf{tr}_{\mathrm{div}}^{\Omega}\boldsymbol{\tau}\rangle_{\partial\Omega} \qquad \forall\boldsymbol{\tau}\in\mathbf{H}(\mathbf{div},\Omega;\mathbb{S})\,. \tag{3.11}$$

However, this revives the difficulties of discretizing $\mathsf{H}(\mathbf{div},\Omega;\mathbb{S})$. To overcome the issue, one must introduce an extra antisymmetric solution variable called the infinitesimal rotation tensor, $\boldsymbol{\omega}$, which satisfies

$$\boldsymbol{\nabla}\boldsymbol{u} = \boldsymbol{\varepsilon}(\boldsymbol{u}) + \boldsymbol{\omega} \qquad \Rightarrow \qquad \mathsf{S}:\boldsymbol{\sigma} - \boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\omega} = 0\,. \tag{3.12}$$

Testing now with $\mathsf{H}(\mathbf{div},\Omega)$ and integrating by parts then yields,

$$(\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega}:\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega = \langle \boldsymbol{u}^{\partial\Omega}, \mathbf{tr}_{\mathrm{div}}^{\Omega}\boldsymbol{\tau}\rangle_{\partial\Omega} \qquad \forall \boldsymbol{\tau}\in\mathsf{H}(\mathbf{div},\Omega)\,, \tag{3.13}$$

where the domain of S is extended trivially from $\mathbb{S}$ to $\mathbb{M}$ (i.e., $\mathsf{S}|_{\mathbb{A}} = 0$). Here, it is natural to consider $\boldsymbol{u}\in\boldsymbol{L}^2(\Omega)$ and $\boldsymbol{\omega}\in\mathsf{L}^2(\Omega;\mathbb{A})$, which are both easy to discretize.

3.2.4 A family of variational formulations

A linear variational formulation is a problem which seeks $\mathfrak{u}\in\mathcal{U}$ such that

$$b(\mathfrak{u},\mathfrak{v}) = \ell(\mathfrak{v}) \qquad \forall \mathfrak{v}\in\mathcal{V}\,, \tag{3.14}$$

where $\mathcal{U}$ and $\mathcal{V}$ are trial and test Hilbert spaces over a fixed field $\mathbb{F}\in\{\mathbb{R},\mathbb{C}\}$, $b:\mathcal{U}\times\mathcal{V}\to\mathbb{F}$ is a continuous bilinear form if $\mathbb{F}=\mathbb{R}$ or sesquilinear form if $\mathbb{F}=\mathbb{C}$, and $\ell\in\mathcal{V}'$ is a continuous linear form if $\mathbb{F}=\mathbb{R}$ or antilinear form if $\mathbb{F}=\mathbb{C}$.

Using the distinct choices to discretize the equations, as shown in the last section, yields seven different variational formulations,

$$\begin{aligned}&\mathcal{U}^{\mathcal{S}_\mathbb{S}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})\times\boldsymbol{H}^1_{\Gamma_u}(\Omega)\,, \qquad \mathcal{V}^{\mathcal{S}_\mathbb{S}} = \mathsf{L}^2(\Omega;\mathbb{S})\times\boldsymbol{L}^2(\Omega)\,,\\ &b^{\mathcal{S}_\mathbb{S}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega - (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\tau})_\Omega - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_\Omega\,,\end{aligned} \tag{3.15}$$

$$\begin{aligned}&\mathcal{U}^{\mathcal{U}_\mathbb{S}} = \mathsf{L}^2(\Omega;\mathbb{S})\times\boldsymbol{L}^2(\Omega)\,, \qquad \mathcal{V}^{\mathcal{U}_\mathbb{S}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})\times\boldsymbol{H}^1_{\Gamma_u}(\Omega)\,,\\ &b^{\mathcal{U}_\mathbb{S}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma},\boldsymbol{\nabla}\boldsymbol{v})_\Omega\,,\end{aligned} \tag{3.16}$$

$$\begin{aligned}&\mathcal{U}^{\mathcal{M}_\mathbb{S}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})\times\boldsymbol{L}^2(\Omega)\,, \qquad \mathcal{V}^{\mathcal{M}_\mathbb{S}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})\times\boldsymbol{L}^2(\Omega)\,,\\ &b^{\mathcal{M}_\mathbb{S}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_\Omega\,,\end{aligned} \tag{3.17}$$

$$\begin{aligned}&\mathcal{U}^{\mathcal{S}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)\times\boldsymbol{H}^1_{\Gamma_u}(\Omega)\,, \qquad \mathcal{V}^{\mathcal{S}} = \mathsf{L}^2(\Omega;\mathbb{S})\times\boldsymbol{L}^2(\Omega)\times\mathsf{L}^2(\Omega;\mathbb{A})\,,\\ &b^{\mathcal{S}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{w})\big) = (\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega - (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\tau})_\Omega - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_\Omega + (\boldsymbol{\sigma},\boldsymbol{w})_\Omega\,,\end{aligned} \tag{3.18}$$

$$\begin{aligned}
&\mathcal{U}^{\mathcal{U}} = \mathsf{L}^2(\Omega;\mathbb{S}) \times \boldsymbol{L}^2(\Omega) \times \mathsf{L}^2(\Omega;\mathbb{A})\,, \qquad \mathcal{V}^{\mathcal{U}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega)\,,\\
&b^{\mathcal{U}}\big((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\omega}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega},\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma},\boldsymbol{\nabla}\boldsymbol{v})_\Omega\,,
\end{aligned}\tag{3.19}$$

$$\begin{aligned}
&\mathcal{U}^{\mathcal{M}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega) \times \boldsymbol{L}^2(\Omega) \times \mathsf{L}^2(\Omega;\mathbb{A})\,, \qquad \mathcal{V}^{\mathcal{M}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega) \times \boldsymbol{L}^2(\Omega) \times \mathsf{L}^2(\Omega;\mathbb{A})\,,\\
&b^{\mathcal{M}}\big((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\omega}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{w})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega},\boldsymbol{\tau})_\Omega + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_\Omega + (\boldsymbol{\sigma},\boldsymbol{w})_\Omega\,,
\end{aligned}\tag{3.20}$$

$$\begin{aligned}
&\mathcal{U}^{\mathcal{D}} = \mathsf{L}^2(\Omega;\mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega)\,, \qquad \mathcal{V}^{\mathcal{D}} = \mathsf{L}^2(\Omega;\mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega)\,,\\
&b^{\mathcal{D}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega - (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma},\boldsymbol{\nabla}\boldsymbol{v})_\Omega\,,
\end{aligned}\tag{3.21}$$

$$\begin{aligned}
&\mathcal{U}^{\mathcal{P}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega)\,, \qquad \mathcal{V}^{\mathcal{P}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega)\,,\\
&b^{\mathcal{P}}\big(\boldsymbol{u},\boldsymbol{v}\big) = (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{v})_\Omega\,.
\end{aligned}\tag{3.22}$$

The eighth formulation at the end was derived directly from the second-order equation in (3.4). Here, $\mathcal{S}$ stands for strong, $\mathcal{U}$ stands for ultraweak, $\mathcal{M}$ stands for mixed, $\mathcal{D}$ stands for dual-mixed, and $\mathcal{P}$ stands for primal, with the subscript $\mathbb{S}$ denoting the cases where the "problematic" $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})$ is involved.

With homogeneous boundary conditions, $\boldsymbol{u}^{\Gamma_u} = 0$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma} = 0$, the linear forms, $\ell^{\mathcal{F}}$, always take the form $\ell^{\mathcal{F}}(\mathfrak{v}) = (\boldsymbol{f},\boldsymbol{v})_\Omega$, where $\boldsymbol{v}$ is a component of $\mathfrak{v} \in \mathcal{V}^{\mathcal{F}}$ with $\mathcal{F}$ being one of the formulations defined above. With nonhomogeneous boundary conditions, $\ell^{\mathcal{F}}$ will have terms involving extensions of the boundary conditions, $\boldsymbol{u}^{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$, to $\partial\Omega$ and Ω. For example, $\ell^{\mathcal{U}}\big((\boldsymbol{\tau},\boldsymbol{v})\big) = (\boldsymbol{f},\boldsymbol{v})_\Omega + \langle\check{\boldsymbol{u}}^{\Gamma_u},\mathbf{tr}^{\Omega}_{\mathrm{div}}\boldsymbol{\tau}\rangle_{\partial\Omega} + \langle\check{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma},\mathbf{tr}^{\Omega}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\Omega}$, where $\check{\boldsymbol{u}}^{\Gamma_u}$ and $\check{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma}$ are some extension to $\partial\Omega$ of $\boldsymbol{u}^{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$ respectively. As these expressions suggest, it is assumed that $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$, $\boldsymbol{u}^{\Gamma_u} \in \mathbf{tr}^{\Omega}_{\mathrm{grad}}\big(\boldsymbol{H}^1(\Omega)\big)\big|_{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma} \in \mathbf{tr}^{\Omega}_{\mathrm{div}}\big(\mathsf{H}(\mathbf{div},\Omega)\big)\big|_{\Gamma_\sigma}$. Moreover, as mentioned in the last section, whenever necessary, C and S are assumed to act on $\mathbb{M}$ (as opposed to merely $\mathbb{S}$) via the trivial extensions $\mathsf{C}|_{\mathbb{A}} = 0$ and $\mathsf{S}|_{\mathbb{A}} = 0$.

The variational formulations with symmetric functional settings, namely, those with the same trial and test spaces, are (3.17), (3.20), (3.21) and (3.22). These can be discretized using the Bubnov-Galerkin method. In principle, it would unclear how to discretize the other formulations using conventional methods while still retaining numerical stability. The minimum residual methodology described here is possibly the most attractive way to do this in a systematic fashion. On another note, those formulations involving S are expected to avoid volumetric locking by being robustly well-posed in the incompressible limit, and in fact, this has been proved in some of those cases ((3.17) and (3.20)) [7, 10, 110].

It should be noted that the list of variational formulations for the equations of linear elasticity proposed here is by no means exhaustive. Indeed, alternative versions of (3.18) and (3.21) containing the compliance tensor (via use of (3.6)) are possible to construct, while in [190] energy functionals are used to propose up to fourteen different variational formulations for these equations. Moreover, volumetric locking can also be avoided by introducing a pressure term that produces yet another formulation, but it comes at the cost making traction (normal stress) boundary conditions more difficult to handle [154].

Finally, all the formulations (3.15)–(3.22) are mutually well-posed. This is the content of the next theorem, whose proof is relegated to Appendix B (see Section B.1).

Theorem 3.1. *If $\Gamma_u \neq \varnothing$, all the previously defined variational formulations are simultaneously well-posed in the sense of Hadamard. That is, for the problem* (3.14) *with the forms and spaces coming from one of* (3.15)–(3.22), *there exists a unique solution $\mathfrak{u}^{\mathcal{F}} \in \mathcal{U}^{\mathcal{F}}$ satisfying the stability estimate $\|\mathfrak{u}^{\mathcal{F}}\|_{\mathcal{U}^{\mathcal{F}}} \leq \frac{1}{\gamma^{\mathcal{F}}}\|\ell^{\mathcal{F}}\|_{(\mathcal{V}^{\mathcal{F}})'}$ for some $\gamma^{\mathcal{F}} > 0$. Since all variational formulations originate in the same equations, by testing with smooth functions it is made clear that the unique solutions agree among all formulations.*

3.2.5 A smaller family of broken variational formulations

For some numerical methods, mesh-dependent broken spaces can bring advantages. In particular, consider the case where only the test spaces are broken. It is in this setting that broken variational formulations arise and, as it will be seen, this is fundamental in order to localize certain computations in the DPG methodology.

Consider a mesh, $\mathcal{T}$, of Ω, containing elements $K \in \mathcal{T}$. Instead of following the original approach of formally multiplying by a test function as in Section 3.2.3, the idea in this case is to integrate over each $K \in \mathcal{T}$ and then sum the contributions, as in Section 2.2. This differs from the former scenario in that the test functions can now be broken, so that they may have trace discontinuities along the boundaries of adjacent elements in the mesh. Thus, when integration by parts is performed, some mesh boundary terms seize to cancel and have to be explicitly considered. Apart from these terms, the resulting formulations are the same as before, where unbroken test functions were being used. However, if they are to retain as much mathematical structure from the original unbroken variational formulations, one finds that the new mesh boundary terms must have a life of their own and become additional independent variables. That is, the price of using broken test functions is that one sometimes needs to define new mesh-dependent *interface* variables along the boundary of the mesh, as described in Section 2.2.

Broken variational formulations are precisely those formulations with broken test spaces constructed as described above. They are clearly related to the original unbroken variational formulations, which do not require the test spaces to be broken. In fact, the bilinear forms of broken variational formulations can be decoupled into two bilinear forms as,

$$b(\mathfrak{u},\mathfrak{v}) = b_0(\mathfrak{u}_0,\mathfrak{v}) + \hat{b}(\hat{\mathfrak{u}},\mathfrak{v})\,, \tag{3.23}$$

where $\mathfrak{u} = (\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathcal{U} = \mathcal{U}_0 \times \hat{\mathcal{U}}$ and $\mathfrak{v} \in \mathcal{V}$, with $\mathcal{U}_0$ being the space associated to the original unbroken formulation, $\hat{\mathcal{U}}$ being a space of interface variables, and $\mathcal{V}$ being the broken test space directly associated to the test space $\mathcal{V}_0 \subseteq \mathcal{V}$ coming from the original unbroken formulation. When the test space is restricted from $\mathcal{V}$ to $\mathcal{V}_0$ the variational formulation collapses to the original unbroken formulation. More precisely, $b_0|_{\mathcal{U}_0\times\mathcal{V}_0}$ is the bilinear form from the unbroken formulation and $\hat{b}|_{\hat{\mathcal{U}}\times\mathcal{V}_0} = 0$, while $\ell|_{\mathcal{V}_0}$ is the linear form from the unbroken formulation. In this sense, a broken variational formulation can be interpreted as an extension to broken test spaces of an unbroken variational formulation. It can be shown that the well-posedness of broken variational formulations depends on the well-posedness of the original unbroken variational formulation and that of $\hat{b}$. Moreover, the unique solution $\mathfrak{u}_0 \in \mathcal{U}_0$ to the unbroken formulation is the $\mathcal{U}_0$ component of the unique solution $(\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathcal{U}_0 \times \hat{\mathcal{U}}$ to the broken variational formulation. This is the content of Theorem 2.1.

We choose only a subset of the formulations in the last section. The reason is that we will perform computations with these formulations. In this sense, we discard (3.15)–(3.17), because they involve the space $\mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})$, which is difficult to discretize. We also discarded (3.21), because it is essentially the same as (3.22), but much more expensive computationally.

The chosen subset of broken variational formulations associated to (3.18), (3.19), (3.20) and (3.22) are deduced to be

$$\begin{aligned}&\mathcal{U}_0^{\mathcal{S}\mathcal{T}} = \mathcal{U}^{\mathcal{S}}\,, \quad \hat{\mathcal{U}}^{\mathcal{S}\mathcal{T}} = \varnothing\,, \qquad \mathcal{V}^{\mathcal{S}\mathcal{T}} = \mathbf{L}^2(\mathcal{T};\mathbb{S}) \times \boldsymbol{L}^2(\mathcal{T}) \times \mathbf{L}^2(\mathcal{T};\mathbb{A})\,,\\ &b_0^{\mathcal{S}\mathcal{T}}\big((\boldsymbol{\sigma},\boldsymbol{u}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{w})\big) = (\boldsymbol{\sigma},\boldsymbol{\tau})_{\mathcal{T}} - (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\tau})_{\mathcal{T}} - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_{\mathcal{T}} + (\boldsymbol{\sigma},\boldsymbol{w})_{\mathcal{T}}\,,\end{aligned}\tag{3.24}$$

$$\begin{aligned}&\mathcal{U}_0^{\mathcal{U}\mathcal{T}} = \mathcal{U}^{\mathcal{U}}\,, \quad \hat{\mathcal{U}}^{\mathcal{U}\mathcal{T}} = \boldsymbol{H}^{1/2}_{\Gamma_u}(\partial\mathcal{T}) \times \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})\,, \qquad \mathcal{V}^{\mathcal{U}\mathcal{T}} = \mathbf{H}(\mathbf{div},\mathcal{T}) \times \boldsymbol{H}^1(\mathcal{T})\,,\\ &b_0^{\mathcal{U}\mathcal{T}}\big((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\omega}),(\boldsymbol{\tau},\boldsymbol{v})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_{\mathcal{T}} + (\boldsymbol{\omega},\boldsymbol{\tau})_{\mathcal{T}} + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_{\mathcal{T}} + (\boldsymbol{\sigma},\boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}}\,,\\ &\hat{b}^{\mathcal{U}\mathcal{T}}\big((\hat{\boldsymbol{u}},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}),(\boldsymbol{\tau},\boldsymbol{v})\big) = -\langle\hat{\boldsymbol{u}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}\boldsymbol{\tau}\rangle_{\partial\mathcal{T}} - \langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\mathcal{T}}\,,\end{aligned}\tag{3.25}$$

$$\begin{aligned}&\mathcal{U}_0^{\mathcal{M}\mathcal{T}} = \mathcal{U}^{\mathcal{M}}\,, \quad \hat{\mathcal{U}}^{\mathcal{M}\mathcal{T}} = \boldsymbol{H}^{1/2}_{\Gamma_u}(\partial\mathcal{T})\,, \qquad \mathcal{V}^{\mathcal{M}\mathcal{T}} = \mathbf{H}(\mathbf{div},\mathcal{T}) \times \boldsymbol{L}^2(\mathcal{T}) \times \mathbf{L}^2(\mathcal{T};\mathbb{A})\,,\\ &b_0^{\mathcal{M}\mathcal{T}}\big((\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\omega}),(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{w})\big) = (\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\tau})_{\mathcal{T}} + (\boldsymbol{\omega},\boldsymbol{\tau})_{\mathcal{T}} + (\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_{\mathcal{T}} - (\mathbf{div}\,\boldsymbol{\sigma},\boldsymbol{v})_{\mathcal{T}} + (\boldsymbol{\sigma},\boldsymbol{w})_{\mathcal{T}}\,,\\ &\hat{b}^{\mathcal{M}\mathcal{T}}\big(\hat{\boldsymbol{u}},(\boldsymbol{\tau},\boldsymbol{v},\boldsymbol{w})\big) = -\langle\hat{\boldsymbol{u}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}\boldsymbol{\tau}\rangle_{\partial\mathcal{T}}\,,\end{aligned}\tag{3.26}$$

$$\begin{aligned}&\mathcal{U}_0^{\mathcal{P}\mathcal{T}} = \mathcal{U}^{\mathcal{P}}\,, \quad \hat{\mathcal{U}}^{\mathcal{P}\mathcal{T}} = \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})\,, \qquad \mathcal{V}^{\mathcal{P}\mathcal{T}} = \boldsymbol{H}^1(\mathcal{T})\,,\\ &b_0^{\mathcal{P}\mathcal{T}}\big(\boldsymbol{u},\boldsymbol{v}\big) = (\mathsf{C}:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}}\,,\\ &\hat{b}^{\mathcal{P}\mathcal{T}}\big(\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\boldsymbol{v}\big) = -\langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\mathcal{T}}\,,\end{aligned}\tag{3.27}$$

where $\mathcal{U}^{\mathcal{F}\mathcal{T}} = \mathcal{U}_0^{\mathcal{F}\mathcal{T}} \times \hat{\mathcal{U}}^{\mathcal{F}\mathcal{T}}$ with $\mathcal{F}$ being a placeholder for one of the preceding formulations, and where $b^{\mathcal{F}\mathcal{T}}:\mathcal{U}^{\mathcal{F}\mathcal{T}}\times\mathcal{V}^{\mathcal{F}\mathcal{T}}\to\mathbb{R}$ is defined in terms of $b_0^{\mathcal{F}\mathcal{T}}$ and $\hat{b}^{\mathcal{F}\mathcal{T}}$ by (3.23). As before, the linear forms $\ell^{\mathcal{F}\mathcal{T}}$ always have the term $(\boldsymbol{f},\boldsymbol{v})_{\mathcal{T}}$ and additionally may include terms involving extensions of the boundary conditions $\boldsymbol{u}^{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$, to Ω and the boundary of the mesh (by use of $\mathbf{tr}_{\mathrm{grad}}$ and $\mathbf{tr}_{\mathrm{div}}$ on an extension to Ω). For example, $\ell^{\mathcal{U}\mathcal{T}}\big((\boldsymbol{\tau},\boldsymbol{v})\big) = (\boldsymbol{f},\boldsymbol{v})_{\mathcal{T}} + \langle\breve{\boldsymbol{u}}^{\Gamma_u},\mathbf{tr}_{\mathrm{div}}\boldsymbol{\tau}\rangle_{\partial\mathcal{T}} + \langle\breve{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma},\mathbf{tr}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\mathcal{T}}$, where $\breve{\boldsymbol{u}}^{\Gamma_u}$ and $\breve{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma}$ are some extension to $\boldsymbol{H}^{1/2}(\partial\mathcal{T})$ and $\boldsymbol{H}^{-1/2}(\partial\mathcal{T})$ of $\boldsymbol{u}^{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$ respectively. As expected, $b^{\mathcal{F}\mathcal{T}}$ and $\ell^{\mathcal{F}\mathcal{T}}$ can be viewed as extensions to the original forms $b^{\mathcal{F}}$ and $\ell^{\mathcal{F}}$, because they collapse to the latter when testing against unbroken test functions in $\mathcal{V}^{\mathcal{F}}\subseteq\mathcal{V}^{\mathcal{F}\mathcal{T}}$. That is, $b_0^{\mathcal{F}\mathcal{T}}|_{\mathcal{U}^{\mathcal{F}}\times\mathcal{V}^{\mathcal{F}}} = b^{\mathcal{F}}$, $\hat{b}^{\mathcal{F}\mathcal{T}}|_{\hat{\mathcal{U}}^{\mathcal{F}\mathcal{T}}\times\mathcal{V}^{\mathcal{F}}} = 0$ and $\ell^{\mathcal{F}\mathcal{T}}|_{\mathcal{V}^{\mathcal{F}}} = \ell^{\mathcal{F}}$.

To establish the well-posedness of the broken variational formulations, (3.24)–(3.27), use Theorem 2.1. As a simple example consider (3.27). First, observe that by Theorem A.1,

$$\mathcal{V}_0 = \big\{\boldsymbol{v}\in\boldsymbol{H}^1(\mathcal{T}) \mid \langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\mathcal{T}} = 0 \;\; \forall\hat{\boldsymbol{\sigma}}_{\mathbf{n}}\in\boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})\big\} = \boldsymbol{H}^1_{\Gamma_u}(\Omega) = \mathcal{V}^{\mathcal{P}}\,,\tag{3.28}$$

so condition (γ_0) in Theorem 2.1 is immediately satisfied in view of Theorem 3.1 provided $\Gamma_u \neq \varnothing$. Moreover, condition $(\hat{\gamma})$ is also satisfied with $\hat{\gamma} = 1$ by use of Theorem A.3. This means the broken variational formulation has a positive inf-sup constant, and actually this constant is independent of the mesh (the continuity bound of $b_0^{\mathcal{P}\mathcal{T}}$ is easily seen to be bounded by $M_0 = 1$ regardless of the mesh). Furthermore, if $\Gamma_u \neq \varnothing$, then

$$\mathcal{V}_{00} = \{\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega) \mid (\mathsf{C} : \boldsymbol{\nabla u}, \boldsymbol{\nabla v})_{\mathcal{T}} = 0 \ \ \forall \boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)\} = \{\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega) \mid \boldsymbol{\nabla v} = 0\} = \{0\}\,. \tag{3.29}$$

Hence, by Theorem 2.1, it follows (3.27) leads to a well-posed variational formulation. In general, use the same procedure for the remaining formulations and show that $\ell^{\mathcal{F}}|_{\mathcal{V}_{00}} = 0$ in all cases. This results in the following theorem.

Theorem 3.2. *If $\Gamma_u \neq \varnothing$, then the broken variational formulations associated to* (3.24)–(3.27) *are well-posed with associated stability constants that are independent of the mesh being considered.*

3.2.6 Coupled variational formulations

As mentioned initially, there are multiple reasons that explain why it is desirable to solve the equations of linear elasticity with different variational formulations on distinct subdomains of the initial domain. The challenge in attaining this goal is that one must find a way of communicating solution information across the shared boundaries of the subdomains. For the purpose of illustration, simply consider a domain Ω partitioned into two disjoint subdomains, $\Omega^{\mathcal{U}}$ and $\Omega^{\mathcal{P}}$, with a common interface, Γ_I, such that $\overline{\Omega}^{\mathcal{U}} \cup \overline{\Omega}^{\mathcal{P}} = \overline{\Omega}$ and $\overline{\Omega}^{\mathcal{U}} \cap \overline{\Omega}^{\mathcal{P}} = \overline{\Gamma}_I$. As suggested by the notation, suppose that the equations of linear elasticity are to be solved in $\Omega^{\mathcal{U}}$ via the ultraweak variational formulation in (3.19), and in $\Omega^{\mathcal{P}}$ via the primal variational formulation in (3.22). If a solution is to exist, then it should be compatible in some sense at the common interface Γ_I. This immediately poses a theoretical concern because the displacement variable in the ultraweak variational formulation lies in $\boldsymbol{L}^2(\Omega)|_{\Omega^{\mathcal{U}}}$ and so it does not even have a notion of trace at Γ_I. Thus, it is not compatible with the primal displacement variable which lies in $\boldsymbol{H}^1_{\Gamma_u}(\Omega)|_{\Omega^{\mathcal{P}}}$. A similar issue also arises with the test spaces, which are obviously different on each subdomain. Even though the finite-dimensional trial and test subspaces of any naive discretization generally do have well-defined traces, these difficulties are reasonably expected to be inherited by the discretization, meaning any discrete convergence or stability analysis will probably be laborious, if at all possible. Hence, the goal is to resolve the compatibility concerns at the infinite-dimensional level by developing a globally well-posed variational problem. Once this is done, there will be a clearer hope of producing stable and convergent discretizations.

The claim is that by using broken variational formulations, the theoretical compatibility issues are naturally dealt with. To see this, suppose instead that the equations are to be solved in $\Omega^{\mathcal{U}}$ with the *broken* ultraweak variational formulation in (3.25) and in $\Omega^{\mathcal{P}}$ with the *broken* primal variational formulation in (3.27). The mesh associated to the broken formulations, $\mathcal{T}$, is obviously assumed to be consistent with the subdomain partitioning, meaning that it is a refinement of the subdomain mesh, $\mathcal{T}_0 = \{\Omega^{\mathcal{U}}, \Omega^{\mathcal{P}}\}$, and as such, there exist submeshes $\mathcal{T}^{\mathcal{U}}$ and $\mathcal{T}^{\mathcal{P}}$ of $\Omega^{\mathcal{U}}$ and $\Omega^{\mathcal{P}}$ respectively, such that $\mathcal{T} = \mathcal{T}^{\mathcal{U}} \cup \mathcal{T}^{\mathcal{P}}$. In this scenario, the displacement variable in the ultraweak domain still lies in $\boldsymbol{L}^2(\Omega)|_{\Omega^{\mathcal{U}}}$, but the difference is that now there is an extra *interface* displacement variable, $\hat{\boldsymbol{u}}^{\mathcal{U}} \in \boldsymbol{H}^{1/2}_{\Gamma_u}(\partial\mathcal{T})|_{\mathcal{T}^{\mathcal{U}}}$. This variable is very convenient, as it is theoretically compatible at Γ_I with the well-defined trace of the displacement variable of the primal variational formulation, $\boldsymbol{u}^{\mathcal{P}} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)|_{\Omega^{\mathcal{P}}}$. Similarly, with regard to the stress, there exist two new interface traction variables, $\hat{\boldsymbol{\sigma}}^{\mathcal{U}}_{\mathbf{n}} \in \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})|_{\mathcal{T}^{\mathcal{U}}}$ and $\hat{\boldsymbol{\sigma}}^{\mathcal{P}}_{\mathbf{n}} \in \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})|_{\mathcal{T}^{\mathcal{P}}}$, which are naturally compatible at Γ_I. Meanwhile, the use of broken test spaces relinquishes any compatibility requirements at the level of test spaces. Notice the compatibility is not limited to the broken ultraweak and primal formulations. Indeed, a close observation of the broken variational formulations in (3.24)–(3.27) shows that there is always either an explicit interface variable or sufficient regularity to have well-defined traces of the displacement and stress.

The next task is to more rigorously define the actual *coupled* variational formulations and analyze their well-posedness. Continuing with the basic example, let $\mathcal{U}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$ be the restriction of the trial space to $\Omega^{\mathcal{P}}$ meaning that typical field variables in $\mathcal{U}^{\mathcal{P}\mathcal{T}}_0 = \mathcal{U}^{\mathcal{P}}$ have their domain restricted to $\Omega^{\mathcal{P}}$, while the interface variables in $\hat{\mathcal{U}}^{\mathcal{P}\mathcal{T}}$ are restricted to those components associated to elements in $\mathcal{T}^{\mathcal{P}}$. Therefore, the space is $\mathcal{U}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega)|_{\Omega^{\mathcal{P}}} \times \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})|_{\mathcal{T}^{\mathcal{P}}}$, with the restricted component norms being $\|\cdot\|_{\boldsymbol{H}^1(\Omega^{\mathcal{P}})}$ and $\|\cdot\|_{\boldsymbol{H}^{-1/2}(\partial\mathcal{T}^{\mathcal{P}})}$ respectively. The same applies to $\mathcal{U}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}$ and the broken test spaces $\mathcal{V}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}$ and $\mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$. Then, the trial and test spaces associated to the coupled formulations are

$$\begin{aligned} \mathcal{U}^{\mathcal{C}} = \Big\{ \mathfrak{u}^{\mathcal{C}} = (\mathfrak{u}^{\mathcal{U}}, \mathfrak{u}^{\mathcal{P}}) \,\Big|\, \mathfrak{u}^{\mathcal{U}} = (\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}}, \hat{\boldsymbol{u}}^{\mathcal{U}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}) \in \mathcal{U}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}},\, \mathfrak{u}^{\mathcal{P}} = (\boldsymbol{u}^{\mathcal{P}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}) \in \mathcal{U}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}, \\ \hat{\boldsymbol{u}}^{\mathcal{U}}|_{\Gamma_I} = \boldsymbol{u}^{\mathcal{P}}|_{\Gamma_I},\, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}|_{\Gamma_I} = -\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}|_{\Gamma_I} \Big\}, \\ \mathcal{V}^{\mathcal{C}} = \mathcal{V}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}} \times \mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}} \,. \end{aligned} \tag{3.30}$$

Hence, the trial space is the subspace of $\mathcal{U}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}} \times \mathcal{U}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$ which satisfies transmission conditions for both displacement and stress at Γ_I (see Remark 3.1 for more details). On the other hand, the *broken* test space is oblivious to any transmission conditions. Lastly, the bilinear and linear forms of the coupled variational formulation are

$$\begin{aligned} b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}}, \mathfrak{v}^{\mathcal{C}}) &= b^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}(\mathfrak{u}^{\mathcal{U}}, \mathfrak{v}^{\mathcal{U}}) + b^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}^{\mathcal{P}}, \mathfrak{v}^{\mathcal{P}}) \,, \\ \ell^{\mathcal{C}}(\mathfrak{v}^{\mathcal{C}}) &= \ell^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}(\mathfrak{v}^{\mathcal{U}}) + \ell^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{v}^{\mathcal{P}}) \,, \end{aligned} \tag{3.31}$$

where the restricted forms $b^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}$ and $b^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$ are those formulations in (3.25) and (3.27) but with the inner products and duality pairings only acting over those elements in $\mathcal{T}^{\mathcal{U}}$ and $\mathcal{T}^{\mathcal{P}}$ respectively. The same applies to the linear forms $\ell^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}$ and $\ell^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$. Evidently, by carefully identifying the trial spaces to enforce the compatibility conditions at the interdomain boundaries, coupled variational formulations can be rigorously generalized to any finite partition of the domain into subdomains, wherein each subdomain is endowed with a broken variational formulation among those found in (3.24)–(3.27).

Remark 3.1. There is an abuse of notation when specifying the transmission conditions that enforce the compatibility at the interdomain boundaries in (3.30). More precisely, $\hat{\boldsymbol{u}}^{\mathcal{U}}|_{\Gamma_I} = \boldsymbol{u}^{\mathcal{P}}|_{\Gamma_I}$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}|_{\Gamma_I} = -\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}|_{\Gamma_I}$ denote that there exist global extensions $\widetilde{\boldsymbol{u}} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\widetilde{\boldsymbol{\sigma}} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$ such that $\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\widetilde{\boldsymbol{u}}|_{\mathcal{T}^{\mathcal{U}}} = \hat{\boldsymbol{u}}^{\mathcal{U}}$, $\widetilde{\boldsymbol{u}}|_{\Omega^{\mathcal{P}}} = \boldsymbol{u}^{\mathcal{P}}$, $\mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}\widetilde{\boldsymbol{\sigma}}|_{\mathcal{T}^{\mathcal{U}}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}$ and $\mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}\widetilde{\boldsymbol{\sigma}}|_{\mathcal{T}^{\mathcal{P}}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}$. In fact, these global extensions for the displacement and stress, $\widetilde{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{\sigma}}$, are fundamental in the numerical implementation, where they are considered global variables in the context of a multi-physics domain, whereas the remaining variables only have local support in a particular subdomain. Moreover, the concept of the extensions is also important for specifying the problem boundary conditions, $\boldsymbol{u}^{\Gamma_u}$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$. Indeed, by definition there exist extensions, $\widetilde{\boldsymbol{u}}^{\Gamma_u} \in \boldsymbol{H}^1(\Omega)$ and $\widetilde{\boldsymbol{\sigma}}^{\Gamma_\sigma} \in \mathsf{H}(\mathrm{div}, \Omega)$, whose appropriate restrictions (e.g. $\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\widetilde{\boldsymbol{u}}^{\Gamma_u}|_{\mathcal{T}^{\mathcal{U}}}$ and $\widetilde{\boldsymbol{u}}^{\Gamma_u}|_{\Omega^{\mathcal{P}}}$ for the displacement) play a role in the linear forms $\ell^{\mathcal{U}_{\mathcal{T}}}|_{\Omega^{\mathcal{U}}}$ and $\ell^{\mathcal{P}_{\mathcal{T}}}|_{\Omega^{\mathcal{P}}}$.

It remains to show that the coupled variational formulations are well-posed. The technique is similar in spirit to the one utilized in proving well-posedness of broken variational formulations in Theorem 2.1 (see [55]), where the first step is always to test with unbroken test functions to cancel the boundary terms. The main idea here, however, is to collapse everything to the well-posed ultraweak formulation by testing with more regular test functions and integrating by parts when necessary. This is interesting since the ultraweak formulation is effectively being used as a tool for a proof, due to its attractive property of having all the weight of the derivatives on the test functions. The proof is again left for Appendix B, to keep continuity of the document.

Theorem 3.3. *Let Ω be a domain partitioned into a finite number of subdomains, wherein each subdomain is endowed with a broken variational formulation associated to one of the bilinear forms in* (3.24)–(3.27). *Provided $\Gamma_u \neq \varnothing$, the resulting coupled variational formulation is well-posed with an inf-sup stability constant, $\gamma^{\mathcal{C}} > 0$, independent of the mesh.*

Remark 3.2. The stability constant of the couple formulation, $\gamma^{\mathcal{C}}$, depends on the distribution of the variational formulations and the shape of the subdomains. The constant will remain robust with respect to heterogeneous material properties as long as each formulation is robust when viewed independently. Thus, the stability constant will remain bounded above as long as any near or fully incompressible elastic behavior is limited to subdomains associated to robustly well-posed variational formulations (i.e. broken ultraweak and mixed formulations).

Remark 3.3. As mentioned before, in this work, the variational formulations of linear elasticity avoid the strong imposition of tensor symmetry in some of the spaces (i.e. they avoid $\mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})$). This, among other reasons, adds a layer of complexity to the formulations and the corresponding proofs, which typically need a few extra calculations. However, these difficulties are not present in many other important equations. Indeed, a simpler version of this proof can easily be applied to coupled formulations of Poisson's equation, time-harmonic Maxwell's equations (see [55] for multiple formulations) and the diffusion-convection-reaction equation, among others.

3.3 Numerical results

The four broken variational formulations described in Section 3.2.5, (3.24)–(3.27), were discretized and solved numerically using the DPG methodology as described in Chapter 2. Often the corresponding superscript is added to equation references to facilitate association, $(3.24)^{\mathcal{S}}$, $(3.25)^{\mathcal{U}}$, $(3.26)^{\mathcal{M}}$ and $(3.27)^{\mathcal{P}}$. In particular, $(3.24)^{\mathcal{S}}$ was implemented as a first-order system least-squares (FOSLS) method as described in Section 2.5, while $(3.26)^{\mathcal{M}}$ was implemented as best as possible by inverting part of the Riesz map exactly as described also in Section 2.5. The other two formulations, $(3.25)^{\mathcal{U}}$ and $(3.27)^{\mathcal{P}}$, were implemented as described throughout the rest of Chapter 2 (see Section 2.3, Section 2.9 and Section 2.7). This was possible because all the spaces in question were SdR spaces (see Section A.5 in Appendix A for definition). Hence, the trial spaces were chosen as compatible SdR discretizations of high-order p, while the test spaces (when required) were chosen as SdR discretizations of order $p + \Delta p$.

It should be noted that it would not be difficult to construct Fortin operators for each of these formulations using Theorem 2.3 and the procedure outlined in Section 2.8.1. Each of these Fortin operators would have a minimum value of Δp which would guarantee high-order h-convergence as stated by Theorem 2.4, and a preliminary analysis shows that $\Delta p \geq 3$ should be sufficient for all formulations. However, this is only a conservative criterion, and lower values of Δp can be used to produce computations, which has the advantage that the numerical method is more efficient. In the computations that follow, we always chose $\Delta p = 1$ and all meshes were hexahedral. Regardless, as stipulated by Theorem 2.4, the expected convergence rate under uniform h-refinements was always inherited from the mesh size, h, the order of the trial space, $p \in \mathbb{N}$, and the fractional order of the exact solution, $s \geq 0$, so that the error would be bounded by $Ch^{\min\{s,p\}}$, where C only depends on the exact solution, s and p.

The software `hp3d` was used for all computations (see Section 2.9.1), as it has compatible SdR discretizations for all the conventional element shapes [114]. The solver used in all computations was `MUMPS 5.0.1`.

The four resulting numerical methods discretizing (3.24)–(3.27) were applied to a manufactured solution and the rates for both the relative displacement error and the global residual were compared. They were also compared when solving for a singular solution in an L-shaped domain, both under uniform refinements and adaptive refinements. These results are in Section 3.3.1 and Section 3.3.2 respectively.

The coupled formulations were also implemented as another DPG method. Here, the sophisticated multi-physics support in `hp3d` was instrumental in facilitating the global assembly necessary to make the trial space identification for coupled formulations as suggested by Remark 3.1. The interdomain continuity was enforced as discussed in Remark 3.1.

Two illustrative examples involving these coupled formulations were solved. First, a smooth manufactured solution on a cube with uniform and contrived material data involving four distinct variational formulations in the same domain was considered. Then, a more physically-motivated and challenging example was tackled: a sheathed hose with large material and layer-thickness contrast, and with one layer composed of a fully incompressible material. The parametric transfinite interpolation supported by `hp3d` made these computations possible. Only uniform refinements were considered in these two examples, which are shown in Section 3.3.1 and Section 3.3.3 respectively.

Remark 3.4. Note that $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S})$ is not an SdR space as defined in Section A.5 in Appendix A, because it is a closed subspace of $\mathsf{H}(\mathbf{div},\Omega)$, but not of $\boldsymbol{H}(\mathrm{div},\Omega)$.

Remark 3.5. In the context of the mixed formulations with weakly imposed symmetry, (3.20) and (3.26), some authors choose to nontrivially extend the compliance tensor, S, from $\mathbb{S}$ to $\mathbb{M}$ [10, 36]. They do this to ensure that $(\mathsf{S}:\boldsymbol{\sigma},\boldsymbol{\sigma})_\Omega$ remains positive definite on $\mathsf{L}^2(\Omega)$ (and not only on $\mathsf{L}^2(\Omega;\mathbb{S})$). However, in this work, we chose to extend the compliance tensor trivially, so that $\mathsf{S}|_{\mathbb{A}}=0$ (see Section 3.2.3). This did not pose any limitations in the infinite-dimensional setting while proving the well-posedness of the mixed variational formulations (see Theorem 3.1 and its proof in Appendix B, as well as Theorem 3.2). For the practical DPG methodology, where the test space is designed to approximate the optimal test space, one can show that for a large enough enrichment (i.e. value of Δp) the problem remains well-posed (construct a Fortin operator using Theorem 2.3). Thus, it is valid to extend S trivially, as this does not affect the presence of discrete stability.

3.3.1 Smooth solutions

In Section 3.3.1.1, the four broken variational formulations, $(3.24)^{\mathcal{S}}$, $(3.25)^{\mathcal{U}}$, $(3.26)^{\mathcal{M}}$ and $(3.27)^{\mathcal{P}}$, were solved, while in Section 3.3.1.2, a coupled variational formulation involving precisely those four formulations was considered. In both cases, a smooth manufactured solution was taken for the displacement, $\boldsymbol{u}$. It was a simple sinusoidal vector field,

$$u_i(x_1, x_2, x_3) = \sin(\pi x_1)\sin(\pi x_2)\sin(\pi x_3)\,, \qquad i = 1, 2, 3, \tag{3.32}$$

on a cubic domain. The domain was $\Omega = (0,1)^3$ in Section 3.3.1.1, and $\Omega = (0,2)^3$ in Section 3.3.1.2. The material was considered to be isotropic and homogeneous with nondimensionalized Lamé parameters $\lambda = \mu = 1$. The problem load, $\boldsymbol{f}$, was deduced from the displacement using the equations. The homogeneous displacement boundary data was prescribed along the whole boundary of Ω.

3.3.1.1 Homogeneous variational formulations

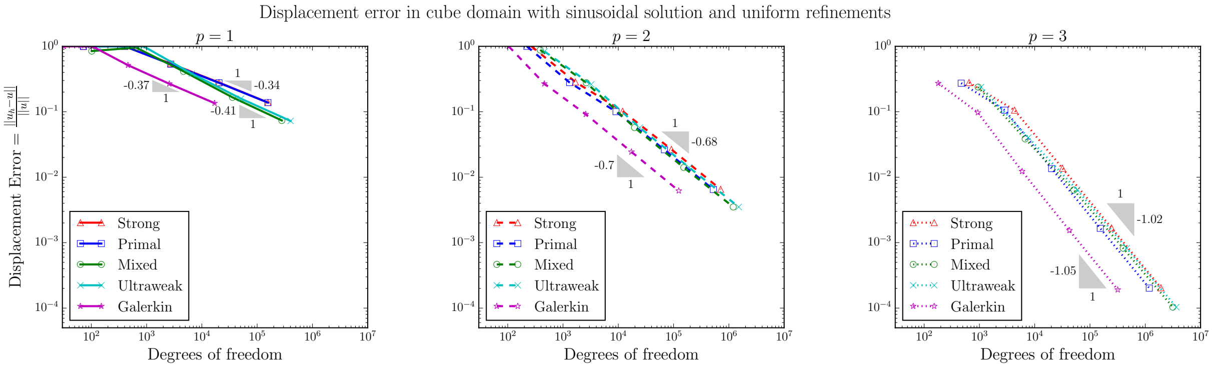


Figure 3.1: Relative displacement error as a function of the degrees of freedom after uniform tetrahedral refinements in the cube domain with a smooth solution.

The domain $\Omega = (0,1)^3$ was initially partitioned into five tetrahedra, and subsequently underwent a series of uniform tetrahedral refinements. The results for the relative error are presented in Figure 3.1 for $p = 1, 2, 3$ and they are shown for the DPG methods discretizing $(3.24)^{\mathcal{S}}$, $(3.25)^{\mathcal{U}}$, $(3.26)^{\mathcal{M}}$ and $(3.27)^{\mathcal{P}}$, alongside results from the classical Bubnov-Galerkin method discretizing (3.22). Here, the displacement error $\frac{\|\boldsymbol{u}-\boldsymbol{u}_h\|}{\|\boldsymbol{u}\|}$ was measured in the norm corresponding to each formulation ($\|\cdot\|_{\boldsymbol{H}^1(\Omega)}$ with $(3.24)^{\mathcal{S}}$ and $(3.27)^{\mathcal{P}}$ and $\|\cdot\|_{\boldsymbol{L}^2(\Omega)}$ with $(3.25)^{\mathcal{U}}$ and $(3.26)^{\mathcal{M}}$), where

$\boldsymbol{u}$ is the exact displacement and $\boldsymbol{u}_h$ is the computed displacement. The results are plotted as a function of the degrees of freedom, N_{dof}, so the expected convergence bound for 3D computations is of the form $Ch^{\min\{s,p\}} = \mathcal{O}\big(N_{\mathrm{dof}}^{-\min\{s/3,p/3\}}\big) = \mathcal{O}\big(N_{\mathrm{dof}}^{-p/3}\big)$, because $h^{-3} = \mathcal{O}(N_{\mathrm{dof}})$ and the solution is smooth (so $s \to \infty$). In this case, the convergence rates are precisely as expected for all methods. All DPG methods seem to behave very similarly, while the Galerkin method stands out for using less degrees of freedom (since it involves no interface variables).

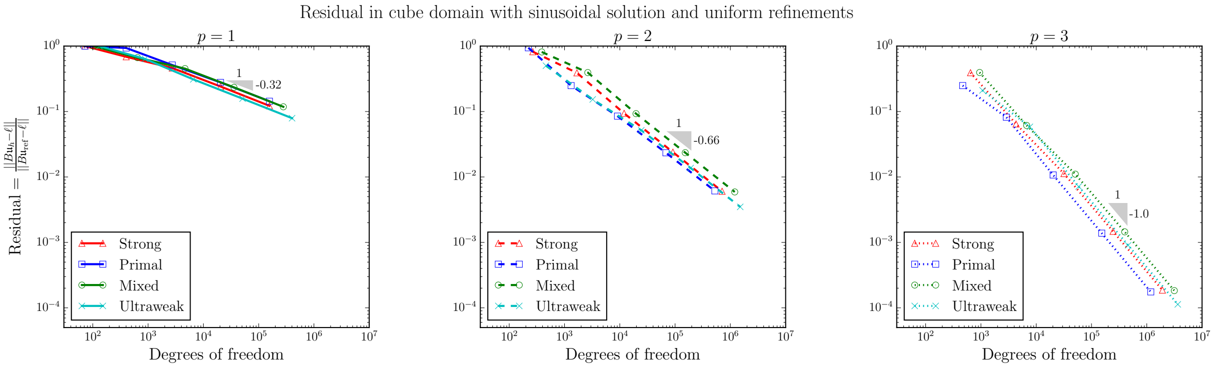


Figure 3.2: Residual as a function of the degrees of freedom after uniform tetrahedral refinements in the cube domain with a smooth solution.

The results in terms of the residual show a similar behavior and are illustrated in Figure 3.2. The global residual, η_h, was computed using (2.50) (see also Remark 2.3 for $(3.24)^{\mathcal{S}}$ and $(3.26)^{\mathcal{M}}$). To facilitate comparison, it was computed with the same level of enrichment ($p + \Delta p = 4$) as described in Remark 2.4, and normalized with a fixed reference residual, η_{ref}. Note there are no results for the residual of the classical Galerkin method because we did not implement a way to calculate it without using broken test spaces.

3.3.1.2 Coupled variational formulations

The domain $\Omega = (0,2)^3$ was partitioned into eight equally sized unit cube subdomains in a configuration in which four distinct broken formulations interact with each other. More specifically, the formulations, $(3.24)^{\mathcal{S}}$, $(3.25)^{\mathcal{U}}$, $(3.26)^{\mathcal{M}}$ and $(3.27)^{\mathcal{P}}$, were organized such that there is at least one face that is a common interface between each of the possible pairs of formulations, as shown in Figure 3.3. The initial mesh had eight hexahedral elements, and was then uniformly refined.

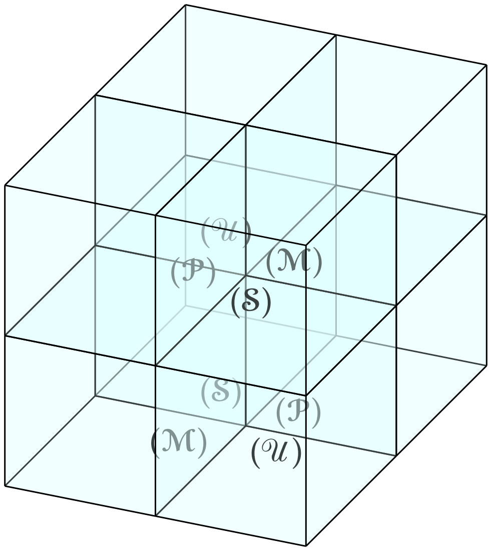


Figure 3.3: Illustration of the geometry and arrangement of subdomains used for a coupled formulation with which the code was verified via manufactured solutions.

To analyze convergence, only the $\boldsymbol{L}^2(\Omega)$ error of the displacement was considered and as Figure 3.4 demonstrates, high-order (or better) convergence rates were witnessed for each $1 \leq p \leq 5$, where the convergence rate was given in terms of the mesh element size, h. Given the smoothness of the solution, this is consistent with the theoretical expectations dictated by Theorem 2.4.

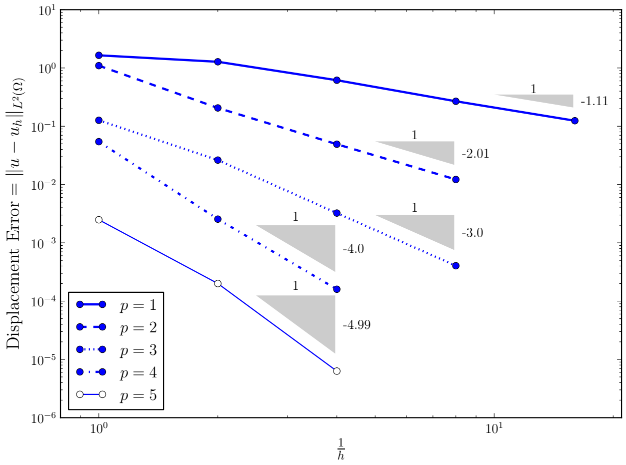


Figure 3.4: Displacement error as a function of the mesh size under uniform hexahedral refinements in the cubic domain $\Omega = (0, 2)^3$ with a sinusoidal manufactured solution.

3.3.2 Singular solution

Perhaps a more interesting test is that of a problem with a singular solution. A typical domain to elicit these solutions is the L-shape domain. A careful presentation in [76, §2.21–26] considers a 3D domain under plane strain or *averaged* plane stress conditions, where in both cases the analysis effectively reduces it to a two dimensional problem. Indeed, the L-shape domain example is prevalent as a 2D singular problem in the literature [225, 36, 135], especially the averaged plane stress case, which is elaborate to reformulate back into 3D [76, §2.26]. For this reason, in this work we consider the plane *strain* case in 3D.

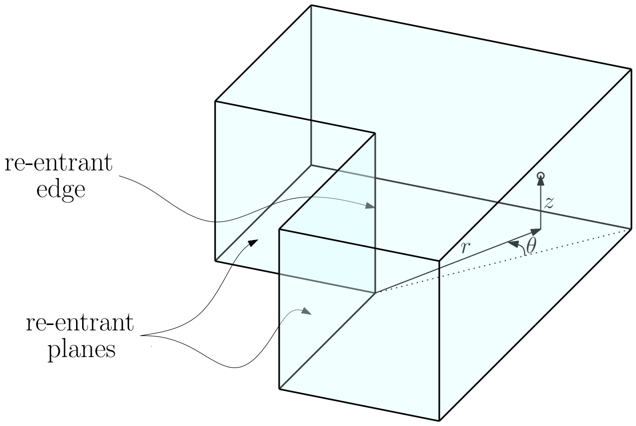


Figure 3.5: L-shape domain in a cylindrical system of coordinates.

As depicted in Figure 3.5, we considered an L-shape domain composed of three unit cubes and a cylindrical system of coordinates, (r, θ, z), such that the re-entrant edge passes through the origin and aligns with the z-axis, while the re-entrant planes align with $\theta = \pm\frac{3}{4}\pi$.

Using Airy functions (see [225]) one can obtain general expressions for the displacement components in polar coordinates of a homogeneous isotropic elastic body in equilibrium, so that $-\,\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u})) = \boldsymbol{f} = 0$. These are

$$\begin{aligned} u_r(r,\theta) &= \frac{1}{2\mu} r^a \Big(-(a+1)F(\theta) + (1-\nu)G'(\theta) \Big)\,, \\ u_\theta(r,\theta) &= \frac{1}{2\mu} r^a \Big(-F'(\theta) + (1-\nu)(a-1)G(\theta) \Big)\,, \\ u_z(r,\theta) &= 0\,, \end{aligned} \tag{3.33}$$

where $\nu = \frac{\lambda}{2(\lambda+\mu)}$ is the Poisson's ratio, a is a constant, and

$$\begin{aligned} F(\theta) &= C_1 \sin\big((a+1)\theta\big) + C_2 \cos\big((a+1)\theta\big) + C_3 \sin\big((a-1)\theta\big) + C_4 \cos\big((a-1)\theta\big)\,, \\ G(\theta) &= -\frac{4}{a-1}\Big(C_3 \cos\big((a-1)\theta\big) - C_4 \sin\big((a-1)\theta\big)\Big)\,. \end{aligned} \tag{3.34}$$

The nonzero stresses in polar coordinates satisfying the constitutive relation (and $\mathbf{div}\,\boldsymbol{\sigma} = 0$) are

$$\begin{aligned} \sigma_{rr}(r,\theta) &= r^{a-1}\big(F''(\theta) + (a+1)F(\theta)\big)\,, \\ \sigma_{\theta\theta}(r,\theta) &= a(a+1)r^{a-1}F(\theta)\,, \\ \sigma_{r\theta}(r,\theta) &= -ar^{a-1}F'(\theta)\,, \\ \sigma_{zz}(r,\theta) &= \lambda \mathrm{tr}_{\mathbb{M}}\big(\boldsymbol{\varepsilon}(\boldsymbol{u})\big)\,. \end{aligned} \tag{3.35}$$

Next, consider zero displacement boundary conditions at the re-entrant planes meaning that we want $u_r(r, \pm\frac{3}{4}\pi) = u_\theta(r, \pm\frac{3}{4}\pi) = 0$. The values of C_1, C_2, C_3, C_4 and a are essentially *chosen* to satisfy these boundary conditions. Indeed, choosing $C_2 = C_4 = 0$, $C_3 = 1$ and

$$C_1 = \frac{\big(4(1-\nu) - (a+1)\big)\sin\big((a-1)\frac{3}{4}\pi\big)}{(a+1)\sin\big((a+1)\frac{3}{4}\pi\big)} \tag{3.36}$$

guarantees that $u_r(r, \pm\frac{3}{4}\pi) = 0$ regardless of the value of a. After making this choice, the condition $u_\theta(r, \pm\frac{3}{4}\pi) = 0$ becomes

$$C_1(a+1)\cos\big((a+1)\tfrac{3}{4}\pi\big) + \big(4(1-\nu) + (a-1)\big)\cos\big((a-1)\tfrac{3}{4}\pi\big) = 0\,. \tag{3.37}$$

Moreover, since $\boldsymbol{\sigma}$ has a common factor of r^{a-1} it follows that $a > 0$ is required to have $\boldsymbol{\sigma} \in \mathbf{L}^2(\Omega; \mathbb{S})$, which in turn implies $\boldsymbol{\sigma} \in \mathbf{H}(\mathbf{div}, \Omega; \mathbb{S})$ in view of the intrinsic expression $\mathbf{div}\,\boldsymbol{\sigma} = 0$. Furthermore, to have an actual singularity in the strains and stresses it is necessary for $a < 1$. Hence, a is chosen to satisfy (3.37), with $a \in (0, 1)$.

For steel, the Lamé parameters are $\lambda = 123\,\mathrm{GPa}$ and $\mu = 79.3\,\mathrm{GPa}$. They yield $\nu \approx 0.304$ and a constant $a \approx 0.5946 \in (0, 1)$. These values are used in our computations. Additionally, we impose displacement boundary conditions at the re-entrant planes, and stress (traction) boundary conditions at the other faces parallel to the z-axis. The remaining two faces perpendicular to the z-axis are equipped with mixed boundary conditions where the displacement is restricted in the normal direction ($u_z = 0$) and where the tangential components of the traction vanish.

Remark 3.6. Under averaged plane stress conditions the problem is extremely similar to the plane strain case. The major difference is that the 2D displacements and stresses, u_r, u_θ, σ_{rr}, $\sigma_{r\theta}$ and $\sigma_{\theta\theta}$, are actually *averaged* quantities over the z direction. To solve the 2D problem for the averages simply consider the same equations as the plane strain case, but ignore u_z and σ_{zz}, and change ν to $\frac{\nu}{1+\nu}$ in (3.33), (3.35), (3.36) and (3.37) (see [225]). Recovering a 3D solution from the averaged quantities involves several calculations and is described in [76, §2.26].

3.3.2.1 Uniform refinements

The common factor of the stresses, r^{a-1}, actually implies that $\boldsymbol{\sigma}$ is in a space of fractional order s, which roughly speaking corresponds to $s = 1 + (a-1) - \delta = a - \delta$, where $\delta > 0$. Since $a \in (0,1)$, it follows that under *uniform* refinements the expected convergence rate with respect to h is approximately a, meaning the expected convergence rate with respect to degrees of freedom is $-\frac{a}{3} \approx -0.1982$ (since $h^{-3} = \mathcal{O}(N_{\mathrm{dof}})$), regardless of the value of p (see Theorem 2.4).

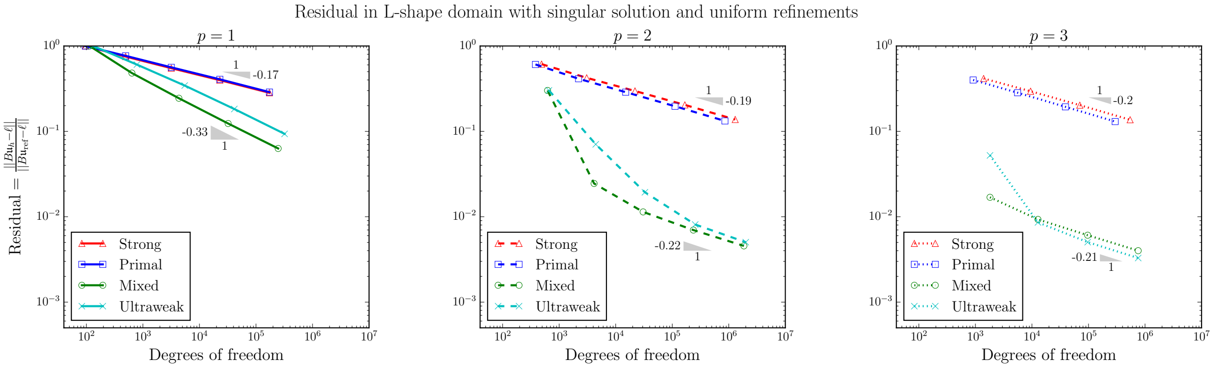


Figure 3.6: Residual as a function of the degrees of freedom after uniform hexahedral refinements in the L-shape domain with a singular solution.

The uniform refinement results for the variational formulations $(3.24)^{\mathcal{S}}$, $(3.25)^{\mathcal{U}}$, $(3.26)^{\mathcal{M}}$ and $(3.27)^{\mathcal{P}}$, are presented in Figure 3.6. As expected, the rates are very close to $-\frac{a}{3} \approx -0.1982$ when $p = 2$ and $p = 3$. When $p = 1$, the mixed and ultraweak methods seem to be converging at a higher rate (about 0.33), but this is probably because it has not reached the asymptotic regime where it stabilizes to the expected rate. The global residual, η_h, was computed using (2.50) (see also Remark 2.3 for $(3.24)^{\mathcal{S}}$ and $(3.26)^{\mathcal{M}}$). It was computed with the same level of enrichment

($p + \Delta p = 4$) as described in Remark 2.4, and normalized with a fixed reference residual, η_{ref}. For each formulation, as expected from the theory of minimum residual methods, the residual decreased both when the mesh was refined for a fixed p and also when p was refined for a fixed mesh. For example, the latter case is observed by looking at how the first point in the strong formulation (corresponding to the fixed initial three-element mesh) decreases in value as the order grows from $p = 1$ (left plot) to $p = 3$ (right plot). This comparison is valid in the discrete setting, only because a fixed value of $p + \Delta p = 4$ was used to compute the residual in all cases.

3.3.2.2 Adaptive refinements

To prevent the proliferation of degrees of freedom and to have some form of theoretical footing we use anisotropic refinements such that no refinements are done in the z direction, where $u_z = 0$. The element a posteriori error estimators, η_K, are calculated for each element separately as in (2.50) (see also Remark 2.3 for $(3.24)^{\mathcal{S}}$ and $(3.26)^{\mathcal{M}}$). The criteria for adaptivity is the one proposed in Remark 2.2 with $\alpha_\eta = 0.5$, where the marked elements are refined in the directions perpendicular to z. With these anisotropic adaptive refinements in place it is possible to use the 2D theory on point singularities from [14], which implies that in the asymptotic limit the expected rate should be equivalent to that coming from a smooth solution. That is, the rate with respect to N_{dof} is expected to approach $-\frac{p}{3}$ in the limit.

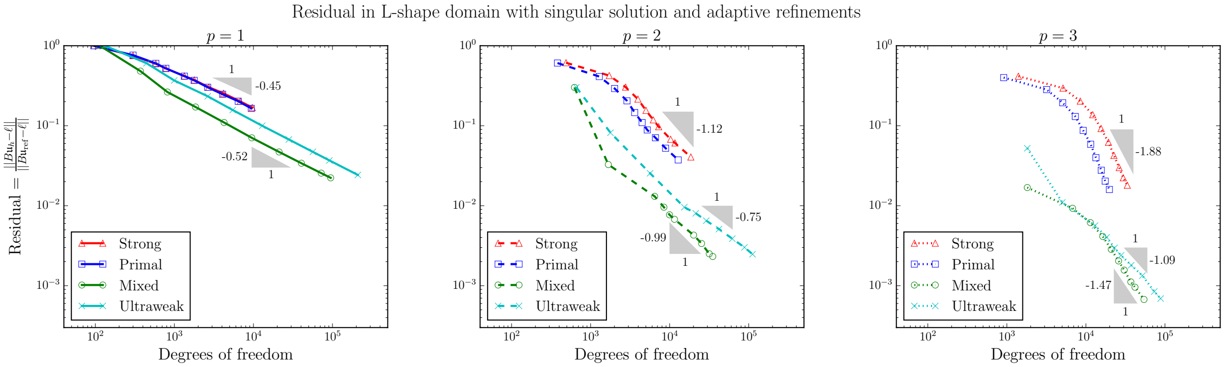


Figure 3.7: Residual as a function of the degrees of freedom after adaptive anisotropic hexahedral refinements in the L-shape domain with a singular solution.

The problem is solved successively through nine adaptive refinements with all formulations. The results are illustrated in Figure 3.7. For $p = 1$ the rates initially oscillate at around 0.5, which is much better than the expected 0.33. This is a desirable quality, because the pre-asymptotic rates are faster than the expected rates. Nevertheless, the rate would probably eventually approach the expected rate if more refinements had been taken. Similar assertions hold for $p = 2$ and $p = 3$. It is worth noting that the formulations in $(3.27)^{\mathcal{P}}$ and $(3.24)^{\mathcal{S}}$ have very similar and consistent behaviors with respect to convergence. On the other hand, for $p = 2$ and $p = 3$, the formulations in $(3.26)^{\mathcal{M}}$ and $(3.25)^{\mathcal{U}}$ seem to have a less consistent behavior with adaptive refinements.

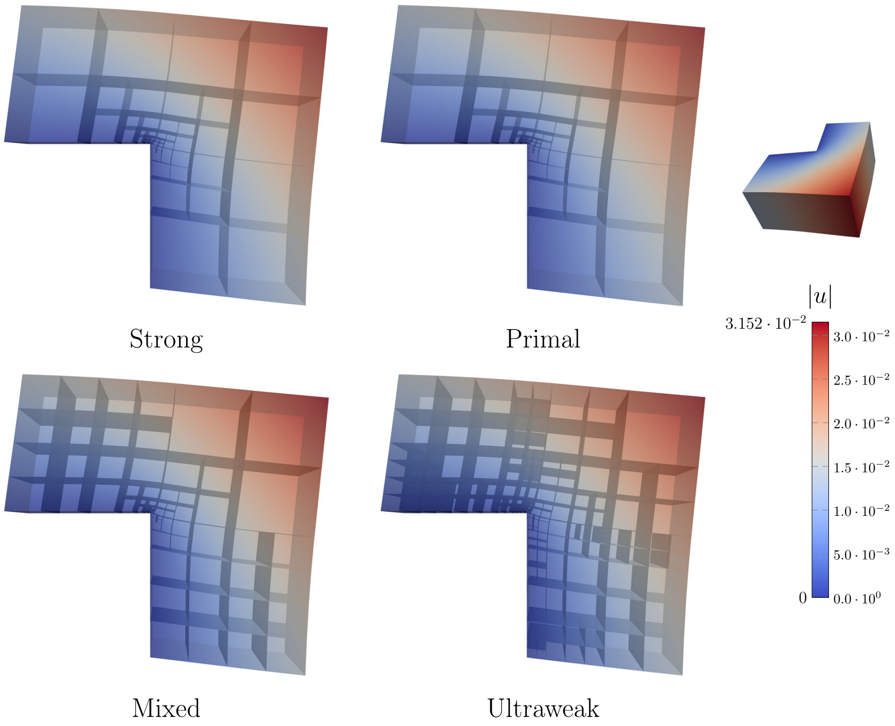


Figure 3.8: The adaptive meshes for each method after five successive refinements. The domains are colored by the displacement magnitude, $|\boldsymbol{u}|$, and warped by a factor of 10.

The adaptive refinement patterns for each of the different methods under this singular problem is interesting to analyze. Indeed, note that for Figure 3.7 the mixed and ultraweak formulations evidence a greater growth in degrees of freedom with each adaptive step. Figure 3.8 complements this by showing the resulting meshes for each of the methods after five refinements were performed.

As can be clearly seen, more elements have been refined with the mixed and ultraweak formulations than with the strong and primal formulations. This is especially evident far from the re-entrant edge (where the singularity lies). There could be many reasons for these refinement patterns, including the nature of the formulation itself and the choice of the test norm. Indeed, $(3.27)^{\mathcal{P}}$ and $(3.24)^{\mathcal{S}}$ have the displacement variable, whose gradient is singular, lying in $\boldsymbol{H}^1(\Omega)$, while the $(3.26)^{\mathcal{M}}$ and $(3.25)^{\mathcal{U}}$ have it lying in $\boldsymbol{L}^2(\Omega)$. This could imply that the residual is affected by those gradient terms, which leads to a much more focused pattern of refinements toward the singularity. On the other hand, the choice of test norm is completely fundamental and can have a profound effect on the computations. Here, we chose the standard norms. However, other choice of norms, such as graph norms for the ultraweak formulation, might lead to radically different refinement patterns.

3.3.3 Sheathed hose

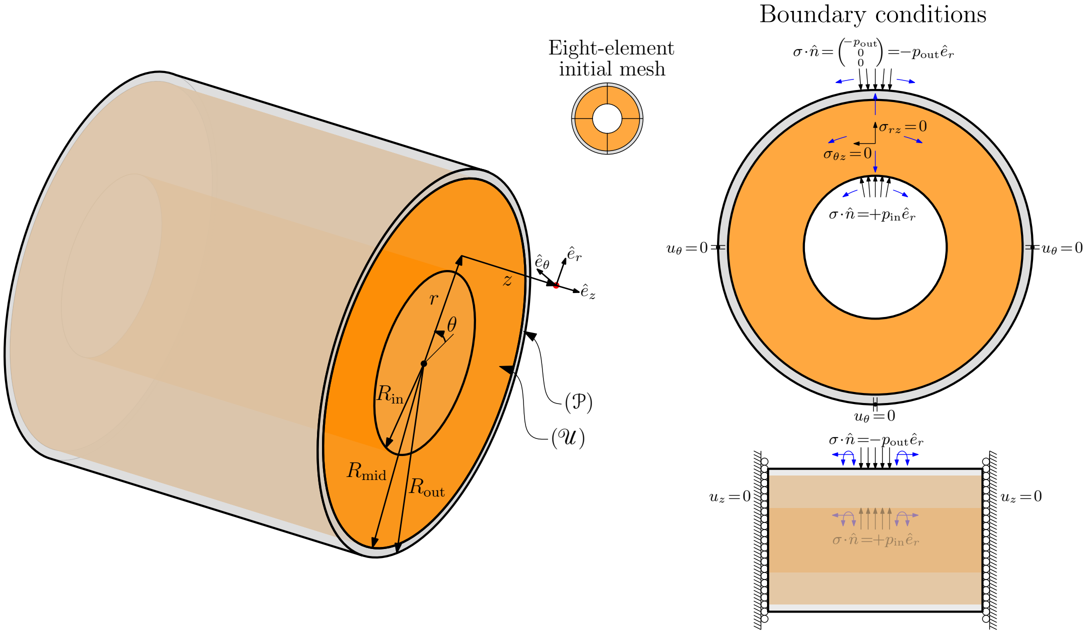

Eight-element initial mesh
Boundary conditions
$\sigma\cdot\hat{n}=\begin{pmatrix}-p_{\text{out}}\\0\\0\end{pmatrix}=-p_{\text{out}}\hat{e}_r$
$\sigma_{rz}=0$
$\sigma_{\theta z}=0$
$\sigma\cdot\hat{n}=+p_{\text{in}}\hat{e}_r$
$u_\theta=0$
$u_\theta=0$
$u_\theta=0$
$\sigma\cdot\hat{n}=-p_{\text{out}}\hat{e}_r$
$u_z=0$
$u_z=0$
$\sigma\cdot\hat{n}=+p_{\text{in}}\hat{e}_r$
$\hat{e}_\theta$
$\hat{e}_r$
$\hat{e}_z$
z
r
θ
R_{in}
R_{mid}
R_{out}
$(\mathcal{P})$
$(\mathcal{U})$

Figure 3.9: Diagram of the sheathed hose problem with the configuration of variational formulations per subdomain and a schematic of the boundary conditions used.

As a second example using coupled variational formulations, a rubber hose (hollow cylinder) sheathed by a layer of steel was considered as illustrated in Figure 3.9. This is a more physically-relevant example, because similar configurations are used in an array of applications including high-performance racing engine hoses and high-pressure hydraulic oil hoses (see SAE hydraulic hose standards), which have a steel braided outer sleeve. A variation of the example may also be pertinent to stents inside an artery.

To simulate balanced axial stresses which would appear in an infinite tube, the axial faces were confined by vanishing normal direction (axial) displacement boundary conditions and zero traction boundary conditions in the tangential directions as depicted in Figure 3.9. Moreover, normal pressure distributions were placed on the inner rubber surface, $p_{\mathrm{in}}(\theta,z)$ at R_{in}, and on the outer steel surface, $p_{\mathrm{out}}(\theta,z)$ at R_{out}, where θ represents the azimuthal direction and z represents the axial direction. Lastly, to have a supported structure with a fixed origin, three non-collinear points of one of the axial faces had their azimuthal displacement set to zero.

The thickness of the outer steel layer, $R_{\mathrm{out}}-R_{\mathrm{mid}}$, was assumed to be much smaller than the thickness of the rubber layer, $R_{\mathrm{mid}}-R_{\mathrm{in}}$. This implied the use of very thin elements provided each layer was discretized by the same number of elements in the radial direction, so that shear locking could have been a concern. Additionally, the rubber was taken to be the demanding case of a fully incompressible material, so that volumetric locking also had to be avoided. In principle, this makes the problem particularly challenging to solve, and therefore constitutes an ideal testing ground for the method being analyzed. Fortunately, the use of coupled variational formulations was extremely convenient, because, at least with regard to volumetric locking, all one needed to do was to choose a robustly well-posed variational formulation for the rubber subdomain coupled with a more efficient formulation in the steel subdomain, where robustness with respect to the material properties was not an issue. In fact, as shown in Figure 3.9, the broken ultraweak formulation, $(3.25)^{\mathcal{U}}$, was chosen for the rubber, while the broken primal formulation, $(3.27)^{\mathcal{P}}$, was chosen for the steel.

For reference, the steel had a Young's modulus of $E_{\mathrm{S}}=200\,\mathrm{GPa}$ and a Poisson's ratio of $\nu_{\mathrm{S}}=0.285$, and the rubber had a Young's modulus of $E_{\mathrm{R}}=0.01\,\mathrm{GPa}$ and a Poisson's ratio of $\nu_{\mathrm{R}}=0.5$. Then, the Lamé parameters were easily calculated using the formulas $\lambda=\frac{E\nu}{(1+\nu)(1-2\nu)}$ and $\mu=\frac{E}{2(1+\nu)}$. Meanwhile, the radii used were $R_{\mathrm{in}}=0.5\,\mathrm{m}$, $R_{\mathrm{mid}}=0.99\,\mathrm{m}$ and $R_{\mathrm{out}}=1.0\,\mathrm{m}$.

3.3.3.1 Uniform pressure distribution

For code verification, a 1D problem was essentially solved in 3D. Indeed, uniform pressure distributions were assumed to hold inside and outside, with values $p_{\text{in}} = 1\,\text{MPa}$ and $p_{\text{out}} = 0\,\text{MPa}$ respectively, so that they were independent of the azimuthal and axial directions, along with all the mechanics of the problem (i.e. $\frac{\partial u}{\partial \theta} = 0$ and $\frac{\partial u}{\partial z} = 0$). The remaining boundary conditions at the axial faces also implied that $u_\theta = 0$ and $u_z = 0$. Thus, the exact solution was derived from the ansatz that all nonvanishing physical variables were functions only of the radial direction, r. With these assumptions, the linear elasticity problem with no external volumetric forces reduces to the scalar equation

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}(r\sigma_{rr}) - \frac{1}{r}\sigma_{\theta\theta} = 0\,, \tag{3.38}$$

with boundary conditions, $(\boldsymbol{\sigma}\cdot\hat{\mathbf{n}}(R_{\text{in}}))_r = -\sigma_{rr}(R_{\text{in}}) = p_{\text{in}}$ and $(\boldsymbol{\sigma}\cdot\hat{\mathbf{n}}(R_{\text{out}}))_r = \sigma_{rr}(R_{\text{out}}) = -p_{\text{out}}$. For the steel, the nonzero stress components are

$$\begin{gathered}\sigma_{rr} = (2\mu_{\text{S}} + \lambda_{\text{S}})\frac{\mathrm{d}u_r}{\mathrm{d}r} + \lambda_{\text{S}}\frac{u_r}{r}\,, \qquad \sigma_{\theta\theta} = (2\mu_{\text{S}} + \lambda_{\text{S}})\frac{u_r}{r} + \lambda_{\text{S}}\frac{\mathrm{d}u_r}{\mathrm{d}r}\,,\\ \sigma_{zz} = \lambda_{\text{S}}\Big(\frac{\mathrm{d}u_r}{\mathrm{d}r} + \frac{u_r}{r}\Big)\,,\end{gathered} \tag{3.39}$$

while for the rubber there is the additional incompressibility equation $\text{div}(\boldsymbol{u}) = \frac{\mathrm{d}u_r}{\mathrm{d}r} + \frac{u_r}{r} = 0$ and the stress components are

$$\begin{gathered}\sigma_{rr} = 2\mu_{\text{R}}\frac{\mathrm{d}u_r}{\mathrm{d}r} - p_0\,, \qquad \sigma_{\theta\theta} = 2\mu_{\text{R}}\frac{u_r}{r} - p_0\,,\\ \sigma_{zz} = -p_0\,,\end{gathered} \tag{3.40}$$

for some constant, $p_0 \in \mathbb{R}$.

This boundary value problem has the general solution

$$u_r(r) = \begin{cases} Ar^{-1} & \text{if } R_{\text{in}} \leq r \leq R_{\text{mid}}\,,\\ Br + Cr^{-1} & \text{if } R_{\text{mid}} \leq r \leq R_{\text{out}}\,,\end{cases} \tag{3.41}$$

where the range $R_{\text{in}} \leq r \leq R_{\text{mid}}$ represents the rubber and the range $R_{\text{mid}} \leq r \leq R_{\text{out}}$ represents the steel. Upon matching displacements and tractions at the interface, R_{mid}, and applying the boundary conditions, the constants A, B, C and p_0 in their general form are given by the following expressions,

$$
\begin{aligned}
A &= \frac{1}{2d}\Big(- p_{\text{in}}\big(\mu_{\text{S}}R_{\text{mid}}^2 + (\lambda_{\text{S}} + \mu_{\text{S}})R_{\text{out}}^2\big) + p_{\text{out}}(\lambda_{\text{S}} + 2\mu_{\text{S}})R_{\text{out}}^2\Big) R_{\text{mid}}^2 R_{\text{in}}^2 \,, \\
B &= \frac{1}{2d}\Big(- p_{\text{in}}\mu_{\text{S}}R_{\text{in}}^2R_{\text{mid}}^2 - p_{\text{out}}\big((\mu_{\text{R}} - \mu_{\text{S}})R_{\text{in}}^2 - \mu_{\text{R}}R_{\text{mid}}^2\big)R_{\text{out}}^2\Big) \,, \\
C &= \frac{1}{2d}\Big(- p_{\text{in}}(\lambda_{\text{S}} + \mu_{\text{S}})R_{\text{in}}^2 + p_{\text{out}}\big((\lambda_{\text{S}} + \mu_{\text{R}} + \mu_{\text{S}})R_{\text{in}}^2 - \mu_{\text{R}}R_{\text{mid}}^2\big)\Big) R_{\text{mid}}^2 R_{\text{out}}^2 \,, \\
p_0 &= \frac{1}{d}\Big(p_{\text{in}}\big((\mu_{\text{R}} - \mu_{\text{S}})(\lambda_{\text{S}} + \mu_{\text{S}})R_{\text{out}}^2 + \mu_{\text{S}}(\lambda_{\text{S}} + \mu_{\text{R}} + \mu_{\text{S}})R_{\text{mid}}^2\big)R_{\text{in}}^2 \\
&\qquad\qquad - p_{\text{out}}\mu_{\text{R}}(\lambda_{\text{S}} + 2\mu_{\text{S}})R_{\text{out}}^2R_{\text{mid}}^2\Big) \,, \\
d &= \big((\mu_{\text{R}} - \mu_{\text{S}})(\lambda_{\text{S}} + \mu_{\text{S}})R_{\text{out}}^2 + \mu_{\text{S}}(\lambda_{\text{S}} + \mu_{\text{R}} + \mu_{\text{S}})R_{\text{mid}}^2\big)R_{\text{in}}^2 \\
&\qquad\qquad - \mu_{\text{R}}\big(\mu_{\text{S}}R_{\text{mid}}^2 + (\lambda_{\text{S}} + \mu_{\text{S}})R_{\text{out}}^2\big)R_{\text{mid}}^2 \,.
\end{aligned}
\tag{3.42}
$$

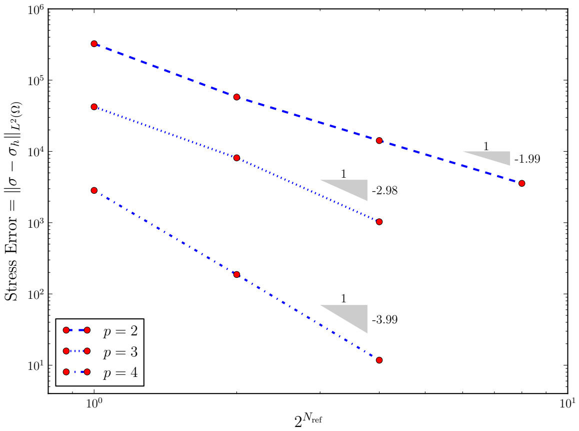


Figure 3.10: Stress error (in Pa) as a function of the number of uniform refinements, N_{ref}. The value of $p = 1$ was not shown because the isoparametric geometry was too inaccurate for the initial meshes.

In this example the convergence of the stress was presented. For this, the $\mathsf{L}^2(\Omega)$ error of the variable $\boldsymbol{\sigma}_h$ was reported, where $\boldsymbol{\sigma}_h$ is the $\mathsf{L}^2(\Omega^{\mathcal{U}})$ ultraweak stress solution variable inside the rubber and $\boldsymbol{\sigma}_h = \mathsf{C}\colon\boldsymbol{\nabla}\boldsymbol{u}_h$ inside the steel, with $\boldsymbol{u}_h$ being the $\boldsymbol{H}^1(\Omega^{\mathcal{P}})$ primal displacement solution variable. Order p convergence rates were expected for order p discretizations as stipulated by Theorem 2.4, because $\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathsf{L}^2(\Omega)} \leq \|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{\mathsf{L}^2(\Omega^{\mathcal{U}})} + \|\mathsf{C}\|\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{H}^1(\Omega^{\mathcal{P}})}$. This was corroborated numerically under uniform refinements for $2 \leq p \leq 4$ as observed in Figure 3.10.

3.3.3.2 Varying pressure distribution

Lastly, a nonuniform internal pressure distribution of $p_{\text{in}}(\theta) = \cos^2(\theta)\,\text{MPa}$ was prescribed on the inside, while the the external pressure was uniformly kept at $p_{\text{out}} = 0\,\text{MPa}$. After a few uniform refinements the solution is displayed in Figure 3.11 in each separate layer. Note that the discontinuity of the stress component, $\sigma_{\theta\theta}$, which is useful in some applications, was amicably reproduced.

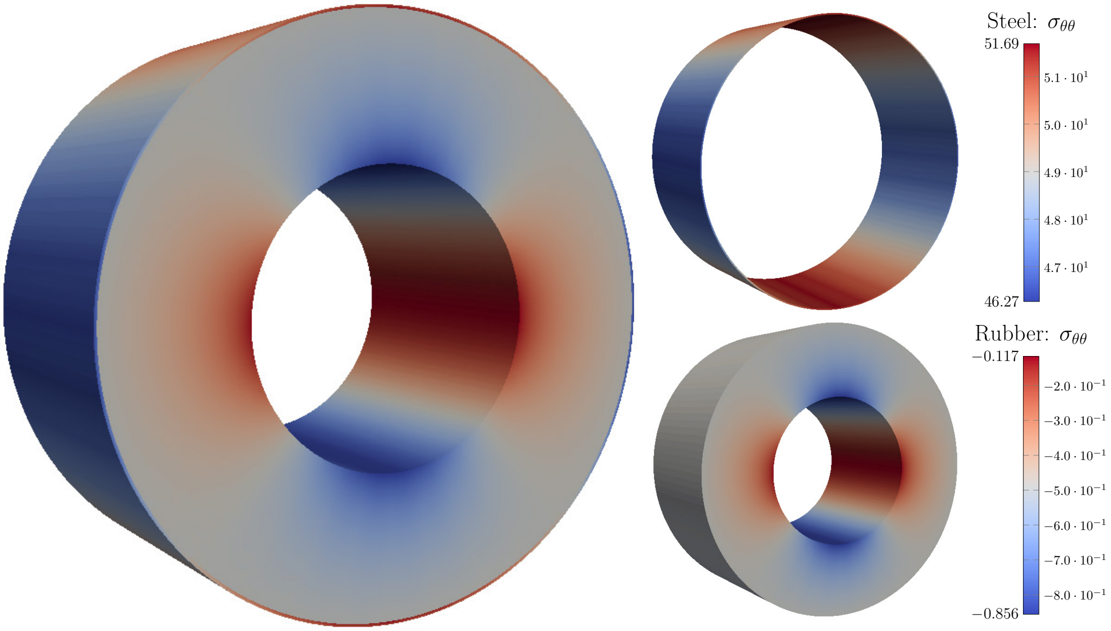


Figure 3.11: Stress component $\sigma_{\theta\theta}$ (in MPa) from computed solution with $p = 2$ and nonuniform internal pressure loading after three uniform refinements of the eight-element initial mesh. Note the discontinuity across the material interface.

3.4 Discussion

This chapter fulfilled two roles. First, it was a proof of concept in showing the adaptability of the DPG methodology to discretize different variational formulations. Second, it showed how to take advantage of the mathematical structure present in broken variational formulations to produce coupled variational formulations, where the variational setting changes heterogenously across the domain.

Eight different variational formulations for the equations of linear elasticity were introduced and showed to be mutually well-posed. The proof is interesting, as it uses the closed range theorem in its different forms to show how all the formulations are simultaneously well-posed, and this can be an interesting technique when applied to other problems (e.g. proving inequalities). Then, a subset of four broken variational formulations was derived, and also proved to be well-posed. High-order DPG methods were then constructed by discretizing the four broken variational formulations, and the resulting methods all evidenced the expected h-convergence rates coming from the theory. This was true for both smooth solutions and singular solutions (coming from an L-shaped domain). Moreover, a natural computation of the residual (in the context of DPG methods) was implemented for use in adaptive refinements. This allowed to solve the singular problem with adaptivity, and interesting results were observed in relation to the adaptive refinement patterns produced by the different formulations. Overall, the dexterity and versatility of DPG methods to successfully discretize variational formulations, including those with non-symmetric functional settings, was made evident.

Meanwhile, coupled variational formulations for linear elasticity were constructed using the family of four broken variational formulations, where each subdomain of a partitioned domain was solved with a distinct formulation from this family. The broken variational formulations that are commonplace in the context of the DPG methodology proved to be ideal in the theory and practice of the coupled formulations due to the presence of interface variables which served as a perfect vessel to transmit the solution information along the shared interdomain boundaries. In fact, the coupled formulations were also proved to be well-posed, and the proof is intriguing in its use of the ultraweak formulation as tool within the proof. The formulations were successfully implemented and solved using the DPG methodology. Expected convergence rates for various values of p were observed for different variables in several well-crafted examples.

The coupled formulations are useful in cases where one might want to exploit the properties of a particular formulation in a certain part of the domain. For example, it is useful to have a robust formulation when a part of the domain is composed of a nearly incompressible material, and certain formulations are more convenient when singular behavior of the solution is expected.

In this work, an example of a sheathed hose with high material contrast was used to illustrate the former point. This included the derivation of a nontrivial and physically-relevant exact solution which can be used as a benchmark by other researchers. Regarding the near singular behavior in the latter point, it would be interesting to study some examples with Maxwell's equations in the future. Additionally, from both a theoretical and practical standpoint, the approach presented in this chapter can be extended with the help of the existing literature to many other equations such as Poisson's equation, Maxwell's equations, the diffusion-convection-reaction equation, and more.

An important point to emphasize about solving coupled formulations with the DPG methodology is that depending on the subdomain formulation, the mode of convergence, which is given by the minimization of the residual (see (2.27)), is different because the residual is directly related to the bilinear form along with its associated trial and broken test spaces and their corresponding norms. Hence, there is a potential subdomain bias in the convergence of the solution variables depending on the variational formulations associated to each subdomain. This bias is even more stark when having a multi-material domain, which expectedly introduces different scales throughout the domain. If the focus is to be centered around optimality in solution estimation, as research in the DPG methodology often has [60, 93, 102], then this subdomain bias in the convergence is a crucial aspect to ponder. Indeed, it directly affects the (local) residual that is used as an a posteriori error estimator to drive adaptivity (see (2.50)). The bias itself is not necessarily undesirable, since it may align with preferences by the end user, but it may be important to understand and modify so that it further aligns with those preferences. For example, in the example of the sheathed hose (Section 3.3.3) one might want to prioritize the values of stress in the steel over those in the rubber, or the values of strain in the rubber over those in the steel. Thus, the ideal scenario is to be able to control the bias in accordance with a desired objective. With this in mind, the lucid approach is to attempt to design objective-biased test norms that satisfy this very purpose. In the examples in this chapter, the use of the standard test norms of the trial and test spaces introduced a natural yet unoptimized subdomain bias, but designing these objective-biased norms would most certainly constitute an improvement to the scope of the present work. It would be compelling to explore more exotic test norms eventually.

Another remark to make is that the coupled method used is akin to domain-decomposition methods and may even serve as an alternative. In fact, many steps of the DPG numerical method can be made parallel too, and the resulting connectivities along the interdomain boundaries are of a similar nature as those of other domain-decomposition methods. Conceptually, compared to domain-decomposition methods, the use of the DPG methodology in the current method has the advantage of providing a solid ground of theory which practically guarantees stability and convergence of the solution with successive refinements, but has the disadvantage of possibly coming at a slightly higher computational cost. In the future, it would be interesting to more rigorously investigate these connections with domain-decomposition methods. Lastly, the interface variables in the broken formulations are not only natural to couple distinct formulations, but also suggest a natural way to couple with entirely different finite element methods or with boundary element methods. The latter has already been further investigated in the context of elliptic transmission problems [117, 119, 120].

Chapter 4

Linear viscoelasticity: DMA experiments and calibration

This chapter is essentially composed of the material published as a research article in [111][‡]. It describes solving the equations of linear viscoelasticity for real problems using DPG methods. It is included in this dissertation, because it shows the applicability of the residual-based high-order a posteriori error estimator that is intrinsic to DPG methods. Indeed, it was useful in solving a real problem involving the calibration of material constants in dynamic mechanical analysis (DMA) experiments of viscoelastic materials. The contributions of the author to the multi-authored article were making all the computations, producing the mathematical proofs and writing the manuscript.

4.1 Introduction

The aim of this chapter is twofold. First, it is to implement a primal formulation of time-harmonic linear viscoelasticity with a DPG method. Second, it is to use such an implementation to validate calibration data directly from dynamic mechanical analysis (DMA) experimental results of the dynamic Young's modulus of two different thermoset resins. Many problems in viscoelasticity have local solution features in the stress or displacement, and the a posteriori error estimator is a very useful trait of the general DPG methodology which can be exploited in those cases. In fact, such local solution features will be observed when simulating the experimental results.

The chapter is outlined as follows. In Section 4.2 the equations of viscoelasticity along with the relevant variational formulations are introduced and proved to be well-posed. In Section 4.3 the DPG discretization is rigorously shown to be h-convergent by constructing a Fortin operator that implies its discrete stability. In Section 4.4, numerical results are presented. Both h- and p-convergence are analyzed in order verify the numerical scheme. Moreover, using experimental data, the validation of calibration models from DMA experiments is studied.

[‡] Fuentes, F., Demkowicz, L., and Wilder, A. (2017a). Using a DPG method to validate DMA experimental calibration of viscoelastic materials. *Comput. Methods Appl. Mech. Engrg.*, 325:748–765.

4.2 Primal variational formulations for viscoelasticity

4.2.1 Equations of linear viscoelasticity

The classical linear viscoelasticity equations are solved in this chapter. The constitutive model was originally developed by Boltzmann [35] and Volterra [231], but later recast more rigorously as a linearization of equations arising in nonlinear continuum mechanics under the additional assumption of a dependence of the stress on the deformation history [83, 78, 186, 188] (see also Appendix C). In the time domain, the first-order system describing a viscoelastic material with constant density $\rho > 0$ in a domain $\Omega \subseteq \mathbb{R}^3$ is

$$\begin{cases} \rho\ddot{\boldsymbol{u}} = \mathbf{div}\,\boldsymbol{\sigma} + \boldsymbol{f}\,, \\ \boldsymbol{\sigma} = \dot{\mathsf{C}} \ast \boldsymbol{\varepsilon} = \displaystyle\int_{-\infty}^{\infty} \dot{\mathsf{C}}(s) : \boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s\,, \end{cases} \tag{4.1}$$

where the displacement $\boldsymbol{u}$ and stress $\boldsymbol{\sigma}$ are unknown, $\boldsymbol{f}$ is a known body force, and the engineering strain is defined in terms of $\boldsymbol{u}$ as $\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^{\mathsf{T}})$. Meanwhile, the viscoelastic stiffness tensor C is in general not only a function in space, but also in time. With the typical assumption of $\mathsf{C}(t) = 0$ for times $t < 0$, this leads to the distributional derivative $\dot{\mathsf{C}}(t) = \mathsf{C}(0)\delta_0(t) + \dot{\mathsf{C}}^{+}(t)H_0(t)$, where $\dot{\mathsf{C}}^{+} = \frac{\mathrm{d}\mathsf{C}|_{(0,\infty)}}{\mathrm{d}t}$, δ_0 is the Dirac delta, and H_0 is the Heaviside step function. This yields the expression $\boldsymbol{\sigma} = \mathsf{C}(0) : \boldsymbol{\varepsilon} + \int_0^\infty \dot{\mathsf{C}}(s) : \boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s$, which is commonly found in the literature [142, 83, 105]. The relaxation or equilibrium stiffness tensor is $\mathsf{C}^\infty = \lim_{t\to\infty} \mathsf{C}(t)$. The classical case of linear elasticity occurs when $\mathsf{C}(t) = \mathsf{C}^\infty H_0(t)$ leading to $\boldsymbol{\sigma} = \mathsf{C}^\infty : \boldsymbol{\varepsilon}$.

