Finite-element methods for fourth-order problems and smectic A liquid crystals

by

© **Abdalaziz Hamdan**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Mathematics and Statistics

Memorial University

August 2022

St. John's, Newfoundland and Labrador, Canada

Abstract

In recent years, energy-minimization finite-element methods have been proposed for the computational modelling of equilibrium states of several types of liquid crystals (LCs) [4, 34, 110]. This thesis is particularly interested in the models of smectic A liquid crystals, based on the free-energy functionals proposed by Pevnyi, Selinger, and Sluckin [112], and by Xia et al. [138]. The Euler-Lagrange equations for these models include fourth-order terms acting on the smectic order parameter (or density variation of the LC) and second-order terms acting on the $\boldsymbol{Q}$-tensor or director field. Thus, we first focus extensively on finite-element methods for fourth-order problems. These methods include (i) C^1-continuous elements with a nonsymmetric Nitsche-type penalty method to weakly impose the essential boundary conditions, (ii) a nonsymmetric version of the C^0 interior penalty method, where the nonsymmetric forms are used to guarantee optimal convergence rates in terms of $h \leq 1$ and $q \approx 40$, where h and q are the refinement level and the smectic wavenumber that prescribes a preferred wavelength for the solution of $2\pi/q$ respectively, and (iii) mixed finite-element methods based on introducing the gradient of the solution as an explicit variable and constraining its value using a Lagrange multiplier, that are symmetric and allow us to strongly impose the essential boundary conditions. Preliminary experiments show that the mixed formulations may be advantageous over the other methods, in the sense that we can construct efficient preconditioners for these discretizations. Therefore, we consider a four-field formulation for models of smectic A liquid crystals, approximating the smectic order parameter, its gradient, the Lagrange multiplier, and the $\boldsymbol{Q}$-tensor. Then, we focus on the construction of solvers for the nonlinear systems that result from the discretization of these models. We consider a Newton-Krylov-Multigrid approach, using Newton's method to linearize the systems, and developing monolithic geometric multigrid preconditioners for the resulting saddle-point systems with vertex-based patch relaxation schemes.

In memory of my mother

To my father

Lay summary

Many physical systems in the world around us can be presented in terms of differential equations. Solving these equations can help us understand physical and natural phenomena, which can save people's lives, efforts, and money. Such systems include, but are not limited to, weather forecasting, earthquake prediction, and the behaviour of liquid crystalline materials.

Because of the special properties of liquid crystalline materials, and their widespread use in TVs, laptop screens, and navigation systems, mathematical modelling of liquid crystals has been extensively studied in the last few decades. These models primarily take the form of complicated (nonlinear) energy functions, and analytically finding their extremizers is expensive and inefficient (and often not feasible), especially when we need to change the parameters that describe the liquid crystal under consideration. Thus, numerical simulations are used to study the behavior of these complex systems.

We investigate simulating one of the most recent models that describe smectic-A liquid crystals, presented by Xia et al. [138], where the optimality conditions for the energy function lead to a nonlinear coupled "multi-physics" system of partial differential equations, including a difficult fourth-order term on the density variation of the smectic crystal. We modify the energy from [138], by introducing additional variables that result in larger linear systems to be solved in the minimization process, but these systems are more amenable to efficient, parallel numerical methods. As a result, we can simulate at much higher resolutions than was possible in [138]. The main contributions of the thesis are analysis of this transformation in comparison with classical discretization techniques and the development of efficient numerical methods for solution of these systems of equations.

Acknowledgements

First of all, the great continuous thank to God who gave me the ability and patience to handle the years I spent during my study. I would like to express my deepest appreciation to my supervisor Prof. Scott MacLachlan for the generosity, guidance and helpful suggestions throughout the preparation of my thesis. This work would not have been possible without him. I am also thankful to our collaborator, Prof. Patrick Farrell. It is a great opportunity to work with him. I am grateful to Prof. Timothy Atherton for answering a lot of questions related to liquid crystalline materials.

My great gratitude to the Department of Mathematics and Statistics for the chance to be a teaching assistant during my Ph.D. studies and to the School of Graduate Studies (SGS) for their financial support.

My sincere thanks to the examination committee for their careful reading of my thesis and valuable feedback.

I will never forget my parents, sisters, sisters-in-law and brothers for their unlimited support.

Statement of contribution

The work represented in Chapters 3, 4, and 5, is the result of collaborative research between Patrick Farrell, Abdalaziz Hamdan, and Scott MacLachlan, with its intellectual property equally co-owned by all. The order of authors for all manuscripts is alphabetical, but Abdalaziz Hamdan is the primary author of all of this work. They all were written by Abdalaziz Hamdan with the guidance of Scott MacLachlan.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

For many reasons, many scientific and engineering questions that are related to our real life cannot be answered by classical theory or experimentation. For example, some phenomenon are either very complicated or contain numerous variables that characterize the system being studied. Other experiments are too expensive to do in the physical laboratories, including those related to studying liquid crystals. It is even sometimes infeasible or unethical to do some biological experiments that, for example, predict drug side effects. These have led to the development of computational modelling as a research discipline within Mathematics and Computer Science.

Computational modelling of physical phenomena can be presented in three steps [111]. The first is to define an idealization of the problem of interest in terms of the quantities in which we are interested. The second step is to obtain a mathematical model that represents the idealization of the physical phenomena. Equations representing the model are usually called the governing equations of the problem. For example, fluid motion can be accurately represented using the Navier-Stokes equation [2], and the deformation of a solid due to applied external forces can be represented using the equations of elasticity [56]. It is preferred that the governing equations be well-posed, which means that the mathematical problem has a unique solution. However, it is possible to get a problem that has many solutions, which is usually the case of governing equations obtained from modelling liquid crystals, or has no solution in complex environments, such as when modeling nuclear reactors where obtaining measurements is difficult. Some well-posed governing equations are

very complicated to solve either analytically or computationally. Therefore, simplifying assumptions to reduce the complexity of the model are given with the hope of discovering methods to solve the original model. Boundary or initial conditions are necessary to be given at this stage. The final step of computational modelling is to analytically or computationally solve the governing equations. These appear to be high-order partial differential equations (PDEs) in a wide range of applications in science, finance, economics, and fluid dynamics [126]. Some of these models, such as for thin films, beams, and liquid crystalline materials, contain fourth-order PDEs. In these cases, applying analytical means like Fourier and Laplace transforms or series solutions becomes inefficient or even impossible. As a result, numerical solutions for these models are developed. The most common and well-known numerical schemes are finite-difference methods, finite-volume methods, and finite-element methods.

1.1 Liquid crystals

Liquid crystals were first discovered by the Austrian chemist Reinitzer in 1888 [116]. They are substances with intermediate properties between liquids and solid crystals. For example, a liquid crystal can flow like liquid while its molecules are oriented in a crystal-like manner. There are a lot of examples of liquid crystals around us, such as in soap, detergents, even in the human body, like some proteins and cell membranes. As temperature and electric fields can affect the orientational order of liquid crystals, they have been widely used in technological materials, such as electric display devices. These applications have naturally led to an increased interest in studying and modelling liquid crystals. For an overview of liquid crystal physics, we refer to [49, 54, 85].

Liquid crystals are commonly characterized by their *phases*, classified as nematic, smectic, and cholesteric liquid crystals [49, 125]. Molecules in the nematic phase have locally similar orientations (that can be described by a bulk parameter, known as the director field), but generally exhibit no layered behaviour and, therefore, the molecules can either rotate or slide past one another [133]. Smectic liquid crystal phases, which usually exist at temperatures lower than those of nematic phases, have well-formed layers with crystals pointing in the same direction [71]. The type of the smectic liquid crystal is determined by the molecular arrangement within each layer. While there

are many such types, the most common smectic liquid crystals are smectic-A, where the molecules are oriented along the normal direction in each layer, and smectic-C for which the molecules have a tilted angle between the layer normal and the director. Finally, cholesteric molecules have a helical structure with layers rotated through different angles. As the temperature of a given material increases, it can exhibit phase changes from a solid, to cholesteric, smectic, nematic, and liquid phases.