In practice, many applications occur in a vibrating environment, so considering the time-harmonic case is natural. This also has the advantage of avoiding the computation of any convolutions, since $\dot{\mathsf{C}} \ast \boldsymbol{\varepsilon}$ becomes a product after using the Fourier transform. As usual, the stiffness tensor is assumed to have minor and major symmetries, so that $\mathsf{C}_{ijkl} = \mathsf{C}_{ijlk} = \mathsf{C}_{jikl} = \mathsf{C}_{klij}$ and as a result $\mathsf{C}_{ijkl}\boldsymbol{\tau}_{kl} = \mathsf{C}_{ijkl}\frac{1}{2}(\boldsymbol{\tau}_{kl} + \boldsymbol{\tau}_{lk})$ for any second-order tensor $\boldsymbol{\tau}$. In particular $\dot{\mathsf{C}} \ast \boldsymbol{\varepsilon} = \dot{\mathsf{C}} \ast \boldsymbol{\nabla u}$. Thus, substituting the constitutive model for the stress into the conservation of momentum, and considering the time-harmonic case at angular frequency ω, yields the second-order equation,

$$-\,\omega^2\rho\boldsymbol{u} - \mathbf{div}(\mathsf{C}^* : \boldsymbol{\nabla u}) = \boldsymbol{f}\,, \tag{4.2}$$

where, given $\boldsymbol{x} \in \Omega$, the complex-valued $\boldsymbol{u}(\boldsymbol{x},\omega)$, $\boldsymbol{f}(\boldsymbol{x},\omega)$ and $\mathsf{C}(\boldsymbol{x},\omega)$ are the corresponding Fourier transforms of the time-dependent displacement, force and stiffness tensor; and where the dynamic stiffness tensor is defined as $\mathsf{C}^*(\boldsymbol{x},\omega) = \mathrm{i}\omega\mathsf{C}(\boldsymbol{x},\omega)$. Note that in the limiting case of linear elasticity, $\mathsf{C}^* = \mathsf{C}^\infty$, so C^* is no longer complex-valued or ω-dependent. For isotropic materials, the stiffness tensor explicitly takes the form,

$$\mathsf{C}_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\,, \tag{4.3}$$

and similarly with C^* in terms of $\lambda^* = \mathrm{i}\omega\lambda$ and $\mu^* = \mathrm{i}\omega\mu$, where in the latter expression, λ and μ are the Fourier transforms of the time-dependent Lamé parameters. Moreover, $G^* = \mu^*$ is the dynamic shear modulus, $K^* = \lambda^* + \frac{2}{3}\mu^*$ is the dynamic bulk modulus, $E^* = \frac{\mu^*(3\lambda^*+2\mu^*)}{\lambda^*+\mu^*}$ is the dynamic Young's modulus, and $\nu^* = \frac{\lambda^*}{2(\lambda^*+\mu^*)}$ is the dynamic Poisson's ratio. Notably, E^* is a nonlinear function of λ^* and μ^*, implying that in general it is not the Fourier transform of $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mu(t)(3\lambda(t)+2\mu(t))}{\lambda(t)+\mu(t)}$. A similar assertion follows for ν^*. Thus, one should be careful when speaking of the time-dependent Young's modulus and Poisson's ratio in three-dimensional viscoelasticity as even different definitions derived from physical principles exist in the literature [164, §5.7].

The goal is to solve the second-order equation in (4.2) for the unknown displacement, provided the forcing and the dynamic stiffness tensor of the material are known throughout the domain $\Omega \subseteq \mathbb{R}^3$ at the angular frequency ω. For this to be possible, boundary conditions need to be specified. Thus, it will be assumed that the boundary is partitioned into relatively open subsets Γ_u and Γ_σ satisfying $\overline{\Gamma_u \cup \Gamma_\sigma} = \partial\Omega$ and $\Gamma_u \cap \Gamma_\sigma = \varnothing$, where displacement and traction boundary conditions are set by the known functions $\boldsymbol{u} = \boldsymbol{u}^{\Gamma_u}$ and $(\mathsf{C}^* : \boldsymbol{\nabla u})\cdot\hat{\mathbf{n}} = \boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$ on Γ_u and Γ_σ respectively, with $\hat{\mathbf{n}}$ being the outward normal at $\partial\Omega$. From now on it will be assumed that $\Gamma_u \neq \varnothing$ and Ω is bounded and Lipschitz.

Remark 4.1. As stated here, the functions will now be complex-valued, so they take values in $\mathbb{C}$. This means that the inner products (see (2.3)), and all the definitions of the Hilbert spaces in Appendix A and Section 3.2.1 have to account for this change. In particular, the inner product will now be sesquilinear. The theory from Chapter 2 will apply, but after the necessary modifications are made.

4.2.2 Classical primal formulation

The usual approach to solve the second-order equation is to multiply by a smooth enough test function that vanishes at Γ_u, and then integrate by parts once. For every test function $\boldsymbol{v}$ this yields the expression

$$b_0(\boldsymbol{u},\boldsymbol{v}) = -\omega^2\rho(\boldsymbol{u},\boldsymbol{v})_\Omega + (\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v})_\Omega = (\boldsymbol{f},\boldsymbol{v})_\Omega + \langle \boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}, \boldsymbol{v}\rangle_{\partial\Omega} = \ell(\boldsymbol{v})\,. \tag{4.4}$$

To prove the convergence and stability of any numerical method aiming to solve (4.4), usually determining well-posedness of the underlying non-discrete equations is either necessary or extremely useful. For this, a deeper understanding of the functional spaces used as trial and test spaces is required. Indeed, when $\boldsymbol{u}^{\Gamma_u} = 0$, the natural choice of space for $\boldsymbol{u}$ and $\boldsymbol{v}$ is $\boldsymbol{H}^1_{\Gamma_u}(\Omega)$. When, $\boldsymbol{u}^{\Gamma_u} \neq 0$, the final displacement takes the form $\boldsymbol{u}_f = \boldsymbol{u} + \widetilde{\boldsymbol{u}}^{\Gamma_u}$, where $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\widetilde{\boldsymbol{u}}^{\Gamma_u} \in \boldsymbol{H}^1(\Omega)$ is an extension of $\boldsymbol{u}^{\Gamma_u}$ to Ω. For simplicity consider $\boldsymbol{u}^{\Gamma_u} = 0$, let $\mathcal{U} = \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and assume that for some $C > 0$, $|\ell(\boldsymbol{v})| \leq C\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)}$ for all $\boldsymbol{v} \in \mathcal{U}$, so that $\ell \in \mathcal{U}'$. Then, solving (4.4) can be rewritten as the problem that aims to find $\boldsymbol{u} \in \mathcal{U}$ such that

$$b_0(\boldsymbol{u},\boldsymbol{v}) = \ell(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathcal{U}\,. \tag{4.5}$$

This is referred to as the primal variational formulation, and the goal is to prove that it is well-posed in the sense of Hadamard, so that there is a guaranteed existence of a unique solution depending continuously upon the forcing and boundary conditions (encoded in ℓ). The proof is presented in what remains of the section, where a bounded $\Omega \subseteq \mathbb{R}^3$ and $\Gamma_u \neq \varnothing$ are assumed throughout. It is based on the use of the Fredholm alternative and the theory of Gelfand triples in the same spirit as [144, 189, 155].

Lemma 4.1. *Let $b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{v}) = (\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v})_\Omega$, for $\boldsymbol{u},\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and with C^* being a fourth-order tensor with major and minor symmetries satisfying $|\bar{\boldsymbol{\varepsilon}} : \mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon}| > 0$ for all symmetric second-order tensors $\boldsymbol{\varepsilon} \neq 0$. Then, $|b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{u})| \geq \alpha\|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)}$ for some $\alpha > 0$.*

Proof. First note that $\overline{\boldsymbol{\nabla}\boldsymbol{u}} : \mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u} = \bar{\boldsymbol{\varepsilon}} : \mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon} + \mathrm{i}\bar{\boldsymbol{\varepsilon}} : \mathfrak{Im}(\mathsf{C}^*) : \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^{\mathsf{T}})$. The major symmetry of C^* clearly implies that both $\bar{\boldsymbol{\varepsilon}} : \mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon}$ and $\bar{\boldsymbol{\varepsilon}} : \mathfrak{Im}(\mathsf{C}^*) : \boldsymbol{\varepsilon}$ are real-valued, so that

$$|b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{u})|^2 = |(\mathfrak{Re}(\mathsf{C}^*)\!:\!\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega|^2 + |(\mathfrak{Im}(\mathsf{C}^*)\!:\!\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega|^2 \geq |(\mathfrak{Re}(\mathsf{C}^*)\!:\!\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega|^2 \,.$$

Due to the symmetries, $\mathfrak{Re}(\mathsf{C}^*)$ and $\boldsymbol{\varepsilon}$ may be reinterpreted in Voigt notation as a symmetric 6×6 matrix and a vector in $\mathbb{C}^6$ respectively, so that the Rayleigh quotient of the Voigt-matrix $\mathfrak{Re}(\mathsf{C}^*)$ takes the form $\bar{\boldsymbol{\varepsilon}} : \mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon}/(\bar{\boldsymbol{\varepsilon}} : \boldsymbol{\varepsilon} + 2|\boldsymbol{\varepsilon}_{12}|^2 + 2|\boldsymbol{\varepsilon}_{13}|^2 + 2|\boldsymbol{\varepsilon}_{23}|^2)$. By hypothesis, 0 is not in the Rayleigh quotient's range, implying $|(\mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega| = \int_\Omega |\bar{\boldsymbol{\varepsilon}} : \mathfrak{Re}(\mathsf{C}^*) : \boldsymbol{\varepsilon}|\, \mathrm{d}\Omega$, because its range is either fully positive or fully negative. If the range is positive, the Rayleigh quotient yields $|\bar{\boldsymbol{\varepsilon}}\!:\!\mathfrak{Re}(\mathsf{C}^*)\!:\!\boldsymbol{\varepsilon}| \geq \lambda_{\min}(\bar{\boldsymbol{\varepsilon}}\!:\!\boldsymbol{\varepsilon} + 2|\boldsymbol{\varepsilon}_{12}|^2 + 2|\boldsymbol{\varepsilon}_{13}|^2 + 2|\boldsymbol{\varepsilon}_{23}|^2) \geq \lambda_{\min}\bar{\boldsymbol{\varepsilon}}\!:\!\boldsymbol{\varepsilon}$, where $\lambda_{\min} > 0$ is the smallest eigenvalue of the Voigt-matrix $\mathfrak{Re}(\mathsf{C}^*)$. Similarly if the range is negative, so that in any case $|b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{u})| \geq \alpha(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega$ for some $\alpha > 0$. The result follows because Korn's and Poincaré inequalities ($\Gamma_u \neq \varnothing$) imply that for all $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$, $(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega \geq \alpha\|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)}$ for some $\alpha > 0$. □

Remark 4.2. In the case of isotropic materials, the conditions on the dynamic stiffness tensor, C^*, are equivalent to $\mathfrak{Re}(G^*)\mathfrak{Re}(K^*) > 0$. The physically-relevant case is when both the storage shear and bulk moduli are positive, $\mathfrak{Re}(G^*) > 0$ and $\mathfrak{Re}(K^*) > 0$, but exotic exceptions do exist where the storage bulk modulus may be negative [165]. Curiously, if $\mathfrak{Re}(G^*) \neq 0$ and $\mathfrak{Re}(K^*) = 0$, then $\mathbf{I} : \mathfrak{Re}(\mathsf{C}^*) : \mathbf{I} = 0$, but the coercive inequality $|b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{u})| \geq \alpha\|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)}$ still holds, because $|(\mathfrak{Re}(\mathsf{C}^*)\!:\!\boldsymbol{\varepsilon},\boldsymbol{\varepsilon})_\Omega| = 2\mathfrak{Re}(G^*)(\boldsymbol{\varepsilon}^D,\boldsymbol{\varepsilon}^D)_\Omega$, where $\boldsymbol{\varepsilon}^D = \boldsymbol{\varepsilon} - \frac{1}{3}\mathrm{tr}(\boldsymbol{\varepsilon})\mathbf{I}$ is the deviatoric part of the strain. Then, all that remains is to apply a recently proved and more general version of Korn's inequality, $(\boldsymbol{\varepsilon}^D,\boldsymbol{\varepsilon}^D)_\Omega \geq \alpha(\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{u})_\Omega$ for all $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\alpha > 0$ [182].

Remark 4.3. In the particular case of static linear elasticity, $\mathsf{C}^* = \mathfrak{Re}(\mathsf{C}^*) = \mathsf{C}^\infty$ and the primal formulation is that of finding $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ such that $b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{v}) = \ell(\boldsymbol{v})$ for all $\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$. Thus, a straightforward application of the Lax-Milgram theorem yields the well-posedness of the static linear elasticity equation provided $|\boldsymbol{\varepsilon}\!:\!\mathsf{C}^\infty\!:\!\boldsymbol{\varepsilon}| > 0$ for all symmetric strains $\boldsymbol{\varepsilon} \neq 0$. If the material is isotropic this implies $G^\infty K^\infty > 0$, and in particular, the equations are well-posed for positive shear and bulk moduli. For even more general conditions (in terms of the compliance tensor, $\mathsf{S}^\infty = (\mathsf{C}^\infty)^{-1}$) under which static linear elasticity remains well-posed, see [8].

Theorem 4.1. *Let $\mathcal{U}=\boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and consider the problem of finding $\boldsymbol{u}\in\mathcal{U}$ such that $b_0(\boldsymbol{u},\boldsymbol{v})=\ell(\boldsymbol{v})$ for all $\boldsymbol{v}\in\mathcal{U}$, where $b_0(\boldsymbol{u},\boldsymbol{v})=-\omega^2\rho(\boldsymbol{u},\boldsymbol{v})_\Omega+(\mathsf{C}^*:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{v})_\Omega$ and $\ell\in\mathcal{U}'$, and where it is assumed $|\bar{\boldsymbol{\varepsilon}}:\mathfrak{Re}(\mathsf{C}^*):\boldsymbol{\varepsilon}|>0$ for all symmetric second-order tensors $\boldsymbol{\varepsilon}\neq 0$. Then, for each value of ω, either there exists $0\neq\boldsymbol{u}\in\mathcal{U}$ such that $b_0(\boldsymbol{u},\boldsymbol{v})=0$ for all $\boldsymbol{v}\in\mathcal{U}$, or, given any $\ell\in\mathcal{U}'$, there exists a unique solution $\boldsymbol{u}\in\mathcal{U}$ solving $b_0(\boldsymbol{u},\boldsymbol{v})=\ell(\boldsymbol{v})$ for all $\boldsymbol{v}\in\mathcal{U}$ which satisfies $\|\boldsymbol{u}\|_{\mathcal{U}}\leq C\|\ell\|_{\mathcal{U}'}$ for a $C>0$ independent of the choice of ℓ. Furthermore, the former case, where infinitely many solutions of the form $\beta\boldsymbol{u}\in\mathcal{U}$ for $\beta\in\mathbb{C}$ exist, only holds for a countable set of values of ω which has no accumulation points.*

Proof. First define the linear operator $\mathcal{B}:\mathcal{U}\to\mathcal{U}'$ as $\langle\mathcal{B}\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}'\times\mathcal{U}}=b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{v})=(\mathsf{C}^*:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{v})_\Omega$ for all $\boldsymbol{v}\in\mathcal{U}$. Lemma 4.1 implies that $\mathcal{B}$ is bounded below, $\|\mathcal{B}\boldsymbol{u}\|_{\mathcal{U}'}\geq\alpha\|\boldsymbol{u}\|_{\mathcal{U}}$, with some $\alpha>0$, so that $\mathcal{B}$ is injective and its range, $\mathsf{R}(\mathcal{B})=\big\{\ell\in\mathcal{U}'\mid\ell|_{\mathcal{U}_{00}}=0,\,\mathcal{U}_{00}=\{\boldsymbol{v}\in\mathcal{U}\mid b_{\mathsf{C}^*}(\boldsymbol{u},\boldsymbol{v})=0\;\forall\boldsymbol{u}\in\mathcal{U}\}\big\}$, is closed. Again by Lemma 4.1, $\mathcal{U}_{00}=\{0\}$ and $\mathsf{R}(\mathcal{B})=\mathcal{U}'$, so the open mapping theorem implies $\mathcal{B}^{-1}:\mathcal{U}'\to\mathcal{U}$ is bounded. Assume the embedding $\iota:\mathcal{U}\to\mathcal{U}'$, defined naturally as $\langle\iota\boldsymbol{u},\boldsymbol{v}\rangle=(\boldsymbol{u},\boldsymbol{v})_\Omega$ for all $\boldsymbol{v}\in\mathcal{U}$, is compact, so that the operator $\mathcal{K}=\iota\mathcal{B}^{-1}:\mathcal{U}'\to\mathcal{U}'$ is a compact operator, with range $\mathsf{R}(\mathcal{K})=\iota(\mathcal{U})$. Given $\omega\neq 0$ (so $\omega^2\rho\neq 0$), the Fredholm alternative applies to $\mathcal{K}-\frac{1}{\omega^2\rho}\mathrm{id}$. Hence, either there exists $0\neq\iota\boldsymbol{u}=\boldsymbol{v}\in\mathsf{R}(\mathcal{K})$ such that $\mathcal{K}\boldsymbol{v}-\frac{1}{\omega^2\rho}\boldsymbol{v}=0$, or $\mathcal{K}-\frac{1}{\omega^2\rho}\mathrm{id}:\mathcal{U}'\to\mathcal{U}'$ is a homeomorphism. In the first case, this would imply $-\omega^2\rho\mathcal{B}\iota^{-1}(\mathcal{K}\boldsymbol{v}-\frac{1}{\omega^2\rho}\boldsymbol{v})=-\omega^2\rho\iota\boldsymbol{u}+\mathcal{B}\boldsymbol{u}=0$ for such $\boldsymbol{u}\in\mathcal{U}$. In the second case this implies $-\omega^2\rho(\mathcal{K}-\frac{1}{\omega^2\rho}\mathrm{id})\mathcal{B}:\mathcal{U}\to\mathcal{U}'$ is a homeomorphism, so there exists a unique solution $\boldsymbol{u}\in\mathcal{U}$ to the equation $-\omega^2\rho(\mathcal{K}-\frac{1}{\omega^2\rho}\mathrm{id})\mathcal{B}\boldsymbol{u}=-\omega^2\rho\iota\boldsymbol{u}+\mathcal{B}\boldsymbol{u}=\ell$ for any $\ell\in\mathcal{U}'$ which satisfies that $\|\boldsymbol{u}\|_{\mathcal{U}}\leq\|(-\omega^2\rho(\mathcal{K}-\frac{1}{\omega^2\rho}\mathrm{id})\mathcal{B})^{-1}\|\|\ell\|_{\mathcal{U}'}$. When $\omega=0$ and for any $\ell\in\mathcal{U}'$, obviously $\|\boldsymbol{u}\|_{\mathcal{U}}\leq\|\mathcal{B}^{-1}\|\|\ell\|_{\mathcal{U}'}$, where $\boldsymbol{u}=\mathcal{B}^{-1}\ell$ is the unique solution to $\mathcal{B}\boldsymbol{u}=\ell$.

From the theory of compact operators the set of eigenvalues of $\mathcal{K}$ is countable, bounded, and can only accumulate at 0. Since the eigenvalues considered are of the form $\frac{1}{\omega^2\rho}$, it follows that their inverses, $\omega^2\rho$, are also countable and have no accumulation point.

It remains to show the embedding $\iota:\mathcal{U}\to\mathcal{U}'$ is compact. This is due to the fact that $(\mathcal{U},\mathcal{V},\mathcal{U}')$ is a Gelfand triple, with $\mathcal{V}=\boldsymbol{L}^2(\Omega)$. More precisely, the natural embedding $\iota_{\mathcal{V}}:\mathcal{U}\to\mathcal{V}$, $\iota_{\mathcal{V}}\boldsymbol{u}=\boldsymbol{u}$, is continuous by the Sobolev embedding theorem, and moreover $\overline{\mathcal{U}}^{\mathcal{V}}=\mathcal{V}$ since $\mathcal{U}$ contains all smooth functions vanishing in $\partial\Omega$ which are well known to be dense in $\mathcal{V}$. Thus, the transpose $\iota_{\mathcal{V}}^{\mathsf{T}} : \mathcal{V}' \to \mathcal{U}'$ is continuous, takes the form $\iota_{\mathcal{V}}^{\mathsf{T}}\boldsymbol{v} = \boldsymbol{v}|_{\mathcal{U}}$, and is injective by the density of $\mathcal{U}$ in $\mathcal{V}$. Let $\mathcal{R}_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ be the Riesz map, which explicitly takes the form $\langle \mathcal{R}_{\mathcal{V}}\boldsymbol{u}, \boldsymbol{v}\rangle = (\boldsymbol{u}, \boldsymbol{v})_\Omega$, and is known to be continuous and bijective by the Riesz representation theorem. Thus, the original embedding $\iota_{\mathcal{V}}^{\mathsf{T}}\mathcal{R}_{\mathcal{V}}\iota_{\mathcal{V}} = \iota : \mathcal{U} \to \mathcal{U}'$ is injective and compact, because $\iota_{\mathcal{V}}$ is compact by the Rellich-Kondrachov theorem. □

Remark 4.4. The theorem can be generalized to spatially heterogeneous (but constant in time) densities, as long as $\rho_{\min} < \rho(\boldsymbol{x}) < \rho_{\max}$ for all $\boldsymbol{x} \in \Omega$, where $\rho_{\min} > 0$ and $\rho_{\max} > 0$ are constants.

Remark 4.5. Theorem 4.1 shows that (4.5) is well-posed for almost every value of ω, with the exception of some critical values which are essentially spread out in the real-number line. At these critical values the system is said to be in resonance, and a unique solution does not exist. Indeed, the constant C in the statement of the theorem, which is ω-dependent, blows up as these resonant frequencies are approached. Thus, when close to these frequencies, numerical schemes discretizing these equations, even if theoretically stable, are usually very ill-conditioned and round-off error may play an undesirable role (see [160]).

4.2.3 Broken primal formulation

In the study of discontinuous finite element methods it is common to merely consider functions that element-wise have a particular regularity and are possibly discontinuous at the boundaries of the elements, instead of requiring those functions to have the regularity at a global level. This leads to broken spaces dependent on a mesh (a relatively open partition of Ω), $\mathcal{T}$ (see Appendix A and Section 3.2.1 for definitions). Proceeding as with the classical case, but this time multiplying (4.2) by a broken test function $\boldsymbol{v} \in \boldsymbol{H}^1(\mathcal{T})$ yields,

$$\begin{aligned} &-\omega^2\rho(\boldsymbol{u}, \boldsymbol{v})_{\mathcal{T}} + (\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}} - \langle(\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u})\cdot\hat{\mathbf{n}}, \boldsymbol{v}\rangle_{\partial\mathcal{T}} = (\boldsymbol{f}, \boldsymbol{v})_{\mathcal{T}}\,, \\ &(\boldsymbol{u}, \boldsymbol{v})_{\mathcal{T}} = \sum_{K\in\mathcal{T}} (\boldsymbol{u}|_K, \boldsymbol{v}|_K)_K\,, \qquad \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\partial\mathcal{T}} = \sum_{K\in\mathcal{T}} \langle \boldsymbol{u}_K, \boldsymbol{v}_K\rangle_{\partial K}\,. \end{aligned} \tag{4.6}$$

This is still not a well-defined formulation because $\langle\,\cdot\,,\,\cdot\,\rangle_{\partial K}$ needs to be interpreted rigorously and the expression $(\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u})\cdot\hat{\mathbf{n}}$ needs to be replaced by an appropriate interface variable along the mesh skeleton. The variable itself represents a traction, so it must come from a stress. As seen in Chapter 3 (and Section 2.2 as well), the appropriate space for these tractions is $\boldsymbol{H}_{\Gamma_\sigma}^{-1/2}(\partial\mathcal{T})$, which in turn come from stresses in $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)$ (see Section 3.2.1 for the definition of those spaces). The corresponding trace operators for variables in $\boldsymbol{H}^1(K)$ and $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},K)$ are $\mathbf{tr}_{\text{grad}}^K$ and $\mathbf{tr}_{\text{div}}^K$, while for variables in $\boldsymbol{H}^1(\mathcal{T})$ and $\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\mathcal{T})$ they are $\mathbf{tr}_{\text{grad}}^{\mathcal{T}}$ and $\mathbf{tr}_{\text{div}}^{\mathcal{T}}$ (see Appendix A for definitions).

Assuming vanishing boundary conditions, $\boldsymbol{u}^{\Gamma_u}=0$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}=0$, the broken primal variational formulation is defined by

$$\begin{aligned} b_{\mathcal{T}}\big((\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}),\boldsymbol{v}\big)&=b_0(\boldsymbol{u},\boldsymbol{v})+\hat{b}(\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\boldsymbol{v})\,, & \ell_{\mathcal{T}}(\boldsymbol{v})&=(\boldsymbol{f},\boldsymbol{v})_{\mathcal{T}}\,,\\ b_0(\boldsymbol{u},\boldsymbol{v})&=-\omega^2\rho(\boldsymbol{u},\boldsymbol{v})_{\mathcal{T}}+(\mathsf{C}^*:\boldsymbol{\nabla}\boldsymbol{u},\boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}}\,, & \hat{b}(\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\boldsymbol{v})&=-\langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}_{\text{grad}}^{\mathcal{T}}\boldsymbol{v}\rangle_{\partial\mathcal{T}}\,,\end{aligned}\tag{4.7}$$

where $\boldsymbol{u}\in\mathcal{U}=\boldsymbol{H}_{\Gamma_u}^1(\Omega)$, $\hat{\boldsymbol{\sigma}}_{\mathbf{n}}\in\hat{\mathcal{U}}=\boldsymbol{H}_{\Gamma_\sigma}^{-1/2}(\partial\mathcal{T})$, $\boldsymbol{v}\in\mathcal{V}_{\mathcal{T}}=\boldsymbol{H}^1(\mathcal{T})$, and the trial space $\mathcal{U}_{\mathcal{T}}=\mathcal{U}\times\hat{\mathcal{U}}$ is equipped with its Hilbert norm. Note that in relation to (4.4), the domain of the test space of b_0 was extended from $\boldsymbol{H}_{\Gamma_u}^1(\Omega)$ to $\boldsymbol{H}^1(\mathcal{T})$. Meanwhile, it is clear $\ell_{\mathcal{T}}\in\mathcal{V}_{\mathcal{T}}'$. Thus, the broken primal formulation that solves (4.2) is equivalent to seeking $\boldsymbol{u}\in\mathcal{U}$ such that,

$$b_{\mathcal{T}}\big((\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}),\boldsymbol{v}\big)=\ell_{\mathcal{T}}(\boldsymbol{v})\qquad\forall\boldsymbol{v}\in\mathcal{V}_{\mathcal{T}}\,.\tag{4.8}$$

When the boundary conditions are nontrivial, terms involving extensions of $\boldsymbol{u}^{\Gamma_u}\in\boldsymbol{H}^{1/2}(\Gamma_u)$ and $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}\in\boldsymbol{H}^{-1/2}(\Gamma_\sigma)$ to $\boldsymbol{H}^1(\Omega)$ and $\boldsymbol{H}^{-1/2}(\partial\mathcal{T})$ respectively, become part of $\ell_{\mathcal{T}}$.

Theorem A.1 implies that $\hat{b}|_{\hat{\mathcal{U}}\times\mathcal{V}_0}=0$, and that $b_{\mathcal{T}}|_{\mathcal{U}\times\mathcal{V}_0}$ and $\ell_{\mathcal{T}}|_{\mathcal{V}_0}$ are effectively the forms of the classical primal formulation (since $\hat{\mathcal{U}}$ ceases to play a role). Moreover, this fact and Theorem A.3 yield the well-posedness of the broken primal formulation via a straightforward application of Theorem 2.1, provided the classical primal formulation is well-posed. Thus, under the assumption of $|\bar{\boldsymbol{\varepsilon}}:\mathfrak{Re}(\mathsf{C}^*):\boldsymbol{\varepsilon}|>0$ for all symmetric second-order tensors $\boldsymbol{\varepsilon}\neq0$, the broken primal formulation is well-posed for most values of ω as established by Theorem 4.1.

Theorem 4.2. *Let $\mathcal{U}=\boldsymbol{H}_{\Gamma_u}^1(\Omega)$, $\hat{\mathcal{U}}=\boldsymbol{H}_{\Gamma_\sigma}^{-1/2}(\partial\mathcal{T})$, $\mathcal{U}_{\mathcal{T}}=\mathcal{U}\times\hat{\mathcal{U}}$, $\mathcal{V}_{\mathcal{T}}=\boldsymbol{H}^1(\mathcal{T})$, and consider the problem in* (4.8)*, with $b_{\mathcal{T}}$ defined in* (4.7) *in terms of b_0 and $\hat{b}$. Then,* (4.8) *is well-posed if and only if the problem in* (4.5) *is well-posed. In case of being well-posed, given any $\ell_{\mathcal{T}}\in\mathcal{V}_{\mathcal{T}}'$, there exists a unique solution $(\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}})\in\mathcal{U}_{\mathcal{T}}$ solving $b_{\mathcal{T}}\big((\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}),\boldsymbol{v}\big)=\ell_{\mathcal{T}}(\boldsymbol{v})$ for all $\boldsymbol{v}\in\mathcal{V}_{\mathcal{T}}$ which satisfies $\|(\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}})\|_{\mathcal{U}_{\mathcal{T}}}\leq C\|\ell_{\mathcal{T}}\|_{\mathcal{V}_{\mathcal{T}}'}$ for a $C>0$ independent of the choice of $\ell_{\mathcal{T}}$ and mesh $\mathcal{T}$.*

4.3 Discretization and convergence analysis

The broken variational formulation was discretized as described throughout Chapter 2. The choice of spaces was precisely that proposed in Section 2.7, because all the spaces are SdR spaces (see Section A.5 in Appendix A for definitions), so there exist compatible SdR discretizations of order p for the trial spaces and SdR discretizations of order $p + \Delta p$ for the enriched test spaces. These discretizations exist for all conventional element shapes [114]. More explicitly, the trial and test spaces are,

$$\begin{aligned} \mathcal{U}_h &= \big\{(\boldsymbol{\phi},\hat{\boldsymbol{\tau}}_{\mathbf{n}})\in\mathcal{U}=\boldsymbol{H}^1_{\Gamma_u}(\Omega)\times\boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T}) \;\big|\; \boldsymbol{\phi}|_K\in\big(W^p(K)\big)^3,\ (\hat{\boldsymbol{\tau}}_{\mathbf{n}})_K\in\big(\mathrm{tr}^K_{\mathrm{div}}(\boldsymbol{V}^p(K))\big)^3\big\}\,,\\ \mathcal{V}_r &= \big\{\boldsymbol{w}\mid \boldsymbol{w}|_K\in\big(W^{p+\Delta p}(K)\big)^3\big\}\subseteq\mathcal{V}=\boldsymbol{H}^1(\mathcal{T})\,, \end{aligned} \tag{4.9}$$

where the spaces $W^p(K)$ and $\boldsymbol{V}^p(K)$ come from a local SdR discretization of a particular element.

Next, assume $\mathcal{T}$ is a shape-regular tetrahedral mesh. Then, the discrete trial and enriched test spaces in (4.9) are explicitly

$$W^p(K)=\mathcal{P}^p\,,\qquad \boldsymbol{V}^p(K)=\mathcal{RT}^p=(\mathcal{P}^{p-1})^3+\boldsymbol{x}\mathcal{P}^{p-1}\,, \tag{4.10}$$

for every $K\in\mathcal{T}$. These spaces come from the classical Nédélec sequence of the first type, where $\mathcal{RT}^p$ is called the Raviart-Thomas space of order p. Then, define a Fortin operator $\Pi_F : H^1(\mathcal{T})\to\mathcal{V}_r$ by $(\Pi_F\boldsymbol{v})|_K=\Pi^{p,\Delta p}_{F,\mathrm{grad},K}\boldsymbol{v}|_K$ for $K\in\mathcal{T}$ with the local Fortin operator coming from Theorem 2.3. Take $(\boldsymbol{\phi},\hat{\boldsymbol{\tau}}_{\mathbf{n}})\in\mathcal{U}_h$, and assume C^* is piecewise constant across the mesh. Hence, for every tetrahedral element $K\in\mathcal{T}$,

$$\begin{aligned} \boldsymbol{\phi}|_K &\in\big(\mathcal{P}^p\big)^3\subseteq\big(\mathcal{P}^{p+\Delta p-4}\big)^3\,,\\ \mathsf{C}^*:\boldsymbol{\nabla}(\boldsymbol{\phi}|_K) &\in\big(\mathcal{P}^{p-1}\big)^{3\times3}\subseteq\big(\mathcal{P}^{p+\Delta p-5}\big)^{3\times3}\subseteq\big(\mathcal{P}^{p+\Delta p-3}\big)^{3\times3}\,,\\ (\hat{\boldsymbol{\tau}}_{\mathbf{n}})_K &\in\big(\mathrm{tr}^K_{\mathrm{div}}(\boldsymbol{V}^p(K))\big)^3\subseteq\big(\mathrm{tr}^K_{\mathrm{div}}\big(\boldsymbol{V}^{p+\Delta p-4}(K)\big)\big)^3\subseteq\big(\mathrm{tr}^K_{\mathrm{div}}\big(\boldsymbol{V}^{p+\Delta p-2}(K)\big)\big)^3\,, \end{aligned} \tag{4.11}$$

for $\Delta p\geq 4$. Therefore, using (2.70), (2.72) and (2.78) it follows

$$\begin{aligned} b_{\mathcal{T}}\big((\boldsymbol{\phi},\hat{\boldsymbol{\tau}}_{\mathbf{n}}),\boldsymbol{v}-\Pi_F\boldsymbol{v}\big) &= -\omega^2\rho\big(\boldsymbol{\phi},\boldsymbol{v}-\Pi_F\boldsymbol{v}\big)_{\mathcal{T}}+\big(\mathsf{C}^*:\boldsymbol{\nabla}\boldsymbol{\phi},\boldsymbol{\nabla}(\boldsymbol{v}-\Pi_F\boldsymbol{v})\big)_{\mathcal{T}}\\ &\qquad -\big\langle\hat{\boldsymbol{\tau}}_{\mathbf{n}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}(\boldsymbol{v}-\Pi_F\boldsymbol{v})\big\rangle_{\partial\mathcal{T}}=0\quad\forall\boldsymbol{v}\in\boldsymbol{H}^1(\mathcal{T})\,. \end{aligned} \tag{4.12}$$

This means that as long as $\Delta p\geq 4$ and $p\in\mathbb{N}$, an existence of a Fortin operator is guaranteed.

Remark 4.6. Due to the nature of the equations, and more specifically to the dynamic term $(\boldsymbol{u},\boldsymbol{v})_{\mathcal{T}}$ in (4.7), the requirement of $\Delta p \geq 4$ is more stringent than that proved for linear elasticity and Poisson's equation, which is $\Delta p \geq 3$ for ultraweak formulations [133] and even $\Delta p \geq 2$ for primal formulations (see Section 2.8.1).

From Theorem 2.1, it is known that the underlying exact solution $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ is the same, regardless of the mesh being considered, and it is easily observed that $\hat{\boldsymbol{\sigma}}_{\mathbf{n}} = \mathbf{tr}^{\mathcal{T}}_{\mathrm{div}}(\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u})$ for every mesh $\mathcal{T}$. Therefore, all the exact solutions $(\boldsymbol{u}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}})$ are attached to the mesh-independent element $\mathfrak{u}_\Omega = (\boldsymbol{u}, \mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}) \in \boldsymbol{H}^1_{\Gamma_u}(\Omega) \times \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)$ through the corresponding family of meshes (see Definition A.4 in Appendix A). With all these facts, the next theorem holds.

Theorem 4.3. *Let $\Omega \subseteq \mathbb{R}^3$ be a polyhedral domain and $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ be family of polyhedral meshes of Ω comprised of shape-regular tetrahedral elements $K \in \mathcal{T}_{\mathfrak{h}}$. For every $\mathcal{T}_{\mathfrak{h}}$, consider the DPG discretization of the variational formulation in* (4.8)*, which has discrete trial and enriched test spaces explicitly written in* (4.9)*, with $\Delta p \geq 4$. Let $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ be the mesh-independent element exact displacement solution of all the formulations, and assume $(\boldsymbol{u}, \mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}) \in \boldsymbol{H}^{1+s}(\Omega) \times \mathsf{H}^s(\mathbf{div},\Omega)$ for some $s > \frac{1}{2}$. Then, if $(\boldsymbol{u}_h, \hat{\boldsymbol{\sigma}}_{\mathbf{n},h})$ is the discrete solution computed at every mesh, and $(\boldsymbol{u}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}})$ is the exact solution, it follows*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{\boldsymbol{H}^1(\Omega)} + \|\hat{\boldsymbol{\sigma}}_{\mathbf{n}} - \hat{\boldsymbol{\sigma}}_{\mathbf{n},h}\|^2_{\boldsymbol{H}^{-1/2}(\partial\mathcal{T}_{\mathfrak{h}})} \leq C^2 h_{\mathfrak{h}}^{2\min\{s,p\}} \left(\|\boldsymbol{u}\|^2_{\boldsymbol{H}^{1+s}(\Omega)} + \|\mathsf{C}^* : \boldsymbol{\nabla}\boldsymbol{u}\|^2_{\mathsf{H}^s(\mathbf{div},\Omega)}\right), \quad (4.13)$$

where $C = C(s,p) > 0$ and $h_{\mathfrak{h}} = \max_{K\in\mathcal{T}_{\mathfrak{h}}} \mathrm{diam}(K)$. Additionally, if there exists a continuous Fortin operator from the test space to the enriched test space with a continuity bound independent of p, then $C = C_s(\ln p)^2 p^{-s}$ with $C_s = C(s)$ being independent of p.

In practice, the result is also expected to hold for other element shapes and for lower values of Δp. The construction of a Fortin operator with a p-independent continuity bound is still an open question at the moment.

4.4 Numerical results

First, verification studies confirming the convergence theory were done in a cube. Then, a validation study was completed using results from dynamic mechanical analysis (DMA) calibration experiments on different viscoelastic polymers. The in-house software `hp3d` was used for all computations and `MUMPS 5.0.1` was utilized as the solver.

4.4.1 Code verification

To verify the convergence results, a cube, $\Omega = (0,1)^3$, was discretized initially with five tetrahedra. A manufactured smooth solution for the displacement, $u_i(x) = \prod_{k=1}^{3} \sin(\pi x_k)$ for $i = 1,2,3$, was utilized to determine the stress, force and boundary data, where the dynamic stiffness tensor, C^*, was defined by $\lambda^* = \mu^* = 1 + \mathrm{i}$. The results are shown in Figure 4.1, where $\Delta p = 1$ in all cases.

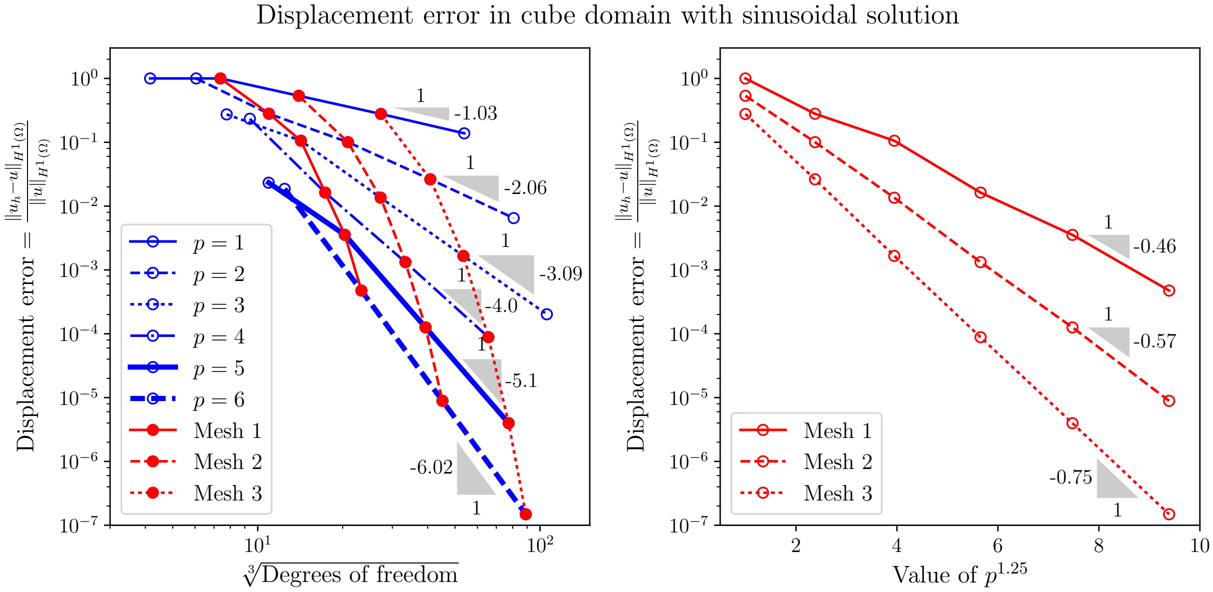


Figure 4.1: Relative displacement error in the $\boldsymbol{H}^1(\Omega)$ norm. Uniform h-refinements yield expected h^p convergence rates for $1 \leq p \leq 6$. Moreover, p-refinements of the same mesh do exhibit exponential convergence of the form $\exp(-b\,p^{1.25})$ with $b > 0$ depending on the mesh (the finer the mesh, the higher the b).

Clearly, uniform mesh refinements confirm the h-convergence theoretical estimates in (4.13), since the rate of convergence is of the type h^p due to the analyticity of the solution (so $s = \infty$). When $p \geq s$, where s is the regularity of the solution, Theorem 4.3 establishes an asymptotic hp-convergence estimate of the form $\|\mathsf{u} - \mathsf{u}_h\|_U \leq C_s(\ln p)^2 \big(\frac{h}{p}\big)^s$, where C_s is independent of h and p.

This is a quasi-algebraic form of convergence. However, when the solution is analytic, this estimate is expected to improve in some sense, but the explicit form cannot be deduced from the previous hp-estimate, since C_s depends on s and its behavior is unknown as $s \to \infty$. Figure 4.1 seems to indicate an exponential p-convergence estimate of the form $\|\mathfrak{u} - \mathfrak{u}_h\|_{\mathfrak{U}} \leq C \exp(-b\, p^{1.25})$, where C and b are independent of p (but not of h) and where $b > 0$ is larger if the mesh is finer. This result can be compared with exponential convergence results found in the literature [151, 152, 209] (it is also better than related hp-exponential rates in geometric meshes [13, 138, 139, 207, 206]).

It should be noted that $\Delta p = 1$ was used in the computations, but numerical experiments were done with higher values of Δp as well (including $\Delta p = 4$), and the resulting data points were almost exactly the same. Thus, for this particular equation it seems $\Delta p = 1$ is preferable, since the results are the same and the local computational cost is much lower. However, this merits further theoretical study to be certain, perhaps by finding a Fortin operator which is valid for $\Delta p \geq 1$. Having said that, there are equations and solution schemes where higher values of Δp provide advantages (see [86]), so this possibility should not be discarded either.

4.4.2 Validation of DMA experiments

Characterization of viscoelastic material properties in the frequency domain is done through dynamic mechanical analysis (DMA) experiments, where the material is subjected to oscillations. More precisely, to find the dynamic Young's modulus, E^*, a clamped material sample at a given temperature is made to vibrate at a particular amplitude and frequency. Thus, the temperature, vibration amplitude and frequency are controlled by the experimenter. A certain force is then measured in the experiment (the dependent variable), and using the appropriate beam theory one can find an inverse model for E^*. Experiments were done at the J. J. Pickle Research Campus of the University of Texas at Austin using the Q800 DMA instrument from TA Instruments. The experimental setup purposefully resembles cantilever beams. Indeed, Figure 4.2 shows a material sample in cantilever, clamped at both ends, where one clamp is static while the other clamp is free to move and vibrate at a given amplitude and frequency. It is at this moving clamp that the force is measured by the instrument.

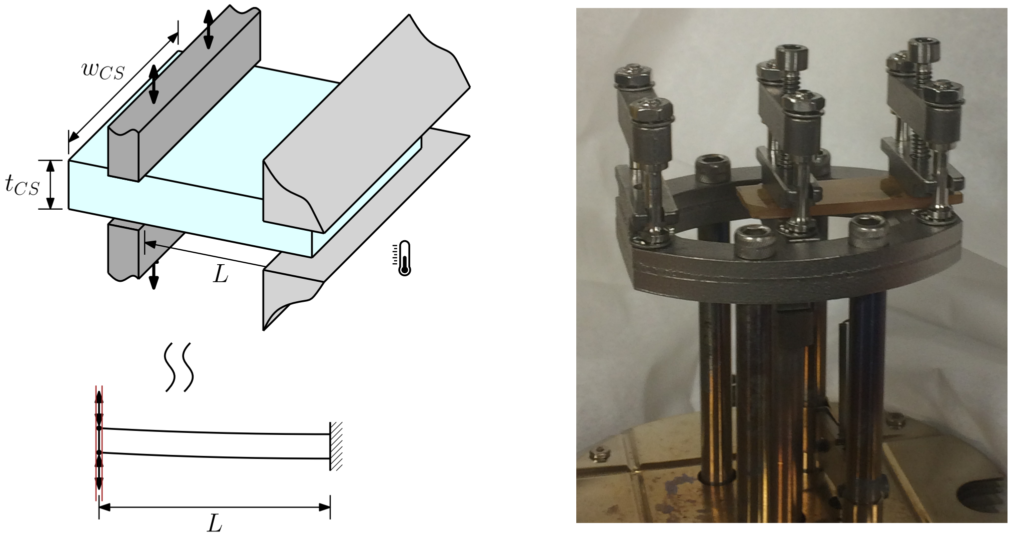


Figure 4.2: The single cantilever DMA experimental setup. The external clamp is statically fixed, while the central clamp, where a force is measured, moves vertically with a known amplitude and frequency. This whole setup lies inside a closed oven that carefully controls the temperature.

An inverse model for E^* can be derived using Timoshenko beam theory. Consider a static linear elastic beam clamped at one end and with a point force applied at the other end, where additionally the cross-section remains parallel to the force (see Figure 4.2). This last condition represents the moving clamp where the force is being measured. Hence, this is *not* a typical cantilever beam (where one of the ends is free), but for simplicity it is still referred as such. Using Timoshenko beam theory [222, 223], the vertical displacement can be determined using the zero-angle boundary conditions at *both* ends and a zero-displacement in the clamped end. The resulting maximum displacement occurs where the force is applied and takes the value,

$$u_{\max} = \frac{FL^3}{12EI} + \frac{FL}{\kappa A_{CS} G} = \frac{FL^3}{12EI}\Big(1 + \frac{2}{\kappa}(1+\nu)\Big(\frac{t_{CS}}{L}\Big)^2\Big)\,, \tag{4.14}$$

where $u_{\max}$ is the maximum vertical displacement of the beam, F is the force applied, L is the length between the clamped end and where the force is applied; E and $G = \frac{E}{2(1+\nu)}$ are the Young's and shear moduli of the linear elastic material while ν is its Poisson's ratio; $A_{CS} = w_{CS} t_{CS}$, w_{CS} and t_{CS} are the cross-sectional area, width and thickness respectively; $I = \frac{w_{CS} t_{CS}^3}{12}$ is the second moment of area of the rectangular cross-section, and κ is the Timoshenko shear coefficient. This equation obeys a correspondence principle with the time-harmonic equations of linear viscoelasticity [164], so that an inverse model of the form,

$$\begin{gathered}E^* = \frac{1}{\alpha_c}\frac{F^*_{exp}}{u^*_{\max}}\frac{L^3}{\beta_c I}\Big(1 + \frac{12}{5}(1+\nu^*)\Big(\frac{t_{CS}}{L}\Big)^2\Big)\,,\\ \alpha_c = 0.7616 - 0.02713\sqrt{\tfrac{L}{t_{CS}}} + 0.1083\ln\big(\tfrac{L}{t_{CS}}\big)\,,\end{gathered} \tag{4.15}$$

is utilized, where $\beta_c = 12$ in this single cantilever setting, and α_c is a correction factor accounting for 3D clamping effects, which is given by the manufacturer. For a rectangular cross-section, the Timoshenko shear coefficient is taken from the literature as $\kappa = \frac{5}{6}$ [156]. Here, E^* is the dynamic Young's modulus, and note that both the experimental force and vibration amplitude, F^*_{exp} and $u^*_{\max}$, are now complex numbers. Note that $\frac{F^*_{exp}}{u^*_{\max}} = \big|\frac{F^*_{exp}}{u^*_{\max}}\big|e^{\mathrm{i}\delta_{ph}}$, where δ_{ph} is an angle that represents the phase change between the oscillations of the force and the driving mechanical vibrations of the displacement. The values of temperature, vibration frequency, $|u^*_{\max}|$, $|F^*_{exp}|\cos(\delta_{ph})$ and $\tan(\delta_{ph})$ are reported by the instrument. The distance L here is the distance between the clamps themselves, *not* the distance between the midpoints of the clamps. The only limitation with this inverse model is that it assumes that the dynamic Poisson's ratio, ν^*, is known. The ideal scenario is that either ν^* or the dynamic shear modulus, G^*, are known from a separate preceding experiment. In the latter case, where G^* is known, note that $\nu* = \frac{E^*}{2G^*} - 1$, so an analogous expression for E^* only in terms of G^* can easily be derived from (4.15). If neither ν^* nor G^* are experimentally known, it is usually assumed that G^* has the same phase as E^*, so that ν^* is real-valued, and then an educated guess is made for $\nu^* \in \mathbb{R}$.

There is a second experimental setup which involves the same instrument, but with the beam arranged in a double cantilever, with two external static clamps at both ends and a middle moving clamp. This can be seen in Figure 4.3. The inverse model is actually the same as that given in (4.15), but with $\beta_c = 24$ in the double cantilever setting, and where the distance L is the same as in the single cantilever case since it represents the distance between the edge of the external clamp and closest edge of the middle clamp (as seen in Figure 4.3). Hence, the actual distance over which the material is being deformed is $L_d = 2L$.

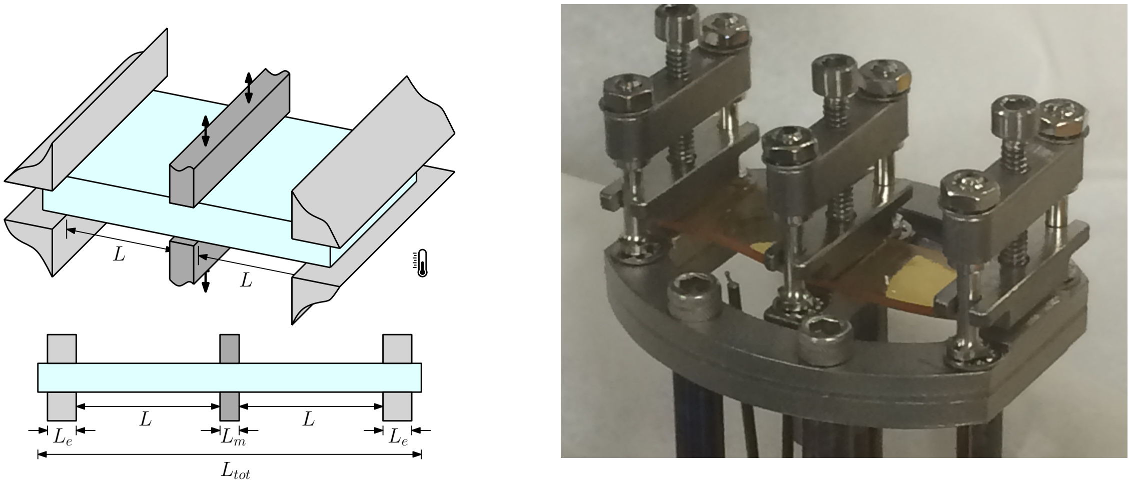


Figure 4.3: The double cantilever DMA experimental setup. The two external clamps are statically fixed, while a force is measured at the central clamp which moves at a controlled amplitude and frequency.

For the sake of brevity, results of only one example of each setup will be shown here. In the single cantilever case, silicone at 30.0 °C was tested at 4 Hz with a controlled amplitude of vibration of $|u^*_{\max}| = 15\,\mu\text{m}$, where the relevant part of the sample measured $L = 17.5\,\text{mm}$, $w_{CS} = 11.8\,\text{mm}$ and $t_{CS} = 1.63\,\text{mm}$. The measured force from the experiment was $|F^*_{exp}|\cos(\delta_{ph}) = 0.1064\,\text{N}$ with $\tan(\delta_{ph}) = 0.0384$. In the double cantilever case, epoxy at 22.4 °C was tested at 40 Hz with an amplitude of vibration of $|u^*_{\max}| = 15\,\mu\text{m}$, where the relevant part of the sample had dimensions of $L_d = 2L = 35.0\,\text{mm}$, $w_{CS} = 13.2\,\text{mm}$ and $t_{CS} = 2.05\,\text{mm}$. The measured force from the experiment was $|F^*_{exp}|\cos(\delta_{ph}) = 0.7248\,\text{N}$ with $\tan(\delta_{ph}) = 0.00869$. In both experiments it was assumed that $\nu^* = 0.33$ (see [128], but higher values are also found in [210]), so using the inverse model in (4.15) with $\beta_c = 12$ and $\beta_c = 24$ respectively, it was possible to calculate E^*.

Next, the dynamic stiffness tensor, C^*, was computed using the values of E^* and ν^*, and the experiments were then simulated computationally. Here, it is important to mention that the the middle clamp measures $L_m = 6.35\,\text{mm}$, while the two external clamps measure $L_e = 7.625\,\text{mm}$ each, as observed from Figure 4.3. Thus the samples themselves (both in the experiment and the simulated geometry) are typically longer than $L_e + L + L_m = 31.475\,\text{mm}$ in the single cantilever case and $2L_e + 2L + L_m = 56.6\,\text{mm}$ in the double cantilever case. The samples used for the numerical results were 40 mm for the single cantilever and $L_{tot} = 60\,\text{mm}$ for the double cantilever.

The densities of the silicone and epoxy resins were assumed to be $1134\,\mathrm{kg{\cdot}m^{-3}}$ and $1250\,\mathrm{kg{\cdot}m^{-3}}$ respectively. The force, which is the quantity of interest, was calculated a posteriori by integrating the vertical traction, $(\hat{\sigma}_{\mathbf{n},h})_3$, over the area where the moving clamp made contact with the sample. The numerically computed force, F_h^*, was then compared with the actual measured force from the experiment, F_{exp}^*. The results for different values of p and with $\Delta p = 1$ are shown in Figure 4.4.

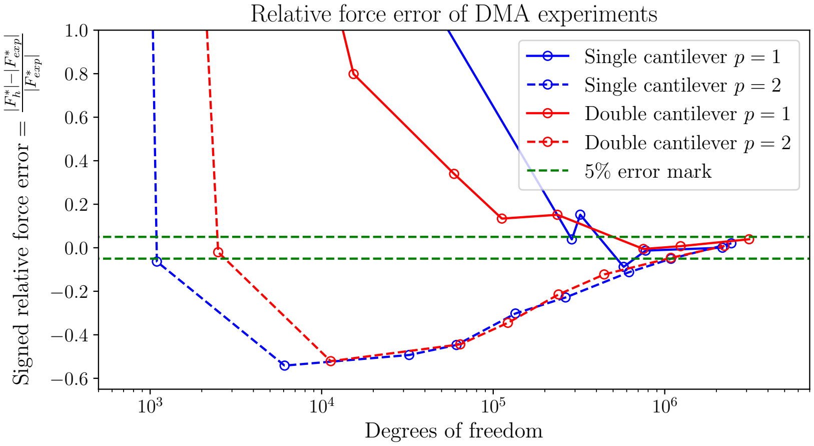


Figure 4.4: Convergence of the magnitude of the computed force, F_h^*, to the real experimental value measured from DMA experiments on different setups, F_{exp}^*. The single cantilever results correspond to a silicone sample, while those of the double cantilever correspond to an epoxy sample.

The magnitude of the force appears to converge to within 5% of the experimental value with both the single and double cantilever setups. This is as good as one can hope for from the validation point of view, and it confirms that the equations do indeed model the actual physical behavior observed experimentally. These results seem to suggest that the value of $p = 2$ does not offer a significant advantage over $p = 1$ to obtain the desired outcome, but further research on this matter might be necessary, as a different quantity of interest might produce very different results. With respect to the phase error in $\tan(\delta_{ph})$, the simulations show virtually no error even from the first computation. This is probably due to the assumption that $\nu^* \in \mathbb{R}$ is real-valued, but perhaps a less trivial convergence behavior would be observed if this hypothesis were to be dropped.

The results in Figure 4.4 were obtained with adaptivity driven by the arbitrary-p residual-based a posteriori error estimator described in (2.50), which is innate to the DPG methodology.

Otherwise, it would have been prohibitively expensive to obtain the same results via uniform refinements. Indeed, from the physics of the problem, it is intuitive to notice that most of the stress will be concentrated in the areas close to where the clamps are holding the material. The computations confirm this, as can be observed from Figure 4.5, where it is clear that not only the stress is localized there, but that the adaptivity scheme is refining in precisely that area, which is where the force will be computed from. Thus, adaptivity is fundamental for this problem which has localized solution features, and this justifies to a degree the use of this DPG method.

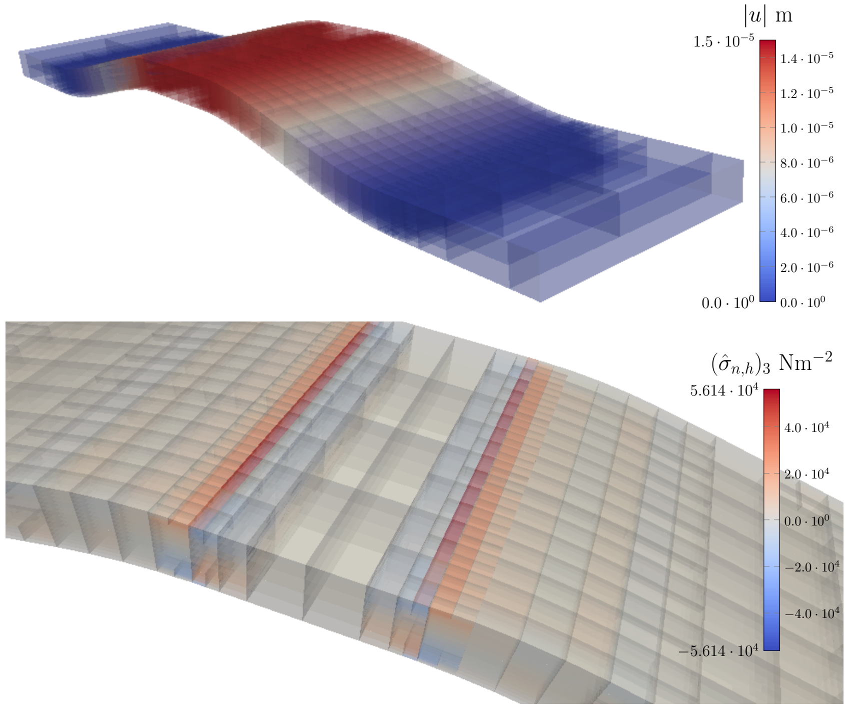


Figure 4.5: Numerical results with the double cantilever setup with $p = 1$ and after 4 isotropic adaptive refinements. The displacement is warped by a factor of 4000 for clarity. The vertical traction seems to be concentrated at the edges of the middle clamp, and adaptive refinements do seem to focus on that area.

4.5 Discussion

A DPG finite element method was implemented for the time-harmonic equations of linear viscoelasticity. The method discretizes a broken primal variational formulation of the equation, which was proved to be well-posed in the infinite-dimensional setting. As part of this proof, the well-posedness of the classical primal variational formulation of linear viscoelasticity was also rigorously established. Moreover, the numerical method itself was shown to be stable and convergent under certain conditions, and this included analyzing both h- and p-convergence estimates. A completely natural a posteriori error estimator for arbitrary-p which is used to drive adaptivity is also included as part of the method. The method was verified using a smooth manufactured solution, where the expected h-convergence rates of the form h^p where corroborated for various values of p. Moreover, the verification tests displayed exponential p-convergence estimates of the form $\exp(-b\,p^{1.25})$.

Additionally, DMA experiments to determine the dynamic Young's modulus, E^*, were performed on different materials and with distinct experimental setups: single and double cantilever. The computational results validated the calibration model to within 5% error of the quantity of interest. Moreover, the simulated stress was very concentrated on certain parts of the domain, so having a good adaptivity scheme was crucial to obtain the desired result. In this sense, the numerical DPG method was extremely convenient, since it already came with its own a posteriori error estimator.

Looking forward, more complicated validation studies could be tackled, where the quantities of interest may vary in nature. The built-in a posteriori error estimator is designed to drive down the residual, but may not be optimal in accelerating the convergence of a particular quantity of interest. In this sense, this could lead to investigating goal-driven adaptivity schemes within the context of the DPG methodology. When the linear system size is large, computations may become prohibitive, so it would be useful to make improvements to reduce the system size as much as possible and to support parallel computing within the solvers. Finally, for more interesting cases closer to the glass transition temperature of the materials in question, the results from the computations might improve if the actual value of the dynamic Poisson's ratio, ν^*, or the dynamic shear modulus, G^*, are used in the calibration inverse model, but this requires separate DMA experiments to be completed, which might be a future endeavor.

Chapter 5

Case study: resins in form-wound medium-voltage coils

This chapter aims to focus on a case study involving resins found in form-wound medium-voltage stator coils sitting inside large electric machinery. These constitute the (electric) insulation of the machinery, and the failure of such insulation is not yet well understood. This chapter is an attempt at a basic understanding the underlying mechanics of certain scenarios that could eventually lead to such failure. It is mostly based on modeling of an idealized geometry under distinct loading profiles occurring at different frequency regimes. Once again, the chapter is included in this dissertation, because it shows the benefits of the residual-based high-order a posteriori error estimator that is intrinsic to DPG methods.

5.1 Introduction

The insulation failure of form-wound medium-voltage stator coils inside electric machinery is still not well understood, even though it is an important problem in the context of electric machinery, as it demands that the machine be stopped and the insulation be replaced. Otherwise, the machine could fail completely. Most efforts have concentrated on the diagnosis of insulation failure (typically associated to the presence and formation of internal voids) through the detection of physical phenomena such as partial discharge [215, 214, 180], and these efforts have had moderate success. However, it would be useful to be able to detect insulation failure earlier, either to prevent it, or to make a replacement with sufficient time before a critical and relatively lengthy period of continuous operation of the machine is foreseen. Given the little knowledge about the problem in the literature, we will focus on a very simple “turn-to-turn” geometry and analyze three hypothetical mechanical scenarios, each associated to a different frequency regime and physical features. That will be the objective of this chapter.

5.2 Model geometry and preliminary assumptions

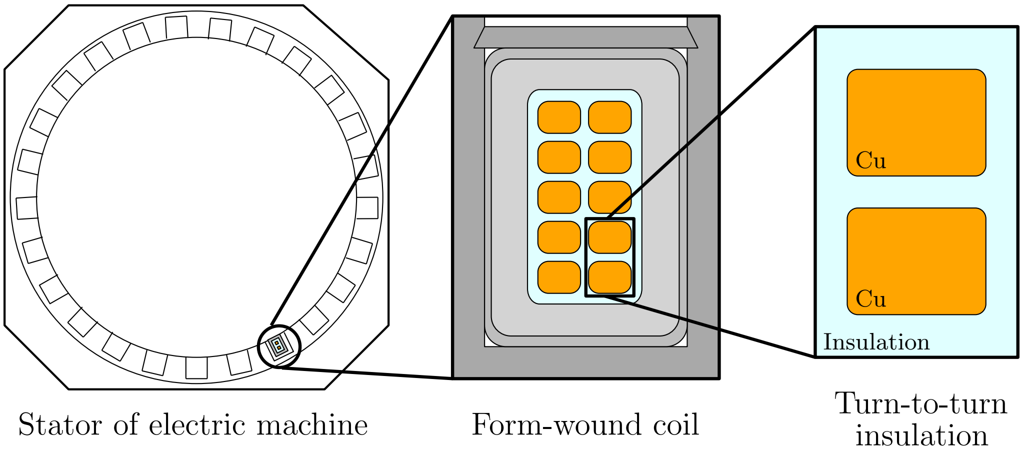
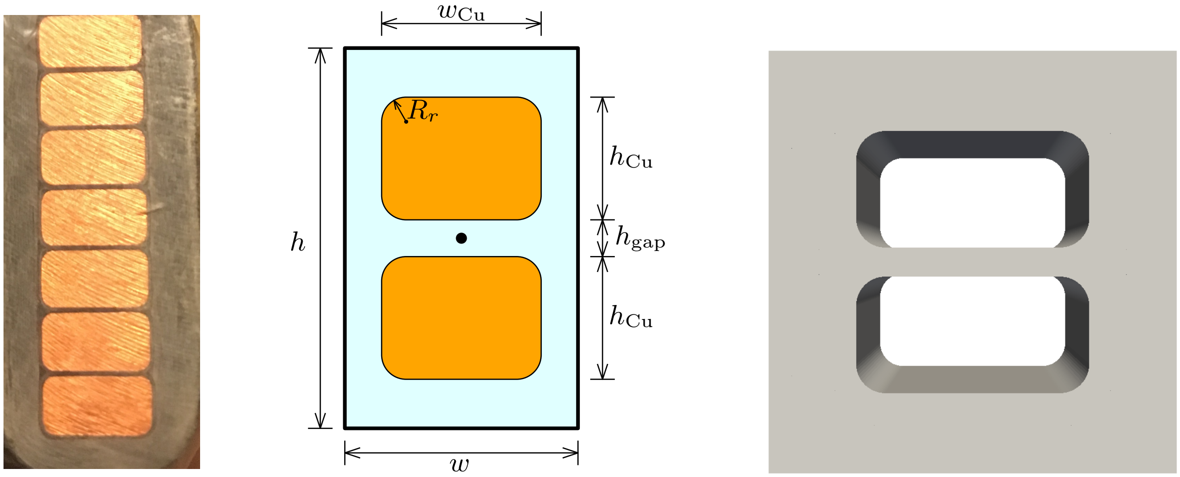


Figure 5.1: Schematic of turn-to-turn region within a form-wound medium-voltage coil sitting inside the stator of a large electric machine.

Form-wound stator coils are carefully organized copper square-like wires surrounded by a composite laminate material that acts as electrical insulation. In reality, the insulation is typically composed of mica or glass tape held together by a matrix of viscoelastic resin (see Figure 5.2). Thus, it is anisotropic and heterogeneous in nature. However, to simplify the physics, the insulation will be assumed to be simply the resin matrix, which will be taken to be a homogeneous isotropic viscoelastic material. Meanwhile, the coils will be made of copper. The interface between the materials will be assumed to have continuous displacement between the two materials. We will additionally focus on the simplest "turn-to-turn" region involving only two copper coils. This is depicted in Figure 5.1. Meanwhile, the parametric geometry adopted is illustrated in Figure 5.2, with the values of the parameters taken as $h = 14.5\,\mathrm{mm}$, $w = 14.0\,\mathrm{mm}$, $h_{\mathrm{Cu}} = 4\,\mathrm{mm}$, $w_{\mathrm{Cu}} = 8\,\mathrm{mm}$, $h_{\mathrm{gap}} = 1.0\,\mathrm{mm}$ and a rounding radius of $R_r = 1.0\,\mathrm{mm}$.

Power-dense electric machines are in a completely oscillatory environment due to their rotational nature, and as a result it makes sense to analyze their mechanics in the frequency domain. To simplify further, the underlying physics is assumed to be linear, so it is assumed that the displacements, temperature variations and their gradients all take very small values. Depending on the frequency regime, the physical phenomena that is predominant in the domain changes completely.



Idealized model geometry

Figure 5.2: Real form-wound stator coil, followed by blueprint of geometry, and actual geometry modeled. The parameters of the modeled geometry were $h = 14.5\,\mathrm{mm}$, $w = 14.0\,\mathrm{mm}$, $h_{\mathrm{Cu}} = 4\,\mathrm{mm}$, $w_{\mathrm{Cu}} = 8\,\mathrm{mm}$, $h_{\mathrm{gap}} = 1.0\,\mathrm{mm}$ and a rounding radius of $R_r = 1.0\,\mathrm{mm}$.

At very low frequencies temperature effects are important, at mid-range frequencies the overall kinematics of the machine are important, whereas at very high frequencies the electromagnetic interaction produces Lorentz forces that affect the system. These three scenarios will be analyzed in what follows. Additionally, two different insulating resins, epoxy and "silicone" (a silicone-based resin), will be compared throughout.

5.3 Low frequency: thermoviscoelasticity

5.3.1 Description and problem setup

Thermal oscillations occur at relatively low frequencies of about 0.05 Hz. These variations are due to Joule heating (i.e. Ohmic heating, resistive heating) in the copper coil and the counteracting effect of rapid cooling in the stator assuming the existence of forced cooling in the power-dense machine design. These produce a natural thermal expansion of all materials involved in the domain (the copper and the viscoelastic insulation). However, at these low frequencies, we expect the deformations of the stator and the supporting structure to be negligible. Thus, the material does not deform from the outside, yet it wants to expand inside, as illustrated in Figure 5.3. This produces stresses which might be interesting to investigate.

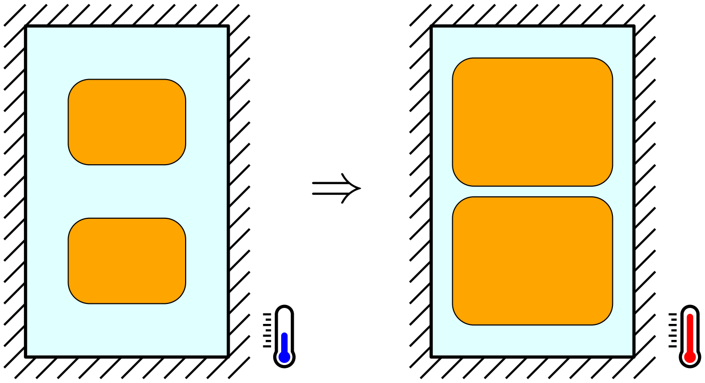


Figure 5.3: Simplified schematic of the effects of temperature in the displacement field (and possibly stress field) assuming no external deformation.

The interaction of thermal effects and viscoelastic solids (previously referred to as "materials with memory") in the time domain can be modeled using classic nonlinear continuum mechanics as done in [78]. To analyze the material in the frequency domain it is convenient to linearize the underlying equations, and this is valid provided the variations in the displacement gradient, temperature, temperature gradient and their time derivatives are small. This procedure is explained rigorously in Appendix C. Under the assumption of no residual stress at the linearization temperature and strain, and that the infinitesimal entropy production in each closed process is "invariant under time-reversal" (see [141]), the resulting linear thermoviscoelastic equations in the time domain are as follows,

$$\begin{cases} \rho\ddot{\boldsymbol{u}} = \mathbf{div}\,\boldsymbol{\sigma} + \boldsymbol{f}\,, \\ \rho\dot{c}_v \ast \dot{\vartheta} = -\operatorname{div}\boldsymbol{q} + \bar{\theta}\dot{\mathsf{M}} \mathbin{\underset{\cdot}{\overset{\cdot}{\ast}}} \dot{\boldsymbol{\varepsilon}} + r\,, \\ \boldsymbol{\sigma} = \dot{\mathsf{C}} \mathbin{\underset{\cdot}{\overset{\cdot}{\ast}}} \boldsymbol{\varepsilon} + \dot{\mathsf{M}} \ast \vartheta\,, \\ \boldsymbol{q} = -\boldsymbol{\kappa}\cdot\nabla(\vartheta + \bar{\theta})\,, \end{cases} \tag{5.1}$$

where $\boldsymbol{u}$ is the displacement, θ is the absolute temperature, $\vartheta = \theta - \bar{\theta}$ is the temperature difference with respect to an initial time-independent temperature distribution $\bar{\theta}$, $\boldsymbol{\sigma}$ is the Cauchy stress tensor, $\boldsymbol{q}$ is the heat flux, and $\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^{\mathsf{T}})$ is the engineering strain as a function of $\boldsymbol{u}$. The source terms in the first two equations are the body force per unit volume, $\boldsymbol{f}$, and the body heat generation per unit volume, r. Regarding the material properties, ρ is the density, C is the viscoelastic stiffness tensor, c_v is the viscoelastic specific heat capacity at constant volume, $\mathsf{M} = -\dot{\mathsf{C}} \mathbin{\ast\!\!\!:} \boldsymbol{\alpha}$ is the viscoelastic stress-temperature tensor, $\boldsymbol{\alpha}$ is the viscoelastic tensor of coefficients of linear thermal expansion, and $\boldsymbol{\kappa}$ is the heat conductivity tensor. The symbols $\ast$ and $\mathbin{\ast\!\!\!:}$ are convolutions in the time domain,

$$\dot{\mathsf{C}} \mathbin{\ast\!\!\!:} \boldsymbol{\varepsilon} = \int_{-\infty}^{\infty} \dot{\mathsf{C}}(s) : \boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s\,, \qquad \dot{\mathsf{M}} \ast \vartheta = \int_{-\infty}^{\infty} \dot{\mathsf{M}}(s)\vartheta(\cdot - s)\,\mathrm{d}s\,, \tag{5.2}$$

and so on. As such, note that with the exception of $\boldsymbol{\kappa}$ and ρ, all material properties are in fact time-dependent in the viscoelastic case, and the resulting system in (5.1) is a set of integro-differential equations (not merely differential equations).

The aforementioned assumptions imply that C has minor and major symmetries, and that $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are symmetric, so that $\mathsf{C}_{ijkl} = \mathsf{C}_{ijlk} = \mathsf{C}_{jikl} = \mathsf{C}_{klij}$, $\boldsymbol{\alpha}_{ij} = \boldsymbol{\alpha}_{ji}$ and $\boldsymbol{\kappa}_{ij} = \boldsymbol{\kappa}_{ji}$. The relaxation or equilibrium stiffness tensor is $\mathsf{C}^\infty = \lim_{t\to\infty} \mathsf{C}(t)$, and the simplest case is $\mathsf{C}(t) = \mathsf{C}^\infty H_0(t)$, where H_0 is the Heaviside step function, so that (distributionally) $\dot{\mathsf{C}} = \mathsf{C}^\infty \delta_0$ with δ_0 being the Dirac delta distribution. The same holds for the relaxation variables $c_v^\infty = \lim_{t\to\infty} c_v(t)$, $\mathsf{M}^\infty = \lim_{t\to\infty} \mathsf{M}(t)$ and $\boldsymbol{\alpha}^\infty = \lim_{t\to\infty} \boldsymbol{\alpha}(t)$, with the simplest cases being $c_v(t) = c_v^\infty H_0(t)$, $\mathsf{M}(t) = \mathsf{M}^\infty H_0(t)$ and $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}^\infty H_0(t)$ respectively. With the simplest expressions it follows that $\boldsymbol{\sigma} = \mathsf{C}^\infty : (\boldsymbol{\varepsilon} - \vartheta\boldsymbol{\alpha}^\infty)$ and $\rho c_v^\infty \dot{\vartheta} = -\operatorname{div} \boldsymbol{q} - \bar{\theta}\alpha^\infty : \mathsf{C}^\infty : \dot{\boldsymbol{\varepsilon}} + r$, which coupled with the remaining two equations are precisely the equations of linear thermoelasticity [53]. Lastly, in the isotropic case it follows that,

$$\mathsf{C}_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\,, \qquad \boldsymbol{\alpha} = \alpha\mathbf{I}\,, \qquad \boldsymbol{\kappa} = \kappa\mathbf{I}\,, \tag{5.3}$$

where λ and μ are the time-dependent Lamé parameters, α is the time-dependent coefficient of linear thermal expansion, and κ is the heat conductivity.

With these linear equations it is now possible to pass to the frequency domain. Let ω be an angular frequency, suppose $\bar{\theta}$ is constant (so $\nabla\bar{\theta} = 0$), and substitute $\boldsymbol{\sigma}$ and $\boldsymbol{q}$ into the other two expressions, so that the frequency-domain thermoviscoelastic equations written as a second-order system are,

$$\begin{cases} (\mathrm{i}\omega)^2\rho\boldsymbol{u} - \mathbf{div}\left(\mathsf{C}^* : \boldsymbol{\varepsilon} - \mathsf{C}^* : \boldsymbol{\alpha}^*\vartheta\right) = \boldsymbol{f}\,, \\ \rho(\mathrm{i}\omega)c_v^*\vartheta + (\mathrm{i}\omega)\bar{\theta}\boldsymbol{\alpha}^* : \mathsf{C}^* : \boldsymbol{\varepsilon} - \operatorname{div}\left(\boldsymbol{\kappa}\cdot\nabla\vartheta\right) = r\,, \end{cases} \tag{5.4}$$

where now the frequency-specific unknown variables $\boldsymbol{u}(\omega)$ and $\vartheta(\omega)$, and known variables $\boldsymbol{f}(\omega)$, $r(\omega)$, $\mathsf{C}^*(\omega)$, $c_v^*(\omega)$ and $\boldsymbol{\alpha}^*(\omega)$, are the Fourier transforms of the old time-dependent variables $\boldsymbol{u}(t)$, $\vartheta(t)$, $\boldsymbol{f}(t)$, $r(t)$, $\dot{\mathsf{C}}(t)$, $\dot{c}_v(t)$ and $\dot{\boldsymbol{\alpha}}(t)$ respectively. Here, $\mathsf{C}^*(\omega) = \mathrm{i}\omega\mathsf{C}(\omega)$ is called the dynamic stiffness tensor, $c_v^*(\omega) = \mathrm{i}\omega c_v(\omega)$ is the dynamic specific heat capacity, and $\boldsymbol{\alpha}^*(\omega) = \mathrm{i}\omega\boldsymbol{\alpha}(\omega)$ is the dynamic tensor of coefficients of linear thermal expansion, where $\mathsf{C}(\omega)$, $c_v(\omega)$ and $\boldsymbol{\alpha}(\omega)$ are the Fourier transforms of $\mathsf{C}(t)$, $c_v(t)$ and $\boldsymbol{\alpha}(t)$ respectively. In the case of linear thermoelasticity it will follow that $\mathsf{C}^* = \mathsf{C}^\infty$, $c_v^* = c_v^\infty$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^\infty$. In the isotropic case, it follows analogously that $\lambda^*(\omega) = \mathrm{i}\omega\lambda(\omega)$, $\mu^*(\omega) = \mathrm{i}\omega\mu(\omega)$ and $\alpha^*(\omega) = \mathrm{i}\omega\alpha(\omega)$, are the Fourier transforms of $\dot{\lambda}(t)$, $\dot{\mu}(t)$ and $\dot{\alpha}(t)$ respectively, and $\mathsf{C}^* = \lambda^*\delta_{ij}\delta_{kl} + \mu^*(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$ and $\boldsymbol{\alpha}^* = \alpha^*\mathbf{I}$. Moreover, $G^* = \mu^*$ is the dynamic shear modulus, $K^* = \lambda^* + \frac{2}{3}\mu^*$ is the dynamic bulk modulus, $E^* = \frac{\mu^*(3\lambda^*+2\mu^*)}{\lambda^*+\mu^*}$ is the dynamic Young's modulus, and $\nu^* = \frac{\lambda^*}{2(\lambda^*+\mu^*)}$ is the dynamic Poisson's ratio. Notably, E^* is a nonlinear function of λ^* and μ^*, implying that in general it is not the Fourier transform of $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\mu(t)(3\lambda(t)+2\mu(t))}{\lambda(t)+\mu(t)}$. A similar assertion follows for ν^*. Thus, one should be careful when speaking of the time-dependent Young's modulus and Poisson's ratio in three-dimensional thermoviscoelasticity as different definitions exist in the literature [164, §5.7]. Finally, note that the convolutions became products, so the equations returned to being differential equations for each fixed frequency ω (as opposed to integro-differential equations), and note that $\boldsymbol{u}(\omega)$, $\vartheta(\omega)$, $\boldsymbol{f}(\omega)$, $r(\omega)$, $\mathsf{C}^*(\omega)$, $c_v^*(\omega)$ and $\boldsymbol{\alpha}^*(\omega)$ are now complex-valued functions (as opposed to real-valued).

Remark 5.1. If $g(t)$ is a function in the time domain, the definition of its Fourier transform, $g(\omega)$ is in general $g(\omega) = \int_{-\infty}^{\infty} g(t)e^{-\mathrm{i}\omega t}\,\mathrm{d}t$, and this definition applies to $\boldsymbol{u}(\omega)$, $\vartheta(\omega)$, $\boldsymbol{f}(\omega)$, $r(\omega)$, $\mathsf{C}^*(\omega)$, $c_v^*(\omega)$ and $\boldsymbol{\alpha}^*(\omega)$. A usual ansatz for $\boldsymbol{u}(\omega)$, $\vartheta(\omega)$, $\boldsymbol{f}(\omega)$ and $r(\omega)$ (not the viscoelastic material properties) is $g(t) = A\cos(\omega_0 t + \psi_{\mathrm{ph}})$, so $g(\omega) = Ae^{\mathrm{i}\omega\cdot(\psi_{\mathrm{ph}}/\omega_0)}\frac{1}{2}\big(\delta_0(\omega-\omega_0) + \delta_0(\omega+\omega_0)\big)$. Typically $\frac{1}{2}\big(\delta_0(\omega-\omega_0) + \delta_0(\omega+\omega_0)\big)$ is factored out of the expressions for $\boldsymbol{u}(\omega)$, $\vartheta(\omega)$, $\boldsymbol{f}(\omega)$ and $r(\omega)$ and cancelled in view of the linearity of (5.4). Thus, under this simple ansatz, instead of writing $g(\omega_0)$, one simply writes $g = A\big(\cos(\psi_{\mathrm{ph}}) + \mathrm{i}\sin(\psi_{\mathrm{ph}})\big)$ and this is understood from the context. For example, $g(t) = A\sin(\omega_0 t)$ would be written as $g = -\mathrm{i}A$ in the frequency domain.

Multiplying the two equations in (5.4) by test functions $\boldsymbol{u}$ and χ and integrating by parts over a Lipschitz domain $\Omega \subseteq \mathbb{R}^3$ yields the equations,

$$\begin{aligned}
(\mathrm{i}\omega)^2\big(\rho\boldsymbol{u},\boldsymbol{v}\big)_\Omega+\big(\mathsf{C}^*\!:\!(\boldsymbol{\varepsilon}-\boldsymbol{\alpha}^*\vartheta),\boldsymbol{\nabla v}\big)_\Omega&=\big(\boldsymbol{f},\boldsymbol{v}\big)_\Omega+\big\langle\mathsf{C}^*\!:\!(\boldsymbol{\varepsilon}-\boldsymbol{\alpha}^*\vartheta)\cdot\hat{\mathbf{n}},\boldsymbol{v}\big\rangle_{\partial\Omega}\,,\\
(\mathrm{i}\omega)\big(\rho c_v^*\vartheta,\chi\big)_\Omega+(\mathrm{i}\omega)\big(\bar{\theta}\boldsymbol{\alpha}^*\!:\!\mathsf{C}^*\!:\!\boldsymbol{\varepsilon},\chi\big)_\Omega+\big(\boldsymbol{\kappa}\cdot\nabla\vartheta,\nabla\chi\big)_\Omega&=\big(r,\chi\big)_\Omega+\big\langle\boldsymbol{\kappa}\cdot\nabla\vartheta\cdot\hat{\mathbf{n}},\chi\big\rangle_{\partial\Omega}\,,
\end{aligned}\tag{5.5}$$

where $\hat{\mathbf{n}}$ is the outward normal to $\partial\Omega$. Here, $(\cdot,\cdot)_\Omega$ is the sesquilinear complex-valued inner product in L^2 (see (2.3)), while $\langle\cdot,\cdot\rangle_{\partial\Omega}$ is a sesquilinear inner product that for now can be interpreted as a boundary integral (for smooth enough inputs). At the moment there has been no mention of boundary conditions. With this in mind, consider two different partitions of $\partial\Omega$, $\{\Gamma_u,\Gamma_\sigma\}$ and $\{\Gamma_\theta,\Gamma_q\}$, such that they are relatively open in $\partial\Omega$ and satisfy $\overline{\Gamma_u\cup\Gamma_\sigma}=\overline{\Gamma_\theta\cup\Gamma_q}=\partial\Omega$ and $\Gamma_u\cap\Gamma_\sigma=\Gamma_\theta\cap\Gamma_q=\varnothing$. In this way, there are displacement boundary conditions, $\boldsymbol{u}=\boldsymbol{u}^{\Gamma_u}$, over Γ_u, normal stress boundary conditions, $\boldsymbol{\sigma}\cdot\hat{\mathbf{n}}=\mathsf{C}^*\!:\!(\boldsymbol{\varepsilon}-\boldsymbol{\alpha}^*\vartheta)\cdot\hat{\mathbf{n}}=\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}$, over Γ_σ, temperature boundary conditions, $\theta=\theta^{\Gamma_\theta}$, over Γ_θ, and surface heat flow boundary conditions, $\boldsymbol{q}\cdot\hat{\mathbf{n}}=-\boldsymbol{\kappa}\cdot\nabla\vartheta\cdot\hat{\mathbf{n}}=q_{\mathbf{n}}^{\Gamma_q}$, over Γ_q. The Hilbert spaces where these variables lie are important and they can be deduced from the original system of equations in (5.4) and the variational equations above (see Section 3.2.1 for the definitions). Indeed, $\boldsymbol{u}\in\widetilde{\boldsymbol{u}}^{\Gamma_u}+\boldsymbol{H}^1_{\Gamma_u}(\Omega)$, where $\widetilde{\boldsymbol{u}}^{\Gamma_u}$ is an extension of the boundary condition $\boldsymbol{u}^{\Gamma_u}\in\boldsymbol{H}^{1/2}(\Gamma_u)$ to $\boldsymbol{H}^1(\Omega)$; and $\vartheta\in\widetilde{\vartheta}^{\Gamma_\theta}+H^1_{\Gamma_\theta}(\Omega)$, where $\widetilde{\vartheta}^{\Gamma_\theta}$ is an extension of the boundary condition $\vartheta^{\Gamma_\theta}\in H^{1/2}(\Gamma_\theta)$ to $H^1(\Omega)$, with $\vartheta^{\Gamma_\theta}=\theta^{\Gamma_\theta}-\bar{\theta}$. For simplicity it will be assumed that $\boldsymbol{f}\in\boldsymbol{L}^2(\Omega)$ and $r\in L^2(\Omega)$ (in reality these assumptions can be relaxed later). Meanwhile, ρ, $\boldsymbol{\kappa}$, C^*, $\boldsymbol{\alpha}^*$ and c_v^*, can in principle be heterogeneous in Ω but should remain bounded. Then, provided $\boldsymbol{v}\in\boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\chi\in H^1_{\Gamma_\theta}(\Omega)$, (5.5) may be recast as the problem of seeking $\boldsymbol{u}_0\in\boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\vartheta_0\in H^1_{\Gamma_\theta}(\Omega)$ such that,

$$\begin{aligned}
b_0\big((\boldsymbol{u}_0,\vartheta_0),(\boldsymbol{v},\chi)\big)&=\ell_0\big((\boldsymbol{v},\chi)\big)\qquad\forall\boldsymbol{v}\in\boldsymbol{H}^1_{\Gamma_u}(\Omega),\ \forall\chi\in H^1_{\Gamma_\theta}(\Omega)\,,\\
b_0\big((\boldsymbol{u},\vartheta),(\boldsymbol{v},\chi)\big)&=(\mathrm{i}\omega)^2\big(\rho\boldsymbol{u},\boldsymbol{v}\big)_\Omega+\big(\mathsf{C}^*\!:\!(\boldsymbol{\varepsilon}-\boldsymbol{\alpha}^*\vartheta),\boldsymbol{\nabla v}\big)_\Omega\\
&\quad+(\mathrm{i}\omega)\big(\rho c_v^*\vartheta,\chi\big)_\Omega+(\mathrm{i}\omega)\big(\bar{\theta}\boldsymbol{\alpha}^*\!:\!\mathsf{C}^*\!:\!\boldsymbol{\varepsilon},\chi\big)_\Omega+\big(\boldsymbol{\kappa}\cdot\nabla\vartheta,\nabla\chi\big)_\Omega\,,\\
\ell_0\big((\boldsymbol{v},\chi)\big)&=\big(\boldsymbol{f},\boldsymbol{v}\big)_\Omega+\langle\check{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma},\mathbf{tr}^{\partial\Omega}_{\mathrm{grad}}\boldsymbol{v}\rangle_{\partial\Omega}\\
&\quad+\big(r,\chi\big)_\Omega-\langle\check{q}_{\mathbf{n}}^{\Gamma_q},\mathrm{tr}^{\partial\Omega}_{\mathrm{grad}}\chi\rangle_{\partial\Omega}-b_0\big((\widetilde{\boldsymbol{u}}^{\Gamma_u},\widetilde{\vartheta}^{\Gamma_\theta}),(\boldsymbol{v},\chi)\big)\,,
\end{aligned}\tag{5.6}$$

where the final solution takes the form $\boldsymbol{u}=\boldsymbol{u}_0+\widetilde{\boldsymbol{u}}^{\Gamma_u}$ and $\vartheta=\vartheta_0+\widetilde{\vartheta}^{\Gamma_\theta}$, and where $\check{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma}$ is an extension of the boundary condition $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma}\in\boldsymbol{H}^{-1/2}(\Gamma_\sigma)$ to $\boldsymbol{H}^{-1/2}(\partial\Omega)$; and $\check{q}_{\mathbf{n}}^{\Gamma_q}$ is an extension of the boundary condition $q_{\mathbf{n}}^{\Gamma_q}\in H^{-1/2}(\Gamma_q)$ to $H^{-1/2}(\partial\Omega)$. Here, $\langle\cdot,\cdot\rangle_{\partial\Omega}$ is the duality pairing between

$H^{1/2}(\partial\Omega)$ and $H^{-1/2}(\partial\Omega)$ and vice versa (or their vector counterparts). This is known as the primal variational formulation.

To implement certain numerical methods, such as discontinuous Petrov-Galerkin (DPG) finite element methods, it is more convenient to have test functions that are discontinuous across a mesh, $\mathcal{T}$, of Ω, which is comprised of open elements $K \in \mathcal{T}$. Thus, one would test with broken test functions $\boldsymbol{v} \in \boldsymbol{H}^1(\mathcal{T})$ and $\chi \in H^1(\mathcal{T})$ (see Section 3.2.1 and Appendix A for the definitions). The resulting inner products look like,

$$\big(\boldsymbol{u},\boldsymbol{v}\big)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} \big(\boldsymbol{u}|_K,\boldsymbol{v}|_K\big)_K\,, \qquad \big\langle \hat{\boldsymbol{u}},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}\big\rangle_{\partial\mathcal{T}} = \sum_{K\in\mathcal{T}} \big\langle \hat{\boldsymbol{u}}_K,(\hat{\boldsymbol{\sigma}}_{\mathbf{n}})_K\big\rangle_{\partial K}\,, \tag{5.7}$$

for the functions $\boldsymbol{u},\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)$, and the $\mathcal{T}$-tuples, $\hat{\boldsymbol{u}} \in \boldsymbol{H}^{1/2}(\partial\mathcal{T})$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{n}} \in \boldsymbol{H}^{-1/2}(\partial\mathcal{T})$ (again see Section 3.2.1 and Appendix A for definitions). Obviously, the notation is the same for scalar-valued functions. Proceeding as in [111], the resulting broken primal variational formulation seeks $\mathfrak{u}_0 = (\boldsymbol{u}_0,\vartheta_0) \in \boldsymbol{H}^1_{\Gamma_u}(\Omega) \times H^1_{\Gamma_\theta}(\Omega)$ and $\hat{\mathfrak{u}}_0 = (\hat{\boldsymbol{\sigma}}_{\mathbf{n},0},\hat{q}_{\mathbf{n},0}) \in \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T}) \times H^{-1/2}_{\Gamma_q}(\partial\mathcal{T})$ such that

$$\begin{aligned}
b_{\mathcal{T}}\big((\mathfrak{u}_0,\hat{\mathfrak{u}}_0),\mathfrak{v}\big) &= \ell_{\mathcal{T}}\big(\mathfrak{v}\big) \qquad \forall \mathfrak{v} = (\boldsymbol{v},\chi) \in \boldsymbol{H}^1(\mathcal{T}) \times H^1(\mathcal{T})\,,\\
b_{\mathcal{T}}\big((\boldsymbol{u},\vartheta,\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\hat{q}_{\mathbf{n}}),(\boldsymbol{v},\chi)\big) &= b_0\big((\boldsymbol{u},\vartheta),(\boldsymbol{v},\chi)\big) + \hat{b}\big((\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\hat{q}_{\mathbf{n}}),(\boldsymbol{v},\chi)\big)\,,\\
b_0\big((\boldsymbol{u},\vartheta),(\boldsymbol{v},\chi)\big) &= (\mathrm{i}\omega)^2\big(\rho\boldsymbol{u},\boldsymbol{v}\big)_{\mathcal{T}} + \big(\mathsf{C}^*\!:\!(\boldsymbol{\varepsilon}-\boldsymbol{\alpha}^*\vartheta),\boldsymbol{\nabla}\boldsymbol{v}\big)_{\mathcal{T}}\\
&\quad + (\mathrm{i}\omega)\big(\rho c_v^*\vartheta,\chi\big)_{\mathcal{T}} + (\mathrm{i}\omega)\big(\bar{\theta}\boldsymbol{\alpha}^*\!:\!\mathsf{C}^*\!:\!\boldsymbol{\varepsilon},\chi\big)_{\mathcal{T}} + \big(\boldsymbol{\kappa}\cdot\nabla\vartheta,\nabla\chi\big)_{\mathcal{T}}\,,\\
\hat{b}\big((\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\hat{q}_{\mathbf{n}}),(\boldsymbol{v},\chi)\big) &= -\big\langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\boldsymbol{v}\big\rangle_{\partial\mathcal{T}} + \big\langle\hat{q}_{\mathbf{n}},\mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}}\chi\big\rangle_{\partial\mathcal{T}}\,,\\
\ell_{\mathcal{T}}\big((\boldsymbol{v},\chi)\big) &= \big(\boldsymbol{f},\boldsymbol{v}\big)_{\mathcal{T}} + \big(r,\chi\big)_{\mathcal{T}} - b_{\mathcal{T}}\big((\widetilde{\boldsymbol{u}}^{\Gamma_u},\widetilde{\vartheta}^{\Gamma_\theta},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma},\hat{q}_{\mathbf{n}}^{\Gamma_q}),(\boldsymbol{v},\chi)\big)\,,
\end{aligned} \tag{5.8}$$

where the final solution takes the form $\boldsymbol{u} = \boldsymbol{u}_0 + \widetilde{\boldsymbol{u}}^{\Gamma_u}$, $\vartheta = \vartheta_0 + \widetilde{\vartheta}^{\Gamma_\theta}$, $\hat{\boldsymbol{\sigma}}_{\mathbf{n}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n},0} + \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma}$ and $\hat{q}_{\mathbf{n}} = \hat{q}_{\mathbf{n}} + \hat{q}_{\mathbf{n}}^{\Gamma_q}$, with $\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\Gamma_\sigma}$ being an extension of the boundary condition $\boldsymbol{\sigma}_{\mathbf{n}}^{\Gamma_\sigma} \in \boldsymbol{H}^{-1/2}(\Gamma_\sigma)$ to $\boldsymbol{H}^{-1/2}(\partial\mathcal{T})$ and $\hat{q}_{\mathbf{n}}^{\Gamma_q}$ being an extension of the boundary condition $q_{\mathbf{n}}^{\Gamma_q} \in H^{-1/2}(\Gamma_q)$ to $H^{-1/2}(\partial\mathcal{T})$.

In the particular case of Figure 5.3, the domain contains two isotropic materials, so that the material constants ρ, $\boldsymbol{\kappa}$, C^*, $\boldsymbol{\alpha}^*$ and c_v^* will be discontinuous across the domain. The source terms $\boldsymbol{f}$ and r will be assumed to vanish. Moreover, viewed as a 2D domain, $\Gamma_u = \Gamma_\theta = \partial\Omega$, and the displacement boundary conditions are assumed to vanish, $\boldsymbol{u}^{\Gamma_u} = 0$, whereas the temperature boundary conditions are taken as $\theta(t) = 348 + 25\sin(\omega t)\,\mathrm{K}$, with $\bar{\theta} = 348\,\mathrm{K} = 75\,^\circ\mathrm{C}$ uniformly in Ω, so that $\vartheta(t) = 25\sin(\omega t)\,\mathrm{K}$ (so $\vartheta = -25\mathrm{i}$ in the frequency domain as shown in Remark 5.1). As a 3D domain, the boundary conditions in the "2D faces" (the faces normal to the direction where there is no variation) impose vanishing normal displacement and vanishing tangential stresses, whereas full heat flux surface boundary conditions are used on those faces. Note that the assumed variation in temperature (namely, 25 K) might not be sufficiently small to justify using the *linear* thermoviscoelasticity equations, but the risk is taken with the hope of at least getting an estimated solution. Lastly, the angular frequency was taken as $\omega = 2\pi{\cdot}0.05\,\mathrm{rad/s}$, and at that frequency (and $\bar{\theta} = 348\,\mathrm{K}$) the (isotropic) material properties, ρ, κ, E^*, ν^*, α^* and c_v^* (determining $\boldsymbol{\kappa}$, C^*, $\boldsymbol{\alpha}^*$), for an epoxy resin, a silicone resin, and copper are given in Table 5.1.

$\omega{=}2\pi{\cdot}0.05\,\mathrm{rad/s}$ $\bar{\theta}{=}348\,\mathrm{K}$	Epoxy	Silicone	Copper
ρ (kg/m^3)	1247	1128	8909
κ (W/(kg·K))	0.12	0.16	390
E^* (MPa)	$1976 + \mathrm{i}38$	$184 + \mathrm{i}43$	117600
ν^*	$0.2883 + \mathrm{i}0.1923$	$0.4629 + \mathrm{i}0.0264$	0.343
α^* (1/MK)	$38 + \mathrm{i}42$	$118 + \mathrm{i}91$	17.78
c_v^* (J/(kg·K))	$1261 - \mathrm{i}880$	$1560 - \mathrm{i}814$	378

Table 5.1: Viscoelastic material properties of an epoxy resin, silicone resin and copper for an angular frequency $\omega = 2\pi{\cdot}0.05\,\mathrm{rad/s}$ and a temperature $\bar{\theta} = 348\,\mathrm{K}$.

5.3.2 Results

The broken variational formulation in (5.8) was solved numerically using a DPG method as in [111]. Indeed, the trial space, $\mathcal{U} = \boldsymbol{H}^1_{\Gamma_u}(\Omega) \times H^1_{\Gamma_\theta}(\Omega) \times \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T}) \times H^{-1/2}_{\Gamma_q}(\partial\mathcal{T})$, and test space, $\mathcal{V} = \boldsymbol{H}^1(\mathcal{T}) \times H^1(\mathcal{T})$, were SdR spaces (see Section A.5 in Appendix A for definitions). Thus, the discrete trial space, $\mathcal{U}_h \subseteq \mathcal{U}$ was chosen as a compatible SdR discretizations of order p, while the enriched test space, $\mathcal{V}_r \subseteq \mathcal{V}$, was chosen as an SdR discretization of order $p + \Delta p$. The software package used to solve the resulting DPG method was `hp3d`, which is an in-house code. It has support for SdR discretizations for all the conventional element shapes [114], local h and p anisotropic refinements via constrained approximations, and the possibility of using transfinite interpolation to implement curvilinear elements (for more details see Section 2.9.1). This last feature was important for the rounded corners present in the interfaces of the two-material model geometry seen in Figure 5.2, which was discretized using both hexahedra and triangular prisms.

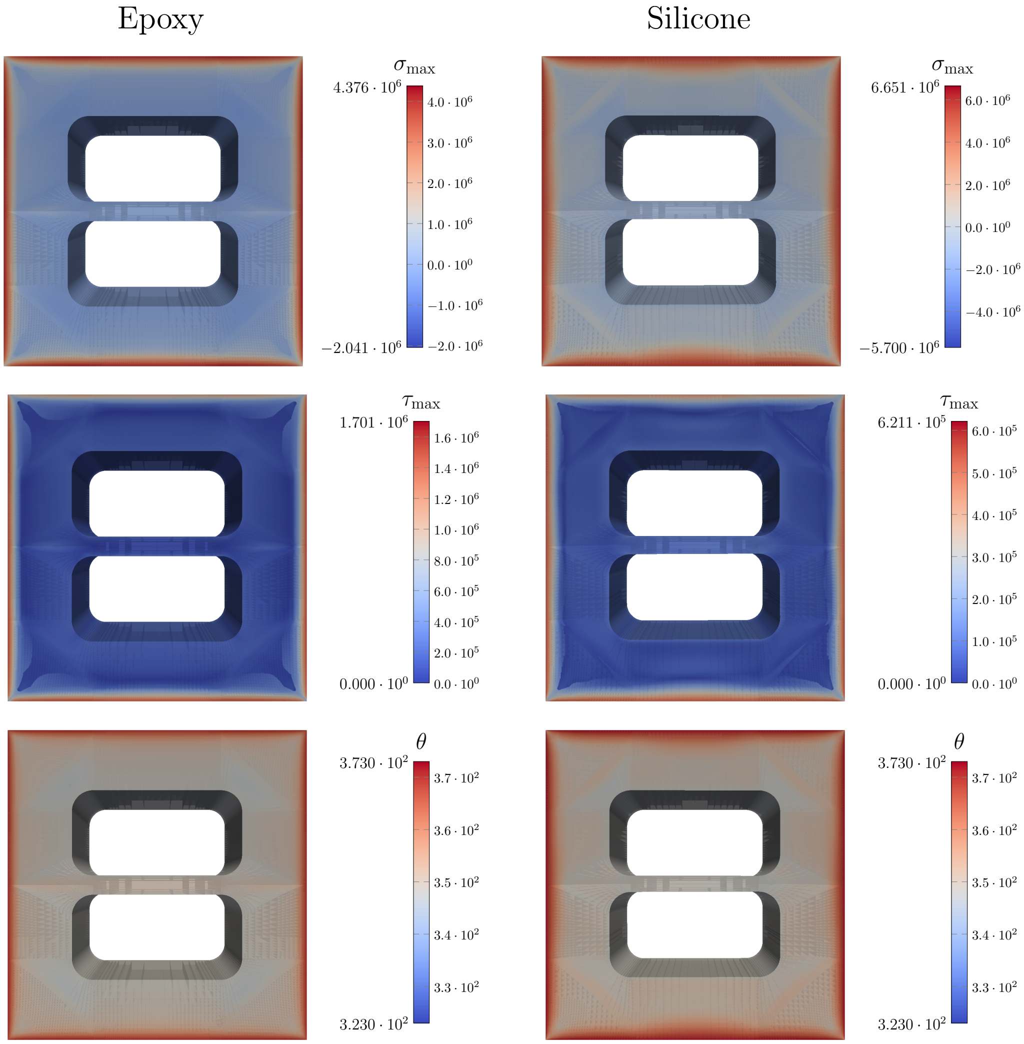


Figure 5.4: Maximum principal stress ($\sigma_{\max}$ in Pa), maximum shear stress ($\tau_{\max}$ in Pa), and temperature (θ in K) in a form-wound coil (see Figure 5.2). The thermoviscoelastic solution of (5.8) was obtained using an adaptive DPG method (with $p = 2$ and $\Delta p = 1$). Both an epoxy resin and a silicone resin were considered, while the coils were made of copper (not illustrated). The relatively low frequency was $\omega = 2\pi{\cdot}0.05\,\mathrm{rad/s}$ and the average temperature was $\bar{\theta} = 348\,\mathrm{K}$, with the material properties taken from Table 5.1. For visual purposes, the domain is warped by a scaled solution displacement.

The results of the simulation are illustrated in Figure 5.4, where $p = 2$ and $\Delta p = 1$ were used for the DPG discretizations. The results are in the time domain plotted at the time step where the highest stresses were observed and at the time step where the highest temperatures were observed for each of the two materials (the copper coils are not shown). The use of the natural residual-based a posteriori error estimator that comes with the DPG method was fundamental in resolving all the localized solution features present (especially in the stresses). From a qualitative perspective, the behavior was similar in both thermoviscoelastic resins. Namely, their maximum principal stress (largest eigenvalue of $\boldsymbol{\sigma}$), σ_{max}, and maximum shear stress ($\frac{\sigma_{\text{max}}-\sigma_{\text{min}}}{2}$, where σ_{max} and σ_{min} are the largest and smallest eigenvalues of $\boldsymbol{\sigma}$), τ_{max}, occurred near the outer boundary. However, the quantitative behavior was different. The maximum principal stress was higher for the silicone than for the epoxy (by about 50%), but the shear stress was higher in the epoxy (almost by 300%). The difference in quantitative behavior might be related to the different thermal properties of both resins. Indeed, the epoxy has a lower heat conductivity than the silicone (see Table 5.1), and the result is that the temperature does not appear to have penetrated as much as in the silicone resin (see distribution of temperature in Figure 5.4). Having said that, it should be noted that it could be more realistic if some form of heat would come from the copper coils themselves, instead of the boundaries. In fact, Joule heating in the copper is what drives the heating in reality, while the cooling comes from the outer boundaries. In the future, it could be possible to incorporate such Joule heating via the body source term, r.

Next, the work done in one second was calculated for both resins. The work is the integral of the power developed in the resin over one second. It is assumed that the material repeats its behavior periodically in time with an angular frequency ω. In that case, it is defined as,

$$W_{\text{ins}} = \int_{\Omega_{\text{ins}}} \frac{\omega}{2\pi} \int_0^{\frac{2\pi}{\omega}} \boldsymbol{\sigma}(\boldsymbol{x},t) : \dot{\boldsymbol{\varepsilon}}(\boldsymbol{x},t)\,\mathrm{d}t\;\mathrm{d}\Omega_{\text{ins}}\,, \tag{5.9}$$

with Ω_{ins} being the domain associated to the insulation (where the thermoviscoelastic resin is present). Note that the power developed typically includes a term for the time derivative of the kinetic energy [187, §4.4], $\frac{\mathrm{d}}{\mathrm{d}t}\big(\frac{1}{2}\rho\dot{\boldsymbol{u}}\cdot\dot{\boldsymbol{u}}\big)$, but under the periodicity assumption its integral will vanish, $\frac{1}{2}\rho\dot{\boldsymbol{u}}\cdot\dot{\boldsymbol{u}}\big|_{t=0}^{t=2\pi/\omega} = 0$. Under the assumption that $\boldsymbol{u}_i(t) = \boldsymbol{A}_{\boldsymbol{u}_i}\cos(\omega t+\psi_{\boldsymbol{u}_i})$ and $\vartheta(t) = A_\vartheta\cos(\omega t+\psi_\vartheta)$, it follows $\boldsymbol{u}_i = \boldsymbol{A}_{\boldsymbol{u}_i} e^{\mathrm{i}\psi_{\boldsymbol{u}_i}}$ and $\vartheta = A_\vartheta e^{\mathrm{i}\psi_\vartheta}$ in the frequency domain as argued in Remark 5.1, so that the dissipative work can be deduced to be,

$$W_{\text{ins}} = \int_{\Omega_{\text{ins}}} \frac{\omega}{2\pi}\Big(\frac{\pi}{\omega}\mathfrak{Re}(\boldsymbol{\sigma}:\bar{\dot{\boldsymbol{\varepsilon}}})\Big)\,\mathrm{d}\Omega_{\text{ins}} = \int_{\Omega_{\text{ins}}} \frac{1}{2}\mathfrak{Re}(-\mathrm{i}\boldsymbol{\sigma}:\bar{\boldsymbol{\varepsilon}})\,\mathrm{d}\Omega_{\text{ins}}\,, \tag{5.10}$$

where the bar denotes complex conjugation. Recall from (5.1) and (5.4) that, in the frequency domain, $\boldsymbol{\sigma} = \mathsf{C}^* : (\boldsymbol{\varepsilon} - \boldsymbol{\alpha}^*\vartheta)$, so $\boldsymbol{\sigma} : \bar{\boldsymbol{\varepsilon}} = (\boldsymbol{\nabla u} - \boldsymbol{\alpha}^*\vartheta) : \mathsf{C}^* : \overline{\boldsymbol{\nabla u}}$. The values for both the epoxy resin and the silicone resin are $W_{\text{ins}}^{\text{Ep}} = -3.870 \cdot 10^{-5}\,\text{J}$ and $W_{\text{ins}}^{\text{Si}} = -1.765 \cdot 10^{-4}\,\text{J}$ respectively.

5.4 Mid-range frequency: stator ovalization

5.4.1 Description and problem setup

During normal operation of an electric machine, certain vibrations (along with acoustic noise) can develop in both the stator and the rotor of the machine. These vibrations can even become large if one of the structure's natural frequencies (either in the stator or rotor) coincides with a multiple of the operating rotational speed of the machine, and they can naturally induce a deformation of the "turn-to-turn" domain of interest. It is possible to calculate such natural frequencies directly [127], but for the purposes of this work it will be assumed that such natural frequencies do coincide with a multiple of the machine's speed. That is, $f_{\text{vib}} = n_{\text{vib}} f_s$, where f_{vib} is such natural frequency in Hz, $n_{\text{vib}} \in \mathbb{N}$ and f_s is the synchronous speed of the machine in Hz (i.e. rotations per second). In [51] it was determined experimentally for a synchronous AC motor operating at $7000\,\text{rpm} \approx 117\,\text{Hz} = f_s$ that a stator vibrational mode of elliptical type, referred to here as "ovalization", occurs at about $f_{\text{vib}} = 2.6\,\text{kHz}$, so that, roughly speaking, n_{vib} lies in the range 21–23. The ovalization is sketched in Figure 5.5 along with the induced displacement boundary conditions. In this case, a synchronous AC single-phase permanent magnet motor with a 6-pole stator and current supplied at $f_{\text{AC}} = 400\,\text{Hz}$ will be considered. Thus, its synchronous operating speed is $f_s = \frac{f_{\text{AC}}}{P_{\text{pole-pairs}}} = \frac{400\,\text{Hz}}{3} \approx 133\,\text{Hz}$ (where $P_{\text{pole-pairs}}$ is half the number of stator poles per phase). Assuming n_{vib} lies in the same range as that of the motor in [51], then implies that f_{vib} lies in the range between $2.8\,\text{kHz}$ and $3.07\,\text{kHz}$. For the purposes of this section, f_{vib} will be taken as $f_{\text{vib}} = 2.8\,\text{kHz}$.

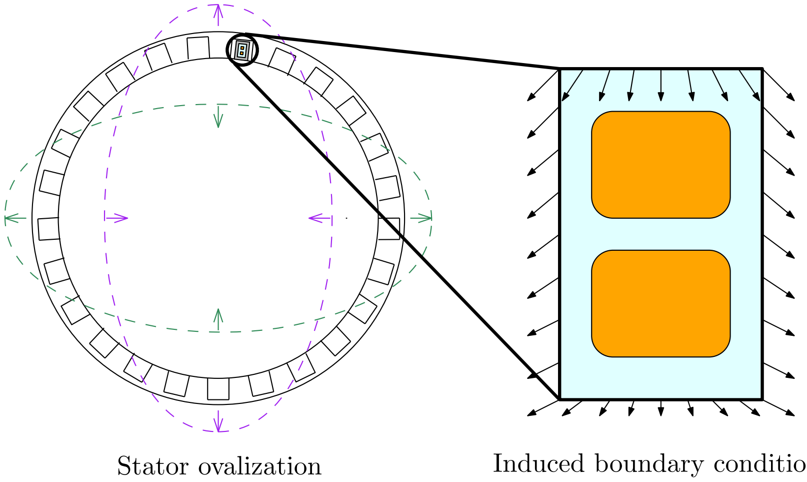


Figure 5.5: Stator ovalization during machine operation leads to deformation of the form-wound coils.

The induced time-dependent linearized displacement boundary conditions are of the type shown in Figure 5.5. More specifically, looking at Figure 5.2 for the definitions of the geometrical parameters, and taking the middle of the domain as the origin, the displacement boundary conditions (viewing the domain as a 2D domain) are,

$$\begin{aligned} u_1(\boldsymbol{x},t) &= A\big(\tfrac{2}{w}x_1\big)\big(\tfrac{1}{2h}x_2 + \tfrac{3}{4}\big)\sin(\omega t)\,, \\ u_2(\boldsymbol{x},t) &= -A\big(\tfrac{1}{2h}x_2 + \tfrac{3}{4}\big)\sin(\omega t)\,, \end{aligned} \tag{5.11}$$

where $\omega = 2\pi f_{\mathrm{vib}} = 2\pi{\cdot}2800\,\mathrm{rad/s}$, $\boldsymbol{x} = (x_1, x_2) \in \partial\Omega$ with $\Omega = (-\frac{w}{2}, \frac{w}{2}) \times (-\frac{h}{2}, \frac{h}{2})$, and h and w being the height and width of the model domain. Here, A is a reasonable vibration displacement (note that the maximum boundary displacement occurs at $\boldsymbol{x} = (\pm\frac{w}{2}, \frac{h}{2})$ and has magnitude $\sqrt{2}A$). In this work, it will be assumed that $A = 100\,\mu\mathrm{m}$. Viewed as a 3D domain, $u_3(\boldsymbol{x},t) = 0$ everywhere, and vanishing tangential stresses are assumed at the faces normal to the “2D domain”. Obviously, these boundary conditions can easily be expressed in the frequency domain as suggested in Remark 5.1.

As implied by the specification of boundary conditions, no thermal effects are assumed to be relevant, with the temperature being constant at $\bar{\theta} = 373\,K = 100\,^{\circ}\mathrm{C}$. Therefore, the physics are modeled with the linear viscoelasticity equations as in Chapter 4. Indeed, assuming $\boldsymbol{\alpha} = 0$, the equations in (5.1) decouple and yield (in their second-order form),

$$\begin{cases} \rho\ddot{\boldsymbol{u}} - \mathbf{div}(\dot{\mathsf{C}} \mathbin{\ddagger} \boldsymbol{\varepsilon}) = \boldsymbol{f}\,, \\ \rho\dot{c}_v \ast \dot{\theta} - \operatorname{div}(\boldsymbol{\kappa}\cdot\nabla\theta) = r\,. \end{cases} \tag{5.12}$$

For the second equation one can take $r = 0$ and place uniform constant temperature boundary conditions, which will yield that the temperature remains constant throughout the domain. In the frequency domain, the relevant second-order linear viscoelasticity equation is,

$$(\mathrm{i}\omega)^2\rho\boldsymbol{u} - \mathbf{div}\left(\mathsf{C}^*\!:\!\boldsymbol{\varepsilon}\right) = \boldsymbol{f}\,. \tag{5.13}$$

As in the previous section, the broken variational formulation this time seeks $\mathfrak{u}_0 = \boldsymbol{u}_0 \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\hat{\mathfrak{u}}_0 = \hat{\boldsymbol{\sigma}}_{\mathbf{n},0} \in \boldsymbol{H}^{-1/2}_{\Gamma_\sigma}(\partial\mathcal{T})$ such that,

$$\begin{aligned} b_{\mathcal{T}}\big((\mathfrak{u}_0,\hat{\mathfrak{u}}_0),\mathfrak{v}\big) &= \ell_{\mathcal{T}}\big(\mathfrak{v}\big) \qquad \forall\mathfrak{v} = \boldsymbol{v} \in \boldsymbol{H}^1(\mathcal{T})\times H^1(\mathcal{T})\,, \\ b_{\mathcal{T}}\big((\boldsymbol{u},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}),\boldsymbol{v}\big) &= b_0\big(\boldsymbol{u},\boldsymbol{v}\big) + \hat{b}\big(\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\boldsymbol{v}\big)\,, \\ b_0\big(\boldsymbol{u},\boldsymbol{v}\big) &= (\mathrm{i}\omega)^2\big(\rho\boldsymbol{u},\boldsymbol{v}\big)_{\mathcal{T}} + \big(\mathsf{C}^*\!:\!\boldsymbol{\varepsilon},\boldsymbol{\nabla}\boldsymbol{v}\big)_{\mathcal{T}}\,, \\ \hat{b}\big(\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\boldsymbol{v}\big) &= -\big\langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}},\mathbf{tr}^{\mathcal{T}}_{\mathrm{grad}}\boldsymbol{v}\big\rangle_{\partial\mathcal{T}}\,, \\ \ell_{\mathcal{T}}\big((\boldsymbol{v},\chi)\big) &= \big(\boldsymbol{f},\boldsymbol{v}\big)_{\mathcal{T}} - b_{\mathcal{T}}\big((\widetilde{\boldsymbol{u}}^{\Gamma_u},\hat{\boldsymbol{\sigma}}^{\Gamma_\sigma}_{\mathbf{n}}),\boldsymbol{v}\big)\,, \end{aligned} \tag{5.14}$$

where the final solution takes the form $\boldsymbol{u} = \boldsymbol{u}_0 + \widetilde{\boldsymbol{u}}^{\Gamma_u}$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{n}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n},0} + \hat{\boldsymbol{\sigma}}^{\Gamma_\sigma}_{\mathbf{n}}$, with $\widetilde{\boldsymbol{u}}^{\Gamma_u}$ being an extension of the boundary condition $\boldsymbol{u}^{\Gamma_u} \in \boldsymbol{H}^{1/2}(\Gamma_u)$ to $\boldsymbol{H}^1(\Omega)$ and $\hat{\boldsymbol{\sigma}}^{\Gamma_\sigma}_{\mathbf{n}}$ being an extension of the boundary condition $\boldsymbol{\sigma}^{\Gamma_\sigma}_{\mathbf{n}} \in \boldsymbol{H}^{-1/2}(\Gamma_\sigma)$ to $\boldsymbol{H}^{-1/2}(\partial\mathcal{T})$.

Lastly, the material properties at $\omega = 2\pi\cdot 2800\,\mathrm{rad/s}$ and $\bar{\theta} = 373\,K = 100\,°\mathrm{C}$ are provided in Table 5.2. These values were obtained experimentally, and are used in the simulations that follow.

$\omega = 2\pi\cdot 2800\,\mathrm{rad/s}$ $\bar{\theta} = 373\,\mathrm{K}$	Epoxy	Silicone	Copper
ρ (kg/m^3)	1245	1125	8909
E^* (MPa)	$1905 + \mathrm{i}37$	$340 + \mathrm{i}44$	117600
ν^*	$0.29 + \mathrm{i}0.12$	$0.433 + \mathrm{i}0.052$	0.343

Table 5.2: Viscoelastic material properties of an epoxy resin, silicone resin and copper for an angular frequency $\omega = 2\pi\cdot 2800\,\mathrm{rad/s}$ and a temperature $\bar{\theta} = 373\,\mathrm{K}$.

5.4.2 Results

The equations were solved using adaptive DPG methods to discretize the broken variational formulation (5.14) as described in Section 5.3.2. The values of $p = 2$ and $\Delta p = 1$ were used for the computations and the results are illustrated in Figure 5.6 (in the time domain) at the time step where the highest stresses were observed (the copper coils are excluded).

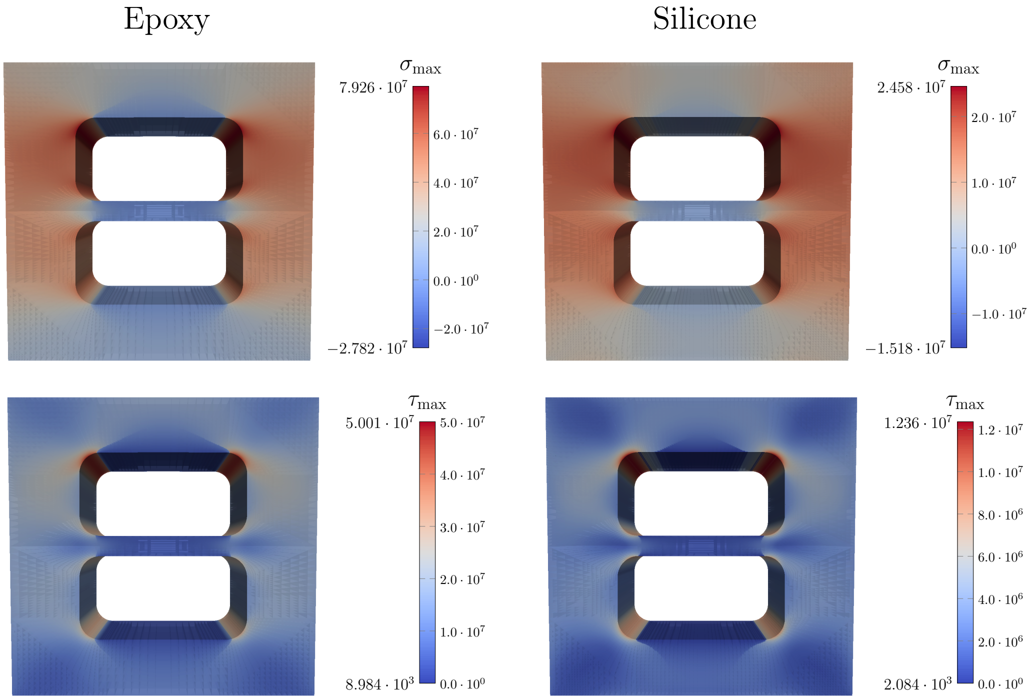


Figure 5.6: Maximum principal stress ($\sigma_{\max}$ in Pa) and maximum shear stress ($\tau_{\max}$ in Pa) in a form-wound coil (see Figure 5.2). The viscoelastic solution of (5.14) was obtained using an adaptive DPG method (with $p = 2$ and $\Delta p = 1$). Both an epoxy resin and a silicone resin were considered, while the coils were made of copper (not illustrated). The mid-range ovalization frequency was $\omega = 2\pi{\cdot}2800\,\mathrm{rad/s}$ and the average temperature was $\bar{\theta} = 373\,\mathrm{K}$, with the material properties taken from Table 5.2. For visual purposes, the domain is warped by a scaled solution displacement.

In both resins, the results were qualitatively very similar, with the maximum principal stress and shear stress occurring near the top corners of the interface of the top coil and the resin. Once again, the DPG a posteriori error estimator was instrumental in resolving and adapting to such localized solution features. Quantitatively, the both the principal and shear stresses were higher for the epoxy resin than the silicone resin by about a factor of 3. Lastly, using (5.10) with $\boldsymbol{\sigma} = \mathsf{C}^* : \boldsymbol{\varepsilon}$ (so $\boldsymbol{\sigma} : \bar{\boldsymbol{\varepsilon}} = \boldsymbol{\nabla u} : \mathsf{C}^* : \overline{\boldsymbol{\nabla u}}$), it follows that the work done in one second for each resin is $W_{\text{ins}}^{\text{Ep}} = 1.141 \cdot 10^5$ J and $W_{\text{ins}}^{\text{Si}} = 1.077 \cdot 10^5$ J for epoxy and silicone respectively.

5.5 High frequency: Lorentz forces

5.5.1 Description and problem setup

As alluded to in the previous section, assume an electric motor with a base current supplied at $f_{\text{AC}} = 400\,\text{Hz}$, which is controlled via pulse-width modulation (PWM) techniques [217] associated to a much higher switching frequency of $f_{\text{sw}} = 200\,\text{kHz}$. The end result is that the current supplied can be modeled as a sum of two alternating cosine waves: one with a base frequency of $f_{\text{AC}} = 400\,\text{Hz}$ and amplitude of about $I_{\text{AC}} = 20\,\text{A}$, and a second "ripple" with a much higher frequency of $f_{\text{sw}} = 200\,\text{kHz}$ (500 times more than f_{AC}) and much lower amplitude of $I_{\text{sw}} = 0.04\,\text{A}$ (500 times less than the base current amplitude).

The current supplied produces a current density field, $\boldsymbol{J}$, that travels through each coil and is modeled as an impressed current. This in turn produces magnetic and electric fields, $\boldsymbol{B}$ and $\boldsymbol{E}$, via Maxwell's equations, as well as a charge density field, ρ_{ch}. Then, the interaction of these fields, in particular the magnetic field and the current density field, produces a Lorentz force [134], $\boldsymbol{f} = \rho_{\text{ch}}\boldsymbol{E} + \boldsymbol{J} \times \boldsymbol{B}$. Lastly, this Lorentz force can be used as the sole loading that drives the deformation of the "turn-to-turn" model domain. Thus, assume vanishing displacement boundary conditions everywhere if the domain is viewed in 2D (vanishing tangential stresses and normal displacement at the faces normal to the "2D domain" if viewed in 3D). Once again, thermal effects are ignored (temperature assumed constant at $\bar{\theta} = 373\,K = 100\,^\circ\text{C}$), so that the linear viscoelasticity equations are considered in the form of the broken variational formulation in (5.14). Note that the expression for the Lorentz force includes the term $\boldsymbol{J} \times \boldsymbol{B}$, where both $\boldsymbol{J}$ and $\boldsymbol{B}$ have frequency components in f_{AC} and f_{sw}, meaning that the final Lorentz force combines these frequencies. The highest frequency involved in the product is $2f_{\text{sw}} = 400\,\text{kHz}$, and it is this frequency that will be analyzed for this loading scenario. The next section is devoted to explaining how to model and calculate this Lorentz force. In the meantime, the material properties of the resins at such high frequencies are given in Table 5.3.

$\omega = 2\pi{\cdot}400000\,\text{rad/s}$ $\bar{\theta} = 373\,\text{K}$	Epoxy	Silicone	Copper
ρ (kg/m^3)	1245	1125	8909
E^* (MPa)	$1945 + \text{i}33$	$480 + \text{i}39$	117600
ν^*	$0.289 + \text{i}0.121$	$0.422 + \text{i}0.071$	0.343

Table 5.3: Viscoelastic material properties of an epoxy resin, silicone resin and copper for an angular frequency $\omega = 2\pi{\cdot}400000\,\text{rad/s}$ and a temperature $\bar{\theta} = 373\,\text{K}$.

5.5.2 Electromagnetic model for surface Lorentz forces

In what follows, the model for the Lorentz force will be described. The first assumption is the medium will be assumed to be uniform (even though it is composed of at least copper and a resin). The second assumption is that the coils are assumed to be of infinite length. The third assumption is that due to such high frequencies, the current distribution is fully concentrated at the interface between the copper and the resin. The fourth assumption is that the current is distributed evenly along the interface. Lastly, the fifth assumption is that the current can be approximated by a finite number of infinite electric line sources lying at the interface between the copper and the resin (see Figure 5.7).

The first four assumptions along with the assumed form of the current means that the impressed current density, $\boldsymbol{J}(\boldsymbol{x},t)$, can be written as

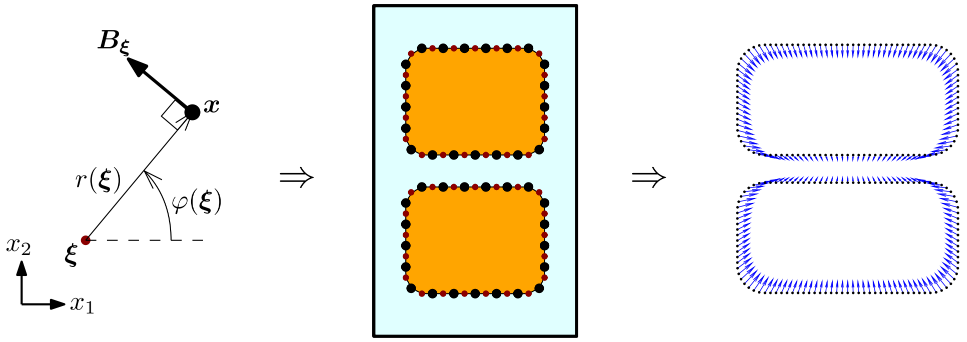
$$\begin{gathered}\boldsymbol{J}(\boldsymbol{x},t) = \boldsymbol{J}_{\text{AC}}(\boldsymbol{x})\cos(\omega_{\text{AC}}t) + \boldsymbol{J}_{\text{sw}}(\boldsymbol{x})\cos(\omega_{\text{sw}}t)\,,\\ \boldsymbol{J}_{\text{AC}}(\boldsymbol{x}) = \frac{I_{\text{AC}}}{l_{\text{coil}}}\delta_{\Gamma_{\text{Cu}}}(\boldsymbol{x})\begin{pmatrix}0\\0\\1\end{pmatrix}\,, \qquad \boldsymbol{J}_{\text{sw}}(\boldsymbol{x}) = \frac{I_{\text{sw}}}{l_{\text{coil}}}\delta_{\Gamma_{\text{Cu}}}(\boldsymbol{x})\begin{pmatrix}0\\0\\1\end{pmatrix}\,,\end{gathered} \tag{5.15}$$

where l_{coil} is the perimeter of one of the rounded coils (i.e. $l_{\text{coil}} = 2\big(\pi R_r + (w_{\text{Cu}} - 2R_r) + (h_{\text{Cu}} - 2R_r)\big)$ with the coil geometry shown in Figure 5.2), $\omega_{\text{AC}} = 2\pi f_{\text{AC}}$, $\omega_{\text{sw}} = 2\pi f_{\text{sw}}$, and $\delta_{\Gamma_{\text{Cu}}}(\boldsymbol{x})$ is a surface Dirac delta with distributional support at the interfaces between the copper and resin, Γ_{Cu}. Indeed, this means that the total current at $f_{\text{AC}} = 400\,\text{Hz}$ flowing through each coil is $I_{\text{AC}} = 20\,\text{A}$, while that at $f_{\text{sw}} = 200\,\text{kHz}$ is $I_{\text{sw}} = 0.04\,\text{A}$.

The distributional divergence of $\boldsymbol{J}_{\mathrm{sw}}$ is,

$$\int_{\Omega} (\operatorname{div} \boldsymbol{J}_{\mathrm{sw}}) \phi \, \mathrm{d}\Omega = - \int_{\Omega} \boldsymbol{J}_{\mathrm{sw}} \cdot \nabla \phi \, \mathrm{d}\Omega = \int_{\Gamma_{\mathrm{Cu}}} \frac{I_{\mathrm{sw}}}{l_{\mathrm{coil}}} \int_{-\infty}^{\infty} \frac{\partial \phi}{\partial x_3}(x_1, x_2, x_3) \, \mathrm{d}x_3 \, \mathrm{d}\Gamma_{\mathrm{Cu}} = 0 \,, \tag{5.16}$$

where $\Omega = (-\frac{w}{2}, \frac{w}{2}) \times (-\frac{h}{2}, \frac{h}{2}) \times (-\infty, \infty)$ and ϕ is smooth with compact support in Ω. The same holds for $\boldsymbol{J}_{\mathrm{AC}}$, so that $\operatorname{div} \boldsymbol{J}_{\mathrm{AC}} = 0$, and overall $\operatorname{div} \boldsymbol{J} = 0$. Note that, as the expression above suggests, $\operatorname{div} \boldsymbol{J} = 0$ also if the current is independent of x_3 but not necessarily evenly distributed along the interface (i.e. $\frac{I_{\mathrm{sw}}}{l_{\mathrm{coil}}}$ is replaced by $J_{\mathrm{sw}}(x_1, x_2)$ with $(x_1, x_2) \in \Gamma_{\mathrm{Cu}}$ and the same with $\frac{I_{\mathrm{AC}}}{l_{\mathrm{coil}}}$). Conservation of charge then results in $\frac{\partial \rho_{\mathrm{ch}}}{\partial t} = -\operatorname{div} \boldsymbol{J} = 0$, so charge is conserved at $\rho_{\mathrm{ch}}(0)$. Typically, $\rho_{\mathrm{ch}}(0) = 0$ too, since ρ_{ch} is assumed to have only the frequency components in $\boldsymbol{J}$. In any case, this implies that in the frequency domain, $\rho_{\mathrm{ch}}(\omega_{\mathrm{AC}}) = 0$ and $\rho_{\mathrm{ch}}(\omega_{\mathrm{sw}}) = 0$, and leads to the conclusion that at those frequencies charge density (from a Eulerian standpoint) vanishes despite a current flowing. Physically this can be explained by taking a fixed control volume with overall zero charge (equal number of positive and negative charges), so $\rho_{\mathrm{ch}} = 0$, and considering an electron flowing into the volume at the same time as one flows out of the volume. Thus, the charge is maintained at $\rho_{\mathrm{ch}} = 0$ for every time step, while the current is being carried by the flow of electrons entering and leaving the control volume.



Magnetic field due to single line source

Source discretization of interface

Resulting Lorentz force

Figure 5.7: First, scattering theory is used to calculate the induced magnetic field contribution, $\boldsymbol{B}_{\boldsymbol{\xi}}$, at a point $\boldsymbol{x}$ due to a single infinite electric line source located at $\boldsymbol{\xi}$. The interface between the copper and the resin is discretized into a series of line sources (red points) and sample points (black points). Lastly, adding all the magnetic field source contributions at each sample point yields a final magnetic field which can be used to calculate a resulting Lorentz force acting at that interface.

Next, the magnetic and electric fields, $\boldsymbol{B}$ and $\boldsymbol{E}$, induced by the impressed current, $\boldsymbol{J}$, are described. For this, recall the fifth assumption, so the current can be approximated by a finite number of infinite electric line sources distributed along Γ_{Cu} as evenly as possible. These are shown in Figure 5.7 as small red dots. The magnetic and electric field induced by each of these electric line sources can be calculated using scattering theory as described in [16, Chapter 11]. With this in mind, place a source at $\boldsymbol{\xi} \in \Gamma_{\mathrm{Cu}}$ carrying an impressed current of the form $\boldsymbol{J}_{\boldsymbol{\xi}} = (0,0,1)^{\mathsf{T}} \frac{I}{N_{\mathrm{coil}}}$, where $I(t) = I_{\mathrm{AC}} \cos(\omega_{\mathrm{AC}} t) + I_{\mathrm{sw}} \cos(\omega_{\mathrm{sw}} t)$ and N_{coil} is the number of sources placed along the interface of each coil. In the frequency domain the frequency components are simply $I(\omega_{\mathrm{AC}}) = I_{\mathrm{AC}}$ and $I(\omega_{\mathrm{sw}}) = I_{\mathrm{sw}}$. At the frequency ω and at a point $\boldsymbol{x} \in \Omega$ (with same x_3 coordinate as $\boldsymbol{\xi}$), the magnetic and electric fields induced by the source are (see [16, Chapter 11]),

$$\begin{aligned} \boldsymbol{B}_{\boldsymbol{\xi}}(\boldsymbol{x},\omega) &= -\mathrm{i}\frac{I(\omega)}{N_{\mathrm{coil}}}\frac{\mu_p \beta_\omega}{4} H_1^{(2)}(\beta_\omega r(\boldsymbol{\xi})) \begin{pmatrix} -\sin(\varphi(\boldsymbol{\xi})) \\ \cos(\varphi(\boldsymbol{\xi})) \\ 0 \end{pmatrix}, \\ \boldsymbol{E}_{\boldsymbol{\xi}}(\boldsymbol{x},\omega) &= -\frac{I(\omega)}{N_{\mathrm{coil}}}\frac{\beta_\omega^2}{4\omega\varepsilon_p} H_0^{(2)}(\beta_\omega r(\boldsymbol{\xi})) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \end{aligned} \tag{5.17}$$

where μ_p is the permeability of the the medium, ε_p is the permittivity of the medium, $\beta_\omega = \omega\sqrt{\mu_p \varepsilon_p}$, and as shown in the left of Figure 5.7, $r(\boldsymbol{\xi}) = |\boldsymbol{\xi} - \boldsymbol{x}|$ and $\varphi(\boldsymbol{\xi})$ is the angle between $\boldsymbol{\xi}$ and $\boldsymbol{x}$ taking $\boldsymbol{\xi}$ as the origin. Here, $H_0^{(2)}$ and $H_1^{(2)}$ are Hankel functions of the second type. Next, let $\boldsymbol{x} \in \Omega$ be the same sample point, but this time consider the contribution of a set of sources located in the set Ξ, so that the electric and magnetic fields become,

$$\boldsymbol{B}(\boldsymbol{x},\omega) = \sum_{\boldsymbol{\xi} \in \Xi} \boldsymbol{B}_{\boldsymbol{\xi}}(\boldsymbol{x},\omega), \qquad \boldsymbol{E}(\boldsymbol{x},\omega) = \sum_{\boldsymbol{\xi} \in \Xi} \boldsymbol{E}_{\boldsymbol{\xi}}(\boldsymbol{x},\omega). \tag{5.18}$$

These fields can be computed at several sample points along the interface, preferably not coinciding with the source locations (to avoid numerical singularities), as shown in Figure 5.7 (the large black dots). Lastly, in the time domain, each component of $\boldsymbol{B}(\boldsymbol{x},t)$ and $\boldsymbol{E}(\boldsymbol{x},t)$ can be expressed as the sum of two (possibly shifted) cosine waves with angular frequencies ω_{AC} and ω_{sw}.

The last step is to calculate the Lorentz force [134], which in a continuum is given by,

$$\boldsymbol{f}(\boldsymbol{x},t) = \rho_{\mathrm{ch}}(\boldsymbol{x},t)\boldsymbol{E}(\boldsymbol{x},t) + \boldsymbol{J}(\boldsymbol{x},t) \times \boldsymbol{B}(\boldsymbol{x},t). \tag{5.19}$$

The Lorentz force only has distributional support at the interface between the copper and resin, Γ_{Cu}, which is where the current and charge flows. Since $\rho_{\mathrm{ch}}(\boldsymbol{x},t)=\rho_{\mathrm{ch}}(\boldsymbol{x},0)$ is time-independent, it follows the term $\rho_{\mathrm{ch}}\boldsymbol{E}$ produces forces in the same direction as the current and which have frequency components only at angular frequencies ω_{AC} and ω_{sw} (the frequency components in $\boldsymbol{E}$), unless $\rho_{\mathrm{ch}}(\boldsymbol{x},0)=0$. Meanwhile, the term $\boldsymbol{J}\times\boldsymbol{B}$ results in forces that are normal to the current direction and which have frequency components at angular frequencies 0, $\omega_{\mathrm{AC}}+\omega_{\mathrm{sw}}$, $\omega_{\mathrm{AC}}-\omega_{\mathrm{sw}}$, $2\omega_{\mathrm{AC}}$ and $2\omega_{\mathrm{sw}}$. The highest among all those frequencies is $2\omega_{\mathrm{sw}}$. In the most extreme case it happens that,

$$\boldsymbol{f}(\boldsymbol{x},2\omega_{\mathrm{sw}})=\boldsymbol{J}(\boldsymbol{x},\omega_{\mathrm{sw}})\times\boldsymbol{B}(\boldsymbol{x},\omega_{\mathrm{sw}})=\sum_{\boldsymbol{\xi}\in\Xi}\mathrm{i}\frac{I_{\mathrm{sw}}^2\mu_p\beta_{\omega_{\mathrm{sw}}}}{4N_{\mathrm{coil}}l_{\mathrm{coil}}}\delta_{\Gamma_{\mathrm{Cu}}}(\boldsymbol{x})H_1^{(2)}(\beta_{\omega_{\mathrm{sw}}}r(\boldsymbol{\xi}))\begin{pmatrix}\cos(\varphi(\boldsymbol{\xi}))\\ \sin(\varphi(\boldsymbol{\xi}))\\ 0\end{pmatrix}, \quad (5.20)$$

with $\beta_{\omega_{\mathrm{sw}}}=\omega_{\mathrm{sw}}\sqrt{\mu_p\varepsilon_p}$. This will be the underlying assumption in this work, because we are interested at looking at the behavior of the materials at very high frequencies. Potential improvements to this model would be to actually consider all the frequency components explicitly (not only at $2\omega_{\mathrm{sw}}$), or even to relax the assumption of evenly distributed current, thereby partly incorporating the effects of the phenomenon known as current crowding [178, 208].

Thus, the viscoelasticity problem will be considered in the frequency domain solely at the highest frequency, $2\omega_{\mathrm{sw}}=2\pi\cdot 400000\,\mathrm{rad/s}$, where the loading will be given by the Lorentz force described above. A `MATLAB`® script solved for such Lorentz force at each of the sample points along the interface, and the results are shown in the right of Figure 5.7 (with $N_{\mathrm{coil}}=110$). The electromagnetic properties used were the permeability (of copper), $\mu_p=1.26\cdot 10^{-6}\,\mathrm{N/A^2}$, and the permittivity (taken as vacuum), $\varepsilon_p=8.854\cdot 10^{-12}\,\mathrm{C^2/(N{\cdot}m^2)}$. With these parameters, the resulting surface Lorentz force had a magnitude in the order of $\mathrm{N/mm^2}$.

5.5.3 Results

The broken variational formulation in (5.14) solving the linear viscoelasticity equations once again was discretized using an adaptive DPG method as in Section 5.4.2 with the relevant parameters, $p=2$ and $\Delta p=1$. The surface Lorentz forces drove the problem and were computed as described in the previous section. The results of the computations are illustrated in Figure 5.8 for each of the two resins at the time step where the highest stresses were observed.

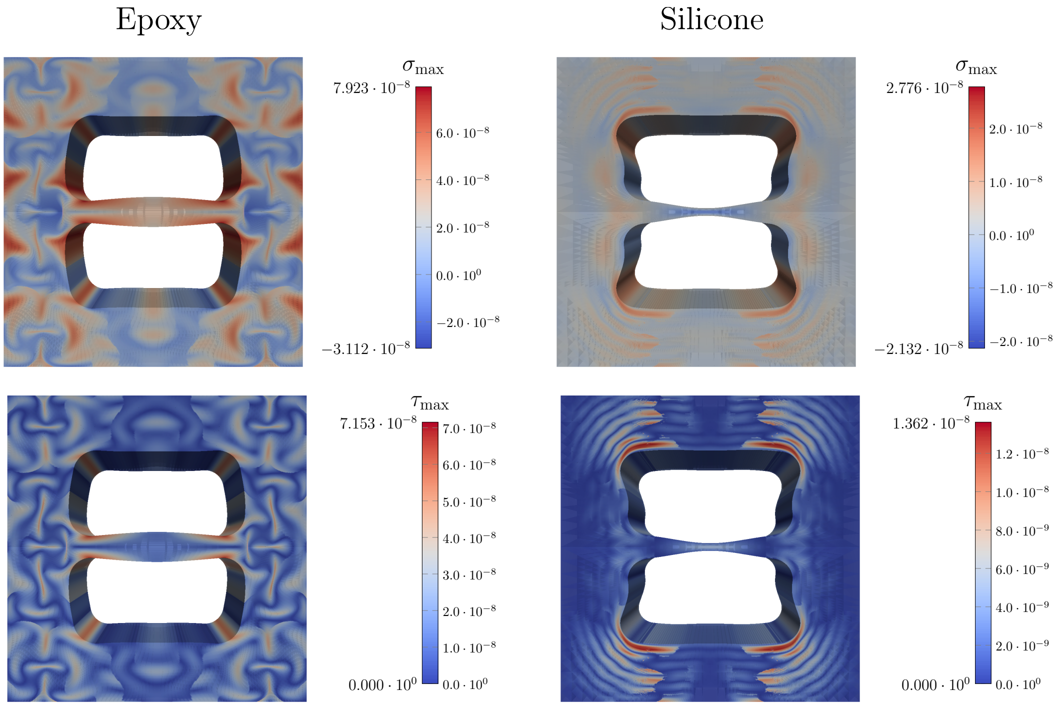


Figure 5.8: Maximum principal stress ($\sigma_{\max}$ in Pa) and maximum shear stress ($\tau_{\max}$ in Pa) in a form-wound coil (see Figure 5.2). The viscoelastic solution of (5.14) was obtained using an adaptive DPG method (with $p = 2$ and $\Delta p = 1$). Both an epoxy resin and a silicone resin were considered, while the coils were made of copper (not illustrated). The high frequency was $\omega = 2\pi{\cdot}400000\,\mathrm{rad/s}$ and the average temperature was $\bar{\theta} = 373\,\mathrm{K}$, with the material properties taken from Table 5.3. For visual purposes, the domain is warped by a scaled solution displacement.

Clearly, the results were completely different for both resins. Qualitatively, the stresses for the epoxy shows a spatially periodic pattern with the largest stresses located at different parts of the resin domain, including at the interface with the copper and near the outer boundaries, and even some large stresses occurring in the interior. The stresses in the silicone resin had an outward wave-like pattern and attained their maximum value at the outermost corners of the interface between the copper and the silicone. From a quantitative perspective, both the maximum principal stress and the maximum shear stress were higher in the epoxy than the silicone by factors of about

2.5 and 5 respectively. The localized solution features were successfully resolved using adaptivity via the DPG a posteriori error estimator. Finally, the work done in one second was calculated using (5.10) by taking $\boldsymbol{\sigma}$ as described in Section 5.4.2. For the epoxy resin it was $W_{\text{ins}}^{\text{Ep}} = 1.551 \cdot 10^{-24}\,\text{J}$ and for the silicone resin it was $W_{\text{ins}}^{\text{Si}} = 2.855 \cdot 10^{-24}\,\text{J}$.

5.6 Discussion

Three scenarios distinguished by a frequency regime and involving a "turn-to-turn" domain representing the basic unit of a form-wound coil were considered. The domain was assumed to be comprised of a viscoelastic resin surrounding two stacked copper coils as shown in Figure 5.2. Different physical phenomena were involved in each scenario. For very low frequencies of about 0.05 Hz thermal effects were assumed to be relevant, so the linear thermoviscoelasticity equations were solved subject to variations in temperature of 25 K in amplitude. At mid-range frequencies of about 2.8 kHz, the normal operation of the machine was assumed to activate vibrations of elliptic nature, referred to as ovalizations, which induced displacement boundary conditions in the order of 100 µm associated to the equations of linear viscoelasticity. At high frequencies of about 400 kHz, the interaction of the current at the switching frequency and its induced magnetic field were found to produce a surface Lorentz force in the order of N/mm at the interface of the copper and the resin, which was then used to load the linear viscoelasticity equations.

For all three scenarios the work done by the resin in one second was calculated. For an epoxy resin in the low-frequency scenario it was in the order of $10^{-5}\,\text{J}$, in the mid-range-frequency scenario it was around $10^{5}\,\text{J}$, and in the high-frequency scenario it was about $10^{-24}\,\text{J}$. The same is true for a silicone resin. Similar differences were found by looking at the order of magnitude of the stresses, where in the scenarios involving low and mid-range frequencies, the stresses were in the order of MPa, while in the high-frequency scenario the stresses were in the order of nPa. Thus, from a mechanical standpoint, this suggests that the Lorentz forces might have negligible effects, that the temperature variations have moderate effects, while the stator ovalizations may have the most significant effect. Having said that, it should be noted that the material properties vary significantly across different frequencies, so for example the resin may be much more sensitive to fatigue at high frequencies, and the effects of the Lorentz force at high frequencies may ultimately be relevant.

The models developed here are by no means exhaustive, and improvements as well as other physical effects can also be incorporated in the future. In particular, it would be more realistic to include body heat produced at the copper to simulate Joule heating more accurately in the low-frequency scenario. For the mid-range frequencies, it could be possible to include other vibration modes associated to the natural frequencies of the underlying stator structure, or even possibly calculate them directly. This, along with a more precise estimate of the vibration amplitude could produce more accurate results. Lastly, in the high-frequency scenario it could be possible to analyze other frequencies components of the Lorentz force besides the highest one, and maybe even incorporate an unevenly distributed current along the interface to better approximate the effects of current crowding.

Chapter 6

Discrete least-squares (DLS) finite element methods

This chapter is a reduced subset of a research article published by the author [160][§]. It is about a new family of finite element methods which is an outgrowth of DPG methods, and which we have called discrete least-squares (DLS) finite element methods. They are based on exploiting the rich algebraic structure provided by DPG discretizations. As their name suggests, they pose the overall problem as a discrete least-squares problem associated to an overdetermined system of equations which can be solved with QR-based algorithms. This has the advantage of having the conditioning reduced by a square root, so these methods are particularly useful in handling very ill-conditioned problems. This chapter is included in this dissertation because it shows how taking advantage of the algebraic structure in DPG methods, a new family of finite element methods was devised. The contributions of the author to the multi-authored article were doing some of the computations, participating in discussions about the numerical method and the mathematical derivations, and writing part of the manuscript.

6.1 Introduction

This chapter is meant to introduce a new family of finite element methods, which we refer to as discrete least-squares (DLS) finite element methods. They are an outgrowth of discontinuous Petrov-Galerkin (DPG) methods, and in particular of their attractive algebraic structure that underlies the discretizations of those methods. DLS methods can be associated with a discrete least-squares minimization corresponding to a rectangular overdetermined system coming from the discretization of a linear well-posed variational formulation. This is in contrast with typical finite element methods, such as Bubnov-Galerkin methods, which always solve a linear system associated to a square system. The approach proposed here is very useful for ill-conditioned problems, since

[§] Keith, B., Petrides, S., Fuentes, F., and Demkowicz, L. (2017c). Discrete least-squares finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 327:226–255.

DLS methods are associated to a condition number that is the square root of the condition number of the associated square system. This means that it is possible to observe a growth of the condition number of the order of $\mathcal{O}(h^{-1})$, instead of the more typical $\mathcal{O}(h^{-2})$, where h represents the element size. Also, when viewed as a discretization of DPG methods, DLS methods have a natural inherent a posteriori error estimator for use in adaptivity, and, more importantly, are crafted for numerical stability.

As expected, the DLS methods have many connections with other minimization problems both at the level of functional analysis and at the level of the discrete equations. In particular, they are related to least-squares finite element methods. Least-squares finite element methods have been demonstrated to be an auspicious class of methods for a wide variety of boundary value problems of engineering interest. These methods are attractive for many challenging problems because of their simple implementation, their numerical stability, and their built-in a posteriori error estimator. For a thorough study of least-squares finite element methods and the most significant references, we refer to [33]. Having said that, DLS methods have a much wider range of applicability, mostly due to the fact that they are identified with the discretization of arbitrary variational formulations, including those posed over non-symmetric functional settings. Indeed, DLS methods are identified with discretizations of general minimum residual methods (of which DPG methods are a subset). More generally, they apply to any system of the form

$$\begin{bmatrix} \mathsf{G} & \mathsf{B} \\ \mathsf{B}^* & 0 \end{bmatrix} \begin{bmatrix} \psi_r \\ \mathsf{u}_h \end{bmatrix} = \begin{bmatrix} \mathsf{l} \\ 0 \end{bmatrix} , \tag{6.1}$$

where for $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $\mathsf{G} \in \mathbb{F}^{M\times M}$ is Hermitian positive (semi-)definite and $\mathsf{B} \in \mathbb{F}^{M\times N}$ is rectangular, $M \geq N$. The discrete saddle-point system in (6.1) can appear in several non-DPG finite element settings as well [227, 75, 173, 59, 47, 177, 15], so we consider this entire class of methods from a comprehensive perspective. However, in practice, only when certain computations are localized does the method become much more effective, and this is precisely what happens when using DPG methods, which are associated to variational formulations with broken (discontinuous) test spaces. Thus, our principal examples will pertain to a specific class of DLS methods: those arising from DPG methods.

The chapter is organized as follows. In Section 6.2, DLS methods are described and some of their connections are elucidated. More specifically, it is mentioned that the conditioning of the system will be much improved when compared with competing alternatives. Furthermore, it will be shown how to perform static condensation of matrices in a manner consistent with DLS methods. The special assembly procedure associated to DLS methods will also be illustrated. In Section 6.3, we present several engaging examples. In particular, DLS methods are compared to both the classical Bubnov-Galerkin method as well the first-order system least-squares (FOSLS) finite element methods. Moreover, an example clearly shows the applicability of DLS methods by looking at a very ill-conditioned problem which other methods fail to accurately solve. Lastly, our final discussion is left to Section 6.4.

6.2 Discrete least-squares (DLS) finite element methods

6.2.1 Exploiting linear algebra

As alluded previously, we start with a system of the form

$$\begin{bmatrix} \mathsf{G} & \mathsf{B} \\ \mathsf{B}^* & 0 \end{bmatrix} \begin{bmatrix} \psi_r \\ \mathsf{u}_h \end{bmatrix} = \begin{bmatrix} \mathsf{l} \\ 0 \end{bmatrix}, \tag{6.2}$$

where $\mathsf{G} \in \mathbb{F}^{M\times M}$ is Hermitian positive (semi-)definite, $\mathsf{B} \in \mathbb{F}^{M\times N}$ is rectangular, $M \geq N$, and where $\mathbb{F}$ is either $\mathbb{C}$ or $\mathbb{R}$. The solution vector has a very particular structure due to the presence of 0, so the Schur complement to (6.2) becomes,

$$\mathsf{B}^*\mathsf{G}^{-1}\mathsf{B}\mathsf{u}_h = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{l} \quad \Leftrightarrow \quad \mathsf{A}\mathsf{u}_h = \mathsf{f}\,, \qquad \mathsf{A} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{B}\,, \quad \mathsf{f} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{l}\,. \tag{6.3}$$

Next, consider a factorization of the form $\mathsf{G} = \mathsf{W}\mathsf{W}^*$, for some $\mathsf{W} \in \mathbb{F}^{M\times M}$. There is an infinite number of solutions, W, to this equation provided $M > 1$, so for now consider one of those. Then, $\mathsf{G}^{-1} = \mathsf{W}^{-\mathsf{T}}\mathsf{W}^{-1}$

$$\begin{aligned} \mathsf{A}\mathsf{u}_h = \mathsf{f} \;\Leftrightarrow\; & \mathsf{B}^*\mathsf{W}^{-\mathsf{T}}\mathsf{W}^{-1}\mathsf{B}\mathsf{u}_h = \mathsf{B}^*\mathsf{W}^{-\mathsf{T}}\mathsf{W}^{-1}\mathsf{l} \;\Leftrightarrow\; (\mathsf{W}^{-1}\mathsf{B})^*(\mathsf{W}^{-1}\mathsf{B})\mathsf{u}_h = (\mathsf{W}^{-1}\mathsf{B})^*(\mathsf{W}^{-1}\mathsf{l}) \\ \Leftrightarrow\; & \mathsf{u}_h = \arg\min_{\mathsf{u}\in\mathbb{F}^N} \left\|\mathsf{W}^{-1}(\mathsf{B}\mathsf{u}-\mathsf{l})\right\|_2^2, \end{aligned} \tag{6.4}$$

where $\|\cdot\|_2$ is the usual Eucledian norm in $\mathbb{F}^M$. Thus, the original problem in (6.2) is equivalent to the discrete least-squares problem above. One particular solution, W, is the unique Cholesky factorization of $\mathsf{G} = \mathsf{L}\mathsf{L}^*$, where L is lower triangular. In this case, (6.4) is rewritten as

$$\mathsf{u}_h = \underset{\mathsf{u}\in\mathbb{F}^N}{\arg\min} \left\|\widetilde{\mathsf{B}}\mathsf{u} - \widetilde{\mathsf{l}}\right\|_2^2 , \qquad \widetilde{\mathsf{B}} = \mathsf{L}^{-1}\mathsf{B}\,, \quad \widetilde{\mathsf{l}} = \mathsf{L}^{-1}\mathsf{l}\,, \quad \mathsf{G} = \mathsf{L}\mathsf{L}^*\,. \tag{6.5}$$

As is common in discrete least-squares problems, (6.3) is referred to as the *normal equation* to the problem in (6.5). In the context of the discrete least-squares literature, the matrix $\mathsf{A} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{B}$ is referred to as the Gram or (Gramian) matrix, but in this chapter, as we will see soon, we safeguard the term of a Gram matrix to G, while A, B and $\widetilde{\mathsf{B}}$ are all called stiffness matrices, and f, l, and $\widetilde{\mathsf{l}}$ are called load vectors.

Remark 6.1. The choice of $\mathsf{W} = \mathsf{L}$ solving $\mathsf{G} = \mathsf{W}\mathsf{W}^*$, is made throughout this paper, but it should be noted that other choices are possible (including $\mathsf{W} = \mathsf{G}^{1/2}$). In fact, W can be interpreted as a change-of-basis matrix, and in the statistics and signal processing communities, the procedure described above is often known as “whitening” or “sphering”. We refer the reader to [162] to explore the benefits of the other possibilities for the change-of-basis matrices W in that context.

6.2.2 Connections with finite element methods

Up to this point only linear algebra has been referred to. However, as described throughout Chapter 2, especially Section 2.3 and Section 2.4, both (6.2) (see (2.44)) and (6.3) (see (2.28)) are equivalent discretizations of general minimum residual finite element methods. Indeed, for a given well-posed linear variational formulation, where a solution in a trial space, $\mathfrak{u} \in \mathcal{U}$, is sought such that,

$$b(\mathfrak{u},\mathfrak{v}) = \ell(\mathfrak{v}) \qquad \forall \mathfrak{v} \in \mathcal{V}\,, \tag{6.6}$$

its corresponding minimum residual discretization involving a discrete trial space $\mathcal{U}_h \subseteq \mathcal{U}$ and enriched test space $\mathcal{V}_r \subseteq \mathcal{V}$ is given by

$$\mathsf{B}^{\text{n-opt}}\mathsf{u}_h = \mathsf{l}^{\text{n-opt}}\,, \qquad \mathsf{A} = \mathsf{B}^{\text{n-opt}} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{B}\,, \quad \mathsf{f} = \mathsf{l}^{\text{n-opt}} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{l}\,, \tag{6.7}$$

where given bases $\{\mathfrak{u}_j\}_{j=1}^N$ and $\{\mathfrak{v}_i\}_{i=1}^M$ for $\mathcal{U}_h$ and $\mathcal{V}_r$, the enriched stiffness matrix and load are defined by $\mathsf{B}_{ij} = b(\mathfrak{u}_j,\mathfrak{v}_i)$ and $\mathsf{l}_i = \ell(\mathfrak{v}_i)$, while the Gram matrix $\mathsf{G}_{ij} = (\mathfrak{v}_i,\mathfrak{v}_j)_{\mathcal{V}}$ is a discretization of a Riesz map over $\mathcal{V}_r$. The solution $\mathfrak{u}_h = \sum_{j=1}^{N}(\mathsf{u}_h)_j\mathfrak{u}_j \in \mathcal{U}_h$ is solves the discrete variational formulation,

$$b(\mathfrak{u}_h, \mathfrak{v}_h) = \ell(\mathfrak{v}_h) \qquad \forall \mathfrak{v}_h \in \mathcal{V}_h = \mathcal{V}^{\text{n-opt}} = \operatorname{span}\left\{\mathfrak{v}_j^{\text{n-opt}}\right\}_{j=1}^{N}, \tag{6.8}$$

where the near-optimal test functions are defined by

$$\mathfrak{v}_j^{\text{n-opt}} = \sum_{i=1}^{M}(\mathsf{G}^{-1}\mathsf{B}\mathsf{e}_j)_i\mathfrak{v}_i \quad \forall j = 1, \dots, N, \tag{6.9}$$

with $\mathsf{e}_j \in \mathbb{F}^N$ being the Eucledian basis vectors, so that $(\mathsf{e}_j)_i = \delta_{ij}$ is the Kronecker delta. Clearly, (6.7) is the same as the normal equation in (6.3), where A and f are identified with the near-optimal stiffness matrix and load, $\mathsf{B}^{\text{n-opt}}$ and $\mathsf{l}^{\text{n-opt}}$. For the purposes of this chapter $\mathsf{A} = \mathsf{B}^{\text{n-opt}}$ and $\mathsf{f} = \mathsf{l}^{\text{n-opt}}$ and their label "near-optimal" is dropped when it is clear from the context, so they are simply referred to as the stiffness matrix and load. Thus, this establishes a direct connection between any minimum residual finite element method and a discrete least-squares problem associated with an overdetermined rectangular system, as written in (6.5). Hence, it makes sense to refer to those discretizations of variational formulations that result in a system like (6.2) or (6.3) as *discrete least-squares* (DLS) finite element methods.

As pointed out in the literature of DPG methods, minimum residual methods may be impractical from the computational standpoint, as they require finding the inverse, G^{-1}, at a global level. However, if the test spaces, $\mathcal{V}$, are broken along the mesh, this is equivalent to providing B, l and G with a decoupled structure. These are referred to as DPG methods. They allow to localize all the computations in A and f, which are ultimately assembled as in any other finite element method. In fact, as we will see shortly, $\widetilde{\mathsf{B}}$ and $\widetilde{\mathsf{l}}$ in the discrete least-squares setting (see (6.5)) can also be assembled from the local contributions, with the Cholesky factorization of the local Gram matrix being a viable calculation. Indeed, note that the observation made in (2.87) in the local context, is equivalent with the global factorization $\mathsf{G} = \mathsf{L}\mathsf{L}^*$ when such a decoupling of the test spaces is being contemplated. Therefore, DLS methods are the most practical when discretizing variational formulations with broken test spaces, like those considered in DPG methods. Moreover, in those cases they inherit other attractive features like residual-based a posteriori estimators for adaptivity and near-guaranteed discrete well-posedness (i.e. numerical stability). Having said that, there are other situations were they might be useful. Namely, when the global computation of G^{-1} is viable, such as when it is diagonal or has very sparse structure (and so does its inverse). This sometimes happens when the spaces lend themselves to being discretized by orthogonal basis functions [26, 114], as we will see in some of the numerical experiments later.

Regarding its connection with least-squares finite element methods, note that as mentioned in Section 2.5, least-squares methods are ideal minimum residual methods. More explicitly, they take the form,

$$\mathsf{A}^{\text{LS}}\mathsf{u}_h^{\text{opt}} = \mathsf{f}^{\text{LS}}\,, \qquad \mathsf{A}_{ij}^{\text{LS}} = \mathsf{B}_{ij}^{\text{opt}} = (\mathcal{L}\mathfrak{u}_j, \mathcal{L}\mathfrak{u}_i)_{L^2}\,, \quad \mathsf{f}_i^{\text{LS}} = \mathsf{l}_i^{\text{opt}} = (f, \mathcal{L}\mathfrak{u}_i)_{L^2}\,, \tag{6.10}$$

where $\mathfrak{u}_h^{\text{opt}} = \sum_{j=1}^N (\mathsf{u}_h^{\text{opt}})_j \mathfrak{u}_j \in \mathcal{U}_h \subseteq \mathcal{U}$ is the solution that exactly minimizes the residual,

$$\mathfrak{u}_h^{\text{opt}} = \underset{\delta\mathfrak{u}_h \in \mathcal{U}_h}{\arg\min} \|\mathcal{B}\delta\mathfrak{u}_h - \ell\|_{\mathcal{V}'}^2 = \|\mathcal{L}\delta\mathfrak{u}_h - f\|_{L^2}\,, \tag{6.11}$$

since loosely speaking $\mathcal{V} = L^2$. Here, $\mathcal{B}$ is the operator associated to the variational formulation, so that it is defined by $b(\mathfrak{u}, \mathfrak{v}) = \langle \mathcal{B}\mathfrak{u}, \mathfrak{v}\rangle_{(L^2)'\times L^2} = (\mathcal{L}\mathfrak{u}, \mathfrak{v})_{L^2}$ and easily identified with $\mathcal{L} : \mathcal{U} \to L^2$, while $\ell(\mathfrak{v}) = (f, \mathfrak{v})_{L^2}$ for some $f \in L^2$. In non-ideal minimum residual methods, the resulting near-optimal matrices A and f in (6.7) depend on an enriched test space $\mathcal{V}_r \subseteq \mathcal{V}$, so for all intents and purposes refer to them as A^r and f^r. As the enriched test space grows it will better approximate $\mathcal{V}$, and we would expect the same to occur with A^r and f^r as they approximate their optimal counterparts, which are precisely A^{LS} and f^{LS} (when the Riesz map is inverted exactly). Thus, $\mathsf{A}^r \to \mathsf{A}^{\text{LS}}$ and $\mathsf{f}^r \to \mathsf{f}^{\text{LS}}$ as $\mathcal{V}_r \subseteq \mathcal{V}$ grows. In some cases, like when $\mathcal{L}\mathcal{U}_h \subseteq \mathcal{V}_r$ (see Remark 2.6), the Riesz map is inverted exactly, so it will happen that $\mathsf{A}^r = \mathsf{A}^{\text{LS}}$ and $\mathsf{f}^r = \mathsf{f}$, and the least-squares finite element method will be exactly the same as the DPG discretization. However, the difference is that only A^r admits a DLS decoupling of the form $\mathsf{A}^r = (\widetilde{\mathsf{B}}^r)^*\widetilde{\mathsf{B}}^r$, while A^{LS} in (6.10) is a monolithic expression that does not obviously admit such a factorization. As we will see shortly, the decoupling will allow the problem to be solved with a better condition number, so when $\mathcal{L}\mathcal{U}_h \subseteq \mathcal{V}_r$ the DLS discretization provides a clear advantage (if ill-conditioning is a problem) while attaining the same solution as least-squares methods.

6.2.3 Solution algorithms

This section will remark on the very simple observation that the condition number resulting from globally solving (6.3) and (6.5) is completely different, even though both systems are equivalent when solved with infinite precision. First, the condition number of a rectangular matrix B is defined as $\mathrm{cond}(\mathsf{B}) = \frac{\sigma_{\max}}{\sigma_{\min}}$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the the maximum and minimum singular values of B respectively (i.e. the square roots of the maximum and minimum eigenvalues of $\mathsf{B}^*\mathsf{B}$). Therefore, it immediately follows that $\mathrm{cond}(\mathsf{B}^*\mathsf{B}) = \mathrm{cond}(\mathsf{B})^2$. The original system considered in (6.2) can be solved in many different ways. We mention here solution methods based on (6.3) and (6.5), but this list is not exhaustive and other alternatives exist in the literature [30, 131]. Our focus with DLS methods on this chapter will be to avoid the conventional approach of solving the normal equation in (6.3), and instead exploit the benefits of other alternatives that relate to the discrete least-squares formulation in (6.5).

6.2.3.1 Normal equations

Obviously, one could solve the normal equation in (6.3) with a direct solver. Note this is advantageous since $\mathsf{A} = \mathsf{B}^*\mathsf{G}^{-1}\mathsf{B}$ would be Hermitian positive definite, so A has a structure amenable to efficient linear solvers not usually available for many challenging problems. Moreover, from computational experience, we have found the normal equation has demonstrated to be adequate when solving systems that result from DPG discretizations. Indeed, in many reasonable circumstances, the round-off error in the solution from the associated linear solve cannot be expected to be nearly as large as the truncation error due to the finite element discretization, so in these cases solving the normal equation is reliable. From experience we can also say direct solvers for Hermitian positive definite systems are usually very fast. In fact, they are faster than the competing alternative that we are about to describe. Lastly, storing A and f usually requires less memory than storing $\widetilde{\mathsf{B}}$ and $\widetilde{\mathsf{l}}$ in (6.5).

Having said all this, solving the normal equation carries disadvantages too. Most notably, the $\mathrm{cond}(\mathsf{A})$ will grow quadratically with $\mathrm{cond}(\widetilde{\mathsf{B}})$ and, likewise, so will the upper bound on the round-off error of the normal equation solution. Furthermore, the scaling constant controlling the condition number of the stiffness matrix is often large in a first-order system setting, which is common in DPG discretizations, due to the additional equations and unknowns. Hence, there are many anticipated circumstances where other solution methods would be especially useful. Archetypal examples include, but are not limited to, singular perturbation problems, problems with large material contrast, high-order PDEs, penalty methods, and nonlinear problems where the linear approximation may become singular or very ill-conditioned. In summary, in some situations it is convenient to consider other approaches which deal explicitly with the matrices B, L, and l, and avoid the normal equation altogether.

6.2.3.2 Orthogonal decomposition methods

The most practical alternative to the normal equation when solving for u_h is to deal directly with the matrices $\widetilde{\mathsf{B}}$ and $\widetilde{\mathsf{l}}$ coming from the (sparsely weighted) linear least-squares problem in (6.5). The most common of these approaches is the orthogonalization algorithm called QR-factorization (Householder, Givens, modified Gram-Schmidt) first introduced for least-squares problems in [129]. Other direct approaches are SVD, complete orthogonal decomposition, and Peter-Wilkinson as well as various hybrid methods [30]. Each of these approaches are usually less efficient (and requires more storage) than solving the normal equation, but are often preferred because they are more numerically stable.

For our intended purposes, the matrix $\widetilde{\mathsf{B}}$ will be large and sparse, and so, because not all of the methods above are well suited for sparse matrices or amenable to parallel computing, we will focus only on the QR approach. As shown in various textbooks [30, 131, 224], the relative error in the solution from a least-squares QR solve is controlled by $\mathrm{cond}(\widetilde{\mathsf{B}}) + \rho(\widetilde{\mathsf{B}},\widetilde{\mathsf{l}}) \cdot \mathrm{cond}(\widetilde{\mathsf{B}})^2$, where

$$\rho(\widetilde{\mathsf{B}},\widetilde{\mathsf{l}}) = \frac{\|\widetilde{\mathsf{B}}\mathsf{u}_h - \widetilde{\mathsf{l}}\|_2}{\|\widetilde{\mathsf{B}}\|_2\|\mathsf{u}_h\|_2}\,, \tag{6.12}$$

and therefore depends upon the load vector. Due to the assumed well-posedness of the problem, the residual (the numerator in (6.12)) is expected to tend to zero as the mesh is refined. That is, $\rho(\widetilde{\mathsf{B}},\widetilde{\mathsf{l}}) \to 0$ as $h \to 0$, where h represents the mesh size. Indeed, the reader may observe that the $\rho(\widetilde{\mathsf{B}},\widetilde{\mathsf{l}})$ even vanishes if $\widetilde{\mathsf{l}} \in \mathsf{R}(\widetilde{\mathsf{B}})$. Of course, the validity of this convergence as well as its rate will be determined by the interpolation spaces used in the discretization. However, in many common scenarios the a priori bound can be proven to decrease at a rate of at least $\mathcal{O}(h)$. Indeed, for many cases we have in mind, this rate is of the form $\mathcal{O}(h^p)$ where $p \in \mathbb{N}$ is a polynomial order used in the trial space discretization. Therefore, the intuition is that the quadratic condition number term controlling the round-off error in a QR solve will be offset by a converging solution.

For instance, recall that $\mathrm{cond}(\widetilde{\mathsf{B}}) = \mathrm{cond}(\mathsf{A})^{1/2}$, since $\mathsf{A} = \widetilde{\mathsf{B}}^*\widetilde{\mathsf{B}}$. For many DPG discretizations and least-squares methods, $\mathrm{cond}(\mathsf{A})$ can be proved to grow as $\mathcal{O}(h^{-2})$ [133, 32, 33]. Thus, we would expect to see $\mathrm{cond}(\widetilde{\mathsf{B}})$ growing as $\mathcal{O}(h^{-1})$, which is much better. Moreover, if the residual converges to zero as described above, then $\rho(\widetilde{\mathsf{B}}, \widetilde{\mathsf{l}}) \cdot \mathrm{cond}(\widetilde{\mathsf{B}})^2$ can be no worse than $\mathcal{O}(h^{-1})$. In such conventional scenarios, the numerical sensitivity of the least-squares solution is controlled only by the inverse of the mesh size and will be far more accurate than any approach involving the normal equation. Precisely, in the typical first-order system scenario, we expect a QR-based algorithm will deliver an error bound of

$$\|\mathsf{e}_h\|_2 \leq \epsilon_{mach.}\|\mathsf{u}\|_2 C h^{-1} \,, \tag{6.13}$$

where e_h is the round-off error in the computation of the least-squares solution, $\epsilon_{mach.}$ is machine precision, and C is a mesh-independent constant.

Unfortunately, although QR-based algorithms are guaranteed to deliver a more accurate solution than solving the normal equation, there is potentially still a concealed obstacle. As many authors have pointed out, explicitly forming a product of two matrices before solving a least-squares problem posed with the matrix product $\widetilde{\mathsf{B}} = \mathsf{L}^{-1}\mathsf{B}$ is still not backwards stable [30]. Indeed, even when B is sparse and the matrix L is diagonal, yet extremely ill-conditioned, this can be a potential issue [31, 150]. Because of this concern, several algorithms exist in the numerical linear algebra literature for this very class of problems [191, 4, 137, 226]. Nevertheless, we believe that such precautions are unwarranted in all but the most exceptional problems that can be expected when implementing DLS methods. Indeed, we implicitly assume that the conditioning of L should not be badly behaved as the problem size grows, and in fact, in the cases where the Gram matrix comes from a DPG method (i.e. is block-diagonal) or from some other technique (perhaps a preconditioner estimate), G^{-1} can usually be generated using local element or patch information. This motivates us to assume that some measure of its local rank structure should stay constant or be uniformly bounded (with respect to the mesh size, h) as the mesh is refined. If this is true, preconditioning the Gram matrix, which itself often acts like a preconditioner, with its diagonal entries,

$$\mathsf{G} \mapsto \mathsf{D}^{-1/2}\mathsf{G}\mathsf{D}^{-1/2}\,, \quad \mathsf{B} \mapsto \mathsf{D}^{1/2}\mathsf{B}\,, \quad \mathsf{l} \mapsto \mathsf{D}^{1/2}\mathsf{l}\,, \tag{6.14}$$

where $\mathsf{D} = \mathrm{diag}(\mathsf{G})$, before locally factoring into $\mathsf{L}\mathsf{L}^*$ and performing back-substitution, should lead to robust results. This diagonal preconditioning procedure has been more than adequate in all of our experiments thus far. However, another possibility for improving the condition of the Gram matrix is the modified Lagrangian approach suggested in [130].

6.2.4 Static condensation

A common procedure which is often used to improve the solving time of linear systems is called static condensation. Here, the degrees of freedom associated with the element interior nodes (bubbles) are eliminated from the linear system. In practice, using a Schur complement procedure, small and independent blocks of the original stiffness matrix and load vectors are removed, and the original system is changed into a smaller-but-modified linear system with fewer unknowns. Often, this procedure of condensing, solving, and then recovering the global solution is much faster than solving the original system outright with standard means.

With this in mind, the idea is to develop the procedure of static condensation, but in the framework of a discrete least-squares problem. The first step is to, without loss of generality, separate the bubble and interface components of the relevant variables, so that $\mathsf{u}_h = \left[\begin{smallmatrix}\mathsf{u}_{\mathrm{bubb.}}\\ \mathsf{u}_{\mathrm{interf.}}\end{smallmatrix}\right]$ and $\mathsf{L}^{-1}\mathsf{B} = \widetilde{\mathsf{B}} = \big[\widetilde{\mathsf{B}}_{\mathrm{bubb.}}\,\widetilde{\mathsf{B}}_{\mathrm{interf.}}\big]$. Then, the normal equation in (6.3) becomes,

$$\begin{bmatrix} \widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}} & \widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{B}}_{\mathrm{interf.}} \\ \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}} & \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{B}}_{\mathrm{interf.}} \end{bmatrix} \begin{bmatrix} \mathsf{u}_{\mathrm{bubb.}} \\ \mathsf{u}_{\mathrm{interf.}} \end{bmatrix} = \widetilde{\mathsf{B}}^*\widetilde{\mathsf{B}}\mathsf{u}_h = \mathsf{A}\mathsf{u}_h = \widetilde{\mathsf{B}}^*\widetilde{\mathsf{l}} = \begin{bmatrix} \widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{l}} \\ \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{l}} \end{bmatrix}\,. \tag{6.15}$$

Then, writing the Schur complement for $\mathsf{u}_{\mathrm{interf.}}$ yields,

$$\begin{aligned} \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}(\mathbf{I} - \mathsf{P}_{\mathrm{bubb.}})\widetilde{\mathsf{B}}_{\mathrm{interf.}}\mathsf{u}_{\mathrm{interf.}} &= \big(\widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{B}}_{\mathrm{interf.}} - \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}}(\widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}})^{-1}\widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{B}}_{\mathrm{interf.}}\big)\mathsf{u}_{\mathrm{interf.}} \\ &= \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{l}} - \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}}(\widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{B}}_{\mathrm{bubb.}})^{-1}\widetilde{\mathsf{B}}^*_{\mathrm{bubb.}}\widetilde{\mathsf{l}} = \widetilde{\mathsf{B}}^*_{\mathrm{interf.}}(\mathbf{I} - \mathsf{P}_{\mathrm{bubb.}})\widetilde{\mathsf{l}}\,, \end{aligned} \tag{6.16}$$

where $\mathsf{P}_{\text{bubb.}} = \widetilde{\mathsf{B}}_{\text{bubb.}}\big(\widetilde{\mathsf{B}}_{\text{bubb.}}^*\widetilde{\mathsf{B}}_{\text{bubb.}}\big)^{-1}\widetilde{\mathsf{B}}_{\text{bubb.}}^*$. Next, note that $\mathsf{P}_{\text{bubb.}} = \mathsf{P}_{\text{bubb.}}^*$ is an orthogonal projection and it can easily be verified that $\mathbf{I} - \mathsf{P}_{\text{bubb.}} = (\mathbf{I} - \mathsf{P}_{\text{bubb.}})^2$, so that

$$\big((\mathbf{I} - \mathsf{P}_{\text{bubb.}})\widetilde{\mathsf{B}}_{\text{interf.}}\big)^*(\mathbf{I} - \mathsf{P}_{\text{bubb.}})\widetilde{\mathsf{B}}_{\text{interf.}}\mathsf{u}_{\text{interf.}} = \big((\mathbf{I} - \mathsf{P}_{\text{bubb.}})\widetilde{\mathsf{B}}_{\text{interf.}}\big)^*(\mathbf{I} - \mathsf{P}_{\text{bubb.}})\widetilde{\mathsf{l}}\,. \tag{6.17}$$

This can easily be seen to be equivalent to an overdetermined system that can be written in the form of a discrete least-squares problem as,

$$\mathsf{u}_{\text{interf.}} = \underset{\mathsf{u}_{\mathsf{S}}\in\mathbb{F}^{N_{\text{interf.}}}}{\arg\min}\ \big\|(\mathbf{I} - \mathsf{P}_{\text{bubb.}})(\widetilde{\mathsf{B}}_{\text{interf.}}\mathsf{u}_{\mathsf{S}} - \widetilde{\mathsf{l}})\big\|_2^2\,, \tag{6.18}$$

where $N_{\text{interf.}}$ is the dimension of the space where the interface components lie. Meanwhile, the remaining degrees of freedom can be recovered a posteriori by the expression,

$$\mathsf{u}_{\text{bubb.}} = \widetilde{\mathsf{B}}_{\text{bubb.}}^{+}\big(\widetilde{\mathsf{l}} - \widetilde{\mathsf{B}}_{\text{interf.}}\mathsf{u}_{\text{interf.}}\big)\,, \qquad \widetilde{\mathsf{B}}_{\text{bubb.}}^{+} = \big(\widetilde{\mathsf{B}}_{\text{bubb.}}^*\widetilde{\mathsf{B}}_{\text{bubb.}}\big)^{-1}\widetilde{\mathsf{B}}_{\text{bubb.}}^*\,. \tag{6.19}$$

The issue then becomes that of computing $\mathsf{P}_{\text{bubb.}}$ and $\widetilde{\mathsf{B}}_{\text{bubb.}}^{+}$, since they involve $\big(\widetilde{\mathsf{B}}_{\text{bubb.}}^*\widetilde{\mathsf{B}}_{\text{bubb.}}\big)^{-1}$. These have to be computed in a way which does not affect the global conditioning the QR-based methods already provide in the discrete least-squares framework. Otherwise, the conditioning would not be controlled as predicted in Section 6.2.3.2. To overcome this, introduce the full QR decomposition of $\widetilde{\mathsf{B}}_{\text{bubb.}}$, which can be computed stably,

$$\widetilde{\mathsf{B}}_{\text{bubb.}} = \begin{bmatrix}\mathsf{Q}_{\text{bubb.}} & \mathsf{Q}_{\text{interf.}}\end{bmatrix}\begin{bmatrix}\mathsf{R}_{\text{bubb.}}\\ 0\end{bmatrix}\,, \tag{6.20}$$

where $\begin{bmatrix}\mathsf{Q}_{\text{bubb.}} & \mathsf{Q}_{\text{interf.}}\end{bmatrix}$ is unitary and $\mathsf{R}_{\text{bubb.}}$ is upper-triangular. Then, it follows,

$$\mathsf{P}_{\text{bubb.}} = \mathsf{Q}_{\text{bubb.}}\mathsf{Q}_{\text{bubb.}}^*\,, \qquad \widetilde{\mathsf{B}}_{\text{bubb.}}^{+} = \mathsf{R}_{\text{bubb.}}^{-1}\mathsf{Q}_{\text{bubb.}}^*\,. \tag{6.21}$$

These computations can naturally be localized if necessary, so this concludes the procedure of static condensation in the context of discrete least-squares methods.

6.2.5 Assembly

In this section, we describe the construction of the global linear systems for DPG methods. As outlined in Section 6.2.3, we are primarily interested in two different procedures. First, directly assembling the normal equation in (6.3), and second, assembling the overdetermined discrete least-squares system in (6.5). To date, forming the normal equation has been the primary assembly procedure for DPG linear systems. The main advantages of this approach is that the assembly algorithm is identical to all traditional conforming finite element methods and it will involve the least storage. Moreover, many efficient direct and iterative solvers specialized for Hermitian positive definite systems can be employed. Nonetheless, there is an important disadvantage: the condition number of this global stiffness matrix A is the square of the condition number of the alternative, which is the (global) enriched stiffness matrix $\widetilde{\mathsf{B}}$. The enriched stiffness matrix is, however, rectangular, and so other solvers, which are generally more expensive, have to be used to solve the overdetermined linear system it is involved with. Be that as it may, this second approach can be applied to ill-conditioned problems, where forming the normal equation becomes an unsatisfactory option. We proceed by giving a brief description of the two different assembly procedures.

6.2.5.1 The normal equations

The assembly of the normal equation for the DPG method can easily be incorporated into any finite element code supporting exact sequence conforming shape functions [9, 114]. Recall that the DPG Gram matrix G is block diagonal. Therefore, it can be inverted element-wise and, therefore, G^{-1} is also block diagonal. Let B_K denote the enriched stiffness matrix for element K and l_K the corresponding load vector. Additionally, let G_K be the element Gram matrix. Then, the DPG element stiffness matrix A_K and load vector f_K are given by $\mathsf{A}_K = \mathsf{B}_K^*\mathsf{G}_K^{-1}\mathsf{B}_K$ and $\mathsf{f}_K = \mathsf{B}_K^*\mathsf{G}_K^{-1}\mathsf{l}_K$, respectively. Using the Cholesky factorization of $\mathsf{G}_K = \mathsf{L}_K\mathsf{L}_K^*$, we obtain

$$\mathsf{A}_K = \mathsf{B}_K^*\mathsf{G}_K^{-1}\mathsf{B}_K = (\mathsf{L}_K^{-1}\mathsf{B}_K)^*(\mathsf{L}_K^{-1}\mathsf{B}_K)\,, \qquad \mathsf{f}_K = \mathsf{B}_K^*\mathsf{G}_K^{-1}\mathsf{l}_K = (\mathsf{L}_K^{-1}\mathsf{B}_K)^*(\mathsf{L}_K^{-1}\mathsf{l}_K)\,. \tag{6.22}$$

We note that one may wish to precondition before the above operations, as in (6.14). The computation of the element DPG stiffness matrix and load vector is given by Algorithm 1.

The assembly of the global DPG stiffness matrix and load vector can be implemented by following the common algorithm of any standard finite element code [89, 103]. Note that there are two modifications that one should make to the element stiffness matrices before the

Algorithm 1 Element stiffness matrix and load vector for the DPG normal equation.

```
$\mathsf{L}_K \leftarrow$ Cholesky($\mathsf{G}_K$)
$\widetilde{\mathsf{B}}_K \leftarrow$ Triangular solve($\mathsf{L}_K \widetilde{\mathsf{B}}_K = \mathsf{B}_K$)
$\widetilde{\mathsf{l}}_K \leftarrow$ Triangular solve($\mathsf{L}_K \widetilde{\mathsf{l}}_K = \mathsf{l}_K$)
$\mathsf{A}_K \leftarrow \widetilde{\mathsf{B}}_K^* \widetilde{\mathsf{B}}_K$                      // DPG element stiffness matrix
$\mathsf{f}_K \leftarrow \widetilde{\mathsf{B}}_K^* \widetilde{\mathsf{l}}_K$                      // DPG element load vector
```

global assembly: account for Dirichlet boundary conditions; and (optional) accommodate degrees of freedom associated to constrained nodes for adaptive mesh refinement (hanging nodes are possibly created after adaptive h-refinements). We refer the reader to [89, 103] for detailed discussion on both of these modifications.

After pre-processing the element stiffness matrices, $\mathsf{A}_K \mapsto \mathsf{A}_K^{\mathtt{mod}}$, one should proceed with static condensation, $\mathsf{A}_K^{\mathtt{mod}} \mapsto \mathsf{A}_K^{\mathsf{c}}$, to reduce the complexity of the global system. For these square and symmetric matrices, this operation is described in Section 6.2.4. Additionally, as in standard FEM, the assembly is driven by the so-called "local-to-global connectivity maps". These maps assign to the local degrees of freedom their corresponding global degrees of freedom. The construction of these maps is based on the "donor strategy" and is implemented as in [89]. A description of the assembly procedure is given in Algorithm 2.

Algorithm 2 Assembly of DPG normal equation.

```
Initialize global stiffness matrix and load vector $\mathsf{A}$ and $\mathsf{f}$.
for $K \leftarrow 1$ to $N_K$ do                              // for each element in the mesh
    Compute $\mathsf{A}_K$ and $\mathsf{f}_K$                            // element stiffness matrix and load vector
    Compute $\mathsf{A}_K^{\mathtt{mod}}$ and $\mathsf{f}_K^{\mathtt{mod}}$                       // modified element matrix and load vector
    Compute $\mathsf{A}_K^{\mathsf{c}}$ and $\mathsf{f}_K^{\mathsf{c}}$                           // condensed element matrix and load vector
    Get $\mathtt{Con}_K$                                     // local-to-global connectivity map
    for $k_1 \leftarrow 1$ to $\mathtt{ndof}_K$ do                  // for each element degree of freedom (DOF)
        $i \leftarrow \mathtt{Con}_K(k_1)$                            // global index for local DOF
        $\mathsf{f}(i) \leftarrow \mathsf{f}(i) + \mathsf{f}_K(k_1)$                    // accumulate for the global load vector
        for $k_2 \leftarrow 1$ to $\mathtt{ndof}_K$ do              // for each element DOF
            $j \leftarrow \mathtt{Con}_K(k_2)$                        // global index for local DOF
            $\mathsf{A}(i,j) \leftarrow \mathsf{A}(i,j) + \mathsf{A}_K(k_1,k_2)$      // accumulate for the global stiffness matrix
```

6.2.5.2 The overdetermined system

Constructing the global overdetermined system requires some modifications to the assembly algorithms above. First, in order to deliver rectangular element stiffness matrices $\widetilde{\mathsf{B}}$ and load vectors $\widetilde{\mathsf{l}}$, one should only perform the first three steps of Algorithm 1. Note that the column size of the element stiffness matrix $\widetilde{\mathsf{B}}$ corresponds to the number of trial degrees of freedom and the row size to the number of test degrees of freedom. Similarly, the size of the load vector $\widetilde{\mathsf{l}}$ corresponds to the number of the test degrees of freedom.

As with square stiffness matrices, after the rectangular element matrices and load vectors have been computed, they need to be modified in order to accommodate Dirichlet boundary conditions and constrained nodes. The Dirichlet boundary conditions can be accounted for locally, like in the assembly algorithm for the normal equation, so that $\widetilde{\mathsf{B}}_K \mapsto \widetilde{\mathsf{B}}_K^{\mathtt{mod}}$ and $\widetilde{\mathsf{l}}_K \mapsto \widetilde{\mathsf{l}}_K^{\mathtt{mod}}$. For constrained nodes, the procedure is similar to the one for the normal equation, with the difference being that now the modifications are performed only on the trial space because the test space is broken. Note that there are no modifications needed for the load vector. The final local step is static condensation, $\widetilde{\mathsf{B}}_K^{\mathtt{mod}} \mapsto \widetilde{\mathsf{B}}_K^{\mathtt{c}}$ and $\widetilde{\mathsf{l}}_K^{\mathtt{mod}} \mapsto \widetilde{\mathsf{l}}_K^{\mathtt{c}}$ (see Section 6.2.4).

The global assembly algorithm then proceeds in a similar manner as for the normal equation. However, there is one important difference because the test space is broken: the need for accumulation of the contributions from different elements in both the (global) enriched stiffness matrix $\widetilde{\mathsf{B}}$ and the (global) enriched load vector $\widetilde{\mathsf{l}}$ has been eliminated. Therefore, in the global stiffness matrix, every row is independent. Note that this allows for a fully parallel assembly algorithm. This entire global assembly procedure is summarized in Algorithm 3.