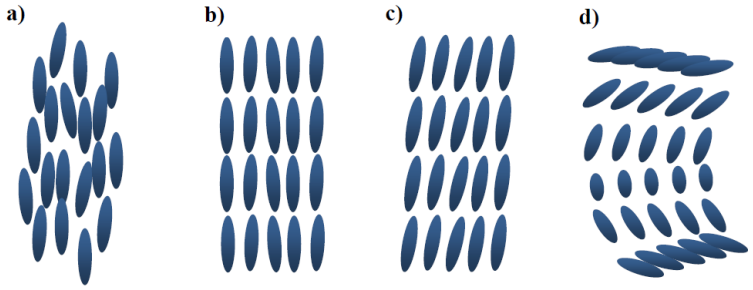

Figure 1.1: Liquid crystal phases: (a) Nematic, (b) Smectic-A, (c) Smectic-C, and (d) Cholesteric phases [139].

1.1.1 Modeling smectic-A liquid crystals

de Gennes model

In this theory, the free energy of smectic A liquid crystals can be modeled as follows [55]

$$J(\nu, \psi) = \int_{\Omega} E_1(\vec{\nu}) + E_2(\vec{\nu}, \psi),$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ is the domain that the liquid crystals occupy, ψ is a complex-valued order parameter for which $|\psi|$ describes the magnitude and $\nabla\psi$ describes the phase of the liquid crystals. The real vector field $\vec{\nu}$ is known as the director, and satisfies the pointwise contraint $\vec{\nu} \cdot \vec{\nu} - 1 = 0$. The energy $E_1(\vec{\nu})$ is the Frank-Oseen energy that is usually used to model nematic liquid crystals,

$$\begin{aligned} E_1(\vec{\nu}) &= \frac{1}{2}\left(K_1 \left(\nabla \cdot \vec{\nu}\right)^2 + K_2 \left(\vec{\nu} \cdot \nabla \times \vec{\nu}\right)^2 + K_3 \left(\vec{\nu} \times \nabla \times \vec{\nu}\right)\right) \\ &+ \frac{1}{2}\left(K_2 + K_4\right)\left(\left(tr(\nabla\vec{\nu})^2\right) - \left(\nabla \cdot \vec{\nu}\right)^2\right). \end{aligned} \tag{1.1}$$

Here, the constants $\{K_i\}_{i=1}^4$ are positive constants called the Frank constants, with K_1, K_2, and K_3 identified as the splay, twist, and bend constants, respectively, see Figure (1.2). In addition, $E_2(\vec{\nu}, \psi)$ is the smectic energy density, given by

$$E_2(\vec{\nu}, \psi) = |\nabla\psi - iq\vec{\nu}\psi|^2 + r|\psi|^2 + \frac{g}{2}|\psi|^4$$

where i is the imaginary unit, q and g are positive constants, while r is a negative constant.

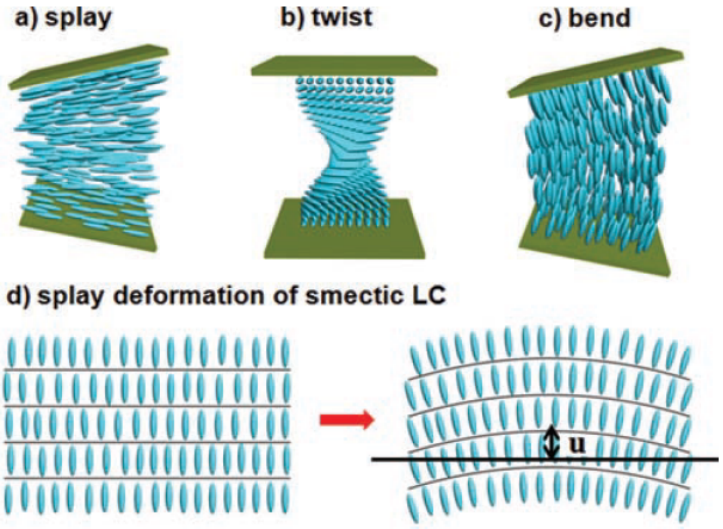

Figure 1.2: The effect of splaying, twisting, and bending liquid crystals [89].

1.1.2 The Pevnyi, Selinger and Sluckin model (PSS)

Using a complex order parameter in the de Gennes model has some disadvantages. First, Im(ψ) does not have a physical interpretation. Secondly, this model is formed on a coarse-grained basis, which means that the model does not represent the local free energy density on the length scale of the smectic layers themselves. Therefore, it is suitable for macroscopic calculations, but not for nanoscale calculations of the positions of defects with respect to smectic layers, or the positions of smectic layers with respect to boundaries. To overcome these difficulties, Pevnyi, Selinger and Sluckin presented the following model with director $\vec{\nu} : \Omega \to \mathbb{R}^d$ (still satisfying the constraint

$\vec{\nu} \cdot \vec{\nu} = 1$) and the real-valued density variation $u : \Omega \to \mathbb{R}$ [112],

$$J(u, \vec{\nu}) = \int_\Omega \frac{a}{2}u^2 + \frac{b}{3}u^3 + \frac{c}{4}u^4 + B\left(\nabla\nabla u + q^2 \vec{\nu} \otimes \vec{\nu} u\right)^2 + \frac{K}{2}|\nabla\vec{\nu}|^2, \tag{1.2}$$

where $a,\ b,\ c,\ B,\ q$, and K are constants determined by the liquid crystal under consideration. We are mostly interested in $c > 0$ to keep $\frac{a}{2}u^2 + \frac{b}{3}u^3 + \frac{c}{4}u^4$ bounded from below. Additionally, we choose $a < 0$ to avoid the trivial solution for u. While the PSS model resolves the main issues related to the use of complex order parameters, it still has some difficulties. Note that the PSS model uses the one-constant (simplest) approximation of the Frank-Oseen energy $\frac{K}{2}|\nabla\vec{\nu}|$, but it can be generalized to different Frank constants such as $E_1(\vec{\nu})$ defined in Equation (1.1). A key limitation is that it fails to reproduce so-called "half charge" defects, due to the presence of director discontinuities in these defects where the director field rotates by 180 degrees (1/2 of a full rotation) around a point in the domain [112, Figure 1]. To overcome this, Pevnyi et al. only approximate $\vec{\nu}$ through the tensor field $\boldsymbol{N} = \vec{\nu} \otimes \vec{\nu}$, which allows them to represent half-charge defects. However, numerical difficulties arise when enforcing $\boldsymbol{N}$ to be of the form $\vec{\nu} \otimes \vec{\nu}$ numerically, for some unit vector $\vec{\nu}$ [33].

1.1.3 The Xia et al. model

Ball and Bedford [22] modified the PSS model, replacing $\boldsymbol{N}$ by $\boldsymbol{Q}/s + I_d/d$. Here, $\boldsymbol{Q}$ is a tensor-valued order parameter, s is a scalar order parameter, and I_d is the identity matrix. While existence of minimizers is proved theoretically in [22], practical use of this model leads to numerical difficulties when s is near zero. Xia et al. [138] proposed the alternative model

$$E(u, \boldsymbol{Q}) = \int_\Omega \frac{a}{2}u^2 + \frac{b}{3}u^3 + \frac{c}{4}u^4 + B\left|\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right|^2 + \frac{K}{2}|\nabla\boldsymbol{Q}|^2 + f_n(\boldsymbol{Q}), \tag{1.3}$$

where $f_n(\boldsymbol{Q}) = -l\,\mathrm{tr}(\boldsymbol{Q}^2) + l\left(\mathrm{tr}(\boldsymbol{Q}^2)\right)^2$ for $d = 2$ and $f_n(\boldsymbol{Q}) = -l\,\mathrm{tr}(\boldsymbol{Q}^2) - \frac{l}{3}\,\mathrm{tr}(\boldsymbol{Q}^3) + \frac{l}{2}\left(\mathrm{tr}(\boldsymbol{Q}^2)\right)^2$ in three dimensions. Here, the functions $f_n(\boldsymbol{Q})$ are chosen so that the minimizers of $\int_\Omega f_n(\boldsymbol{Q})$ are of the form $\boldsymbol{Q} = \vec{\nu} \otimes \vec{\nu} - \frac{I_d}{d}$ (see [101, Proposition 15]), and are included in the energy to weakly enforce the rank-one condition implied by Pevnyi et al.'s model, without the potential singularity when including a scalar order parameter, as in [22]. This model still has some difficulties related to existence of the

Hessian, which requires $u \in H^2(\Omega)$. Addressing these difficulties is one of the focuses of this thesis.

1.2 Discretization Methods

1.2.1 Finite-difference methods [111]

Finite-difference methods (FDMs) are one of the oldest and simplest numerical schemes, based on applying a local Taylor expansion to approximate the PDE. Continuous domains are discretized and the PDE is converted into a system of algebraic equations. FDMs are stable, and generally converge rapidly for 1D problems on simple domains. However, FDMs start to have difficulties when one is interested in solving multidimensional PDEs with variable coefficients, and domains with complicated geometry. These difficulties can be overcome using staggered meshes and meshfree finite-difference ideas [121], but these introduce their own complications, with many open research questions for complex systems of equations. A good option to avoid these complications is the use of integral (variational) forms of the PDEs, which leads to the development of finite-volume and finite-element methods.

1.2.2 Finite-volume methods [64]

Finite-volume methods (FVMs) overcome some of the difficulties with FDMs, as they are readily applicable to problems on complex geometries and to PDEs with variable or discontinuous coerfficients. They are based on writing the differential equation in conservative form, i.e $\int_K \nabla \cdot F = \int_K G$ for some functions F and G, and volume K, then converting the volume integral of F into a surface integral using the divergence theorem. These terms are then approximated using approximate fluxes at each surface of the finite volume. These methods are *conservative* as the flux entering a given volume is constrained to be identical to that leaving the adjacent volume through their common face. One common use of FVMs is for time-dependent problems, because of their natural conservative properties. However, finite-element methods can be more accurate when using high-order basis functions and the solution is smooth enough, particularly for time-steady problems and coupled systems of PDEs.

1.2.3 Finite-element methods [35]

Finite-element methods (FEMs) are based on a variational form that is derived from the continuum PDE and used for its discrete approximation (known as conforming methods). The variational formulation is obtained by multiplying the PDE by a test function and then integrating by parts, giving what is called the weak form. At the discrete level, the unknown is approximated using a linear combination of basis functions. Different types of basis functions can be used, depending on the differential operator(s) in the equation (gradients, divergences, and curls), and the order of the PDE. While FEMs overcome some of the given difficulties for FDMs and FVMs, they still have their own difficulties. Deriving stable discretizations for coupled systems is challenging, as many natural discretizations lead to ill-posed discrete problems. Finite-element discretizations are used in this thesis and discussed in more detail in Section (2.2).

1.3 Preconditioners [29]

Applying any of the numerical schemes in the previous subsection leads to the need for solving large ill-conditioned linear and/or nonlinear systems of algebraic equations. Solving the underlying linear systems using direct methods (LU factorization) typically requires significant computational time and memory. On the other hand, iterative methods (for example, Krylov subspace methods) can be very slow to converge due to the high condition numbers of these systems (or their linearizations). Preconditioning means transforming the linear system $\mathcal{A}\vec{u} = \vec{b}$ into another system with better properties, i.e. the preconditioned matrix, $\mathcal{A}\mathcal{P}^{-1}$ or $\mathcal{P}^{-1}\mathcal{A}$, has a (signficantly) smaller condition number, and/or eigenvalues clustered around 1. One might also be interested in the situation where the minimum polynomial of the preconditioned matrix is of small degree, also leading to a situation where the iteration counts that are required to converge within a given tolerance are small. Clever choices of the preconditioner matrix, $\mathcal{P}$, can lead to the development of iterative methods that dramatically outperform direct methods, especially for 3D problems. For linear systems that arise from discretizing complicated PDEs, it is common to use Krylov subspace methods (see Section 2.3), with multigrid preconditioners (see Section 2.4).

1.4 Literature review

This thesis develops finite-element methods for the model in (1.3), along with some simplifications. In this section, we give a brief summary of what has been done in the field of finite-element methods for H^2 elliptic problems, preconditioners for saddle-point systems, and finite-element discretizations (and associated linear and nonlinear solvers) for liquid crystals.

1.4.1 Finite element methods for H^2 elliptic problems

The Euler-Lagrange equations for (1.3) lead to a coupled system of PDEs, with a fourth-order operator applied to u, and a second-order operator acting on Q. Finite-element discretizations for fourth-order H^2-elliptic problems have been studied widely. These include conforming methods, such as the use of Argyris elements, nonconforming methods [41, 52, 131], C^0 interior penalty methods (C0IP) [18, 40, 42, 127], and mixed-finite element methods, including two-field [51, 52, 102], three-field [24, 66], and four-field discretizations [27, 50, 97], and mixed-nonconforming methods (the HHJ mixed formulation) [50, 82, 95, 113]. Here, we focus on developing three field finite-element formulations for H^2-elliptic problems. The goal is to avoid the difficulties encountered when using conforming and C0IP methods. At the same time, we try to avoid using "too many" variables, which makes the arising linear systems very large and more difficult to solve. A key point here is that, to our knowledge, there are no existing methods that offer provably good discretization for (1.3) for which we also have fast solution algorithms. The C0IP methods used in [138] restricted the simulations in that paper to unreasonably low resolutions, particularly for 3D systems, due to the lack of scalable solvers. This motivates the work of this thesis.

1.4.2 Preconditioners for saddle-point systems

Systems arising from the three-field discretizations we develop are of saddle-point type with condition numbers that usually grow like h^{-p} for $p > 0$, resulting in increasingly ill-conditioned systems as the mesh size, h, goes to zero. Therefore, Krylov subspace methods [76] alone for these systems are not efficient. As a result, we have to apply preconditioned Krylov subspace methods. Two common families of preconditioners

are block factorization [59, 69, 105] and monolithic multigrid preconditioners [4, 6, 7, 130]. A key part of this work is to propose effective monolithic multigrid solvers for the arising saddle-point systems [6, 7, 68].

1.4.3 Modeling liquid crystals

Recent years have seen significant and successful effort in developing numerical models of various liquid crystalline materials [4, 5, 9, 23, 34, 53, 100, 110, 112, 114]. In these models, equilibrium states of liquid crystals usually correspond to minimizers of a given energy functional, which can be directly discretized using finite-element (or other) variational techniques. We are interested in smectic-A liquid crystals, which are characterized by their natural propensity to form layers with periodic variation in the density of the liquid crystal along lines orthogonal to the orientation of the crystals. While some models make use of a complex order parameter as a model of the energy of liquid crystals [55], several recent papers have proposed models based directly on the (real-valued) density variation [22, 112, 138]. In this work, we apply mixed finite-element formulations that we have developed to the fourth-order operators that appear in such models of smectic-A liquid crystals, and then solve the corresponding nonlinear systems. For this, we linearize using Newton's method, and solve the arising linear systems using preconditioned Krylov suspace methods.

1.5 Thesis overview

A mathematical model for smectic A liquid crystals based on the free-energy was presented in [138]. Using finite-element methods (or any other discretization method) for finding extremizers of this model is challenging as the optimality conditions lead to a nonlinear coupled multiphysics system with fourth-order operator on the smectic density variable and a second-order operator on the tensor orientation variable. In this thesis, we propose a mathematical modification to the smectic A model that was presented in [138] and prove that the extremizers of both models are equivalent. This modification has been made to allow discretizing the new energy using mixed finite-element methods that offer some advantages. The main goal of these discretizations is the ability to develop efficient preconditioners and, thus, solve the arising nonlinear

systems faster (with cheaper computational cost) than direct methods. This gives us the ability to perform higher resolution simulations, especially in three dimensions, where solving the arising nonlinear systems using direct methods becomes very expensive even for low-resolution simulations. We do this in steps; we focus on developing stable mixed finite-element methods for a range of linear fourth-order problems that are related to the nonlinear fourth-order equation that is a part of the full smectic A model, and provide efficient preconditioners for the arising linear systems in Chapters 3 and (4). Then, we employ these techniques for the ultimate smectic A model in Chapter 5. This thesis is a manuscript-based thesis and contains six chapters organised as follows.