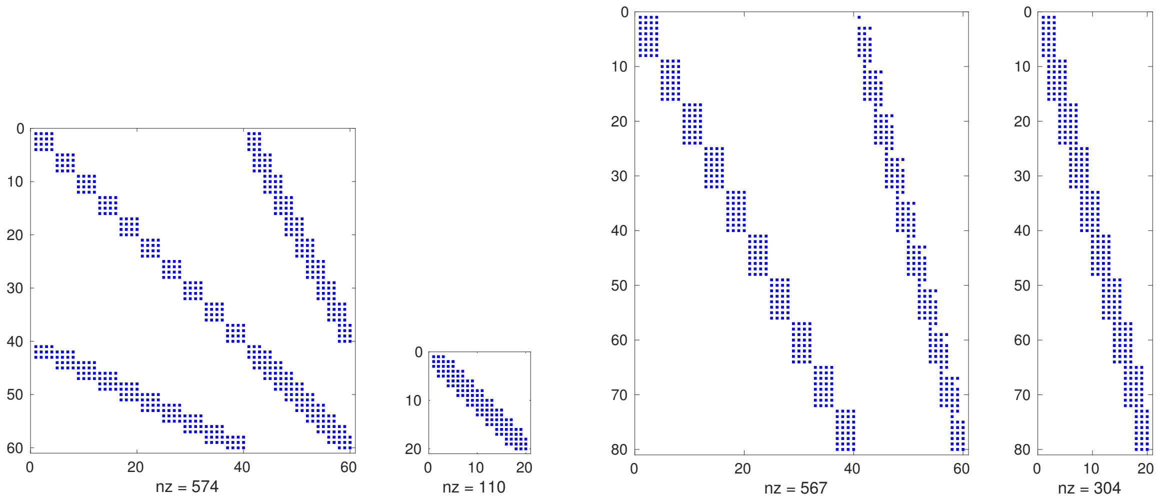
6.2.5.3 Comparison

For the DPG ultraweak variational formulation of Poisson's equation in one dimension (see Section 6.3.3), Figure 6.1 depicts the global DPG stiffness matrix for the normal equation (left in (a)) and the overdetermined system (left in (b)). The mesh used consisted of ten quadratic elements and the polynomial order used for the enriched test space was three. In a broken ultraweak formulation, continuity is enforced with the introduction of new interface unknowns, as will be

Algorithm 3 Assembly of DPG overdetermined system.

```
Initialize global stiffness matrix and load vector $\widetilde{\mathsf{B}}$ and $\widetilde{\mathsf{f}}$.
Initialize global test degreee of freedom (DOF) counter $i$.
for $K \leftarrow 1$ to $N_K$ do                          // for each element in the mesh
    Compute $\widetilde{\mathsf{B}}_K$ and $\widetilde{\mathsf{l}}_K$                 // element stiffness matrix and load vector
    Compute $\widetilde{\mathsf{B}}_K^{\mathtt{mod}}$ and $\widetilde{\mathsf{l}}_K^{\mathtt{mod}}$         // modified element matrix and load vector
    Compute $\widetilde{\mathsf{B}}_K^{\mathsf{c}}$ and $\widetilde{\mathsf{l}}_K^{\mathsf{c}}$             // condensed element matrix and load vector
    Get $\mathtt{Con}_K$                                // local-to-global connectivity map
    for $k_1 \leftarrow 1$ to $\mathtt{ndofT}_K$ do            // for each element test degree of freedom
        $i \leftarrow i+1$                              // global test DOF counter
        $\widetilde{\mathsf{l}}(i) \leftarrow \widetilde{\mathsf{l}}_K(k_1)$                    // fill in the global load vector
        for $k_2 \leftarrow 1$ to $\mathtt{ndof}_K$ do          // for each element DOF
            $j \leftarrow \mathtt{Con}_K(k_2)$                  // global index for local DOF
            $\widetilde{\mathsf{B}}(i,j) \leftarrow \widetilde{\mathsf{B}}_K(k_1,k_2)$          // fill in the global stiffness matrix
```

seen shortly. This explains the structure of the matrix for the total system seen in the left of Figure 6.1(a). After static condensation of the interior degrees of freedom, the resulting linear system involved only the interface unknowns, and, therefore, the matrix on the right of Figure 6.1(a) consists of only one band of overlapping blocks.



(a) A total (left) and condensed (right) systems.

(b) $\widetilde{\mathsf{B}}$ total (left) and condensed (right) systems.

Figure 6.1: DPG stiffness matrices A (normal equation) and $\widetilde{\mathsf{B}}$ (overdetermined system).

The situation is slightly different in the case of the discrete least-squares overdetermined system (see Figure 6.1(b)). For instance, because the test space is broken, there is no overlap between rows. Similar to the case of the normal equation, static condensation led to a linear system involving only the interface unknowns. However, here, the size of test space remained the same and therefore only the column dimension was reduced (see right of Figure 6.1(b)).

6.3 Results

The software used for all computations was the in-house `hp3d` which has support for SdR discretizations for all the element shapes [114] (see Section A.5 in Appendix A). The linear systems associated to the normal equation were solved with `MUMPS 5.0.1`, while the overdetermined system was solved with QR-based algorithms via the solver `qr_mumps 1.2`.We will now compare the condition number behavior, numerical sensitivity, discretization error, and round-off error incurred in DPG methods. As in [198], in each of our experiments, the stiffness matrix was diagonally preconditioned before the condition number was reported,

$$\mathsf{A} \mapsto \mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2}\,, \quad \mathsf{f} \mapsto \mathsf{D}^{-1/2}\mathsf{f}\,, \qquad \Leftrightarrow \qquad \widetilde{\mathsf{B}} \mapsto \widetilde{\mathsf{B}}\mathsf{D}^{-1/2}\,, \quad \widetilde{\mathsf{l}} \mapsto \widetilde{\mathsf{l}}\,, \tag{6.23}$$

where $\mathsf{D} = \operatorname{diag}(\mathsf{A})$. The cost of this procedure is computationally negligible and it is common practice to scale the matrix in this way before iterative solution methods. Meanwhile, it is performed implicitly in most direct solvers. Therefore, we presume no offense in this action. For additional perspective on several topics we do not cover, related to the condition number of DPG stiffness matrices, but with a focus on Stokes equation, we refer the interested reader to [198, Chapter 9].

6.3.1 FOSLS vs. DLS

For illustration, consider Poisson's equation in $\mathbb{R}^2$ with body force f and Dirichlet boundary condition given by $u^{\partial\Omega} \in H^{1/2}(\partial\Omega)$. Here, the aim is to find $(u, \boldsymbol{\sigma}) \in (\widetilde{u} + H_0^1(\Omega)) \times \boldsymbol{H}(\operatorname{div}, \Omega)$ such that

$$\begin{cases} -\operatorname{div}\boldsymbol{\sigma} + \alpha u = f\,, \\ \quad \boldsymbol{\sigma} - \operatorname{grad} u = 0\,, \end{cases} \tag{6.24}$$

where, $\alpha = 0$, $f \in L^2(\Omega)$, and $\widetilde{u} \in H^1(\Omega)$ is an extension of $u^{\partial\Omega}$ to Ω, so that $\operatorname{tr}^{\Omega}_{\text{grad}}\widetilde{u} = u^{\partial\Omega} \in H^{1/2}(\partial\Omega)$.

The simplest variational formulation of (6.24) seeks $(u,\boldsymbol{\sigma})\in H_0^1(\Omega)\times\boldsymbol{H}(\mathrm{div},\Omega)=\mathcal{U}$ so

$$\begin{aligned} b\big((u,\boldsymbol{\sigma}),(v,\boldsymbol{\tau})\big)&=\ell\big((v,\boldsymbol{\tau})\big)\quad\forall(v,\boldsymbol{\tau})\in L^2(\Omega)\times\boldsymbol{L}^2(\Omega)=\mathcal{V}\,,\\ b\big((u,\boldsymbol{\sigma}),(v,\boldsymbol{\tau})\big)&=-(\mathrm{div}\,\boldsymbol{\sigma},v)_\Omega+(\alpha u,v)_\Omega+(\boldsymbol{\sigma},\boldsymbol{\tau})_\Omega-(\nabla u,\boldsymbol{\tau})_\Omega\,,\\ \ell\big((v,\boldsymbol{\tau})\big)&=(f,v)_\Omega-(\alpha\widetilde{u},v)_\Omega+(\nabla\widetilde{u},\boldsymbol{\tau})_\Omega\,. \end{aligned}\tag{6.25}$$

And, where $(\cdot,\cdot)_\Omega=(\cdot,\cdot)_{L^2(\Omega)}$, $\boldsymbol{L}^2(\Omega)=\big(L^2(\Omega)\big)^2$, and $(u+\widetilde{u},\boldsymbol{\sigma})$ is the solution to (6.24). This is known as the (first-order) "strong variational formulation" of the Poisson equation [55]. These equations can be solved using the first-order system least-squares (FOSLS) method [49, 50] by directly discretizing the trial space $\mathcal{U}=H_0^1(\Omega)\times\boldsymbol{H}(\mathrm{div},\Omega)$, as in (6.10). The appropriate discretization $\mathcal{U}_h\subseteq\mathcal{U}$ should be a compatible SdR discretization of order p (see Section A.5 in Appendix A), such as the ones found in [114] for triangles and quadrilaterals. A DLS discretization can be found by using the same finite-dimensional spaces $\mathcal{U}_h$ as in the FOSLS method above, as well as an L^2-conforming test space $\mathcal{V}_r$ derived from an SdR discretization of order $p+\Delta p$. Notice that, for quadrilaterals at the master element level, the test functions are in $Y^{p+\Delta p}=\mathcal{Q}^{p+\Delta p-1,p+\Delta p-1}$, where $\mathcal{Q}^{p,q}=\mathcal{P}^p(x)\otimes\mathcal{P}^q(y)$, whereas $\mathrm{div}\,\boldsymbol{\sigma}\in Y^p$ and $\boldsymbol{\sigma}-\nabla u\in(Y^{p+1})^2$. Thus, it follows that $\mathcal{B}(\mathcal{U}_h)\subset\mathcal{R}_\mathcal{V}\,\mathcal{V}_r$ whenever $\Delta p\geq 1$, and all the comments in Remark 2.6 and Section 6.2.2 would be valid.

Let $\Omega=(0,1)^2$ be partitioned by a uniform quadrilateral mesh of side length h, and assume $u(x,y)=\sin(\pi x)\sin(\pi y)$. Define $\hat{f}=0$ and $f=-\mathrm{div}(\nabla u)$. With the extension $\widetilde{u}=0$, we separately solved (6.24) using both least-squares (i.e. FOSLS) and DLS methods (with $\Delta p\geq 1$). Moreover, because of the uniformity of the mesh and from using L^2-orthogonal shape functions at the master element level [114], we produced a diagonal Gram matrix, $\mathsf{G}_{ij}=(\mathfrak{v}_i,\mathfrak{v}_j)_{\boldsymbol{L}^2(\Omega)}$.

The condition number we computed for each of the three possible stiffness matrices is presented in Figure 6.2. Note that the condition number of the FOSLS stiffness matrix A^{LS} was verified to grow as $\mathcal{O}(h^{-2})$ (same as with a DLS stiffness matrix A), but the condition number of the DLS enriched stiffness matrix $\widetilde{\mathsf{B}}$ was verified to grow only as $\mathcal{O}(h^{-1})$.

Recall (6.10) and (6.7). The FOSLS solution would be $\mathfrak{u}_h^{\mathrm{opt}}$, while that coming from the DLS discretization would be $\mathfrak{u}_{h,r}$. As expected from Remark 2.6 and Section 6.2.2, numerical results also

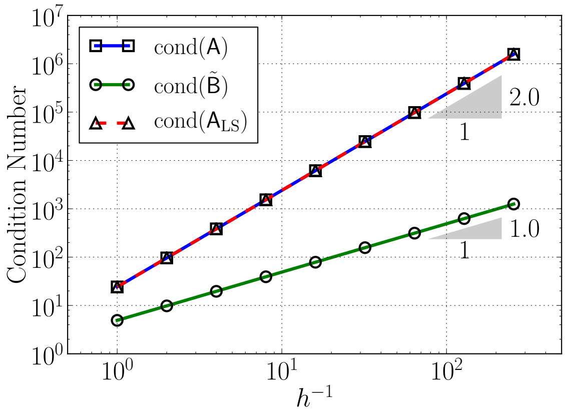
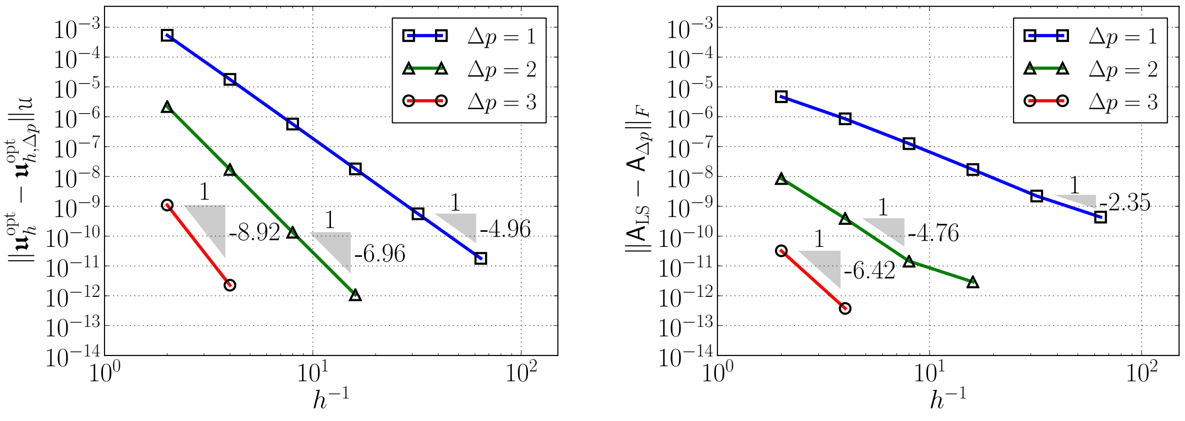


Figure 6.2: Comparison with Poisson's equation of the condition number of the FOSLS stiffness matrix A^{LS} to the condition number of the DLS stiffness matrices A and $\widetilde{\mathsf{B}}$ coming from the strong formulation, (6.25). Exact sequence polynomial order $p = 2$ for $\mathcal{U}_h$ and, in the DLS setting, $\Delta p = 1$. Observe that $\mathrm{cond}(\mathsf{A}^{\mathrm{LS}}) = \mathrm{cond}(\mathsf{A}) = \mathrm{cond}(\widetilde{\mathsf{B}})^2$. All reported results are for the statically condensed and diagonally preconditioned matrices.

confirmed that when $\Delta p \geq 1$ (so $\mathcal{B}(\mathcal{U}_h) \subset \mathcal{R}_{\mathcal{V}}\, \mathcal{V}_r$), $\mathfrak{u}_h^{\mathrm{opt}} = \mathfrak{u}_{h,r}$ and $\mathsf{A}^{\mathrm{LS}} = \mathsf{A}$, up to floating-point precision.



(a) Solution difference

(b) Matrix difference

Figure 6.3: Distance between the FOSLS and DLS solutions, $\mathfrak{u}_h^{\mathrm{opt}}$ and $\mathfrak{u}_{h,\Delta p}$, and the stiffness matrices A^{LS} and A. The solution $\mathfrak{u}_{h,\Delta p}$ was computed, in each experiment, by way of the normal equation (6.3). Exact sequence polynomial order $p = 2$ was used for $\mathcal{U}_h$, for all computations. The norms are defined as $\|(u, \boldsymbol{\sigma})\|_{\mathcal{U}}^2 = \|u\|_{H^1(\Omega)}^2 + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}^2$, whereas $\|\cdot\|_F^2$ is the Frobenius norm.

When $\mathcal{B}(\mathcal{U}_h) \not\subseteq \mathcal{R}_{\mathcal{V}} \mathcal{V}_r$, the solutions $\mathfrak{u}_h^{\mathrm{opt}}$ and $\mathfrak{u}_{h,r}$, and matrices A^{LS} and A will no longer be equal. Naturally, however, the distance between $\mathfrak{u}_{h,r}$ to $\mathfrak{u}_h^{\mathrm{opt}}$ is expected to decrease as $\mathcal{V}_r$ is enriched (i.e. as Δp is increased). Specifically, if the enriched test spaces are nested, $\mathcal{V}_{r_1} \subsetneq \mathcal{V}_{r_2} \subsetneq \mathcal{V}_{r_3} \subsetneq \cdots$, we expect $\|\mathfrak{u}_h^{\mathrm{opt}} - \mathfrak{u}_{h,r_k}\|_{\mathcal{U}} \to 0$ as $k \in \mathbb{N}$ increases. Indeed, this was observed when considering $\alpha(x,y) = \sin(\pi x)\sin(\pi y)$ in (6.24) and (6.25) and comparing the FOSLS solution, $\mathfrak{u}_h^{\mathrm{opt}}$, to the DLS normal equation solution, $\mathfrak{u}_{h,\Delta p}$, for increasing values of Δp. The results in Figure 6.3(a) show that the rate of h-convergence between the two discrete solutions grows with Δp. Moreover, Figure 6.3(b) shows that the matrix A converges to A^{LS}. These numerical results suggest that the error incurred in discretizing $\mathcal{V}$ to $\mathcal{V}_r$ can be made very small, and we expect this to be true even with non-trivial variational formulations (i.e. after integrating by parts).

6.3.2 Bubnov-Galerkin vs. DLS

In order to explore less trivial variational formulations, consider the broken primal formulation of the Poisson equation in (6.24) (with $\alpha = 0$) [94]. In this setting, we seek a solution $(u, \hat{\sigma}_{\mathbf{n}}) \in H_0^1(\Omega) \times H^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$ (see Appendix A for definitions of the trace Sobolev spaces) such that

$$\begin{aligned}
b\big((u,\hat{\sigma}_{\mathbf{n}}),v\big) &= \ell(v)\,, \quad \forall v \in H^1(\mathcal{T}) = \mathcal{V}\,,\\
b\big((u,\hat{\sigma}_{n}),v\big) &= (\nabla u, \nabla v)_{\mathcal{T}} - \langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}_{\mathrm{grad}}^{\mathcal{T}} v\rangle_{\partial\mathcal{T}}\,, \qquad (6.26)\\
\ell(v) &= (f,v)_{\mathcal{T}} - (\nabla\widetilde{u}, \nabla v)_{\mathcal{T}}\,.
\end{aligned}$$

Here, $H^1(\mathcal{T})$ is a broken Sobolev space with norm $\|v\|_{H^1(\mathcal{T})}^2$, and the restriction of each member of this space to any single $K \in \mathcal{T}$ is in $H^1(K)$. Likewise, $(\cdot,\cdot)_{\mathcal{T}} = \sum_{K\in\mathcal{T}}(\cdot,\cdot)_K$ and similarly with $\langle\cdot,\cdot\rangle_{\partial\mathcal{T}}$. This second pairing, $\langle\cdot,\cdot\rangle_{\partial\mathcal{T}}$, can be understood, intuitively, as a mesh-boundary integral, however, the inquisitive reader may wish to examine Appendix A for further detail. The exension $\widetilde{u} \in H^1(\Omega)$ in (6.26) is identical to that from the least-squares setting presented before.

For the discretization, let the trial space $\mathcal{U}_h$ be a compatible SdR discretization of order p. For the enriched test space $\mathcal{V}_r$, it is sufficient to choose an SdR discretization of order $p + \Delta p$. The enriched test space, $\mathcal{V}_r$, need not be continuous across elements.

Observe that (6.26) is similar to the standard Bubnov-Galerkin problem, in which, the aim is to find $u \in H_0^1(\Omega) = \mathcal{U}^{\mathrm{BG}}$ such that

$$\begin{gathered} b^{\mathrm{BG}}(u,v) = \ell^{\mathrm{BG}}(v) \quad \forall v \in H_0^1(\Omega) = \mathcal{V}^{\mathrm{BG}} = \mathcal{U}^{\mathrm{BG}}\,, \\ b^{\mathrm{BG}}(u,v) = (\nabla u, \nabla v)_\Omega\,, \qquad \ell^{\mathrm{BG}}(v) = (f,v)_\Omega - (\nabla \widetilde{u}, \nabla v)_\Omega\,. \end{gathered} \tag{6.27}$$

Here, both trial and test spaces need to be H^1-conforming (continuous) across the mesh.

As far as we are aware, the condition number of the DPG (or DLS) stiffness matrix A coming from (6.26) has never been derived analytically. We leave that work to another researcher and so do not derive it here, either. Nevertheless, we expect it to grow like h^{-2}, as the Bubnov-Galerkin stiffness matrix, A^{BG}, does [37]. This hypothesis was confirmed with experiments on a square domain $\Omega = (0,1)^2$ starting from a uniform mesh of four square elements and exact solution $u(x,y) = \sin(10\pi x)\sin(10\pi y)$ (see Figure 6.4). Likewise, similar to the least-squares scenario, $\mathrm{cond}(\widetilde{\mathsf{B}}) = \mathrm{cond}(\mathsf{A})^{1/2}$ was confirmed to grow only as h^{-1}, and so it eventually became less than the condition number of A_{BG}, which started out the smallest. Notably, in contrast to least-squares finite element methods, this shows that a first-order system formulation is not required to achieve $\mathcal{O}(h^{-2})$ (or even $\mathcal{O}(h^{-1})$) condition number growth with a DLS method.

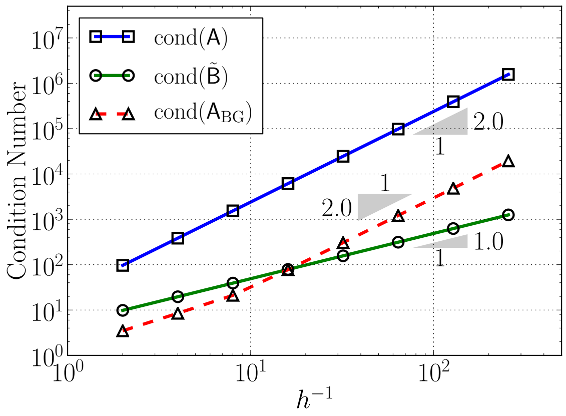


Figure 6.4: Comparison with Poisson's equation of the condition number of the Bubnov-Galerkin stiffness matrix A_{BG} to the condition number of the DPG stiffness matrices A and $\widetilde{\mathsf{B}}$ coming from the broken primal formulation, (6.26). Exact sequence polynomial order $p = 2$ for $\mathcal{U}_h$ and $\mathcal{U}_h^{\mathrm{BG}}$, and, in the DPG setting, $\Delta p = 1$. Observe that $\mathrm{cond}(\mathsf{A}_{\mathrm{BG}}) \neq \mathrm{cond}(\mathsf{A})$ and that $\mathrm{cond}(\widetilde{\mathsf{B}}) < \mathrm{cond}(\mathsf{A}_{\mathrm{BG}})$, eventually, for small enough h. All reported results are for the statically condensed and diagonally preconditioned matrices.

The optimal solution of the discrete minimum residual problem (6.7) coming from the variational formulation (6.26) would be $\mathfrak{u}_{h,\Delta p} = (u_{h,\Delta p}, (\hat{\sigma}_{\mathbf{n}})_{h,\Delta p})$. If we define the Bubnov-Galerkin problem's exact solution to be u^{BG}, and the exact solution of (6.26) to be $(u^{\mathrm{opt}}, \hat{\sigma}_{\mathbf{n}}^{\mathrm{opt}})$ then, it can be shown that $u^{\mathrm{BG}} = u^{\mathrm{opt}}$ [55]. However, for any given Δp, there is no reason to expect u_h^{BG} to be equal to $u_{h,\Delta p}^{\mathrm{opt}}$. Indeed, these two solutions did not always agree as is demonstrated in Figure 6.5. Nevertheless, the two different solutions clearly converged to each other, rapidly.

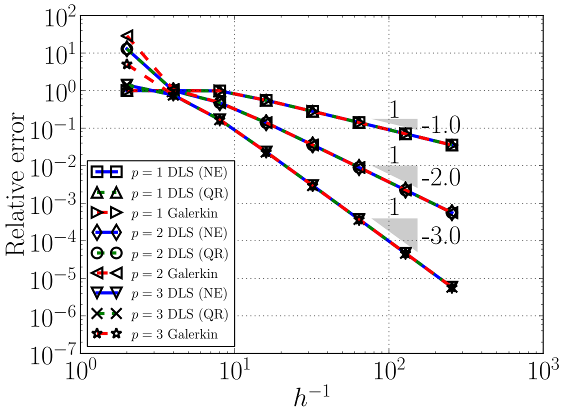


Figure 6.5: Relative error in the discrete solutions $u_{h,\Delta p}$ and u_h^{BG}. In the DPG solution, $\Delta p = 1$. Observe that, up to the smallest mesh size considered, the solutions $u_{h,\Delta p}$ obtained by the normal equation (6.3) do not differ noticeably from those computed with QR factorization of the least-squares problem (6.5).

6.3.3 Ultraweak DLS

Lastly, consider the broken ultraweak formulation of the Poisson equation in (6.24) (with $\alpha = 0$). This seeks a solution $(u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_{\mathbf{n}}) \in L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times H_0^{1/2}(\partial\mathcal{T}) \times H^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$ (again, see Appendix A for definitions of the Sobolev spaces) such that

$$\begin{aligned}
b\big((u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_{\mathbf{n}}), (v, \boldsymbol{\tau})\big) &= \ell\big((v, \boldsymbol{\tau})\big) \quad \forall (v, \boldsymbol{\tau}) \in H^1(\mathcal{T}) \times \boldsymbol{H}(\mathrm{div}, \mathcal{T}) = \mathcal{V}\,, \\
b\big((u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_{\mathbf{n}}), (v, \boldsymbol{\tau})\big) &= (\boldsymbol{\sigma}, \mathrm{grad}\, v + \boldsymbol{\tau})_{\mathcal{T}} + (u, \mathrm{div}\, \boldsymbol{\tau})_{\mathcal{T}} - \langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}_{\mathrm{grad}}^{\mathcal{T}} \rangle_{\partial\mathcal{T}} - \langle \hat{u}, \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{\tau} \rangle_{\partial\mathcal{T}}\,, \qquad (6.28) \\
\ell\big((v, \boldsymbol{\tau})\big) &= (f, v)_{\mathcal{T}} + \langle \check{u}, \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{\tau} \rangle_{\partial\mathcal{T}}\,.
\end{aligned}$$

Again, $H^1(\mathcal{T})$ and $\boldsymbol{H}(\mathrm{div}, \mathcal{T})$ are broken Sobolev spaces which essentially means that the restriction of each member to any single $K \in \mathcal{T}$ is in $H^1(K)$ and $\boldsymbol{H}(\mathrm{div}, K)$, respectively. We assume that the norm for $\mathcal{V}$ is then $\|\mathfrak{v}\|_{\mathcal{V}}^2 = \|v\|_{H^1(\mathcal{T})}^2 + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},\mathcal{T})}^2$, although other choices are possible [60]. Lastly, $\check{u} \in H^{1/2}(\partial\mathcal{T})$ is an extension of $\hat{u}^{\partial\Omega} \in H^{1/2}(\partial\Omega)$ to $\partial\mathcal{T}$.

As before, consider an SdR discretization of order p for $\mathcal{U}_h \subseteq \mathcal{U}$. Let the enriched test space $\mathcal{V}_r$ be an SdR discretization of order $p+\Delta p$. In this situation, it was proven in [133] that, provided $\Delta p \geq 2$ and that the mesh is made of triangles, the condition number of A would grow with h^{-2}. This $\mathcal{O}(h^{-2})$ growth was indeed confirmed with the hexahedral elements we were using as can be observed in Figure 6.6. More importantly, the enriched stiffness matrix $\widetilde{\mathsf{B}}$ was, therefore, verified to have a much improved condition number growth of $\mathcal{O}(h^{-1})$.

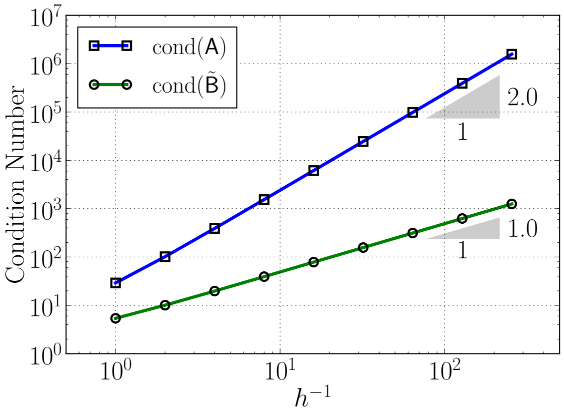
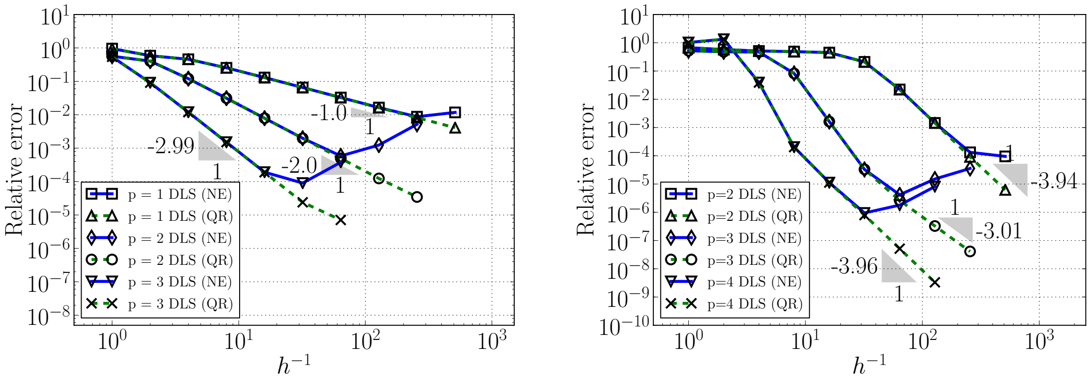


Figure 6.6: Condition number growth of the DPG stiffness matrices A and $\widetilde{\mathsf{B}}$ coming from broken ultraweak formulation (6.28). Here, $p = 2$ and $\Delta p = 1$. All reported results are for the statically condensed and diagonally preconditioned matrices.

6.3.4 Ill-conditioned failure study

In some circumstances, finite element stiffness matrices can be so poorly conditioned that the round-off error in solving the discrete equations will compete with, or even surpass, the truncation error coming from the method itself and interpolation spaces being used. In such scenarios, exploiting the overdetermined system of equations with QR is very attractive.

Due to time and space limitation, we will illustrate this behavior only for ultraweak DPG methods for problems of the form (6.24). We have chosen the ultraweak setting because it is the most actively researched variational setting for DPG methods, at this time. For the normal equation, we solved the system with `MUMPS 5.0.1` [168, 3]; and for the overdetermined system, we used `qr_mumps 1.2` [48]. Our results are reported in Figure 6.7.



(a) Poisson's equation (single precision).

(b) Linear acoustics (double precision).

Figure 6.7: Divergence of the discrete solution is observed for various polynomial orders p in two standard ultraweak DPG methods when the normal equation (6.3) is constructed, statically condensed, and then solved. Notice that, instead, when QR factorization was used to directly solve the (statically condensed) least-squares problem (6.5), the convergence of the discrete solution was maintained, at least for longer. In the $p = 2$ run for (b), had more refinements been performed, we expect that the anticipated rate of convergence (i.e. $2 \neq 3.94$) would have been recovered. $\Delta p = 1$ in all experiments.

First, we performed a *single-precision* floating-point computation with Poisson's equation (6.24) to verify that round-off error would eventually overwhelm truncation error, even in the most well-behaved of problems. Here, we chose $\Omega = (0,1)^2$ and an exact solution of the form $u(x,y) = x^2(1-x)^2y^2(1-y)^2$, and imposed (homogeneous) Dirichlet boundary conditions around the entire boundary $\partial\Omega$. Uniform h-refinements of quadrilateral elements, starting from a single element, were then performed until our computer ran out of memory. In Figure 6.7(a), we report the loss of convergence to this polynomial exact solution for $p = 1, 2, 3$ encountered after several mesh refinements when solving the normal equation. Notice that, however, for each polynomial order, `qr_mumps` applied to the overdetermined system continued to produce the expected rates of convergence after the normal equation approach failed.

In the $p = 2$ case, we can use Figure 6.6 to corroborate this outcome. For example, note that, the two solutions began to visibly diverge at the 7th mesh. Here, the relative error was just below 10^{-3}. Also, note that, from inspecting Figure 6.6, the corresponding condition number of the stiffness matrix A was approximately 10^5. Since machine single precision is approximately 10^{-7}, the round-off error would have been, at most, approximately 10^{-2}. It was, therefore, large enough to compete with, or surpass, the truncation error, and this is indeed what we began to see.

Our second example is with the linear acoustics problem in $\Omega = (0,1)^2$, which we chose to solve with the ultraweak formulation with broken test spaces, *near resonance*, in *double precision*. The equations of linear acoustics are:

$$\begin{cases} \mathrm{i}\omega p + \operatorname{div}\mathbf{u} = f\,, \\ \quad \mathrm{i}\omega\mathbf{u} + \nabla p = 0\,, \end{cases} \tag{6.29}$$

where p is the pressure and $\mathbf{u}$ is the velocity. We imposed a hard boundary for this problem. That is Neumann-type boundary conditions $\hat{g} \in H^{-1/2}(\partial\Omega)$ were specified on the entire boundary.

Similar to (6.28), we can derive the corresponding ultraweak formulation. In it, we seek a solution $(p, \mathbf{u}, \hat{p}, \hat{u}_{\mathbf{n}}) \in L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times H^{1/2}(\partial\mathcal{T}) \times H_0^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$ such that

$$\begin{aligned} b\big((p, \mathbf{u}, \hat{p}, \hat{u}_{\mathbf{n}}), (q, \boldsymbol{v})\big) &= \ell\big((q, \boldsymbol{v})\big) \qquad (q, \boldsymbol{v}) \in H^1(\mathcal{T}) \times \boldsymbol{H}(\mathrm{div}, \mathcal{T}) = \mathcal{V}\,, \\ b\big((p, \mathbf{u}, \hat{p}, \hat{u}_{\mathbf{n}}), (q, \boldsymbol{v})\big) &= -(p, \mathrm{i}\omega q + \operatorname{div}\boldsymbol{v}) - (\mathbf{u}, \mathrm{i}\omega\boldsymbol{v} + \nabla q) + \langle \hat{u}_{\mathbf{n}}, \mathrm{tr}_{\mathrm{grad}}^{\mathcal{T}} q\rangle_{\partial\mathcal{T}} + \langle \hat{p}, \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{v}\rangle_{\partial\mathcal{T}}\,, \\ \ell\big((q, \boldsymbol{v})\big) &= (f, q)_{\mathcal{T}} - \langle \check{u}_{\mathbf{n}}, q\rangle_{\partial\mathcal{T}}\,, \end{aligned} \tag{6.30}$$

where $\check{u}_{\mathbf{n}} \in H^{-1/2}(\partial\mathcal{T})$ is an extension of $\hat{g} \in H^{-1/2}(\partial\Omega)$ to $\partial\mathcal{T}$.

The discretization is also an SdR discretization of order p, similar to (6.28), except in $\mathbb{F} = \mathbb{C}$ instead of $\mathbb{R}$. We chose to solve the problem for $\omega = 0.5001 \cdot 2\pi$ which is very close to a resonance frequency. Note that, the first eigenvalue of the Laplacian in this setting is $\omega_1 = \pi$. Therefore, we can expect that the stiffness matrix will become very badly-conditioned as the mesh is refined. Using a Gaussian beam for the exact solution and a corresponding discrete extension $\check{u}_{\mathbf{n}}$ constructed using projection-based interpolation [90], we clearly received more robust convergence with the QR approach. The results are presented in Figure 6.7(b).

6.4 Discussion

In this chapter we presented a general framework for discrete least-squares finite element methods and illustrated features of this special class of methods. In particular every minimum residual method was posed equivalently as a discrete least-squares problem associated to an overdetermined rectangular system. This was done by exploiting the underlying linear algebra. Ultimately, this allowed to solve the overdetermined problem directly with QR-based algorithms instead of handling the normal equation, which is the conventional approach. In particular, this allows to solve the problem with a condition number that is the square root of that associated to the normal equation. DLS methods work best when they are applied to DPG discretizations (a subset of minimum residual methods) because these methods localize all the computations and allow for a much more practical method. When solving the overdetermined system, they are more costly, but are able to support very ill-conditioned problems.

More specifically, a procedure for static condensation was described in the discrete least-squares framework, which allows for the computations to remain accurate. Moreover, the assembly algorithms were discussed in detail. We provided several examples that elucidated the benefits of this new family of methods. In fact, the growth of the condition number of the associated stiffness matrix, after only diagonal preconditioning, was demonstrated to be $\mathcal{O}(h^{-1})$ when solving Poisson's equation in all of our examples. In other experiments which compared sensitivity to round-off error, we demonstrated the associated QR-factorization approach is particularly well-suited for ill-conditioned problems.

Although the key results we present are directed towards DPG methods, we have maintained a comprehensive and broad perspective in our presentation which has allowed us to make several connections with other methods in the literature. Extending this work to iterative solvers would be an interesting endeavor.

Chapter 7

Polygonal DPG (PolyDPG) methods with ultraweak formulations

This chapter is the content of a research publication by the author [229][¶]. It explains the use of ultraweak formulations and DPG methods to handle meshes with polygonal elements in 2D. The resulting methods are labeled polygonal DPG (PolyDPG) methods. They are high-order methods capable of discretizing arbitrary polygonal elements, including non-convex polygons. As other DPG methods, they are crafted to be stable and they have an built-in polygonal a posteriori error estimator to be used in adaptive refinements. Provided certain assumptions, PolyDPG methods can be shown to be numerically stable and convergent. This chapter is included in this dissertation because it shows how DPG methods are able to successfully exploit features in ultraweak formulations to produce novel numerical methods that are useful in certain applications. The contributions of the author to the multi-authored article included devising the method, advising related to the computations, producing the mathematical proofs, and writing of the manuscript.

7.1 Introduction

Numerical solutions of boundary value problems with meshes of general polytopes were first proposed by Wachspress [232], who introduced rational barycentric coordinates that formed a finite element basis over convex polygons, leading to a conforming finite element method (FEM) with new types of elements. Over the last two decades, there has been a growing collection of numerical methods using general polytopes which extend well beyond the original ideas of Wachspress. Among the reasons for this group of methods to thrive is a handful of advantages that polytopes offer over traditionally-shaped elements (simplices, hexahedra, etc.). These include: matching complex interfaces (see e.g. [175, 62]); greater flexibility to mesh complex geometries and their role as transition elements [216]; avoiding the limitations of parametric elements for highly distorted or

[¶] Vaziri Astaneh, A., Fuentes, F., Mora, J., and Demkowicz, L. (2018a). High-order polygonal discontinuous Petrov-Galerkin (PolyDPG) methods using ultraweak formulations. *Comput. Methods Appl. Mech. Engrg.*, 332:686–711.

ill-shaped elements (see e.g. [64, 166]); handling multiple hanging nodes in local h-refinements [218]; and allowing for greater deformations and less tendency to mesh-locking in incompressible media [65].

The features just mentioned give polytopal FEMs a wide range of applicability, especially where conventional methods do not fare well. In fact, they are useful for resolving problems involving the deformation of materials with heterogeneous microstructure [124], modeling complex materials like elastomers and biomaterials [65, 85], creating meshes where interface fitting is required [62], and modeling fractured media [22]. Promising results have also been obtained in crack propagation modeling [211, 167, 27, 29] and in topology optimization [219, 122, 5, 221], since polygonal meshes combine the ability to mesh complex geometries with a reasonable number of elements while reducing mesh-induced bias in particular directions (which occurs in structured meshes of triangles or quadrilaterals) [219, 167, 5].

Many methods still utilize different types of generalized barycentric coordinates (including some valid in non-convex polytopes), which have proliferated since Wachspress originally introduced them, as well as other choices of shape functions (see e.g. [28]). These methods are usually H^1-conforming Galerkin FEMs [216], but there are some extensions to mixed methods (see e.g. [65]). They mostly allow very flexible refinement schemes while avoiding constrained approximations [218], but they are typically limited by first order h-convergence. Some families of high-order shape functions have been proposed, but only for convex polytopes (see e.g. [196, 126]). As the barycentric coordinates are in general rational polynomials, another challenge is the choice of the quadrature scheme used for integration [174, 66].

Mimetic finite difference (MFD) methods are based on another discretization technique which also supports polygonal elements. The technique consists of designing discrete differential operators such that fundamental vector calculus identities and physical laws can be reproduced in a discrete context [163, 40, 39]. Later, the ideas of MFDs led to the development of virtual element methods (VEMs) [18]. In VEMs, appropriate spaces are tailored for each polytopal element, such that their functions have continuous and piecewise polynomial traces over the boundaries. In the lowest order case, the integrals over the cells can be computed exactly (i.e. up to machine precision) with quadrature points only on the boundary [169]. The power of VEMs lies partly in eliminating the need of explicitly constructing the shape functions in the element, and yet resulting in a FEM-like variational setting [21]. They are also high-order methods [17], and recent work has resulted in the construction of $\boldsymbol{H}(\mathrm{div})$- and $\boldsymbol{H}(\mathrm{curl})$-conforming spaces [20]. VEMs have been used for different problems like linear elasticity, plate bending, and second-order elliptic problems [19, 41, 21]. But it must be noted that VEMs need a problem-dependent stability operator to guarantee their convergence [169], and the solution at interior points of the elements is not accessible directly, so it has to be approximated [21].

Another method is the polytopal interior penalty *hp* discontinuous Galerkin (IPDG) method [52]. It is a nonconforming high-order method, which uses restrictions of standard FE spaces associated to a bounding box of each element. Due to its nonconformity, the method has a thorough but nonstandard equation-dependent error analysis, and like VEMs, it needs adding extra terms to ensure stability. Lastly, other recent methods include hybrid mimetic mixed methods [107, 106], PFEM-VEM [169], the weak Galerkin (WG) method [175, 176, 234], hybrid high-order (HHO) methods [104], and hybridizable discontinuous Galerkin (HDG) methods [71, 73]. More details on the historical development can be found in the thorough review [169].

The objective of this chapter is to present a completely new family of high-order methods termed polygonal discontinuous Petrov-Galerkin (PolyDPG) methods. They are based on so-called "broken" ultraweak variational formulations discretized using the discontinuous Petrov-Galerkin (DPG) methodology [95]. These formulations, despite being well-defined at the infinite-dimensional level, admit a very large degree of discontinuities in both the trial and test spaces, since their test spaces are broken (i.e. they may be discontinuous across element interfaces) and part of their trial spaces is in L^2. In fact, the only communication between elements happens through the so-called interface (or skeleton) variables that live on the element boundaries. These nonstandard formulations can be systematically discretized in a conforming fashion (i.e., with discrete trial and test spaces that are subspaces of the infinite-dimensional ones) and solved using the variationally versatile DPG methodology, which always produces a positive definite finite element stiffness matrix. The DPG methodology is essentially crafted to produce stability by using optimal test functions and without resorting to additional stabilization terms. DPG methods have been successfully used for equations involving numerical stability issues [75, 102, 60, 184, 160], and applied to various physical problems such as wave propagation [237, 132, 96, 194], transmission problems [145, 117], electromagnetism [55], elasticity [158, 36, 113, 111], fluid flow [202, 58, 109, 159] and optical fibers via Schrödinger's equation [97].

In this chapter we consider 2D problems, where the element boundaries are merely line segments, so high-order discretization of the interface variables is straightforward. As we will show, this makes the broken ultraweak formulations an ideal framework for defining polygonal elements, and it results in the conforming FEMs we refer to as PolyDPG methods. PolyDPG methods are competitive with other existing polygonal methods, since they arise from very different ideas and they inherit many advantages from the DPG methodology. For example, they can be easily generalized to different linear equations; they have a solid mathematical background in terms of proving stability and high-order convergence; they allow for discontinuous material properties while retaining stability; they result in positive definite stiffness matrices; and they carry a completely natural arbitrary-order a posteriori error estimator, which facilitates implementation of adaptive refinement strategies. The last feature is particularly desirable when combined with polygonal elements, because there is no need for the constrained approximation technology to treat hanging nodes, paving the way for use in applications like dynamic fracture [211, 167, 27, 29] and topology optimization [219, 122, 5, 221]. We complement this work by providing an open-source software in `MATLAB`®, also named `PolyDPG` [228].

The outline of the chapter is as follows. In Section 7.2 we describe a PolyDPG method for a model problem (Poisson's equation), along with the DPG solution scheme and the convergence theory (with the proof relegated to Appendix D). In Section 7.3 several illustrative examples are presented. High-order convergence for different p is verified for both convex and highly distorted non-convex elements. Then, a physically relevant problem involving discontinuous material properties along an arbitrary interface is solved. Finally, an adaptive refinement strategy is described, successfully implemented, and compared to traditional adaptive schemes. Our concluding discussion is presented in Section 7.4.

7.2 PolyDPG methods

Typical FEMs map elements from the actual physical space to a known fixed master element space corresponding to the same element type. For example, in 2D a general quadrilateral in $\mathbb{R}^2$ is mapped to a master quadrilateral (typically $(0,1)^2$ or $(-1,1)^2$). This requires defining a master element for each element type, which is possible for limited types of elements (e.g. quadrilaterals and triangles in 2D, or hexahedra, tetrahedra, triangular prisms and pyramids in 3D), but is usually nonviable when dealing with general polytopes. Thus, as with any polytopal FEM, the idea is to circumvent any master elements by shifting the focus directly to the physical space.

The main issue in doing so is satisfying inter-element continuity of the basis functions, which is required for discretizing Sobolev spaces such as H^1. This is partly resolved by using generalized barycentric coordinates, but these techniques are usually limited to first order methods (in terms of convergence), and it becomes difficult to discretize other Sobolev spaces such as $\boldsymbol{H}(\mathrm{curl})$ and $\boldsymbol{H}(\mathrm{div})$ even for the lowest order cases [63]. Indeed, even with the "traditional" pyramid element, having high-order discretizations for different spaces is challenging to achieve [185, 114, 125, 1], and so is the case for 2D non-affine quadrilaterals [6]. To overcome this, VEMs concentrate on the boundaries while nonconforming polytopal discontinuous methods, like IPDG, HHO, WG, and HDG (which are closely related [72, 71]), remove the continuity requirements altogether. However, all of these methods need to carefully add (equation-dependent) stabilization or penalty terms [18, 52, 104, 234, 73], and they must account for these in the error analysis, leading to a nonstandard theory of convergence [68].

As will be seen, the discontinuous Petrov-Galerkin (DPG) methodology is very general from a variational standpoint, so it is not limited to the traditional primal and mixed formulations. Thus, without sacrificing any desirable stability properties, it is able to discretize "broken" ultraweak variational formulations, which avoid most inter-element continuity requirements. The only continuity requirements are met by interface variables which live on the element boundaries. Technically speaking, the resulting method is still a conforming FEM, and the "standard" error analysis can be applied. This is very useful, because it allows to generalize the method to any well-posed linear equation formulated with traditional functional spaces (i.e. H^1, $\boldsymbol{H}(\mathrm{curl})$, $\boldsymbol{H}(\mathrm{div})$ and L^2 as in Appendix A).

In 2D, the polygonal element boundaries are simply line segments, so it is easy to define high-order discretizations along the mesh skeleton. Given that this is less trivial for polyhedra in 3D, we only analyze 2D problems in this introductory paper. We now proceed by introducing the model problem and its corresponding ultraweak formulations in the next section.

7.2.1 Model problem and ultraweak formulations

As a model problem, consider Poisson's equation coming from the steady-state heat equation in a (heterogeneous) domain $\Omega \subseteq \mathbb{R}^2$, where u is the temperature, $\boldsymbol{q}$ is the heat flux, $k > 0$ is the variable thermal conductivity, and r is the internal heat source,

$$-\operatorname{div}(k\nabla u) = r\,, \qquad \Leftrightarrow \qquad \begin{cases} \operatorname{div}\boldsymbol{q} = r\,, \\ \frac{1}{k}\boldsymbol{q} + \nabla u = 0\,. \end{cases} \tag{7.1}$$

Note that the equation can be written directly as a second order system (left) or as a first order system (right). For simplicity, we assume temperature boundary conditions along all of $\partial\Omega$, so that $u = g$ at $\partial\Omega$, where g is a known function.

To solve the equation using FEMs, a variational form is required, and in this respect, there are many possibilities. For now assume vanishing temperature boundary conditions so that $g = 0$. The ultraweak formulation is derived from the first-order system, so that both equations are integrated by parts to pass the derivatives to the test functions. The resulting ultraweak formulation seeks $(u, \boldsymbol{q}) = \mathfrak{u}_0 \in \mathcal{U}_0 = L^2(\Omega) \times \boldsymbol{L}^2(\Omega)$ satisfying

$$\begin{gathered} b_0(\mathfrak{u}_0, \mathfrak{v}_0) = \ell(\mathfrak{v}_0) \qquad \forall (v, \boldsymbol{\tau}) = \mathfrak{v}_0 \in \mathcal{V}_0 = H_0^1(\Omega) \times \boldsymbol{H}(\mathrm{div}, \Omega)\,, \\ b_0\big((u, \boldsymbol{q}), (v, \boldsymbol{\tau})\big) = -(\boldsymbol{q}, \nabla v)_\Omega + (\tfrac{1}{k}\boldsymbol{q}, \boldsymbol{\tau})_\Omega - (u, \operatorname{div}\boldsymbol{\tau})_\Omega, \qquad \ell\big((v, \boldsymbol{\tau})\big) = (r, v)_\Omega\,, \end{gathered} \tag{7.2}$$

where $\boldsymbol{L}^2(\Omega) = (L^2(\Omega))^2$. Clearly the trial and test spaces in this case are completely different, $\mathcal{U}_0 \neq \mathcal{V}_0$. Thus, to solve this system it is necessary to drift away from the traditional Bubnov-Galerkin method. As we will see, a discretization via minimum residual FEMs is a viable option. It is worth remarking that the primal and ultraweak formulations are mutually well-posed in the infinite-dimensional setting (see Chapter 3 and Appendix B). Since the primal formulation is known to be well-posed in view of the Lax-Milgram theorem and Poincaré's inequality, so is the ultraweak formulation. This guarantees the existence of a unique solution in the trial space satisfying a stability estimate.

The ultraweak formulation has copies of $L^2(\Omega)$ as a trial space, thus its discretization does not require satisfying any inter-element continuity, which is very desirable for polygons. However, all the difficulties are passed to the test space for which inter-element continuity requirements are essential. Fortunately, it is possible to remove these requirements in the test space as well, but at the cost of introducing interface variables, as has been shown in Chapter 2 and Chapter 3. Consider a mesh (i.e. an open partition), $\mathcal{T}$, of Ω comprised of (disjoint) elements $K \in \mathcal{T}$. Then, *element-wise*, multiply by broken test functions $(v, \boldsymbol{\tau}) = \mathfrak{v} \in \mathcal{V} = H^1(\mathcal{T}) \times \boldsymbol{H}(\mathrm{div}, \mathcal{T})$, integrate by parts, and sum across all elements. The result is very similar to the ultraweak formulation, but has new terms on the boundaries of the elements involving $u|_{\partial K}$ and $\boldsymbol{q}|_{\partial K} \cdot \hat{\mathbf{n}}_K$, where $\hat{\mathbf{n}}_K$ is the outward normal to the element K. These terms vanish if the test space is not broken (i.e. $\mathcal{V}_0$), but if we want $u \in L^2(\Omega)$ and $\boldsymbol{q} \in \boldsymbol{L}^2(\Omega)$, then the traces $u|_{\partial K}$ and $\boldsymbol{q}|_{\partial K} \cdot \mathbf{n}_K$ technically do not exist and to incorporate them it is necessary to add new interface variables in the spaces $H_0^{1/2}(\partial\mathcal{T})$ and $H^{-1/2}(\partial\mathcal{T})$. Therefore, the resulting broken ultraweak variational formulation seeks

$$\begin{gathered}(\mathfrak{u}_0, \hat{\mathfrak{u}}) = \mathfrak{u} \in \mathcal{U} = \mathcal{U}_0 \times \hat{\mathcal{U}}\,, \\ (u, \boldsymbol{q}) = \mathfrak{u}_0 \in \mathcal{U}_0 = L^2(\Omega) \times \boldsymbol{L}^2(\Omega)\,, \qquad (\hat{u}, \hat{q}_{\mathbf{n}}) = \hat{\mathfrak{u}} \in \hat{\mathcal{U}} = H_0^{1/2}(\partial\mathcal{T}) \times H^{-1/2}(\partial\mathcal{T})\,,\end{gathered} \tag{7.3}$$

such that

$$\begin{aligned} &b(\mathfrak{u}, \mathfrak{v}) = \ell(\mathfrak{v}) \qquad \forall (v, \boldsymbol{\tau}) = \mathfrak{v} \in \mathcal{V} = H^1(\mathcal{T}) \times \boldsymbol{H}(\mathrm{div}, \mathcal{T})\,, \\ &b\big((\mathfrak{u}_0, \hat{\mathfrak{u}}), \mathfrak{v}\big) = b_0(\mathfrak{u}_0, \mathfrak{v}) + \hat{b}(\hat{\mathfrak{u}}, \mathfrak{v})\,, \qquad \ell\big((v, \boldsymbol{\tau})\big) = (r, v)_{\mathcal{T}}\,, \\ &b_0\big((u, \boldsymbol{q}), (v, \boldsymbol{\tau})\big) = -(\boldsymbol{q}, \nabla v)_{\mathcal{T}} + (\tfrac{1}{k}\boldsymbol{q}, \boldsymbol{\tau})_{\mathcal{T}} - (u, \mathrm{div}\,\boldsymbol{\tau})_{\mathcal{T}}\,, \\ &\hat{b}\big((\hat{u}, \hat{q}_{\mathbf{n}}), (v, \boldsymbol{\tau})\big) = \langle \hat{q}_{\mathbf{n}}, \mathrm{tr}_{\mathrm{grad}}^{\mathcal{T}} v \rangle_{\partial\mathcal{T}} + \langle \hat{u}, \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{\tau} \rangle_{\partial\mathcal{T}}\,. \end{aligned} \tag{7.4}$$

This formulation can also be proved to be well-posed, with stability properties independent of the choice of the mesh (same technique as in Chapter2 and Chapter 3 via use of Theorem 2.1). With nontrivial boundary conditions, $g \neq 0$, simply consider $\ell(\mathfrak{v}) = (r, v)_{\mathcal{T}} - \langle \mathrm{tr}_{\mathrm{grad}}^{\mathcal{T}} \widetilde{g}, \mathrm{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{\tau} \rangle_{\partial\mathcal{T}}$ instead, where $\widetilde{g} \in H^1(\Omega)$ is an extension of $g \in H^{1/2}(\partial\Omega)$, and add $\widetilde{g}$ to the solution u of (7.4) to obtain the final temperature.

Despite looking intricate, the broken ultraweak variational formulation has the advantage of removing much of the inter-element compatibility conditions, since some of its trial variables are in $L^2(\Omega)$ and its test variables are discontinuous along the elements. The only inter-element compatibility is due to the interface variables, which reside solely on the element boundaries. In 2D, as we mentioned before, this is extremely convenient since the element boundaries are simply 1D line segments.

7.2.2 Choice of trial and test spaces

For details on the DPG methodology, the reader is invited to read Section 2.3. The usual choice of trial and test spaces from DPG methods are SdR discretizations (see Section A.5 in Appendix A), but these are not directly available for general polygonal elements. Therefore, a different choice must be made, and this is the content of this section.

The choice of trial and test spaces is important to establish the method's convergence. As mentioned before, strict inter-element compatibility requirements leaves very limited options. Particularly, the problem seems to be extremely complicated for general polygons with high-order discretizations. Fortunately, the $\mathcal{U}_0$ trial space component of the broken ultraweak formulation in (7.3) consists of copies of L^2, so its discretization can be discontinuous across the elements. Moreover, the test spaces are broken, so their discretization should be discontinuous across elements too. This is what makes the ultraweak formulation a natural candidate to develop a DPG method for polygonal elements. Indeed, this freedom allows one to create bases locally, disregarding the neighboring elements. In particular, bases may be defined by restriction (to the polygonal element of interest), as we will see next.

Our procedure is similar to that in [52] where a bounding box was utilized, but we use a bounding triangle instead. First, the centroid of the polygon and the furthest vertex from the centroid are determined. Next, a bounding circle centered at the centroid and passing through the furthest vertex is defined. Then, the bounding equilateral triangle inscribing the circle is computed such that one of its edge-midpoints is the polygon's furthest vertex. This is shown in Figure 7.1. Lastly, the "usual" high-order polynomial shape functions for the triangle are used and then restricted to the polygon. We use the term "usual" liberally, but to clarify, we include further details below.

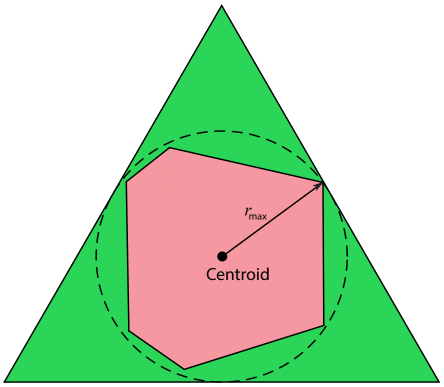


Figure 7.1: Bounding triangle of a polygonal element. The equilateral triangle is defined such that the bounding circle centered at the polygon's centroid is inscribed.

There are several spaces at the infinite-dimensional level which we want to discretize using this technique. Namely, the test space components, $H^1(\mathcal{T})$ and $\boldsymbol{H}(\mathrm{div},\mathcal{T})$, and the $\mathcal{U}_0$ trial space component, which may be represented by $L^2(\Omega)$. Following our technique, the procedure reduces to finding the local discretizations of $H^1(T_K)$, $\boldsymbol{H}(\mathrm{div},T_K)$ and $L^2(T_K)$, where T_K is the bounding triangle of the polygonal element $K\in\mathcal{T}$. These three spaces actually form a differential de Rham exact sequence, and it is convenient that their respective discretizations do too. For triangles, this is satisfied by the classical Nédélec sequence of the first type [103, 114],

$$\begin{tikzcd}
H^1(T_K) \arrow[r, "\mathrm{curl}"] \arrow[d, phantom, "\cup|"] & \boldsymbol{H}(\mathrm{div},T_K) \arrow[r, "\nabla\cdot"] \arrow[d, phantom, "\cup|"] & L^2(T_K) \arrow[d, phantom, "\cup|"] \\
\mathcal{P}^p(T_K) \arrow[r, "\mathrm{curl}"] & \mathcal{RT}^p(T_K) \arrow[r, "\nabla\cdot"] & \mathcal{P}^{p-1}(T_K)\,,
\end{tikzcd} \tag{7.5}$$

where $\mathcal{P}^p(T_K)$ are the polynomials in $\boldsymbol{x}=(x_1,x_2)$ of total order less than or equal to $p\in\mathbb{N}$, the 2D Raviart-Thomas space is $\mathcal{RT}^p(T_K)=(\mathcal{P}^{p-1}(T_K))^2+\boldsymbol{x}\mathcal{P}^{p-1}(T_K)$ (a rotation of the 2D Nédélec space), and the 2D scalar-to-vector curl operator is defined as $\mathrm{curl}(u)=\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)\nabla u$ for any $u\in H^1(T_K)$. Notice that the parameter p represents the order of the discrete *sequence* and does not necessarily coincide with the order of the polynomials of a particular discretization. For example if $p = 3$, the discretization of $L^2(T_K)$ are the polynomials of at most total order $p - 1 = 2$. Instead, the parameter p is intended to coincide with the order of h-convergence.

This sequence has many desirable properties, and precisely because of these, we prefer to use a bounding triangle instead of a bounding box. In particular, the spaces are invariant under affine transformations (the spaces remain the same even if the bounding triangle is arbitrarily rotated about the polygon centroid); the overall drop of polynomial order across the sequence is one (from $\mathcal{P}^p(T_K)$ to $\mathcal{P}^{p-1}(T_K)$); the approximation properties are suitable (see Appendix A); and they are the smallest possible spaces with all these properties (see [9, §3.4]).

Having said that, a similar procedure can be carried out for a bounding box, Q_K of $K \in \mathcal{T}$, where the spaces become

$$\begin{tikzcd}
H^1(Q_K) \arrow[r, "\mathrm{curl}"] & \boldsymbol{H}(\mathrm{div}, Q_K) \arrow[r, "\nabla\cdot"] & L^2(Q_K) \\
\mathcal{Q}^{p,p}(Q_K) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\mathrm{curl}"] & \mathcal{Q}^{p,p-1}(Q_K) \times \mathcal{Q}^{p-1,p}(Q_K) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\nabla\cdot"] & \mathcal{Q}^{p-1,p-1}(Q_K)\,, \arrow[u, phantom, sloped, "\subseteq"]
\end{tikzcd} \tag{7.6}$$

with $\mathcal{Q}^{p,q}(Q_K) = \mathcal{P}^p(x_1) \otimes \mathcal{P}^q(x_2)$.

In either case, the final spaces for the polygon $K \subseteq T_K$ (or $K \subseteq Q_K$) are defined by restricting the domain to $K \in \mathcal{T}$, so we denote them by $\mathcal{P}^p(K)$ and $\mathcal{RT}^p(K)$ (or $\mathcal{Q}^{p,p}(K)$) instead.

The only remaining spaces to specify are those of the interface variables lying in the $\hat{\mathcal{U}}$ trial space component (see (7.3)). These can also be deduced using the same philosophy of exact sequences, but utilizing the traces instead. Indeed, the spaces $H_0^{1/2}(\partial\mathcal{T})$ and $H^{-1/2}(\partial\mathcal{T})$ are merely $\mathcal{T}$-tuples of compatible traces of $H^1(K)$ and normal-traces of $\boldsymbol{H}(\mathrm{div}, K)$ respectively. If two elements of different type (a triangle and a quadrilateral) share an edge, the discrete spaces should be compatible across that edge. This is the case when considering the $H^1(K)$-discretizations of triangles and quadrilaterals: even though the discretizations themselves are different ($\mathcal{P}^p$ and $\mathcal{Q}^{p,p}$), their restrictions to edges are exactly the same, $\mathcal{P}^p(e)$, where e represents an edge parametrized linearly by t_e. The same occurs with the $\boldsymbol{H}(\mathrm{div}, K)$-discretizations, which have $\mathcal{P}^{p-1}(e)$ as normal-trace along the edges. Additionally, the $H^1(K)$-discretizations should be compatible at vertices. This is consistent with 1D discretizations of H^1 and L^2, which also form an exact sequence, but instead occurring along the boundary of each element and being edge-parametrized along all edges (see [114, §1.6] or definitions of SdR discretizations in Section A.5 of the Appendix A). This pattern should hold for arbitrary polygons as well. For this, let $\mathcal{E}(K)$ be the set of edges of a polygon $K \in \mathcal{T}$, and define the local discretizations,

$$\begin{aligned} \mathcal{P}^{p-1}(\partial K) &= \{\hat{w}_K \mid \hat{w}_K|_e \in \mathcal{P}^{p-1}(e), \forall e \in \mathcal{E}(K)\} \subseteq H^{-1/2}(\partial K)\,, \\ \mathcal{P}^p_C(\partial K) &= \mathcal{P}^p(\partial K) \cap C^0(\partial K) \subseteq H^{1/2}(\partial K)\,, \end{aligned} \tag{7.7}$$

where $C^0(\partial K)$ are the continuous functions in ∂K (the intersection ensures that values of neighboring edges coincide at a common vertex), and the infinite-dimensional local trace spaces are $H^{1/2}(\partial K) = \{\hat{u}_K = u|_{\partial K} \mid u \in H^1(K)\}$ and $H^{-1/2}(\partial K) = \{(\hat{q}_{\mathbf{n}})_K = \boldsymbol{q}|_{\partial K} \cdot \mathbf{n}_K \mid \boldsymbol{q} \in \boldsymbol{H}(\mathrm{div}, K)\}$.

Now we have enough information to actually globally define the discrete trial space. For a value of $p \in \mathbb{N}$, it is

$$\begin{aligned} \mathcal{U}_h = \big\{(u, \boldsymbol{q}, \hat{u}, \hat{q}_{\mathbf{n}}) \in \mathcal{U} \mid u|_K \in \mathcal{P}^{p-1}(K),\, \boldsymbol{q}|_K \in \big(\mathcal{P}^{p-1}(K)\big)^2, \\ \hat{u}_K \in \mathcal{P}^p_C(\partial K),\, (\hat{q}_{\mathbf{n}})_K \in \mathcal{P}^{p-1}(\partial K),\, \forall K \in \mathcal{T}\big\}\,. \end{aligned} \tag{7.8}$$

Notice that the condition $(u, \boldsymbol{q}, \hat{u}, \hat{q}_{\mathbf{n}}) \in \mathcal{U}$ (so $(\hat{u}, \hat{q}_{\mathbf{n}}) \in \hat{\mathcal{U}}$) implies that $\hat{u}$ vanishes at the boundaries, that $\hat{u}_{K_1}|_e = \hat{u}_{K_2}|_e$, and that $(\hat{q}_{\mathbf{n}})_{K_1}|_e = -(\hat{q}_{\mathbf{n}})_{K_2}|_e$, where e is a common edge between the elements K_1 and K_2. No such compatibility implications exist for $(u, \boldsymbol{q}) \in \mathcal{U}_0$.

For the enriched test space, the discretizations are chosen from a sequence of order $p + \Delta p$, and we say the space is p-enriched, so that

$$\mathcal{V}_r = \big\{(v, \boldsymbol{\tau}) \mid v|_K \in \mathcal{P}^{p+\Delta p_K}(K),\, \boldsymbol{\tau}|_K \in \mathcal{RT}^{p+\Delta p_K}(K),\, \forall K \in \mathcal{T}\big\}\,. \tag{7.9}$$

The notation Δp_K indicates that this value is element-dependent. In fact, recall that for minimum residual methods to work (Section 2.3), $M = \dim(\mathcal{V}_r) \geq \dim(\mathcal{U}_h) = N$, and if this analogous restriction on the dimensionality holds locally, then it will hold globally as well. Thus, Δp_K has to be chosen such that this condition holds. This is important for the polygonal element methods, because when a polygon has many sides, the size of the local trial space may be quite large and a large value of Δp_K must be chosen for that particular element.

To elaborate, consider an interior n-sided polygonal element K (so that $\partial K \cap \partial\Omega = \varnothing$). Its local trial and test space dimensions would be

$$\begin{aligned}
\dim\big(\mathcal{U}_h(K)\big) &= \overbrace{\tfrac{1}{2}p(p+1)}^{u|_K} + \overbrace{p(p+1)}^{\boldsymbol{q}|_K} + \overbrace{n+n(p-1)}^{\hat{u}_K} + \overbrace{np}^{(\hat{q}_{\mathbf{n}})_K}\,,\\
\dim\big(\mathcal{V}_r(K)\big) &= \underbrace{\tfrac{1}{2}(p+\Delta p_K+1)(p+\Delta p_K+2)}_{v|_K} + \underbrace{(p+\Delta p_K)(p+\Delta p_K+2)}_{\boldsymbol{\tau}|_K}\,.
\end{aligned} \tag{7.10}$$

Thus, for $p = 2$ and $n = 3$ (a triangle), $\dim(\mathcal{U}_h(K)) = 21$, so that a value of $\Delta p_K = 1$ is sufficient ($\dim(\mathcal{V}_r(K)) = 25$); but if $p = 2$ and $n = 8$ (an octagon), $\dim(\mathcal{U}_h(K)) = 41$, a value of at least $\Delta p_K = 3$ (so that $\dim(\mathcal{V}_r(K)) = 56$) is required. Having said that, sometimes for simplicity a valid value of Δp is chosen uniformly throughout the mesh (this is the case for all of our examples in Section 7.3).

To illustrate, some representative shape functions of the components of $\mathcal{U}_h(K)$ and $\mathcal{V}_r(K)$ are shown in Figure 7.2 for the different energy spaces and multiple values of p.

We refer to the high-order polygonal DPG method resulting from this choice of trial and enriched test spaces as a PolyDPG method for Poisson's equation. However, it can easily be generalized to ultraweak formulations coming from other linear equations (see Remark 7.2 later), so it is more appropriate to allude to a *family* of PolyDPG methods. Note that the methods seem to be very expensive due to the large number of variables in the trial space $\mathcal{U}_h$, but this is deceiving. In fact, all of the $\mathcal{U}_0$ trial space components can be statically condensed locally for ultraweak formulations, meaning that this part of the near-optimal stiffness matrix, $\mathsf{B}^{\text{n-opt}}$, can be effectively removed by taking Schur complements. Thus, the only remaining connectivity is that coming from the interface variables in $\hat{\mathcal{U}}$. So computationally speaking, solving with these variational formulations is not as costly as one might initially imagine. Lastly, it should be noted that usually the choice of Δp is made globally in other DPG methods. However, PolyDPG methods embrace the idea of local values of Δp instead, as it may even be a necessary criterion for the numerical method to work.

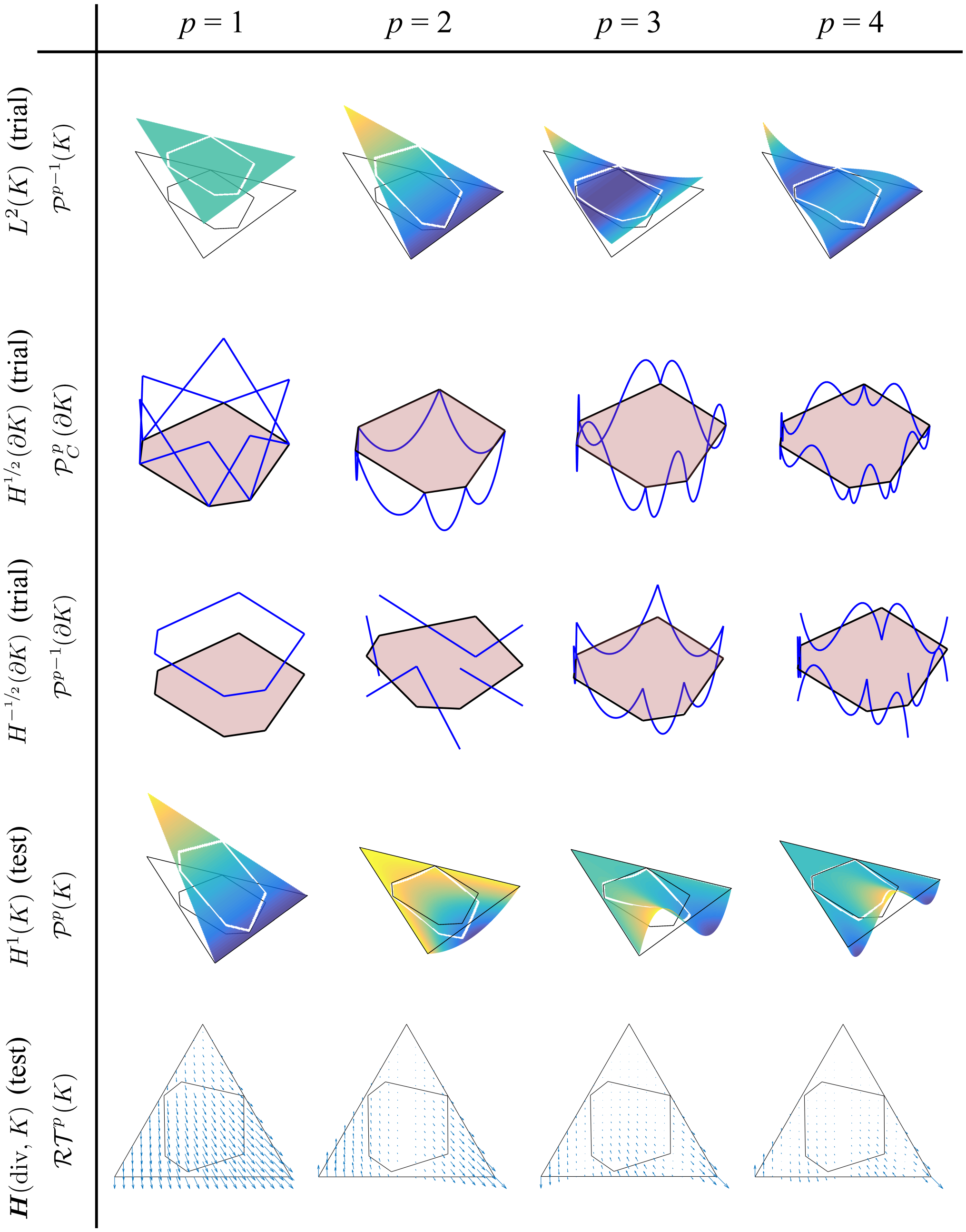

$p = 1$
$p = 2$
$p = 3$
$p = 4$
$L^2(K)$ (trial)
$\mathcal{P}^{p-1}(K)$
$H^{1/2}(\partial K)$ (trial)
$\mathcal{P}^p_C(\partial K)$
$H^{-1/2}(\partial K)$ (trial)
$\mathcal{P}^{p-1}(\partial K)$
$H^1(K)$ (test)
$\mathcal{P}^p(K)$
$\boldsymbol{H}(\mathrm{div}, K)$ (test)
$\mathcal{RT}^p(K)$

Figure 7.2: Some of the shape functions on a polygonal element used either as trial or test variables in the PolyDPG method. They are classified by the energy space ($H^1(K)$, $\boldsymbol{H}(\mathrm{div}, K)$, $L^2(K)$ and their respective traces) and shown for different values of the parameter $p \in \mathbb{N}$ denoting the order of the differential *sequence*. The underlying hierarchical shape functions for the bounding triangle and edges are taken from [114].

7.2.3 Convergence analysis

Since the subspaces used to discretize the ultraweak variational formulation are, rigorously speaking, subsets of the infinite dimensional trial and test spaces, PolyDPG methods are conforming FEMs. Thus, the "standard" convergence theory can be applied. However, this is an understatement because the interface variables are not standard, so they require a careful treatment, and the same holds true for the restricted shape functions in the remaining spaces. The standard interpolation estimates are presented in Section A.6 of Appendix A, and the necessary modifications to account for the choice of trial spaces in PolyDPG methods is covered in Appendix D, which the reader should consult for the details. Combining the results in Appendix D with the theory of Fortin operators in Section 2.8.1, it is not difficult to obtain an analogous convergence result to Theorem 2.4. Here, we only display the main result along with the key assumptions.

Definition 7.1. *A collection of subsets of $\mathbb{R}^2$, $\mathcal{T}_{\mathcal{K}}$, is said to have the finite overlap condition if*

$$\operatorname{ov}(\mathcal{T}_{\mathcal{K}}) = \sup_{\boldsymbol{x}\in\mathbb{R}^2} \operatorname{ov}(\boldsymbol{x}) < \infty \,, \qquad \operatorname{ov}(\boldsymbol{x}) = |\{\mathcal{K} \in \mathcal{T}_{\mathcal{K}} \mid \boldsymbol{x} \in \mathcal{K}\}| \,. \tag{7.11}$$

For a family of such collections given by a parameter $\mathfrak{h} \in \mathfrak{H}$, $\{\mathcal{T}_{\mathcal{K},\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$, the finite overlap condition is said to be robust in $\mathfrak{h}$ if there exists an integer $M_{\mathrm{ov}} > 0$, independent of $\mathfrak{h}$, such that $\operatorname{ov}(\mathcal{T}_{\mathcal{K},\mathfrak{h}}) \leq M_{\mathrm{ov}}$ for any $\mathfrak{h} \in \mathfrak{H}$.

Definition 7.2. *A triangulation $\mathfrak{T}(K) = \{\mathfrak{T}_i(K)\}_{i\in I_K}$ (with I_K finite) of a (simple) polygonal element K is said to be edge-compatible if for each edge of K, only one $\mathfrak{T}_i(K)$ shares that edge. For any polygon such a triangulation is known to exist* [171, 61, 2]. *The triangulation is additionally said to be shape-regular if all $\mathfrak{T}_i(K)$ satisfy a kind of uniform shape-regularity condition (e.g. they satisfy a minimum angle condition or the ratio of their diameters to their incircle radii remains bounded).*

Theorem 7.1. *Let Ω be a polygonal domain and $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ be family of meshes of Ω comprised of general polygonal elements with shape-regular edge-compatible triangulations for all $K \in \mathcal{T}_{\mathfrak{h}}$, and with a robust shape-regularity condition independent of $\mathcal{T}_{\mathfrak{h}}$. Assume that the associated collections of bounding triangles as described in Section 7.2.2, $\{\mathcal{T}_{T,\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}} = \{\{T_K\}_{K\in\mathcal{T}_{\mathfrak{h}}}\}_{\mathfrak{h}\in\mathfrak{H}}$, where T_K is* *the bounding triangle of a polygonal element K, satisfy a robust finite overlap condition. Consider linear well-posed variational formulations associated to $b_{\mathfrak{h}} : \mathcal{U}_{\mathfrak{h}} \times \mathcal{V}_{\mathfrak{h}} \to \mathbb{R}$ and $\ell_{\mathfrak{h}} : \mathcal{V}_{\mathfrak{h}} \to \mathbb{R}$, where $\mathcal{U}_{\mathfrak{h}}$ and $\mathcal{V}_{\mathfrak{h}}$ are SdR spaces. Let $\{\mathfrak{u}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$ be the exact solutions to the corresponding formulations, and assume they are attached to some $\mathfrak{u}_{\Omega} \in \mathcal{U}_{\Omega}$ through $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$, where $\mathcal{U}_{\Omega}$ is a compatible SdR space. Let $p \in \mathbb{N}$, and let $\mathcal{U}_{h,\mathfrak{h}} \subseteq \mathcal{U}_{\mathfrak{h}}$ and $\mathcal{V}_{r,\mathfrak{h}} \subseteq \mathcal{V}_{\mathfrak{h}}$ be PolyDPG discretizations of the trial and test spaces as described in Section 7.2.2. Suppose there exists a continuous linear Fortin operator $\Pi_{F,\mathfrak{h}} : \mathcal{V}_{\mathfrak{h}} \to \mathcal{V}_{r,\mathfrak{h}}$ such that for all $\mathfrak{v} \in \mathcal{V}_{\mathfrak{h}}$ and $\delta\mathfrak{u}_h \in \mathcal{U}_{h,\mathfrak{h}}$, $\|\Pi_{F,\mathfrak{h}}\mathfrak{v}\|_{\mathcal{V}_{\mathfrak{h}}} \leq C_{\Pi}\|\mathfrak{v}\|_{\mathcal{V}_{\mathfrak{h}}}$ and $b_{\mathfrak{h}}(\delta\mathfrak{u}_h, \mathfrak{v} - \Pi_{F,\mathfrak{h}}\mathfrak{v}) = 0$, for some $C_{\Pi} = C_{\Pi}(p) > 0$ that does not depend on the family of meshes $\{\mathcal{T}_{\mathfrak{h}}\}_{\mathfrak{h}\in\mathfrak{H}}$. Then, there exists a unique solution, $\mathfrak{u}_{h,\mathfrak{h}} \in \mathcal{U}_{h,\mathfrak{h}}$, solving the discrete variational formulation,*

$$b_{\mathfrak{h}}(\mathfrak{u}_{h,\mathfrak{h}}, \mathfrak{v}_{h,\mathfrak{h}}) = \ell_{\mathfrak{h}}(\mathfrak{v}_{h,\mathfrak{h}}) \qquad \forall \mathfrak{v}_{h,\mathfrak{h}} \in \mathcal{V}_{\mathfrak{h}}^{\text{n-opt}} = \mathcal{R}_{\mathcal{V}_{r,\mathfrak{h}}}^{-1}\mathcal{B}_{\mathfrak{h}}\mathcal{U}_{h,\mathfrak{h}}\,, \tag{7.12}$$

where $\mathcal{R}_{\mathcal{V}_{r,\mathfrak{h}}}^{-1}$ is defined in (2.39) *and $\mathcal{B}_{\mathfrak{h}}$ is defined in* (2.25)*. If the attached exact solution $\mathfrak{u}_{\Omega} \in \mathcal{U}_{\Omega}^{s}$ for some $s > \frac{1}{2}$, where $\mathcal{U}_{\Omega}^{s}$ is the fractional counterpart to $\mathcal{U}_{\Omega}$, then*

$$\|\mathfrak{u}_{\mathfrak{h}} - \mathfrak{u}_{h,\mathfrak{h}}\|_{\mathcal{U}_{\mathfrak{h}}} \leq C h_{\mathfrak{h}}^{\min\{s,p\}} \|\mathfrak{u}_{\Omega}\|_{\mathcal{U}_{\Omega}^{s}}\,, \tag{7.13}$$

where $h_{\mathfrak{h}} = \sup_{K\in\mathcal{T}_{\mathfrak{h}}} \operatorname{diam}(K)$ and $C = C(s,p) > 0$. Moreover, if C_{Π} is p-independent as well, then in the p-asymptotic limit an hp-convergence estimate holds with $C = C_s(\ln p)^2 p^{-s}$ where $C_s = C(s)$ is independent of p.

Remark 7.1. The robust finite overlap condition is also assumed in [52], and is not a very restrictive assumption. It is used in the proof to establish a robust finite constant for the global $L^2(\Omega)$ convergence estimates (details are in Appendix D). On the other hand, the robust shape-regular edge-compatible triangulation of all elements is a more restrictive assumption, but it is necessary to prove the convergence estimates of the interface variables.

Remark 7.2. As shown in Appendix A, the theorem actually holds for any well-posed broken ultraweak variational formulation with trial variables in $L^2(\Omega)$ and interface (also trial) variables in subsets of $H^{1/2}(\partial\mathcal{T})$ and $H^{-1/2}(\partial\mathcal{T})$. Thus, this result also holds for other equations such as linear elasticity, acoustics, and convection-dominated diffusion.

Remark 7.3. The arguments can be easily extended to a 3D mesh with polyhedral elements provided all the faces of the polyhedra are triangular. Then, the proof would even hold for equations involving interface variables representing the traces of $\boldsymbol{H}(\mathrm{curl},\Omega)$ spaces, like an ultraweak formulation of Maxwell's equations (see [55]). However, the problem (and the corresponding numerical implementation) is more challenging for general polyhedra in 3D.

7.3 Numerical results

In this section we consider several examples to examine the performance of the PolyDPG method. In all cases, Poisson's equation representing the nondimensionalized steady-state heat equation was solved in the domain $\Omega = (0,1)^2$. Unless otherwise stated, bounding triangles were utilized (as opposed to bounding boxes) and the (nondimensional) conductivity was taken as $k = 1$. Also, a default uniform value of $\Delta p = 1$ was used, but was increased (uniformly across the mesh, for the sake of simplicity) if deemed necessary (see (7.10) in Section 7.2.2).

For the broken ultraweak formulations, the adjoint graph norm has interesting properties [95]. Using the ultraweak formulation in (7.2), the first two terms in this norm can be derived as,

$$\|(v,\boldsymbol{\tau})\|_{\mathcal{V}}^2 = \|\tfrac{1}{k}\boldsymbol{\tau} - \nabla v\|_{\boldsymbol{L}^2(\mathcal{T})}^2 + \|-\operatorname{div}\boldsymbol{\tau}\|_{L^2(\mathcal{T})}^2 + \varepsilon_0^2\big(\|v\|_{L^2(\mathcal{T})}^2 + \|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(\mathcal{T})}^2\big)\,. \tag{7.14}$$

The third term, which has the ε_0^2 factor, makes the norm localizable, because otherwise (7.14) would not be a norm for arbitrary broken functions $v \in H^1(\mathcal{T})$ (although it would be a norm for $v \in H_0^1(\Omega)$). One can choose an arbitrary value for $\varepsilon_0 > 0$, but using small values of ε_0 (with the caveat of ill-conditioned local problems) is of particular interest for certain equations, such as Helmholtz [132]. Note that the corresponding inner products for the (real-valued) Hilbert space $\mathcal{V}$ can be derived from the polarization identity, $(\mathfrak{v}_1,\mathfrak{v}_2)_{\mathcal{V}} = \frac{1}{4}\big(\|\mathfrak{v}_1+\mathfrak{v}_2\|_{\mathcal{V}}^2 - \|\mathfrak{v}_1-\mathfrak{v}_2\|_{\mathcal{V}}^2\big)$. For all computations, the adjoint graph norm written in (7.14) with $\varepsilon = 1$ was used as the test space norm.

In the first example, we studied nontrivial meshes with n-sided convex polygons. In the second example, we considered highly distorted non-convex elements in the mesh. The third example was inspired by problems in geoscience, where arbitrary faults separating different material properties occur. To model this, we cut a uniform grid at an angle, so that the resulting mesh had different polygons (pentagons, quadrilaterals and triangles) with discontinuous material properties at each side of the cut. In these three examples, "uniform" refinements were analyzed for different values of $p \in \mathbb{N}$, in the sense that the largest element diameter was roughly cut in half with each refinement.

Adaptivity in its own right is a very interesting subject of study for polygonal elements, as they provide great flexibility for the implementation of such strategies without resorting to constrained approximations to deal with hanging nodes. The natural arbitrary-order a posteriori error estimator inherent to all DPG methods, computed from (2.50), also applies to the polygons. Therefore, in the final example we described a polygonal adaptivity scheme by using the PolyDPG arbitrary-order a posteriori error estimator, and compared it with conventional adaptive methods (using standard element shapes). This is particularly important since adaptive refinement algorithms applied to polygonal elements have applications in topology optimization [219, 122, 5, 221] and crack propagation [211, 167].

Note that in all examples we only report the relative error in the $\mathcal{U}_0$ trial space component. This is because a rigorous computation of the norms in the $\hat{\mathcal{U}}$ trial space component is simply not viable. The $\mathcal{U}_0$ relative error is defined as

$$\text{Relative error} = \frac{\|\mathfrak{u}_0 - (\mathfrak{u}_0)_h\|_{\mathcal{U}_0}}{\|\mathfrak{u}_0\|_{\mathcal{U}_0}}\,, \qquad \|(u, \boldsymbol{q})\|_{\mathcal{U}_0}^2 = \|u\|_{L^2(\Omega)}^2 + \|\boldsymbol{q}\|_{\boldsymbol{L}^2(\Omega)}^2\,, \tag{7.15}$$

where $\mathfrak{u}_0$ is the exact solution and $(\mathfrak{u}_0)_h$ is the computed solution from the PolyDPG method.

Remark 7.4 (**`PolyDPG` software**)**.** Implementation of PolyDPG methods may deceptively appear difficult when compared to typical FEM algorithms, so we developed an open-source code written in MATLAB$^{®}$ also called `PolyDPG` [228]. It can be run sequentially or in parallel, and it supports both conventional and polygonal elements. We hope this removes some qualms related to the implementation and makes DPG methods more accessible to other researchers. The shape functions used in the code were originally described in [114] (see Figure 7.2). The numerical integration was carried out by splitting the polygons into triangles (through Delaunay triangulation), followed by using Gaussian quadrature for each triangle (the Gaussian quadrature points and weights were carefully mapped back from a square), so that polynomial integrands of a certain order were computed up to machine precision.

7.3.1 Mesh with convex polygons

In this example, we investigated meshes with n-sided convex polygonal elements. The software `PolyMesher` [220] was used to generate the polygonal meshes. In Figure 7.3 an initial mesh and three subsequent refinements are displayed. The elements are colored according to their number of sides, ranging from 4 (quadrilaterals) to 7 (heptagons). We used the manufactured solution,

$$u(x, y) = \sin(\pi x) \sin(\pi y) \,, \tag{7.16}$$

for $(x, y) \in \Omega = (0, 1)^2$ to determine the forcing, i.e. the internal heat source r in (2.1), and the boundary conditions of u at $\partial\Omega$.

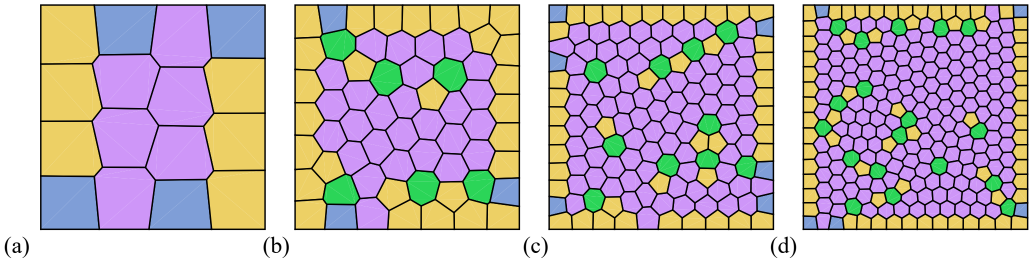


Figure 7.3: Four refinements of a mesh with n-sided convex polygonal elements. The elements are colored according to their number of sides.

As mentioned before, given a trial space associated to a parameter p, the corresponding (uniform) value of Δp was calculated from (7.10) (using the polygon with the greatest number of sides). Given the presence of hexagons and heptagons, this meant that $\Delta p = 2$ was required when $p = 1, 2$, while $\Delta p = 3$ was needed when $p = 3, 4$. The numerical results are plotted and presented in Figure 7.4 for $p = 4$, including the interface temperature, temperature, and heat flux. Additionally, the relative error, calculated using (7.15), is shown in Figure 7.5, where the expected h-convergence rates can be observed for all values of p (the behavior is of the form h^p as established by Theorem 7.1). Note that the number of degrees of freedom, N_{dof}, is proportional to h^2. Thus, the log-log slope indicators in Figure 7.5 display a 2 in the N_{dof}-direction, while the other label corresponds to the h-convergence rate, $\widetilde{p}$ (so that $\frac{\widetilde{p}}{2}$ is the N_{dof}-convergence rate).

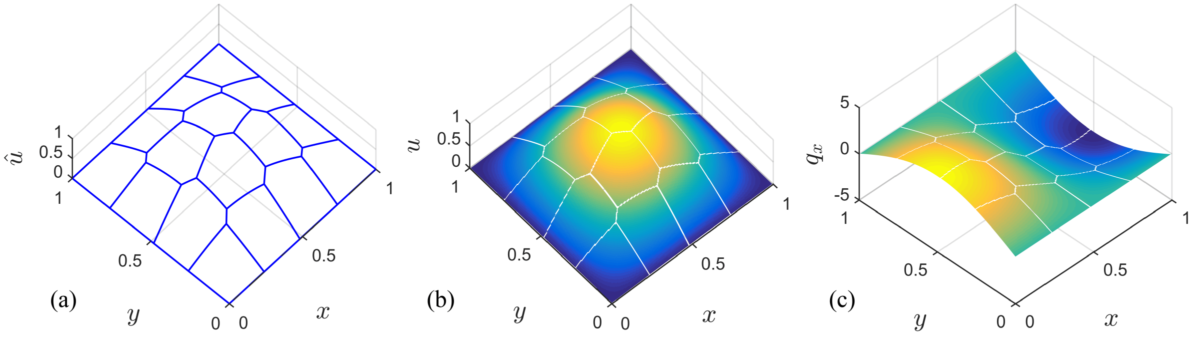


Figure 7.4: Numerical results using the solution in (7.16) on the coarse mesh from Figure 7.3(a) using $p = 4$ and $\Delta p = 3$: (a) interface temperature, (b) temperature, (c) first component of the heat flux.

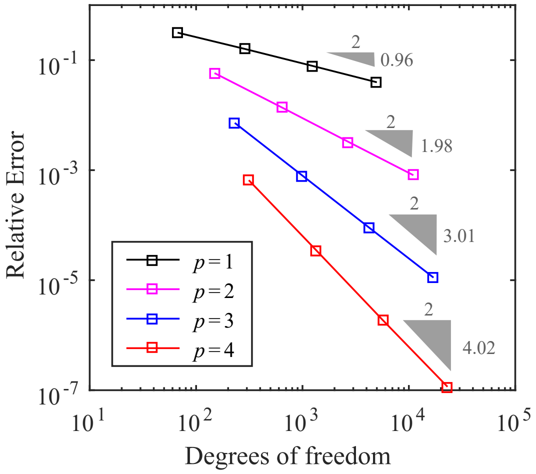


Figure 7.5: Convergence study of the PolyDPG method in terms of degrees of freedom. The h-convergence behavior is displayed for different values of p using the polygonal meshes in Figure 7.3.

7.3.2 Mesh with distorted elements

To demonstrate the distortion tolerance of PolyDPG methods, we considered a mesh with highly distorted quadrilaterals, including non-convex elements. The pattern was then scaled and tessellated to produce the refinements shown in Figure 7.6. This example is challenging in the sense that other numerical methods likely fail due to the degeneration of either the parametric mapping or the barycentric coordinates associated with the highly distorted elements [166, 149]. The same problem as in Section 7.3.1 was solved (see (7.16) for manufactured solution). The solution values and h-convergence rates for $1 \leq p \leq 4$ are shown in Figures 7.7 and 7.8 respectively. The expected convergence behavior was observed, showing the flexibility of PolyDPG methods to deal with irregular elements.

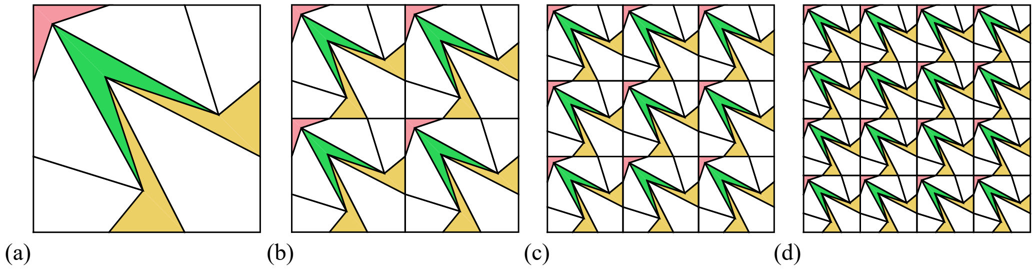


Figure 7.6: Four refinements using the tessellation of a mesh with highly distorted quadrilaterals. The non-convex elements are colored.

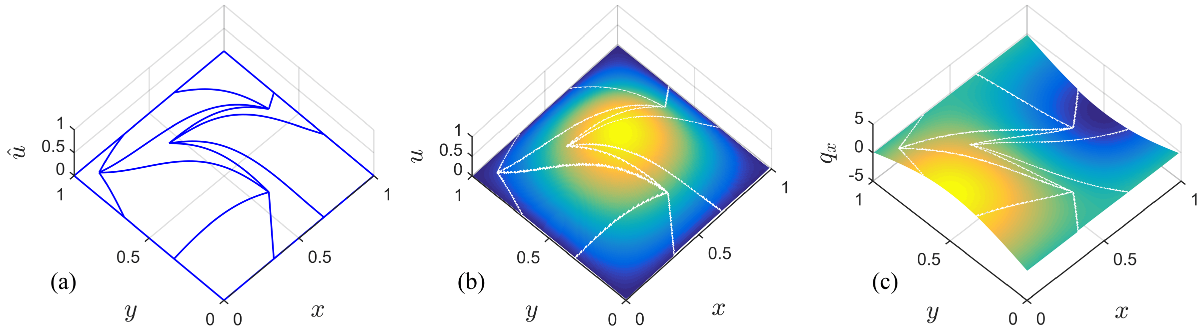


Figure 7.7: Numerical results using the solution in (7.16) on the coarse mesh from Figure 7.6(a) using $p = 4$ and $\Delta p = 2$: (a) interface temperature, (b) temperature, (c) first component of the heat flux.

7.3.3 Interface problem

The inspiration behind this example came from applications in geoscience where faults abruptly separate the material properties within a domain. Here we considered a domain composed of two materials with different heat conductivities, which share an interface (for simplicity a straight line at an arbitrary angle dividing the square). The heat conductivities are assumed to be uniform on each side of the interface, taking values k_1 and k_2, as depicted in Figure 7.9.

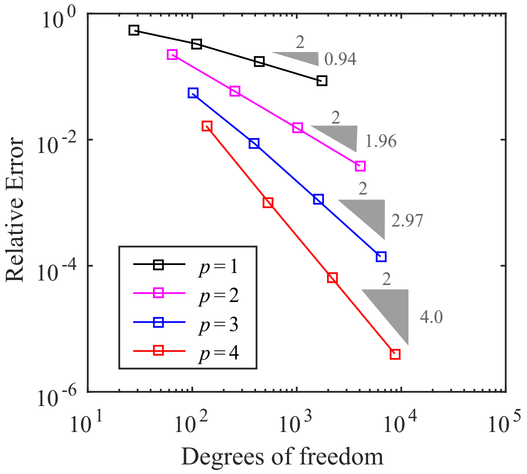


Figure 7.8: Convergence study of the PolyDPG method in terms of degrees of freedom. The h-convergence behavior is displayed for different p using the meshes with highly distorted quadrilaterals in Figure 7.6.

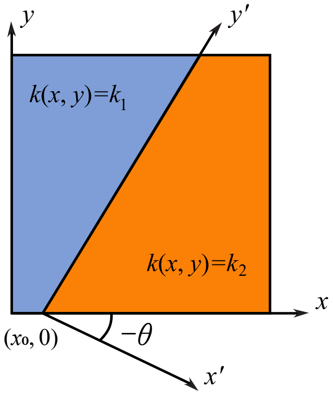


Figure 7.9: Material properties and rotated coordinates of the interface problem.

To model certain interfaces one would need unstructured grids. However, by using PolyDPG methods we are able to consider a uniform background grid and simply cut the elements through the interface, leading to the creation of triangles, right trapezoids and pentagons near the interface. In fact, to refine the mesh, first the background mesh was uniformly refined, and then the elements were cut by the interface line. There is one caveat which is only evident for high values of p or small values of h: when extremely small triangles (compared to their neighbors) are formed, the assembled stiffness matrix becomes ill conditioned (so the *infinite*-precision result in Theorem 7.1 seizes to hold). Thus, it is necessary to either relocate the nodes along the interface or to collapse the nodes of the small triangle into a single node on the interface. We chose to implement the latter approach whenever the area of the small triangle was less than 1% of the area of the background grid elements. The meshes obtained are shown in Figure 7.10.

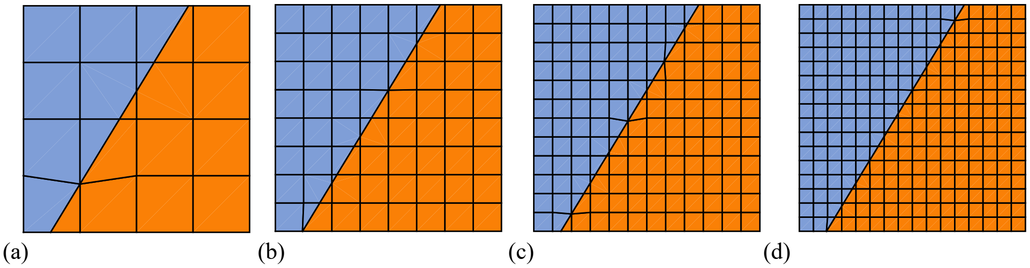


Figure 7.10: Four refinements of a mesh with an interface between two materials. Notice that some nodes are collapsed to a node on the interface. This is due to eliminating the undesired tiny triangles that cause ill conditioning.

For this problem we designed a manufactured solution that guarantees continuity of the temperature and the heat flux across the interface, taking into account the finite jump in the conductivity coefficient. By means of a translated and rotated system of coordinates, and following the notation in Figure 7.9, the exact solution is given by,

$$u(x',y') = \left\{ \begin{array}{ll} k_2 \sin(\pi x')\sin(\pi y'), & \text{for } x' \leq 0\,, \\ k_1 \sin(\pi x')\sin(\pi y'), & \text{for } x' > 0\,, \end{array} \right. \tag{7.17}$$

where the coordinates x' and y' come from a translation and rotation of the reference system defined by the following transformation,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x - x_0 \\ y \end{pmatrix} . \tag{7.18}$$

The values of conductivity and the geometric data used for the numerical computation are $k_1 = 1$, $k_2 = 5$, $x_0 = 0.12$ and $\theta = \tan^{-1}(1/0.65)$. The nonzero boundary conditions were imposed using projection-based interpolation of the manufactured solution on the boundary edges [90, 103].

Figure 7.11 shows the appearance of the computed ultraweak solution. As it can be observed in Figure 7.12, the expected convergence rates were verified once again. It is remarkable that without collapsing any nodes in these meshes, the same data points were observed for $1 \leq p \leq 3$, but the last data point for $p = 4$ did behave unexpectedly, so collapsing the nodes is still recommended in general.

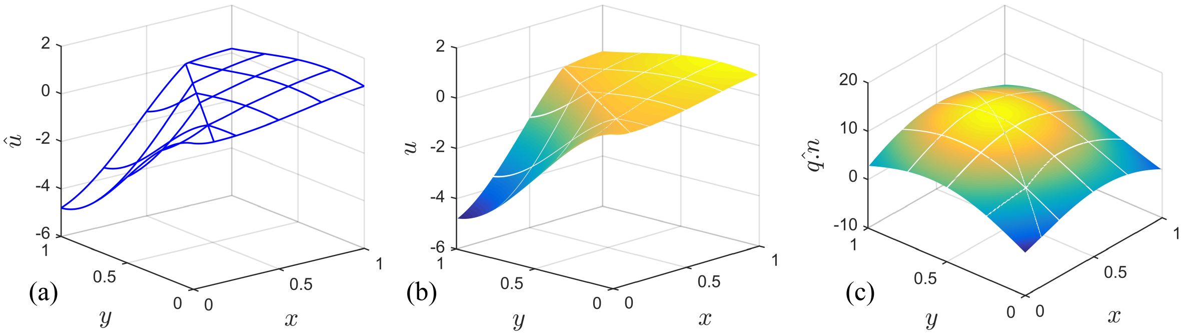


Figure 7.11: Numerical results using the manufactured solution in (7.17) and (7.18) on the coarse mesh from Figure 7.10(a) using $p = 4$ and $\Delta p = 2$: (a) interface temperature, (b) temperature, (c) first component of the heat flux.

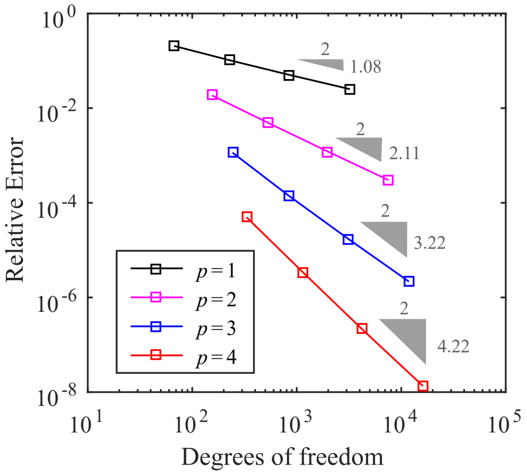


Figure 7.12: Convergence study of the PolyDPG method in terms of degrees of freedom. The h-convergence behavior is displayed for different p using the meshes with an interface in Figure 7.10.

7.3.4 Adaptivity

In the last example, we aimed to present a polygonal adaptive strategy. This is of interest as it has direct applications in fracture dynamics [211, 167] and topology optimization [219, 122]. Implementing such a strategy was possible, because the DPG methodology carries a natural arbitrary-order a posteriori error estimator (see (2.50) and Section 2.6). The criterion used to mark an element for refinement was given by Remark 2.2 with $\alpha_\eta = 0.5$.

In order to refine traditional quadrilateral elements, typically hanging nodes arise in the mesh. But in practice, only one "level" of refinement is possible per element (often edges cannot have more than one hanging node), resulting in so-called quadtree meshes [218]. To implement this strategy a constrained approximation technology is necessary to handle the hanging nodes. Additionally, under anisotropic refinements, sometimes dead-lock scenarios arise (where it is logically impossible to continue refining) and these must be avoided [89]. In short, it may be challenging to implement conventional refinement strategies used for adaptivity.

An important advantage of the polygonal elements is that they naturally embrace hanging nodes, because they merely represent that a polygon has an extra edge collinear with another edge. Thus, the polygonal methods do not require an extra level of difficulty in terms of implementing the adaptive refinements. We devised a practical convex polygonal refinement strategy as illustrated in Figure 7.13: (a) shows the initial mesh in which an element of interest is picked and split into quadrilaterals by using the centroid and edge midpoints as depicted in (b); next, any of the resulting elements can be subsequently refined into finer quadrilaterals as shown in (c); and lastly, as shown in (d), if a neighbor element needs to be refined too, it is split into quadrilaterals assuming all adjacent collinear edges constitute a single edge (i.e. the vertices of this combined edge are used in the calculation of the centroid and its midpoint used to place the new quadrilateral node).

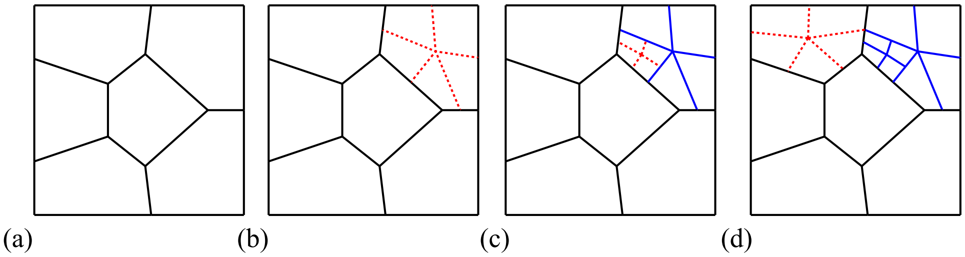


Figure 7.13: A practical local refinement strategy for convex polygons: (a) initial coarse polygonal mesh; (b) line segments are projected from the centroid to every edge midpoint in the element of interest; (c) the same approach is used to refine sub-elements; (d) the strategy can be re-applied to any other coarser element by assuming all collinear vertices constitute a single combined edge.

The manufactured solution for this problem is the sum of two Gaussian surfaces, given by the function,

$$u(x,y) = \frac{1}{2\pi\sigma^2}\left[e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{y-\mu_1}{\sigma}\right)^2} + e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma}\right)^2}\right] , \tag{7.19}$$

where the standard deviation is $\sigma = \sqrt{10^{-3}}$ and the two means are $\mu_1 = 0.25$ and $\mu_2 = 0.75$. Again, projection-based interpolation [90, 103] was used to approximate the nearly vanishing temperature boundary conditions.

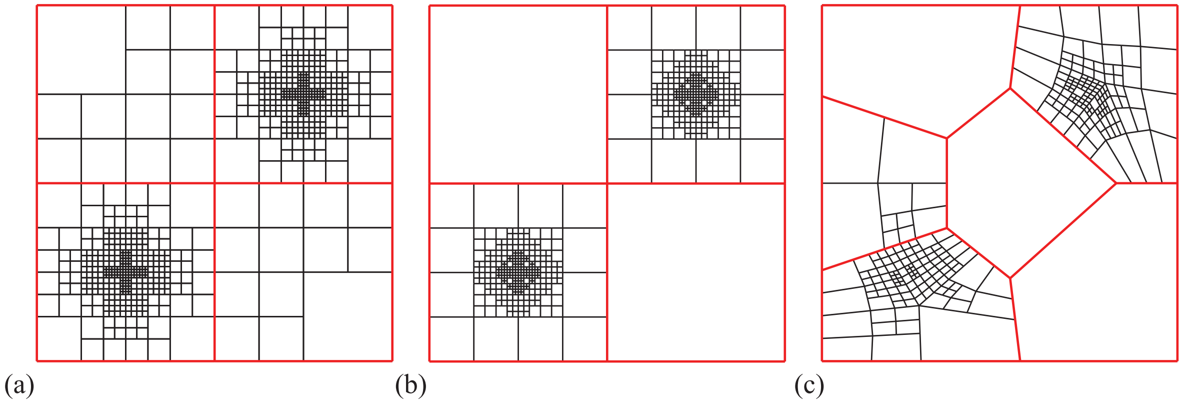


Figure 7.14: Three h-adaptively refined meshes (the thick red lines represent the initial mesh) for the manufactured solution in (7.19): (a) traditional quadtree meshes via constrained nodes; (b) quadrilateral mesh using the polygonal adaptive strategy; (c) polygonal mesh using the polygonal adaptive strategy.

In order to compare with other adaptive schemes, a traditional adaptive strategy using quadtree meshes and constrained hanging nodes via quadrilateral elements was considered here [89]. Starting with the same initial mesh, the traditional refinement strategy and the polygonal refinement strategy were allowed to refine accordingly. When using the polygonal strategy on these quadrilateral meshes, we used the more natural choice of bounding boxes instead of the bounding triangles. Additionally, the same polygonal refinement strategy was applied to an initial polygonal mesh (using bounding triangles as usual). Figure 7.14 shows the results of the three different scenarios after several refinements. Clearly, the traditional adaptive strategy produces quadtree meshes (see Figure 7.14(a)), so it is forced to refine and create new elements in areas of the domain where the solution is nearly constant. However, the polygonal adaptive strategy applied to the same initial mesh produces a more localized refinement pattern which is not a quadtree mesh (see Figure 7.14(b)). Lastly, the polygonal adaptive strategy applied to a polygonal mesh produces a completely nonstandard, yet localized mesh (see Figure 7.14(c)).

The numerical solution for $p = 6$ and $\Delta p = 2$ using the mesh in Figure 7.14(c) is presented in Figure 7.15. The error convergence curves corresponding to the three refinement schemes in Figure 7.14 are also displayed in Figure 7.16. The proposed polygonal refinement technique generates more edges (each new sub-segment becomes an edge) resulting in more degrees of freedom. However, in the end the additional cost is compensated by producing less elements than traditional quadtree refinement schemes (compare (b) and (c) with (a) in Figure 7.14). It can be seen from Figure 7.16 that the convergence behavior in terms of degrees of freedom is very similar using both approaches. Therefore, the polygonal adaptive strategy proposed here is competitive with the existing strategies for traditional elements, whilst being more general in its applicability as it also works for polygonal elements.

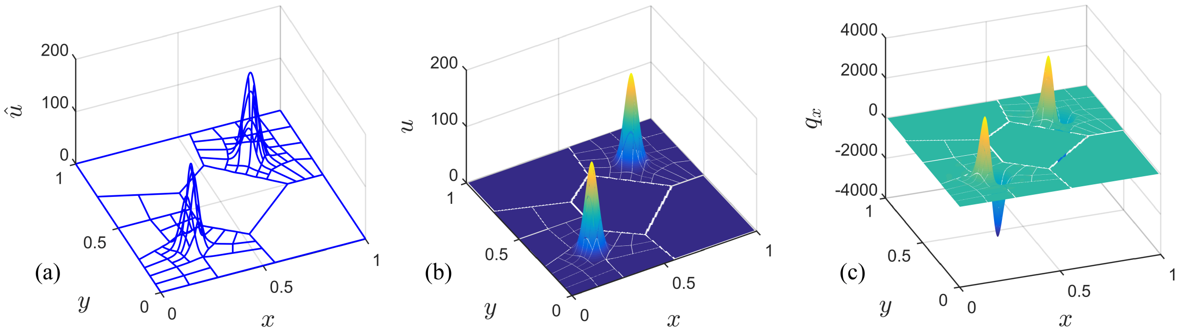


Figure 7.15: Numerical results using the manufactured solution in (7.19) on the mesh from Figure 7.14(c) using $p = 6$ and $\Delta p = 2$: (a) skeleton temperature, (b) temperature, (c) first component of the heat flux.

7.4 Discussion

A PolyDPG method discretized with high-order polygonal elements was successfully implemented using ultraweak formulations and the DPG methodology. Here, the PolyDPG method solves Poisson's equation. However, like with the DPG methodology, the discretization and theory is quite general. Thus, it can be applied to a large family of equations including acoustics,

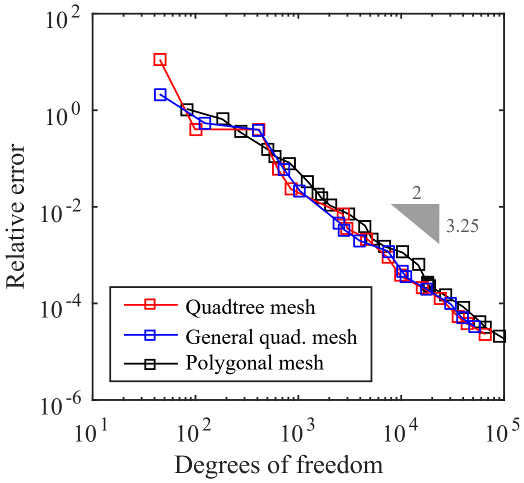


Figure 7.16: Convergence study of the PolyDPG method in terms of degrees of freedom. The h-convergence behavior is displayed using $p = 3$ for several successive refinements associated with the refinement strategies in Figure 7.14.

convection-dominated diffusion and linear elasticity. PolyDPG methods are conforming FEMs, and as with many other polytopal methods, the spaces and integration schemes are defined directly in the physical space. Indeed, given that the ultraweak formulations avoid interelement compatibility conditions, it is relatively straightforward to obtain many of the shape functions by restricting them from a bounding (triangular or quadrilateral) element to the polygonal element. Despite the greater computational cost compared to conventional methods, the resulting PolyDPG methods are naturally high-order, carry their own residual-based a posteriori error estimator, have no need of ad hoc stabilization terms, and always produce positive definite stiffness matrices. Moreover, under reasonable assumptions, a rigorous proof demonstrating the convergence of PolyDPG methods was included. To complement this chapter, the `PolyDPG` software [228] written in MATLAB® is provided. We hope this will prove to be a practical tool for other researchers interested in polygonal FEMs and in DPG methods.

Different illustrative examples corroborated the expected results. In the first example, n-sided convex polygons were investigated, while in the second example, highly distorted concave elements were examined. In both cases, as predicted by the theory, convergence rates of the form h^p were observed for different values of p, confirming that PolyDPG methods are distortion-tolerant.

The third example was relevant to the field of geosciences, where faults cause heterogeneity in the domain. This was simulated by irregularly cutting a uniform grid with an interface and assigning different material properties on each side. Once again, the method converged as expected, displaying its robustness in resolving heterogeneous material properties. The final example explored a polygonal adaptivity scheme driven by the arbitrary-order a posteriori error estimator of PolyDPG methods. Even though polygonal and standard refinement strategies led to practically identical convergence curves, polygonal techniques are more general since they apply to polygonal elements and avoid the typical approaches of constrained approximations via hanging nodes. These techniques may be useful in applications such as crack propagation and topology optimization.

Extension of the presented technique to arbitrary 3D polyhedral elements would be an interesting endeavor. In principle, the current numerical method can be extended naturally to polyhedral elements, as long as all the faces are triangular, but the case of arbitrary faces is much more challenging and might lead to analyzing nonconforming numerical methods.

Chapter 8

Conclusions

While it is true that some finite element methods are much more effective than DPG methods at solving a specific PDE, they usually succumb to ad hoc modifications whenever the PDE is fundamentally changed. The power of the DPG methodology lies in its ability to effectively solve almost any linear PDE without altering the underlying approach. In this dissertation such versatility was shown by solving four distinct variational formulations of the equations of linear elasticity. Moreover, by using insights from functional analysis it was proved those formulations were mutually well-posed. Additionally, it was shown how interface variables present in broken variational formulations serve as an ideal vessel to couple different DPG methods within the same domain. Similar ideas have been applied by other researchers to couple DPG methods with other numerical methods. From a physical standpoint, a relevant example was presented where such coupling is convenient.

The natural high-order residual-based a posteriori error estimator attached to DPG methods was also found to be very useful in physical applications involving the equations of linear viscoelasticity and thermoviscoelasticity. More specifically, DMA calibration experiments of two different polymers were reproduced and the quantity of interest was found to be within 5% of the experimental values. The adaptivity was key in helping to reproduce the results, since the stresses were concentrated in a very small area within the domain. Similarly, three different scenarios modeling form-wound medium-voltage stator coils were simulated. These scenarios were distinguished by the physical phenomena occurring at different frequency regimes. At low frequencies thermoviscoelastic behavior was assumed to dominate, at mid-range frequencies the ovalization of the stator determined the deformation, while at high frequencies the Lorentz forces resulting from the interaction of electromagnetic variables drove the problem. For this last scenario, a new model based on scattering theory was developed to calculate the Lorentz forces occurring at the interface between the copper coils and the surrounding resin. The mechanical work done in one second was computed for all three physical scenarios and it was found to be higher in the mid-range frequency scenario.

Exploiting the algebraic properties in the DPG methodology led to some computational optimizations, but more importantly it guided the development of a new family of numerical methods labeled as discrete least-squares (DLS) finite element methods. They rely on using QR-based algorithms to solve overdetermined rectangular discrete least-squares problems that discretize a particular PDE, and are especially advantageous when dealing with very ill-conditioned problems. Some illuminating examples were presented to show their applicability.

Finally, the features of ultraweak formulations were exploited to construct polygonal DPG (PolyDPG) methods, along with a proof of convergence subject to some sensible assumptions. Numerical experiments showed the methods were distortion-tolerant and able to handle complicated non-convex polygonal elements in the mesh. Moreover, polygonal adaptive strategies were developed and found to be competitive with adaptive strategies involving traditional elements. Lastly, interesting practical examples pertinent to geophysical applications were illustrated.

Appendices

Appendix A

Sobolev spaces: simplest de Rham sequence

This chapter is designed to be a short reference to Sobolev spaces that are part of the most traditional differential de Rham exact sequence. The spaces, their traces and norms are all explicitly defined, both in usual domains and in a mesh. The latest results of relevance are stated as theorems. Moreover, a family of hypothetical discretizations of these spaces is considered, and their approximation and interpolation properties are stated.

A.1 Exact sequence spaces

Depending on the number of spatial dimensions, the simplest differential de Rham exact sequence involving the "standard" Sobolev spaces is different,

$$\text{1D}: \quad H^1(K) \xrightarrow{\nabla} L^2(K) \tag{A.1}$$

$$\text{2D}: \begin{cases} H^1(K) \xrightarrow{\nabla} \boldsymbol{H}(\text{curl}, K) \xrightarrow{\nabla\times} L^2(K) \\ H^1(K) \xrightarrow{\text{curl}} \boldsymbol{H}(\text{div}, K) \xrightarrow{\text{div}} L^2(K) \end{cases} \tag{A.2}$$

$$\text{3D}: \quad H^1(K) \xrightarrow{\nabla} \boldsymbol{H}(\text{curl}, K) \xrightarrow{\text{curl}} \boldsymbol{H}(\text{div}, K) \xrightarrow{\text{div}} L^2(K) \tag{A.3}$$

where $K \subseteq \mathbb{R}^{n_d}$ is a contractible Lipschitz domain (e.g. a simply connected domain with connected complement) and n_d is the number of spatial dimensions. These sequences should be prepended by $\mathbb{R}$ and appended by $\{0\}$ (with the first operation being $\text{id}(\cdot)$ and the last being the zero operator, 0), but for simplicity only the relevant part is shown. The definitions of the spaces and operations will be explained below.

First, let the L^2 inner product in K be defined as

$$(u, v)_K = \int_K \text{tr}_{\mathbb{M}}(u^{\mathsf{T}} v) \, \mathrm{d}K \,, \tag{A.4}$$

where $\mathrm{tr}_{\mathbb{M}}$ is the usual algebraic trace of a matrix, so that depending on whether u and v take scalar, vector or matrix values, $\mathrm{tr}_{\mathbb{M}}(u^{\mathsf{T}}v)$ will be uv, $u\cdot v$ or $u:v$, respectively. Next, let

$$L^2(K)=\big\{u:K\to\mathbb{R}\mid \|u\|_{L^2(K)}<\infty\big\}\,,\qquad \|u\|^2_{L^2(K)}=(u,u)_K\,. \tag{A.5}$$

Then, for every $u\in L^2(K)$, $\boldsymbol{v}\in\big(L^2(K)\big)^{n_d}$ and $\boldsymbol{E}\in\big(L^2(K)\big)^3$, the distributional gradient, divergence and 3D curl are uniquely characterized by

$$\begin{aligned}(\nabla u,\boldsymbol{\phi})_K&=-(u,\operatorname{div}\boldsymbol{\phi})_K &&\forall\boldsymbol{\phi}\in\big(C_0^\infty(K)\big)^{n_d}\,,\\ (\operatorname{div}\boldsymbol{v},\phi)_K&=-(\boldsymbol{v},\nabla\phi)_K &&\forall\phi\in C_0^\infty(K)\,,\\ (\operatorname{curl}\boldsymbol{E},\boldsymbol{\phi})_K&=(\boldsymbol{E},\operatorname{curl}\boldsymbol{\phi})_K &&\forall\boldsymbol{\phi}\in\big(C_0^\infty(K)\big)^3\,,\end{aligned} \tag{A.6}$$

whenever $\nabla u\in\big(L^2(K)\big)^{n_d}$, $\operatorname{div}\boldsymbol{v}\in L^2(K)$ and $\operatorname{curl}\boldsymbol{E}\in\big(L^2(K)\big)^3$. Otherwise these concepts do exist, but are identified directly as distributions, and the definitions are the same, except the inner product is replaced by the duality pairing between distributions and the space of test functions, $C_0^\infty(K)$ (the smooth functions with compact support in K endowed with the topology of test functions [203]). For this document the theory of distributions will not be needed. Note that in 1D, $\nabla=\operatorname{div}$. Lastly, the only operations that remain to be defined are the 2D curl and $\nabla\times$, which for $u\in L^2(K)$ and $\boldsymbol{E}\in\big(L^2(K)\big)^2$ are uniquely characterized by

$$\begin{aligned}\Big(\big(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\big)\operatorname{curl}u,\boldsymbol{\phi}\Big)_K&=-\big(u,\operatorname{div}\boldsymbol{\phi}\big)_K &&\forall\boldsymbol{\phi}\in\big(C_0^\infty(K)\big)^2\,,\\ \big(\nabla\times\boldsymbol{E},\phi\big)_K&=-\Big(\big(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\big)\boldsymbol{E},\nabla\phi\Big)_K &&\forall\phi\in C_0^\infty(K)\,,\end{aligned} \tag{A.7}$$

when $\operatorname{curl}u\in\big(L^2(K)\big)^2$ and $\nabla\times\boldsymbol{E}\in L^2(K)$.

From now on, assume that $n_d=3$ is the number of spatial dimensions. It should be clear how to define everything for $n_d=2$ and $n_d=1$. The definitions of the remaining 3D spaces in the exact sequence are

$$H^1(K)=\big\{u\in L^2(K)\mid\nabla u\in L^2(K)\big\}\,,\qquad \|u\|^2_{W^1(K)}=\|u\|^2_{L^2(K)}+\|\nabla u\|^2_{\boldsymbol{L}^2(K)}\,, \tag{A.8}$$

$$\boldsymbol{H}(\mathrm{curl},K)=\big\{\boldsymbol{E}\in\boldsymbol{L}^2(K)\mid\operatorname{curl}\boldsymbol{E}\in\boldsymbol{L}^2(K)\big\}\,,\quad \|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathrm{curl},K)}=\|\boldsymbol{E}\|^2_{\boldsymbol{L}^2(K)}+\|\operatorname{curl}\boldsymbol{E}\|^2_{\boldsymbol{L}^2(K)}\,, \tag{A.9}$$

$$\boldsymbol{H}(\mathrm{div},K)=\big\{\boldsymbol{v}\in\boldsymbol{L}^2(K)\mid\operatorname{div}\boldsymbol{v}\in L^2(K)\big\}\,,\qquad \|\boldsymbol{v}\|^2_{\boldsymbol{H}(\mathrm{div},K)}=\|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(K)}+\|\operatorname{div}\boldsymbol{v}\|^2_{L^2(K)}\,, \tag{A.10}$$

where $\|\cdot\|^2_{\boldsymbol{L}^2(K)}=(\cdot\,,\cdot)_K$ is the norm of $\boldsymbol{L}^2(K)=\big(L^2(K)\big)^3$.

Given a domain $\Omega \subseteq \mathbb{R}^3$, let $\mathcal{T}$ be a mesh (i.e. an open partition) of elements (i.e. subdomains) $K \in \mathcal{T}$. The broken spaces are simply

$$H^1(\mathcal{T}) = \left\{u \in L^2(\Omega) \mid u|_K \in H^1(K)\ \forall K \in \mathcal{T}\right\}, \qquad \|u\|^2_{W^1(\mathcal{T})} = \sum_{K\in\mathcal{T}} \left\|u|_K\right\|^2_{W^1(K)}, \tag{A.11}$$

$$\boldsymbol{H}(\mathrm{curl},\mathcal{T}) = \left\{\boldsymbol{E} \in \boldsymbol{L}^2(\Omega) \mid \boldsymbol{E}|_K \in \boldsymbol{H}(\mathrm{curl},K)\ \forall K \in \mathcal{T}\right\}, \quad \|\boldsymbol{E}\|^2_{\boldsymbol{H}(\mathrm{curl},\mathcal{T})} = \sum_{K\in\mathcal{T}} \left\|\boldsymbol{E}|_K\right\|^2_{\boldsymbol{H}(\mathrm{curl},K)}, \tag{A.12}$$

$$\boldsymbol{H}(\mathrm{div},\mathcal{T}) = \left\{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \mid \boldsymbol{v}|_K \in \boldsymbol{H}(\mathrm{div},K)\ \forall K \in \mathcal{T}\right\}, \qquad \|\boldsymbol{v}\|^2_{\boldsymbol{H}(\mathrm{div},\mathcal{T})} = \sum_{K\in\mathcal{T}} \left\|\boldsymbol{v}|_K\right\|^2_{\boldsymbol{H}(\mathrm{div},K)}, \tag{A.13}$$

$$L^2(\mathcal{T}) = L^2(\Omega)\,. \tag{A.14}$$

The $\mathcal{T}$-broken L^2 inner product is

$$(u, v)_{\mathcal{T}} = \sum_{K\in\mathcal{T}} (u|_K, v|_K)_K\,, \tag{A.15}$$

where u and v can take scalar, vector, or matrix values, as stated previously.

Note that all functions have been assumed to take values in $\mathbb{R}$ throughout this section, but everything can be easily generalized to functions taking values in $\mathbb{C}$. For simplicity, we will continue to assume the field is $\mathbb{R}$, but all results will hold in $\mathbb{C}$ as well.

A.2 Fractional spaces

For the purposes of this work, it is only required to know that a set of spaces $H^s(K) \subseteq L^2(K)$ exist for all $s \geq 0$, such that $H^0(K) = L^2(K)$ and that $H^s(K)$ coincides with (A.8) when $s = 1$. However, for the sake of completeness, they are defined explicitly. Indeed, the fractional Sobolev spaces for $s \geq 0$ are

$$\begin{gathered} H^s(K) = \left\{u \in L^2(\Omega) \mid \|u\|_{H^s(K)} < \infty\right\}, \\ \|u\|_{H^s(K)} = \min_{U|_K = u, U \in L^2(\mathbb{R}^3)} \left\| \lim_{R\to\infty} \int_{|\boldsymbol{x}|<R} (1 + |\cdot|^2)^{s/2} e^{-\mathrm{i}2\pi(\cdot)\cdot\boldsymbol{x}} U(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right\|_{L^2(\mathbb{R}^3)}. \end{gathered} \tag{A.16}$$

This definition looks complicated, but can be written in simpler terms by using the Fourier transform and the Bessel potential, which were concepts we did not want to introduce here. For more details, refer to [170]. It is not difficult to see that $H^0(K) = L^2(K)$ (with the same norms) and it can be shown that when $s = 1$, $H^1(K)$ coincides with (A.8), so the notation is justified. The norms

$\|\cdot\|_{W^1(K)}$ and $\|\cdot\|_{H^1(K)}$ are equivalent, but sometimes in the literature the notation $\|\cdot\|_{H^1(K)}$ is meant to refer to $\|\cdot\|_{W^1(K)}$. Here, we distinguish between the two for the sake of clarity and consistency. It can be shown $H^{s_1}(K) \subseteq H^{s_2}(K)$ for $s_1 \geq s_2$, so in particular $H^s(K) \subseteq L^2(K)$ for all $s \geq 0$.

Let $\boldsymbol{H}^s(K) = \big(H^s(K)\big)^3$ be equipped with its natural Hilbert norm. Then, for all $s \geq 0$, the fractional counterparts to (A.9) and (A.10) are

$$\boldsymbol{H}^s(\mathrm{curl}, K) = \big\{\boldsymbol{E} \in \boldsymbol{H}^s(K) \mid \mathrm{curl}\,\boldsymbol{E} \in \boldsymbol{H}^s(K)\big\},\ \ \|\boldsymbol{E}\|^2_{\boldsymbol{H}^s(\mathrm{curl},K)} = \|\boldsymbol{E}\|^2_{\boldsymbol{H}^s(K)} + \|\,\mathrm{curl}\,\boldsymbol{E}\|^2_{\boldsymbol{H}^s(K)}\,, \tag{A.17}$$

$$\boldsymbol{H}^s(\mathrm{div}, K) = \big\{\boldsymbol{v} \in \boldsymbol{H}^s(K) \mid \mathrm{div}\,\boldsymbol{v} \in H^s(K)\big\},\qquad \|\boldsymbol{v}\|^2_{\boldsymbol{H}^s(\mathrm{div},K)} = \|\boldsymbol{v}\|^2_{\boldsymbol{H}^s(K)} + \|\,\mathrm{div}\,\boldsymbol{v}\|^2_{H^s(K)}\,, \tag{A.18}$$

so that $\boldsymbol{H}^0(\mathrm{curl}, K) = \boldsymbol{H}(\mathrm{curl}, K)$ and $\boldsymbol{H}^0(\mathrm{div}, K) = \boldsymbol{H}(\mathrm{div}, K)$.

Lastly, for a mesh, $\mathcal{T}$, of a domain Ω, define the fractional broken spaces $H^s(\mathcal{T})$, $\boldsymbol{H}^s(\mathrm{curl}, \mathcal{T})$ and $\boldsymbol{H}^s(\mathrm{div}, \mathcal{T})$ analogously to (A.11), (A.12) and (A.13), and once again it is obvious that $\boldsymbol{H}^0(\mathrm{curl}, \mathcal{T}) = \boldsymbol{H}(\mathrm{curl}, \mathcal{T})$, $\boldsymbol{H}^0(\mathrm{div}, \mathcal{T}) = \boldsymbol{H}(\mathrm{div}, \mathcal{T})$ and when $s = 1$, $H^s(\mathcal{T})$ is consistent with (A.11) and has equivalent norms.

A.3 Traces and boundary restrictions

Next, for a bounded Lipschitz domain $K \subseteq \mathbb{R}^3$ with outward normal $\hat{\mathbf{n}}_K$, consider the well-defined surjective trace operators,

$$\begin{aligned}
\mathrm{tr}^K_{\mathrm{grad}} = \mathrm{tr}_{H^1(K)} &: H^1(K) \to H^{1/2}(\partial K)\,,\\
\mathbf{tr}^K_{\mathrm{curl},\top} = \mathbf{tr}_{\boldsymbol{H}_\top(\mathrm{curl},K)} &: \boldsymbol{H}(\mathrm{curl}, K) \to \boldsymbol{H}^{-1/2}(\mathrm{curl}, \partial K)\,,\\
\mathbf{tr}^K_{\mathrm{curl},\dashv} = \mathbf{tr}_{\boldsymbol{H}_\dashv(\mathrm{curl},K)} &: \boldsymbol{H}(\mathrm{curl}, K) \to \boldsymbol{H}^{-1/2}(\mathrm{div}, \partial K)\,,\\
\mathrm{tr}^K_{\mathrm{div}} = \mathrm{tr}_{\boldsymbol{H}(\mathrm{div},K)} &: \boldsymbol{H}(\mathrm{div}, K) \to H^{-1/2}(\mathrm{div}, \partial K)\,,
\end{aligned} \tag{A.19}$$

which for $\mathcal{D}(\overline{K}) = \{\phi|_K \mid \phi \in C_0^\infty(\mathbb{R}^3)\}$ are defined as,

$$\begin{aligned}
\mathrm{tr}_{H^1(K)} u &= u|_{\partial K} && \forall u \in \mathcal{D}(\overline{K})\,,\\
\mathbf{tr}_{\boldsymbol{H}_\top(\mathrm{curl},K)} \boldsymbol{E} &= \big(\hat{\mathbf{n}}_K \times \boldsymbol{E}|_{\partial K}\big) \times \hat{\mathbf{n}}_K && \forall \boldsymbol{E} \in \big(\mathcal{D}(\overline{K})\big)^3\,,\\
\mathbf{tr}_{\boldsymbol{H}_\dashv(\mathrm{curl},K)} \boldsymbol{F} &= \big(\hat{\mathbf{n}}_K \times \boldsymbol{F}|_{\partial K}\big) && \forall \boldsymbol{F} \in \big(\mathcal{D}(\overline{K})\big)^3\,,\\
\mathrm{tr}_{\boldsymbol{H}(\mathrm{div},K)} \boldsymbol{v} &= \boldsymbol{v}|_{\partial K} \cdot \hat{\mathbf{n}}_K && \forall \boldsymbol{v} \in \big(\mathcal{D}(\overline{K})\big)^3\,.
\end{aligned} \tag{A.20}$$

Here, $\mathrm{tr}_{\boldsymbol{H}_{\top}(\mathrm{curl},K)}$ is a tangential trace, while $\mathrm{tr}_{\boldsymbol{H}_{\dashv}(\mathrm{curl},K)}$ is a rotated tangential trace. Before defining the norms, consider the following placeholder of ordered pairs,

$$\begin{array}{lcllll} \mathcal{H}(K) & \to & H^1(K)\,, & \boldsymbol{H}_{\top}(\mathrm{curl},K)\,, & \boldsymbol{H}_{\dashv}(\mathrm{curl},K)\,, & \boldsymbol{H}(\mathrm{div},K)\,, \\ \mathcal{H}(\partial K) & \to & H^{1/2}(\partial K)\,, & \boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)\,, & \boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)\,, & H^{-1/2}(\partial K)\,, \end{array} \tag{A.21}$$

where $\boldsymbol{H}_{\top}(\mathrm{curl},K)=\boldsymbol{H}_{\dashv}(\mathrm{curl},K)=\boldsymbol{H}(\mathrm{curl},K)$, and where $H^1(K)$ is given the norm $\|\cdot\|_{W^1(K)}$. Thus, (A.19) is simply $\mathrm{tr}_{\mathcal{H}(K)}:\mathcal{H}(K)\to\mathcal{H}(\partial K)$, and the norms for all the $\mathcal{H}(\partial K)$ spaces are

$$\|\hat{w}\|_{\mathcal{H}(\partial K)}=\min_{w\in\mathrm{tr}^{-1}_{\mathcal{H}(K)}\{\hat{w}\}}\|w\|_{\mathcal{H}(K)} \qquad \forall\hat{w}\in\mathcal{H}(\partial K)\,. \tag{A.22}$$

These norms are usually referred to as minimum energy extension norms and they obviously make the trace operators automatically continuous. Note the trace spaces themselves, $\mathcal{H}(\partial K)$, have not been defined. In fact, a possibility is to define them as a completion using the trace maps: let $\mathcal{D}_{\overline{K}}$ be $\mathcal{D}(\overline{K})$ or $\big(\mathcal{D}(\overline{K})\big)^3$ depending on $\mathcal{H}(K)$; note $\mathcal{D}_{\overline{K}}$ is dense in $\mathcal{H}(K)$, so consider a sequence $\{w_i\}_{i\in\mathbb{N}}\subseteq\mathcal{D}_{\overline{K}}$ converging to $w\in\mathcal{H}(K)$ in $\|\cdot\|_{\mathcal{H}(K)}$; let $\hat{w}_i=\mathrm{tr}_{\mathcal{H}(K)}w_i$ using (A.20) and temporarily redefine the minimum in (A.22) as an infimum with $\big(\mathrm{tr}_{\mathcal{H}(K)}|_{\mathcal{D}_{\overline{K}}}\big)^{-1}$ instead of $\mathrm{tr}^{-1}_{\mathcal{H}(K)}$, so that $\|\hat{w}_i-\hat{w}_j\|_{\mathcal{H}(\partial K)}\leq\|w_i-w_j\|_{\mathcal{H}(K)}$ is identified as a Cauchy sequence that must converge to some $\hat{w}$ in $\mathcal{H}(\partial K)=\overline{\mathrm{tr}_{\mathcal{H}(K)}(\mathcal{D}_{\overline{K}})}^{\|\cdot\|_{\mathcal{H}(\partial K)}}$ and is naturally identified with $\hat{w}=\mathrm{tr}_{\mathcal{H}(K)}w$. Alternatively, there are explicit characterizations in the literature. We refer to [170] for $H^{1/2}(\partial K)$ and $H^{-1/2}(\partial K)$, and to [45] for $\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)$ and $\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$.

The trace spaces are naturally dual to each other. Indeed, $\big(H^{1/2}(\partial K)\big)'=H^{-1/2}(\partial K)$ and $\big(\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)\big)'=\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$. Their duality pairing can be explicitly characterized for every $\hat{u}_K\in H^{1/2}(\partial K)$, $\hat{v}_{\mathbf{n}_K}\in H^{-1/2}(\partial K)$, $\hat{\boldsymbol{E}}_{\top_K}\in\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)$ and $\hat{\boldsymbol{F}}_{\dashv_K}\in\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$,

$$\begin{aligned} \langle\hat{u}_K,\hat{v}_{\mathbf{n}_K}\rangle_{\partial K}&=(u,\mathrm{div}\,\boldsymbol{v})_K+(\nabla u,\boldsymbol{v})_K\,,\\ \langle\hat{v}_{\mathbf{n}_K},\hat{u}_K\rangle_{\partial K}&=(\mathrm{div}\,\boldsymbol{v},u)_K+(\boldsymbol{v},\nabla u)_K\,,\\ \langle\hat{\boldsymbol{E}}_{\top_K},\hat{\boldsymbol{F}}_{\dashv_K}\rangle_{\partial K}&=(\boldsymbol{E},\mathrm{curl}\,\boldsymbol{F})_K-(\mathrm{curl}\,\boldsymbol{E},\boldsymbol{F})_K\,,\\ \langle\hat{\boldsymbol{F}}_{\dashv_K},\hat{\boldsymbol{E}}_{\top_K}\rangle_{\partial K}&=(\mathrm{curl}\,\boldsymbol{F},\boldsymbol{E})_K-(\boldsymbol{F},\mathrm{curl}\,\boldsymbol{E})_K\,, \end{aligned} \tag{A.23}$$

where the identities hold for any $u\in\mathrm{tr}^{-1}_{H^1(K)}\{\hat{u}_K\}$, $\boldsymbol{v}\in\mathrm{tr}^{-1}_{\boldsymbol{H}(\mathrm{div},K)}\{\hat{v}_{\mathbf{n}_K}\}$, $\boldsymbol{E}\in\mathbf{tr}^{-1}_{\boldsymbol{H}_{\top}(\mathrm{curl},K)}\{\hat{\boldsymbol{E}}_{\top_K}\}$ and $\boldsymbol{F}\in\mathbf{tr}^{-1}_{\boldsymbol{H}_{\dashv}(\mathrm{curl},K)}\{\hat{\boldsymbol{F}}_{\dashv_K}\}$. Note that all the duality pairings are encompassed by the same notation,

$\langle\cdot,\cdot\rangle_{\partial K}$, so the relevant duality pairing should be deduced from the context. The global notation, $\langle\cdot,\cdot\rangle_{\partial K}$, is partly justified because for smooth enough inputs it becomes a boundary integral, and intuitively this is a useful association.

Returning to the placeholder notation, $\mathcal{H}(K)$ and $\mathcal{H}(\partial K)$, the traces for fractional Sobolev spaces will now be defined. For $s > 0$, the new placeholders that complement the spaces in (A.21) in the same order are,

$$\begin{array}{ccccccc} \mathcal{H}^s(K) & \to & H^{1+s}(K)\,, & \boldsymbol{H}^s_{\top}(\mathrm{curl},K)\,, & \boldsymbol{H}^s_{\dashv}(\mathrm{curl},K)\,, & \boldsymbol{H}^s(\mathrm{div},K)\,, \\ \mathcal{H}^s(\partial K) & \to & H^{1/2+s}(\partial K)\,, & \boldsymbol{H}^{-1/2+s}(\mathrm{curl},\partial K)\,, & \boldsymbol{H}^{-1/2+s}(\mathrm{div},\partial K)\,, & H^{-1/2+s}(\partial K)\,, \end{array} \tag{A.24}$$

where $\boldsymbol{H}^s_{\top}(\mathrm{curl},K) = \boldsymbol{H}^s_{\dashv}(\mathrm{curl},K) = \boldsymbol{H}^s(\mathrm{curl},K)$. Note that these spaces were defined for $s > 0$. For $s = 0$ define $\mathcal{H}^0(K) = \mathcal{H}(K)$ and $\mathcal{H}^0(\partial K) = \mathcal{H}(\partial K)$, where the only detail to be aware is that the norm of $\mathcal{H}^0(K)$ will be $\|\cdot\|_{W^1(K)}$ (and not $\|\cdot\|_{H^1(K)}$) in the case of $H^1(K)$. It is useful to think about a space as belonging to the family $\mathcal{H}^s(\partial K)$ of order s, instead of thinking of them independently, since this avoids confusions that arise due to abuse of notation (e.g. $H^{1/2+s}(\partial K)$ when $s = 0$ is different from $H^{-1/2+s}(\partial K)$ when $s = 1$). To define $\mathcal{H}^s(\partial K)$, note that for all $s > 0$, $\mathcal{H}^s(K) \subseteq \mathcal{H}(K)$, so

$$\begin{gathered} \mathrm{tr}_{\mathcal{H}^s(K)} = \mathrm{tr}_{\mathcal{H}(K)}\big|_{\mathcal{H}^s(K)} : \mathcal{H}^s(K) \to \mathcal{H}^s(\partial K) = \mathrm{tr}_{\mathcal{H}(K)}\big(\mathcal{H}^s(K)\big)\,, \\ \mathrm{tr}_{\mathcal{H}^s(K)}w = \mathrm{tr}_{\mathcal{H}(K)}w \qquad \forall w \in \mathcal{H}^s(K)\,. \end{gathered} \tag{A.25}$$

Clearly $\mathcal{H}^s(\partial K) \subseteq \mathcal{H}(\partial K)$, and the associated norm is once again,

$$\|\hat{w}\|_{\mathcal{H}^s(\partial K)} = \min_{w\in\mathrm{tr}^{-1}_{\mathcal{H}^s(K)}\{\hat{w}\}} \|w\|_{\mathcal{H}^s(K)} \qquad \forall \hat{w} \in \mathcal{H}^s(\partial K)\,. \tag{A.26}$$

To consider traces in only part of a boundary, let $K = \Omega \subseteq \mathbb{R}^3$ be a bounded domain, and let $\Gamma \subseteq \partial\Omega$ be relatively open in $\partial\Omega$ (i.e. there exists an open set $\Gamma_{\mathbb{R}^3} \subseteq \mathbb{R}^3$ such that $\Gamma = \Gamma_{\mathbb{R}^3} \cap \partial\Omega$). Then, the spaces $\mathcal{H}^s(\Gamma)$ can be defined by restriction, and their functions characterized uniquely by testing with $\phi_\Gamma \in \mathcal{D}(\Gamma) = \{\phi_\Gamma = \widetilde{\phi}|_\Gamma \mid \widetilde{\phi} = \mathrm{tr}_{\mathrm{rev}(\mathcal{H}(\Omega))}\phi,\ \phi \in \mathcal{D}_{\overline{K}},\ (\mathrm{supp}\,\phi) \cap \partial\Omega \subseteq \Gamma\}$, where $\mathrm{rev}(\mathcal{H}(\Omega))$ runs over the list in (A.21) in reverse, so that $\mathrm{rev}(\mathcal{H}(\Omega)) = \boldsymbol{H}(\mathrm{div},\Omega)$ if $\mathcal{H}(\Omega) = H^1(\Omega)$, $\mathrm{rev}(\mathcal{H}(\Omega)) = \boldsymbol{H}_{\dashv}(\mathrm{curl},\Omega)$ if $\mathcal{H}(\Omega) = \boldsymbol{H}_{\top}(\mathrm{curl},\Omega)$, and so on. Indeed, for $s \geq 0$,

$$\mathcal{H}^s(\Gamma) = \{\hat{w}_\Gamma = \hat{w}|_\Gamma \mid \hat{w} \in \mathcal{H}^s(\partial\Omega)\}\,, \qquad \langle \hat{w}|_\Gamma, \phi\rangle_\Gamma = \langle \hat{w}, \widetilde{\phi}\rangle_{\partial\Omega} \quad \forall \phi \in \mathcal{D}(\Gamma)\,, \tag{A.27}$$

where $\widetilde{\phi}$ is the zero extension of ϕ to $\partial\Omega$, and where $\langle\cdot,\cdot\rangle_{\partial\Omega}$ is the duality pairing in (A.23) (so those identities could be used if necessary). The functions vanishing at $\Gamma \subseteq \partial\Omega$ are defined for $s \geq 0$ as

$$\mathcal{H}^s_\Gamma(\Omega) = \big\{ w \in \mathcal{H}^s(\Omega) \mid (\mathrm{tr}_{\mathcal{H}^s(K)} w)\big|_\Gamma = 0 \big\} \,, \tag{A.28}$$

and finally, when $\Gamma = \partial\Omega$ or $\Gamma = \varnothing$, the usual notation $\mathcal{H}^s_0(\Omega) = \mathcal{H}^s_{\partial\Omega}(\Omega)$ and $\mathcal{H}^s(\Omega) = \mathcal{H}^s_\varnothing(\Omega)$ is justifiably adopted.

Lastly, consider a mesh, $\mathcal{T}$, of the domain Ω with elements $K \in \mathcal{T}$. The mesh traces are $\mathcal{T}$-tuples defined as

$$\begin{gathered} \mathrm{tr}_{\mathcal{H}^s(\mathcal{T})} : \mathcal{H}^s(\mathcal{T}) \to \prod_{K\in\mathcal{T}} \mathcal{H}^s(\partial K) = \mathcal{H}^s_\Pi(\partial\mathcal{T}) \,, \\ \big(\mathrm{tr}_{\mathcal{H}^s(\mathcal{T})} w\big)_K = \mathrm{tr}_{\mathcal{H}^s(K)}(w|_K) \quad \forall w \in \mathcal{H}^s(\mathcal{T}) \,, \end{gathered} \tag{A.29}$$

where the space $\mathcal{H}^s_\Pi(\partial\mathcal{T})$ is given its natural Hilbert norm, and it follows

$$\|\hat{w}\|^2_{\mathcal{H}^s_\Pi(\partial\mathcal{T})} = \sum_{K\in\mathcal{T}} \|\hat{w}_K\|^2_{\mathcal{H}^s(\partial K)} = \sum_{K\in\mathcal{T}} \min_{w_K \in \mathrm{tr}^{-1}_{\mathcal{H}^s(K)}\{\hat{w}_K\}} \|w_K\|^2_{\mathcal{H}^s(K)} = \min_{w\in\mathrm{tr}^{-1}_{\mathcal{H}^s(\mathcal{T})}\{\hat{w}\}} \|w\|^2_{\mathcal{H}^s(\mathcal{T})} \,. \tag{A.30}$$

The spaces that are often of interest, are not those in the $\mathcal{H}^s_\Pi(\partial\mathcal{T})$ family, which are decoupled across elements, but rather those in $\mathcal{H}^s(\partial\mathcal{T})$, $\mathcal{H}^s_\Gamma(\partial\mathcal{T})$ and $\mathcal{H}^s_0(\partial\mathcal{T})$, which have interelement compatibility and even vanishing boundary conditions,

$$\mathcal{H}^s(\partial\mathcal{T}) = \mathrm{tr}_{\mathcal{H}^s(\mathcal{T})}\big(\mathcal{H}^s(\Omega)\big) \,, \quad \mathcal{H}^s_\Gamma(\partial\mathcal{T}) = \mathrm{tr}_{\mathcal{H}^s(\mathcal{T})}\big(\mathcal{H}^s_\Gamma(\Omega)\big) \,, \quad \mathcal{H}^s_0(\partial\mathcal{T}) = \mathrm{tr}_{\mathcal{H}^s(\mathcal{T})}\big(\mathcal{H}^s_0(\Omega)\big) \,, \tag{A.31}$$

where $\Gamma \subseteq \partial\Omega$ is relatively open in $\partial\Omega$. In all the spaces above, the norm is relabeled for cosmetic purposes as $\|\cdot\|_{\mathcal{H}^s(\partial\mathcal{T})} = \|\cdot\|_{\mathcal{H}^s_\Pi(\partial\mathcal{T})}$. All the subspaces of $\mathcal{H}^s_\Pi(\partial\mathcal{T})$ are often referred to as interface or skeleton spaces. When $s = 0$ (drop the superscript) and $\hat{w} \in \mathcal{H}(\partial\mathcal{T})$, then

$$\|\hat{w}\|_{\mathcal{H}(\partial\mathcal{T})} = \min_{w\in\mathrm{tr}^{-1}_{\mathcal{H}(\mathcal{T})}\{\hat{w}\}} \|w\|_{\mathcal{H}(\Omega)} \qquad \forall \hat{w} \in \mathcal{H}(\partial\mathcal{T}) \,, \tag{A.32}$$

since Theorem A.1 implies $\mathrm{tr}^{-1}_{\mathcal{H}(\mathcal{T})}\{\hat{w}\} \subseteq \mathcal{H}(\Omega) \subseteq \mathcal{H}(\mathcal{T})$. Moreover, when $s = 0$, there is a duality within the spaces in the family $\mathcal{H}_\Pi(\partial\mathcal{T})$ analogous to (A.23), so that the mesh duality pairings are

$$\big\langle \hat{w}_1, \hat{w}_2 \big\rangle_{\partial\mathcal{T}} = \sum_{K\in\mathcal{T}} \big\langle (\hat{w}_1)_K, (\hat{w}_2)_K \big\rangle_{\partial K} \,, \tag{A.33}$$

where $(\hat{w}_1, \hat{w}_2)$ is in $H^{1/2}_\Pi(\partial\mathcal{T}) \times H^{-1/2}_\Pi(\partial\mathcal{T})$, $\boldsymbol{H}^{-1/2}_\Pi(\mathrm{curl}, \partial\mathcal{T}) \times \boldsymbol{H}^{-1/2}_\Pi(\mathrm{div}, \partial\mathcal{T})$ or their permutations. Finally, when $s = 0$ another notation often used to replace $\mathrm{tr}_{\mathcal{H}(\partial\mathcal{T})}$ is

$$\mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} = \mathrm{tr}_{H^1(\partial\mathcal{T})} \,, \quad \mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\top} = \mathbf{tr}_{\boldsymbol{H}_\top(\mathrm{curl},\partial\mathcal{T})} \,, \quad \mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\dashv} = \mathbf{tr}_{\boldsymbol{H}_\dashv(\mathrm{curl},\partial\mathcal{T})} \,, \quad \mathrm{tr}^{\mathcal{T}}_{\mathrm{div}} = \mathrm{tr}_{\boldsymbol{H}(\mathrm{div},\partial\mathcal{T})} \,. \tag{A.34}$$

A.4 Duality theorems

The following theorem was first proved in [55], and was slightly generalized in [158], whose proof we present below.

Theorem A.1. *Let Γ_0 and Γ_1 be relatively open in $\partial\Omega$, and satisfy $\overline{\Gamma_0 \cup \Gamma_1} = \partial\Omega$ and $\Gamma_0 \cap \Gamma_1 = \varnothing$.*

(i) *Let $u \in H^1(\mathcal{T})$. Then $u \in H^1_{\Gamma_0}(\Omega)$ if and only if $\langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} u \rangle_{\partial\mathcal{T}} = 0$ for all $\hat{\sigma}_{\mathbf{n}} \in H^{-1/2}_{\Gamma_1}(\partial\mathcal{T})$.*

(ii) *Let $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}, \mathcal{T})$. Then $\boldsymbol{\sigma} \in \boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega)$ if and only if $\langle \hat{u}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{div}} \boldsymbol{\sigma} \rangle_{\partial\mathcal{T}} = 0$ for all $\hat{u} \in H^{1/2}_{\Gamma_0}(\partial\mathcal{T})$.*

(iii) *Let $\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \mathcal{T})$. Then $\boldsymbol{E} \in \boldsymbol{H}_{\top,\Gamma_0}(\mathrm{curl}, \Omega)$ if and only if $\langle \hat{\boldsymbol{F}}_{\dashv}, \mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\top} \boldsymbol{E} \rangle_{\partial\mathcal{T}} = 0$ for all $\hat{\boldsymbol{F}}_{\dashv} \in \boldsymbol{H}^{-1/2}_{\dashv,\Gamma_1}(\mathrm{div}, \partial\mathcal{T})$.*

(iv) *Let $\boldsymbol{F} \in \boldsymbol{H}(\mathrm{curl}, \mathcal{T})$. Then $\boldsymbol{F} \in \boldsymbol{H}_{\dashv,\Gamma_1}(\mathrm{curl}, \Omega)$ if and only if $\langle \hat{\boldsymbol{E}}_{\top}, \mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\dashv} \boldsymbol{F} \rangle_{\partial\mathcal{T}} = 0$ for all $\hat{\boldsymbol{E}}_{\top} \in \boldsymbol{H}^{-1/2}_{\top,\Gamma_0}(\mathrm{curl}, \partial\mathcal{T})$.*

Proof. Only the first equivalence is proved, because the other three follow similarly.

Let $u \in H^1_{\Gamma_0}(\Omega)$ and $\hat{\sigma}_{\mathbf{n}} \in H^{-1/2}_{\Gamma_1}(\partial\mathcal{T})$. By definition of $H^{-1/2}_{\Gamma_1}(\partial\mathcal{T})$, there exists some $\boldsymbol{\sigma} \in \boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega)$ such that $\mathrm{tr}^{\mathcal{T}}_{\mathrm{div}} \boldsymbol{\sigma} = \hat{\sigma}_{\mathbf{n}}$. Given a domain K, for all $u \in H^1(K)$ and $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}, K)$, the following identity holds,

$$(\boldsymbol{\sigma}, \nabla u)_K + (\mathrm{div}\, \boldsymbol{\sigma}, u)_K = \langle \mathrm{tr}^K_{\mathrm{div}} \boldsymbol{\sigma}, \mathrm{tr}^K_{\mathrm{grad}} u \rangle_{\partial K}\,, \tag{A.35}$$

and in particular if $u \in H^1_{\Gamma_0}(\Omega)$ and $\boldsymbol{\sigma} \in \boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega)$ the following identity holds,

$$(\boldsymbol{\sigma}, \nabla u)_\Omega + (\mathrm{div}\, \boldsymbol{\sigma}, u)_\Omega = \langle \mathrm{tr}^\Omega_{\mathrm{div}} \boldsymbol{\sigma}, \mathrm{tr}^\Omega_{\mathrm{grad}} u \rangle_{\partial\Omega} = 0\,. \tag{A.36}$$

Hence, rewriting the integral $(\boldsymbol{\sigma}, \nabla u)_\Omega + (\mathrm{div}\, \boldsymbol{\sigma}, u)_\Omega = 0$ as a sum of integrals over each element in the mesh and using the first identity yields the result,

$$0 = \sum_{K \in \mathcal{T}} (\boldsymbol{\sigma}, \nabla u)_K + (\mathrm{div}\, \boldsymbol{\sigma}, u)_K = \sum_{K \in \mathcal{T}} \langle \mathrm{tr}^K_{\mathrm{div}} \boldsymbol{\sigma}, \mathrm{tr}^K_{\mathrm{grad}} u \rangle_{\partial K} = \langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} u \rangle_{\partial\mathcal{T}}\,. \tag{A.37}$$

For the converse assume $u \in H^1(\mathcal{T})$, so that $u|_K \in H^1(K)$ for any $K \in \mathcal{T}$. Next, let $\hat{\sigma}_{\mathbf{n}} \in H^{-1/2}_{\Gamma_1}(\partial\mathcal{T})$, so that there exists $\boldsymbol{\sigma} \in \boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega)$ satisfying $\mathrm{tr}^{\mathcal{T}}_{\mathrm{div}} \boldsymbol{\sigma} = \hat{\sigma}_{\mathbf{n}}$. Define v such that $v|_K = \nabla(u|_K)$, meaning that $v \in L^2(\Omega)$. Then, using the hypothesis and the distributional identities gives,

$$0 = \langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}} u \rangle_{\partial\mathcal{T}} = \sum_{K\in\mathcal{T}} (\boldsymbol{\sigma}, \nabla(u|_K))_K + (\operatorname{div}\boldsymbol{\sigma}, u|_K)_K = (\boldsymbol{\sigma}, v)_\Omega + (\operatorname{div}\boldsymbol{\sigma}, u)_\Omega\,. \tag{A.38}$$

In particular, for any smooth test function $\boldsymbol{\sigma}$, it holds that $(v, \boldsymbol{\sigma})_\Omega = -(u, \operatorname{div}\boldsymbol{\sigma})_\Omega$. This means $v = \nabla u$ is the distributional derivative of u, so that $u \in H^1(\Omega)$. Next, let $\phi \in \mathcal{D}(\Gamma_0)$ (i.e. a test function defined on Γ_0 with support in Γ_0). Then, its zero extension to $\partial\Omega$, $\widetilde{\phi}$, satisfies that $\widetilde{\phi} = \mathrm{tr}^{\Omega}_{\mathrm{div}} \widetilde{\boldsymbol{\phi}} \in \mathrm{tr}^{\Omega}_{\mathrm{div}}(\boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega))$, for some smooth $\widetilde{\boldsymbol{\phi}} \in \boldsymbol{H}_{\Gamma_1}(\mathrm{div}, \Omega)$ so that $\mathrm{tr}^{\Omega}_{\mathrm{div}} \widetilde{\boldsymbol{\phi}}\big|_{\Gamma_0} = \phi$. By definition of distributional restriction and the previous equality, it follows

$$\langle \mathrm{tr}^{\Omega}_{\mathrm{grad}} u\big|_{\Gamma_0}, \phi \rangle_{\Gamma_0} = \langle \mathrm{tr}^{\Omega}_{\mathrm{grad}} u, \mathrm{tr}^{\Omega}_{\mathrm{div}} \widetilde{\boldsymbol{\phi}} \rangle_{\partial\Omega} = (\nabla u, \widetilde{\boldsymbol{\phi}})_\Omega + (u, \operatorname{div}\widetilde{\boldsymbol{\phi}})_\Omega = 0\,, \tag{A.39}$$

where the first distributional identity was utilized. This is true for all smooth test functions ϕ, implying $\mathrm{tr}^{\Omega}_{\mathrm{grad}} u\big|_{\Gamma_0} = 0$, so that $u \in H^1_{\Gamma_0}(\Omega)$. □

The following two theorems were proved in [55]. They show that the minimum energy extension norms of the the trace spaces are actually dual to each other.

Theorem A.2. *Let $K \subseteq \mathbb{R}^3$ be a bounded Lipschitz domain. Then, for every $\hat{u} \in H^{1/2}(\partial K)$, $\hat{\sigma}_{\mathbf{n}} \in H^{-1/2}(\partial K)$, $\hat{\boldsymbol{E}}_\top \in \boldsymbol{H}^{-1/2}(\mathrm{curl}, \partial K)$ and $\hat{\boldsymbol{F}}_\dashv \in \boldsymbol{H}^{-1/2}(\mathrm{div}, \partial K)$,*

$$\begin{aligned} \|\hat{u}\|_{H^{1/2}(\partial K)} = \min_{u\in(\mathrm{tr}^K_{\mathrm{grad}})^{-1}\{\hat{u}\}} \|u\|_{W^1(K)} &= \sup_{\hat{\sigma}_{\mathbf{n}}\in H^{-1/2}(\partial K)} \frac{|\langle \hat{u}, \hat{\sigma}_{\mathbf{n}} \rangle_{\partial K}|}{\|\hat{\sigma}_{\mathbf{n}}\|_{H^{-1/2}(\partial K)}} \\ &= \sup_{\boldsymbol{\sigma}\in\boldsymbol{H}(\mathrm{div},K)} \frac{|\langle \hat{u}, \mathrm{tr}^K_{\mathrm{div}}\boldsymbol{\sigma} \rangle_{\partial K}|}{\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},K)}}\,, \end{aligned} \tag{A.40}$$

$$\begin{aligned} \|\hat{\sigma}_{\mathbf{n}}\|_{H^{-1/2}(\partial K)} = \min_{\boldsymbol{\sigma}\in(\mathrm{tr}^K_{\mathrm{div}})^{-1}\{\hat{\sigma}_{\mathbf{n}}\}} \|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},K)} &= \sup_{\hat{u}\in H^{1/2}(\partial K)} \frac{|\langle \hat{\sigma}_{\mathbf{n}}, \hat{u} \rangle_{\partial K}|}{\|\hat{u}\|_{H^{1/2}(\partial K)}} \\ &= \sup_{u\in H^1(K)} \frac{|\langle \hat{\sigma}_{\mathbf{n}}, \mathrm{tr}^K_{\mathrm{grad}} u \rangle_{\partial K}|}{\|u\|_{W^1(K)}}\,, \end{aligned} \tag{A.41}$$

$$\begin{aligned} \|\hat{\boldsymbol{E}}_\top\|_{\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)} = \min_{\boldsymbol{E}\in(\mathbf{tr}^K_{\mathrm{curl},\top})^{-1}\{\hat{\boldsymbol{E}}_\top\}} \|\boldsymbol{E}\|_{\boldsymbol{H}(\mathrm{curl},K)} &= \sup_{\hat{\boldsymbol{F}}_\dashv\in\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)} \frac{|\langle \hat{\boldsymbol{E}}_\top, \hat{\boldsymbol{F}}_\dashv \rangle_{\partial K}|}{\|\hat{\boldsymbol{F}}_\dashv\|_{\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)}} \\ &= \sup_{\boldsymbol{F}\in\boldsymbol{H}(\mathrm{curl},K)} \frac{|\langle \hat{\boldsymbol{E}}_\top, \mathbf{tr}^K_{\mathrm{curl},\dashv}\boldsymbol{F} \rangle_{\partial K}|}{\|\boldsymbol{F}\|_{\boldsymbol{H}(\mathrm{curl},K)}}\,, \end{aligned} \tag{A.42}$$

$$\begin{aligned}\|\hat{\boldsymbol{F}}_{\dashv}\|_{\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)} = \min_{\boldsymbol{F}\in(\mathbf{tr}^K_{\mathrm{curl},\dashv})^{-1}\{\hat{\boldsymbol{F}}_\dashv\}} \|\boldsymbol{F}\|_{\boldsymbol{H}(\mathrm{curl},K)} &= \sup_{\hat{\boldsymbol{E}}_\top\in\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)} \frac{|\langle\hat{\boldsymbol{F}}_\dashv,\hat{\boldsymbol{E}}_\top\rangle_{\partial K}|}{\|\hat{\boldsymbol{E}}_\top\|_{\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)}}\\ &= \sup_{\boldsymbol{E}\in\boldsymbol{H}(\mathrm{curl},K)} \frac{|\langle\hat{\boldsymbol{F}}_\dashv,\mathbf{tr}^K_{\mathrm{curl},\top}\boldsymbol{E}\rangle_{\partial K}|}{\|\boldsymbol{E}\|_{\boldsymbol{H}(\mathrm{curl},K)}}\,.\end{aligned} \tag{A.43}$$

Theorem A.3. *Let $\mathcal{T}$ be a mesh of elements $K\in\mathcal{T}$ discretizing a bounded Lipschitz domain $\Omega\subseteq\mathbb{R}^3$. Then, for every $\hat{u}\in H^{1/2}_{\Pi}(\partial\mathcal{T})$, $\hat{\sigma}_{\mathbf{n}}\in H^{-1/2}_{\Pi}(\partial\mathcal{T})$, $\hat{\boldsymbol{E}}_\top\in\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{curl},\partial\mathcal{T})$ and $\hat{\boldsymbol{F}}_\dashv\in\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{div},\partial\mathcal{T})$,*

$$\|\hat{u}\|_{H^{1/2}_{\Pi}(\partial\mathcal{T})} = \min_{u\in(\mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}})^{-1}\{\hat{u}\}} \|u\|_{W^1(\mathcal{T})} = \sup_{\boldsymbol{\sigma}\in\boldsymbol{H}(\mathrm{div},\mathcal{T})} \frac{|\langle\hat{u},\mathrm{tr}^{\mathcal{T}}_{\mathrm{div}}\boldsymbol{\sigma}\rangle_{\partial\mathcal{T}}|}{\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},\mathcal{T})}}\,, \tag{A.44}$$

$$\|\hat{\sigma}_{\mathbf{n}}\|_{H^{-1/2}_{\Pi}(\partial\mathcal{T})} = \min_{\boldsymbol{\sigma}\in(\mathrm{tr}^{\mathcal{T}}_{\mathrm{div}})^{-1}\{\hat{\sigma}_{\mathbf{n}}\}} \|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div},\mathcal{T})} = \sup_{u\in H^1(\mathcal{T})} \frac{|\langle\hat{\sigma}_{\mathbf{n}},\mathrm{tr}^{\mathcal{T}}_{\mathrm{grad}}u\rangle_{\partial\mathcal{T}}|}{\|u\|_{W^1(\mathcal{T})}}\,, \tag{A.45}$$

$$\|\hat{\boldsymbol{E}}_\top\|_{\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{curl},\partial\mathcal{T})} = \min_{\boldsymbol{E}\in(\mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\top})^{-1}\{\hat{\boldsymbol{E}}_\top\}} \|\boldsymbol{E}\|_{\boldsymbol{H}(\mathrm{curl},\mathcal{T})} = \sup_{\boldsymbol{F}\in\boldsymbol{H}(\mathrm{curl},\mathcal{T})} \frac{|\langle\hat{\boldsymbol{E}}_\top,\mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\dashv}\boldsymbol{F}\rangle_{\partial\mathcal{T}}|}{\|\boldsymbol{F}\|_{\boldsymbol{H}(\mathrm{curl},\mathcal{T})}}\,, \tag{A.46}$$

$$\|\hat{\boldsymbol{F}}_\dashv\|_{\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{div},\partial\mathcal{T})} = \min_{\boldsymbol{F}\in(\mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\dashv})^{-1}\{\hat{\boldsymbol{F}}_\dashv\}} \|\boldsymbol{F}\|_{\boldsymbol{H}(\mathrm{curl},\mathcal{T})} = \sup_{\boldsymbol{E}\in\boldsymbol{H}(\mathrm{curl},\mathcal{T})} \frac{|\langle\hat{\boldsymbol{F}}_\dashv,\mathbf{tr}^{\mathcal{T}}_{\mathrm{curl},\top}\boldsymbol{E}\rangle_{\partial\mathcal{T}}|}{\|\boldsymbol{E}\|_{\boldsymbol{H}(\mathrm{curl},\mathcal{T})}}\,. \tag{A.47}$$

If $\hat{u}\in H^{1/2}(\partial\mathcal{T})$, $\hat{\sigma}_{\mathbf{n}}\in H^{-1/2}(\partial\mathcal{T})$, $\hat{\boldsymbol{E}}_\top\in\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial\mathcal{T})$ and $\hat{\boldsymbol{F}}_\dashv\in\boldsymbol{H}^{-1/2}(\mathrm{div},\partial\mathcal{T})$, then the Π-subcript can be dropped from the expressions above, and the norms $\|\cdot\|_{W^1(\Omega)}$, $\|\cdot\|_{\boldsymbol{H}(\mathrm{div},\Omega)}$ and $\|\cdot\|_{\boldsymbol{H}(\mathrm{curl},\Omega)}$ can be used in the expressions for the minimum energy extension norms.

A.5 Sobolev-de Rham spaces and discretizations

The aim of this section is to define the concept of Sobolev-de Rham spaces and discretizations. The definition of the former is found below.

Definition A.1. *Let $\Omega\subseteq\mathbb{R}^3$ be a domain partitioned into elements, $K\in\mathcal{T}$, of a mesh, $\mathcal{T}$. If a Hilbert space $\mathcal{U}$ is comprised of products of closed subspaces of $H^1(\mathcal{T})$, $\boldsymbol{H}(\mathrm{curl},\mathcal{T})$, $\boldsymbol{H}(\mathrm{div},\mathcal{T})$, $L^2(\mathcal{T})$, $H^{1/2}_{\Pi}(\partial\mathcal{T})$, $\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{curl},\partial\mathcal{T})$, $\boldsymbol{H}^{-1/2}_{\Pi}(\mathrm{div},\partial\mathcal{T})$ and $H^{-1/2}_{\Pi}(\partial\mathcal{T})$, then $\mathcal{U}$ is said to be a 3D Sobolev-de Rham space, or simply and an SdR space. If it is only comprised of products of closed subspaces of $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl},\Omega)$, $\boldsymbol{H}(\mathrm{div},\Omega)$, $L^2(\Omega)$, then it is said to be a compatible SdR space. The definition is analogous for domains in 2D and 1D.*

Remark A.1. Compatible SdR spaces only depend on Ω, so are independent of any mesh.

Next, consider a simply connected polyhedron $K \subseteq \mathbb{R}^3$ with simply connected faces. Assume there exists a family of high-order finite-dimensional discretizations of the spaces $H^1(K)$, $\boldsymbol{H}(\mathrm{curl}, K)$, $\boldsymbol{H}(\mathrm{div}, K)$ and $L^2(K)$ forming a differential de Rham exact sequence (or complex) as follows,

$$\begin{tikzcd}
H^1(K) & \boldsymbol{H}(\mathrm{curl}, K) & \boldsymbol{H}(\mathrm{div}, K) & L^2(K) \\
W^p(K) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\nabla"] & \boldsymbol{Q}^p(K) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\mathrm{curl}"] & \boldsymbol{V}^p(K) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\mathrm{div}"] & Y^p(K) \arrow[u, phantom, sloped, "\subseteq"] \\
\mathcal{P}^p \arrow[u, phantom, sloped, "\subseteq"] & (\mathcal{P}^{p-1})^3 \arrow[u, phantom, sloped, "\subseteq"] & (\mathcal{P}^{p-1})^3 \arrow[u, phantom, sloped, "\subseteq"] & \mathcal{P}^{p-1} \arrow[u, phantom, sloped, "\subseteq"]
\end{tikzcd} \tag{A.48}$$

where $\mathcal{P}^p$ are the high-order polynomials in $\boldsymbol{x} = (x_1, x_2, x_3)$ of total order at most p. Moreover, for each polygnal face $F \subseteq \partial K$ of K, consider its image in 2D via an affine mapping, $\hat{F} \subseteq \mathbb{R}^2$, and assume that the 2D affine pullbacks also form differential de Rham exact sequences,

$$\begin{tikzcd}
H^1(\hat{F}) & \boldsymbol{H}(\mathrm{curl}, \hat{F}) & L^2(\hat{F}) & H^1(\hat{F}) & \boldsymbol{H}(\mathrm{curl}, \hat{F}) & L^2(\hat{F}) \\
W^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\nabla"] & \boldsymbol{Q}^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\nabla\times"] & Y^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] & W^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\mathrm{curl}"] & \boldsymbol{V}^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\mathrm{div}"] & Y^p(\hat{F}) \arrow[u, phantom, sloped, "\subseteq"] \\
\mathcal{P}^p \arrow[u, phantom, sloped, "\subseteq"] & (\mathcal{P}^{p-1})^2 \arrow[u, phantom, sloped, "\subseteq"] & \mathcal{P}^{p-1} \arrow[u, phantom, sloped, "\subseteq"] & \mathcal{P}^p \arrow[u, phantom, sloped, "\subseteq"] & (\mathcal{P}^{p-1})^2 \arrow[u, phantom, sloped, "\subseteq"] & \mathcal{P}^{p-1} \arrow[u, phantom, sloped, "\subseteq"]
\end{tikzcd} \tag{A.49}$$

where $W^p(\hat{F})$, $\boldsymbol{Q}^p(\hat{F})$, $\boldsymbol{V}^p(\hat{F})$ and $Y^p(\hat{F})$ are 2D affine pullbacks of the spaces

$$\begin{aligned}
W^p(F) &= \left\{ \hat{u} = (\mathrm{tr}^K_{\mathrm{grad}} u)\big|_F \mid u \in W^p(K) \right\}, \\
\boldsymbol{Q}^p(F) &= \left\{ \hat{\boldsymbol{E}} = (\mathbf{tr}^K_{\mathrm{curl},\top} \boldsymbol{E})\big|_F \mid \boldsymbol{E} \in \boldsymbol{Q}^p(K) \right\}, \\
\boldsymbol{V}^p(F) &= \left\{ \hat{\boldsymbol{v}} = (\mathbf{tr}^K_{\mathrm{curl},\dashv} \boldsymbol{F})\big|_F \mid \boldsymbol{F} \in \boldsymbol{Q}^p(K) \right\}, \\
Y^p(F) &= \left\{ \hat{v} = (\mathrm{tr}^K_{\mathrm{div}} \boldsymbol{v})\big|_F \mid \boldsymbol{v} \in \boldsymbol{V}^p(K) \right\},
\end{aligned} \tag{A.50}$$

to the pulled back polygonal face $\hat{F} \subseteq \mathbb{R}^2$, and where $\mathcal{P}^p$ refers to polynomials in $\boldsymbol{x} = (x_1, x_2)$ instead. Note that if one of the sequences in (A.49) is exact, then the other automatically is too. Finally, for every edge $E \subseteq F \subseteq \partial K$ of each face F of K, consider its image in 1D via an affine mapping, $\hat{E} \subseteq \mathbb{R}$, and assume that the 1D affine pullback forms a differential de Rham exact sequence,

$$\begin{tikzcd}
H^1(\hat{E}) & L^2(\hat{E}) \\
W^p(\hat{E}) \arrow[u, phantom, sloped, "\subseteq"] \arrow[r, "\nabla"] & Y^p(\hat{E}) \arrow[u, phantom, sloped, "\subseteq"] \\
\mathcal{P}^p \arrow[u, phantom, sloped, "\subseteq"] & \mathcal{P}^{p-1} \arrow[u, phantom, sloped, "\subseteq"]
\end{tikzcd} \tag{A.51}$$

where $W^p(\hat{E})$ and $Y^p(\hat{E})$ are 1D affine pullbacks of the spaces

$$\begin{aligned} W^p(E) &= \big\{\hat{u} = \big(\mathrm{tr}_{\mathrm{grad}}^F(\mathrm{tr}_{\mathrm{grad}}^K u)|_F\big)\big|_E \mid u \in W^p(K)\big\}\,, \\ Y^p(E) &= \big\{\hat{v} = \big(\mathrm{tr}_{\mathrm{div}}^F(\mathbf{tr}_{\mathrm{curl},\dashv}^K \boldsymbol{F})|_F\big)\big|_E \mid \boldsymbol{F} \in \boldsymbol{Q}^p(K)\big\}\,, \end{aligned} \tag{A.52}$$

to the pulled back segment $\hat{E} \subseteq \mathbb{R}$, and where $\mathcal{P}^p$ are now polynomials in x.

Discretizations of the traces of these spaces, $H^{1/2}(\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)$, $\boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)$ and $H^{-1/2}(\partial K)$, are defined naturally as

$$\begin{aligned} W^p(\partial K) &= \big\{\hat{u}_K = \mathrm{tr}_{\mathrm{grad}}^K u \mid u \in W^p(K)\big\} \subseteq H^{1/2}(\partial K)\,, \\ \boldsymbol{Q}_{\top}^p(\partial K) &= \big\{\hat{\boldsymbol{E}}_{\top_K} = \mathbf{tr}_{\mathrm{curl},\top}^K \boldsymbol{E} \mid \boldsymbol{E} \in \boldsymbol{Q}^p(K)\big\} \subseteq \boldsymbol{H}^{-1/2}(\mathrm{curl},\partial K)\,, \\ \boldsymbol{Q}_{\dashv}^p(\partial K) &= \big\{\hat{\boldsymbol{F}}_{\dashv_K} = \mathbf{tr}_{\mathrm{curl},\top}^K \boldsymbol{F} \mid \boldsymbol{F} \in \boldsymbol{Q}^p(K)\big\} \subseteq \boldsymbol{H}^{-1/2}(\mathrm{div},\partial K)\,, \\ V^p(\partial K) &= \big\{\hat{v}_{\mathbf{n}_K} = \mathrm{tr}_{\mathrm{grad}}^K \boldsymbol{v} \mid \boldsymbol{v} \in \boldsymbol{V}^p(K)\big\} \subseteq H^{-1/2}(\partial K)\,. \end{aligned} \tag{A.53}$$

Definition A.2. *Let $\Omega \subseteq \mathbb{R}^3$ be a polyhedral domain partitioned into a mesh, $\mathcal{T}$, comprised of simply connected polyhedral elements $K \in \mathcal{T}$ with simply connected faces, and let $\mathfrak{U}$ be an SdR space. Without loss of generality, let $\mathfrak{U}=\mathfrak{U}_0\times\hat{\mathfrak{U}}$, where $\mathfrak{U}_0$ is comprised of copies of closed subspaces of $H^1(\mathcal{T})$, $\boldsymbol{H}(\mathrm{curl},\mathcal{T})$, $\boldsymbol{H}(\mathrm{div},\mathcal{T})$ and $L^2(\mathcal{T})$, whereas $\hat{\mathfrak{U}}$ is comprised of copies of closed subspaces of $H_{\Pi}^{1/2}(\partial\mathcal{T})$, $\boldsymbol{H}_{\Pi}^{-1/2}(\mathrm{curl},\partial\mathcal{T})$, $\boldsymbol{H}_{\Pi}^{-1/2}(\mathrm{div},\partial\mathcal{T})$ and $H_{\Pi}^{-1/2}(\partial\mathcal{T})$. Then, every $\mathfrak{u} = (\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathfrak{U}$ is naturally associated to a $\mathcal{T}$-tuple, $(\mathfrak{u}|_K)_{K\in\mathcal{T}}$, where $\mathfrak{u}|_K = \big(\mathfrak{u}_0|_K,\hat{\mathfrak{u}}_K\big)$ with $\mathfrak{u}_0|_K$ being a restriction from Ω to K, and $\hat{\mathfrak{u}}_K$ being the natural K-component of the $\mathcal{T}$-tuple. Assume that for every $K \in \mathcal{T}$ there exists a high-order discretization satisfying* (A.48)–(A.52)*, and let $\mathfrak{U}_h^p(K) = \mathfrak{U}_{0,h}^p(K)\times\hat{\mathfrak{U}}_h^p(K)$ be defined so that $\mathfrak{U}_{0,h}^p(K)$ has the corresponding copies of $W^p(K)$, $\boldsymbol{Q}^p(K)$, $\boldsymbol{V}^p(K)$ and $Y^p(K)$ subordinated to $\mathfrak{U}_0$, while $\hat{\mathfrak{U}}_h^p(K)$ has the corresponding copies of $W^p(\partial K)$, $\boldsymbol{Q}_{\top}^p(\partial K)$, $\boldsymbol{Q}_{\dashv}^p(\partial K)$ and $V^p(\partial K)$ subordinated to $\hat{\mathfrak{U}}$. Then,*

$$\mathfrak{U}_h^p = \big\{\mathfrak{u} = (\mathfrak{u}_0,\hat{\mathfrak{u}}) \in \mathfrak{U} \mid \mathfrak{u}|_K \in \mathfrak{U}_h^p(K)\big\}\,, \tag{A.54}$$

is said to be a Sobolev-de Rham discretization of order p, or simply an SdR discretization. The SdR discretization is additionally said to be compatible if the relevant traces of the $\mathfrak{U}_h^p(K)$ match exactly on common faces between elements. The definition is analogous for relevant domains in 2D and 1D.

There are several SdR discretizations associated to the affine conventional element shapes, such as triangles and quadrilaterals (parallelograms) in 2D; and tetrahedra, hexahedra (parallelepipeds), triangular prisms (parallelogram-based prisms) and pyramids (parallelogram-based pyramids) in 3D. In particular, the classical Nédélec sequences of the first type for triangles, quadrilaterals, tetrahedra and hexahedra [181, 103, 34] are SdR discretizations, but others exist as well [205, 236]. There are also examples for triangular prisms and pyramids [103, 185, 114, 74, 125, 1], and even some efforts made for (lowest-order) general polygons [63]. Of particular interest are unified constructions that allow for compatible SdR discretizations across a mesh with elements of different types [114, 74, 103, 24, 25, 57, 1], because they allow to construct globally conforming discretizations of the subspaces of $H^1(\Omega)$, $\boldsymbol{H}(\mathrm{curl},\Omega)$, $\boldsymbol{H}(\mathrm{div},\Omega)$, $L^2(\Omega)$, $H^{1/2}(\partial\mathcal{T})$, $\boldsymbol{H}^{-1/2}(\mathrm{curl},\partial\mathcal{T})$, $\boldsymbol{H}^{-1/2}(\mathrm{div},\partial\mathcal{T})$ and $H^{-1/2}(\partial\mathcal{T})$ (as opposed to their broken counterparts).

A.6 Interpolation estimates

Let $\Omega\subseteq\mathbb{R}^3$ be a polyhedral domain partitioned into polyhedral elements $K\in\mathcal{T}$ from the mesh $\mathcal{T}$, let $\mathcal{U}$ be an SdR space, and let $\mathcal{U}_h^p$ be a corresponding SdR discretization of order p. Return to the placeholder notation, so that if $\mathcal{U}=\mathcal{U}_0\times\hat{\mathcal{U}}$, $\mathcal{U}_0$ is comprised of copies of subspaces of the spaces in $\mathcal{H}(\mathcal{T})$ (append $L^2(\mathcal{T})=H^0(\mathcal{T})$ to (A.24) and elsewhere when it makes sense), while $\hat{\mathcal{U}}$ of copies of subspaces of the spaces in $\mathcal{H}_\Pi(\partial\mathcal{T})$. For $s\geq 0$ recall the fractional counterparts are $\mathcal{H}^s(\mathcal{T})$ and $\mathcal{H}_\Pi^s(\partial\mathcal{T})$, and using them it is easy to construct the fractional counterpart to $\mathcal{U}$ as $\mathcal{U}^s\subseteq\mathcal{U}$. Meanwhile, denote their discretization by $\mathcal{H}_h^p(\mathcal{T})$ and $\mathcal{H}_h^p(\partial\mathcal{T})$ respectively, which in turn can be viewed as $\mathcal{T}$-tuples of $\mathcal{H}_h^p(K)$ (taken from (A.48)) and $\mathcal{H}_h^p(\partial K)$ (taken from (A.53)). These are the building blocks of $\mathcal{U}_h^p$. The idea is to construct an interpolation operator for a range of s,

$$\Pi_{\mathcal{U}^s}:\mathcal{U}^s\to\mathcal{U}_h^p\,. \tag{A.55}$$

This operator will be constructed from its components, which in general take the form,

$$\begin{aligned}
&\Pi_{\mathcal{H}^s(\mathcal{T})}:\mathcal{H}^s(\mathcal{T})\to\mathcal{H}_h^p(\mathcal{T})\,, &&\Big(\Pi_{\mathcal{H}^s(\mathcal{T})}w\Big)\Big|_K=\Pi_{\mathcal{H}^s(K)}(w|_K) &&\forall K\in\mathcal{T}\,,\\
&\Pi_{\mathcal{H}_\Pi^s(\partial\mathcal{T})}:\mathcal{H}_\Pi^s(\partial\mathcal{T})\to\mathcal{H}_h^p(\partial\mathcal{T})\,, &&\Big(\Pi_{\mathcal{H}_\Pi^s(\partial\mathcal{T})}\hat{w}\Big)_K=\Pi_{\mathcal{H}^s(\partial K)}\hat{w}_K &&\forall K\in\mathcal{T}\,,
\end{aligned} \tag{A.56}$$

where the second operator can be defined explicitly for any $\hat{w}_K \in \mathcal{H}^s(\partial K)$ as

$$\Pi_{\mathcal{H}^s(\partial K)}\hat{w}_K = \mathrm{tr}_{\mathcal{H}^s(\partial K)}\big(\Pi_{\mathcal{H}^s(K)}w\big) \qquad w \in \mathrm{tr}^{-1}_{\mathcal{H}^s(\partial K)}\{\hat{w}_K\}\,. \tag{A.57}$$

This operator is only well-defined if $\Pi_{\mathcal{H}^s(K)}w$ at ∂K only depends on the trace of w at ∂K. Therefore, provided that condition is held, $\Pi_{\mathfrak{U}^s}$ is completely determined by the local operators $\Pi_{\mathcal{H}^s(K)} : \mathcal{H}^s(K) \to \mathcal{H}^p_h(K)$ at each $K \in \mathcal{T}$, and the task at hand is to define them.

For $s > \frac{1}{2}$, one such general construction is provided by projection-based interpolation [90],

$$\begin{tikzcd} H^{1+s}(K) \arrow[r, "\nabla"] \arrow[d, "\Pi_{H^{1+s}(K)}"] & \boldsymbol{H}^s(\mathrm{curl}, K) \arrow[r, "\mathrm{curl}"] \arrow[d, "\Pi_{\boldsymbol{H}^s(\mathrm{curl},K)}"] & \boldsymbol{H}^s(\mathrm{div}, K) \arrow[r, "\mathrm{div}"] \arrow[d, "\Pi_{\boldsymbol{H}^s(\mathrm{div},K)}"] & H^s(K) \arrow[d, "\Pi_{H^s(K)}"] \\ W^p(K) \arrow[r, "\nabla"] & \boldsymbol{Q}^p(K) \arrow[r, "\mathrm{curl}"] & \boldsymbol{V}^p(K) \arrow[r, "\mathrm{div}"] & Y^p(K) \end{tikzcd} \tag{A.58}$$

where $K \in \mathcal{T}$ can be any polyhedral element with an SdR discretization. As most interpolation operators, they are designed to satisfy interelement compatibility, in the sense that the interpolation of a function on each face and edge depends only on the relevant trace of the function in that face or edge. Hence, (A.57) will be well-defined, and the resulting mesh interpolants $\Pi_{\mathcal{H}^s(\mathcal{T})}$ and $\Pi_{\mathcal{H}^s(\partial\mathcal{T})}$ are applicable to globally conforming discretizations involving closed subspaces of $\mathcal{H}(\Omega) \subseteq \mathcal{H}(\mathcal{T})$ and $\mathcal{H}(\partial\mathcal{T}) \subseteq \mathcal{H}_\Pi(\partial\mathcal{T})$. More importantly, these interpolation operators commute, so

$$\nabla\Pi_{H^{1+s}(K)}{=}\Pi_{\boldsymbol{H}^s(\mathrm{curl},K)}\nabla\,,\ \mathrm{curl}\,\Pi_{\boldsymbol{H}^s(\mathrm{curl},K)}{=}\Pi_{\boldsymbol{H}^s(\mathrm{div},K)}\mathrm{curl}\,,\ \mathrm{div}\,\Pi_{\boldsymbol{H}^s(\mathrm{div},K)}{=}\Pi_{H^s(K)}\mathrm{div}\ . \tag{A.59}$$

Additionally, the operators are continuous (the range is measured in the $L^2(K)$ norm) and they satisfy $\Pi_{\mathcal{H}^s(K)}w_h = w_h$ for all $w_h \in \mathcal{H}^p_h(K)$. Lastly, if $\kappa : K \to \hat{K}$, and $\kappa_{\mathcal{H}} : \mathcal{H}(K) \to \mathcal{H}(\hat{K})$ is the induced pullback, then the property $\kappa_{\mathcal{H}}\big(\Pi_{\mathcal{H}^s(K)}w\big) = \Pi_{\mathcal{H}^s(\hat{K})}\kappa_{\mathcal{H}}w$ holds (by definition) for κ being an affine mapping.

Next, assume that $s > \frac{1}{2}$ and $p \in \mathbb{N}$. Using the properties of projection-based interpolation, it can be shown that [91],

$$\|w - \Pi_{\mathcal{H}^s(K)}w\|_{\mathcal{H}(K)} \leq C_{\mathcal{H}(\hat{K})}h_K^{\min\{s,p\}}\|w\|_{\mathcal{H}^s(K)} \qquad \forall w \in \mathcal{H}^s(K)\,, \tag{A.60}$$

where $C_{\mathcal{H}(\hat{K})} = C_{\mathcal{H}(\hat{K})}\big(s, p, \mathcal{H}(\hat{K})\big) > 0$, $h_K = \mathrm{diam}(K)$ and $\hat{K}$ is an affine master element version of K such that $\mathrm{diam}(\hat{K}) = 1$. A sketch of the procedure to get this estimate is: consider $p \geq r \in \mathbb{N}$;

transform K to $\hat{K}$ so the pullback is preserved as noted above; use commutativity, invariance under the space $\mathcal{H}_h^p(\hat{K})$ and boundedness of the $\Pi_{\mathcal{H}^r(K)}$ to get a bound in terms of the best approximation error; define relevant seminorms $|\cdot|_{W^r(\hat{K})}$ and use the approximation properties of the space (the polynomials that are contained in it) to bound the best approximation error in terms of the $|\cdot|_{W^r(\hat{K})}$ seminorms of the important quantities in $\mathcal{H}^r(\hat{K})$ and a constant dependent on r and $\mathcal{H}^r(\hat{K})$; transform back all the estimates to K, which involves powers of h_K, and rearrange to get the bound in terms of h_K; then generalize to s using the K-method to interpolate [170, 23].