In Chapter 2, we review the mathematical tools and concepts that will be used in the following chapters. We first review general results on Sobolev and finite-element spaces. Then, finite-element methods and the associated well-posedness theory, including conforming, nonconforming, interior penalty, and mixed methods are presented. Finally, we present a brief discussion of iterative solvers.

In Chapter 3, a three-field mixed finite-element method is presented for d-dimensional H^2-elliptic problems with essential boundary enforced weakly using Nitsche-type penalty methods where required. Efficient monolithic-multigrid preconditioners are developed for the resulting saddle-point systems.

While a particular family of fourth-order operator (the biharmonic, Δ^2 form) is considered in Chapter 3, the smectic model of interest includes the Hessian-squared operator, with a wrong-sign shift, making it somehow closer to the Helmholtz operator than the elliptic case. In Chapter 4, we consider the fourth-order PDE in Equation (4.4), with a focus on developing finite-element methods that generate optimal convergence rates in both $q \approx 40$ and $h < 1$. We have developed a nonsymmetric version of conforming and C0IP methods, as well as a mixed finite-element method similar to the one proposed in Chapter 3 which allows the construction of similar preconditioners.

In Chapter 5, we discretize the smectic model of interest, using the mixed formulations proposed in Chapter 4. In addition, a Nested Iteration-Newton-Krylov-Multigrid solver for the arising nonlinear systems is presented.

Finally, in Chapter 6, we present conclusions and some directions for future work.

Chapter 2

Background

2.1 Continuous and discrete function spaces

2.1.1 Sobolev spaces

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ be an open, connected set with Lipschitz boundary, and $\Gamma \subset \partial\Omega$. We use the following standard Sobolev spaces [32, 63]. The space of square integrable functions is denoted

$$L^2(\Omega) := \left\{ u \middle| \int_\Omega |u|^2 = ||u||^2_{0,\Omega} < \infty \right\}.$$

For integer $m \geq 0$, the Sobolev space $H^m(\Omega)$ is defined as

$$H^m(\Omega) := \left\{ u \middle| \ D^\alpha u \in L^2(\Omega), \forall |\alpha| \leq m \right\},$$

where the weak derivative $D^\alpha u = v$ is defined by $\int_\Omega u \ D^\alpha \phi \ = (-1)^{|\alpha|} \ \int_\Omega v\phi$, for all test functions $\phi \in C_c^\infty(\Omega)$, with $|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_d$ for the multi-index $\alpha = (\alpha_1, \ \alpha_2, ..., \ \alpha_d)$ with non-negative integer entries. In this space, the $H^k(\Omega)$ seminorm is defined as

$$|u|^2_{k,\Omega} := \sum_{|\alpha|=k} \|D^\alpha u\|^2_{0,\Omega}, \qquad k = 0, 1, ..., m,$$

and the $H^m(\Omega)$ norm is given by

$$||u||^2_{m,\Omega} := \sum_{k\leq m} |u|^2_{k,\Omega}.$$

Since the domain, Ω, is generally fixed in what we consider, we typically drop it from the norm subscript. For $m = 1$, we introduce the trace operator $H^1(\Omega) \to H^{1/2}(\partial\Omega)$, which is surjective and the norm of $H^{1/2}(\partial\Omega)$ is defined by

$$\|u\|_{1/2,\partial\Omega} = \inf_{w\in H^1(\Omega),\ w=u \text{ on } \partial\Omega} \|w\|_1.$$

For $m > 1$ and $0 \leq i \leq m-1$, similar trace operators can be defined, but with compatibility conditions if the Lipschitz domain Ω has corners. For better understanding of the compatibility conditions, see for example [74, Remark 1.1], where the case when $m = 2$ is discussed.

The subspace $H^m_\Gamma(\Omega)$ is defined as

$$H^m_\Gamma(\Omega) := \left\{ u \;\middle|\; u \in H^m(\Omega) \;\text{ s.t }\; \frac{\partial^i u}{\partial n^i} = 0 \;\text{ on }\; \Gamma, \forall i = 0, 1, 2, ..., m-1 \right\}.$$

Given that $\Gamma = \partial\Omega$, then $H^m_{\partial\Omega}(\Omega)$ is the closure of $C^\infty_0(\Omega)$ in $H^m(\Omega)$, where $C^\infty_0(\Omega)$ denotes the space of infinitely differentiable functions of compact support in Ω. In addition, if $\sigma \in (0, 1)$, we define the space $H^{m+\sigma}(\Omega)$ as follows [77]

$$H^{m+\sigma}(\Omega) = \{u \in H^m(\Omega)|\ \|u\|_{m+\sigma,\Omega} < \infty\},$$

where

$$\|u\|^2_{m+\sigma,\Omega} = \|u\|^2_{m,\Omega} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{\left(D^\alpha u(x) - D^\alpha u(y)\right)^2}{|x-y|^{d+2\sigma}},$$

The vector-valued function space $H(\text{div};\Omega)$ is defined as

$$H(\text{div};\Omega) := \left\{ \vec{v} \;\middle|\; \vec{v} \in [L^2(\Omega)]^d, \nabla\cdot\vec{v} \in L^2(\Omega) \right\},$$

with the norm

$$||\vec{v}||^2_{\text{div},\Omega} := ||\vec{v}||^2_{0,\Omega} + ||\nabla \cdot \vec{v}||^2_{0,\Omega}.$$

The normal component $\vec{v} \cdot \vec{n} \in H^{-1/2}(\partial\Omega)$, the dual space of $H^{1/2}(\partial\Omega)$, with the norm

$$\|\vec{v} \cdot \vec{n}\|_{-1/2,\partial\Omega} = \sup_{u \neq 0 \in H^{1/2}(\partial\Omega)} \frac{< u\vec{v} \cdot \vec{n} >}{\|u\|_{1/2,\partial\Omega}}.$$

The subspace $H_0^{\Gamma}(\text{div}; \Omega)$ is given by

$$H_0^{\Gamma}(\text{div}; \Omega) := \{\vec{v} \mid \vec{v} \in H(\text{div}; \Omega),\ \vec{v} \cdot \vec{n} = 0 \ \text{ on } \ \Gamma\}.$$

The vector-valued function space $H(\text{curl}; \Omega)$ is defined as

$$H(\text{curl}; \Omega) := \left\{ \vec{v} \;\middle|\; \vec{v} \in \left[L^2(\Omega)\right]^d, \nabla \times \vec{v} \in L^2(\Omega) \right\},$$

where in $2D$, $\nabla \times \vec{v} = v_{2x} - v_{1y}$, and its norm is given by

$$||\vec{v}||^2_{\text{curl},\Omega} := ||\vec{v}||^2_{0,\Omega} + ||\nabla \times \vec{v}||^2_{0,\Omega}.$$

The tangential component $\vec{v} \times \vec{n} \in [H^{-1/2}(\partial\Omega)]^3$, for $\{d = 3\}$, and one can use the trace results for $H(\text{div}; \Omega)$ to obtain results for $H(\text{curl}; \Omega)$ in the 2D case. The subspace $H_0^{\Gamma}(\text{curl}; \Omega)$ has the form

$$H_0^{\Gamma}(\text{curl}; \Omega) := \{\vec{v} \mid \vec{v} \in H(\text{curl}; \Omega),\ \vec{v} \times \vec{n} = 0 \ \text{ on } \ \Gamma\}.$$

Remark 2.1.1. If $\Gamma = \partial\Omega$, it is common to write $H_0^m(\Omega),\ H_0(\text{div}; \Omega),\ H_0(\text{curl}; \Omega)$ instead of $H^m_{\partial\Omega}(\Omega),\ H_0^{\partial\Omega}(\text{div}; \Omega),\ H_0^{\partial\Omega}(\text{curl}; \Omega)$.