Next, notice the traces will also satisfy (A.60): let $\hat{w}_K \in \mathcal{H}^s(\partial K)$, then use (A.57), (A.25) ($\mathrm{tr}_{\mathcal{H}^s(K)} = \mathrm{tr}_{\mathcal{H}(K)}$ on $\mathcal{H}^s(K)$), (A.22), and (A.60) to obtain that

$$\begin{aligned}\|\hat{w}_K - \Pi_{\mathcal{H}^s(\partial K)}\hat{w}_K\|_{\mathcal{H}(\partial K)} &= \big\|\mathrm{tr}_{\mathcal{H}^s(K)}\big(w - \Pi_{\mathcal{H}^s(K)}w\big)\big\|_{\mathcal{H}(\partial K)} \leq \|w - \Pi_{\mathcal{H}^s(K)}w\|_{\mathcal{H}(K)} \\ &\leq C_{\mathcal{H}(\hat{K})} h_K^{\min\{s,p\}} \|w\|_{\mathcal{H}^s(K)} \quad \forall w \in \mathrm{tr}_{\mathcal{H}^s(K)}^{-1}\{\hat{w}_K\}\,.\end{aligned} \tag{A.61}$$

This is true for every $w \in \mathrm{tr}_{\mathcal{H}^s(K)}^{-1}\{\hat{w}_K\}$, so take the infimum to get (see (A.26)),

$$\|\hat{w}_K - \Pi_{\mathcal{H}^s(\partial K)}\hat{w}_K\|_{\mathcal{H}(\partial K)} \leq C_{\mathcal{H}(\hat{K})} h_K^{\min\{s,p\}} \|\hat{w}_K\|_{\mathcal{H}^s(\partial K)} \qquad \forall \hat{w}_K \in \mathcal{H}^s(\partial K)\,. \tag{A.62}$$

In what follows, for a given K, choose the maximum $C_{\mathcal{H}(\hat{K})}$ among all four $\mathcal{H}(\hat{K})$, and call it $C_{\hat{K}} = C_{\hat{K}}(s, p, \hat{K})$ so it no longer depends on $\mathcal{H}(\hat{K})$, but still depends on the domain. To obtain the estimates across the whole mesh, $\mathcal{T}$, consider (A.56) for $w \in \mathcal{H}^s(\mathcal{T})$ and $\hat{w} \in \mathcal{H}_\Pi^s(\partial\mathcal{T})$,

$$\begin{aligned}\|w - \Pi_{\mathcal{H}^s(\mathcal{T})}w\|_{\mathcal{H}(\mathcal{T})}^2 &= \sum_{K\in\mathcal{T}} \|w|_K - \Pi_{\mathcal{H}^s(K)}(w|_K)\|_{\mathcal{H}(K)}^2 \leq C_{\mathcal{T}}^2 h_{\mathcal{T}}^{2\min\{s,p\}} \|w\|_{\mathcal{H}^s(\mathcal{T})}^2\,, \\ \|\hat{w} - \Pi_{\mathcal{H}_\Pi^s(\partial\mathcal{T})}\hat{w}\|_{\mathcal{H}_\Pi(\partial\mathcal{T})}^2 &= \sum_{K\in\mathcal{T}} \|\hat{w}_K - \Pi_{\mathcal{H}^s(\partial K)}\hat{w}_K\|_{\mathcal{H}(\partial K)}^2 \leq C_{\mathcal{T}}^2 h_{\mathcal{T}}^{2\min\{s,p\}} \|\hat{w}\|_{\mathcal{H}_\Pi^s(\partial\mathcal{T})}^2\,,\end{aligned} \tag{A.63}$$

where $C_{\mathcal{T}} = C_{\mathcal{T}}(s, p) = \max_{K\in\mathcal{T}} C_{\hat{K}}$ and $h_{\mathcal{T}} = \max_{K\in\mathcal{T}} \mathrm{diam}(K)$. Due to the compatibility of the interpolation, the result holds true even if $w \in \mathcal{H}^s(\Omega) \subseteq \mathcal{H}^s(\mathcal{T})$ and $\hat{w} \in \mathcal{H}^s(\partial\mathcal{T}) \subseteq \mathcal{H}_\Pi^s(\partial\mathcal{T})$ or their closed subspaces, provided the SdR discretization is compatible. Thus, adding all the variables in $\mathcal{U}$ yields,

$$\|\mathfrak{u} - \Pi_{\mathcal{U}^s}\mathfrak{u}\|_{\mathcal{U}} \leq C_{\mathcal{T}} h_{\mathcal{T}}^{\min\{s,p\}} \|\mathfrak{u}\|_{\mathcal{U}^s} \qquad \forall \mathfrak{u} \in \mathcal{U}^s\,. \tag{A.64}$$

Note the norm $\|\mathfrak{u}\|_{\mathcal{U}^s}$ technically depends on $\mathcal{T}$, since in general $\mathcal{U}$ depends on $\mathcal{T}$. If a series of meshes is considered, one would be interested in having an estimate that does not depend on the

$\mathcal{T}$ beyond the element size, $h_{\mathcal{T}}$ (so $C_{\mathcal{T}}$ and $\|\mathfrak{u}\|_{\mathfrak{U}^s}$ should lose their dependence on $\mathcal{T}$). With this in mind, take into account the following definitions.

Definition A.3. *Let $K \subseteq \mathbb{R}^3$ be a simply connected polyhedral domain with simply connected faces, $s > \frac{1}{2}$ and $p \in \mathbb{N}$. Let $C(s,p) > 0$ be a fixed shape-regularity function of s and p. Then K is said to be $C(s,p)$ shape-regular, or simply shape-regular, if its Sobolev-de Rham domain constant, $C_{\hat{K}}(s,p,\hat{K})$, shown above, is always bounded above by $C(s,p)$ for all s and p.*

Definition A.4. *Let $\Omega \subseteq \mathbb{R}^3$ be a domain and $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ be family of meshes of Ω with elements $K \in \mathcal{T}_\mathfrak{h}$. Let $\mathfrak{U}_\Omega = \mathfrak{U}_{0,\Omega} \times \widetilde{\mathfrak{U}}_\Omega$ be a compatible SdR space, and for every $\mathfrak{h} \in \mathfrak{H}$, let $\mathfrak{U}_\mathfrak{h} = \mathfrak{U}_{0,\mathfrak{h}} \times \hat{\mathfrak{U}}_\mathfrak{h}$ be SdR spaces. Define $\mathrm{tr}_{\widetilde{\mathfrak{U}}^s_\Omega(\mathcal{T}_\mathfrak{h})}$ naturally for each mesh. If $\mathfrak{u}_\mathfrak{h} = (\mathfrak{u}_0, \mathrm{tr}_{\widetilde{\mathfrak{U}}^s_\Omega(\mathcal{T}_\mathfrak{h})}\widetilde{\mathfrak{u}}) \in \mathfrak{U}_\mathfrak{h}$ for some fixed $\mathfrak{u} = (\mathfrak{u}_0, \widetilde{\mathfrak{u}}) \in \mathfrak{U}_\Omega$, then $\{\mathfrak{u}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ is said to be attached to $\mathfrak{u}$ through $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$, and it is clear $\|\mathfrak{u}_\mathfrak{h}\|_{\mathfrak{U}_\mathfrak{h}} \leq \|\mathfrak{u}\|_{\mathfrak{U}_\Omega}$ for all such $\mathfrak{u}_\mathfrak{h}$.*

Theorem A.4. *Let $\Omega \subseteq \mathbb{R}^3$ be a polyhedral domain and $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ be family of polyhedral meshes of Ω comprised of simply connected polyhedral elements $K \in \mathcal{T}_\mathfrak{h}$ with simply connected faces. For every $\mathfrak{h} \in \mathfrak{H}$, let $\mathfrak{U}_\mathfrak{h}$ be an SdR space, and $\mathfrak{U}^p_\mathfrak{h}$ be a corresponding compatible SdR discretization of order $p \in \mathbb{N}$. Let $s > \frac{1}{2}$ and $\mathfrak{U}^s_\mathfrak{h} \subseteq \mathfrak{U}_\mathfrak{h}$ be the fractional counterpart to $\mathfrak{U}_\mathfrak{h}$. Then,*

$$\|\mathfrak{u} - \Pi_{\mathfrak{U}^s_\mathfrak{h}}\mathfrak{u}\|_{\mathfrak{U}_\mathfrak{h}} \leq C_\mathfrak{h} h_\mathfrak{h}^{\min\{s,p\}} \|\mathfrak{u}\|_{\mathfrak{U}^s_\mathfrak{h}} \qquad \forall \mathfrak{u} \in \mathfrak{U}^s_\mathfrak{h}\,, \tag{A.65}$$

where $C_\mathfrak{h} = C_\mathfrak{h}(s,p) = \max_{K\in\mathcal{T}_\mathfrak{h}} C_{\hat{K}}(s,p,\hat{K})$ and $h_\mathfrak{h} = \max_{K\in\mathcal{T}_\mathfrak{h}} \mathrm{diam}(K)$. Let $\{\mathfrak{u}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ be attached to $\mathfrak{u}_\Omega \in \mathfrak{U}_\Omega$ through $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$, where $\mathfrak{U}_\Omega$ is a compatible SdR space. Then, if $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ is a shape-regular family of meshes, it follows

$$\|\mathfrak{u}_\mathfrak{h} - \Pi_{\mathfrak{U}^s_\mathfrak{h}}\mathfrak{u}_\mathfrak{h}\|_{\mathfrak{U}_\mathfrak{h}} \leq C h_\mathfrak{h}^{\min\{s,p\}} \|\mathfrak{u}_\Omega\|_{\mathfrak{U}^s_\Omega}\,, \tag{A.66}$$

where $C = C(s,p)$ is a uniform shape-regularity bound, and $h_\mathfrak{h} = \max_{K\in\mathcal{T}_\mathfrak{h}} \mathrm{diam}(K)$. Additionally, if all the elements in the meshes $\{\mathcal{T}_\mathfrak{h}\}_{\mathfrak{h}\in\mathfrak{H}}$ are tetrahedra or hexahedra, then there is an hp-convergence estimate where $C = C_s(\ln p)^2 p^{-s}$ with $C_s = C(s)$ being independent of p.

Remark A.2. The main reason why the p-convergence estimates is only valid for tetrahedra and hexahedra is that polynomial preserving extension operators with a continuity bound independent of p have been proved for those two elements across the whole sequence [90, 99, 100, 101, 84].

Appendix B

Well-posedness in linear elasticity

This chapter aims to prove two theorems associated to the well-posedness of variational formulations of the equations of linear elasticity. The first one is Theorem 3.1, whose proof will use techniques from functional analysis to prove the mutual well-posedness of a family of eight variational formulations. The particular tool that will be utilized is the closed range theorem, in both of its settings (as it applies to closed operators and continuous operators). This technique is interesting, as it could be used in other frameworks, like proving inequalities. The second one is Theorem 3.3, which proves that the coupled variational formulations introduced in Section 3.2.6 are also well-posed. The compelling aspect of this proof is its use of the ultraweak formulations as a tool within the proof. In Section B.1, Theorem 3.1 is proved, while in Section B.2, Theorem 3.3 is proved.

B.1 Mutual well-posedness

The goal is to prove Theorem 3.1. Throughout this section we assume $\Omega \subseteq \mathbb{R}^3$ is a three-dimensional bounded simply connected domain with a Lipschitz boundary $\partial\Omega = \overline{\Gamma_u \cup \Gamma_\sigma}$, where Γ_u and Γ_σ are disjoint and relatively open in $\partial\Omega$. Note the results hold in two and one-dimensional domains as well. Well-posedness and stability estimates are proved using the well-known result by Babuška and Nečas.

Theorem B.1 (Babuška-Nečas). *Let $\mathcal{U}$ and $\mathcal{V}$ be Hilbert spaces over a fixed field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, $b : \mathcal{U} \times \mathcal{V} \to \mathbb{F}$ be a continuous bilinear form if $\mathbb{F} = \mathbb{R}$ or sesquilinear form if $\mathbb{F} = \mathbb{C}$, and $\ell : \mathcal{V} \to \mathbb{F}$ be a continuous linear form if $\mathbb{F} = \mathbb{R}$ or antilinear form if $\mathbb{F} = \mathbb{C}$. If there exists an inf-sup constant $\gamma > 0$ such that for all $\mathfrak{u} \in \mathcal{U}$,*

$$\sup_{\mathfrak{v} \in \mathcal{V} \setminus \{0\}} \frac{|b(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}}} \geq \gamma \|\mathfrak{u}\|_{\mathcal{U}} \,, \tag{B.1}$$

and ℓ satisfies the compatibility condition

$$\ell(\mathfrak{v}) = 0 \qquad \forall \mathfrak{v} \in \mathcal{V}_{00} = \big\{ \mathfrak{v} \in \mathcal{V} \mid b(\mathfrak{u}, \mathfrak{v}) = 0 \;\; \forall \mathfrak{u} \in \mathcal{U} \big\} \,, \tag{B.2}$$

then the problem of finding $\mathfrak{u} \in \mathcal{U}$ such that

$$b(\mathfrak{u}, \mathfrak{v}) = \ell(\mathfrak{u}) \qquad \forall \mathfrak{v} \in \mathcal{V} \,, \tag{B.3}$$

is well-posed in the sense of Hadamard, so that there exists a unique solution $\mathfrak{u} \in \mathcal{U}$ satisfying the stability estimate $\|\mathfrak{u}\|_{\mathcal{U}} \leq \frac{1}{\gamma} \|\ell\|_{\mathcal{V}'}$.

For ease of reference, the variational formulations of linear elasticity which will be shown to be well-posed, (3.15)–(3.22), are repeated here again,

$$\begin{aligned} &\mathcal{U}^{\mathcal{S}_{\mathbb{S}}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \qquad \mathcal{V}^{\mathcal{S}_{\mathbb{S}}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \,, \\ &b^{\mathcal{S}_{\mathbb{S}}}\big((\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\tau}, \boldsymbol{v})\big) = (\boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega - (\mathsf{C} : \boldsymbol{\nabla} \boldsymbol{u}, \boldsymbol{\tau})_\Omega - (\mathbf{div}\, \boldsymbol{\sigma}, \boldsymbol{v})_\Omega \,, \end{aligned} \tag{B.4}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{U}_{\mathbb{S}}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \,, \qquad \mathcal{V}^{\mathcal{U}_{\mathbb{S}}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \\ &b^{\mathcal{U}_{\mathbb{S}}}\big((\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\tau}, \boldsymbol{v})\big) = (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (\boldsymbol{u}, \mathbf{div}\, \boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma}, \boldsymbol{\nabla} \boldsymbol{v})_\Omega \,, \end{aligned} \tag{B.5}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{M}_{\mathbb{S}}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \,, \qquad \mathcal{V}^{\mathcal{M}_{\mathbb{S}}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \,, \\ &b^{\mathcal{M}_{\mathbb{S}}}\big((\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\tau}, \boldsymbol{v})\big) = (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (\boldsymbol{u}, \mathbf{div}\, \boldsymbol{\tau})_\Omega - (\mathbf{div}\, \boldsymbol{\sigma}, \boldsymbol{v})_\Omega \,, \end{aligned} \tag{B.6}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{S}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \qquad \mathcal{V}^{\mathcal{S}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \times \mathbf{L}^2(\Omega; \mathbb{A}) \,, \\ &b^{\mathcal{S}}\big((\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{w})\big) = (\boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega - (\mathsf{C} : \boldsymbol{\nabla} \boldsymbol{u}, \boldsymbol{\tau})_\Omega - (\mathbf{div}\, \boldsymbol{\sigma}, \boldsymbol{v})_\Omega + (\boldsymbol{\sigma}, \boldsymbol{w})_\Omega \,, \end{aligned} \tag{B.7}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{U}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{L}^2(\Omega) \times \mathbf{L}^2(\Omega; \mathbb{A}) \,, \qquad \mathcal{V}^{\mathcal{U}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \\ &b^{\mathcal{U}}\big((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\omega}), (\boldsymbol{\tau}, \boldsymbol{v})\big) = (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (\boldsymbol{\omega}, \boldsymbol{\tau})_\Omega + (\boldsymbol{u}, \mathbf{div}\, \boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma}, \boldsymbol{\nabla} \boldsymbol{v})_\Omega \,, \end{aligned} \tag{B.8}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{M}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \times \boldsymbol{L}^2(\Omega) \times \mathbf{L}^2(\Omega; \mathbb{A}) \,, \qquad \mathcal{V}^{\mathcal{M}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \times \boldsymbol{L}^2(\Omega) \times \mathbf{L}^2(\Omega; \mathbb{A}) \,, \\ &b^{\mathcal{M}}\big((\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\omega}), (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{w})\big) = (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (\boldsymbol{\omega}, \boldsymbol{\tau})_\Omega + (\boldsymbol{u}, \mathbf{div}\, \boldsymbol{\tau})_\Omega - (\mathbf{div}\, \boldsymbol{\sigma}, \boldsymbol{v})_\Omega + (\boldsymbol{\sigma}, \boldsymbol{w})_\Omega \,, \end{aligned} \tag{B.9}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{D}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \qquad \mathcal{V}^{\mathcal{D}} = \mathbf{L}^2(\Omega; \mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \\ &b^{\mathcal{D}}\big((\boldsymbol{\sigma}, \boldsymbol{u}), (\boldsymbol{\tau}, \boldsymbol{v})\big) = (\boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega - (\mathsf{C} : \boldsymbol{\nabla} \boldsymbol{u}, \boldsymbol{\tau})_\Omega + (\boldsymbol{\sigma}, \boldsymbol{\nabla} \boldsymbol{v})_\Omega \,, \end{aligned} \tag{B.10}$$

$$\begin{aligned} &\mathcal{U}^{\mathcal{P}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \qquad \mathcal{V}^{\mathcal{P}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega) \,, \\ &b^{\mathcal{P}}\big(\boldsymbol{u}, \boldsymbol{v}\big) = (\mathsf{C} : \boldsymbol{\nabla} \boldsymbol{u}, \boldsymbol{\nabla} \boldsymbol{v})_\Omega \,. \end{aligned} \tag{B.11}$$

The proof of mutual well-posedness is discussed in two parts. First, the mutual satisfaction of the compatibility conditions is analyzed. Second, the inf-sup constants are also shown to be mutually satisfied.

Throughout, note that the proofs only hold in the compressible regime. Here, C and S are inverse to each other over $\mathbb{S}$. This is no longer true in the incompressible case (in the limit of $\lambda \to \infty$), where only the variational formulations that make use of S can be proved to remain well-posed.

B.1.1 Compatibility conditions

Well-posedness of the variational formulations depends on the nature of Γ_u and Γ_σ. The first lemma shows that $\Gamma_u \neq \varnothing$ is a necessary condition for all variational formulations to be well-posed. The condition is also sufficient, and this is the content of Theorem 3.1.

Lemma B.1. *Suppose one of the variational formulations among* (B.4)–(B.11) *is well-posed. Then* $\Gamma_u \neq \varnothing$*.*

Proof. Assume the hypothesis so that the well-posed variational formulation has a unique solution $\mathfrak{u}$, whose component $\boldsymbol{u}$ is the displacement solution variable. By contradiction assume $\Gamma_u = \varnothing$. Then any translation (constant) $\boldsymbol{u}_C$ satisfies the boundary conditions vacuously and $\boldsymbol{\nabla}\boldsymbol{u}_C = 0$. For the variational formulations $(\text{B.4})^{\mathcal{S}_\mathbb{S}}$, $(\text{B.7})^{\mathcal{S}}$, $(\text{B.10})^{\mathcal{D}}$ and $(\text{B.11})^{\mathcal{P}}$ it is straightforward that, ceteris paribus, the solution $\mathfrak{u}_C$ with displacement component $\boldsymbol{u} + \boldsymbol{u}_C$ is a different solution (provided $\boldsymbol{u}_C \neq 0$) to the original problem. Similarly, since $\boldsymbol{u}_C \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\boldsymbol{\nabla}\boldsymbol{u}_C = 0$, the distributional identity yields $(\boldsymbol{u}_C, \mathbf{div}\,\boldsymbol{\tau})_\Omega = -(\boldsymbol{\nabla}\boldsymbol{u}_C, \boldsymbol{\tau})_\Omega = 0$ for all $\boldsymbol{\tau} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$, so that $\mathfrak{u}_C$ is also a different solution to the variational formulations $(\text{B.5})^{\mathcal{U}_\mathbb{S}}$, $(\text{B.8})^{\mathcal{U}}$, $(\text{B.6})^{\mathcal{M}_\mathbb{S}}$ and $(\text{B.9})^{\mathcal{M}}$. This contradicts that the original solution was unique. □

The next lemma shows that the solution to the original elasticity equation with homogeneous forcing and boundary conditions is $\boldsymbol{u} = 0$ and is unique provided $\Gamma_u \neq \varnothing$.

Lemma B.2. *Suppose* $\Gamma_u \neq \varnothing$ *and consider the equation* $-\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u})) = 0$ *in* Ω*, where* $\boldsymbol{u}$ *is sought in* $\boldsymbol{H}^1_{\Gamma_u}(\Omega)$ *and* $\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$*. Then* $\boldsymbol{u} = 0$ *is the unique solution to the problem.*

Proof. Multiplying the equation by a test function $\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$, integrating and using a distributional identity yields the equation $(\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v})_\Omega = 0$ for all $\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$, which is precisely the formulation $(\text{B.11})^{\mathcal{P}}$ with $\boldsymbol{f} = 0$. Using Korn's inequality and that $\Gamma_u \neq \varnothing$, the bilinear form can be shown to be coercive, meaning $b^{\mathcal{P}}(\boldsymbol{u}, \boldsymbol{u}) = (\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{u})_\Omega \geq \alpha \|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)}$ for some $\alpha > 0$ [70]. Taking $\boldsymbol{v} = \boldsymbol{u}$, the equation becomes $b^{\mathcal{P}}(\boldsymbol{u}, \boldsymbol{u}) = 0$, and using coercivity it implies $\|\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)} = 0$, so that $\boldsymbol{u} = 0$ is the only solution. □

Finally, it is shown that given $\Gamma_u \neq \varnothing$, the compatibility condition is satisfied trivially for every variational formulation.

Lemma B.3. *Let* $\Gamma_u \neq \varnothing$*. Then the variational formulations* (B.4)–(B.11) *all have a trivial compatibility space, implying that the compatibility conditions,* (B.2)*, are satisfied trivially.*

Proof. First consider $(\text{B.7})^{\mathcal{S}}$. The aim is to prove $\mathcal{V}^{\mathcal{S}}_{00} = \{0\}$. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathcal{U}^{\mathcal{S}}$, with $\boldsymbol{u} = 0$ and $\boldsymbol{\sigma}$ being any smooth symmetric matrix field vanishing at the boundary. The condition $b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}) = 0$ then becomes $(\boldsymbol{\tau}, \boldsymbol{\sigma})_\Omega - (\boldsymbol{v}, \mathbf{div}\,\boldsymbol{\sigma})_\Omega = 0$, which yields the distributional equality $-\boldsymbol{\varepsilon}(\boldsymbol{v}) = \boldsymbol{\tau} \in \mathsf{L}^2(\Omega; \mathbb{S})$. By Korn's inequality, $\boldsymbol{v} \in \boldsymbol{H}^1(\Omega)$, and further testing against $\boldsymbol{\sigma} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S})$ yields additionally that $\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$. Next, test with $\boldsymbol{\sigma} = 0$ and $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$, so that $b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}) = 0$ now implies that $(\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v})_\Omega = 0$, which can be rewritten as $-\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{v})) = 0$. By Lemma B.2, $\boldsymbol{v} = 0$, meaning $\boldsymbol{\tau} = -\boldsymbol{\varepsilon}(\boldsymbol{v}) = 0$. Finally, $b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}) = 0$ becomes $(\boldsymbol{\sigma}, \boldsymbol{w})_\Omega = 0$ when testing with $\boldsymbol{\sigma} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$ (non-symmetric), which results in $\boldsymbol{w} = 0$ as well. Therefore $\mathfrak{u} = (0, 0, 0)$ is the only element of $\mathcal{V}^{\mathcal{S}}_{00}$.

Next consider $(\text{B.8})^{\mathcal{U}}$ and the condition $b^{\mathcal{U}}(\mathfrak{u}, \mathfrak{v}) = 0$ for all $\mathfrak{u} = (\boldsymbol{u}, \boldsymbol{\sigma}, \boldsymbol{\omega}) \in \mathcal{U}^{\mathcal{U}}$. First let $\boldsymbol{\sigma} = 0$ and $\boldsymbol{u} = 0$, so that the condition becomes $(\boldsymbol{\omega}, \boldsymbol{\tau})_\Omega = 0$. Therefore, the antisymmetric part of $\boldsymbol{\tau}$ vanishes, meaning $\boldsymbol{\tau} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega; \mathbb{S})$. Then, with $\boldsymbol{\sigma} = 0$, the condition becomes $(\boldsymbol{u}, \mathbf{div}\,\boldsymbol{\tau})_\Omega = 0$, so that $\mathbf{div}\,\boldsymbol{\tau} = 0$. Finally, test with $\boldsymbol{u} = 0$, so that the condition yields the equation $\mathsf{S} : \boldsymbol{\tau} + \boldsymbol{\varepsilon}(\boldsymbol{v}) = 0$, which can be rewritten as $\boldsymbol{\tau} = -\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{v})$. Taking the divergence and using $\mathbf{div}\,\boldsymbol{\tau} = 0$ gives $-\mathbf{div}\big(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{v})\big) = 0$, which by Lemma B.2 results in $\boldsymbol{v} = 0$ and $\boldsymbol{\tau} = -\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{v}) = 0$. Hence, $\mathcal{V}_{00}^{\mathcal{U}} = \{0\}$.

Similar calculations follow for $\text{(B.10)}^{\mathcal{D}}$, $\text{(B.9)}^{\mathcal{M}}$, $\text{(B.11)}^{\mathcal{P}}$, $\text{(B.4)}^{\mathcal{S}_{\mathbb{S}}}$, $\text{(B.5)}^{\mathcal{U}_{\mathbb{S}}}$ and $\text{(B.6)}^{\mathcal{M}_{\mathbb{S}}}$. □

When $\Gamma_u = \varnothing$ it is possible to redefine some spaces by considering the quotient over a particular null space (e.g. rigid body motions). This essentially produces new closely related, yet modified variational formulations which are sometimes well-posed even when $\Gamma_u = \varnothing$. Indeed, after redefining these spaces, a relevant version of Korn's inequality can be proved to hold [69, Theorem 2.3]. However, in this work we will not be dealing with those cases.

B.1.2 Boundedness-below constants

Before proceeding to the main result, the most challenging results are proved as three independent lemmas. The closed range theorem in the closed operator setting plays a key role in two of these lemmas, while the remaining lemma uses the Rellich-Kondrachov theorem to prove a relevant Poincaré-type inequality.

Lemma B.4. *The formulations* $\text{(B.4)}^{\mathcal{S}_{\mathbb{S}}}$ *and* $\text{(B.5)}^{\mathcal{U}_{\mathbb{S}}}$ *are mutually ill or well-posed.*

Proof. Assume $\text{(B.4)}^{\mathcal{S}_{\mathbb{S}}}$ is well-posed, so the compatibility conditions are satisfied and $\gamma^{\mathcal{S}_{\mathbb{S}}} > 0$ exists. Then by Lemma B.1, $\Gamma_u \neq \varnothing$. Using Lemma B.3, it follows $\mathcal{V}_{00}^{\mathcal{S}_{\mathbb{S}}} = \{0\}$ and $\mathcal{V}_{00}^{\mathcal{U}_{\mathbb{S}}} = \{0\}$ so the compatibility conditions are satisfied for $\text{(B.4)}^{\mathcal{S}_{\mathbb{S}}}$ and $\text{(B.5)}^{\mathcal{U}_{\mathbb{S}}}$. It remains to show the existence of $\gamma^{\mathcal{U}_{\mathbb{S}}} > 0$.

The first step is to recognize the underlying linear operators in $\text{(B.4)}^{\mathcal{S}_{\mathbb{S}}}$ and $\text{(B.5)}^{\mathcal{U}_{\mathbb{S}}}$. Indeed, for $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u})$ and $\mathfrak{v} = (\boldsymbol{\tau}, \boldsymbol{v})$,

$$\begin{aligned}
b^{\mathcal{S}_{\mathbb{S}}}(\mathfrak{u},\mathfrak{v}) &= (\mathcal{A}_{\mathcal{S}}\mathfrak{u},\mathfrak{v})_\Omega\,, \qquad \mathcal{A}_{\mathcal{S}} : \mathcal{U}^{\mathcal{S}_{\mathbb{S}}} \to \mathcal{V}^{\mathcal{S}_{\mathbb{S}}}\,, \quad \mathcal{A}_{\mathcal{S}}\mathfrak{u} = \begin{pmatrix} \mathbf{id} & -\mathsf{C} : \boldsymbol{\varepsilon} \\ -\,\mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\sigma} \\ \boldsymbol{u} \end{pmatrix}, \\
b^{\mathcal{U}_{\mathbb{S}}}(\mathfrak{u},\mathfrak{v}) &= (\mathfrak{u},\mathcal{A}_{\mathcal{U}}\mathfrak{v})_\Omega\,, \qquad \mathcal{A}_{\mathcal{U}} : \mathcal{V}^{\mathcal{U}_{\mathbb{S}}} \to \mathcal{U}^{\mathcal{U}_{\mathbb{S}}}\,, \quad \mathcal{A}_{\mathcal{U}}\mathfrak{v} = \begin{pmatrix} \mathsf{S} & \boldsymbol{\varepsilon} \\ \mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{v} \end{pmatrix}.
\end{aligned} \tag{B.12}$$

Define $\mathcal{L} = \mathcal{V}^{\mathcal{S}_{\mathbb{S}}} = \mathcal{U}^{\mathcal{U}_{\mathbb{S}}} = \mathbf{L}^2(\Omega;\mathbb{S}) \times \boldsymbol{L}^2(\Omega)$ and $\mathcal{D} = \mathcal{U}^{\mathcal{S}_{\mathbb{S}}} = \mathcal{V}^{\mathcal{U}_{\mathbb{S}}} = \mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S}) \times \boldsymbol{H}^1_{\Gamma_u}(\Omega)$. In the topology of $\mathcal{L}$, $\mathcal{D}$ is dense in $\mathcal{L}$, and $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{U}}$ are well-defined closed operators. Meanwhile, if $\mathcal{D}$

is suited with a graph norm such as $\|\mathfrak{u}\|_{\mathcal{A}_{\mathcal{S}}}^2 = \|\mathfrak{u}\|_{\mathcal{L}}^2 + \|\mathcal{A}_{\mathcal{S}}\mathfrak{u}\|_{\mathcal{L}}^2$ or $\|\mathfrak{v}\|_{\mathcal{A}_{\mathcal{U}}}^2 = \|\mathfrak{v}\|_{\mathcal{L}}^2 + \|\mathcal{A}_{\mathcal{U}}\mathfrak{v}\|_{\mathcal{L}}^2$ or with the standard norm $\|(\boldsymbol{\sigma}, \boldsymbol{u})\|_{\mathcal{D}}^2 = \|\boldsymbol{\sigma}\|_{\mathsf{H}(\mathbf{div},\Omega)}^2 + \|\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)}^2$, the operators $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{U}}$ are well-defined continuous operators. It can be shown that the norms $\|\cdot\|_{\mathcal{A}_{\mathcal{S}}}$, $\|\cdot\|_{\mathcal{A}_{\mathcal{U}}}$ and $\|\cdot\|_{\mathcal{D}}$ are equivalent.

Moreover, both operators are injective. Indeed, suppose $\mathcal{A}_{\mathcal{S}}\mathfrak{u} = 0$ and $\mathcal{A}_{\mathcal{U}}\mathfrak{v} = 0$ for $\mathfrak{u} = (\boldsymbol{\sigma}, \mathfrak{u})$ and $\mathfrak{v} = (\boldsymbol{\tau}, \boldsymbol{v})$, so that $\boldsymbol{\sigma} - \mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) = 0$, $-\mathbf{div}\,\boldsymbol{\sigma} = 0$, $\mathsf{S} : \boldsymbol{\tau} + \boldsymbol{\varepsilon}(\boldsymbol{v}) = 0$ and $\mathbf{div}\,\boldsymbol{\tau} = 0$. These can be rewritten as $-\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u})) = 0$ and $-\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{v})) = 0$ respectively, and, since $\Gamma_u \neq \varnothing$, by Lemma B.2, $\boldsymbol{u} = \boldsymbol{v} = 0$ and $\boldsymbol{\sigma} = \boldsymbol{\tau} = 0$, so that $\mathfrak{u} = \mathfrak{v} = 0$.

Next, notice the closed operator adjoints of $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{U}}$ are closely related, since $\mathcal{A}_{\mathcal{S}}^* = \mathcal{A}_{\mathcal{U}}\mathcal{M}$ and $\mathcal{A}_{\mathcal{U}}^* = \mathcal{M}^{-1}\mathcal{A}_{\mathcal{S}}$, where $\mathcal{M} = \left(\begin{smallmatrix} \mathsf{C} & 0 \\ 0 & \mathbf{id} \end{smallmatrix}\right) : \mathcal{L} \to \mathcal{L}$ is an invertible bounded linear operator with continuous inverse $\mathcal{M}^{-1} = \left(\begin{smallmatrix} \mathsf{S} & 0 \\ 0 & \mathbf{id} \end{smallmatrix}\right) : \mathcal{L} \to \mathcal{L}$. Therefore, both $\mathcal{A}_{\mathcal{S}}^*$ and $\mathcal{A}_{\mathcal{U}}^*$ are injective.

Now, using that $(\text{B.4})^{\mathcal{S}_{\mathbb{S}}}$ is well-posed, it follows there exists $\gamma^{\mathcal{S}_{\mathbb{S}}} > 0$ such that for all $\mathfrak{u} \in \mathcal{D}$

$$\|\mathcal{A}_{\mathcal{S}}\mathfrak{u}\|_{\mathcal{L}} = \sup_{\mathfrak{v}\in\mathcal{L}} \frac{|(\mathcal{A}_{\mathcal{S}}\mathfrak{u}, \mathfrak{v})_{\Omega}|}{\|\mathfrak{v}\|_{\mathcal{L}}} = \sup_{\mathfrak{v}\in\mathcal{L}} \frac{|b^{\mathcal{S}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{L}}} \geq \gamma^{\mathcal{S}_{\mathbb{S}}}\|\mathfrak{u}\|_{\mathcal{D}} \geq \gamma^{\mathcal{S}_{\mathbb{S}}}\|\mathfrak{u}\|_{\mathcal{L}}\,. \tag{B.13}$$

Using the closed range theorem for closed operators along with the injectivity of $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{S}}^*$, it follows that $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{S}}^*$ are surjective, so that $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{U}} = \mathcal{A}_{\mathcal{S}}^*\mathcal{M}^{-1}$ are bijective, and for all $\mathfrak{v} \in \mathcal{D}$

$$\|\mathcal{A}_{\mathcal{U}}\mathfrak{v}\|_{\mathcal{L}} = \|\mathcal{A}_{\mathcal{S}}^*\mathcal{M}^{-1}\mathfrak{v}\|_{\mathcal{L}} \geq \gamma^{\mathcal{S}_{\mathbb{S}}}\|\mathcal{M}^{-1}\mathfrak{v}\|_{\mathcal{L}} \geq \tfrac{\gamma^{\mathcal{S}_{\mathbb{S}}}}{\|\mathcal{M}\|}\|\mathfrak{v}\|_{\mathcal{L}}\,, \tag{B.14}$$

where it was used that $\|\mathfrak{v}\|_{\mathcal{L}} \leq \|\mathcal{M}\|\|\mathcal{M}^{-1}\mathfrak{v}\|_{\mathcal{L}}$, where $\|\mathcal{M}\|$ is the operator norm of $\mathcal{M}$. Squaring the inequality and adding $C_{\mathcal{S}}^2\|\mathcal{A}_{\mathcal{U}}\mathfrak{v}\|_{\mathcal{L}}^2$ on both sides, where $C_{\mathcal{S}} = \frac{\gamma^{\mathcal{S}_{\mathbb{S}}}}{\|\mathcal{M}\|}$, yields for all $\mathfrak{v} \in \mathcal{D}$

$$\|\mathcal{A}_{\mathcal{U}}\mathfrak{v}\|_{\mathcal{L}} \geq \sqrt{\tfrac{C_{\mathcal{S}}^2}{1+C_{\mathcal{S}}^2}}\|\mathfrak{v}\|_{A_{\mathcal{U}}} \geq C_{\mathcal{U}\mathcal{D}}\sqrt{\tfrac{C_{\mathcal{S}}^2}{1+C_{\mathcal{S}}^2}}\|\mathfrak{v}\|_{\mathcal{D}}\,, \tag{B.15}$$

where $C_{\mathcal{U}\mathcal{D}}$ is the relevant equivalence constant between the norms $\|\cdot\|_{\mathcal{A}_{\mathcal{U}}}$ and $\|\cdot\|_{\mathcal{D}}$. Let $\gamma^{\mathcal{U}_{\mathbb{S}}} > 0$ be defined by $(\gamma^{\mathcal{U}_{\mathbb{S}}})^2 = \frac{C_{\mathcal{U}\mathcal{D}}^2 C_{\mathcal{S}}^2}{1+C_{\mathcal{S}}^2}$. Since $\mathcal{A}_{\mathcal{U}} : \mathcal{D} \to \mathcal{L}$ is bijective, it follows it is invertible with inverse $\mathcal{A}_{\mathcal{U}}^{-1} : \mathcal{L} \to \mathcal{D}$, which is continuous (by the open mapping theorem) when $\mathcal{D}$ is viewed as a normed space. The continuous operator transpose $(\mathcal{A}_{\mathcal{U}}^{-1})' : \mathcal{D}' \to \mathcal{L}' = \mathcal{L}$ therefore exists, and by its properties it follows its operator norm is $\|(\mathcal{A}_{\mathcal{U}}^{-1})'\| = \|\mathcal{A}_{\mathcal{U}}^{-1}\|$. Moreover, $(\mathcal{A}_{\mathcal{U}}^{-1})' = (\mathcal{A}_{\mathcal{U}}')^{-1}$ where $\mathcal{A}_{\mathcal{U}}' : \mathcal{L}' = \mathcal{L} \to \mathcal{D}'$ is the continuous operator transpose of $\mathcal{A}_{\mathcal{U}}$ satisfying $(\mathfrak{u}, \mathcal{A}_{\mathcal{U}}\mathfrak{v})_{\Omega} = \langle \mathcal{A}_{\mathcal{U}}'\mathfrak{u}, \mathfrak{v}\rangle_{\mathcal{D}'\times\mathcal{D}}$

for all $\mathfrak{u} \in \mathcal{L}$ and $\mathfrak{v} \in \mathcal{D}$. Hence,

$$\begin{aligned} \gamma^{\mathscr{U}_{\mathbb{S}}} &\leq \inf_{\mathfrak{v}\in\mathcal{D}} \frac{\|\mathcal{A}_{\mathscr{U}}\mathfrak{v}\|_{\mathcal{L}}}{\|\mathfrak{v}\|_{\mathcal{D}}} = \Big(\sup_{\mathfrak{v}\in\mathcal{D}} \frac{\|\mathfrak{v}\|_{\mathcal{D}}}{\|\mathcal{A}_{\mathscr{U}}\mathfrak{v}\|_{\mathcal{L}}}\Big)^{-1} = \Big(\sup_{\mathfrak{u}\in\mathcal{L}} \frac{\|\mathcal{A}_{\mathscr{U}}^{-1}\mathfrak{u}\|_{\mathcal{D}}}{\|\mathfrak{u}\|_{\mathcal{L}}}\Big)^{-1} = \frac{1}{\|\mathcal{A}_{\mathscr{U}}^{-1}\|} \\ &= \frac{1}{\|(\mathcal{A}_{\mathscr{U}}^{-1})'\|} = \frac{1}{\|(\mathcal{A}'_{\mathscr{U}})^{-1}\|} = \inf_{\mathfrak{u}\in\mathcal{L}} \frac{\|\mathcal{A}'_{\mathscr{U}}\mathfrak{u}\|_{\mathcal{D}'}}{\|\mathfrak{u}\|_{\mathcal{L}}} = \inf_{\mathfrak{u}\in\mathcal{L}} \sup_{\mathfrak{v}\in\mathcal{D}} \frac{|\langle \mathcal{A}'_{\mathscr{U}}\mathfrak{u}, \mathfrak{v}\rangle_{\mathcal{D}'\times\mathcal{D}}|}{\|\mathfrak{u}\|_{\mathcal{L}}\|\mathfrak{v}\|_{\mathcal{D}}} \\ &= \inf_{\mathfrak{u}\in\mathcal{L}} \sup_{\mathfrak{v}\in\mathcal{D}} \frac{|(\mathfrak{u}, \mathcal{A}_{\mathscr{U}}\mathfrak{v})_{\Omega}|}{\|\mathfrak{u}\|_{\mathcal{L}}\|\mathfrak{v}\|_{\mathcal{D}}} = \inf_{\mathfrak{u}\in\mathcal{L}} \sup_{\mathfrak{v}\in\mathcal{D}} \frac{|b^{\mathscr{U}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{u}\|_{\mathcal{L}}\|\mathfrak{v}\|_{\mathcal{D}}} \,. \end{aligned} \tag{B.16}$$

This shows the existence of $\gamma^{\mathscr{U}_{\mathbb{S}}} > 0$ satisfying the desired property, meaning $\text{(B.5)}^{\mathscr{U}_{\mathbb{S}}}$ is well-posed.

Similar calculations show that if $\text{(B.5)}^{\mathscr{U}_{\mathbb{S}}}$ is well-posed then $\text{(B.4)}^{\mathscr{S}_{\mathbb{S}}}$ is well-posed. □

Lemma B.5. *Let $\Gamma_u \neq \varnothing$. There exists a constant $C_P > 0$ such that for all $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\boldsymbol{\omega} \in \mathsf{L}^2(\Omega; \mathbb{A})$,*

$$\sqrt{\|\boldsymbol{u}\|^2_{\boldsymbol{L}^2(\Omega)} + \|\boldsymbol{\omega}\|^2_{\mathsf{L}^2(\Omega;\mathbb{A})}} \leq C_P \| -\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\omega}\|_{\mathsf{L}^2(\Omega)} \,. \tag{B.17}$$

Proof. Let $\mathcal{L} = \boldsymbol{L}^2(\Omega) \times \mathsf{L}^2(\Omega; \mathbb{A})$ and $\|\cdot\|_{\mathcal{L}}$ be its Hilbert norm. Suppose by contradiction that such constant C_P does not exist. Then, for every $n \in \mathbb{N}$ there exists $(\widetilde{\boldsymbol{u}}_n, \widetilde{\boldsymbol{\omega}}_n) \in \boldsymbol{H}^1_{\Gamma_u}(\Omega) \times \mathsf{L}^2(\Omega; \mathbb{A})$ such that

$$\|(\widetilde{\boldsymbol{u}}_n, \widetilde{\boldsymbol{\omega}}_n)\|_{\mathcal{L}} > n\| - \nabla\widetilde{\boldsymbol{u}}_n + \widetilde{\boldsymbol{\omega}}_n\|_{\mathsf{L}^2(\Omega)} \,. \tag{B.18}$$

Let $(\boldsymbol{u}_n, \boldsymbol{\omega}_n) = \frac{1}{\|(\widetilde{\boldsymbol{u}}_n, \widetilde{\boldsymbol{\omega}}_n)\|_{\mathcal{L}}}(\widetilde{\boldsymbol{u}}_n, \widetilde{\boldsymbol{\omega}}_n)$ so that $\|(\boldsymbol{u}_n, \boldsymbol{\omega}_n)\|_{\mathcal{L}} = 1$ and $\| -\boldsymbol{\nabla}\boldsymbol{u}_n + \boldsymbol{\omega}_n\|_{\mathsf{L}^2(\Omega)} < \frac{1}{n}$ for all $n \in \mathbb{N}$. Note $(\boldsymbol{\omega}_n)_{n\in\mathbb{N}} \subseteq \mathsf{L}^2(\Omega; \mathbb{A})$ is antisymmetric so taking the symmetric part of the previous inequality yields $\|\boldsymbol{\varepsilon}(\boldsymbol{u}_n)\|_{\mathsf{L}^2(\Omega;\mathbb{S})} \leq \| -\boldsymbol{\nabla}\boldsymbol{u}_n + \boldsymbol{\omega}_n\|_{\mathsf{L}^2(\Omega)} < \frac{1}{n}$ for all $n \in \mathbb{N}$. Moreover, it is clear that $\|\boldsymbol{\omega}_n\|_{\mathsf{L}^2(\Omega;\mathbb{A})} \leq \|(\boldsymbol{u}_n, \boldsymbol{\omega}_n)\|_{\mathcal{L}} = 1$, $\|\boldsymbol{\nabla}\boldsymbol{u}_n\|_{\mathsf{L}^2(\Omega)} \leq \| -\boldsymbol{\nabla}\boldsymbol{u}_n + \boldsymbol{\omega}_n\|_{\mathsf{L}^2(\Omega)} + \|\boldsymbol{\omega}_n\|_{\mathsf{L}^2(\Omega;\mathbb{A})} \leq 2$ and $\|\boldsymbol{u}_n\|_{\boldsymbol{L}^2(\Omega)} \leq 1$, so $\|\boldsymbol{u}_n\|_{\boldsymbol{H}^1(\Omega)} \leq \sqrt{5}$ for all $n \in \mathbb{N}$ and by the Rellich-Kondrachov theorem it follows there exists a subsequence convergent to some $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$, $\lim_{k\to\infty} \|\boldsymbol{u}_{n_k} - \boldsymbol{u}\|_{\boldsymbol{L}^2(\Omega)} = 0$. Then $\boldsymbol{\varepsilon}(\boldsymbol{u}_{n_k})$ converges to $\boldsymbol{\varepsilon}(\boldsymbol{u})$ as distributions, which in turn implies $\boldsymbol{\varepsilon}(\boldsymbol{u}) = 0$. Thus, $-\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u})) = 0$ and by Lemma B.2 it follows $\boldsymbol{u} = 0$. Using Korn's inequality yields $\lim_{k\to\infty} \|\boldsymbol{u}_{n_k}\|_{\boldsymbol{H}^1(\Omega)} = 0$, so that in particular $(\boldsymbol{\nabla}\boldsymbol{u}_{n_k})_{k\in\mathbb{N}}$ converges to $\boldsymbol{\nabla}\boldsymbol{u} = 0$ in $\mathsf{L}^2(\Omega)$ and as a result $(\boldsymbol{\omega}_{n_k})_{k\in\mathbb{N}}$ converges to $\boldsymbol{\omega} = 0$ in $\mathsf{L}^2(\Omega; \mathbb{A})$ as well. Lastly, $\|(\boldsymbol{u}_{n_k}, \boldsymbol{\omega}_{n_k}) - (\boldsymbol{u}, \boldsymbol{\omega})\|_{\mathcal{L}} \geq |\|(\boldsymbol{u}_{n_k}, \boldsymbol{\omega}_{n_k})\|_{\mathcal{L}} - \|(\boldsymbol{u}, \boldsymbol{\omega})\|_{\mathcal{L}}| = 1$ for all $k \in \mathbb{N}$, because $\|(\boldsymbol{u}_{n_k}, \boldsymbol{\omega}_{n_k})\|_{\mathcal{L}} = 1$. This contradicts that $(\boldsymbol{u}_{n_k}, \boldsymbol{\omega}_{n_k})_{k\in\mathbb{N}}$ is convergent to $(\boldsymbol{u}, \boldsymbol{\omega}) = (0, 0) \in \mathcal{L}$. □

The next lemma proves an inf-sup condition which is the same as one of the Brezzi conditions for (B.9)$^{\mathcal{M}}$ [10]. It presents an alternate proof to that provided in [9, 110] and uses the closed range theorem as opposed to differential forms.

Lemma B.6. *Let $\Gamma_u \neq \varnothing$. There exists a constant $C_B > 0$ such that for all $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{\omega} \in \mathsf{L}^2(\Omega;\mathbb{A})$,*

$$C_B\sqrt{\|\boldsymbol{u}\|^2_{\boldsymbol{L}^2(\Omega)} + \|\boldsymbol{\omega}\|^2_{\mathsf{L}^2(\Omega;\mathbb{A})}} \leq \sup_{\boldsymbol{\tau}\in\mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)} \frac{|(\boldsymbol{u},\mathbf{div}\,\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega},\boldsymbol{\tau})_\Omega|}{\|\boldsymbol{\tau}\|_{\mathsf{H}(\mathbf{div},\Omega)}}\,. \tag{B.19}$$

Proof. The proof is very similar to that of Lemma B.4. First consider

$$\begin{aligned} &\mathcal{A}_{\mathcal{W}} : \mathcal{D}_{\mathcal{W}} \to \mathcal{L}_{\mathbb{M}}\,, \qquad &&\mathcal{A}_{\mathcal{W}}(\boldsymbol{u},\boldsymbol{\omega}) = -\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\omega}\,,\\ &\mathcal{A}_{\mathcal{V}} : \mathcal{D}_{\mathcal{V}} \to \mathcal{L}\,, &&\mathcal{A}_{\mathcal{V}}\boldsymbol{\tau} = \big(\mathbf{div}\,\boldsymbol{\tau}, \tfrac{1}{2}(\boldsymbol{\tau} - \boldsymbol{\tau}^{\mathsf{T}})\big)\,, \end{aligned} \tag{B.20}$$

where $\mathcal{D}_{\mathcal{W}} = \boldsymbol{H}^1_{\Gamma_u}(\Omega) \times \mathsf{L}^2(\Omega;\mathbb{A})$, $\mathcal{D}_{\mathcal{V}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)$, $\mathcal{L}_{\mathbb{M}} = \mathsf{L}^2(\Omega)$ and $\mathcal{L} = \boldsymbol{L}^2(\Omega) \times \mathsf{L}^2(\Omega;\mathbb{A})$. Clearly in the topologies of $\mathcal{L}$ and $\mathcal{L}_{\mathbb{M}}$, the domains $\mathcal{D}_{\mathcal{W}}$ and $\mathcal{D}_{\mathcal{V}}$ are dense in $\mathcal{L}$ and $\mathcal{L}_{\mathbb{M}}$ respectively. With these topologies $\mathcal{A}_{\mathcal{W}}$ and $\mathcal{A}_{\mathcal{V}}$ are well-defined closed operators. If $\mathcal{D}_{\mathcal{W}}$ is given the graph norm $\|(\boldsymbol{u},\boldsymbol{\omega})\|^2_{\mathcal{A}_{\mathcal{W}}} = \|(\boldsymbol{u},\boldsymbol{\omega})\|^2_{\mathcal{L}} + \|\mathcal{A}_{\mathcal{W}}(\boldsymbol{u},\boldsymbol{\omega})\|^2_{\mathcal{L}_{\mathbb{M}}}$ or the standard norm $\|(\boldsymbol{u},\boldsymbol{\omega})\|^2_{\mathcal{D}_{\mathcal{W}}} = \|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{\omega}\|^2_{\mathcal{L}_{\mathbb{M}}}$, then $\mathcal{A}_{\mathcal{W}}$ is a well-defined continuous operator. Note $\|\cdot\|_{\mathcal{A}_{\mathcal{W}}}$ and $\|\cdot\|_{\mathcal{D}_{\mathcal{W}}}$ are equivalent norms. Similarly, if $\mathcal{D}_{\mathcal{V}}$ is given the graph norm $\|\boldsymbol{\tau}\|^2_{\mathcal{A}_{\mathcal{V}}} = \|\boldsymbol{\tau}\|^2_{\mathcal{L}_{\mathbb{M}}} + \|\mathcal{A}_{\mathcal{V}}\boldsymbol{\tau}\|^2_{\mathcal{L}}$ or the standard norm $\|\boldsymbol{\tau}\|_{\mathcal{D}_{\mathcal{V}}} = \|\boldsymbol{\tau}\|_{\mathsf{H}(\mathbf{div},\Omega)}$ then $\mathcal{A}_{\mathcal{V}}$ is a well-defined continuous operator. Note $\|\cdot\|_{\mathcal{A}_{\mathcal{V}}}$ and $\|\cdot\|_{\mathcal{D}_{\mathcal{V}}}$ are equivalent norms.

As closed operators, $\mathcal{A}_{\mathcal{W}}$ and $\mathcal{A}_{\mathcal{V}}$ are clearly adjoint to each other, so that $\mathcal{A}^*_{\mathcal{W}} = \mathcal{A}_{\mathcal{V}}$. Moreover, if $\mathcal{A}_{\mathcal{W}}(\boldsymbol{u},\boldsymbol{\omega}) = 0$, then $\boldsymbol{\nabla}\boldsymbol{u} = \boldsymbol{\omega} \in \mathsf{L}^2(\Omega;\mathbb{A})$, so that $\boldsymbol{\varepsilon}(\boldsymbol{u}) = 0$. This implies that $-\,\mathbf{div}(\mathsf{C} : \boldsymbol{\varepsilon}(\boldsymbol{u})) = 0$ and by Lemma B.2, $\boldsymbol{u} = 0$ and $\boldsymbol{\omega} = \boldsymbol{\nabla}\boldsymbol{u} = 0$, so that $\mathcal{A}_{\mathcal{W}}$ is injective. On the other hand, if $\mathcal{A}_{\mathcal{V}}\boldsymbol{\tau} = (0,0)$, then $\boldsymbol{\tau} \in \mathsf{N}(\mathcal{A}_{\mathcal{V}}) = \{\boldsymbol{\tau}_0 \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega;\mathbb{S}) \mid \mathbf{div}\,\boldsymbol{\tau}_0 = 0\}$, so $\mathcal{A}_{\mathcal{V}}$ has a nontrivial null space.

By Lemma B.5, it follows that for all $(\boldsymbol{u},\boldsymbol{\omega}) \in \mathcal{D}_{\mathcal{W}}$, $\|\mathcal{A}_{\mathcal{W}}(\boldsymbol{u},\boldsymbol{\omega})\|_{\mathcal{L}_{\mathbb{M}}} \geq \frac{1}{C_P}\|(\boldsymbol{u},\boldsymbol{\omega})\|_{\mathcal{L}}$. By the closed range theorem, $\|\widetilde{\mathcal{A}}_{\mathcal{V}}[\boldsymbol{\tau}]\|_{\mathcal{L}} = \|\mathcal{A}_{\mathcal{V}}\boldsymbol{\tau}\|_{\mathcal{L}} \geq \frac{1}{C_P}\|[\boldsymbol{\tau}]\|_{\mathcal{L}_{\mathbb{M}}/\mathsf{N}(\mathcal{A}_{\mathcal{V}})}$ for all $\boldsymbol{\tau} \in \mathcal{D}_{\mathcal{V}}$, where $\widetilde{\mathcal{A}}_{\mathcal{V}} : \widetilde{\mathcal{D}}_{\mathcal{V}} \to \mathcal{L}$ is defined by $\widetilde{\mathcal{A}}_{\mathcal{V}}[\boldsymbol{\tau}] = \mathcal{A}_{\mathcal{V}}\boldsymbol{\tau}$, with $\widetilde{\mathcal{D}}_{\mathcal{V}} = \mathcal{D}_{\mathcal{V}}/\mathsf{N}(\mathcal{A}_{\mathcal{V}})$ and $\|[\boldsymbol{\tau}]\|_{\mathcal{L}_{\mathbb{M}}/\mathsf{N}(\mathcal{A}_{\mathcal{V}})} = \inf_{\boldsymbol{\tau}_0\in\mathsf{N}(\mathcal{A}_{\mathcal{V}})}\|\boldsymbol{\tau} + \boldsymbol{\tau}_0\|_{\mathcal{L}_{\mathbb{M}}}$. Then, $\|\widetilde{\mathcal{A}}_{\mathcal{V}}[\boldsymbol{\tau}]\|_{\mathcal{L}} \geq C_B\|[\boldsymbol{\tau}]\|_{\widetilde{\mathcal{D}}_{\mathcal{V}}}$ for all $\boldsymbol{\tau} \in \mathcal{D}_{\mathcal{V}}$, where $\|[\boldsymbol{\tau}]\|_{\widetilde{\mathcal{D}}_{\mathcal{V}}} = \inf_{\boldsymbol{\tau}_0\in\mathsf{N}(\mathcal{A}_{\mathcal{V}})}\|\boldsymbol{\tau} + \boldsymbol{\tau}_0\|_{\mathcal{D}_{\mathcal{V}}}$ and where $C_B = C_{\mathcal{V}\mathcal{D}}\frac{1}{1+C_P^2}$, with $C_{\mathcal{V}\mathcal{D}}$ being the relevant equivalence constant between $\|\cdot\|_{\mathcal{A}_{\mathcal{V}}}$ and $\|\cdot\|_{\mathcal{D}_{\mathcal{V}}}$.

The closed range theorem also implies $\mathsf{R}(\widetilde{\mathcal{A}}_{\mathcal{V}}) = \mathsf{R}(\mathcal{A}_{\mathcal{V}}) = \mathcal{L}$ is closed, so that $\widetilde{\mathcal{A}}_{\mathcal{V}}$ is bijective and by the open mapping theorem it is a homeomorphism with continuous inverse $\widetilde{\mathcal{A}}_{\mathcal{V}}^{-1} : \mathcal{L} \to \widetilde{\mathcal{D}}_{\mathcal{V}}$. Using the continuous operator transpose of $\widetilde{\mathcal{A}}_{\mathcal{V}}$ and $\widetilde{\mathcal{A}}_{\mathcal{V}}^{-1}$ as in the proof of Lemma B.4 yields for all $\mathfrak{u} = (\boldsymbol{u}, \boldsymbol{\omega}) \in \mathcal{L}$,

$$\begin{aligned} C_B \|\mathfrak{u}\|_{\mathcal{L}} &\leq \sup_{[\boldsymbol{\tau}] \in \widetilde{\mathcal{D}}_{\mathcal{V}}} \frac{|(\mathfrak{u}, \widetilde{\mathcal{A}}_{\mathcal{V}}[\boldsymbol{\tau}])_{\Omega}|}{\|[\boldsymbol{\tau}]\|_{\widetilde{\mathcal{D}}_{\mathcal{V}}}} = \sup_{\boldsymbol{\tau}^{\perp} \in \mathcal{Z}} \sup_{\boldsymbol{\tau}_0 \in \mathsf{N}(\mathcal{A}_{\mathcal{V}})} \frac{|(\mathfrak{u}, \mathcal{A}_{\mathcal{V}} \boldsymbol{\tau}^{\perp})_{\Omega}|}{\inf_{\boldsymbol{\tau}_0' \in \mathsf{N}(\mathcal{A}_{\mathcal{V}})} \|\boldsymbol{\tau}^{\perp} + \boldsymbol{\tau}_0 + \boldsymbol{\tau}_0'\|_{\mathcal{D}_{\mathcal{V}}}} \\ &= \sup_{\boldsymbol{\tau}^{\perp} \in \mathcal{Z}} \sup_{\boldsymbol{\tau}_0 \in \mathsf{N}(\mathcal{A}_{\mathcal{V}})} \sup_{\boldsymbol{\tau}_0' \in \mathsf{N}(\mathcal{A}_{\mathcal{V}})} \frac{|(\mathfrak{u}, \mathcal{A}_{\mathcal{V}} \boldsymbol{\tau}^{\perp})_{\Omega}|}{\|\boldsymbol{\tau}^{\perp} + \boldsymbol{\tau}_0 + \boldsymbol{\tau}_0'\|_{\mathcal{D}_{\mathcal{V}}}} = \sup_{\boldsymbol{\tau} \in \mathcal{D}_{\mathcal{V}}} \frac{|(\mathfrak{u}, \mathcal{A}_{\mathcal{V}} \boldsymbol{\tau})_{\Omega}|}{\|\boldsymbol{\tau}\|_{\mathcal{D}_{\mathcal{V}}}}, \end{aligned} \tag{B.21}$$

where $\mathcal{Z}$ is any algebraic complement to $\mathsf{N}(\mathcal{A}_{\mathcal{V}})$ so that $\mathcal{D}_{\mathcal{V}} = \mathsf{N}(\mathcal{A}_{\mathcal{V}}) \oplus \mathcal{Z}$. The result follows because $\big((\boldsymbol{u}, \boldsymbol{\omega}), \mathcal{A}_{\mathcal{V}} \boldsymbol{\tau}\big)_{\Omega} = (\boldsymbol{u}, \operatorname{\mathbf{div}} \boldsymbol{\tau})_{\Omega} + (\boldsymbol{\omega}, \boldsymbol{\tau})_{\Omega}$. □

Finally, we can proceed to proving the main result, Theorem 3.1.

Theorem B.2. *The variational formulations* (B.4)$^{\mathcal{S}_{\mathbb{S}}}$*,* (B.5)$^{\mathcal{U}_{\mathbb{S}}}$*,* (B.6)$^{\mathcal{M}_{\mathbb{S}}}$*,* (B.7)$^{\mathcal{S}}$*,* (B.8)$^{\mathcal{U}}$*,* (B.9)$^{\mathcal{M}}$*,* (B.10)$^{\mathcal{D}}$ *and* (B.11)$^{\mathcal{P}}$*, are mutually ill or well-posed. That is, if any single formulation is well-posed, then all others are also well-posed. In particular, if $\Gamma_u \neq \varnothing$, then all formulations are well-posed.*

Proof. From now on, the formulations will be referred to by their label, ($\mathcal{S}_{\mathbb{S}}$), ($\mathcal{U}_{\mathbb{S}}$), ($\mathcal{M}_{\mathbb{S}}$), ($\mathcal{S}$), ($\mathcal{U}$), ($\mathcal{M}$), ($\mathcal{D}$) and ($\mathcal{P}$). Assume one of the variational formulations is well-posed. Then by Lemma B.1, $\Gamma_u \neq \varnothing$. Using Lemma B.3, it follows that for all formulations the compatibility space is trivial so the compatibility conditions are satisfied immediately for any linear form.

It remains to show that the positive inf-sup constants exist for the remaining formulations. This is proved according to the following implication diagram.

$$\begin{tikzcd}
 & & (\mathcal{M}) \arrow[dll, Rightarrow] & & \\
(\mathcal{S}) \arrow[dr, Rightarrow] & & (\mathcal{M}_{\mathbb{S}}) \arrow[dl, Rightarrow] & & (\mathcal{U}) \arrow[ull, Rightarrow] \arrow[ddl, Rightarrow] \\
 & (\mathcal{S}_{\mathbb{S}}) \arrow[rr, Rightarrow] & & (\mathcal{U}_{\mathbb{S}}) \arrow[ul, Rightarrow] \arrow[ur, Rightarrow] & \\
 & (\mathcal{P}) \arrow[uul, Rightarrow] & & (\mathcal{D}) \arrow[ll, Rightarrow] &
\end{tikzcd} \tag{B.22}$$

The last statement in the theorem will hold because ($\mathcal{P}$) is well-known to be well-posed using Korn's inequality and the Lax-Milgram theorem provided $\Gamma_u \neq \varnothing$.

$(\mathcal{S}_{\mathbb{S}}) \Rightarrow (\mathscr{U}_{\mathbb{S}})$: This is the content of Lemma B.4.

$(\mathcal{S}) \Rightarrow (\mathcal{S}_{\mathbb{S}})$: The inf-sup constant $\gamma^{\mathcal{S}} > 0$ is assumed to exist. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathcal{U}^{\mathcal{S}_{\mathbb{S}}} \subseteq \mathcal{U}^{\mathcal{S}}$, $\mathfrak{v} = (\boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathcal{S}_{\mathbb{S}}}$ and $\widetilde{\mathfrak{v}} = (\mathfrak{v}, \boldsymbol{w}) \in \mathcal{V}^{\mathcal{S}}$, so that $\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{S}_{\mathbb{S}}}} \leq \|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}$. Due to the symmetry of $\boldsymbol{\sigma}$ it follows $b^{\mathcal{S}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v}) = b^{\mathcal{S}}(\mathfrak{u}, \widetilde{\mathfrak{v}})$. Hence,

$$\gamma^{\mathcal{S}}\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{S}_{\mathbb{S}}}} = \gamma^{\mathcal{S}}\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{S}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}} = \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}} \leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{S}_{\mathbb{S}}}} \frac{|b^{\mathcal{S}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{S}_{\mathbb{S}}}}}\,, \tag{B.23}$$

so that the desired inf-sup constant $\gamma^{\mathcal{S}_{\mathbb{S}}} = \gamma^{\mathcal{S}} > 0$ exists and $(\mathcal{S}_{\mathbb{S}})$ is well-posed.

$(\mathscr{U}_{\mathbb{S}}) \Rightarrow (\mathscr{U})$: Let the constant $\gamma^{\mathscr{U}_{\mathbb{S}}} > 0$ exist. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathcal{U}^{\mathscr{U}_{\mathbb{S}}}$, $\widetilde{\mathfrak{u}} = (\mathfrak{u}, \boldsymbol{\omega}) \in \mathcal{U}^{\mathscr{U}}$, $\mathfrak{v}_{\mathbb{S}} = (\boldsymbol{\tau}_{\mathbb{S}}, \boldsymbol{v}) \in \mathcal{V}^{\mathscr{U}_{\mathbb{S}}} \subseteq \mathcal{V}^{\mathscr{U}}$ and $\mathfrak{v} = (\boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathscr{U}}$. Clearly, $b^{\mathscr{U}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v}_{\mathbb{S}}) = b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_{\mathbb{S}})$, and

$$\gamma^{\mathscr{U}_{\mathbb{S}}}\|\mathfrak{u}\|_{\mathcal{U}^{\mathscr{U}_{\mathbb{S}}}} \leq \sup_{\mathfrak{v}_{\mathbb{S}} \in \mathcal{V}^{\mathscr{U}_{\mathbb{S}}}} \frac{|b^{\mathscr{U}_{\mathbb{S}}}(\mathfrak{u}, \mathfrak{v}_{\mathbb{S}})|}{\|\mathfrak{v}_{\mathbb{S}}\|_{\mathcal{V}^{\mathscr{U}_{\mathbb{S}}}}} = \sup_{\mathfrak{v}_{\mathbb{S}} \in \mathcal{V}^{\mathscr{U}_{\mathbb{S}}}} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_{\mathbb{S}})|}{\|\mathfrak{v}_{\mathbb{S}}\|_{\mathcal{V}^{\mathscr{U}_{\mathbb{S}}}}} \leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathscr{U}}} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathscr{U}}}}\,. \tag{B.24}$$

Due to $\|\widetilde{\mathfrak{u}}\|^2_{\mathcal{U}^{\mathscr{U}}} = \|\mathfrak{u}\|^2_{\mathcal{U}^{\mathscr{U}_{\mathbb{S}}}} + \|\boldsymbol{\omega}\|^2_{\mathbf{L}^2(\Omega;\mathbb{A})}$, it remains to find a bound for $\|\boldsymbol{\omega}\|_{\mathbf{L}^2(\Omega;\mathbb{A})}$. Let $\mathfrak{v}_0 = (\boldsymbol{\tau}, 0) \in \mathcal{V}^{\mathscr{U}}$, so that $\|\mathfrak{v}_0\|_{\mathcal{V}^{\mathscr{U}}} = \|\boldsymbol{\tau}\|_{\mathbf{H}(\mathbf{div},\Omega)} \geq \|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)}$ and $(\boldsymbol{u}, \mathbf{div}\,\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega}, \boldsymbol{\tau})_\Omega = b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_0) - (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega$. Then, by Lemma B.6 it follows

$$\begin{aligned} C_B\|\boldsymbol{\omega}\|_{\mathbf{L}^2(\Omega;\mathbb{A})} &\leq \sup_{\boldsymbol{\tau} \in \mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)} \frac{|(\boldsymbol{u}, \mathbf{div}\,\boldsymbol{\tau})_\Omega + (\boldsymbol{\omega}, \boldsymbol{\tau})_\Omega|}{\|\boldsymbol{\tau}\|_{\mathbf{H}(\mathbf{div},\Omega)}} = \sup_{\boldsymbol{\tau} \in \mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_0) - (\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega|}{\|\boldsymbol{\tau}\|_{\mathbf{H}(\mathbf{div},\Omega)}} \\ &\leq \sup_{\boldsymbol{\tau} \in \mathbf{H}_{\Gamma_\sigma}(\mathbf{div},\Omega)} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_0)|}{\|\mathfrak{v}_0\|_{\mathcal{V}^{\mathscr{U}}}} + \sup_{\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)} \frac{|(\mathsf{S} : \boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega|}{\|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)}} \\ &\leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathscr{U}}} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathscr{U}}}} + \|\mathsf{S}\|\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega;\mathbb{S})} \leq \Big(1 + \frac{\|\mathsf{S}\|}{\gamma^{\mathscr{U}_{\mathbb{S}}}}\Big) \sup_{\mathfrak{v} \in \mathcal{V}^{\mathscr{U}}} \frac{|b^{\mathscr{U}}(\widetilde{\mathfrak{u}}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathscr{U}}}}\,, \end{aligned} \tag{B.25}$$

since $\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega;\mathbb{S})} \leq \|\mathfrak{u}\|_{\mathcal{U}^{\mathscr{U}_{\mathbb{S}}}}$. Therefore, the existence of the desired inf-sup constant $\gamma^{\mathcal{S}} > 0$ defined by $(\gamma^{\mathscr{U}})^{-2} = \frac{1}{C_B^2}\big(1 + \frac{\|\mathsf{S}\|}{\gamma^{\mathscr{U}_{\mathbb{S}}}}\big)^2 + \big(\frac{1}{\gamma^{\mathscr{U}_{\mathbb{S}}}}\big)^2$ is ensured.

$(\mathscr{U}) \Rightarrow (\mathcal{M})$: The inf-sup constant $\gamma^{\mathscr{U}} > 0$ is assumed to exist. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\omega}) \in \mathcal{U}^{\mathcal{M}}$ and $\mathfrak{u}_{\mathbb{S}} = (\boldsymbol{\sigma}_{\mathbb{S}}, \boldsymbol{u}, \boldsymbol{\omega}) \in \mathcal{U}^{\mathscr{U}}$, where $\boldsymbol{\sigma}_{\mathbb{S}} = \frac{1}{2}(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{\mathsf{T}})$ and $\boldsymbol{\sigma}_{\mathbb{A}} = \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\mathsf{T}})$. Since $(\boldsymbol{\sigma}_{\mathbb{S}}, \boldsymbol{\sigma}_{\mathbb{A}})_\Omega = 0$, it follows $\|\mathfrak{u}\|^2_{\mathcal{U}^{\mathcal{M}}} = \|\mathfrak{u}_{\mathbb{S}}\|^2_{\mathcal{U}^{\mathscr{U}}} + \|\boldsymbol{\sigma}_{\mathbb{A}}\|^2_{\mathbf{L}^2(\Omega;\mathbb{A})} + \|\,\mathbf{div}\,\boldsymbol{\sigma}\|^2_{\boldsymbol{L}^2(\Omega)}$. Let $\widetilde{\mathfrak{v}}_{\boldsymbol{w}} = (0, 0, \boldsymbol{w}) \in \mathcal{V}^{\mathcal{M}}$ and $\widetilde{\mathfrak{v}} = (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{V}^{\mathcal{M}}$ so that $\|\widetilde{\mathfrak{v}}_{\boldsymbol{w}}\|_{\mathcal{V}^{\mathcal{M}}} = \|\boldsymbol{w}\|_{\mathbf{L}^2(\Omega;\mathbb{A})}$. Then, it is clear $b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_{\boldsymbol{w}}) = (\boldsymbol{\sigma}_{\mathbb{A}}, \boldsymbol{w})_\Omega$, and

$$\|\boldsymbol{\sigma}_{\mathbb{A}}\|_{\mathbf{L}^2(\Omega;\mathbb{A})} = \sup_{\boldsymbol{w} \in \mathbf{L}^2(\Omega;\mathbb{A})} \frac{|(\boldsymbol{\sigma}_{\mathbb{A}}, \boldsymbol{w})_\Omega|}{\|\boldsymbol{w}\|_{\mathbf{L}^2(\Omega;\mathbb{A})}} = \sup_{\boldsymbol{w} \in \mathbf{L}^2(\Omega;\mathbb{A})} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_{\boldsymbol{w}})|}{\|\widetilde{\mathfrak{v}}_{\boldsymbol{w}}\|_{\mathcal{V}^{\mathcal{M}}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{M}}} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{M}}}}\,. \tag{B.26}$$

Next, let $\mathfrak{v} = (\boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathcal{U}}$ and $\widetilde{\mathfrak{v}}_0 = (\mathfrak{v}, 0) \in \mathcal{V}^{\mathcal{M}}$, so that $\|\widetilde{\mathfrak{v}}_0\|_{\mathcal{V}^{\mathcal{M}}} \leq \|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{U}}}$ and $\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)} \leq \|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{U}}}$. The distributional identity $-(\mathbf{div}\,\boldsymbol{\sigma}, \boldsymbol{v})_\Omega = (\boldsymbol{\sigma}, \nabla \boldsymbol{v})_\Omega$ holds for $\boldsymbol{\sigma} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$ and $\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$. A careful calculation shows that $b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_0) = b^{\mathcal{U}}(\mathfrak{u}_{\mathbb{S}}, \mathfrak{v}) + (\boldsymbol{\sigma}_{\mathbb{A}}, \nabla \boldsymbol{v})_\Omega$. Therefore,

$$\begin{aligned}\gamma^{\mathcal{U}} \|\mathfrak{u}_{\mathbb{S}}\|_{\mathcal{U}^{\mathcal{U}}} &\leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{U}}} \frac{|b^{\mathcal{U}}(\mathfrak{u}_{\mathbb{S}}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{U}}}} \leq \sup_{\widetilde{\mathfrak{v}}_0 \in (\mathcal{V}^{\mathcal{U}}) \times \{0\}} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_0)|}{\|\widetilde{\mathfrak{v}}_0\|_{\mathcal{V}^{\mathcal{M}}}} + \sup_{\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)} \frac{|(\boldsymbol{\sigma}_{\mathbb{A}}, \nabla \boldsymbol{v})_\Omega|}{\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)}} \\ &\leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{M}}} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{M}}}} + \|\boldsymbol{\sigma}_{\mathbb{A}}\|_{\mathsf{L}^2(\Omega;\mathbb{A})} \sup_{\boldsymbol{v} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)} \frac{\|\nabla \boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)}}{\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)}} \leq 2 \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{M}}} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{M}}}}\,.\end{aligned} \tag{B.27}$$

Finally, let $\widetilde{\mathfrak{v}}_{\boldsymbol{v}} = (0, \boldsymbol{v}, 0) \in \mathcal{V}^{\mathcal{M}}$ so that $\|\widetilde{\mathfrak{v}}_{\boldsymbol{v}}\|_{\mathcal{V}}^{\mathcal{M}} = \|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)}$ and $-(\mathbf{div}\,\boldsymbol{\sigma}, \boldsymbol{v})_\Omega = b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_{\boldsymbol{v}})$. Then,

$$\|\,\mathbf{div}\,\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(\Omega)} = \sup_{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)} \frac{|(\mathbf{div}\,\boldsymbol{\sigma}, \boldsymbol{v})_\Omega|}{\|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)}} = \sup_{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}}_{\boldsymbol{v}})|}{\|\widetilde{\mathfrak{v}}_{\boldsymbol{v}}\|_{\mathcal{V}^{\mathcal{M}}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{M}}} \frac{|b^{\mathcal{M}}(\mathfrak{u}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{M}}}}\,, \tag{B.28}$$

which implies that $\gamma^{\mathcal{M}} > 0$ defined by $(\gamma^{\mathcal{M}})^2 = \frac{(\gamma^{\mathcal{U}})^2}{4 + 2(\gamma^{\mathcal{U}})^2}$ is the desired inf-sup constant.

$(\mathcal{U}_{\mathbb{S}}) \Rightarrow (\mathcal{M}_{\mathbb{S}})$: This is proved analogously to $(\mathcal{U}) \Rightarrow (\mathcal{M})$, but ignoring the calculations associated to the term $\boldsymbol{\sigma}_{\mathbb{A}}$, which vanishes in this symmetric setting.

$(\mathcal{M}) \Rightarrow (\mathcal{S})$: The inf-sup constant of $(\mathcal{M})$, $\gamma^{\mathcal{M}} > 0$, is assumed to exist. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathcal{U}^{\mathcal{S}}$, $\widetilde{\mathfrak{u}} = \big(\mathfrak{u}, \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} - \boldsymbol{\nabla}\boldsymbol{u}^{\mathsf{T}})\big) \in \mathcal{U}^{\mathcal{M}}$, $\mathfrak{v}_{\mathcal{M}} = (\boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{V}^{\mathcal{M}}$ and $\mathfrak{v}_{\mathcal{S}} = (\mathsf{S} : \boldsymbol{\tau}, \boldsymbol{v}, \boldsymbol{w}) \in \mathcal{V}^{\mathcal{S}}$. Then, notice that $\|\mathfrak{u}\|^2_{\mathcal{U}^{\mathcal{S}}} = \|\widetilde{\mathfrak{u}}\|^2_{\mathcal{U}^{\mathcal{M}}} + \|\boldsymbol{\varepsilon}(\boldsymbol{u})\|^2_{\mathsf{L}^2(\Omega;\mathbb{S})}$ and $\|\mathsf{S} : \boldsymbol{\tau}\|_{\mathsf{L}^2(\Omega;\mathbb{S})} \leq \|\mathsf{S}\|\|\boldsymbol{\tau}\|_{\mathsf{L}^2(\Omega)}$, so $\|\mathfrak{v}_{\mathcal{S}}\|_{\mathcal{V}^{\mathcal{S}}} \leq M_{\mathsf{S}} \|\mathfrak{v}_{\mathcal{M}}\|_{\mathcal{V}^{\mathcal{M}}}$ where $M_{\mathsf{S}} = \max\{\|\mathsf{S}\|, 1\}$. The distributional identity $(\boldsymbol{u}, \mathbf{div}\,\boldsymbol{\tau})_\Omega = -(\boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\tau})_\Omega$ holds because $\boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_u}(\Omega)$ and $\boldsymbol{\tau} \in \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega)$, and implies that $b^{\mathcal{M}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_{\mathcal{M}}) = b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}_{\mathcal{S}})$. Hence,

$$\gamma^{\mathcal{M}} \|\widetilde{\mathfrak{u}}\|_{\mathcal{U}^{\mathcal{M}}} \leq \sup_{\mathfrak{v}_{\mathcal{M}} \in \mathcal{V}^{\mathcal{M}}} \frac{|b^{\mathcal{M}}(\widetilde{\mathfrak{u}}, \mathfrak{v}_{\mathcal{M}})|}{\|\mathfrak{v}_{\mathcal{M}}\|_{\mathcal{V}^{\mathcal{M}}}} \leq M_{\mathsf{S}} \sup_{\mathfrak{v}_{\mathcal{M}} \in \mathcal{V}^{\mathcal{M}}, \mathfrak{v}_{\mathcal{S}} \neq 0} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}_{\mathcal{S}})|}{\|\mathfrak{v}_{\mathcal{S}}\|_{\mathcal{V}^{\mathcal{S}}}} \leq M_{\mathsf{S}} \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{S}}}}\,. \tag{B.29}$$

It remains to find a bound for $\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{\mathsf{L}^2(\Omega;\mathbb{S})}$. Let $\mathfrak{v}_0 = (\mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}}, 0, 0) \in \mathcal{V}^{\mathcal{S}}$ for $\boldsymbol{\tau}_{\mathbb{S}} \in \mathsf{L}^2(\Omega; \mathbb{S})$, so that $\|\mathfrak{v}_0\|_{\mathcal{V}^{\mathcal{S}}} \leq \|\mathsf{S}\|\|\boldsymbol{\tau}_{\mathbb{S}}\|_{\mathsf{L}^2(\Omega;\mathbb{S})}$. Notice $(\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\tau}_{\mathbb{S}})_\Omega = (\boldsymbol{\sigma}, \mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}})_\Omega - b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}_0)$. Therefore,

$$\begin{aligned}\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{\mathsf{L}^2(\Omega;\mathbb{S})} &= \sup_{\boldsymbol{\tau}_{\mathbb{S}} \in \mathsf{L}^2(\Omega;\mathbb{S})} \frac{|(\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\tau}_{\mathbb{S}})_\Omega|}{\|\boldsymbol{\tau}_{\mathbb{S}}\|_{\mathsf{L}^2(\Omega;\mathbb{S})}} = \sup_{\boldsymbol{\tau}_{\mathbb{S}} \in \mathsf{L}^2(\Omega;\mathbb{S})} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}_0) - (\boldsymbol{\sigma}, \mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}})_\Omega|}{\|\boldsymbol{\tau}_{\mathbb{S}}\|_{\mathsf{L}^2(\Omega;\mathbb{S})}} \\ &\leq \|\mathsf{S}\| \sup_{\boldsymbol{\tau}_{\mathbb{S}} \in \mathsf{L}^2(\Omega;\mathbb{S})} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v}_0)|}{\|\mathfrak{v}_0\|_{\mathcal{V}^{\mathcal{S}}}} + \|\mathsf{S}\| \sup_{\boldsymbol{\tau}_{\mathbb{S}} \in \mathsf{L}^2(\Omega;\mathbb{S})} \frac{|(\boldsymbol{\sigma}, \mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}})_\Omega|}{\|\mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}}\|_{\mathsf{L}^2(\Omega;\mathbb{S})}} \\ &\leq \|\mathsf{S}\| \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{S}}}} + \|\mathsf{S}\|\|\boldsymbol{\sigma}\|_{\mathsf{L}^2(\Omega;\mathbb{S})}\,,\end{aligned} \tag{B.30}$$

where it is used that S is bijective on $\mathbf{L}^2(\Omega;\mathbb{S})$. Using that $\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega;\mathbb{S})} \leq \|\widetilde{\mathfrak{u}}\|_{\mathcal{U}^{\mathcal{M}}}$, the existence of the inf-sup constant $\gamma^{\mathcal{S}} > 0$ defined by $(\gamma^{\mathcal{S}})^{-2} = \|\mathsf{S}\|^2\big(1+\frac{M_{\mathsf{S}}}{\gamma^{\mathcal{M}}}\big)^2 + \big(\frac{M_{\mathsf{S}}}{\gamma^{\mathcal{M}}}\big)^2$ is ensured.

$(\mathcal{M}_{\mathbb{S}}) \Rightarrow (\mathcal{S}_{\mathbb{S}})$: This is proved analogously to $(\mathcal{M}) \Rightarrow (\mathcal{S})$.

$(\mathscr{U}) \Rightarrow (\mathcal{D})$: The inf-sup constant of $(\mathscr{U})$, $\gamma^{\mathscr{U}} > 0$, is assumed to exist. Let $\mathfrak{u} = (\boldsymbol{\sigma}, \boldsymbol{u}) \in \mathcal{U}^{\mathcal{D}}$, $\widetilde{\mathfrak{u}} = \big(\mathfrak{u}, \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} - \boldsymbol{\nabla}\boldsymbol{u}^{\mathsf{T}})\big) \in \mathcal{U}^{\mathscr{U}}$, $\mathfrak{v}_{\mathcal{U}} = (\boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathscr{U}}$, $\mathfrak{v}_{\mathcal{D}} = (\mathsf{S} : \boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathcal{D}}$, and $\mathfrak{v}_0 = (\mathsf{S} : \boldsymbol{\tau}_{\mathbb{S}}, 0) \in \mathcal{V}^{\mathcal{D}}$ for $\boldsymbol{\tau}_{\mathbb{S}} \in \mathbf{L}^2(\Omega;\mathbb{S})$. The proof is then the same as for $(\mathcal{M}) \Rightarrow (\mathcal{S})$, but replacing $\mathcal{M}$ by $\mathscr{U}$ and $\mathcal{S}$ by $\mathcal{D}$.

$(\mathcal{D}) \Rightarrow (\mathcal{P})$: Assume the constant $\gamma^{\mathcal{D}} > 0$ exists. Let $\mathfrak{u} = \boldsymbol{u} \in \mathcal{U}^{\mathcal{P}}$, $\widetilde{\mathfrak{u}} = (\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \mathfrak{u}) \in \mathcal{U}^{\mathcal{D}}$, $\mathfrak{v} = \boldsymbol{v} \in \mathcal{V}^{\mathcal{P}}$ and $\widetilde{\mathfrak{v}} = (\boldsymbol{\tau}, \mathfrak{v}) \in \mathcal{V}^{\mathcal{D}}$, so that $\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{P}}} \leq \|\widetilde{\mathfrak{u}}\|_{\mathcal{U}^{\mathcal{D}}}$ and $\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{P}}} \leq \|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{D}}}$. Clearly it holds that $b^{\mathcal{D}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}) = b^{\mathcal{P}}(\mathfrak{u}, \mathfrak{v})$. Then,

$$\gamma^{\mathcal{D}}\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{P}}} \leq \gamma^{\mathcal{D}}\|\widetilde{\mathfrak{u}}\|_{\mathcal{U}^{\mathcal{D}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{D}}} \frac{|b^{\mathcal{D}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{D}}}} = \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{D}}} \frac{|b^{\mathcal{P}}(\mathfrak{u}, \mathfrak{v})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{D}}}} \leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{P}}} \frac{|b^{\mathcal{P}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{P}}}}\,, \tag{B.31}$$

so that the desired inf-sup constant $\gamma^{\mathcal{P}} = \gamma^{\mathcal{D}} > 0$ exists and $(\mathcal{P})$ is well-posed.

$(\mathcal{P}) \Rightarrow (\mathcal{S})$: The constant $\gamma^{\mathcal{P}} > 0$ is assumed to exist. Let $\mathfrak{u} = \boldsymbol{u} \in \mathcal{U}^{\mathcal{P}}$, $\widetilde{\mathfrak{u}} = (\boldsymbol{\sigma}, \mathfrak{u}) \in \mathcal{U}^{\mathcal{S}}$, $\mathfrak{v} = \boldsymbol{v} \in \mathcal{V}^{\mathcal{P}}$ and $\widetilde{\mathfrak{v}}_{\boldsymbol{v}} = \big(-\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{v}, \frac{1}{2}(\nabla\boldsymbol{v} - \nabla\boldsymbol{v}^{\mathsf{T}})\big) \in \mathcal{V}^{\mathcal{S}}$. Then, notice that $\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{P}}} = \|\widetilde{\mathfrak{v}}_{\boldsymbol{v}}\|_{\mathcal{V}^{\mathcal{S}}}$ and that $\|\widetilde{\mathfrak{u}}\|_{\mathcal{U}^{\mathcal{S}}}^2 = \|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{P}}}^2 + \|\boldsymbol{\sigma}\|_{\mathbf{H}(\mathbf{div},\Omega)}^2$. A careful calculation yields $b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_{\boldsymbol{v}}) = b^{\mathcal{P}}(\mathfrak{u}, \mathfrak{v})$. Therefore,

$$\gamma^{\mathcal{P}}\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{P}}} \leq \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{P}}} \frac{|b^{\mathcal{P}}(\mathfrak{u}, \mathfrak{v})|}{\|\mathfrak{v}\|_{\mathcal{V}^{\mathcal{P}}}} = \sup_{\mathfrak{v} \in \mathcal{V}^{\mathcal{P}}} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_v)|}{\|\widetilde{\mathfrak{v}}_v\|_{\mathcal{V}^{\mathcal{S}}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}}\,. \tag{B.32}$$

Next, consider $\boldsymbol{\xi} \in \mathbf{L}^2(\Omega)$ which is decomposed into $\boldsymbol{\xi}_{\mathbb{S}} = \frac{1}{2}(\boldsymbol{\xi} + \boldsymbol{\xi}^{\mathsf{T}})$ and $\boldsymbol{\xi}_{\mathbb{A}} = \frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^{\mathsf{T}})$, and let $\widetilde{\mathfrak{v}}_{\boldsymbol{\xi}} = (\boldsymbol{\xi}_{\mathbb{S}}, 0, \boldsymbol{\xi}_{\mathbb{A}})$, so that $\|\boldsymbol{\xi}\|_{\mathbf{L}^2(\Omega)} = \|\widetilde{\mathfrak{v}}_{\boldsymbol{\xi}}\|_{\mathcal{V}^{\mathcal{S}}}$. Notice that $b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_{\boldsymbol{\xi}}) = (\boldsymbol{\sigma}, \boldsymbol{\xi})_\Omega - (\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\xi}_{\mathbb{S}})_\Omega$. Hence,

$$\begin{aligned}
\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} &= \sup_{\boldsymbol{\xi} \in \mathbf{L}^2(\Omega)} \frac{|(\boldsymbol{\sigma}, \boldsymbol{\xi})_\Omega|}{\|\boldsymbol{\xi}\|_{\mathbf{L}^2(\Omega)}} = \sup_{\boldsymbol{\xi} \in \mathbf{L}^2(\Omega)} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_{\boldsymbol{\xi}}) + (\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\xi}_{\mathbb{S}})_\Omega|}{\|\widetilde{\mathfrak{v}}_{\boldsymbol{\xi}}\|_{\mathcal{V}^{\mathcal{S}}}} \\
&\leq \sup_{\boldsymbol{\xi} \in \mathbf{L}^2(\Omega)} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_{\boldsymbol{\xi}})|}{\|\widetilde{\mathfrak{v}}_{\boldsymbol{\xi}}\|_{\mathcal{V}^{\mathcal{S}}}} + \sup_{\boldsymbol{\xi}_{\mathbb{S}} \in \mathbf{L}^2(\Omega;\mathbb{S})} \frac{|(\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\xi}_{\mathbb{S}})_\Omega|}{\|\boldsymbol{\xi}_{\mathbb{S}}\|_{\mathbf{L}^2(\Omega;\mathbb{S})}} \\
&\leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}} + \|\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}\|_{\mathbf{L}^2(\Omega;\mathbb{S})} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}} + \|\mathsf{C}\|\|\mathfrak{u}\|_{\mathcal{U}^{\mathcal{P}}}\,.
\end{aligned} \tag{B.33}$$

Finally, let $\widetilde{\mathfrak{v}}_0 = (0, \boldsymbol{v}, 0) \in \mathcal{V}^{\mathcal{S}}$ so that $\|\widetilde{\mathfrak{v}}_0\|_{\mathcal{V}}^{\mathcal{S}} = \|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)}$ and $-(\mathbf{div}\,\boldsymbol{\sigma}, \boldsymbol{v})_\Omega = b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_0)$. Then,

$$\|\,\mathbf{div}\,\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(\Omega)} = \sup_{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)} \frac{|(\mathbf{div}\,\boldsymbol{\sigma}, \boldsymbol{v})_\Omega|}{\|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Omega)}} = \sup_{\boldsymbol{v} \in \boldsymbol{L}^2(\Omega)} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}}_0)|}{\|\widetilde{\mathfrak{v}}_0\|_{\mathcal{V}^{\mathcal{S}}}} \leq \sup_{\widetilde{\mathfrak{v}} \in \mathcal{V}^{\mathcal{S}}} \frac{|b^{\mathcal{S}}(\widetilde{\mathfrak{u}}, \widetilde{\mathfrak{v}})|}{\|\widetilde{\mathfrak{v}}\|_{\mathcal{V}^{\mathcal{S}}}}\,. \tag{B.34}$$

Therefore, the inf-sup constant $\gamma^{\mathcal{S}} > 0$ exists and is defined by $(\gamma^{\mathcal{S}})^{-2} = \big(1 + \frac{\|\mathsf{C}\|}{\gamma^{\mathcal{P}}}\big)^2 + \big(\frac{1}{\gamma^{\mathcal{P}}}\big)^2 + 1$. $\square$

B.2 Well-posedness of coupled formulations

The aim of this section is to prove Theorem 3.3, which is about the well-posedness of the coupled variational formulations of the equations of linear elasticity described in Section 3.2.6.

Theorem B.3. *Let Ω be a domain partitioned into a finite number of subdomains, wherein each subdomain is endowed with a broken variational formulation of linear elasticity among those found in* (3.24)–(3.27)*. Provided $\Gamma_u \neq \varnothing$, the resulting coupled variational formulation is well-posed.*

Proof. For the sake of consistency and simplicity the main body of the proof applies to the two-subdomain example described in Section 3.2.6, which involves the ultraweak and primal formulations. Then, a few observations will clarify the more general case.

The goal is to prove that there exists a $\gamma^{\mathcal{C}} > 0$ such that for every $\mathfrak{u}^{\mathcal{C}} = (\mathfrak{u}^{\mathcal{U}}, \mathfrak{u}^{\mathcal{P}}) \in \mathcal{U}^{\mathcal{C}}$,

$$\gamma^{\mathcal{C}} \|\mathfrak{u}^{\mathcal{C}}\|_{\mathcal{U}^{\mathcal{C}}} \leq \sup_{\mathfrak{v}^{\mathcal{C}} \in \mathcal{V}^{\mathcal{C}}} \frac{|b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}}, \mathfrak{v}^{\mathcal{C}})|}{\|\mathfrak{v}^{\mathcal{C}}\|_{\mathcal{V}^{\mathcal{C}}}} = \|\mathfrak{u}^{\mathcal{C}}\|_E \,, \tag{B.35}$$

where $\|\mathfrak{u}^{\mathcal{C}}\|_E = \|\mathcal{B}^{\mathcal{C}}\mathfrak{u}^{\mathcal{C}}\|_{(\mathcal{V}^{\mathcal{C}})'}$ with $\mathcal{B}^{\mathcal{C}} : \mathcal{U}^{\mathcal{C}} \to (\mathcal{V}^{\mathcal{C}})'$ defined by $\langle \mathcal{B}^{\mathcal{C}}\mathfrak{u}^{\mathcal{C}}, \mathfrak{v}^{\mathcal{C}} \rangle_{(\mathcal{V}^{\mathcal{C}})' \times \mathcal{V}^{\mathcal{C}}} = b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}}, \mathfrak{v}^{\mathcal{C}})$.

As usual, the approach is to prove this bound for each of the components in $\mathfrak{u}^{\mathcal{C}} = (\mathfrak{u}^{\mathcal{U}}, \mathfrak{u}^{\mathcal{P}})$, where $\mathfrak{u}^{\mathcal{U}} = (\mathfrak{u}_0^{\mathcal{U}}, \hat{\mathfrak{u}}^{\mathcal{U}})$, $\mathfrak{u}_0^{\mathcal{U}} = (\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}}) \in \mathcal{U}^{\mathcal{U}}|_{\Omega^{\mathcal{U}}}$, $\hat{\mathfrak{u}}^{\mathcal{U}} = (\hat{\boldsymbol{u}}^{\mathcal{U}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}) \in \hat{\mathcal{U}}^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}$, $\mathfrak{u}^{\mathcal{P}} = (\mathfrak{u}_0^{\mathcal{P}}, \hat{\mathfrak{u}}^{\mathcal{P}})$, $\mathfrak{u}_0^{\mathcal{P}} = \boldsymbol{u}^{\mathcal{P}} \in \mathcal{U}^{\mathcal{P}}|_{\Omega^{\mathcal{P}}}$ and $\hat{\mathfrak{u}}^{\mathcal{P}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}} \in \hat{\mathcal{U}}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$. The first step is to find the bounds for the field variables $\mathfrak{u}_0^{\mathcal{U}}$ and $\mathfrak{u}_0^{\mathcal{P}}$ by somehow avoiding the terms involving the interface variables. The main idea to achieve this is to collapse *all* formulations to the ultraweak formulation via careful testing and integration by parts, yielding a *global* ultraweak formulation. This formulation has all the weight of the derivatives on the test function, so it makes sense to consider the *global* ultraweak test functions $\mathfrak{v}^{\Omega} = (\boldsymbol{\tau}, \boldsymbol{v}) \in \mathcal{V}^{\mathcal{U}} = \mathsf{H}_{\Gamma_\sigma}(\mathbf{div}, \Omega) \times \boldsymbol{H}_{\Gamma_u}^1(\Omega)$.

From now on, given any tensor, let the subscripts $\mathbb{S}$ and $\mathbb{A}$ denote its symmetric and antisymmetric parts. Let $\boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}}) = (\boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}})_{\mathbb{A}}$, $\boldsymbol{\varepsilon}(\boldsymbol{u}^{\mathcal{P}}) = (\boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}})_{\mathbb{S}}$, and $\boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}) = \mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}}$, so that $\boldsymbol{\varepsilon}(\boldsymbol{u}^{\mathcal{P}}) = \mathsf{S} : \boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}})$. Thus, $(\boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} = (\mathsf{S} : \boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} + (\boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}}$, and it follows

$$\begin{aligned}
(\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}^{\mathcal{P}}} &= (\mathsf{S} : \boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} + (\boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} - (\boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} + (\mathsf{C} : \boldsymbol{\nabla}\boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}^{\mathcal{P}}} \\
&= (\mathsf{S} : \boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} + (\boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} + (\boldsymbol{u}^{\mathcal{P}}, \mathbf{div}\, \boldsymbol{\tau})_{\mathcal{T}^{\mathcal{P}}} \\
&\qquad + (\boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{\nabla}\boldsymbol{v})_{\mathcal{T}^{\mathcal{P}}} - \langle \mathbf{tr}_{\mathrm{grad}}^{\mathcal{T}} \boldsymbol{u}^{\mathcal{P}}, \mathbf{tr}_{\mathrm{div}}^{\mathcal{T}} \boldsymbol{\tau} \rangle_{\partial\mathcal{T}^{\mathcal{P}}} \,,
\end{aligned} \tag{B.36}$$

where integration by parts was valid due to the high regularity of both $\boldsymbol{u}^{\mathcal{P}}$ and $\boldsymbol{\tau}$. Therefore

$$b^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}^{\mathcal{P}},\mathfrak{v}^{\mathcal{P}}) = b^{\mathcal{U}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}^{\mathcal{U}_{\mathcal{P}}},\mathfrak{v}^{\Omega}) - \langle \mathbf{tr}_{\mathrm{grad}}^{\mathcal{T}}\boldsymbol{u}^{\mathcal{P}}, \mathbf{tr}_{\mathrm{div}}^{\mathcal{T}}\boldsymbol{\tau}\rangle_{\partial\mathcal{T}^{\mathcal{P}}} - \langle \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}, \mathbf{tr}_{\mathrm{grad}}^{\mathcal{T}}\boldsymbol{v}\rangle_{\partial\mathcal{T}^{\mathcal{P}}}\,, \tag{B.37}$$

where $\mathfrak{u}^{\mathcal{P}} = (\boldsymbol{u}^{\mathcal{P}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}})$, $\mathfrak{u}^{\mathcal{U}_{\mathcal{P}}} = (\boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}}))$, $\mathfrak{v}^{\Omega} = (\boldsymbol{\tau}, \boldsymbol{v})$ and $\mathfrak{v}^{\mathcal{P}} = \boldsymbol{v}$. Trivially it holds that

$$b^{\mathcal{U}\mathcal{T}}|_{\Omega^{\mathcal{U}}}(\mathfrak{u}^{\mathcal{U}},\mathfrak{v}^{\mathcal{U}}) = b^{\mathcal{U}}|_{\Omega^{\mathcal{U}}}(\mathfrak{u}^{\mathcal{U}_{\mathcal{U}}},\mathfrak{v}^{\Omega}) - \langle \hat{\boldsymbol{u}}^{\mathcal{U}}, \mathbf{tr}_{\mathrm{div}}^{\mathcal{T}}\boldsymbol{\tau}\rangle_{\partial\mathcal{T}^{\mathcal{U}}} - \langle \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}, \mathbf{tr}_{\mathrm{grad}}^{\mathcal{T}}\boldsymbol{v}\rangle_{\partial\mathcal{T}^{\mathcal{U}}}\,, \tag{B.38}$$

where $\mathfrak{u}^{\mathcal{U}} = \big(\mathfrak{u}_0^{\mathcal{U}}, (\hat{\boldsymbol{u}}^{\mathcal{U}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}})\big)$, $\mathfrak{u}^{\mathcal{U}_{\mathcal{U}}} = \mathfrak{u}_0^{\mathcal{U}} = (\boldsymbol{\sigma}^{\mathcal{U}}, u^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}})$, $\mathfrak{v}^{\Omega} = \mathfrak{v}^{\mathcal{U}} = (\boldsymbol{\tau}, \boldsymbol{v})$. Adding the previous two expressions and using the transmission conditions for the displacement and stress (see Remark 3.1) along with Theorem A.1, it follows that the interface terms vanish, resulting in

$$b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}},\mathfrak{v}^{\mathcal{C}_{\Omega}}) = b^{\mathcal{U}}(\mathfrak{u}^{\Omega},\mathfrak{v}^{\Omega})\,, \tag{B.39}$$

where $\mathfrak{u}^{\mathcal{C}} = \big((\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}}, \hat{\boldsymbol{u}}^{\mathcal{U}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}), (\boldsymbol{u}^{\mathcal{P}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}})\big) \in \mathcal{U}^{\mathcal{C}}$, $\mathfrak{u}^{\Omega} = (\boldsymbol{\sigma}^{\Omega}, \boldsymbol{u}^{\Omega}, \boldsymbol{\omega}^{\Omega})$ is defined by restriction as $\mathfrak{u}^{\Omega}|_{\Omega^{\mathcal{U}}} = (\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}})$ and $\mathfrak{u}^{\Omega}|_{\Omega^{\mathcal{P}}} = \big(\boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}})\big)$, $\mathfrak{v}^{\Omega} = (\boldsymbol{\tau}, \boldsymbol{v})$ and $\mathfrak{v}^{\mathcal{C}_{\Omega}} = \big((\boldsymbol{\tau}, \boldsymbol{v})|_{\Omega^{\mathcal{U}}}, \boldsymbol{v}|_{\Omega^{\mathcal{P}}}\big)$. Thus, when testing appropriately, the coupled formulation is essentially a global ultraweak formulation.

The global ultraweak variational formulation is known to be well-posed when $\Gamma_u \neq \varnothing$ by Theorem 3.1, so that there exists $\gamma^{\mathcal{U}}$ such that

$$\gamma^{\mathcal{U}}\|\mathfrak{u}^{\Omega}\|_{\mathcal{U}^{\mathcal{U}}} \leq \sup_{\mathfrak{v}^{\Omega}\in\mathcal{V}^{\mathcal{U}}} \frac{|b^{\mathcal{U}}(\mathfrak{u}^{\Omega},\mathfrak{v}^{\Omega})|}{\|\mathfrak{v}^{\Omega}\|_{\mathcal{V}^{\mathcal{U}}}} = \sup_{\mathfrak{v}^{\Omega}\in\mathcal{V}^{\mathcal{U}}} \frac{|b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}},\mathfrak{v}^{\mathcal{C}_{\Omega}})|}{\|\mathfrak{v}^{\Omega}\|_{\mathcal{V}^{\mathcal{U}}}} \leq \sup_{\mathfrak{v}^{\Omega}\in\mathcal{V}^{\mathcal{U}}} \frac{|b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}},\mathfrak{v}^{\mathcal{C}_{\Omega}})|}{\|\mathfrak{v}^{\mathcal{C}_{\Omega}}\|_{\mathcal{V}^{\mathcal{C}}}} \leq \|\mathfrak{u}^{\mathcal{C}}\|_{E}\,, \tag{B.40}$$

where it is used that $\|\mathfrak{v}^{\mathcal{C}_{\Omega}}\|_{\mathcal{V}^{\mathcal{C}}} \leq \|\mathfrak{v}^{\Omega}\|_{\mathcal{V}^{\mathcal{U}}}$. Naturally, $\|(\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}})\|_{\mathcal{U}^{\mathcal{U}}|_{\Omega^{\mathcal{U}}}} \leq \|\mathfrak{u}^{\Omega}\|_{\mathcal{U}^{\mathcal{U}}}$. Meanwhile, $\|\boldsymbol{u}^{\mathcal{P}}\|_{\boldsymbol{H}^1(\Omega^{\mathcal{P}})} = \|\big(\boldsymbol{\varepsilon}(\boldsymbol{u}^{\mathcal{P}}), \boldsymbol{u}^{\mathcal{P}}, \boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{P}})\big)\|_{\mathcal{U}^{\mathcal{U}}|_{\Omega^{\mathcal{P}}}}$, so in view of $\|\boldsymbol{\varepsilon}(\boldsymbol{u}^{\mathcal{P}})\|_{\mathsf{L}^2(\Omega^{\mathcal{P}};\mathbb{S})} \leq \|\mathsf{S}\|\|\boldsymbol{\sigma}(\boldsymbol{u}^{\mathcal{P}})\|_{\mathsf{L}^2(\Omega^{\mathcal{P}};\mathbb{S})}$, it follows $\|\boldsymbol{u}^{\mathcal{P}}\|_{\boldsymbol{H}^1(\Omega^{\mathcal{P}})} \leq C_{\mathsf{S}}\|\mathfrak{u}^{\Omega}\|_{\mathcal{U}^{\mathcal{U}}}$, where $C_{\mathsf{S}} = \max\{1, \|\mathsf{S}\|\}$. Thus, there exist constants $C^{\mathcal{U}} > 0$ and $C^{\mathcal{P}} > 0$, such that

$$\|\mathfrak{u}_0^{\mathcal{U}}\|_{\mathcal{U}^{\mathcal{U}}|_{\Omega^{\mathcal{U}}}} \leq C^{\mathcal{U}}\|\mathfrak{u}^{\mathcal{C}}\|_{E}\,, \qquad\qquad \|\mathfrak{u}_0^{\mathcal{P}}\|_{\mathcal{U}^{\mathcal{P}}|_{\Omega^{\mathcal{P}}}} \leq C^{\mathcal{P}}\|\mathfrak{u}^{\mathcal{C}}\|_{E}\,, \tag{B.41}$$

for all $\mathfrak{u}_0^{\mathcal{U}} = (\boldsymbol{\sigma}^{\mathcal{U}}, \boldsymbol{u}^{\mathcal{U}}, \boldsymbol{\omega}^{\mathcal{U}})$ and $\mathfrak{u}_0^{\mathcal{P}} = \boldsymbol{u}^{\mathcal{P}}$.

The last step involves finding the bounds for the interface variables, $\hat{\mathfrak{u}}^{\mathcal{U}} = (\hat{\boldsymbol{u}}^{\mathcal{U}}, \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}})$ and $\hat{\mathfrak{u}}^{\mathcal{P}} = \hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}$. Consider $\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}$ and let $M_0^{\mathcal{P}}$ satisfy $\big|b_0^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}_0^{\mathcal{P}}, \mathfrak{v}^{\mathcal{P}})\big| \leq M_0^{\mathcal{P}}\|\mathfrak{u}_0^{\mathcal{P}}\|_{\mathcal{U}^{\mathcal{P}}|_{\Omega^{\mathcal{P}}}}\|\mathfrak{v}^{\mathcal{P}}\|_{\mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}}$ for all

$\mathfrak{u}_0^{\mathcal{P}} \in \mathcal{U}^{\mathcal{P}}|_{\Omega^{\mathcal{P}}}$ and $\mathfrak{v}^{\mathcal{P}} \in \mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}$. Then, using the identities in Theorem A.3, it follows

$$\begin{aligned}\|\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}\|_{\boldsymbol{H}^{-1/2}(\partial\mathcal{T}^{\mathcal{P}})} &= \sup_{\boldsymbol{v}\in\boldsymbol{H}^1(\mathcal{T}^{\mathcal{P}})} \frac{|\langle\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{P}}, \mathbf{tr}_{\mathrm{grad}}^{\mathcal{T}}\boldsymbol{v}\rangle_{\partial\mathcal{T}^{\mathcal{P}}}|}{\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\mathcal{T}^{\mathcal{P}})}} \leq \sup_{\mathfrak{v}^{\mathcal{P}}\in\mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}} \frac{\left|b^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}^{\mathcal{P}},\mathfrak{v}^{\mathcal{P}}) - b_0^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}(\mathfrak{u}_0^{\mathcal{P}},\mathfrak{v}^{\mathcal{P}})\right|}{\|\mathfrak{v}^{\mathcal{P}}\|_{\mathcal{V}^{\mathcal{P}\mathcal{T}}|_{\Omega^{\mathcal{P}}}}} \\ &\leq \|\mathfrak{u}^{\mathcal{C}}\|_E + M_0^{\mathcal{P}}\|\mathfrak{u}_0^{\mathcal{P}}\|_{\mathcal{U}^{\mathcal{P}}|_{\Omega^{\mathcal{P}}}} \leq \left(1 + M_0^{\mathcal{P}}C^{\mathcal{P}}\right)\|\mathfrak{u}^{\mathcal{C}}\|_E\,.\end{aligned} \tag{B.42}$$

Similar calculations hold for $\hat{\boldsymbol{u}}^{\mathcal{U}}$ and $\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}}$. Summing the contributions from $\mathfrak{u}_0^{\mathcal{U}}$, $\hat{\mathfrak{u}}^{\mathcal{U}}$, $\mathfrak{u}_0^{\mathcal{P}}$ and $\hat{\mathfrak{u}}^{\mathcal{P}}$, yields the desired constant $C^{\mathcal{C}} = \frac{1}{\gamma^{\mathcal{C}}} > 0$, so that for all $\mathfrak{u}^{\mathcal{C}} = \left((\mathfrak{u}_0^{\mathcal{U}},\hat{\mathfrak{u}}^{\mathcal{U}}),(\mathfrak{u}_0^{\mathcal{P}},\hat{\mathfrak{u}}^{\mathcal{P}})\right) \in \mathcal{U}^{\mathcal{C}}$ it holds that $\|\mathfrak{u}^{\mathcal{C}}\|_{\mathcal{U}^{\mathcal{C}}} \leq C^{\mathcal{C}}\|\mathfrak{u}^{\mathcal{C}}\|_E$.

The more general case follows analogously, but some technicalities, mostly arising form the weak imposition of symmetry, are worth mentioning. To observe the changes, it suffices to consider the two-subdomain case involving the strong and ultraweak formulations. In this case, a similar procedure as before yields

$$b^{\mathcal{C}}(\mathfrak{u}^{\mathcal{C}},\mathfrak{v}^{\mathcal{C}_\Omega}) = b^{\mathcal{U}}(\mathfrak{u}^{\Omega},\mathfrak{v}^{\Omega}) + (\boldsymbol{\sigma}_{\mathbb{A}}^{\mathcal{S}},\boldsymbol{\nabla}\boldsymbol{v})_{\Omega^{\mathcal{S}}}\,, \tag{B.43}$$

where $\mathfrak{u}^{\mathcal{C}} = \left((\boldsymbol{\sigma}^{\mathcal{S}},\boldsymbol{u}^{\mathcal{S}}),(\boldsymbol{\sigma}^{\mathcal{U}},\boldsymbol{u}^{\mathcal{U}},\boldsymbol{\omega}^{\mathcal{U}},\hat{\boldsymbol{u}}^{\mathcal{U}},\hat{\boldsymbol{\sigma}}_{\mathbf{n}}^{\mathcal{U}})\right) \in \mathcal{U}^{\mathcal{C}}$, $\mathfrak{u}^{\Omega} = (\boldsymbol{\sigma}^{\Omega},\boldsymbol{u}^{\Omega},\boldsymbol{\omega}^{\Omega})$ is defined by restriction as $\mathfrak{u}^{\Omega}|_{\Omega^{\mathcal{S}}} = \left(\boldsymbol{\sigma}_{\mathbb{S}}^{\mathcal{S}},\boldsymbol{u}^{\mathcal{S}},\boldsymbol{\omega}(\boldsymbol{u}^{\mathcal{S}})\right)$ and $\mathfrak{u}^{\Omega}|_{\Omega^{\mathcal{U}}} = (\boldsymbol{\sigma}^{\mathcal{U}},\boldsymbol{u}^{\mathcal{U}},\boldsymbol{\omega}^{\mathcal{U}})$, whereas the test functions are $\mathfrak{v}^{\Omega} = (\boldsymbol{\tau},\boldsymbol{v})$ and $\mathfrak{v}^{\mathcal{C}_\Omega} = \left((\mathsf{S}:\boldsymbol{\tau},\boldsymbol{v},0)|_{\Omega^{\mathcal{S}}},(\boldsymbol{\tau},\boldsymbol{v})|_{\Omega^{\mathcal{U}}}\right)$. Thus, it is enough to bound $\|\boldsymbol{\sigma}_{\mathbb{A}}^{\mathcal{S}}\|_{\mathbf{L}^2(\Omega^{\mathcal{S}};\mathbb{A})}$, $\|\operatorname{\mathbf{div}}\boldsymbol{\sigma}^{\mathcal{S}}\|_{\boldsymbol{L}^2(\Omega^{\mathcal{S}})}$ and $\|\boldsymbol{\varepsilon}(\boldsymbol{u}^{\mathcal{S}})\|_{\mathbf{L}^2(\Omega^{\mathcal{S}};\mathbb{S})}$ in terms of $\|\mathfrak{u}^{\mathcal{C}}\|_E$ in order to obtain the desired bound of $\|\mathfrak{u}_0^{\mathcal{S}}\|_{\mathcal{U}^{\mathcal{S}}|_{\Omega^{\mathcal{S}}}} = \|(\boldsymbol{\sigma}^{\mathcal{S}},\boldsymbol{u}^{\mathcal{S}})\|_{\mathbf{H}(\mathbf{div},\Omega^{\mathcal{S}})\times\boldsymbol{H}^1(\Omega^{\mathcal{S}})}$ in terms of $\|\mathfrak{u}^{\mathcal{C}}\|_E$. These bounds are easily obtained through a careful choice of test functions in $\Omega^{\mathcal{S}}$. With these facts, the astute reader can deduce the proof for any other relevant and general scenario. □

Appendix C

Derivation of thermoviscoelastic equations

The purpose of this chapter is to provide a rigorous derivation of the equations of linear thermoviscoelasticity from first principles. These equations are solved in the frequency domain in Chapter 5. When the thermal interaction is assumed to vanish they result in the equations of linear viscoelasticity, which are solved in Chapter 4 and Chapter 5. In Section C.1 of this chapter the continuum mechanics equations of thermoviscoelastic materials with memory are presented following the pioneering work of Coleman [78]. In Section C.2, the equations are then rigorously linearized about a certain state. This leads to the linear first order system of thermoviscoelasticity equations akin to that proposed in [141, 197]. The first order system is presented in the time domain in Section C.3, where different simplifications of the equations are discussed.

C.1 Nonlinear thermodynamics of materials with memory

The purpose of this section is to present the thermoviscoelastic equations of solid materials with fading memory in the framework of continuum mechanics. The approach is due to the seminal work of Coleman [78].

Consider a material element domain $\Omega(t) \subseteq \mathbb{R}^3$ for $t > 0$, so that the function $\boldsymbol{x}(\boldsymbol{X}, t) \in \Omega(t)$ provides the motion of a particle initially at $\boldsymbol{X} \in \Omega = \Omega(0)$. Consequently, the displacement of the particle is $\boldsymbol{u}(\boldsymbol{X}, t) = \boldsymbol{x}(\boldsymbol{X}, t) - \boldsymbol{X}$, while the deformation gradient is $\boldsymbol{F} = \boldsymbol{\nabla}\boldsymbol{x} = \mathbf{I} + \boldsymbol{\nabla}\boldsymbol{u}$, where $\boldsymbol{\nabla} = \boldsymbol{\nabla}_{\boldsymbol{X}}$ and $\mathbf{I}$ is the identity transformation. In the fixed configuration Ω, the conservation of mass, momenta and energy, and the second law of thermodynamics (in the form of the Clausius-Duhem inequality) [187] are,

$$\text{conservation of mass:} \qquad \dot{\rho} = 0\,, \tag{C.1}$$

$$\text{conservation of linear momentum:} \qquad \rho\ddot{\boldsymbol{u}} = \mathbf{div}(\boldsymbol{F}\boldsymbol{S}) + \boldsymbol{f}\,, \tag{C.2}$$

conservation of angular momentum: $\boldsymbol{S} = \boldsymbol{S}^{\mathsf{T}}\,,$ (C.3)

conservation of energy: $\rho\dot{\boldsymbol{E}} = \boldsymbol{S}:\dot{\boldsymbol{E}} - \operatorname{div}\boldsymbol{q} + r\,,$ (C.4)

second law of thermodynamics: $\rho\dot{\eta} + \operatorname{div}(\frac{\boldsymbol{q}}{\theta}) - \frac{r}{\theta} \geq 0\,,$ (C.5)

where ρ is the mass density with $\dot{\rho} = \frac{\partial\rho(\boldsymbol{X},t)}{\partial t}$, $\boldsymbol{S}$ is the second Piola-Kirchhoff stress tensor, $\boldsymbol{f}$ represents the body force density, $\ddot{\boldsymbol{u}} = \frac{\partial^2\boldsymbol{u}(\boldsymbol{X},t)}{\partial t^2}$ is the acceleration, e is the specific internal energy density (not due to the motion) with $\dot{e} = \frac{\partial e(\boldsymbol{X},t)}{\partial t}$, $\boldsymbol{E} = \frac{1}{2}(\boldsymbol{F}^{\mathsf{T}}\boldsymbol{F} - \mathbf{I})$ is the Green strain tensor with $\dot{\boldsymbol{E}} = \frac{\partial\boldsymbol{E}(\boldsymbol{X},t)}{\partial t}$, $\boldsymbol{q}$ is the heat flux, r is the heat per unit volume due to internal sources, η is the specific entropy density with $\dot{\eta} = \frac{\partial\eta(\boldsymbol{X},t)}{\partial t}$, and θ is the absolute temperature. Note that all variables are functions of $\boldsymbol{X} \in \Omega$ and $t > 0$ and are actually pullbacks of variables in $\Omega(t)$. Indeed, $\rho = \widetilde{\rho}\det\boldsymbol{F}$, $\boldsymbol{S} = \boldsymbol{F}^{-1}\widetilde{\boldsymbol{\sigma}}\boldsymbol{F}^{-\mathsf{T}}\det\boldsymbol{F}$, $\boldsymbol{f} = \widetilde{\boldsymbol{f}}\det\boldsymbol{F}$, $e = \widetilde{e}$, $\boldsymbol{q} = \boldsymbol{F}^{-1}\widetilde{\boldsymbol{q}}\det\boldsymbol{F}$, $r = \widetilde{r}\det\boldsymbol{F}$, $\eta = \widetilde{\eta}$ and $\theta = \widetilde{\theta}$, where the tilded variables are functions of $\boldsymbol{x} \in \Omega(t)$ and $t > 0$ and are being evaluated at $(\boldsymbol{x}(\boldsymbol{X},t),t)$ in the preceding expressions. Here, $\widetilde{\boldsymbol{\sigma}}$ is the Cauchy stress tensor. Hence, (C.1)–(C.5) are to be interpreted in the fixed configuration Ω, so that $\operatorname{div}(\cdot) = \nabla_{\boldsymbol{X}}\cdot(\cdot)$ and $\mathbf{div}(\cdot) = \boldsymbol{\nabla}_{\boldsymbol{X}}\cdot(\cdot)$.

Introducing the specific Helmholtz free energy density $\psi = e - \eta\theta$ and using the conservation of energy allows the second law of thermodynamics to be rewritten in terms of $\dot{\psi} = \frac{\partial\psi(\boldsymbol{X},t)}{\partial t}$ and $\dot{\theta} = \frac{\partial\theta(\boldsymbol{X},t)}{\partial t}$ as

$$-\rho\dot{\psi} - \rho\eta\dot{\theta} + \boldsymbol{S}:\dot{\boldsymbol{E}} - \tfrac{1}{\theta}\boldsymbol{q}\cdot\nabla\theta \geq 0\,. \tag{C.6}$$

At this point the equations cannot be solved, since constitutive models for ψ, $\boldsymbol{S}$, η and $\boldsymbol{q}$ (which are sufficient to determine e) are unknown. The main physical question lies in establishing which variables determine the behavior of ψ, $\boldsymbol{S}$, η and $\boldsymbol{q}$. That is, what are the dependencies of the constitutive models. In classical theoretical hyperelasticity the assumption is that the models are dependent on $\boldsymbol{X}$ and the current values of $\boldsymbol{F}$, θ and $\nabla\theta$.

However, a much more general yet challenging assumption is that the models for ψ, $\boldsymbol{S}$, η and $\boldsymbol{q}$ depend on $\boldsymbol{X}$ and the *histories* of $\boldsymbol{F}$, θ and $\nabla\theta$ [78, 187] (see [140] for even more generality). The history of $\boldsymbol{F}$ at time t is denoted as $\boldsymbol{F}^t$, and defined as

$$\boldsymbol{F}^t(\boldsymbol{X},s) = \boldsymbol{F}(\boldsymbol{X},t-s)\,, \qquad \forall\boldsymbol{X}\in\Omega\,,\quad \forall s\in[0,\infty)\,. \tag{C.7}$$

Hence, $\boldsymbol{F}^t$ encompasses all the values of $\boldsymbol{F}$ in the past, with small values of s representing the recent past and large values of s representing the distant past. Notice, the history requires the knowledge of $\boldsymbol{F}$ even for all $t < 0$, but this is not an issue since it is usually assumed the material has remained at rest or in a constant state up to $t = 0$. For a fixed $\boldsymbol{X}$, the question arises as to what functional space the history $\boldsymbol{F}^t(\boldsymbol{X}, \cdot)$ lies. The principle of fading memory, introduced in [82, 83, 78], states that the history is at least in a weighted space of square integrable functions, $\boldsymbol{F}^t(\boldsymbol{X}, \cdot) \in \mathsf{L}^2_w(0, \infty; \mathbb{M})$, where the weight function w_{L^2} defined in $(0, \infty)$ is some almost-everywhere positive function satisfying that $\int_0^\infty w_{L^2}(s)\, \mathrm{d}s < \infty$ (see [81] for more generality). More precisely,

$$\begin{aligned} L^2_w(0,\infty) &= \{\theta^t : (0,\infty) \to \mathbb{R} \mid \|\theta^t\|_{L^2_w(0,\infty)} = (\theta^t, \theta^t)_{L^2_w} < \infty\}\,, \\ \mathsf{L}^2_w(0,\infty;\mathbb{U}) &= \{\boldsymbol{A}^t : (0,\infty) \to \mathbb{U} \mid \|\boldsymbol{A}^t\|_{\mathsf{L}^2_w(0,\infty;\mathbb{U})} = (\boldsymbol{A}^t, \boldsymbol{A}^t)_{L^2_w} < \infty\}\,, \end{aligned} \tag{C.8}$$

where $\mathbb{U}$ is a subspace of $\mathbb{M}$, the space of 3×3 matrices. The referenced inner product is simply $(u^t, v^t)_{L^2_w} = \int_0^\infty \mathrm{tr}_{\mathbb{M}}\big((u^t(s))^{\mathsf{T}} v^t(s)\big)\, w_{L^2}(s)\, \mathrm{d}s$, where $\mathrm{tr}_{\mathbb{M}}$ is the usual trace of a matrix so that $\mathrm{tr}_{\mathbb{M}}\big((u^t(s))^{\mathsf{T}} v^t(s)\big)$ is $u^t(s)v^t(s)$ if the range of u^t and v^t is $\mathbb{R}$, or $u^t(s) : v^t(s)$ if the range is $\mathbb{U} \subseteq \mathbb{M}$. The assumption that $\int_0^\infty w_{L^2}(s)\, \mathrm{d}s < \infty$ obviously implies that $\lim_{s\to\infty} w_{L^2}(s) = 0$. Hence, the physical aspect behind the principle of fading memory is that the recent memory (small s) carries more weight than the distant memory (large s) which is essentially forgotten. Obviously, the same assertions apply to the history of θ at time t, denoted as θ^t, and to any other variable with a superscript t, so that in particular $\theta^t(\boldsymbol{X}, \cdot) \in L^2_w(0, \infty)$ for all $\boldsymbol{X} \in \Omega$.

The specific assumption made throughout this work [78] is that the constitutive models are of the form

$$\begin{aligned} \psi(\boldsymbol{X}, t) &= \psi^{\mathrm{r}_0}(\boldsymbol{X}, \boldsymbol{F}^t(\boldsymbol{X}, \cdot), \theta^t(\boldsymbol{X}, \cdot), \nabla\theta(\boldsymbol{X}, t))\,, \\ \boldsymbol{S}(\boldsymbol{X}, t) &= \boldsymbol{S}^{\mathrm{r}_0}(\boldsymbol{X}, \boldsymbol{F}^t(\boldsymbol{X}, \cdot), \theta^t(\boldsymbol{X}, \cdot), \nabla\theta(\boldsymbol{X}, t))\,, \\ \eta(\boldsymbol{X}, t) &= \eta^{\mathrm{r}_0}(\boldsymbol{X}, \boldsymbol{F}^t(\boldsymbol{X}, \cdot), \theta^t(\boldsymbol{X}, \cdot), \nabla\theta(\boldsymbol{X}, t))\,, \\ \boldsymbol{q}(\boldsymbol{X}, t) &= \boldsymbol{q}^{\mathrm{r}_0}(\boldsymbol{X}, \boldsymbol{F}^t(\boldsymbol{X}, \cdot), \theta^t(\boldsymbol{X}, \cdot), \nabla\theta(\boldsymbol{X}, t))\,. \end{aligned} \tag{C.9}$$

The functions ψ^{r_0}, $\boldsymbol{S}^{\mathrm{r}_0}$, η^{r_0} and $\boldsymbol{q}^{\mathrm{r}_0}$ are called response functions. This is by no means the most general hypothesis. Indeed, a dependence on the history $\nabla\theta^t$ is plausible [79, 141], as well as dependencies on the history $\nabla\boldsymbol{E}^t$ [140]. However, this choice is general enough for most physical purposes.

Due to the principle of material frame indifference [187], one can show there exist (different) response functions that depend on $\boldsymbol{E}^t$ (defined by $\boldsymbol{E}^t(\boldsymbol{X},s)=\boldsymbol{E}(\boldsymbol{X},t-s)$) as opposed to $\boldsymbol{F}^t$. Being consistent with the principle of fading memory, it is assumed $\boldsymbol{E}^t(\boldsymbol{X},\cdot)\in\mathsf{L}^2_w(0,\infty;\mathbb{S})$ for all $\boldsymbol{X}\in\Omega$, where $\mathbb{S}$ is the space of all symmetric 3×3 matrices. Next, let the *difference histories* of $\boldsymbol{E}$ and θ at t be $\boldsymbol{E}^t_d$ and θ^t_d, defined by

$$\begin{aligned}\boldsymbol{E}^t_d(\boldsymbol{X},s)&=\boldsymbol{E}^t(\boldsymbol{X},s)-\boldsymbol{E}(\boldsymbol{X},t)=\boldsymbol{E}(\boldsymbol{X},t-s)-\boldsymbol{E}(\boldsymbol{X},t)\,,\\ \theta^t_d(\boldsymbol{X},s)&=\theta^t(\boldsymbol{X},s)-\theta(\boldsymbol{X},t)=\theta(\boldsymbol{X},t-s)-\theta(\boldsymbol{X},t)\,,\end{aligned}\tag{C.10}$$

for all $\boldsymbol{X}\in\Omega$ and $s\in[0,\infty)$. Clearly a dependence on $\boldsymbol{E}^t$ is equivalent to a joint dependence on $\boldsymbol{E}$ and $\boldsymbol{E}^t_d$, but the latter choice is more beneficial to establish future results and to elucidate analogies with hyperelastic materials which are dependent on the current values of $\boldsymbol{E}$ and θ, but not on the histories. Therefore, it is convenient to write the response functions in terms of $\boldsymbol{E}$ and $\boldsymbol{E}^t_d$, as opposed to $\boldsymbol{E}^t$. Due to the properties of the weight w_{L^2}, it is clear that $\boldsymbol{E}^t(\boldsymbol{X},\cdot)\in\mathsf{L}^2_w(0,\infty;\mathbb{S})$ if and only if $\boldsymbol{E}^t_d(\boldsymbol{X},\cdot)\in\mathsf{L}^2_w(0,\infty;\mathbb{S})$. The same applies to a dependence on θ^t, which is better replaced by a joint dependence on θ and θ^t_d. Thus, the constitutive models are written in terms of new response functions as

$$\begin{aligned}\psi(\boldsymbol{X},t)&=\psi^{\mathfrak{r}}(\boldsymbol{X},\boldsymbol{E}(\boldsymbol{X},t),\boldsymbol{E}^t_d(\boldsymbol{X},\cdot),\theta(\boldsymbol{X},t),\theta^t_d(\boldsymbol{X},\cdot),\nabla\theta(\boldsymbol{X},t))\,,\\ \boldsymbol{S}(\boldsymbol{X},t)&=\boldsymbol{S}^{\mathfrak{r}}(\boldsymbol{X},\boldsymbol{E}(\boldsymbol{X},t),\boldsymbol{E}^t_d(\boldsymbol{X},\cdot),\theta(\boldsymbol{X},t),\theta^t_d(\boldsymbol{X},\cdot),\nabla\theta(\boldsymbol{X},t))\,,\\ \eta(\boldsymbol{X},t)&=\eta^{\mathfrak{r}}(\boldsymbol{X},\boldsymbol{E}(\boldsymbol{X},t),\boldsymbol{E}^t_d(\boldsymbol{X},\cdot),\theta(\boldsymbol{X},t),\theta^t_d(\boldsymbol{X},\cdot),\nabla\theta(\boldsymbol{X},t))\,,\\ \boldsymbol{q}(\boldsymbol{X},t)&=\boldsymbol{q}^{\mathfrak{r}}(\boldsymbol{X},\boldsymbol{E}(\boldsymbol{X},t),\boldsymbol{E}^t_d(\boldsymbol{X},\cdot),\theta(\boldsymbol{X},t),\theta^t_d(\boldsymbol{X},\cdot),\nabla\theta(\boldsymbol{X},t))\,.\end{aligned}\tag{C.11}$$

From now on, whenever it is clear from the context, the dependence on $(\boldsymbol{X},t)$, $(\boldsymbol{X},\boldsymbol{E},\boldsymbol{E}^t_d,\theta,\theta^t_d,\nabla\theta)$ and the superscript $\mathfrak{r}$ will be omitted.

Before proceeding, a comment on the partial derivatives of the response functions is necessary. Indeed, sufficient derivatives of the response functionals are assumed to exist to derive the results that will follow. More precisely this is referring to Fréchet derivatives (which collapse to usual derivatives if the variable is in $\mathbb{R}$). This technicality is particularly necessary when looking at partial derivatives with respect to the history variables which lie in nontrivial Hilbert spaces. In this case the partial derivatives are continuous linear functionals acting on the Hilbert space in question, so they belong to the dual of the Hilbert space. Using the Riesz representation theorem, this allows to identify the partial derivative with an element of the Hilbert space itself. Hence, it is valid to write $\frac{\partial\psi}{\partial\boldsymbol{E}}(\Xi) \in \mathbb{S}$, $\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\Xi,\cdot) \in \mathsf{L}_w^2(0,\infty;\mathbb{S})$, $\frac{\partial\psi}{\partial\theta}(\Xi) \in \mathbb{R}$, $\frac{\partial\psi}{\partial\theta_d^t}(\Xi,\cdot) \in L_w^2(0,\infty)$ and $\frac{\partial\psi}{\partial\nabla\theta}(\Xi) \in \mathbb{R}^3$, where $\Xi = (\boldsymbol{X},\boldsymbol{E},\boldsymbol{E}_d^t,\theta,\theta_d^t,\nabla\theta)$. Moreover, consider the following definitions to be used in the coming calculations,

$$\nabla_{\boldsymbol{E}_d^t}\psi(\Xi) = \int_0^\infty \frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\Xi,s)w_{L^2}(s)\,\mathrm{d}s \in \mathbb{S}\,, \qquad \nabla_{\theta_d^t}\psi(\Xi) = \int_0^\infty \frac{\partial\psi}{\partial\theta_d^t}(\Xi,s)w_{L^2}(s)\,\mathrm{d}s \in \mathbb{R}\,. \tag{C.12}$$

The same logic applies to higher order derivatives and derivatives of the other response functions.

As an example, consider the time derivative of ψ,

$$\frac{\partial\psi}{\partial t} = \frac{\partial\psi^{\mathfrak{r}}}{\partial\boldsymbol{X}}\cdot\frac{\partial\boldsymbol{X}}{\partial t} + \frac{\partial\psi^{\mathfrak{r}}}{\partial\boldsymbol{E}} : \frac{\partial\boldsymbol{E}}{\partial t} + \Big(\frac{\partial\psi^{\mathfrak{r}}}{\partial\boldsymbol{E}_d^t},\frac{\partial\boldsymbol{E}_d^t}{\partial t}\Big)_{L_w^2} + \frac{\partial\psi^{\mathfrak{r}}}{\partial\theta}\frac{\partial\theta}{\partial t} + \Big(\frac{\partial\psi^{\mathfrak{r}}}{\partial\theta_d^t},\frac{\partial\theta_d^t}{\partial t}\Big)_{L_w^2} + \frac{\partial\psi^{\mathfrak{r}}}{\partial\nabla\theta}\cdot\frac{\partial\nabla\theta}{\partial t}\,, \tag{C.13}$$

where the partial derivatives of $\psi^{\mathfrak{r}}$ are evaluated at $\Xi(\boldsymbol{X},t)$ with $\Xi = (\boldsymbol{X},\boldsymbol{E},\boldsymbol{E}_d^t,\theta,\theta_d^t,\nabla\theta)$, and all other terms evaluated at $(\boldsymbol{X},t)$. Clearly $\frac{\partial\boldsymbol{X}}{\partial t} = 0$, while $\frac{\partial\boldsymbol{E}_d^t}{\partial t} = \dot{\boldsymbol{E}}_d^t = \dot{\boldsymbol{E}}^t - \dot{\boldsymbol{E}}$ and $\frac{\partial\theta_d^t}{\partial t} = \dot{\theta}_d^t = \dot{\theta}^t - \dot{\theta}$, where $\dot{\boldsymbol{E}} = \frac{\partial\boldsymbol{E}}{\partial t}$, $\dot{\boldsymbol{E}}^t(\boldsymbol{X},s) = \dot{\boldsymbol{E}}(\boldsymbol{X},t-s)$, $\dot{\theta} = \frac{\partial\theta}{\partial t}$, and $\dot{\theta}^t(\boldsymbol{X},s) = \dot{\theta}(\boldsymbol{X},t-s)$. Denoting $\dot{\psi} = \frac{\partial\psi}{\partial t}$ and $\frac{\partial\nabla\theta}{\partial t} = \nabla\frac{\partial\theta}{\partial t} = \nabla\dot{\theta}$ and omitting the superscript $\mathfrak{r}$ leads to

$$\begin{aligned}\dot{\psi} &= \frac{\partial\psi}{\partial\boldsymbol{E}} : \dot{\boldsymbol{E}} + \Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t},\dot{\boldsymbol{E}}_d^t\Big)_{L_w^2} + \frac{\partial\psi}{\partial\theta}\dot{\theta} + \Big(\frac{\partial\psi}{\partial\theta_d^t},\dot{\theta}_d^t\Big)_{L_w^2} + \frac{\partial\psi}{\partial\nabla\theta}\cdot\nabla\dot{\theta}\\ &= \Big(\frac{\partial\psi}{\partial\boldsymbol{E}} - \nabla_{\boldsymbol{E}_d^t}\psi\Big) : \dot{\boldsymbol{E}} + \Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t},\dot{\boldsymbol{E}}^t\Big)_{L_w^2} + \Big(\frac{\partial\psi}{\partial\theta} - \nabla_{\theta_d^t}\psi\Big)\dot{\theta} + \Big(\frac{\partial\psi}{\partial\theta_d^t},\dot{\theta}^t\Big)_{L_w^2} + \frac{\partial\psi}{\partial\nabla\theta}\cdot\nabla\dot{\theta}\,,\end{aligned} \tag{C.14}$$

since $(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t},\dot{\boldsymbol{E}})_{L_w^2} = \nabla_{\boldsymbol{E}_d^t}\psi : \dot{\boldsymbol{E}}$ and $(\frac{\partial\psi}{\partial\theta_d^t},\dot{\theta})_{L_w^2} = \nabla_{\theta_d^t}\psi\,\dot{\theta}$ due to $\dot{\boldsymbol{E}}$ and $\dot{\theta}$ being constant for any history time s.

With these facts in mind, it is possible to state the main conclusions derived by Coleman [78]. Using (C.6), he proved that the response functions for ψ, $\boldsymbol{S}$ and η in (C.11) are independent of $\nabla\theta$ (so that $\frac{\partial\psi}{\partial\nabla\theta} = 0$, $\frac{\partial\boldsymbol{S}}{\partial\nabla\theta} = 0$ and $\frac{\partial\eta}{\partial\nabla\theta} = 0$) and

$$\begin{aligned}\boldsymbol{S} &= \rho\Big(\frac{\partial\psi}{\partial\boldsymbol{E}} - \nabla_{\boldsymbol{E}_d^t}\psi\Big)\,,\\ \eta &= -\Big(\frac{\partial\psi}{\partial\theta} - \nabla_{\theta_d^t}\psi\Big)\,,\\ -\rho\Big(\Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t},\dot{\boldsymbol{E}}^t\Big)_{L_w^2} &+ \Big(\frac{\partial\psi}{\partial\theta_d^t},\dot{\theta}^t\Big)_{L_w^2}\Big) - \frac{1}{\theta}\boldsymbol{q}\cdot\nabla\theta \geq 0\,.\end{aligned} \tag{C.15}$$

Moreover, it was established that of all histories, $\boldsymbol{E}^t$ and θ^t, ending with given values of $\boldsymbol{E}$ and θ respectively, those corresponding to constant values of $\boldsymbol{E}$ and θ for all times result in the least Helmholtz free energy. In other words, given arbitrary strain and temperature fields at time t, $\boldsymbol{E}_0(\boldsymbol{X})$ and $\theta_0(\boldsymbol{X})$, the histories $\boldsymbol{E}^t(\boldsymbol{X},s) = \boldsymbol{E}_0(\boldsymbol{X})$ and $\theta^t(\boldsymbol{X},s) = \theta_0(\boldsymbol{X})$ for all $s \geq 0$, resulting in difference histories $\boldsymbol{E}_d^t(\boldsymbol{X},s) = 0$ and $\theta_d^t(\boldsymbol{X},s) = 0$, produce a minimum in the response function of the Helmholtz free energy with respect to the difference histories of $\boldsymbol{E}_d^t$ and θ_d^t. Therefore,

$$\begin{aligned} &\frac{\partial \psi}{\partial \boldsymbol{E}_d^t}(\Xi_{0^t}, \cdot) = 0\,, \qquad && \nabla_{\boldsymbol{E}_d^t}\psi(\Xi_{0^t}) = 0\,, \\ &\frac{\partial \psi}{\partial \theta_d^t}(\Xi_{0^t}, \cdot) = 0\,, \qquad && \nabla_{\theta_d^t}\psi(\Xi_{0^t}) = 0\,, \end{aligned} \tag{C.16}$$

where $\Xi_{0^t} = (\boldsymbol{X}, \boldsymbol{E}, 0, \theta, 0)$, so that $\boldsymbol{X}$, $\boldsymbol{E}$ and θ are arbitrary, while $\boldsymbol{E}_d^t = 0$ and $\theta_d^t = 0$ (note ψ is not dependent on $\nabla\theta$ anymore). Additionally, as noticed by Coleman and Gurtin [79], in this same scenario of zero difference histories, the dissipation inequality in (C.15) reduces to $-\frac{1}{\theta}\boldsymbol{q}(\Xi_{0^t}^q)\cdot\nabla\theta \geq 0$, where $\Xi_{0^t}^q = (\boldsymbol{X}, \boldsymbol{E}, 0, \theta, 0, \nabla\theta)$. From this point one can derive that at the zero temperature gradient, the heat conductivity is positive semidefinite [80] (it can additionally be shown to be *symmetric* positive semidefinite under certain assumptions of the nature of the anisotropy of the material as in [233] or under "thermally stable" conditions as in [88]), and the heat flux vanishes [79]. That is, at the state given by $\Xi_{00}^q = (\boldsymbol{X}, \boldsymbol{E}, 0, \theta, 0, 0)$, it holds that $-\mathsf{v}^{\mathsf{T}}\frac{\partial \boldsymbol{q}}{\partial \nabla\theta}(\Xi_{00}^q)\mathsf{v} \geq 0$ for all $\mathsf{v} \in \mathbb{R}^3$, and $\boldsymbol{q}(\Xi_{00}^q) = 0$. This implies some partial derivatives at this point are zero,

$$\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{E}}(\Xi_{00}^q) = 0\,, \qquad \frac{\partial \boldsymbol{q}}{\partial \theta}(\Xi_{00}^q) = 0\,. \tag{C.17}$$

The previous results are very important. Firstly, if the response functions prove to be independent of the difference histories, then the statements in (C.15) clearly collapse to the simpler and better known expressions in dynamic nonlinear hyperelasticity. Secondly, even with this dependence on the difference histories, the case of zero difference histories (constant strain an temperature at given values) is essentially an equilibrium state that matches that of static nonlinear hyperelasticity. Alternatively, this is equivalent to saying that in the limit of infinitely slow processes the behavior matches that of static nonlinear hyperelasticity [78]. Thirdly, these results can be used when linearizing about the points of zero difference histories. This is the subject of the next section.

C.2 Linearized thermoviscoelasticity

In this section the rigorous linearization of the physical laws and constitutive models is carried out as in [83, 77, 87, 53] but in more generality. This leads to a system akin to that proposed in [141, 197]. The are several technical details, so the uninterested reader may skip directly to the derived first-order system in Section C.3.

To begin the linearization process the first step is to choose the origin about which the equations and variables will be linearized. Since only small perturbations about this origin will be contemplated, it is natural to consider the origin as the static reference configuration at rest (no deformation) with a fixed constant temperature distribution. Hence, in terms of displacements and temperature, the idea is to linearize around a zero strain ($\boldsymbol{\nabla u} = 0$), a fixed temperature $\bar{\theta}$ ($\theta = \bar{\theta}$), a zero temperature gradient ($\nabla\theta = 0$), a zero strain rate ($\boldsymbol{\nabla}\dot{\boldsymbol{u}} = 0$), and a zero temperature rate ($\dot{\theta} = 0$). Hence, at the origin of the linearization, $\boldsymbol{E} = 0$ and $\dot{\boldsymbol{E}} = 0$, since $\boldsymbol{E} = \frac{1}{2}(\boldsymbol{F}^{\mathsf{T}}\boldsymbol{F} - \mathbf{I})$ and $\dot{\boldsymbol{E}} = \frac{1}{2}(\dot{\boldsymbol{F}}^{\mathsf{T}}\boldsymbol{F} + \boldsymbol{F}^{\mathsf{T}}\dot{\boldsymbol{F}})$ with $\boldsymbol{F} = \mathbf{I} + \boldsymbol{\nabla u}$. This is valid for all times, meaning that at the origin of the linearization $\boldsymbol{E}^t(\boldsymbol{X}, s) = 0$ and $\theta^t(\boldsymbol{X}, s) = \bar{\theta}(\boldsymbol{X})$ for all $s \geq 0$, so that $\boldsymbol{E}^t_d = 0$ and $\theta^t_d = 0$. The tuples at the origin are thus $\bar{\Xi} = (\boldsymbol{X}, \bar{\boldsymbol{E}}, \bar{\boldsymbol{E}}^t_d, \bar{\theta}, \bar{\theta}^t_d) = (\boldsymbol{X}, 0, 0, \bar{\theta}, 0)$ and $\bar{\Xi}^q = (\bar{\Xi}, \overline{\nabla\theta}) = (\boldsymbol{X}, 0, 0, \bar{\theta}, 0, 0)$.

Naturally, small variations of the strains, temperatures, temperature gradients, strain rates and temperature rates are assumed about the linearization origin. To express this mathematically, first let $\mathfrak{k}(\delta)$ be some variable dependent on $\delta > 0$ and lying in a normed space. Then, the notation $\mathfrak{k}(\delta) = \mathcal{O}(\delta^n)$ as $\delta \to 0$ means that there exist $\delta_0 > 0$ and $M_{\mathcal{O}} > 0$ such that $\|\mathfrak{k}(\delta)\| \leq M_{\mathcal{O}}\delta^n$ for every $\delta \in (0, \delta_0)$, where $n \in \mathbb{N}$. Meanwhile, the notation $\mathfrak{k}(\delta) = \mathcal{o}(\delta^n)$ as $\delta \to 0$ means $\lim_{\delta\to 0} \frac{\|\mathfrak{k}(\delta)\|}{\delta^n} = 0$, where $n \in \mathbb{N}$. With this in mind, let

$$\boldsymbol{\nabla u} = \mathcal{O}(\delta)\,, \qquad \theta - \bar{\theta} = \mathcal{O}(\delta)\,, \qquad \nabla\theta = \mathcal{O}(\delta)\,, \qquad \boldsymbol{\nabla}\dot{\boldsymbol{u}} = \mathcal{O}(\delta)\,, \qquad \dot{\theta} = \mathcal{O}(\delta)\,. \tag{C.18}$$

Hence, for a given $\delta > 0$, the norms of the preceding variables at all points $(\boldsymbol{X}, t)$ are assumed to be roughly of order δ. Due to the definition of $\boldsymbol{E}$ and the expression for $\dot{\boldsymbol{E}}$, it easily follows that these variables are of order δ as well, and more importantly, since this is true at all times, it can be shown [83]

$$\boldsymbol{E} = \mathcal{O}(\delta)\,, \quad \dot{\boldsymbol{E}} = \mathcal{O}(\delta)\,, \quad \boldsymbol{E}^t_d = \mathcal{O}(\delta)\,, \quad \dot{\boldsymbol{E}}^t = \mathcal{O}(\delta)\,, \quad \theta^t_d = \mathcal{O}(\delta)\,, \quad \dot{\theta}^t = \mathcal{O}(\delta)\,. \tag{C.19}$$

Note for the histories this implies looking at the norms $\|\boldsymbol{E}_d^t\|_{\mathsf{L}^2_w(0,\infty;\mathbb{S})}$, $\|\dot{\boldsymbol{E}}^t\|_{\mathsf{L}^2_w(0,\infty;\mathbb{S})}$, $\|\theta_d^t\|_{L^2_w(0,\infty)}$ and $\|\dot{\theta}^t\|_{L^2_w(0,\infty)}$. Next, define the tuples $\Xi = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{E}_d^t, \theta, \theta_d^t)$ and $\Xi^q = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{E}_d^t, \theta, \theta_d^t, \nabla\theta)$ satisfying the assumptions of small variations. Then it follows $(\Xi - \bar{\Xi}) = \mathcal{O}(\delta)$ and $(\Xi^q - \bar{\Xi}^q) = \mathcal{O}(\delta)$. Finally, recall the engineering strain is defined as $\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^\mathsf{T})$, from which it easily follows that

$$\begin{aligned}\boldsymbol{E} &= \tfrac{1}{2}(\boldsymbol{F}^\mathsf{T}\boldsymbol{F} - \mathbf{I}) = \boldsymbol{\varepsilon} + \tfrac{1}{2}\boldsymbol{\nabla u}^\mathsf{T}\boldsymbol{\nabla u} = \boldsymbol{\varepsilon} + \mathcal{O}(\delta^2)\,,\\ \dot{\boldsymbol{E}} &= \tfrac{1}{2}(\dot{\boldsymbol{F}}^\mathsf{T}\boldsymbol{F} + \boldsymbol{F}^\mathsf{T}\dot{\boldsymbol{F}}) = \dot{\boldsymbol{\varepsilon}} + \tfrac{1}{2}(\boldsymbol{\nabla}\dot{\boldsymbol{u}}^\mathsf{T}\boldsymbol{\nabla u} + \boldsymbol{\nabla u}^\mathsf{T}\boldsymbol{\nabla}\dot{\boldsymbol{u}}) = \dot{\boldsymbol{\varepsilon}} + \mathcal{O}(\delta^2)\,.\end{aligned} \tag{C.20}$$

Note that the engineering strain is a linear operator of $\boldsymbol{u}$, while $\boldsymbol{E}$ is not. This is enough to proceed to the linearization of the equations themselves.

The first observation to make is that it is only required to solve the equations of conservation of linear momentum and the conservation of energy for the displacement and temperature, provided constitutive models are known for the second Piola-Kirchhoff stress tensor, the entropy and the heat. The first two constitutive models are known through the result (C.15), which is assumed to hold so that the Clausius-Duhem inequality (expressing the second law of thermodynamics) is satisfied. The equation for conservation of mass does not need to be solved, but if required the mass density in the current configuration can be calculated from a knowledge of the displacement. Meanwhile the conservation of angular momentum is satisfied automatically by the natural symmetry of the deduced constitutive model in (C.15).

The conservation of linear momentum and conservation of energy can be rewritten in terms of the Helmholtz free energy as

$$\begin{gathered}\mathbf{div}\,\boldsymbol{P} + \boldsymbol{f} = \rho\ddot{\boldsymbol{u}}\,,\\ \rho\dot{\psi} + \rho\eta\dot{\theta} + \rho\dot{\eta}\theta = \boldsymbol{S}:\dot{\boldsymbol{E}} - \operatorname{div}\boldsymbol{q} + r\,,\end{gathered} \tag{C.21}$$

where $\boldsymbol{P} = \boldsymbol{FS}$ is the first Piola-Kirchhoff stress tensor. Substituting the results in (C.15) into (C.14) and using that $\frac{\partial\psi}{\partial\nabla\theta} = 0$ gives

$$\dot{\psi} = \frac{1}{\rho}\boldsymbol{S}:\dot{\boldsymbol{E}} - \eta\dot{\theta} + \Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}, \dot{\boldsymbol{E}}^t\Big)_{L^2_w} + \Big(\frac{\partial\psi}{\partial\theta_d^t}, \dot{\theta}^t\Big)_{L^2_w}\,. \tag{C.22}$$

Hence, the equation for conservation of energy becomes

$$\rho\dot{\eta}\theta + \Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}, \dot{\boldsymbol{E}}^t\Big)_{L^2_w} + \Big(\frac{\partial\psi}{\partial\theta_d^t}, \dot{\theta}^t\Big)_{L^2_w} = -\operatorname{div}\boldsymbol{q} + r\,. \tag{C.23}$$

The first equation to be linearized in detail is the second Piola-Kirchhoff stress tensor, $\boldsymbol{S}$, which is given by (C.15). Using the definition of Fréchet derivative at $\bar{\Xi}$ (which is assumed to exist), it follows

$$\boldsymbol{S}(\Xi) = \boldsymbol{S}(\bar{\Xi}) + \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) : \boldsymbol{E} + \Big(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}), \boldsymbol{E}_d^t\Big)_{L_w^2} + \frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi})(\theta - \bar{\theta}) + \Big(\frac{\partial \boldsymbol{S}}{\partial \theta_d^t}(\bar{\Xi}), \theta_d^t\Big)_{L_w^2} + o(\delta)\,. \tag{C.24}$$

Next, recall $\boldsymbol{S} = \rho\Big(\frac{\partial \psi}{\partial \boldsymbol{E}} - \nabla_{\boldsymbol{E}_d^t}\psi\Big)$ and the results in (C.16). First notice that

$$\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) = \rho \frac{\partial^2 \psi}{\partial \boldsymbol{E}^2}(\bar{\Xi})\,, \qquad \qquad \frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi}) = \rho \frac{\partial^2 \psi}{\partial \theta \partial \boldsymbol{E}}(\bar{\Xi})\,, \tag{C.25}$$

since $\boldsymbol{E}_d^t$ and θ_d^t remain fixed at zero in these partial derivatives and $\nabla_{\boldsymbol{E}_d^t}\psi(\Xi_{0^t}) = 0$, so essentially $\frac{\partial}{\partial \boldsymbol{E}}\nabla_{\boldsymbol{E}_d^t}\psi(\Xi_{0^t}) = 0$ and $\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{E}_d^t}\psi(\Xi_{0^t}) = 0$. Using a similar reasoning and that mixed derivatives commute, it follows

$$\frac{\partial^2 \psi}{\partial \boldsymbol{E}_d^t \partial \boldsymbol{E}}(\bar{\Xi}, \cdot) = \frac{\partial^2 \psi}{\partial \boldsymbol{E} \partial \boldsymbol{E}_d^t}(\bar{\Xi}, \cdot) = 0\,, \qquad \qquad \frac{\partial^2 \psi}{\partial \theta_d^t \partial \boldsymbol{E}}(\bar{\Xi}, \cdot) = \frac{\partial^2 \psi}{\partial \boldsymbol{E} \partial \theta_d^t}(\bar{\Xi}, \cdot) = 0\,, \tag{C.26}$$

since $\frac{\partial \psi}{\partial \boldsymbol{E}_d^t}(\Xi_{0^t}, \cdot) = 0$ and $\frac{\partial \psi}{\partial \theta_d^t}(\Xi_{0^t}, \cdot) = 0$. This implies

$$\begin{aligned} \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}, \cdot) &= -\rho \int_0^\infty \frac{\partial^2 \psi}{\partial \boldsymbol{E}_d^t \partial \boldsymbol{E}_d^t}(\bar{\Xi}, \tau, \cdot) w_{L^2}(\tau)\,\mathrm{d}\tau\,, \\ \frac{\partial \boldsymbol{S}}{\partial \theta_d^t}(\bar{\Xi}, \cdot) &= -\rho \int_0^\infty \frac{\partial^2 \psi}{\partial \theta_d^t \partial \boldsymbol{E}_d^t}(\bar{\Xi}, \tau, \cdot) w_{L^2}(\tau)\,\mathrm{d}\tau\,. \end{aligned} \tag{C.27}$$

Define the following material properties,

$$\begin{aligned} &\mathsf{C}^{\infty}(\boldsymbol{X}) = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) = \rho \frac{\partial^2 \psi}{\partial \boldsymbol{E}^2}(\bar{\Xi})\,, \\ &\mathsf{C}^{⧖}(\boldsymbol{X}, s) = -\int_s^\infty \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}, \tau') w_{L^2}(\tau')\,\mathrm{d}\tau' = \rho \int_s^\infty \int_0^\infty \frac{\partial^2 \psi}{\partial \boldsymbol{E}_d^t \partial \boldsymbol{E}_d^t}(\bar{\Xi}, \tau, \tau') w_{L^2}(\tau)\,\mathrm{d}\tau\, w_{L^2}(\tau')\,\mathrm{d}\tau'\,, \end{aligned} \tag{C.28}$$

where $s \geq 0$. It follows,

$$\lim_{s \to \infty} \mathsf{C}^{⧖}(\boldsymbol{X}, s) = 0\,, \qquad \dot{\mathsf{C}}^{⧖}(\boldsymbol{X}, s) = \frac{\partial \mathsf{C}^{⧖}}{\partial s}(\boldsymbol{X}, s) = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}, s) w_{L^2}(s)\,, \tag{C.29}$$

for all $s > 0$. Moreover, both C^{∞} and $\mathsf{C}^{⧖}$ have major and minor symmetries because they come from second derivatives of the same variable, so that $\mathsf{C}^{\infty}_{ijkl} = \mathsf{C}^{\infty}_{klij} = \mathsf{C}^{\infty}_{jikl}$ and $\mathsf{C}^{⧖}_{ijkl} = \mathsf{C}^{⧖}_{klij} = \mathsf{C}^{⧖}_{jikl}$

for all $\boldsymbol{X} \in \Omega$, $s \geq 0$ and $i, j, k, l = 1, 2, 3$. Therefore,

$$\begin{aligned}
\Big(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}), \boldsymbol{E}_d^t\Big)_{L_w^2} &= \int_0^\infty \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}, s) : \boldsymbol{E}_d^t(\boldsymbol{X}, s) w_{L^2}(s)\,\mathrm{d}s \\
&= \int_0^\infty \dot{\mathsf{C}}^{⧖}(\boldsymbol{X}, s) : (\boldsymbol{E}^t(\boldsymbol{X}, s) - \boldsymbol{E}(\boldsymbol{X}, t))\,\mathrm{d}s \\
&= -\int_0^\infty \dot{\mathsf{C}}^{⧖}(\boldsymbol{X}, s)\,\mathrm{d}s : \boldsymbol{E}(\boldsymbol{X}, t) + \int_0^\infty \dot{\mathsf{C}}^{⧖}(\boldsymbol{X}, s) : \boldsymbol{E}^t(\boldsymbol{X}, s)\,\mathrm{d}s \\
&= \mathsf{C}^{⧖}(\boldsymbol{X}, 0) : \boldsymbol{E}(\boldsymbol{X}, t) + \int_0^\infty \dot{\mathsf{C}}^{⧖}(\boldsymbol{X}, s) : \boldsymbol{E}(\boldsymbol{X}, t - s)\,\mathrm{d}s\,.
\end{aligned} \tag{C.30}$$

At this point, it is helpful to make a brief pause to dwell with a different subject in order to make a connection with the preceding expression. Indeed, the Stieltjes convolution of $g : [0, \infty) \to \mathbb{R}$ and $h : (-\infty, \infty) \to \mathbb{R}$, denoted by $g * \mathrm{d}\, h$, is defined as [142]

$$g * \mathrm{d}\, h(t) = \int_{-\infty}^t g(t - s) \frac{\mathrm{d}h}{\mathrm{d}t}(s)\,\mathrm{d}s\,, \tag{C.31}$$

for $t \in (-\infty, \infty)$. Provided $\lim_{s \to \infty} g(s) h(t - s) = 0$, some simple manipulations yield,

$$\begin{aligned}
g * \mathrm{d}\, h(t) &= \int_{-\infty}^t g(t - s) \frac{\mathrm{d}h}{\mathrm{d}t}(s)\,\mathrm{d}s = \int_0^\infty g(s) \frac{\mathrm{d}h}{\mathrm{d}t}(t - s)\,\mathrm{d}s \\
&= -g(s) h(t - s)\Big|_{s=0}^{s=\infty} + \int_0^\infty \frac{\mathrm{d}g}{\mathrm{d}t}(s) h(t - s)\,\mathrm{d}s = g(0) h(t) + \int_0^\infty \frac{\mathrm{d}g}{\mathrm{d}t}(s) h(t - s)\,\mathrm{d}s\,.
\end{aligned} \tag{C.32}$$

Additionally, if g is interpreted as having the domain $(-\infty, \infty)$ (as opposed to $[0, \infty)$) and having support in $[0, \infty)$, then its distributional derivative in $(-\infty, \infty)$ is $\frac{\mathrm{d}g}{\mathrm{d}t}(t) = g(0)\delta_0(t) + \frac{\mathrm{d}g^+}{\mathrm{d}t}(t) H_0(t)$, where $g^+ = g|_{(0,\infty)}$, $\delta_0(t)$ is the Dirac delta distribution centered at $t = 0$, and $H_0(t)$ is the Heaviside step function centered at $t = 0$. In this case, $\frac{\mathrm{d}g}{\mathrm{d}t}(t) = \frac{\mathrm{d}g^+}{\mathrm{d}t}(t)$ for $t > 0$, $g(t) = \frac{\mathrm{d}g}{\mathrm{d}t}(t) = 0$ for $t < 0$, and the following relation involving the regular convolution holds

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}t} * h(t) &= \int_{-\infty}^\infty \frac{\mathrm{d}g}{\mathrm{d}t}(s) h(t - s)\,\mathrm{d}s = g(0) h(t) + \int_0^\infty \frac{\mathrm{d}g}{\mathrm{d}t}(s) h(t - s)\,\mathrm{d}s \\
&= g * \mathrm{d}\, h(t) + \lim_{s \to \infty} g(s) h(t - s)\,,
\end{aligned} \tag{C.33}$$

for $t \in (-\infty, \infty)$, where the last equality follows from (C.32). This observation is useful when dealing with Fourier transforms as it involves a convolution in the full real line, $\mathbb{R}$, as opposed to a subset of $\mathbb{R}$. Moreover, if h has support in $[0, \infty)$, then as above, $\frac{\mathrm{d}h}{\mathrm{d}t}(t) = h(0)\delta_0(t) + \frac{\mathrm{d}h^+}{\mathrm{d}t}(t) H_0(t)$

and obviously $\lim_{s\to\infty} g(s)h(t-s) = 0$. It follows that if h has support in $[0,\infty)$, the Stieltjes convolution commutes,

$$\begin{aligned} g \ast\mathrm{d}\, h(t) &= \int_{-\infty}^{t} g(t-s)\frac{\mathrm{d}h}{\mathrm{d}t}(s)\,\mathrm{d}s = g(t)h(0) + \int_{0}^{t} g(t-s)\frac{\mathrm{d}h}{\mathrm{d}t}(s)\,\mathrm{d}s \\ &= g(0)h(t) + \int_{0}^{t} \frac{\mathrm{d}g}{\mathrm{d}t}(s)h(t-s)\,\mathrm{d}s = \int_{-\infty}^{t} h(t-s)\frac{\mathrm{d}g}{\mathrm{d}t}(s)\,\mathrm{d}s = h \ast\mathrm{d}\, g(t)\,. \end{aligned} \tag{C.34}$$

Now returning to (C.30), it is easy to see the remarkable similarities with the regular convolution as written in (C.33) and the Stieltjes convolution as written in (C.32), with $\mathsf{C}^{\text{⧖}}(\boldsymbol{X},\cdot)$ playing the role of g (which does have the domain $[0,\infty)$ and satisfies $\lim_{s\to\infty}\mathsf{C}^{\text{⧖}}(\boldsymbol{X},s) = 0$) and $\boldsymbol{E}(\boldsymbol{X},\cdot)$ playing the role of h (which, consistent with the definition of the history, does have the domain $(-\infty,\infty)$). Indeed, dropping the $\boldsymbol{X}$, it follows,

$$\Big(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}), \boldsymbol{E}_d^t\Big)_{L_w^2} = \mathsf{C}^{\text{⧖}}(0) : \boldsymbol{E}(t) + \int_0^\infty \dot{\mathsf{C}}^{\text{⧖}}(s) : \boldsymbol{E}(t-s)\,\mathrm{d}s = \dot{\mathsf{C}}^{\text{⧖}} \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}} \boldsymbol{E}(t) = \mathsf{C}^{\text{⧖}} \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}}\mathrm{d}\, \boldsymbol{E}(t)\,, \tag{C.35}$$

where the usual convolution and Stieltjes convolution involving a double dot product are denoted by $\overset{\cdot}{\underset{\cdot}{\ast}}$ and $\overset{\cdot}{\underset{\cdot}{\ast}}\mathrm{d}$ respectively. It is important to note that the expression $\dot{\mathsf{C}}^{\text{⧖}} \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}} \boldsymbol{E}$ assumes $\mathsf{C}^{\text{⧖}}$ is the zero extension of the original function to the whole domain $(-\infty,\infty)$, so that $\mathsf{C}^{\text{⧖}}(\boldsymbol{X},s) = 0$ if $s < 0$ and is given by (C.28) if $s \geq 0$. Thus, dropping the $\boldsymbol{X}$, the distributional derivative is $\dot{\mathsf{C}}^{\text{⧖}}(s) = \mathsf{C}^{\text{⧖}}(0)\delta_0(s) + \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi},s)w_{L^2}(s)H_0(s)$. All other expressions in (C.35), including $\mathsf{C}^{\text{⧖}} \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}}\mathrm{d}\, \boldsymbol{E}$, only require that $\mathsf{C}^{\text{⧖}}(\boldsymbol{X},s)$ be defined for $s \geq 0$. Next, recall that C^∞ is constant for $s \geq 0$ so that $\dot{\mathsf{C}}^\infty = 0$ for $s > 0$, and

$$\mathsf{C}^\infty : \boldsymbol{E}(t) = \mathsf{C}^\infty : \boldsymbol{E}(t) + \int_0^\infty \dot{\mathsf{C}}^\infty : \boldsymbol{E}(t-s)\,\mathrm{d}s = \dot{\mathsf{C}}^\infty \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}} \boldsymbol{E}(t) = \mathsf{C}^\infty \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}}\mathrm{d}\, \boldsymbol{E}(t) + \mathsf{C}^\infty : \lim_{s\to\infty} \boldsymbol{E}(t-s)\,, \tag{C.36}$$

where the expression $\dot{\mathsf{C}}^\infty \mathbin{\overset{\cdot}{\underset{\cdot}{\ast}}} \boldsymbol{E}$ implicitly assumes C^∞ is the zero extension to $(-\infty,\infty)$ of the definition in (C.28), so that $\mathsf{C}^\infty(\boldsymbol{X},s) = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi})H_0(s)$ and $\dot{\mathsf{C}}^\infty(\boldsymbol{X},s) = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi})\delta_0(s)$. Therefore, for every $\boldsymbol{X} \in \Omega$ and $s \in (-\infty,\infty)$, define the full stiffness tensor as

$$\mathsf{C} = \mathsf{C}^\infty + \mathsf{C}^{\text{⧖}}\,, \tag{C.37}$$

where C^∞ and $\mathsf{C}^{\text{⧖}}$ are the zero extensions of the functions defined in (C.28) which have support in $[0,\infty)$. Here, C^∞ represents the equilibrium value, while $\mathsf{C}^{\text{⧖}}$ represents the transient effects (thus the use of the hourglass, ⧖) due to the viscosity produced by the memory of the material. Naturally, C inherits the minor and major symmetries from C^∞ and $\mathsf{C}^{\text{⧖}}$, so that $\mathsf{C}_{ijkl} = \mathsf{C}_{klij} = \mathsf{C}_{jikl}$ for all $i, j, k, l = 1, 2, 3$. Clearly, $\mathsf{C}(\boldsymbol{X}, s) = 0$ and $\dot{\mathsf{C}}(\boldsymbol{X}, s) = 0$ for all $s < 0$, and using (C.29), it follows

$$\lim_{s\to\infty} \mathsf{C}(\boldsymbol{X}, s) = \mathsf{C}^\infty(\boldsymbol{X})|_{[0,\infty)}\,, \qquad \dot{\mathsf{C}}(\boldsymbol{X}, s) = \dot{\mathsf{C}}^{\text{⧖}}(\boldsymbol{X}, s) \quad \forall s > 0\,. \tag{C.38}$$

Here, $\mathsf{C}^\infty(\boldsymbol{X})|_{[0,\infty)} = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi})$, which is clearly independent of the time $s \geq 0$, as defined in (C.28). The full distributional time derivative is $\dot{\mathsf{C}}(s) = (\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) + \mathsf{C}^{\text{⧖}}(0))\delta_0(s) + \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}, s) w_{L^2}(s) H_0(s)$. Finally, letting $\boldsymbol{E}^{-\infty}(\boldsymbol{X}) = \lim_{s\to-\infty} \boldsymbol{E}(\boldsymbol{X}, s)$ (which is independent of t), and combining the expressions in (C.35) and (C.36), yields for all $\boldsymbol{X} \in \Omega$ and $t \in (-\infty, \infty)$,

$$\begin{aligned} \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) : \boldsymbol{E}(t) + \Big(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}_d^t}(\bar{\Xi}), \boldsymbol{E}_d^t\Big)_{L_w^2} &= \mathsf{C}(0) : \boldsymbol{E}(t) + \int_0^\infty \dot{\mathsf{C}}(s) : \boldsymbol{E}(t-s)\,\mathrm{d}s \\ &= \dot{\mathsf{C}} \ast \boldsymbol{E}(t) = \mathsf{C} \ast\mathrm{d}\, \boldsymbol{E}(t) + \mathsf{C}^\infty|_{[0,\infty)} : \boldsymbol{E}^{-\infty}\,. \end{aligned} \tag{C.39}$$

Note in the preceding expression that the only term requiring C to be defined for negative times, $s < 0$, is $\dot{\mathsf{C}} \ast \boldsymbol{E}$. Moreover, if $\boldsymbol{E}$ is assumed to vanish for all $t < 0$, which is often the case, then $\boldsymbol{E}^{-\infty} = 0$ and the last term drops, so that $\dot{\mathsf{C}} \ast \boldsymbol{E} = \mathsf{C} \ast\mathrm{d}\, \boldsymbol{E} = \boldsymbol{E} \ast\mathrm{d}\, \mathsf{C}$ as in (C.34).

Proceeding in an entirely analogous fashion, define the following material properties,

$$\begin{aligned} \mathsf{M}^\infty(\boldsymbol{X}) &= \frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi}) = \rho \frac{\partial^2 \psi}{\partial \theta \partial \boldsymbol{E}}(\bar{\Xi})\,, \\ \mathsf{M}^{\text{⧖}\theta}(\boldsymbol{X}, s) &= -\int_s^\infty \frac{\partial \boldsymbol{S}}{\partial \theta_d^t}(\bar{\Xi}, \tau') w_{L^2}(\tau')\,\mathrm{d}\tau' = \rho \int_s^\infty \int_0^\infty \frac{\partial^2 \psi}{\partial \theta_d^t \partial \boldsymbol{E}_d^t}(\bar{\Xi}, \tau, \tau') w_{L^2}(\tau)\,\mathrm{d}\tau\, w_{L^2}(\tau')\,\mathrm{d}\tau'\,, \end{aligned} \tag{C.40}$$

where $s \geq 0$. Additionally, for all $\boldsymbol{X} \in \Omega$ let $\mathsf{M}^\infty = \mathsf{M}^{\text{⧖}\theta} = 0$ whenever $s < 0$. Moreover, define for all $\boldsymbol{X} \in \Omega$ and $s \in (-\infty, \infty)$,

$$\mathsf{M}^\theta = \mathsf{M}^\infty + \mathsf{M}^{\text{⧖}\theta}\,. \tag{C.41}$$

Using the definitions, it follows $\mathsf{M}^\theta(\boldsymbol{X}, s) = 0$ and $\dot{\mathsf{M}}^\theta(\boldsymbol{X}, s) = 0$ for all $s < 0$, and

$$\lim_{s\to\infty} \mathsf{M}^\theta(\boldsymbol{X}, s) = \mathsf{M}^\infty(\boldsymbol{X})|_{[0,\infty)}\,, \qquad \dot{\mathsf{M}}^\theta(\boldsymbol{X}, s) = \dot{\mathsf{M}}^{\text{⧖}\theta}(\boldsymbol{X}, s) = \frac{\partial \boldsymbol{S}}{\partial \theta_d^t}(\bar{\Xi}, s) w_{L^2}(s)\,, \tag{C.42}$$

for all $s > 0$. As before, $\dot{\mathsf{M}}^\theta(s) = (\frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi}) + \mathsf{M}^{\text{⧖}\theta}(0))\delta_0(s) + \frac{\partial \boldsymbol{S}}{\partial \theta_d^t}(\bar{\Xi}, s) w_{L^2}(s) H_0(s)$ is the full distributional time derivative. Moreover, from the definitions, M^∞ and $\mathsf{M}^{\text{⧖}\theta}$ are seen to be symmetric, so that M^θ inherits that symmetry. That is, $\mathsf{M}^\infty_{ij} = \mathsf{M}^\infty_{ji}$, $\mathsf{M}^{⧖_\theta}_{ij} = \mathsf{M}^{⧖_\theta}_{ji}$ and $\mathsf{M}^\theta_{ij} = \mathsf{M}^\theta_{ji}$ for all $i, j = 1, 2, 3$. Proceeding as with (C.30) and (C.35), this gives

$$\Big(\frac{\partial \boldsymbol{S}}{\partial \theta^t_d}(\bar{\Xi}), \theta^t_d\Big)_{L^2_w} = \mathsf{M}^{⧖_\theta}(0)\,\theta(t) + \int_0^\infty \dot{\mathsf{M}}^{⧖_\theta}(s)\,\theta(t-s)\,\mathrm{d}s = \dot{\mathsf{M}}^{⧖_\theta} \ast \theta(t) = \mathsf{M}^{⧖_\theta} \ast\mathrm{d}\,\theta(t)\,. \tag{C.43}$$

Moreover,

$$\begin{aligned} \dot{\mathsf{M}}^{⧖_\theta} \ast \bar{\theta}(t) &= \int_{-\infty}^\infty \dot{\mathsf{M}}^{⧖_\theta}(s)\,\bar{\theta}\,\mathrm{d}s = \mathsf{M}^{⧖_\theta}(0)\,\bar{\theta} + \int_0^\infty \mathsf{M}^{⧖_\theta}(s)\,\bar{\theta}\,\mathrm{d}s \\ &= \mathsf{M}^{⧖_\theta}(0)\,\bar{\theta} + \lim_{s\to\infty} \mathsf{M}^{⧖_\theta}(s)\,\bar{\theta} - \mathsf{M}^{⧖_\theta}(0)\,\bar{\theta} = 0\,. \end{aligned} \tag{C.44}$$

Therefore, (C.43) becomes

$$\begin{aligned} \Big(\frac{\partial \boldsymbol{S}}{\partial \theta^t_d}(\bar{\Xi}), \theta^t_d\Big)_{L^2_w} &= \mathsf{M}^{⧖_\theta}(0)\,(\theta(t) - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^{⧖_\theta}(s)\,(\theta(t-s) - \bar{\theta})\,\mathrm{d}s \\ &= \dot{\mathsf{M}}^{⧖_\theta} \ast (\theta - \bar{\theta})(t) = \mathsf{M}^{⧖_\theta} \ast\mathrm{d}\,(\theta - \bar{\theta})(t)\,. \end{aligned} \tag{C.45}$$

One can then obtain an expression like (C.36) for $\mathsf{M}^\infty\,(\theta - \bar{\theta})$, which in turn leads to the analogous expression to (C.39),

$$\begin{aligned} \frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi})\,(\theta(t) - \bar{\theta}) + \Big(\frac{\partial \boldsymbol{S}}{\partial \theta^t_d}(\bar{\Xi}), \theta^t_d\Big)_{L^2_w} &= \mathsf{M}^\theta(0)\,(\theta(t) - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^\theta(s)\,(\theta(t-s) - \bar{\theta})\,\mathrm{d}s \\ &= \dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta})(t) \\ &= \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta})(t) + \mathsf{M}^\infty|_{[0,\infty)}\,(\theta^{-\infty} - \bar{\theta})\,, \end{aligned} \tag{C.46}$$

where $\theta(\boldsymbol{X})^{-\infty} = \lim_{s\to-\infty}\theta(s)$. As before, the only term requiring M^θ to be defined for negative times, $s < 0$, is $\dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta})$. Moreover, if $\theta = \bar{\theta}$ for all $t < 0$, which is often the case, then $\theta^{-\infty} - \bar{\theta} = 0$ and the last term drops, so that $\dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta}) = \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta}) = (\theta - \bar{\theta}) \ast\mathrm{d}\,\mathsf{M}^\theta$ as in (C.34).

Now, it is possible to return to the original linearized expression in (C.24), and use the results in (C.39) and (C.46), so that

$$\begin{aligned} \boldsymbol{S}(\Xi) &= \boldsymbol{S}(\bar{\Xi}) + \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}(\bar{\Xi}) : \boldsymbol{E} + \Big(\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}^t_d}(\bar{\Xi}), \boldsymbol{E}^t_d\Big)_{L^2_w} + \frac{\partial \boldsymbol{S}}{\partial \theta}(\bar{\Xi})(\theta - \bar{\theta}) + \Big(\frac{\partial \boldsymbol{S}}{\partial \theta^t_d}(\bar{\Xi}), \theta^t_d\Big)_{L^2_w} + o(\delta) \\ &= \overline{\boldsymbol{S}} + \mathsf{C}(0) : \boldsymbol{E}(\cdot) + \int_0^\infty \dot{\mathsf{C}}(s) : \boldsymbol{E}(\cdot - s)\,\mathrm{d}s + \mathsf{M}^\theta(0)(\theta(\cdot) - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^\theta(s)(\theta(\cdot - s) - \bar{\theta})\,\mathrm{d}s + o(\delta) \\ &= \overline{\boldsymbol{S}} + \dot{\mathsf{C}} \mathbin{\underset{\cdot}{\dot{\ast}}} \boldsymbol{E} + \dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta}) + o(\delta) \\ &= \overline{\boldsymbol{S}} + \mathsf{C} \mathbin{\underset{\cdot}{\dot{\ast}}}\mathrm{d}\,\boldsymbol{E} + \mathsf{C}^\infty|_{[0,\infty)} : \boldsymbol{E}^{-\infty} + \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta}) + \mathsf{M}^\infty|_{[0,\infty)}\,(\theta^{-\infty} - \bar{\theta}) + o(\delta)\,, \end{aligned} \tag{C.47}$$

where $\overline{\boldsymbol{S}} = \boldsymbol{S}(\bar{\Xi})$. Next, notice the equation is linear in $\boldsymbol{E}$, θ and their histories. However, it is desired that the equation is linear in $\boldsymbol{u}$, θ and their histories. For this, invoke the engineering strain in (C.20), and note that $\mathcal{O}(\delta^2) = \mathcal{o}(\delta)$, so that

$$\begin{aligned}
\boldsymbol{S} &= \overline{\boldsymbol{S}} + \mathsf{C}(0){:}\,\boldsymbol{\varepsilon}(\cdot) + \int_0^\infty \dot{\mathsf{C}}(s){:}\,\boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s + \mathsf{M}^\theta(0)(\theta(\cdot) - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^\theta(s)(\theta(\cdot - s) - \bar{\theta})\,\mathrm{d}s + \mathcal{o}(\delta) \\
&= \overline{\boldsymbol{S}} + \dot{\mathsf{C}} \mathbin{\ast\!\!:} \boldsymbol{\varepsilon} + \dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta}) + \mathcal{o}(\delta) \qquad\qquad \text{(C.48)} \\
&= \overline{\boldsymbol{S}} + \mathsf{C} \mathbin{\ast\!\!:}\mathrm{d}\,\boldsymbol{\varepsilon} + \mathsf{C}^\infty|_{[0,\infty)} : \boldsymbol{\varepsilon}^{-\infty} + \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta}) + \mathsf{M}^\infty|_{[0,\infty)}\,(\theta^{-\infty} - \bar{\theta}) + \mathcal{o}(\delta)\,,
\end{aligned}$$

with $\boldsymbol{\varepsilon}^{-\infty} = \lim_{s\to-\infty} \boldsymbol{\varepsilon}(s)$. Clearly, (C.48) has the form $\boldsymbol{S} = \overline{\boldsymbol{S}} + \mathcal{O}(\delta)$, while it is already known that $\boldsymbol{F} = \mathbf{I} + \boldsymbol{\nabla}\boldsymbol{u} = \mathbf{I} + \mathcal{O}(\delta)$. Using this, it follows that the first Piola-Kirchhoff stress tensor is,

$$\begin{aligned}
\boldsymbol{P} = \boldsymbol{F}\boldsymbol{S} &= \overline{\boldsymbol{S}} + \boldsymbol{\nabla}\boldsymbol{u}\,\overline{\boldsymbol{S}} + \mathsf{C}(0) : \boldsymbol{\varepsilon}(\cdot) + \int_0^\infty \dot{\mathsf{C}}(s) : \boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s \\
&\qquad + \mathsf{M}^\theta(0)\,(\theta(\cdot) - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^\theta(s)\,(\theta(\cdot - s) - \bar{\theta})\,\mathrm{d}s + \mathcal{o}(\delta) \\
&= \overline{\boldsymbol{S}} + \boldsymbol{\nabla}\boldsymbol{u}\,\overline{\boldsymbol{S}} + \dot{\mathsf{C}} \mathbin{\ast\!\!:} \boldsymbol{\varepsilon} + \dot{\mathsf{M}}^\theta \ast (\theta - \bar{\theta}) + \mathcal{o}(\delta) \qquad\qquad \text{(C.49)} \\
&= \overline{\boldsymbol{S}} + \boldsymbol{\nabla}\boldsymbol{u}\,\overline{\boldsymbol{S}} + \mathsf{C} \mathbin{\ast\!\!:}\mathrm{d}\,\boldsymbol{\varepsilon} + \mathsf{C}^\infty|_{[0,\infty)} : \boldsymbol{\varepsilon}^{-\infty} \\
&\qquad + \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta}) + \mathsf{M}^\infty|_{[0,\infty)}\,(\theta^{-\infty} - \bar{\theta}) + \mathcal{o}(\delta)\,.
\end{aligned}$$

Notice that if $\overline{\boldsymbol{S}} = 0$ (or if $\boldsymbol{\nabla}\boldsymbol{u}\,\overline{\boldsymbol{S}} = \overline{\boldsymbol{S}}\,\boldsymbol{\nabla}\boldsymbol{u}^{\mathsf{T}}$), then $\boldsymbol{P}$ is symmetric up to a "small" error. This concludes the linearization of the equation for conservation of momentum, which is linear in $\boldsymbol{P}$ and $\boldsymbol{u}$ (see (C.21)), with $\boldsymbol{P}$ being linear in $\boldsymbol{u}$ and θ up to a "small" error.

At this point, one can shift the attention to the equation for conservation of energy as written in (C.23). From (C.16), assumptions of Fréchet differentiability, and the small variations about $\bar{\Xi}$, it follows that

$$\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\Xi) = \frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}) + \mathcal{O}(\delta) = \mathcal{O}(\delta)\,, \qquad \frac{\partial\psi}{\partial\theta_d^t}(\Xi) = \frac{\partial\psi}{\partial\theta_d^t}(\bar{\Xi}) + \mathcal{O}(\delta) = \mathcal{O}(\delta)\,. \qquad \text{(C.50)}$$

Using this and that the histories $\dot{\boldsymbol{E}}^t$ and $\dot{\theta}^t$ are small as in (C.19), it follows

$$\begin{aligned}
\Big|\Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\Xi), \dot{\boldsymbol{E}}^t\Big)_{L_w^2}\Big| &\leq \Big\|\frac{\partial\psi}{\partial\boldsymbol{E}_d^t}(\Xi)\Big\|_{\mathbf{L}_w^2(0,\infty;\mathbb{S})} \|\dot{\boldsymbol{E}}^t\|_{\mathbf{L}_w^2(0,\infty;\mathbb{S})} = \mathcal{O}(\delta^2)\,, \\
\Big|\Big(\frac{\partial\psi}{\partial\theta_d^t}(\Xi), \dot{\theta}^t\Big)_{L_w^2}\Big| &\leq \Big\|\frac{\partial\psi}{\partial\theta_d^t}(\Xi)\Big\|_{L_w^2(0,\infty)} \|\dot{\theta}^t\|_{L_w^2(0,\infty)} = \mathcal{O}(\delta^2)\,.
\end{aligned} \qquad \text{(C.51)}$$

Next, as in (C.14), note that $\dot{\eta}$ is

$$\dot{\eta}(\Xi) = \frac{\partial \eta}{\partial \boldsymbol{E}}(\Xi) : \dot{\boldsymbol{E}} + \Big(\frac{\partial \eta}{\partial \boldsymbol{E}^t_d}(\Xi), \dot{\boldsymbol{E}}^t_d\Big)_{L^2_w} + \frac{\partial \eta}{\partial \theta}(\Xi)\dot{\theta} + \Big(\frac{\partial \eta}{\partial \theta^t_d}(\Xi), \dot{\theta}^t_d\Big)_{L^2_w} . \tag{C.52}$$

Assumptions of Fréchet differentiability and small variations about $\bar{\Xi}$ yield

$$\begin{aligned} \theta\frac{\partial \eta}{\partial \boldsymbol{E}}(\Xi) &= \bar{\theta}\frac{\partial \eta}{\partial \boldsymbol{E}}(\bar{\Xi}) + \mathcal{O}(\delta)\,, \quad & \theta\frac{\partial \eta}{\partial \boldsymbol{E}^t_d}(\Xi,\cdot) &= \bar{\theta}\frac{\partial \eta}{\partial \boldsymbol{E}^t_d}(\bar{\Xi},\cdot) + \mathcal{O}(\delta)\,, \\ \theta\frac{\partial \eta}{\partial \theta}(\Xi) &= \bar{\theta}\frac{\partial \eta}{\partial \theta}(\bar{\Xi}) + \mathcal{O}(\delta)\,, \quad & \theta\frac{\partial \eta}{\partial \theta^t_d}(\Xi,\cdot) &= \bar{\theta}\frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi},\cdot) + \mathcal{O}(\delta)\,. \end{aligned} \tag{C.53}$$

Using this and that $\dot{\boldsymbol{E}}$, $\dot{\theta}$ and their difference histories are all small as in (C.19), it follows,

$$\rho\dot{\eta}\theta = \rho\bar{\theta}\frac{\partial \eta}{\partial \theta}(\bar{\Xi})\dot{\theta} + \rho\bar{\theta}\Big(\frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi}), \dot{\theta}^t_d\Big)_{L^2_w} + \rho\bar{\theta}\frac{\partial \eta}{\partial \boldsymbol{E}}(\bar{\Xi}) : \dot{\boldsymbol{E}} + \rho\bar{\theta}\Big(\frac{\partial \eta}{\partial \boldsymbol{E}^t_d}(\bar{\Xi}), \dot{\boldsymbol{E}}^t_d\Big)_{L^2_w} + \mathcal{O}(\delta^2)\,. \tag{C.54}$$

Recalling from (C.15) that $\eta = -(\frac{\partial \psi}{\partial \theta} - \nabla_{\theta^t_d}\psi)$, using the results in (C.16), and proceeding as in (C.25), (C.26) and (C.27), it follows that

$$\begin{aligned} \frac{\partial \eta}{\partial \theta}(\bar{\Xi}) &= -\frac{\partial^2 \psi}{\partial \theta^2}(\bar{\Xi})\,, \qquad \frac{\partial \eta}{\partial \boldsymbol{E}}(\bar{\Xi}) = -\frac{\partial^2 \psi}{\partial \boldsymbol{E}\partial \theta}(\bar{\Xi}) = -\frac{\partial^2 \psi}{\partial \theta \partial \boldsymbol{E}}(\bar{\Xi})\,, \\ \frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi},\cdot) &= \int_0^\infty \frac{\partial^2 \psi}{\partial \theta^t_d \partial \theta^t_d}(\bar{\Xi},\tau,\cdot) w_{L^2}(\tau)\,\mathrm{d}\tau\,, \\ \frac{\partial \eta}{\partial \boldsymbol{E}^t_d}(\bar{\Xi},\cdot) &= \int_0^\infty \frac{\partial^2 \psi}{\partial \boldsymbol{E}^t_d \partial \theta^t_d}(\bar{\Xi},\tau,\cdot) w_{L^2}(\tau)\,\mathrm{d}\tau = \int_0^\infty \frac{\partial^2 \psi}{\partial \theta^t_d \partial \boldsymbol{E}^t_d}(\bar{\Xi},\cdot,\tau) w_{L^2}(\tau)\,\mathrm{d}\tau\,. \end{aligned} \tag{C.55}$$

Before proceeding further, recall that $e = \psi + \eta\theta$. Using that $\eta = -(\frac{\partial \psi}{\partial \theta} - \nabla_{\theta^t_d}\psi)$ and the results in (C.16), it follows

$$\begin{aligned} \frac{\partial e}{\partial \theta}(\bar{\Xi}) &= \frac{\partial \psi}{\partial \theta}(\bar{\Xi}) + \eta(\bar{\Xi}) + \bar{\theta}\frac{\partial \eta}{\partial \theta}(\bar{\Xi}) = \bar{\theta}\frac{\partial \eta}{\partial \theta}(\bar{\Xi})\,, \\ \frac{\partial e}{\partial \theta^t_d}(\bar{\Xi},\cdot) &= \frac{\partial \psi}{\partial \theta^t_d}(\bar{\Xi},\cdot) + \bar{\theta}\frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi},\cdot) = \bar{\theta}\frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi},\cdot)\,. \end{aligned} \tag{C.56}$$

As before, define the following material properties,

$$\begin{aligned} c_v^{\infty}(\boldsymbol{X}) &= \frac{\partial e}{\partial \theta}(\bar{\Xi}) = \bar{\theta}\,\frac{\partial \eta}{\partial \theta}(\bar{\Xi}) = -\bar{\theta}\,\frac{\partial^2 \psi}{\partial \theta^2}(\bar{\Xi})\,, \\ c_v^{\text{⧗}}(\boldsymbol{X},s) &= -\int_s^\infty \frac{\partial e}{\partial \theta^t_d}(\bar{\Xi},\tau') w_{L^2}(\tau')\,\mathrm{d}\tau' = -\bar{\theta}\int_s^\infty \frac{\partial \eta}{\partial \theta^t_d}(\bar{\Xi},\tau') w_{L^2}(\tau')\,\mathrm{d}\tau' \\ &= -\bar{\theta}\int_s^\infty\!\!\int_0^\infty \frac{\partial^2 \psi}{\partial \theta^t_d \partial \theta^t_d}(\bar{\Xi},\tau,\tau') w_{L^2}(\tau)\,\mathrm{d}\tau\, w_{L^2}(\tau')\,\mathrm{d}\tau'\,, \end{aligned} \tag{C.57}$$

$$\begin{aligned}
\mathsf{M}^{\infty}(\boldsymbol{X}) &= -\rho\frac{\partial\eta}{\partial\boldsymbol{E}}(\bar{\Xi}) = \rho\frac{\partial^2\psi}{\partial\boldsymbol{E}\partial\theta}(\bar{\Xi}) = \rho\frac{\partial^2\psi}{\partial\theta\partial\boldsymbol{E}}(\bar{\Xi}) = \frac{\partial\boldsymbol{S}}{\partial\theta}(\bar{\Xi})\,,\\
\mathsf{M}^{\text{⧖}_E}(\boldsymbol{X},s) &= \rho\int_s^\infty \frac{\partial\eta}{\partial\boldsymbol{E}_d^t}(\bar{\Xi},\tau')w_{L^2}(\tau')\,\mathrm{d}\tau' = \rho\int_s^\infty\int_0^\infty \frac{\partial^2\psi}{\partial\boldsymbol{E}_d^t\partial\theta_d^t}(\bar{\Xi},\tau,\tau')w_{L^2}(\tau)\,\mathrm{d}\tau\,w_{L^2}(\tau')\,\mathrm{d}\tau'\,,\\
&= \rho\int_s^\infty\int_0^\infty \frac{\partial^2\psi}{\partial\theta_d^t\partial\boldsymbol{E}_d^t}(\bar{\Xi},\tau',\tau)w_{L^2}(\tau)\,\mathrm{d}\tau\,w_{L^2}(\tau')\,\mathrm{d}\tau'\,,
\end{aligned} \tag{C.58}$$

for all $s \geq 0$. Additionally, for all $\boldsymbol{X} \in \Omega$ let $c_v^\infty = c_v^{\text{⧖}} = 0$ and $\mathsf{M}^\infty = \mathsf{M}^{\text{⧖}_E} = 0$ whenever $s < 0$. Notice, the definition of M^∞ here is equal to that shown in (C.40), so it is valid to call it with the same name. Also, notice that $\mathsf{M}^{\text{⧖}_E}$ and $\mathsf{M}^{\text{⧖}_\theta}$ defined in (C.40) are extremely similar. Indeed, $\mathsf{M}^{\text{⧖}_E}(\boldsymbol{X},0) = \mathsf{M}^{\text{⧖}_\theta}(\boldsymbol{X},0)$, but not for all s, since they differ in the variable that is being integrated over. Next, define for all $\boldsymbol{X} \in \Omega$ and $s \in (-\infty,\infty)$,

$$c_v = c_v^\infty + c_v^{\text{⧖}}\,, \qquad\qquad \mathsf{M}^E = \mathsf{M}^\infty + \mathsf{M}^{\text{⧖}_E}\,. \tag{C.59}$$

Their full distributional time derivatives are $\dot{c}_v(s) = (\bar{\theta}\frac{\partial\eta}{\partial\theta}(\bar{\Xi}) + c_v^{\text{⧖}}(0))\delta_0(s) + \bar{\theta}\frac{\partial\eta}{\partial\theta_d^t}(\bar{\Xi},s)w_{L^2}(s)H_0(s)$ and $\dot{\mathsf{M}}^E(s) = (-\rho\frac{\partial\eta}{\partial\boldsymbol{E}}(\bar{\Xi}) + \mathsf{M}^{\text{⧖}_E}(0))\delta_0(s) - \rho\frac{\partial\eta}{\partial\boldsymbol{E}_d^t}(\bar{\Xi},s)w_{L^2}(s)H_0(s)$ respectively. Hence, it is clear that for all $\boldsymbol{X} \in \Omega$ and $s < 0$, $c_v = \dot{c}_v = 0$ and $\mathsf{M}^E = \dot{\mathsf{M}}^E = 0$. Meanwhile,

$$\begin{aligned}
&\lim_{s\to\infty} c_v(\boldsymbol{X},s) = c_v^\infty(\boldsymbol{X})|_{[0,\infty)}\,, &&\dot{c}_v(\boldsymbol{X},s) = \dot{c}_v^{\text{⧖}}(\boldsymbol{X},s) = \bar{\theta}\frac{\partial\eta}{\partial\theta_d^t}(\bar{\Xi},s)w_{L^2}(s)\,,\\
&\lim_{s\to\infty} \mathsf{M}^E(\boldsymbol{X},s) = \mathsf{M}^\infty(\boldsymbol{X})|_{[0,\infty)}\,, &&\dot{\mathsf{M}}^E(\boldsymbol{X},s) = \dot{\mathsf{M}}^{\text{⧖}_E}(\boldsymbol{X},s) = -\rho\frac{\partial\eta}{\partial\boldsymbol{E}_d^t}(\bar{\Xi},s)w_{L^2}(s)\,,
\end{aligned} \tag{C.60}$$

for all $s > 0$. Moreover, as with M^θ and M^∞, $\mathsf{M}^{\text{⧖}_E}$ and M^E are seen to be symmetric, so $\mathsf{M}_{ij}^{\text{⧖}_E} = \mathsf{M}_{ji}^{\text{⧖}_E}$ and $\mathsf{M}_{ij}^E = \mathsf{M}_{ji}^E$ for all $i,j = 1,2,3$.