2.1.2 Finite-element spaces

Let $\Omega \subset \mathbb{R}^d$ be a bounded, Lipschitz, and connected domain, and let $\{\tau_h\}$ be a quasiuniform family of triangular meshes of Ω, with $0 < h < 1$. For $T \in \tau_h$, $\mathcal{P}_k(T)$ is the space of multivariate polynomials of degree at most k on T. Then, the space of continuous Lagrange elements $CG_k(\Omega, \tau_h) \subset H^1(\Omega)$, $k \geq 1$ is defined as

$$CG_k(\Omega, \tau_h) = \left\{ u \in H^1(\Omega),\ u|_T \in \mathcal{P}_k(T),\ \forall T \in \tau_h \right\}.$$

Note that functions in this space are continuous across each edge of the triangulation. The space of discontinuous Lagrange elements $DG_k(\Omega, \tau_h) \subset L^2(\Omega)$, $k \geq 0$ is defined as

$$DG_k(\Omega, \tau_h) = \left\{ u \in L^2(\Omega),\ u|_T \in \mathcal{P}_k(T),\ \forall T \in \tau_h \right\}.$$

Unlike the continuous Lagrange elements, the degrees of freedom of discontinuous Lagrange elements are considered to be internal. That is, functions in $DG_k(\Omega, \tau_h)$ do not necessarily possess C^0 continuity across each edge in τ_h. We also consider the space of Crouzeix-Raviart elements of first order,

$$CR_1 = \left\{ u \in L^2(\Omega), u|_T \in \mathcal{P}_1(T), u \text{ is continuous at the midpoints of each edge } \epsilon \in \epsilon_h \right\}.$$

Note that $CG_1(\Omega, \tau_h) \subset CR_1(\Omega, \tau_h) \subset DG_1(\Omega, \tau_h)$. We also consider the Raviart-Thomas family, $RT_k(\Omega, \tau_h) \subset H(\mathrm{div}; \Omega)$, $k \geq 1$, that is defined by

$$RT_k(\Omega, \tau_h) = \left\{ \vec{v} \in H(\mathrm{div}; \Omega),\ \vec{v}_T \in [\mathcal{P}_{k-1}]^d + \mathcal{P}_{k-1}(T)\vec{x}, \forall T \in \tau_h \right\}.$$

The $H(\mathrm{div})$-conformity of Raviart-Thomas elements requires that normal components of functions in $RT_k(\Omega, \tau_h)$ are continuous across element faces. We point out that, while the lowest-order Raviart-Thomas element is sometimes denoted $RT_0(\Omega, \tau_h)$, we follow the alternate notation (cf. [92]), where the lowest-order element is denoted $RT_1(\Omega, \tau_h)$, with the property that $RT_k(\Omega, \tau_h) \subset [DG_k(\Omega, \tau_h)]^d$. Finally, we consider subspaces that impose Dirichlet boundary conditions on $\Gamma \subset \partial\Omega$,

$$CG_k^\Gamma(\Omega, \tau_h) = CG_k(\Omega, \tau_h) \cap H_\Gamma^1(\Omega), \quad RT_k^\Gamma = RT_k(\Omega, \tau_h) \cap H_0^\Gamma(\mathrm{div}; \Omega)$$

and $CR_1^\Gamma(\Omega, \tau_h) = \{ u_h \in CR_1(\Omega, \tau_h),\ u_h = 0 \text{ on } \Gamma \}$.

2.2 Variational Formulation of Elliptic Boundary Value Problems

Definition 2.2.1. [35,41] Let V be a Hilbert space. The bilinear form $a : V \times V \to \mathbb{R}$ is said to be continuous if there exists a constant $0 < c_1 < \infty$ such that

$$|a(u, v)| \leq c_1 ||u||_V ||v||_V, \quad \forall u, v \in V, \tag{2.1}$$

and coercive on V if there exists $0 < c_2 < \infty$ such that

$$a(u,u) \geq c_2 ||u||_V^2, \quad \forall u \in V. \tag{2.2}$$

Theorem 1. *[35, 41] Assume that (H, (·,·)) is a Hilbert space, V is a closed suspace of H, and a(·,·) is a continuous bilinear form. If a(·,·) is coercive on V, then the problem*

$$a(u,v) = L(v), \quad \forall v \in V, \tag{2.3}$$

has a unique solution for any continuous linear operator, $L : V \to \mathbb{R}$.

2.2.1 Conforming finite-elements methods

Definition 2.2.2. [41] Let $V_h \subset V$ be a finite-dimensional space. Consider (2.3) restricted to V_h, that is finding $u_h \in V_h$ such that

$$a(u_h, v) = L(v), \quad \forall v \in V_h, \tag{2.4}$$

The solution u_h is called the Ritz-Galerkin Approximation.

Theorem 2. *[41] Under the same conditions as Theorem 1, Problem* (2.4) *has a unique solution.*

Remark 2.2.1. [32, 41] When $a(\cdot,\cdot)$ is symmetric, the solution u_h of Problem (2.4) is a minimizer of the quadratic functional $J(v) = \frac{1}{2}a(v,v) - L(v)$ over $v \in V_h$.

Example 1. *Given $f \in L^2(\Omega)$, consider the minimization problem*

$$\inf_{v \in H_0^1(\Omega)} \left(\frac{1}{2} \int_\Omega \nabla v \cdot \nabla v - \int_\Omega fv \right). \tag{2.5}$$

The solution of this minimization problem can be characterized by: $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega). \tag{2.6}$$

u is the solution of the Poisson equation, $-\Delta u = f$, with $u = 0$ on $\partial\Omega$, in the weak sense.

Example 2. *Given $f \in L^2(\Omega)$, consider the solution characterized by: $u_h \in CG_k^{\partial\Omega}(\Omega, \tau_h)$ such that*

$$\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h, \quad \forall v_h \in CG_k^{\partial\Omega}(\Omega, \tau_h), \tag{2.7}$$

This defines u_h as the Ritz-Galerkin approximation of the solution u in Example (1).

Lemma 1 (Céa's Lemma [41])**.** *Under assumptions of Theorem 1, if u and u_h are the solutions of* (2.3) *and* (2.4) *respectively, then the following estimate holds,*

$$\|u - u_h\|_V \leq \frac{c_1}{c_2} \min_{v \in V_h} \|u - v\|_V, \tag{2.8}$$

where c_1 is the continuity constant and c_2 is the coercivity constant of a.

Corollary 1. *If u and u_h are the solutions of* (2.6) *and* (2.7) *respectively, and $u \in H^{k+1}(\Omega)$, then, $\exists c > 0$ such that*

$$\|u - u_h\|_1 \leq c h^k |u|_{k+1}. \tag{2.9}$$

The benefit of applying conforming methods is, in general, that convergence is guaranteed by straight-forward arguments. However, for higher-order PDEs, complicated finite-element spaces are needed to ensure conformity. For example, to solve fourth-order problems, the finite-element space should be a subspace of the Sobolev space $H^2(\Omega)$. This requires the use of C^1-continuous elements, which can only be realized on simplices with a high number of degrees of freedom per element, requiring multivariate polynomials of degree $2^d + 1$ for d-dimensional problems [52, 141]. Conforming methods become even more complicated for PDEs with orders higher than four, because of the complexity needed in the resulting finite element-spaces. Furthermore, strongly implementing essential boundary conditions becomes difficult using such elements and, therefore, weakly imposing boundary conditions using penalty or Nitsche-type methods becomes necessary.

2.2.2 Nonconforming finite-elements methods [13, 123, 126]

In nonconforming methods, the finite-dimensional space V_h used to define the Ritz-Galerkin Approximation in Definition 2.2.2 is not required to be a subspace of V.

Recall the variational form of the Laplace equation in Example 1, to find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = L(v), \quad \forall v \in H_0^1(\Omega),$$

where $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$, and $L(v) = \int_\Omega fv$. Let $V_h = CR_1^{\partial\Omega}(\Omega,\tau_h) \not\subseteq H_0^1(\Omega)$. Multiplying the differential equation by a test function, $v_h \in CR_1^{\partial\Omega}$, and integrating by parts over each triangle T gives

$$\int_\Omega f v_h = \sum_{T\in\tau_h} \int_T \nabla u \cdot \nabla v_h - \sum_{T\in\tau_h} \int_{\partial T} v_h \nabla u \cdot \vec{n}_T.$$

This gives us the weak form $a_h(u,v_h) = \int_\Omega f v_h + E_h(u,v_h)$, where

$$a_h(u,v_h) = \sum_{T\in\tau_h} \int_T \nabla u \cdot \nabla v_h, \quad E_h(u,v_h) = \sum_{T\in\tau_h} \int_{\partial T} v_h \nabla u \cdot \vec{n}_T,$$

where $E_h(u,v_h)$ quantifies how the exact solution fails to satisfy the finite-element equations, defining a kind of consistency error. The nonconforming finite-element method is to find $u_h \in CR_1^{\partial\Omega}(\Omega,\tau_h)$ such that

$$a_h(u_h,v_h) = \int_\Omega f v_h,$$

where the term $E_h(u_h,v_h)$ is omitted from the weak form because it is $\mathcal{O}(h)$ as can be seen in Inequality (2.10). Define the norm on $CR_1(\Omega,\tau_h)$

$$\|u_h\|_{1,h}^2 = \sum_{T\in\tau_h} \int_T \nabla u_h \cdot \nabla u_h, \quad \forall u_h \in CR_1(\Omega,\tau_h).$$

Note that if $\|u_h\|_{1,h} = 0$, then u_h is a piecewise constant. Since it is continuous at the midpoint of each edge, it is globally constant, and since it vanishes at the midpoint of each boundary edge, it vanishes altogether, showing that $\|u_h\|_{1,h}$ is, indeed, a norm on $CR_1(\Omega,\tau_h)$. The bilinear form $a_h(\cdot,\cdot)$ is continuous and coercive in the $||.||_{1,h}$ norm, and given $u \in H^2(\Omega) \cap H_0^1(\Omega)$, $\exists k > 0$ such that the consistency error satisfies

$$|E(u,v_h)| \le kh||u||_2||v_h||_{1,h}, \quad \forall v_h \in CR_1^{\partial\Omega}(\Omega,\tau_h). \tag{2.10}$$

Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$. For two adjacent elements, T_i and T_j, with $e = T_i \cap T_j$, we have $\nabla u \cdot \vec{n}_i = -\nabla u \cdot \vec{n}_j$, where $\vec{n}_i$ and $\vec{n}_j$ are the outward normal vectors to $T_i \cap e$ and

$T_j \cap e$. Since a general function $v_h \in CR_1^{\partial\Omega}(\Omega, \tau_h)$ is only continuous at the midpoint of e, $E(u, v_h)$ does not vanish. In contrast, if v_h is continuous at e, then

$$\int_e v_h \nabla u \cdot \vec{n}_i + \int_e v_h \nabla u \cdot \vec{n}_j = 0.$$

In particular, $E(u, v_h) = 0$ if $v_h \in H^1(\Omega)$. With the above observations, we can prove the following result.

Theorem 3. *[123] Let $u \in H^2(\Omega)$ solve Poisson's equation and $u_h \in CR_1^{\partial\Omega}(\Omega, \tau_h)$ be the nonconforming finite-element approximation. Then, there exists positive constants, c_1, and c_2, such that,*

$$||u - u_h||_{1,h} \leq c_1 h ||u||_2, \quad ||u - u_h||_0 \leq c_2 h^2 ||u||_2.$$

While nonconforming elements often have simpler definitions than conforming ones, many examples are low-order elements that, therefore, fail to give good approximations to sufficiently smooth solutions. Despite their simplicity, these elements can be complex to implement, and their analysis requires analysis of the consistency error, which sometimes implies suboptimal convergence when the consistency error is larger than the interpolation error of the element. Well-known non-conforming finite elements for second and fourth-order PDEs are Crouziex-Raviart and Morley elements, respectively. We also point out that essential boundary conditions of second-order problems can be strongly imposed by Crouziex-Raviart elements, but no known method exists for implementing essential boundary conditions strongly using Morley elements. Thus, Nitsche-type (or other) penalty methods are required.

2.2.3 Interior penalty methods [117]

Another family of methods that can be classified as nonconforming are the discontinuous Galerkin (interior penalty) methods. Recalling the Laplace equation from Example 1. At the discrete level, the discontinuous Galkerin weak form is to find $u_h \in DG_k(\Omega, \tau_h)$ such that

$$a_\lambda(u_h, v_h) = L(v_h), \quad \forall v_h \in DG_k(\Omega, \tau_h), \tag{2.11}$$

where

$$\begin{aligned} a_\lambda(u, v_h) &= \sum_{T\in\tau_h}\int_T \nabla u_h\cdot\nabla v_h - \sum_{e\in\epsilon_h\setminus\partial\Omega}\int_e \{\nabla u_h\cdot\vec{n}\}[\![v_h]\!] + \lambda\sum_{e\in\epsilon_h\setminus\partial\Omega}\int_e \{\nabla v_h\cdot\vec{n}\}[\![u_h]\!] \\ &+ \frac{\rho}{h^\beta}\sum_{e\in\epsilon_h\setminus\partial\Omega}\int_e [\![u_h]\!][\![v_h]\!] + \frac{\rho}{h^\beta}\int_{\partial\Omega} u_h v_h, \\ L(v_h) &= \int_\Omega f v_h, \end{aligned}$$

where β is a positive number that depends on the problem dimension, d, and $\lambda \in \{-1, 0, 1\}$. Note that a_λ is symmetric for $\lambda = -1$ and nonsymmetric otherwise. For two adjacent elements T_i and T_j with a common side, there are two traces of a function in $DG_k(\Omega, \tau_h)$ along $e = T_i \cap T_j$. We add/subtract these traces to obtain the average/jump for a test function v_h, defining

$$\{v_h\} = \frac{1}{2}(v_h|_{T_i^e}) + \frac{1}{2}(v_h|_{T_j^e}), \quad [\![v_h]\!] = (v_h|_{T_i^e}) - (v_h|_{T_j^e})$$

When evaluating such terms for a normal flux, we assume that the normal vector $\vec{n}_e$ is oriented outward from T_i to T_j for $i < j$ when evaluting $\nabla u_h\cdot\vec{n}_e|_{T_i^e}$.

Theorem 4. *Let $k \geq 1$, $u \in H^{k+1}(\Omega)$ be the solution of Example (1) and $u_h \in DG_k(\Omega, \tau_h)$ be the solution of (2.11) with $\lambda = -1$. For large enough ρ and $\beta(d-1) \geq 1$, we have the estimate*

$$\|u - u_h\|_{IP} \leq Ch^k\|u\|_{k+1},$$

where

$$\|u_h\|_{IP}^2 = \sum_{T\in\tau_h}\int_T \nabla u_h\cdot\nabla u_h + \frac{\rho}{h^\beta}\sum_{e\in\epsilon_h\setminus\partial\Omega}[\![u_h]\!]^2 + \frac{\rho}{h^\beta}\int_{\partial\Omega} u_h^2.$$

While interior penalty methods are attractive, especially when generalized for high-order PDEs, the penalty terms worsen the condition number of the linear systems arising from these discretizations, which makes providing efficient preconditioners for such systems more challenging.

2.2.4 Saddle-point problems

Saddle-point problems arise in many areas of computational science and engineering. Most importantly to us, they naturally arise in the context of mixed finite element approximations of elliptic PDEs [29]. Let V and Q be Hilbert spaces, and given $f \in V'$ and $g \in Q'$, where V' and Q' are the dual spaces of V and Q respectively, we consider the saddle-point problem of finding $(u, p) \in V \times Q$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= (f, v)_{V \times V'}, \quad \forall v \in V \\ b(u, q) &= (g, q)_{Q \times Q'}, \quad \forall q \in Q. \end{aligned} \tag{2.12}$$

Which also can be written as

$$Au + B^T p = f, \quad \text{in } V', \tag{2.13}$$

$$Bu = g, \quad \text{in } Q', \tag{2.14}$$

where the linear operators $A : V \to V'$, $A^T : V \to V'$, $B : V \to Q'$ and $B^T : Q \to V'$ satisfy the following

$$(Au, v)_{V' \times V} = (u, A^T v)_{V \times V'} = a(u, v), \quad \text{and } (Bv, q)_{Q' \times Q} = (v, B^T q)_{V \times V'} = b(u, q).$$

For symmetric bilinear form a, Problem (2.12) finds a minimizer for energy of a physical system subject to a set of constraints. In this case, the variable p plays the role of a Lagrange multiplier, and its computation is of interest especially in the mixed finite-element context. Laplace problem, $-\Delta u = f$ with $u = 0$ on $\partial\Omega$ and $f \in L^2(\Omega)$, can be seen as a system of first-order problems, writing

$$\vec{v} - \nabla u = 0, \quad \text{and } \nabla \cdot \vec{v} = f,$$

Multiplying be the relevant test functions and integrating by parts, the mixed Poisson problem finds $(u, \vec{v}) \in L^2(\Omega) \times H(\text{div}; \Omega)$ such that

$$\int_\Omega \vec{v} \cdot \vec{\psi} + u \nabla \cdot \vec{\psi} = 0, \quad \forall \vec{\psi} \in H(\text{div}; \Omega) \tag{2.15}$$

$$\int_\Omega \phi \nabla \cdot \vec{v} = \int_\Omega f \phi, \quad \forall \phi \in L^2(\Omega), \tag{2.16}$$

the Lagrange multiplier u is very important, as it is the solution of Poisson's equation. More general forms of Saddle-point problems can be found in [29].