With these definitions, invoking the engineering strain in (C.20), using (C.51), and proceeding in the same vein as before, it is possible to rewrite (C.54), so that the equation of energy in (C.23) becomes

$$\begin{aligned}
-\operatorname{div}\boldsymbol{q} + r &= \rho\dot{\eta}\theta + \Big(\frac{\partial\psi}{\partial\boldsymbol{E}_d^t},\dot{\boldsymbol{E}}^t\Big)_{L_w^2} + \Big(\frac{\partial\psi}{\partial\theta_d^t},\dot{\theta}^t\Big)_{L_w^2}\\
&= \rho\bar{\theta}\frac{\partial\eta}{\partial\theta}(\bar{\Xi})\dot{\theta} + \rho\bar{\theta}\Big(\frac{\partial\eta}{\partial\theta_d^t}(\bar{\Xi}),\dot{\theta}_d^t\Big)_{L_w^2} + \rho\bar{\theta}\frac{\partial\eta}{\partial\boldsymbol{E}}(\bar{\Xi}):\dot{\boldsymbol{E}} + \rho\bar{\theta}\Big(\frac{\partial\eta}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}),\dot{\boldsymbol{E}}_d^t\Big)_{L_w^2} + \mathcal{O}(\delta^2)\\
&= \rho c_v(0)\dot{\theta}(\cdot) + \int_0^\infty \rho\dot{c}_v(s)\dot{\theta}(\cdot - s)\,\mathrm{d}s - \bar{\theta}\mathsf{M}^E(0):\dot{\boldsymbol{\varepsilon}}(\cdot) - \int_0^\infty \bar{\theta}\dot{\mathsf{M}}^E(s):\dot{\boldsymbol{\varepsilon}}(\cdot - s)\,\mathrm{d}s + \mathcal{O}(\delta^2)\\
&= \rho\dot{c}_v * \dot{\theta} - \bar{\theta}\dot{\mathsf{M}}^E \mathbin{⁑} \dot{\boldsymbol{\varepsilon}} + \mathcal{O}(\delta^2)\\
&= \rho c_v *\mathrm{d}\,\dot{\theta} + \rho c_v^\infty|_{[0,\infty)}\,\dot{\theta}^{-\infty} - \bar{\theta}\mathsf{M}^E \mathbin{⁑}\mathrm{d}\,\dot{\boldsymbol{\varepsilon}} - \bar{\theta}\mathsf{M}^\infty|_{[0,\infty)}:\dot{\boldsymbol{\varepsilon}}^{-\infty} + \mathcal{O}(\delta^2)\,,
\end{aligned} \tag{C.61}$$

where $\dot{\theta}^{-\infty} = \lim_{s\to-\infty}\dot{\theta}(s)$ and $\dot{\boldsymbol{\varepsilon}}^{-\infty} = \lim_{s\to-\infty}\dot{\boldsymbol{\varepsilon}}(s)$. The only terms requiring c_v and M^E to be defined for negative times, $s<0$, are $\dot{c}_v{\ast}\dot{\theta}$ and $\dot{\mathsf{M}}^E \mathbin{\ast\!\!\!\vert}\, \dot{\boldsymbol{\varepsilon}}$ respectively. Moreover, if $\dot{\theta}=0$ and $\dot{\boldsymbol{\varepsilon}}=0$ for all $t<0$, which is often the case, then $\theta^{-\infty}=0$ and $\boldsymbol{\varepsilon}^{-\infty}=0$, which further implies $\dot{c}_v{\ast}\dot{\theta} = c_v \ast\mathrm{d}\,\dot{\theta}$ and $\dot{\mathsf{M}}^E \mathbin{\ast\!\!\!\vert}\, \dot{\boldsymbol{\varepsilon}} = \mathsf{M}^E \mathbin{\ast\!\!\!\vert}\mathrm{d}\,\dot{\boldsymbol{\varepsilon}}$, but the Stieltjes convolutions will not necessarily commute since $\dot{\theta}$ and $\dot{\boldsymbol{\varepsilon}}$ might have distributional singularities. Clearly, the equation is linear in $\boldsymbol{q}$, $\boldsymbol{u}$ and θ up to a "small" error, so it only remains to investigate the behavior of $\boldsymbol{q}$ as a function of $\boldsymbol{u}$ and θ.

With regard to $\boldsymbol{q}$, proceed as with $\boldsymbol{S}$ in (C.24) by assuming it is Fréchet differentiable at $\bar{\Xi}^q$, and use that $\boldsymbol{q}(\bar{\Xi}^q)=0$ and the results in (C.17) to conclude that

$$\boldsymbol{q}(\Xi^q) = \Big(\frac{\partial\boldsymbol{q}}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}^q), \boldsymbol{E}_d^t\Big)_{L_w^2} + \Big(\frac{\partial\boldsymbol{q}}{\partial\theta_d^t}(\bar{\Xi}^q), \theta_d^t\Big)_{L_w^2} + \frac{\partial\boldsymbol{q}}{\partial\nabla\theta}(\bar{\Xi}^q)\cdot\nabla\theta + o(\delta)\,. \tag{C.62}$$

Then, for all $\boldsymbol{X}\in\Omega$, define the material properties,

$$\begin{aligned}
\boldsymbol{\kappa}(\boldsymbol{X}) &= -\frac{\partial\boldsymbol{q}}{\partial\nabla\theta}(\bar{\Xi}^q)\,,\\
\mathsf{J}^E(\boldsymbol{X},s) &= -\int_s^\infty \frac{\partial\boldsymbol{q}}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}^q,\tau')w_{L^2}(\tau')\,\mathrm{d}\tau'\,,\\
\mathsf{j}^\theta(\boldsymbol{X},s) &= -\int_s^\infty \frac{\partial\boldsymbol{q}}{\partial\theta_d^t}(\bar{\Xi}^q,\tau')w_{L^2}(\tau')\,\mathrm{d}\tau'\,,
\end{aligned} \tag{C.63}$$

for all $s\geq 0$, while $\mathsf{J}^E(\boldsymbol{X},s)=0$ and $\mathsf{j}^\theta(\boldsymbol{X},s)=0$ for all $s<0$. Their full distributional time derivatives are $\dot{\mathsf{J}}^E(s) = \mathsf{J}^E(0)\delta_0(s) + \frac{\partial\boldsymbol{q}}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}^q,s)w_{L^2}(s)H_0(s)$ and $\dot{\mathsf{j}}^\theta(s) = \mathsf{j}^\theta(0)\delta_0(s) + \frac{\partial\boldsymbol{q}}{\partial\theta_d^t}(\bar{\Xi}^q,s)w_{L^2}(s)H_0(s)$ respectively. Thus, for all $\boldsymbol{X}\in\Omega$ and $s<0$, $\mathsf{J}^E=\dot{\mathsf{J}}^E=0$ and $\mathsf{j}^\theta=\dot{\mathsf{j}}^\theta=0$. Also,

$$\begin{aligned}
&\lim_{s\to\infty}\mathsf{J}^E(\boldsymbol{X},s)=0\,, &\qquad \dot{\mathsf{J}}^E(\boldsymbol{X},s) &= \frac{\partial\boldsymbol{q}}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}^q,s)w_{L^2}(s)\,,\\
&\lim_{s\to\infty}\mathsf{j}^\theta(\boldsymbol{X},s)=0\,, &\qquad \dot{\mathsf{j}}^\theta(\boldsymbol{X},s) &= \frac{\partial\boldsymbol{q}}{\partial\theta_d^t}(\bar{\Xi}^q,s)w_{L^2}(s)\,,
\end{aligned} \tag{C.64}$$

for all $s>0$. From the definition it follows that J^E is symmetric in its last two indices, so $\mathsf{J}^E_{ijk}=\mathsf{J}^E_{ikj}$ for all $i,j,k=1,2,3$. Lastly, it is known that $\mathsf{v}^\mathsf{T}\boldsymbol{\kappa}\mathsf{v}\geq 0$ for all $\mathsf{v}\in\mathbb{R}^3$ (see end of Section C.1). Invoking $\boldsymbol{\varepsilon}$ in (C.20) along with these definitions and properties allow (C.62) to be rewritten as

$$\begin{aligned}
\boldsymbol{q} &= \Big(\frac{\partial\boldsymbol{q}}{\partial\boldsymbol{E}_d^t}(\bar{\Xi}^q), \boldsymbol{E}_d^t\Big)_{L_w^2} + \Big(\frac{\partial\boldsymbol{q}}{\partial\theta_d^t}(\bar{\Xi}^q), \theta_d^t\Big)_{L_w^2} + \frac{\partial\boldsymbol{q}}{\partial\nabla\theta}(\bar{\Xi}^q)\cdot\nabla\theta + o(\delta)\\
&= \mathsf{J}^E(0):\boldsymbol{\varepsilon}(\cdot) + \int_0^\infty \dot{\mathsf{J}}^E(s):\boldsymbol{\varepsilon}(\cdot-s)\,\mathrm{d}s + \mathsf{j}^\theta(0)\,(\theta(\cdot)-\bar{\theta}) + \int_0^\infty \dot{\mathsf{j}}^\theta(s)(\theta(\cdot-s)-\bar{\theta})\,\mathrm{d}s - \boldsymbol{\kappa}\cdot\nabla\theta + o(\delta)\\
&= \dot{\mathsf{J}}^E \mathbin{\ast\!\!\!\vert}\, \boldsymbol{\varepsilon} + \dot{\mathsf{j}}^\theta \ast (\theta-\bar{\theta}) - \boldsymbol{\kappa}\cdot\nabla\theta + o(\delta)\\
&= \mathsf{J}^E \mathbin{\ast\!\!\!\vert}\mathrm{d}\,\boldsymbol{\varepsilon} + \mathsf{j}^\theta \ast\mathrm{d}\,(\theta-\bar{\theta}) - \boldsymbol{\kappa}\cdot\nabla\theta + o(\delta)\,.
\end{aligned} \tag{C.65}$$

As usual, the only terms requiring J^E and j^θ to be defined for negative times, $s < 0$, are $\dot{\mathsf{J}}^E \mathbin{\overset{.}{\underset{.}{\ast}}} \boldsymbol{\varepsilon}$ and $\dot{\mathsf{j}}^\theta \ast (\theta - \bar{\theta})$ respectively. Note that in this case, there are no terms $\boldsymbol{\varepsilon}^{-\infty}$ and $\theta^{-\infty}$, because the equilibrium values of J^E and j^θ vanish. Finally, if $\boldsymbol{\varepsilon} = 0$ and $\theta = \bar{\theta}$ for all $t < 0$, then the Stieltjes integrals commute, so that $\mathsf{J}^E \mathbin{\overset{.}{\underset{.}{\ast}}} \mathrm{d}\, \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon} \mathbin{\overset{.}{\underset{.}{\ast}}} \mathrm{d}\, \mathsf{J}^E$ and $\mathsf{j}^\theta \ast \mathrm{d}\,(\theta - \bar{\theta}) = (\theta - \bar{\theta}) \ast \mathrm{d}\, \mathsf{j}^\theta$ as in (C.34). This last expression for $\boldsymbol{q}$ is linear in both $\boldsymbol{u}$ and θ up to a "small" error, so this concludes the linearization of the equation of energy.

C.3 Linear first-order system

In the previous section the equations for conservation of momentum and energy in (C.21) were linearized in $\boldsymbol{u}$ and θ, and their histories. Dropping the terms of the form $o(\delta)$ and $\mathcal{O}(\delta^2) = o(\delta)$ associated to the "small" error, the following first order system emerges,

$$\begin{gathered}
\rho \ddot{\boldsymbol{u}} - \mathbf{div}\, \boldsymbol{P} = \boldsymbol{f}\,, \\
\rho c_v(0)\dot{\theta} + \int_0^\infty \rho \dot{c}_v(s) \dot{\theta}^t(s)\, \mathrm{d}s - \bar{\theta} \mathsf{M}^E(0) : \dot{\boldsymbol{\varepsilon}} - \int_0^\infty \bar{\theta} \dot{\mathsf{M}}^E(s) : \dot{\boldsymbol{\varepsilon}}^t(s)\, \mathrm{d}s + \operatorname{div} \boldsymbol{q} = r\,, \\
\boldsymbol{P} = \overline{\boldsymbol{S}} + \boldsymbol{\nabla} \boldsymbol{u} \overline{\boldsymbol{S}} + \mathsf{C}(0) : \boldsymbol{\varepsilon} + \int_0^\infty \dot{\mathsf{C}}(s) : \boldsymbol{\varepsilon}^t(s)\, \mathrm{d}s + \mathsf{M}^\theta(0)(\theta - \bar{\theta}) + \int_0^\infty \dot{\mathsf{M}}^\theta(s)(\theta^t(s) - \bar{\theta})\, \mathrm{d}s\,, \\
\boldsymbol{q} = \mathsf{J}^E(0) : \boldsymbol{\varepsilon} + \int_0^\infty \dot{\mathsf{J}}^E(s) : \boldsymbol{\varepsilon}^t(s)\, \mathrm{d}s + \mathsf{j}^\theta(0)\, (\theta - \bar{\theta}) + \int_0^\infty \dot{\mathsf{j}}^\theta(s)(\theta^t(s) - \bar{\theta})\, \mathrm{d}s - \boldsymbol{\kappa} \cdot \nabla \theta\,,
\end{gathered} \tag{C.66}$$

with $\mathfrak{k}^t(s) = \mathfrak{k}(t - s)$, where $\mathfrak{k}$ can be either $\boldsymbol{\varepsilon} = \frac{1}{2}(\boldsymbol{\nabla} \boldsymbol{u} + \boldsymbol{\nabla} \boldsymbol{u}^\mathsf{T})$, θ, $\dot{\theta}$ or $\dot{\boldsymbol{\varepsilon}}$. These equations hold for all $\boldsymbol{X} \in \Omega$ and all $t \in (-\infty, \infty)$. Notice they are not typical differential equations, but rather *integro*-differential equations.

The known time-dependent and possibly heterogeneous material properties are $\mathsf{C}(\boldsymbol{X}, s)$, the fourth order stiffness tensor; $\mathsf{M}^\theta(\boldsymbol{X}, s)$, the second order stress-temperature tensor; $\mathsf{M}^E(\boldsymbol{X}, s)$, the second order entropy-strain tensor; $c_v(\boldsymbol{X}, s)$, the (scalar) specific heat capacity; $\mathsf{J}^E(\boldsymbol{X}, s)$, the third order heat-strain tensor; and $\mathsf{j}^\theta(\boldsymbol{X}, s)$, the (vector) heat-temperature distribution. All time-dependent material properties are assumed to vanish when $s < 0$. They attain equilibrium values as $s \to \infty$ of C^∞ for C, M^∞ for M^θ and M^E, c_v^∞ for c_v, and 0 for J^E and j^θ. Moreover, when acting on general second order tensors (possibly non-symmetric), it is known that the stiffness tensor has major and minor symmetries, the stress-temperature and entropy-strain tensors are symmetric, and the heat-strain tensor is symmetric in its last two indices. Additionally, the mass density distribution, $\rho(\boldsymbol{X})$, is assumed to be known, along with the second order heat conductivity tensor, $\boldsymbol{\kappa}(\boldsymbol{X})$, the average temperature distribution, $\bar{\theta}(\boldsymbol{X})$ (about which the equations were linearized), and the residual stress distribution, $\overline{\boldsymbol{S}}(\boldsymbol{X})$. The heat conductivity tensor is positive semidefinite, but not necessarily symmetric. Finally, the time-dependent distributions of body force density, $\boldsymbol{f}$, and heat source density, r, are also assumed to be known.

The unknowns are $\boldsymbol{u}$, θ, $\boldsymbol{P}$ and $\boldsymbol{q}$, where the latter two variables are present so that there are no differentials of second order of θ and $\boldsymbol{u}$. Indeed, $\boldsymbol{P}$ and $\boldsymbol{q}$ could be replaced into the first two equations, leading to a second order system. But a first order system is preferred as one can argue it is more "physical". This argument is particularly valid when specifying boundary conditions, as often stress and heat flux boundary conditions are natural, and these correspond to specifying values of $\boldsymbol{P} \cdot \hat{\mathbf{n}}$ and $\boldsymbol{q} \cdot \hat{\mathbf{n}}$, where $\hat{\mathbf{n}}$ is the external unit normal to $\partial\Omega$.

With this in mind, consider two relatively open partitions of the boundary $\partial\Omega$, $\{\Gamma_u, \Gamma_P\}$ and $\{\Gamma_\theta, \Gamma_q\}$, so that they are relatively open sets in $\partial\Omega$ satisfying that $\partial\Omega = \overline{\Gamma_u \cup \Gamma_P} = \overline{\Gamma_\theta \cup \Gamma_q}$ and $\varnothing = \Gamma_u \cap \Gamma_P = \Gamma_\theta \cap \Gamma_q$. The boundary conditions are specified at these subsets of $\partial\Omega$,

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{X},t) &= \boldsymbol{u}^{\Gamma_u}(\boldsymbol{X},t)\,, \quad \boldsymbol{X} \in \Gamma_u\,, & \boldsymbol{P}(\boldsymbol{X},t) \cdot \hat{\mathbf{n}}(\boldsymbol{X}) &= \boldsymbol{P}_{\mathbf{n}}^{\Gamma_P}(\boldsymbol{X},t)\,, \quad \boldsymbol{X} \in \Gamma_P\,, \\ \theta(\boldsymbol{X},t) &= \theta^{\Gamma_\theta}(\boldsymbol{X},t)\,, \quad \boldsymbol{X} \in \Gamma_\theta\,, & \boldsymbol{q}(\boldsymbol{X},t) \cdot \hat{\mathbf{n}}(\boldsymbol{X}) &= \boldsymbol{q}_{\mathbf{n}}^{\Gamma_q}(\boldsymbol{X},t)\,, \quad \boldsymbol{X} \in \Gamma_q\,, \end{aligned} \tag{C.67}$$

where $\hat{\mathbf{n}}$ is the external unit normal vector at $\partial\Omega$ and $t \in (-\infty, \infty)$.

Note the equations hold for $t \in (-\infty, \infty)$, in which case only the values of $\boldsymbol{u}$ and θ need to be known in the limit of $t \to -\infty$ for the equations to be solved. When interested in solving the problem for all $\boldsymbol{X} \in \Omega$ and $t \in (-\infty, \infty)$, it is convenient to write the equations with convolutions,

$$\begin{gathered} \rho\ddot{\boldsymbol{u}} - \mathbf{div}\,\boldsymbol{P} = \boldsymbol{f}\,, \\ \rho\dot{c}_v * \dot{\theta} - \bar{\theta}\dot{\mathsf{M}}^E \mathbin{\ast\!\!:} \dot{\boldsymbol{\varepsilon}} + \operatorname{div}\boldsymbol{q} = r\,, \\ \boldsymbol{P} = \overline{\boldsymbol{S}} + \boldsymbol{\nabla}\boldsymbol{u}\overline{\boldsymbol{S}} + \dot{\mathsf{C}} \mathbin{\ast\!\!:} \boldsymbol{\varepsilon} + \dot{\mathsf{M}}^\theta * (\theta - \bar{\theta})\,, \\ \boldsymbol{q} = \dot{\mathsf{J}}^E \mathbin{\ast\!\!:} \boldsymbol{\varepsilon} + \dot{\mathsf{j}}^\theta * (\theta - \bar{\theta}) - \boldsymbol{\kappa} \cdot \nabla\theta\,. \end{gathered} \tag{C.68}$$

Here, the convolutions are $g_1 * g_2(t) = \int_{-\infty}^{\infty} g_1(s) g_2(t-s)\,\mathrm{d}s$ and $G_1 \mathbin{\ast\!\!:} G_2(t) = \int_{-\infty}^{\infty} G_1(s) : G_2(t-s)\,\mathrm{d}s$.

Written in this form, the equations are especially useful to transform to the frequency domain in case a full time-harmonic scenario is of interest.

However, for many applications the functions are unknown at $t < 0$, and one often assumes that the problem "starts" at $t = 0$. In this case, it can be assumed that $\boldsymbol{u} = 0$ and $\theta = \bar{\theta}$ for $t < 0$, so that $\boldsymbol{\varepsilon}$, $\dot{\boldsymbol{\varepsilon}}$ and $\dot{\theta}$ all vanish for $t < 0$. Under these vanishing assumptions, the system is conveniently written in terms of Stieltjes convolutions as

$$\begin{gathered}
\rho\ddot{\boldsymbol{u}} - \mathbf{div}\,\boldsymbol{P} = \boldsymbol{f}\,, \\
\rho c_v \ast\mathrm{d}\,\dot{\theta} - \bar{\theta}\mathsf{M}^E \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\dot{\boldsymbol{\varepsilon}} + \operatorname{div}\boldsymbol{q} = r\,, \\
\boldsymbol{P} = \overline{\boldsymbol{S}} + \boldsymbol{\nabla u}\overline{\boldsymbol{S}} + \mathsf{C} \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\boldsymbol{\varepsilon} + \mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta})\,, \\
\boldsymbol{q} = \mathsf{J}^E \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\boldsymbol{\varepsilon} + \mathsf{j}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta}) - \boldsymbol{\kappa}\cdot\nabla\theta\,,
\end{gathered} \tag{C.69}$$

for all $\boldsymbol{X} \in \Omega$ and $t \in (-\infty, \infty)$. Being consistent with the physics, it is desirable to look at the system and the variables only at $t > 0$. However, the Stieltjes convolution is defined as $g \ast\mathrm{d}\,h(t) = \int_{-\infty}^{t} g(t-s)\dot{h}(s)\,\mathrm{d}s$, and it involves considering values of $\dot{h}(s)$ for $s < 0$. To dispose of this requirement, it is opportune to write the Stieltjes convolution in a form which allows to look at the variables and their derivatives only for $t \geq 0$. Indeed, given g and h defined in $(-\infty, \infty)$ and with support in $[0, \infty)$, it follows that

$$\begin{aligned}
g \ast\mathrm{d}\,h(t) &= g(t)h(0) + \int_0^t g(t-s)\dot{h}(s)\,\mathrm{d}s = g(0)h(t) + \int_0^t \dot{g}(s)h(t-s)\,\mathrm{d}s = h \ast\mathrm{d}\,g(t)\,, \\
g \ast\mathrm{d}\,\dot{h}(t) &= g(t)\dot{h}(0) + \dot{g}(t)h(0) + \int_0^t g(t-s)\ddot{h}(s)\,\mathrm{d}s = g(0)\dot{h}(t) + \dot{g}(t)h(0) + \int_0^t \dot{g}(s)\dot{h}(t-s)\,\mathrm{d}s \\
&= g(0)\dot{h}(t) + \dot{g}(0)h(t) + \int_0^t \ddot{g}(s)h(t-s)\,\mathrm{d}s = h \ast\mathrm{d}\,\dot{g}(t)\,,
\end{aligned} \tag{C.70}$$

where $t \in (-\infty, \infty)$. Clearly, the values of g, $\dot{g}$, $\ddot{g}$, h, $\dot{h}$ and $\ddot{h}$ only need to be known at $t > 0$ in order to compute the integrals. Note the former expression commutes, but the latter does not commute (although roughly speaking, in the latter scenario g and h do commute under the relation "$\ast\mathrm{d}^2$"). The former expression applies to $\mathsf{C} \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\boldsymbol{\varepsilon}$, $\mathsf{M}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta})$, $\mathsf{J}^E \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\boldsymbol{\varepsilon}$ and $\mathsf{j}^\theta \ast\mathrm{d}\,(\theta - \bar{\theta})$, while the latter expression applies to $c_v \ast\mathrm{d}\,\dot{\theta}$ and $\mathsf{M}^E \overset{.}{\underset{.}{\ast}}\mathrm{d}\,\dot{\boldsymbol{\varepsilon}}$.

Notwithstanding, there is a caveat to looking only at times $t > 0$. Indeed, consider the term $\ddot{\boldsymbol{u}}(t)$ at $t > 0$, which is evidently a second time derivative of a function $\boldsymbol{u}(t)$ defined for $t > 0$.

However, its zero extension, $\ddot{\boldsymbol{u}}(t)H_0(t)$, defined for all $t \in (-\infty, \infty)$ (where $H_0(t)$ is the Heaviside step function centered at $t = 0$), is *not* a second time derivative of the zero extension $\boldsymbol{u}(t)H_0(t)$ defined for $t \in (-\infty, \infty)$. In fact, $\frac{\partial^2 \boldsymbol{u} H_0}{\partial t^2}(t) = \ddot{\boldsymbol{u}}(t)H_0(t) + \dot{\boldsymbol{u}}(0)\delta_0(t) + \boldsymbol{u}(0)\dot{\delta}_0(t)$ is the second distributional time derivative of $\boldsymbol{u}(t)H_0(t)$, where $\dot{\boldsymbol{u}}(t) = \frac{\partial \boldsymbol{u}}{\partial t}(t)$ and $\ddot{\boldsymbol{u}}(t) = \frac{\partial^2 \boldsymbol{u}}{\partial t^2}(t)$ are defined for $t \geq 0$ and $t > 0$ respectively, and where δ_0 and $\dot{\delta}_0$ are the Dirac delta distribution and its derivative centered at $t = 0$. Fortunately $\frac{\partial^2 \boldsymbol{u} H_0}{\partial t^2}(t) = \ddot{\boldsymbol{u}}(t)H_0(t)$ for $t > 0$, but it is clear the general expression for $\frac{\partial^2 \boldsymbol{u} H_0}{\partial t^2}(t)$ suggests that certain information is needed at $t = 0$.

Put more simply, it must be assumed that initial conditions,

$$\boldsymbol{u}_0(\boldsymbol{X}) = \boldsymbol{u}(\boldsymbol{X}, 0)\,, \qquad \dot{\boldsymbol{u}}_0(\boldsymbol{X}) = \dot{\boldsymbol{u}}(\boldsymbol{X}, 0)\,, \qquad \theta_0(\boldsymbol{X}) - \bar{\theta}(\boldsymbol{X}) = \theta(\boldsymbol{X}, 0) - \bar{\theta}(\boldsymbol{X})\,, \tag{C.71}$$

are known for all $\boldsymbol{X} \in \Omega$. Note that the initial conditions need to be compatible with the boundary conditions at (C.67) when $t = 0$. Also, notice that the initial conditions for the variable $\theta - \bar{\theta}$ were written, since it is this variable (and not θ) that vanishes when $t < 0$.

The integro-differential first order system written in (C.68) and (C.69) merits some further discussion. Indeed, it is so general that under some simple assumptions it collapses to several well-known equations in the literature. First, assume that the material properties are time-independent for $t > 0$, so that they take their equilibrium values, $\mathsf{C}(t) = \mathsf{C}^\infty H_0(t)$, $\mathsf{M}^\theta = \mathsf{M}^E = \mathsf{M}^\infty H_0(t)$, $c_v = c_v^\infty H_0(t)$, $\mathsf{J}^E = 0$ and $\mathsf{j}^\theta = 0$. Essentially, this implies the material no longer depends on its memory. In fact, the integrals disappear and the equations collapse to the differential first order system of linear coupled dynamic thermoelasticity (see [53]) with nonzero residual stress. If additionally $\mathsf{M}^\infty = 0$, then the equations decouple to the dynamic linear elasticity equation with nonzero residual stress, and the anisotropic dynamic heat equation with nonzero forcing. Moreover, it the equations are at a static equilibrium, then the time-dependence of all variables drops, and the system further becomes the static linear elasticity equation with nonzero residual stress and the static heat equation (Poisson equation if heat conductivity is also isotropic). Finally, if the residual stress vanishes, $\overline{\boldsymbol{S}} = 0$, then the linearized $\boldsymbol{P}$ becomes symmetric and can be naturally identified with the Cauchy stress tensor $\boldsymbol{\sigma}$, so the first equation collapses to the classical static linear elasticity equation.

Similarly, when $\mathsf{M}^E = 0$ in (C.68) or (C.69), and the residual stress vanishes, $\overline{\boldsymbol{S}} = 0$, then $\boldsymbol{P}$ is identified with the Cauchy stress $\boldsymbol{\sigma}$ and the constitutive model for the stress decouples from the energy equation, so only the conservation of momentum is relevant to deduce mechanics. Additionally, let $\mathsf{C}(t) = 0$ for $t < 0$. The resulting equations are the classical equations of linear viscoelasticity, which written in convolution form are

$$\begin{gathered} \rho\ddot{\boldsymbol{u}} - \mathbf{div}\,\boldsymbol{\sigma} = \boldsymbol{f}\,, \\ \boldsymbol{\sigma} = \dot{\mathsf{C}} \ast \boldsymbol{\varepsilon} = \mathsf{C}(0)\!:\!\boldsymbol{\varepsilon} + \int_0^\infty \dot{\mathsf{C}}(s)\!:\!\boldsymbol{\varepsilon}(\cdot - s)\,\mathrm{d}s\,. \end{gathered} \tag{C.72}$$

To gain some physical intuition with these equations, consider a step displacement function of the form $\boldsymbol{u}(t) = \boldsymbol{u}_0 H_0(t)$, with $\boldsymbol{\varepsilon}_0 = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u}_0 + \boldsymbol{\nabla}\boldsymbol{u}_0^{\mathsf{T}})$, so the stress is written as (see (C.70)),

$$\boldsymbol{\sigma}(t) = \mathsf{C} \ast \mathrm{d}\,\boldsymbol{\varepsilon}_0(t) = \mathsf{C}(t)\!:\!\boldsymbol{\varepsilon}_0 + \int_0^t \mathsf{C}(t-s)\!:\!\dot{\boldsymbol{\varepsilon}}_0\,\mathrm{d}s = \mathsf{C}(t)\!:\!\boldsymbol{\varepsilon}_0\,. \tag{C.73}$$

Hence, $\lim_{t\to\infty}\boldsymbol{\sigma}(t) = \mathsf{C}^\infty\!:\!\boldsymbol{\varepsilon}_0$, which reflects precisely what occurs in a relaxation test of viscoelastic materials, where the stress has a large discrete jump at first and then "relaxes" until it reaches a nonzero equilibrium. Similarly, one can show that there exists a function $\mathsf{S}(t)$ such that the engineering strain becomes $\boldsymbol{\varepsilon}(t) = \mathsf{S} \ast \mathrm{d}\,\boldsymbol{\sigma}(t)$. Thus, under a creep test with a step stress function $\boldsymbol{\sigma}(t) = \boldsymbol{\sigma}_0 H_0(t)$ the strain is given by $\boldsymbol{\varepsilon}(t) = \mathsf{S}(t) : \boldsymbol{\sigma}_0$ and "creeps" to an equilibrium value of $\lim_{t\to\infty}\boldsymbol{\varepsilon}(t) = \mathsf{S}^\infty\!:\!\boldsymbol{\sigma}_0 = (\mathsf{C}^\infty)^{-1}\!:\!\boldsymbol{\sigma}_0$.

To finalize this section, consider the assumption made by Gurtin in [141], where the infinitesimal entropy production in each closed process was taken to be "invariant under time-reversal" and the residual stress was assumed to vanish, $\overline{\boldsymbol{S}} = 0$. Under those conditions (and since no dependence on the history of the temperature gradient was assumed in the first place) Gurtin showed that in (C.68) or (C.69), $\mathsf{M}^E = \mathsf{M}^\theta$, $\mathsf{J}^E = 0$, $\mathsf{j}^\theta = 0$ and $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\mathsf{T}}$ is symmetric. Thus, let $\mathsf{M} = \mathsf{M}^E = \mathsf{M}^\theta$, and reinterpret $\boldsymbol{P}$ as the Cauchy stress tensor $\boldsymbol{\sigma}$. It can be shown that [53, 143],

$$\mathsf{M}(\boldsymbol{X},t) = -\dot{\mathsf{C}} \ast \boldsymbol{\alpha}(\boldsymbol{X},t) = -\int_{-\infty}^{\infty} \dot{\mathsf{C}}(\boldsymbol{X},s)\!:\!\boldsymbol{\alpha}(\boldsymbol{X},t-s)\,\mathrm{d}s\,, \tag{C.74}$$

where $\boldsymbol{\alpha}(\boldsymbol{X},t)$ is a new time-dependent material property known as the tensor of coefficients of linear thermal expansion. It is symmetric ($\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathsf{T}}$) and relates the engineering strain to dynamic changes in temperature under no external stresses (i.e. when $\boldsymbol{\sigma}(t) = 0$, then $\boldsymbol{\varepsilon} = \dot{\boldsymbol{\alpha}} * (\theta - \bar{\theta})$). Therefore, the linear first-order system in convolution form, (C.68), can be rewritten as,

$$
\begin{aligned}
\rho \ddot{\boldsymbol{u}} - \mathbf{div}\, \boldsymbol{\sigma} &= \boldsymbol{f}\,, \\
\rho \dot{c}_v * \dot{\theta} + \bar{\theta} \dot{\boldsymbol{\alpha}} \mathbin{\overset{.}{\underset{.}{*}}} \dot{\mathsf{C}} \mathbin{\overset{.}{\underset{.}{*}}} \dot{\boldsymbol{\varepsilon}} + \operatorname{div} \boldsymbol{q} &= r\,, \\
\boldsymbol{\sigma} &= \dot{\mathsf{C}} \mathbin{\overset{.}{\underset{.}{*}}} \big(\boldsymbol{\varepsilon} - \dot{\boldsymbol{\alpha}} * (\theta - \bar{\theta}) \big)\,, \\
\boldsymbol{q} &= -\boldsymbol{\kappa} \cdot \nabla \theta\,.
\end{aligned}
\tag{C.75}
$$

This system can then be rewritten in the frequency domain if necessary, where the Fourier transforms of $\dot{\mathsf{C}}$, $\dot{\boldsymbol{\alpha}}$ and $\dot{c}_v$, are C^*, $\boldsymbol{\alpha}^*$ and c_v^*. These are complex-valued and are typically referred to as the dynamic material properties (dynamic stiffness tensor, etc.), with their real part being called the “storage” component, while their imaginary part is called the “loss” component. Obviously, they depend on the frequency being analyzed, and dynamic mechanical analysis (DMA) experiments are often used to determine their values experimentally.

Appendix D

PolyDPG interpolation estimates

The objective of this chapter is to prove important interpolation estimates for polygonal DPG (PolyDPG) methods. It complements Section A.6 of Appendix A, since in this case the choice of discrete trial and test space is different. It is recommended for the reader to first go through the material in Appendix A, especially that in Section A.5 and Section A.6.

D.1 Interface variables

Let $K \in \mathcal{T}$ be a polygonal element. First, consider the local discrete interface spaces defined in (7.7) which form part of the discrete trial space, $\mathcal{U}_h$, defined in (7.8). They are $\mathcal{P}_C^p(\partial K)$ and $\mathcal{P}^{p-1}(\partial K)$. The idea is to show they are equal to $\mathrm{tr}_{\mathrm{grad}}^K\big(W^p(K)\big)$ and $\mathrm{tr}_{\mathrm{div}}^K\big(\boldsymbol{V}^p(K)\big)$ for some spaces $W^p(K)$ and $\boldsymbol{V}^p(\mathrm{div},K)$. To do this, denote by $\mathfrak{T}(K)=\{\mathfrak{T}_i(K)\}_{i\in I_K}$ (with I_K finite) the shape-regular edge-compatible triangulations of each $K\in\mathcal{T}$, and define the spaces,

$$\begin{aligned} W^p(K) &= \big\{u \in H^1(K) \mid u|_{\mathfrak{T}_i(K)} \in \mathcal{P}^p\big(\mathfrak{T}_i(K)\big), \forall i \in I_K\big\}\,,\\ \boldsymbol{V}^p(K) &= \big\{\boldsymbol{q} \in \boldsymbol{H}(\mathrm{div},K) \mid \boldsymbol{q}|_{\mathfrak{T}_i(K)} \in \mathcal{RT}^p\big(\mathfrak{T}_i(K)\big), \forall i \in I_K\big\}\,, \end{aligned} \tag{D.1}$$

where $\mathcal{P}^p\big(\mathfrak{T}_i(K)\big)=\mathcal{P}^p$ and $\mathcal{RT}^p\big(\mathfrak{T}_i(K)\big)=\mathcal{RT}^p$ come from the classical Nédélec sequence of the first type. Then, the fact that they are edge-compatible, and the edge-local definitions of $\mathcal{P}_C^p(\partial K)$ and $\mathcal{P}^{p-1}(\partial K)$ do indeed yield that $\mathcal{P}_C^p(\partial K)=\mathrm{tr}_{\mathrm{grad}}^K\big(W^p(K)\big)$ and $\mathcal{P}^{p-1}(\partial K)=\mathrm{tr}_{\mathrm{div}}^K\big(\boldsymbol{V}^p(K)\big)$.

Next, for each triangle $\mathfrak{T}_i(K)\in\mathfrak{T}(K)$, notice the estimate in (A.60) applies, so that,

$$\|w-\Pi_{\mathcal{H}^s(\mathfrak{T}_i(K))}w\|_{\mathcal{H}(\mathfrak{T}_i(K))} \leq \widetilde{C}_{\hat{\mathfrak{T}}_0} h^s_{\mathfrak{T}_i(K)} \|w\|_{\mathcal{H}^s(\mathfrak{T}_i(K))} \leq \widetilde{C}_{\hat{\mathfrak{T}}_0} h^s_K \|w\|_{\mathcal{H}^s(\mathfrak{T}_i(K))}\,, \tag{D.2}$$

where $\mathcal{H}^s(\mathfrak{T}_i(K))$ stands for $H^{1+s}(\mathfrak{T}_i(K))$ or $\boldsymbol{H}^s(\mathrm{div},\mathfrak{T}_i(K))$, and w stands for elements in those spaces, and where $p \geq s > \frac{1}{2}$. The operators $\Pi_{\mathcal{H}^s(\mathfrak{T}_i(K))}$ are the projection-based interpolation operators for each $\mathfrak{T}_i(K)\in\{\mathfrak{T}_i(K)\}_{i\in I_K}$, defined in Section A.6. Note that it was used that

$\mathrm{diam}(\mathcal{T}_i(K)) = h_{\mathcal{T}_i(K)} \leq h_K = \mathrm{diam}(K)$. Moreover, the constant $\widetilde{C}_{\hat{\mathcal{T}}_0} = \widetilde{C}_{\hat{\mathcal{T}}_0}(s)$ is independent of $\mathcal{T}_i(K)$ (and of K entirely), because the triangulations were assumed to be shape-regular. Therefore, adding among all $\mathcal{T}_i(K) \in \mathcal{T}(K)$ yields,

$$\|w - \Pi_{\mathcal{H}^s(K)} w\|^2_{\mathcal{H}(K)} = \sum_{i \in I_K} \|w - \Pi_{\mathcal{H}^s(\mathcal{T}_i(K))} w\|^2_{\mathcal{H}(\mathcal{T}_i(K))} \leq \widetilde{C}^2_{\hat{\mathcal{T}}_0} h_K^{2s} \|w\|^2_{\mathcal{H}^s(K)} \,. \tag{D.3}$$

where now $\mathcal{H}^s(K)$ stands for $H^{1+s}(K)$ or $\boldsymbol{H}^s(\mathrm{div}, K)$. Here, the element interpolation is defined locally by the projection-based interpolation of the triangulation,

$$\Pi_{\mathcal{H}^s(K)} : \mathcal{H}^s(K) \to \mathcal{H}^p_h(K) \,, \qquad \big(\Pi_{\mathcal{H}^s(K)} w\big)\big|_{\mathcal{T}_i(K)} = \Pi_{\mathcal{H}^s(\mathcal{T}_i(K))}(w|_{\mathcal{T}_i(K)}) \,, \tag{D.4}$$

for each $\mathcal{T}_i(K) \in \mathcal{T}(K)$, where $\mathcal{H}^p_h(K)$ stands for $W^p(K)$ or $\boldsymbol{V}^p(K)$ as defined in (D.1). This is well-defined because the projection-based interpolation is designed to be compatible. Next, following the steps in (A.61) and (A.62) for $p \in \mathbb{N}$ and $s > \frac{1}{2}$ yields,

$$\|\hat{w}_K - \Pi_{\mathcal{H}^s(\partial K)} \hat{w}_K\|_{\mathcal{H}(\partial K)} \leq C_{\hat{\mathcal{T}}_0} h_K^{\min\{s,p\}} \|\hat{w}_K\|_{\mathcal{H}^s(\partial K)} \qquad \forall \hat{w}_K \in \mathcal{H}^s(\partial K) \,. \tag{D.5}$$

where $\mathcal{H}^s(\partial K)$ stands for $H^{1/2+s}(\partial K)$ or $H^{-1/2+s}(\partial K)$ respectively and the $\hat{w}_K$ lie in those spaces. Meanwhile, $C_{\hat{\mathcal{T}}_0} = C_{\hat{\mathcal{T}}_0}(s,p)$ does not depend on K. The boundary interpolation operator here, $\Pi_{\mathcal{H}^s(\partial K)} : \mathcal{H}^s(\partial K) \to \mathcal{H}^p_h(\partial K)$, is defined by (A.57) (but with $\Pi_{\mathcal{H}^s(K)}$ defined by (D.4)), where $\mathcal{H}^p_h(\partial K)$ stands for $\mathcal{P}^p_C(\partial K)$ and $\mathcal{P}^{p-1}(\partial K)$.

D.2 Remaining variables

For the variables in (7.8) associated to $L^2(K)$ which are discretized by $\mathcal{P}^{p-1}(K)$, one could proceed directly to get an estimate of the form in (A.60), but the constant would have an explicit dependence on the polygonal shape, which we want to be as arbitrary as possible. To avoid this, recall that the discretization is actually a restriction of the polynomials in the bounding triangle, T_K of K. Next, let $w \in H^s(\Omega)$ for some $s > \frac{1}{2}$, and use interpolation theory (see [170, 23]) applied to the universal extension operators of Sobolev spaces of differential forms defined in [148] (which is even more general than the universal extension operator defined by Stein in [212]), to establish the existence of a continuous extension operator,

$$E : \mathcal{H}^s(\Omega) \to \mathcal{H}^s(\mathbb{R}^2) \,, \qquad \|Ew\|_{\mathcal{H}^s(\mathbb{R}^2)} \leq C_E \|w\|_{\mathcal{H}^s(\Omega)} \,, \tag{D.6}$$

where $s \geq 0$, Ω is the domain where the equations are being solved, and $C_E = C_E(s, \Omega) > 0$. Then, define a local interpolation operator as,

$$\Pi_{\mathcal{H}^s(K)} w|_K = \left(\Pi_{\mathcal{H}^s(T_K)} Ew|_{T_K}\right)\big|_K \,, \tag{D.7}$$

where $\Pi_{\mathcal{H}^s(T_K)}$ is the projection-based interpolation in the bounding triangle T_K. Note the extension is needed since the bounding triangles might be outside of Ω. In these calculations, obviously $\mathcal{H}^s(K) = H^s(K)$, and similarly with $\mathcal{H}^s(T_K)$. Clearly, (A.60) applies to the bounding equilateral triangle T_K, and the constant will only depend on the triangle of unit diameter $\hat{T}_0$. This means T_K is scaled by $h_{T_K} = \mathrm{diam}(T_K) = \frac{6}{\sqrt{3}} r_{\max} \leq \sqrt{12} h_K$, where $r_{\max}$ is the distance of the centroid to the furthest vertex (see Figure 7.1) and $h_K = \mathrm{diam}(K)$. Thus, (A.60) becomes,

$$\|w - \Pi_{\mathcal{H}^s(K)} w\|_{\mathcal{H}(K)} \leq \|Ew - \Pi_{\mathcal{H}^s(T_K)} Ew\|_{\mathcal{H}(T_K)} \leq C_{\hat{T}_0} h_K^{\min\{s,p\}} \|Ew\|_{\mathcal{H}^s(T_K)} \,, \tag{D.8}$$

for every $w \in \mathcal{H}^s(K)$, where $C_{\hat{T}_0} = C_{\hat{T}_0}(p, s) > 0$ is now independent of K.

D.3 Final interpolation estimates

Define the global interpolation operators as in (A.56) to construct the bounded linear global interpolation operator $\Pi_{\mathfrak{U}^s} : \mathfrak{U}^s \to \mathfrak{U}_h$. Note that adding (D.8) associated with $u \in H^s(\Omega)$ among $K \in \mathcal{T}$, using the robust finite overlap condition, and the extension operator in (D.6), gives:

$$\begin{aligned} \|u - \Pi_{H^s(\Omega)} u\|^2_{L^2(\Omega)} &\leq C^2_{\hat{T}_0} \sum_{K \in \mathcal{T}} h_K^{2\min\{s,p\}} \|Eu\|^2_{H^s(T_K)} \\ &\leq M_{\mathrm{ov}} C^2_{\hat{T}_0} h^{2\min\{s,p\}} \|Eu\|^2_{H^s(\mathbb{R}^2)} \leq C_E^2 M_{\mathrm{ov}} C^2_{\hat{T}_0} h^{2\min\{s,p\}} \|u\|^2_{H^s(\Omega)} \,, \end{aligned} \tag{D.9}$$

where $h = \sup_{K \in \mathcal{T}} h_K$ and $C_E = C_E(s, \Omega)$ is not dependent on p. The same estimate holds for the variable $\boldsymbol{q} \in (H^s(\Omega))^2$. Similar bounds (without using extension operators and the finite overlap condition) hold for $\hat{u} \in H_0^{1/2+s}(\partial\mathcal{T})$ and $\hat{q}_{\mathbf{n}} \in H^{-1/2+s}(\partial\mathcal{T})$ simply by summing the contributions in (D.5), so the result is

$$\begin{aligned} \|\hat{u} - \Pi_{H^{1/2+s}(\partial\mathcal{T})} \hat{u}\|_{H^{1/2}(\partial\mathcal{T})} &\leq C_{H^{1+s}(\hat{\mathfrak{T}}_0)} h^{\min\{s,p\}} \|\hat{u}\|_{H^{1/2+s}(\partial\mathcal{T})} \,, \\ \|\hat{q}_{\mathbf{n}} - \Pi_{H^{-1/2+s}(\partial\mathcal{T})} \hat{q}_{\mathbf{n}}\|_{H^{-1/2}(\partial\mathcal{T})} &\leq C_{\boldsymbol{H}^s(\mathrm{div},\hat{\mathfrak{T}}_0)} h^{\min\{s,p\}} \|\hat{q}_{\mathbf{n}}\|_{H^{-1/2+s}(\partial\mathcal{T})} \,, \end{aligned} \tag{D.10}$$

where the constants $C_{H^{1+s}(\hat{\mathfrak{T}}_0)}$ and $C_{\boldsymbol{H}^s(\mathrm{div},\hat{\mathfrak{T}}_0)}$ only depend on p and s, but not on K (they depend on the uniform shape-regularity of the edge-compatible triangulations of all elements). Finally, since these constants come from triangles, the theory of projection-based interpolation [90] implies that in the p-asymptotic limit,

$$C_{\hat{T}_0} = \widetilde{C}_{\hat{T}_0} (\ln p) p^{-s} \,, \;\; C_{H^{1+s}(\hat{\mathfrak{T}}_0)} = \widetilde{C}_{H^{1+s}(\hat{\mathfrak{T}}_0)} (\ln p)^2 p^{-s} \,, \;\; C_{\boldsymbol{H}^s(\mathrm{div},\hat{\mathfrak{T}}_0)} = \widetilde{C}_{\boldsymbol{H}^s(\mathrm{div},\hat{\mathfrak{T}}_0)} (\ln p) p^{-s} \,, \tag{D.11}$$

where $\widetilde{C}_{\hat{T}_0}$, $\widetilde{C}_{H^{1+s}(\hat{\mathfrak{T}}_0)}$ and $\widetilde{C}_{\boldsymbol{H}^s(\mathrm{div},\hat{\mathfrak{T}}_0)}$ are constants independent of p and of any $K \in \mathcal{T}$ across all possible meshes being considered. Overall,

$$\|\mathfrak{u} - \Pi_{\mathcal{U}^s} \mathfrak{u}\|_{\mathcal{U}} \leq C h^{\min\{s,p\}} \|\mathfrak{u}\|_{\mathcal{U}^s} \,, \tag{D.12}$$

where $\mathfrak{u} = (u, \boldsymbol{q}, \hat{u}, \hat{q}_{\mathbf{n}}) \in \mathcal{U}^s$, $h = \sup_{K \in \mathcal{T}} h_K$, and $C = C(p, s, \Omega) > 0$, but is independent of the meshes being considered. In the p-asymptotic limit, $C = (\ln p)^2 p^{-s} C_s$, where $C_s = C_s(s)$ is independent of p. Using these results and the theory of Fortin operators in Section 2.8, the results summarized in Theorem 7.1 will follow.

Bibliography

[1] Ainsworth, M., Davydov, O., and Schumaker, L. L. (2016). Bernstein-Bézier finite elements on tetrahedral-hexahedral-pyramidal partitions. *Comput. Methods Appl. Mech. Engrg.*, 304:140–170.

[2] Amato, N. M., Goodrich, M. T., and Ramos, E. A. (2001). A randomized algorithm for triangulating a simple polygon in linear time. *Discrete Comput. Geom.*, 26(2):245–265.

[3] Amestoy, P. R., Duff, I. S., L'Excellent, J.-Y., and Koster, J. (2001). A fully asynchronous multifrontal solver using distributeddynamic scheduling. *SIAM J. Matrix Anal. Appl.*, 23(1):15–41.

[4] Anderson, E., Bai, Z., and Dongarra, J. (1992). Generalized QR factorization and its applications. *Linear Algebra Appl.*, 162:243–271.

[5] Antonietti, P., Bruggi, M., Scacchi, S., and Verani, M. (2017). On the virtual element method for topology optimization on polygonal meshes: A numerical study. *Comput. Math. Appl.*, 74(5):1091–1109.

[6] Arbogast, T. and Correa, M. R. (2016). Two families of $H(\text{div})$ mixed finite elements on quadrilaterals of minimal dimension. *SIAM J. Numer. Anal.*, 54(6):3332–3356.

[7] Arnold, D. N., Awanou, G., and Winther, R. (2008). Finite elements for symmetric tensors in three dimensions. *Math. Comput.*, 77:1229–1251.

[8] Arnold, D. N. and Falk, R. S. (1987). Well-posedness of the fundamental boundary value problems for constrained anisotropic elastic materials. *Arch. Rational Mech. Anal.*, 98(2):143–165.

[9] Arnold, D. N., Falk, R. S., and Winther, R. (2006). Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155.

[10] Arnold, D. N., Falk, R. S., and Winther, R. (2007). Mixed finite element methods for linear elasticity with weakly imposed symmetry. *Math. Comput.*, 76:1699–1723.

[11] Babuška, I. and Suri, M. (1987). The *hp* version of the finite element method with quasiuniform meshes. *RAIRO Modél. Math. Anal. Numér.*, 21(2):199–238.

[12] Babuška, I. (1971). Error-bounds for finite element method. *Numer. Math.*, 16(4):322–333.

[13] Babuška, I. and Guo, B. Q. (1988). Regularity of the solution of elliptic problems with piecewise analytic data. Part I. Boundary value problems for linear elliptic equation of second order. *SIAM J. Math. Anal.*, 19(1):172–203.

[14] Babuška, I., Kellogg, R. B., and Pitkäranta, J. (1979). Direct and inverse error estimates for finite elements with mesh refinements. *Numer. Math.*, 33(4):447–471.

[15] Bacuta, C. and Qirko, K. (2015). A saddle point least squares approach to mixed methods. *Comput. Math. Appl.*, 70(12):2920–2932.

[16] Balanis, C. A. (2012). *Advanced Engineering Electromagnetics.* John Wiley & Sons, 2nd edition.

[17] Beirão da Veiga, L., Dassi, F., and Russo, A. (2017). High-order Virtual Element Method on polyhedral meshes. *Comput. Math. Appl.*, 74(5):1110–1122.

[18] Beirão da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L. D., and Russo, A. (2013a). Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, 23(01):199–214.

[19] Beirão da Veiga, L., Brezzi, F., and Marini, L. D. (2013b). Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal.*, 51(2):794–812.

[20] Beirão da Veiga, L., Brezzi, F., Marini, L. D., and Russo, A. (2016a). $H(\text{div})$ and $H(\text{curl})$-conforming virtual element methods. *Numer. Math.*, 133(2):303–332.

[21] Beirão da Veiga, L., Brezzi, F., Marini, L. D., and Russo, A. (2016b). Virtual element method for general second-order elliptic problems on polygonal meshes. *Math. Models Methods Appl. Sci.*, 26(04):729–750.

[22] Benedetto, M. F., Berrone, S., Pieraccini, S., and Scialò, S. (2014). The virtual element method for discrete fracture network simulations. *Comput. Methods Appl. Mech. Engrg.*, 280:135–156.

[23] Bergh, J. and Löfström, J. (1976). *Interpolation Spaces: An Introduction*, volume 223 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin.

[24] Bergot, M., Cohen, G., and Duruflé, M. (2010). Higher-order finite elements for hybrid meshes using new nodal pyramidal elements. *J. Sci. Comput.*, 42:345–381.

[25] Bergot, M. and Duruflé, M. (2013). Approximation of $H(\mathrm{div})$ with high-order optimal finite elements for pyramids, prisms and hexahedra. *Commun. Comput. Phys.*, 14(5):1372–1414.

[26] Beuchler, S., Pillwein, V., Schöberl, J., and Zaglmayr, S. (2012). Sparsity optimized high order finite element functions on simplices. In Langer, U. and Paule, P., editors, *Numerical and Symbolic Scientific Computing*, Texts & Monographs in Symbolic Computation, pages 21–44. Springer, Vienna.

[27] Bishop, J. E. (2009). Simulating the pervasive fracture of materials and structures using randomly close packed Voronoi tessellations. *Comput. Mech.*, 44(4):455–471.

[28] Bishop, J. E. (2014). A displacement-based finite element formulation for general polyhedra using harmonic shape functions. *Int. J. Num. Meth. in Eng.*, 97(1):1–31.

[29] Bishop, J. E., Martinez, M. J., and Newell, P. (2016). Simulating fragmentation and fluid-induced fracture in disordered media using random finite-element meshes. *Int. J. Multiscale Comput. Eng.*, 14(4):349–366.

[30] Björck, Å. (1996). *Numerical methods for least squares problems*, volume 51 of *Other Titles in Applied Mathematics*. SIAM.

[31] Björck, Å. and Duff, I. S. (1980). A direct method for the solution of sparse linear least squares problems. *Linear Algebra Appl.*, 34:43–67.

[32] Bochev, P. B. (2004). Least-squares finite element methods for first-order elliptic systems. *Int. J. Numer. Anal. Model*, 1(1):49–64.

[33] Bochev, P. B. and Gunzburger, M. D. (2009). *Least-squares finite element methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York.

[34] Boffi, D., Brezzi, F., and Fortin, M. (2013). *Mixed Finite Element Methods and Applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Berlin.

[35] Boltzmann, L. (1874). Zur theorie der elastischen nachwirkung. *Wien. Ber.*, 70:275–306.

[36] Bramwell, J., Demkowicz, L., Gopalakrishnan, J., and Weifeng, Q. (2012). A locking-free *hp* DPG method for linear elasticity with symmetric stresses. *Numer. Math.*, 122(4):671–707.

[37] Brenner, S. and Scott, R. (2007). *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer Science & Business Media.

[38] Brezzi, F. and Fortin, M. (1991). *Mixed and Hybrid Finite Element Methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer, New York.

[39] Brezzi, F., Lipnikov, K., and Shashkov, M. (2005a). Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 43(5):1872–1896.

[40] Brezzi, F., Lipnikov, K., and Simoncini, V. (2005b). A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15(10):1533–1551.

[41] Brezzi, F. and Marini, L. D. (2013). Virtual element methods for plate bending problems. *Comput. Methods Appl. Mech. Engrg.*, 253:455–462.

[42] Broersen, D., Dahmen, W., and Stevenson, R. (2018). On the stability of DPG formulations of transport equations. *Math. Comp.*, 87(311):1051–1082.

[43] Broersen, D. and Stevenson, R. (2014). A robust Petrov-Galerkin discretisation of convection-diffusion equations. *Comput. Math. Appl.*, 68(11):1605–1618.

[44] Broersen, D. and Stevenson, R. (2015). A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form. *IMA J. Numer. Anal.*, 35(1):39–73.

[45] Buffa, A., Costabel, M., and Sheen, D. (2002). On traces for $H(\mathrm{curl},\Omega)$ in Lipschitz domains. *J. Math. Anal. Appl.*, 276(2):845–867.

[46] Bui-Thanh, T., Demkowicz, L., and Ghattas, O. (2013). A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. *SIAM J. Numer. Anal.*, 51(4):1933–1958.

[47] Bui-Thanh, T. and Ghattas, O. (2014). A PDE-constrained optimization approach to the discontinuous Petrov-Galerkin method with a trust region inexact Newton-CG solver. *Comput. Methods Appl. Mech. Engrg.*, 278:20–40.

[48] Buttari, A. (2013). Fine-grained multithreading for the multifrontal QR factorization of sparse matrices. *SIAM J. Sci. Comput.*, 35(4):C323–C345.

[49] Cai, Z., Lazarov, R., Manteuffel, T. A., and McCormick, S. F. (1994). First-order system least squares for second-order partial differential equations: Part I. *SIAM J. Numer. Anal.*, 31(6):1785–1799.

[50] Cai, Z., Manteuffel, T. A., and McCormick, S. F. (1997). First-order system least squares for second-order partial differential equations: Part II. *SIAM J. Numer. Anal.*, 34(2):425–454.

[51] Cameron, D. E., Lang, J. H., and Umans, S. D. (1992). The origin and reduction of acoustic noise in doubly salient variable-reluctance motors. *IEEE Trans. Ind. Appl.*, 28(6):1250–1255.

[52] Cangiani, A., Georgoulis, E. H., and Houston, P. (2014). *hp*-version discontinuous Galerkin methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 24(10):2009–2041.

[53] Carlson, D. E. (1973). Linear thermoelasticity. In Truesdell, C., editor, *Linear Theories of Elasticity and Thermoelasticity*, pages 297–346. Springer-Verlag, Berlin.

[54] Carstensen, C., Demkowicz, L., and Gopalakrishnan, J. (2014). A posteriori error control for DPG methods. *SIAM J. Numer. Anal.*, 52(3):1335–1353.

[55] Carstensen, C., Demkowicz, L., and Gopalakrishnan, J. (2016). Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.*, 72(3):494–522.

[56] Carstensen, C. and Hellwig, F. (2016). Low-order discontinuous Petrov-Galerkin finite element methods for linear elasticity. *SIAM J. Numer. Anal.*, 54(6):3388–3410.

[57] Castro, D. A., Devloo, P. R., Farias, A. M., Gomes, S. M., de Siqueira, D., and Durán, O. (2016). Three dimensional hierarchical mixed finite element approximations with enhanced primal variable accuracy. *Comput. Methods Appl. Mech. Engrg.*, 306:479–502.

[58] Chan, J., Demkowicz, L., and Moser, R. (2014a). A DPG method for steady viscous compressible flow. *Comput. Fluids*, 98:69–90.

[59] Chan, J., Evans, J. A., and Qiu, W. (2014b). A dual Petrov-Galerkin finite element method for the convectiondiffusion equation. *Comput. Math. Appl.*, 68(11):1513–1529.

[60] Chan, J., Heuer, N., Bui-Thanh, T., and Demkowicz, L. (2014c). A robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms. *Comput. Math. Appl.*, 67(4):771–795.

[61] Chazelle, B. (1991). Triangulating a simple polygon in linear time. *Discrete Comput. Geom.*, 6(3):485–524.

[62] Chen, L., Wei, H., and Wen, M. (2017). An interface-fitted mesh generator and virtual element methods for elliptic interface problems. *J. Comput. Phys.*, 334(1):327–348.

[63] Chen, W. and Wang, Y. (2017). Minimal degree $H(\text{curl})$ and $H(\text{div})$ conforming finite elements on polytopal meshes. *Math. Comp.*, 86(307):2053–2087.

[64] Chi, H., Beirão da Veiga, L., and Paulino, G. H. (2017). Some basic formulations of the virtual element method (VEM) for finite deformations. *Comput. Methods Appl. Mech. Engrg.*, 318:148–192.

[65] Chi, H., Talischi, C., Lopez-Pamies, O., and Paulino, G. H. (2015). Polygonal finite elements for finite elasticity. *Int. J. Num. Meth. in Eng.*, 101(4):305–328.

[66] Chin, E. B., Lasserre, J. B., and Sukumar, N. (2015). Numerical integration of homogeneous functions on convex and nonconvex polygons and polyhedra. *Comput. Mech.*, 56(6):967–981.

[67] Ciarlet, P. G. (1988). *Mathematical Elasticity, Volume 1: Three Dimensional Elasticity.* North-Holland, Amsterdam.

[68] Ciarlet, P. G. (2002). *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics.* SIAM.

[69] Ciarlet, P. G. (2010). On Korn's inequality. *Chin. Ann. Math. Ser. B*, 31(5):607–618.

[70] Ciarlet, P. G. (2013). *Linear and Nonlinear Functional Analysis with Applications*, volume 130 of *Other Titles in Applied Mathematics.* SIAM.

[71] Cockburn, B. (2016). Static condensation, hybridization, and the devising of the HDG methods. In Barrenechea, G. R., Brezzi, F., Cangiani, A., and Georgoulis, E. H., editors, *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, volume 114 of *Lecture Notes in Computational Science and Engineering*, pages 129–177. Springer, Cham.

[72] Cockburn, B., Di Pietro, D. A., and Ern, A. (2016). Bridging the hybrid high-order and hybridizable discontinuous Galerkin methods. *ESAIM Math. Model. Numer. Anal.*, 50(3):635–650.

[73] Cockburn, B. and Fu, G. (2017a). Superconvergence by M-decompositions. Part II: Construction of two-dimensional finite elements. *ESAIM Math. Model. Numer. Anal.*, 51(1):165–186.

[74] Cockburn, B. and Fu, G. (2017b). A systematic construction of finite element commuting exact sequences. *SIAM J. Numer. Anal.*, 55(4):1650–1688.

[75] Cohen, A., Dahmen, W., and Welper, G. (2012). Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(5):1247–1273.

[76] Coker, E. G. and Filon, L. N. G. (1957). *A Treatise on Photoelasticity.* Cambridge University Press, London, 2nd edition.

[77] Coleman, B. D. (1964a). On thermodynamics, strain impulses, and viscoelasticity. *Arch. Rational Mech. Anal.*, 17(3):230–254.

[78] Coleman, B. D. (1964b). Thermodynamics of materials with memory. *Arch. Rational Mech. Anal.*, 17(1):1–46.

[79] Coleman, B. D. and Gurtin, M. E. (1967). Equipresence and constitutive equations for rigid heat conductors. *Z. Angew. Math. Phys.*, 18(2):199–208.

[80] Coleman, B. D. and Mizel, V. J. (1963). Thermodynamics and departures from Fourier's law of heat conduction. *Arch. Rational Mech. Anal.*, 13(1):245–261.

[81] Coleman, B. D. and Mizel, V. J. (1967). A general theory of dissipation in materials with memory. *Arch. Rational Mech. Anal.*, 27(4):255–274.

[82] Coleman, B. D. and Noll, W. (1960). An approximation theorem for functionals, with applications in continuum mechanics. *Arch. Rational Mech. Anal.*, 6(1):355–370.

[83] Coleman, B. D. and Noll, W. (1961). Foundations of linear viscoelasticity. *Rev. Mod. Phys.*, 33:239–249.

[84] Costabel, M., Dauge, M., and Demkowicz, L. (2008). Polynomial extension operators for H^1, $H(\text{curl})$ and $H(\text{div})$ spaces on a cube. *Math. Comp.*, 77(264):1967–1999.

[85] Cueto, E., Calvo, B., and Doblaré, M. (2002). Modelling three-dimensional piece-wise homogeneous domains using the α-shape-based natural element method. *Int. J. Num. Meth. in Eng.*, 54(6):871–897.

[86] Dahmen, W., Huang, C., Schwab, C., and Welper, G. (2012). Adaptive Petrov-Galerkin methods for first order transport equations. *SIAM J. Numer. Anal.*, 50(5):2420–2445.

[87] Day, W. A. (1972). *The Thermodynamics of Simple Materials with Fading Memory*, volume 22 of *Springer Tracts in Natural Philosophy.* Springer-Verlag, Berlin.

[88] Day, W. A. and Gurtin, M. E. (1969). On the symmetry of the conductivity tensor and other restrictions in the nonlinear theory of heat conduction. *Arch. Rational Mech. Anal.*, 33(1):26–32.

[89] Demkowicz, L. (2006). *Computing with hp Finite Elements. I. One and Two Dimensional Elliptic and Maxwell Problems.* Chapman & Hall/CRC Press, New York.

[90] Demkowicz, L. (2008). Polynomial exact sequences and projection-based interpolation with application to Maxwell equations. In Boffi, D. and Gastaldi, L., editors, *Mixed Finite Elements, Compatibility Conditions, and Applications*, volume 1939 of *Lecture Notes in Mathematics*, pages 101–158. Springer, Berlin.

[91] Demkowicz, L. (2017). Lecture notes of "Advanced Finite Element Analysis".

[92] Demkowicz, L. and Gopalakrishnan, J. (2010). A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Engrg.*, 199(23-24):1558–1572.

[93] Demkowicz, L. and Gopalakrishnan, J. (2011). A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differential Equations*, 27(1):70–105.

[94] Demkowicz, L. and Gopalakrishnan, J. (2013). A primal DPG method without a first-order reformulation. *Comput. Math. Appl.*, 66(6):1058–1064.

[95] Demkowicz, L. and Gopalakrishnan, J. (2014). An overview of the DPG method. In Feng, X., Karakashian, O., and Xing, Y., editors, *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations*, volume 157 of *The IMA Volumes in Mathematics and its Applications*, pages 149–180. Springer.

[96] Demkowicz, L., Gopalakrishnan, J., Muga, I., and Zitelli, J. (2012a). Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 213–216:126–138.

[97] Demkowicz, L., Gopalakrishnan, J., Nagaraj, S., and Sepúlveda, P. (2017). A spacetime DPG method for the Schrödinger equation. *SIAM J. Numer. Anal.*, 55(4):1740–1759.

[98] Demkowicz, L., Gopalakrishnan, J., and Niemi, A. H. (2012b). A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62(4):396–427.

[99] Demkowicz, L., Gopalakrishnan, J., and Schöberl, J. (2008). Polynomial extension operators. Part I. *SIAM J. Numer. Anal.*, 46(6):3006–3031.

[100] Demkowicz, L., Gopalakrishnan, J., and Schöberl, J. (2009). Polynomial extension operators. Part II. *SIAM J. Numer. Anal.*, 47(5):3293–3324.

[101] Demkowicz, L., Gopalakrishnan, J., and Schöberl, J. (2012c). Polynomial extension operators. Part III. *Math. Comp.*, 81(279):1289–1326.

[102] Demkowicz, L. and Heuer, N. (2013). Robust DPG method for convection-dominated diffusion problems. *SIAM J. Numer. Anal.*, 51(5):2514–2537.

[103] Demkowicz, L., Kurtz, J., Pardo, D., Paszyński, M., Rachowicz, W., and Zdunek, A. (2007). *Computing with hp Finite Elements. II. Frontiers: Three Dimensional Elliptic and Maxwell Problems with Applications.* Chapman & Hall/CRC, New York.

[104] Di Pietro, D. A., Ern, A., and Lemaire, S. (2016). A review of Hybrid High-Order methods: formulations, computational aspects, comparison with other methods. In Barrenechea, G. R., Brezzi, F., Cangiani, A., and Georgoulis, E. H., editors, *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, volume 114 of *Lecture Notes in Computational Science and Engineering*, pages 205–236. Springer, Cham.

[105] Dill, E. H. (2006). *Continuum Mechanics: Elasticity, Plasticity, Viscoelasticity.* CRC Press, Boca Raton, FL.

[106] Droniou, J., Eymard, R., Gallouët, T., and Herbin, R. (2010). A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(02):265–295.

[107] Droniou, J., Eymard, R., Gallouët, T., and Herbin, R. (2013). Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Math. Models Methods Appl. Sci.*, 23(13):2395–2432.

[108] Ekeland, I. and Témam, R. (1999). *Convex Analysis and Variational Problems*, volume 28 of *Classics in Applied Mathematics*. SIAM.

[109] Ellis, T., Demkowicz, L., and Chan, J. (2014). Locally conservative discontinuous Petrov-Galerkin finite elements for fluid problems. *Comput. Math. Appl.*, 68(11):1530–1549.

[110] Falk, R. S. (2008). Finite element methods for linear elasticity. In Boffi, D. and Gastaldi, L., editors, *Mixed Finite Elements, Compatibility Conditions, and Applications*, volume 1939 of *Lecture Notes in Mathematics*, pages 159–194. Springer, Berlin.

[111] Fuentes, F., Demkowicz, L., and Wilder, A. (2017a). Using a DPG method to validate DMA experimental calibration of viscoelastic materials. *Comput. Methods Appl. Mech. Engrg.*, 325:748–765.

[112] Fuentes, F., Keith, B., and Demkowicz, L. (2015a). Exact sequence for elements of all shapes (`ESEAS`): an implementation of hierarchical high-order orientation-embedded shape functions for standard elements and Sobolev spaces in 1D, 2D and 3D. `http://github.com/libESEAS/ESEAS`.

[113] Fuentes, F., Keith, B., Demkowicz, L., and Le Tallec, P. (2017b). Coupled variational formulations of linear elasticity and the DPG methodology. *J. Comput. Phys.*, 348:715–731.

[114] Fuentes, F., Keith, B., Demkowicz, L., and Nagaraj, S. (2015b). Orientation embedded high order shape functions for the exact sequence elements of all shapes. *Comput. Math. Appl.*, 70(4):353–458.

[115] Führer, T. (2017). Superconvergent DPG methods for second order elliptic problems. *ArXiv e-prints*, arXiv:1712.07719 [math.NA].

[116] Führer, T. (2018). Superconvergence in a DPG method for an ultra-weak formulation. *Comput. Math. Appl.*, 75(5):1705–1718.

[117] Führer, T. and Heuer, N. (2016). Robust coupling of DPG and BEM for a singularly perturbed transmission problem. *Comput. Math. Appl.*, 74(8):1940–1954.

[118] Führer, T., Heuer, N., and Gupta, J. S. (2017a). A time-stepping DPG scheme for the heat equation. *Comput. Methods Appl. Math.*, 17(2):237–252.

[119] Führer, T., Heuer, N., and Karkulik, M. (2017b). On the coupling of DPG and BEM. *Math. Comp.*, 86(307):2261–2284.

[120] Führer, T., Heuer, N., Karkulik, M., and Rodríguez, R. (2017c). Combining the DPG method with finite elements. *Comput. Methods Appl. Math.*

[121] Führer, T., Heuer, N., and Stephan, E. P. (2017d). On the DPG method for Signorini problems. *IMA J. Numer. Anal.*

[122] Gain, A. L., Paulino, G. H., Duarte, L. S., and Menezes, I. F. (2015). Topology optimization using polytopes. *Comput. Methods Appl. Mech. Engrg.*, 293:411–430.

[123] Gatto, P. and Demkowicz, L. (2010). Construction of H^1-conforming hierarchical shape functions for elements of all shapes and transfinite interpolation. *Finite Elem. Anal. Des.*, 46:474–486.

[124] Ghosh, S. and Moorthy, S. (1995). Elastic-plastic analysis of arbitrary heterogeneous materials with the Voronoi cell finite element method. *Comput. Methods Appl. Mech. Engrg.*, 121(1–4):373–409.

[125] Gillette, A. (2016). Serendipity and tensor product affine pyramid finite elements. *SMAI J. Comput. Math.*, 2:215–228.

[126] Gillette, A., Rand, A., and Bajaj, C. (2016). Construction of scalar and vector finite element families on polygonal and polyhedral meshes. *Comput. Methods Appl. Math.*, 16(4):667–683.

[127] Girgis, R. S. and Vermas, S. P. (1981). Method for accurate determination of resonant frequencies and vibration behaviour of stators of electrical machines. *IEE Proc. B*, 128(1):1–11.

[128] Goggin, P. R. (1973). The elastic constants of carbon-fibre composites. *J. Mater. Sci.*, 8(2):233–244.

[129] Golub, G. (1965). Numerical methods for solving linear least squares problems. *Numer. Math.*, 7(3):206–216.

[130] Golub, G. H. and Greif, C. (2003). On solving block-structured indefinite linear systems. *SIAM J. Sci. Comput.*, 24(6):2076–2092.

[131] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3 of *Johns Hopkins Studies in the Mathematical Sciences.* JHU Press.

[132] Gopalakrishnan, J., Muga, I., and Olivares, N. (2014). Dispersive and dissipative errors in the DPG method with scaled norms for Helmholtz equation. *SIAM J. Sci. Comput.*, 36(1):A20–A39.

[133] Gopalakrishnan, J. and Qiu, W. (2014). An analysis of the practical DPG method. *Math. Comp.*, 83(286):537–552.

[134] Griffiths, D. J. (2017). *Introduction to Electrodynamics.* Cambridge University Press, Padstow, UK, 4th edition.

[135] Grisvard, P. (1992). *Singularities in Boundary Value Problems*, volume 22 of *Research Notes in Applied Mathematics.* Masson, Paris.

[136] Gruber, F., Klewinghaus, A., and Mula, O. (2017). The DUNE-DPG library for solving PDEs with Discontinuous Petrov-Galerkin finite elements. *Arch. Num. Soft.*, 5(1):111–128.

[137] Gulliksson, M. and Wedin, P.-Å. (1992). Modifying the QR-decomposition to constrained and weighted linear least squares. *SIAM J. Matrix Anal. Appl.*, 13(4):1298–1313.

[138] Guo, B. Q. and Babuška, I. (1986a). The *h-p* version of the finite element method. Part 1: The basic approximation results. *Comput. Mech.*, 1(1):21–41.

[139] Guo, B. Q. and Babuška, I. (1986b). The *h-p* version of the finite element method. Part 2: General results and applications. *Comput. Mech.*, 1(3):203–220.

[140] Gurtin, M. E. (1968). On the thermodynamics of materials with memory. *Arch. Rational Mech. Anal.*, 28(1):40–50.

[141] Gurtin, M. E. (1972). Time-reversal and symmetry in the thermodynamics of materials with memory. *Arch. Rational Mech. Anal.*, 44(5):387–399.

[142] Gurtin, M. E. and Sternberg, E. (1962). On the linear theory of viscoelasticity. *Arch. Rational Mech. Anal.*, 11(1):291–356.

[143] Gutierrez-Lemini, D. (2014). *Engineering Viscoelasticity.* Springer US, New York, NY.

[144] Hackbusch, W. (1992). *Elliptic Differential Equations: Theory and Numerical Treatment*, volume 18 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin.

[145] Heuer, N. and Karkulik, M. (2015). DPG method with optimal test functions for a transmission problem. *Comput. Math. Appl.*, 70(5):1070–1081.

[146] Heuer, N. and Karkulik, M. (2017a). Discontinuous Petrov-Galerkin boundary elements. *Numer. Math.*, 135(4):1011–1043.

[147] Heuer, N. and Karkulik, M. (2017b). A robust DPG method for singularly perturbed reaction-diffusion problems. *SIAM J. Numer. Anal.*, 55(3):1218–1242.

[148] Hiptmair, R., Li, J., and Zou, J. (2012). Universal extension for Sobolev spaces of differential forms and applications. *J. Funct. Anal.*, 263(2):364–382.

[149] Hormann, K. and Floater, M. S. (2006). Mean value coordinates for arbitrary planar polygons. *ACM Trans. Graph.*, 25(4):1424–1441.

[150] Hough, P. D. and Vavasis, S. A. (1997). Complete orthogonal decomposition for weighted least squares. *SIAM J. Matrix Anal. Appl.*, 18(2):369–392.

[151] Houston, P., Schwab, C., and Süli, E. (2000). Stabilized *hp*-finite element methods for first-order hyperbolic problems. *SIAM J. Numer. Anal.*, 37(5):1618–1643.

[152] Houston, P., Schwab, C., and Süli, E. (2002). Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163.

[153] Hu, J. (2015). A new family of efficient conforming mixed finite elements on both rectangular and cuboid meshes for linear elasticity in the symmetric formulation. *SIAM J. Numer. Anal.*, 53(3):1438–1463.

[154] Hughes, T. J. R. (1987). *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

[155] Ihlenburg, F. (1998). *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

[156] Kaneko, T. (1975). On Timoshenko's correction for shear in vibrating beams. *J. Phys. D: Appl. Phys.*, 8(16):1927–1936.

[157] Keith, B., Demkowicz, L., and Gopalakrishnan, J. (2017a). DPG* method. *ArXiv e-prints*, arXiv:1710.05223 [math.NA].

[158] Keith, B., Fuentes, F., and Demkowicz, L. (2016). The DPG methodology applied to different variational formulations of linear elasticity. *Comput. Methods Appl. Mech. Engrg.*, 309:579–609.

[159] Keith, B., Knechtges, P., Roberts, N. V., Elgeti, S., Behr, M., and Demkowicz, L. (2017b). An ultraweak DPG method for viscoelastic fluids. *J. Nonnewton. Fluid. Mech.*, 247:107–122.

[160] Keith, B., Petrides, S., Fuentes, F., and Demkowicz, L. (2017c). Discrete least-squares finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 327:226–255.

[161] Keith, B., Vaziri Astaneh, A., and Demkowicz, L. (2017d). Goal-oriented adaptive mesh refinement for non-symmetric functional settings. *ArXiv e-prints*, arXiv:1711.01996 [math.NA].

[162] Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *Amer. Statist.*

[163] Kuznetsov, Y., Lipnikov, K., and Shashkov, M. (2004). The mimetic finite difference method on polygonal meshes for diffusion-type problems. *Computat. Geosci.*, 8(4):301–324.

[164] Lakes, R. (2009). *Viscoelastic Materials*. Cambridge University Press, Cambridge.

[165] Lakes, R. and Wojciechowski, K. W. (2008). Negative compressibility, negative Poisson's ratio, and stability. *Phys. Status Solidi B*, 245(3):545–551.

[166] Lee, N.-S. and Bathe, K.-J. (1993). Effects of element distortions on the performance of isoparametric elements. *Int. J. Num. Meth. in Eng.*, 36(20):3553–3576.

[167] Leon, S. E., Spring, D. W., and Paulino, G. H. (2014). Reduction in mesh bias for dynamic fracture using adaptive splitting of polygonal finite elements. *Int. J. Num. Meth. in Eng.*, 100(8):555–576.

[168] Liu, J. W. H. (1992). The multifrontal method for sparse matrix solution: Theory and practice. *SIAM Review*, 34(1):82–109.

[169] Manzini, G., Russo, A., and Sukumar, N. (2014). New perspectives on polygonal and polyhedral finite element methods. *Math. Models Methods Appl. Sci.*, 24(08):1665–1699.

[170] McLean, W. (2000). *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, Cambridge.

[171] Meisters, G. H. (1975). Polygons have ears. *Amer. Math. Monthly*, 82(6):648–651.

[172] Melenk, J. M. and Sauter, S. (2018). Wavenumber-explicit *hp*-FEM analysis for Maxwell's equations with transparent boundary conditions. *ArXiv e-prints*, arXiv:1803.01619 [math.NA].

[173] Moro, D., Nguyen, N., and Peraire, J. (2012). A hybridized discontinuous Petrov-Galerkin scheme for scalar conservation laws. *Int. J. Numer. Meth. Eng.*, 91(9):950–970.

[174] Mousavi, S. E. and Sukumar, N. (2011). Numerical integration of polynomials and discontinuous functions on irregular convex polygons and polyhedrons. *Comput. Mech.*, 47(5):535–554.

[175] Mu, L., Wang, J., and Ye, X. (2015). Weak Galerkin finite element methods on polytopal meshes. *Int. J. Numer. Anal. Model.*, 12(1):31–53.

[176] Mu, L., Wang, J., Ye, X., and Zhao, S. (2016). A new weak Galerkin finite element method for elliptic interface problems. *J. Comput. Phys.*, 325:157–173.

[177] Muga, I. and van der Zee, K. G. (2015). Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov-Galerkin, and monotone mixed methods. *ArXiv e-prints*, arXiv:1511.04400 [math.NA].

[178] Murrmann, H. and Widmann, D. (1969). Current crowding on metal contacts to planar devices. *IEEE Trans. Electron Devices*, 16(12):1022–1024.

[179] Nagaraj, S., Petrides, S., and Demkowicz, L. (2017). Construction of DPG Fortin operators for second order problems. *Comput. Math. Appl.*, 74(8):1964–1980.

[180] Nassar, O. M. (1987). The use of partial discharge and impulse voltage testing in the evaluation of interturn insulation failure of large motors. *IEEE Trans. Energy Convers.*, EC-2(4):615–621.

[181] Nédélec, J. C. (1980). Mixed finite elements in $\mathbb{R}^3$. *Numer. Math.*, 35:315–341.

[182] Neff, P., Pauly, D., and Witsch, K.-J. (2015). Poincaré meets Korn via Maxwell: Extending Korn's first inequality to incompatible tensor fields. *J. Differential Equations*, 258(4):1267–1302.

[183] Niemi, A. H., Bramwell, J., and Demkowicz, L. (2011). Discontinuous Petrov-Galerkin method with optimal test functions for thin-body problems in solid mechanics. *Comput. Methods Appl. Mech. Engrg.*, 200(9–12):1291–1300.

[184] Niemi, A. H., Collier, N., and Calo, V. M. (2013). Automatically stable discontinuous Petrov-Galerkin methods for stationary transport problems: Quasi-optimal test space norm. *Comput. Math. Appl.*, 66(10):2096–2113.

[185] Nigam, N. and Phillips, J. (2012). High-order conforming finite elements on pyramids. *IMA J. Numer. Anal.*, 32(2):448–483.

[186] Oden, J. T. (1972). *Finite Elements of Nonlinear Continua.* McGraw-Hill, New York. Reprinted in 2006 by Dover Publications, Inc., Mineola, NY.

[187] Oden, J. T. (2011). *An Introduction to Mathematical Modeling: A Course in Mechanics.* John Wiley & Sons, Hoboken, NJ.

[188] Oden, J. T. and Armstrong, W. H. (1971). Analysis of nonlinear, dynamic coupled thermoviscoelasticity problems by the finite element method. *Comput. Struct.*, 1(4):603–621.

[189] Oden, J. T. and Demkowicz, L. (2010). *Applied Functional Analysis.* Chapman & Hall/CRC Press, New York, 2nd edition.

[190] Oden, J. T. and Reddy, J. N. (1983). *Variational Methods in Theoretical Mechanics.* Universitext. Springer-Verlag, Berlin, 2nd edition.

[191] Paige, C. C. (1990). Some aspects of generalized QR factorizations. In Cox, M. G. and Hammarling, S. J., editors, *Reliable Numerical Computation*, pages 71–91. Clarendon Press, Oxford, UK.

[192] Pechstein, A. and Schöberl, J. (2011). Tangential-displacement and normal-normal-stress continuous mixed finite elements for elasticity. *Math. Models Methods Appl. Sci.*, 21(08):1761–1782.

[193] Pechstein, A. and Schöberl, J. (2012). Anisotropic mixed finite elements for elasticity. *Internat. J. Numer. Methods Engrg.*, 90(2):196–217.

[194] Petrides, S. and Demkowicz, L. (2017). An adaptive DPG method for high frequency time-harmonic wave propagation problems. *Comput. Math. Appl.*, 74(8):1999–2017.

[195] Qiu, W. and Demkowicz, L. (2011). Mixed *hp*-finite element method for linear elasticity with weakly imposed symmetry: Stability analysis. *SIAM J. Numer. Anal.*, 49(2):619–641.

[196] Rand, A., Gillette, A., and Bajaj, C. (2014). Quadratic serendipity finite elements on polygons using generalized barycentric coordinates. *Math. Comput.*, 83(290):2691–2716.

[197] Reddy, J. N. (1976). Variational principles for linear coupled dynamic theory of thermoviscoelasticity. *Int. J. Eng. Sci.*, 14(7):605–616.

[198] Roberts, N. V. (2013). *A discontinuous Petrov-Galerkin methodology for incompressible flow problems*. PhD thesis, The University of Texas at Austin, Austin, Texas, U.S.A.

[199] Roberts, N. V. (2014). Camellia: A software framework for discontinuous Petrov-Galerkin methods. *Comput. Math. Appl.*, 68(11):1581–1604.

[200] Roberts, N. V. (2016). Camellia v1.0 manual: Part I. Technical Report ANL/ALCF-16/3, Argonne National Laboratory, Argonne, Illinois.

[201] Roberts, N. V., Bui-Thanh, T., and Demkowicz, L. (2014). The DPG method for the Stokes problem. *Comput. Math. Appl.*, 67(4):966–995.

[202] Roberts, N. V., Demkowicz, L., and Moser, R. (2015). A discontinuous Petrov-Galerkin methodology for adaptive solutions to the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 301:456–483.

[203] Rudin, W. (1991). *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, Singapore, 2nd edition.

[204] Schneidesch, C. R., Deville, M. O., and Mund, E. H. (1994). Domain decomposition method coupling finite elements and preconditioned Chebyshev collocation to solve elliptic problems. In Quarteroni, A., Périaux, J., Kuznetsov, Y. A., and Widlund, O. B., editors, *Domain Decomposition Methods in Science and Engineering*, volume 157 of *Contemporary Mathematics*, pages 293–298. American Mathematical Society, USA.

[205] Schöberl, J. and Zaglmayr, S. (2005). High order Nédélec elements with local complete sequence properties. *COMPEL*, 24(2):374–384.

[206] Schötzau, D. and Schwab, C. (2015). Exponential convergence for *hp*-version and spectral finite element methods for elliptic problems in polyhedra. *Math. Models Methods Appl. Sci.*, 25(09):1617–1661.

[207] Schötzau, D., Schwab, C., and Wihler, T. P. (2013). *hp*-DGFEM for second order elliptic problems in polyhedra II: Exponential convergence. *SIAM J. Numer. Anal.*, 51(4):2005–2035.

[208] Schroder, D. K. (2006). *Semiconductor Material and Device Characterization.* John Wiley & Sons, 3rd edition.

[209] Schwab, C. (1998). *p- and hp- Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics.* Oxford University Press, Oxford.

[210] Smith, A., Wilkinson, S. J., and Reynolds, W. N. (1974). The elastic constants of some epoxy resins. *J. Mater. Sci.*, 9(4):547–550.

[211] Spring, D. W., Leon, S. E., and Paulino, G. H. (2014). Unstructured polygonal meshes with adaptive refinement for the numerical simulation of dynamic cohesive fracture. *Int. J. Fract.*, 189(1):33–57.

[212] Stein, E. M. (1970). *Singular integrals and differentiability properties of functions.* Princeton University Press, Princeton.

[213] Stern, A. (2015). Banach space projections and Petrov-Galerkin estimates. *Numer. Math.*, 130(1):125–133.

[214] Stone, G. C. (2005). Recent important changes in IEEE motor and generator winding insulation diagnostic testing standards. *IEEE Trans. Ind. Appl.*, 41(1):91–100.

[215] Stone, G. C., Culbert, I., Boulter, E. A., and Dhirani, H. (2014). *Electrical insulation for rotating machines: design, evaluation, aging, testing, and repair.* IEEE Press Series on Power Engineering. John Wiley & Sons, 2nd edition.

[216] Sukumar, N. and Tabarraei, A. (2004). Conforming polygonal finite elements. *Int. J. Num. Meth. in Eng.*, 61(12):2045–2066.

[217] Sun, J. (2012). Pulse-width modulation. In Vasca, F. and Iannelli, L., editors, *Dynamics and control of switched electronic systems*, Advances in Industrial Control, pages 25–61. Springer, London.

[218] Tabarraei, A. and Sukumar, N. (2007). Adaptive computations using material forces and residual-based error estimators on quadtree meshes. *Comput. Methods Appl. Mech. Engrg.*, 196(25):2657–2680.

[219] Talischi, C., Paulino, G. H., Pereira, A., and Menezes, I. F. (2010). Polygonal finite elements for topology optimization: a unifying paradigm. *Int. J. Num. Meth. in Eng.*, 82(6):671–698.

[220] Talischi, C., Paulino, G. H., Pereira, A., and Menezes, I. F. (2012a). `PolyMesher`: a general-purpose mesh generator for polygonal elements written in Matlab. *Struct. Multidiscip. Optim.*, 45(3):309–328.

[221] Talischi, C., Paulino, G. H., Pereira, A., and Menezes, I. F. (2012b). `PolyTop`: a Matlab implementation of a general topology optimization framework using unstructured polygonal finite element meshes. *Struct. Multidiscip. Optim.*, 45(3):329–357.

[222] Timoshenko, S. P. (1921). On the correction for shear of the differential equation for transverse vibrations of prismatic bars. *Philos. Mag.*, 41(245):744–746.

[223] Timoshenko, S. P. (1922). On the transverse vibrations of bars of uniform cross-section. *Philos. Mag.*, 43(253):125–131.

[224] Trefethen, L. N. and Bau III, D. (1997). *Numerical linear algebra*, volume 50 of *Other Titles in Applied Mathematics.* SIAM.

[225] Vasilopoulos, D. (1988). On the determination of higher order terms of singular elastic stress fields near corners. *Numer. Math.*, 53(1):51–95.

[226] Vavasis, S. A. (1994). Stable numerical algorithms for equilibrium systems. *SIAM J. Matrix Anal. Appl.*, 15(4):1108–1131.

[227] Vavasis, S. A. (1996). Stable finite elements for problems with wild coefficients. *SIAM J. Numer. Anal.*, 33(3):890–916.

[228] Vaziri Astaneh, A., Fuentes, F., Mora, J., and Demkowicz, L. (2017). `PolyDPG`: a `MATLAB` implementation of discontinuous Petrov-Galerkin (DPG) methods using polygonal elements. `http://www.polydpg.com/`.

[229] Vaziri Astaneh, A., Fuentes, F., Mora, J., and Demkowicz, L. (2018a). High-order polygonal discontinuous Petrov-Galerkin (PolyDPG) methods using ultraweak formulations. *Comput. Methods Appl. Mech. Engrg.*, 332:686–711.

[230] Vaziri Astaneh, A., Keith, B., and Demkowicz, L. (2018b). On perfectly matched layers for discontinuous Petrov-Galerkin methods. *ArXiv e-prints*, arXiv:1804.04496 [math.NA].

[231] Volterra, V. (1912). Sur les équations intégro-différentielles et leurs applications. *Acta Math.*, 35:295–356.

[232] Wachspress, E. (1975). *A Rational Finite Element Basis*, volume 114 of *Mathematics in Science and Engineering*. Academic Press.

[233] Wang, C.-C. (1984). On the symmetry of the heat-conduction tensor. In Truesdell, C., editor, *Rational Thermodynamics*, pages 396–401. Springer, New York, 2nd edition.

[234] Wang, J. and Ye, X. (2014). A weak Galerkin mixed finite element method for second order elliptic problems. *Math. Comp.*, 83(289):2101–2126.

[235] Xu, J. and Zikatanov, L. (2003). Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202.

[236] Zaglmayr, S. (2006). *High Order Finite Element Methods for Electromagnetic Field Computation*. PhD thesis, Johannes Kepler Universität Linz, Linz.

[237] Zitelli, J., Muga, I., Demkowicz, L., Gopalakrishnan, J., Pardo, D., and Calo, V. M. (2011). A class of discontinuous Petrov-Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D. *J. Comput. Phys.*, 230(7):2406–2432.