Theorem 5. *[32] Let bilinear forms $a: V\times V \to \mathbb{R}$ and $b: V\times Q \to \mathbb{R}$ be continuous forms. Assume also that the range of the operator B associated with b is closed in Q'. If $a(\cdot,\cdot)$ is coercive on $Ker(B)$ and there exists a positive constant γ such that*

$$\sup_{v\in V} \frac{b(v,q)}{||v||_V} \geq \gamma ||q||_Q, \quad \forall q \in Q \tag{2.17}$$

then the pair (u,p) is a unique solution of Problem (2.12).

For approximation purposes, we choose $V_h \subset V$, $Q_h \subset Q$ to be finite-dimensional subspaces of V and Q respectively. Let the bilinear forms a and b be restricted to $V_h \times V_h$ and $V_h \times Q_h$, then the pair (u_h, p_h) that solves

$$a(u_h, v_h) + b(v_h, p_h) = (f, v_h), \quad \forall v_h \in V_h, \tag{2.18}$$

$$b(u_h, q_h) = (g, p_h), \quad \forall q_h \in Q_h. \tag{2.19}$$

is an approximation of problem (2.12).

Theorem 6. *[45] Let (u,p) and (u_h,p_h) be the solutions of* (2.12) *and* (2.19) *respectively. If the assumptions of Theorem 5 are satisfied at the discrete level, then the following error estimate holds,*

$$||u-u_h||_V \leq \hat{c} \inf_{v_h\in V_h} ||u-v_h||_V + \tilde{c} \inf_{q_h \in Q_h} ||p-q_h||_Q, \tag{2.20}$$

where $\hat{c} \leq (\frac{1+||a||}{k_1})(\frac{1+||b||}{k_2})$ and $\hat{c} \leq \frac{||b||}{k_1}$. In addition,

$$||p-p_h||_Q \leq \left(1+\frac{||b||}{k_2}\right) \inf_{q_h\in Q_h} ||p-q_h||_Q + \frac{||a||}{k_2} \inf_{v_h \in V_h} ||u-v_h||_V. \tag{2.21}$$

A better estimate is satisfied if $Ker(B_h) \subset Ker(B)$, with

$$||u-u_h||_V \leq \hat{c} \inf_{v_h\in V_h} ||u-v_h||_V. \tag{2.22}$$

When solving a mixed finite-element approximation of an elliptic PDE, the positive definite linear systems that arise from conforming and nonconforming methods

are replaced by saddle-point systems. Developing discrete spaces that inherit the continuum inf-sup condition for each problem is difficult, and smart choices for the spaces at the discrete level have to be made. In particular, if V_h is not rich enough, the inf-sup condition (2.17) at the discrete level might not be satisfied, leading to an unstable discretization. However, enriching V_h more than we should can lead to suboptimal convergence rates because of the unbalanced approximation properties between the discrete spaces V_h and Q_h. The appropriate choice of such spaces in mixed finite-element discretizations is a major theme in the remainder of this thesis.

2.3 Krylov subspace methods

This section is based on notes from [3, 76, 119]. Discretizing linear PDEs leads to linear systems of the form $Ax = b$, and requires solutions of these systems to obtain the finite-element approximation to the solution of the PDE. Gaussian elimination (a *direct* method) is readily applicable for small matrices, but it is inefficient for the huge matrices that arise from many PDE discretizations, because of its expensive computational cost, $\mathcal{O}(n^3)$, where n is the matrix dimension. This cost can be reduced for symmetric positive-definite or banded matrices, but is still very expensive for the discretization matrices that arise from d-dimensional PDEs, for $d \in \{2, 3\}$. Therefore, iterative methods are extensively used to solve linear systems of algebraic equations. This section focuses on general Krylov subspace methods that will be used in the following chapters.

Given an initial guess, x_0, to $x = A^{-1}b$, consider the general polynomial method, defining $x_k = x_{k-1} + \omega_k(b - Ax_{k-1})$, for scalar weights ω_k. The vector x_k can also be expressed as

$$x_k = x_0 + \sum_{i=0}^{k-1} c_i A^i (b - Ax_0),$$

where c_i, $i = 0, 1, ..., k-1$, are constants determined by $\{\omega_t\}_{t=1}^{k}$.

Definition 2.3.1 (Krylov subspace)**.** Given a vector r_0 and a matrix A, the Krylov subspace of dimension m is

$$\mathcal{K}_m(A, r_0) = span\left\{r_0, Ar_0, A^2 r_0, ..., A^{m-1} r_0\right\}.$$

Note that $x_k - x_0 \in \mathcal{K}_k(A, b - Ax_0)$. We note that $\mathcal{K}_m(A, r_0)$ may not be m-dimensional, for example if r_0 is an eigenvector of A or a linear combination of a few eigenvectors. Since we consider the general case where r_0 is assumed to be a linear combination of many more than $m \ll n$ eigenvectors, we still refer to this as the Krylov space of dimension m.

There are several Krylov subspace methods that generate different approximations, x_k, and one can choose the appropriate method depending on the properties of the discretization matrix. For example, the General Minimum Residual (GMRES) method is appropriate for general singular/nonsingular matrices, while the conjugate gradient method can be used only for symmetric and positive-definite matrices. In the next chapters, we will focus on GMRES with variable preconditioning (FGMRES). In the next subsection, we present GMRES preliminaries.

2.3.1 General minimum residual (GMRES)

Within GMRES, we choose x_k to minimize $\|b - Ax_k\|_0$, over $x_k \in x_0 + \mathcal{K}_k(A, r_0)$, where $r_0 = b - Ax_0$. Defining the matrix $P_k = [r_0, Ar_0, A^2r_0, ..., A^{k-1}r_0]$, the vector x_k can be written as

$$x_k = x_0 + P_k y_k, \quad y_k \in \mathbb{R}^k. \tag{2.23}$$

Then, we can define x_k by solving the least-square problem for y_k,

$$\min_{x_k - x_0 \in \mathcal{K}_k(A, r_0)} \|b - Ax_k\|_0 = \min_{y_k} \|r_0 - AP_k y_k\|_0. \tag{2.24}$$

The solution y_k, of Problem (2.24) solves the normal equations

$$P_k^T A^T A P_k y_k = P_k^T A^T r_0, \tag{2.25}$$

Computational difficulties arise when solving (2.25) for y_k, due to ill-conditioning of $P_k^T A^T A P_k$. Instead, the Arnoldi algorithm is usually used to construct an orthonormal basis $\{q_1, q_2, \cdots, q_m\}$ of $\mathcal{K}_m(A, r_0)$ [119, Algorithm 6.1] and one can prove that

$$\text{span}\{q_1, Aq_1, A^2q_1, ..., A^{m-1}q_1\} = \text{span}\{q_1, q_2, q_3, ..., q_m\}, \tag{2.26}$$

for any $m \geq 1$. Define $Q_m = [q_1, q_2, ..., q_m]$, and the upper Hessenberg matrix $\bar{H}_m$,

$$\left(\bar{H}_m\right)_{ij} = \begin{cases} q_i^T A q_j & 1 \leq j \leq m, 1 \leq i \leq \min(j, m), \\ \left\| A q_j - \sum_{i=1}^{j} \left(\bar{H}_m\right)_{ij} q_i \right\| & 1 \leq j \leq m, i = j+1 \\ 0 & \text{otherwise.} \end{cases}$$

These definitions lead to the relation $Aq_j = \sum_{i=1}^{j+1} h_{ij} q_i$, for $1 \leq j \leq m$, which can be rewritten as $AQ_m = \hat{H}_m Q_{m+1}$. Thus, with $q_1 = \frac{1}{\|r_0\|} r_0$, GMRES uses the basis Q_k for $\mathcal{K}_k$ instead of P_k, i.e. we write $x_m = x_0 + Q_k y_k$, and solve the problem

$$\min_{y_k} \|r_0 - AQ_k y_k\|_0 = \min_{y_k} \|r_0 - Q_{k+1} \bar{H}_k y_k\|_0 = \min_{y_k} \|Q_{k+1}^T r_0 - \bar{H}_k y_k\|_0. \tag{2.27}$$

As the first column of Q_{k+1} is $q_1 = \frac{1}{\|r_0\|} r_0$, $Q_{k+1}^T r_0 = \beta e_1^{k+1}$, where e_1^{k+1} is the first column of the identity matrix of size $k+1$. Thus

$$\min_{x_k - x_0 \in \mathcal{K}_k(A, r_0)} \|b - Ax_k\|_0 = \min_{y_k} \|\beta e_1^{k+1} - \bar{H}_k y_k\|_0. \tag{2.28}$$

We solve this problem using the QR factorization of $\bar{H}_k$, rather than the matrix itself, writing $\bar{H}_k = U_k R_k$ for $(k+1) \times (k+1)$ orthogonal matrix, U_k, and $(k+1) \times k$ upper triangular matrix, R_k, giving

$$\min_{x_k - x_0 \in \mathcal{K}_k(A, r_0)} \|b - Ax_k\|_0 = \min_{y_k} \|\beta U_k^T e_1^{k+1} - R_k y_k\|_0. \tag{2.29}$$

Note that the last component of $R_k y_k$ must be zero, since R_k is upper triangular. Thus, we pick y_k so that the first k components of $R_k y_k$ match $\beta U_k^T e_1^{k+1}$. In this case, $\min_{y_k} \|\beta U_k^T e_1^{k+1} - R_k y_k\|_0 = |\eta_{k+1}|$, where η_{k+1} is the last component of $\beta U_k^T e_1^{k+1}$. The GMRES algorithm iterates until $|\eta_{k+1}|$ is small enough to satisfy some given stopping criteria. To reduce the cost of the QR factorization of $\bar{H}_k$, which is $\mathcal{O}(k^3)$, we take advantage of the fact that $\bar{H}_k$ has a nearly upper-triangular form. For example, the

matrix $\bar{H}_5$ has the following structure

$$\bar{H}_5 = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} & h_{15} \\ h_{21} & h_{22} & h_{23} & h_{24} & h_{25} \\ 0 & h_{32} & h_{33} & h_{34} & h_{35} \\ 0 & 0 & h_{43} & h_{44} & h_{45} \\ 0 & 0 & 0 & h_{54} & h_{55} \\ 0 & 0 & 0 & 0 & h_{65} \end{bmatrix}.$$

In order to generate R_k from $\bar{H}_k$, we multiply by a sequence of rotations, where

$$F_k^k F_{k-1}^k F_{k-2}^k \cdots F_1^k \bar{H}_k = R_i$$

$$F_i = \begin{bmatrix} I_1 & & & \\ & c_i & s_i & \\ & -s_i & c_i & \\ & & & I_2 \end{bmatrix},$$

where I_1 and I_2 are the identity matrices of size $i-1$ and $k-1-(i-1)$, $s_i = h_{i+1,i}/\sqrt{(h_{ii}^*)^2 + h_{i+1,i}^2}$, and $c_i = h_{ii}^*/\sqrt{(h_{ii}^*)^2 + h_{i+1,i}^2}$. Here, h_{ii}^* and $h_{i+1,i}$ are the $(i,i)^{\text{th}}$ and $(i+1,i)^{\text{th}}$ components of $F_{i-1}^k F_{i-2}^k \cdots F_1^k \bar{H}_k$. At the i^{th} step, $h_{i+1,i}$ is the same in $\bar{H}_k$ and $F_{i-1}^k F_{i-2}^k \cdots F_1^k \bar{H}_k$ as it is not updated yet. An important observation to reduce the cost is that R_{k+1} can be generated by augmenting F_i^k, $i = \{1, 2, ..., k\}$. That is, if

$$F_k^k F_{k-1}^k \cdots F_1^k = R_k, \tag{2.30}$$

then F_j^{k+1} can be constructed by augmenting F_j^k, for $j = \{1, 2, \cdots, k\}$ by the $(k+1)^{\text{th}}$ row and column of the identity matrix, applying these transformations to the $(k+1)^{\text{th}}$ column of $\bar{H}_{k+1}$ to generate the first k entries of the $(k+1)^{\text{th}}$ column of R_{k+1}. Finally, we can compute F_{k+1}^{k+1}, giving

$$F_{k+1}^{k+1} F_k^{k+1} \cdots F_1^k = R_{k+1}. \tag{2.31}$$

The cost of this is simply $\mathcal{O}(k+1)$, to apply the k known rotations to the last column of $\bar{H}_{k+1}$, plus $\mathcal{O}(1)$ to compute and apply F_{k+1}^{k+1} to a vector. Since we have already computed

$$\beta U_k^T e_1^{k+1} = \beta F_k^k F_{k-1}^k \cdots F_k^1 e_1^{k+1}, \tag{2.32}$$

we only need to extend this by zero to apply F_{k+1}^{k+1} at an $\mathcal{O}(1)$ cost, giving

$$\beta U_{k+1}^T e_1^{k+2} = \beta F_{k+1}^{k+1} F_k^{k+1} \cdots F_{k+1}^1 e_1^{k+2}. \tag{2.33}$$

Thus, the total cost of m steps of GMRES is

$$\left(\sum_{k=1}^{m} c_1 nk + c_2 k + c_3 \right) + c_4 m^2 + c_5 nm,$$

Note that the total cost is $\mathcal{O}(m^2 n)$. In the next chapters, we mainly use preconditioned Krylov subspace methods with multigrid preconditioners to efficiently approximate the solutions arising from our finite-element discretizations. As we use monolithic-multigrid preconditioners for the saddle-point systems arising from our discretizations, we present an introduction to multigrid methods in the following section.

2.4 Multigrid methods

Consider two discretizations of the same problem, on a given "coarse" grid, and a "fine" grid that is a uniform refinement of the coarse grid. Let h and H be the fine and coarse grid discretization parameters, respectively, with $H = 2h$, and let I_H^h and I_h^H be the restriction and prolongation operators, between finite-element spaces on these grids, respectively. To distinguish between discretizations on these two grids, rewrite the fine-grid system $Ax = b$ as $A_h x_h = b_h$, with corresponding coarse-grid system $A_H x_H = b_H$. The two-grid correction scheme [29, 46] is as follows:

1. Relax υ_1 times on $A_h x_h = b_h$ (pre-relaxation),
2. Compute the fine-grid residual $r_h = b_h - A x_h$,
3. Restrict the fine-grid residual $r_H = I_H^h r_h$,
4. Solve the system $A_H e_H = r_H$ using a direct method.
5. Correct the current approximation $x_h = x_h + I_h^H e_H$,
6. Relax υ_2 times on $A_h x_h = b_h$ (post-relaxation).

Here, the choice of the relaxation scheme is critical to achieving an efficient method, with simple schemes (such as weighted Jacobi) being sufficient for simple problems, but more complicated relaxation schemes needed for more difficult problems.

A multigrid V-cycle is obtained by recursively applying the above procedure for the solution of $A_H e_H = r_H$ in Step 4. In the context of finite-element discretizations, we usually choose the restriction and the prolongation operators to be the the natural finite-element operators, but other choices are possible. Similarly, there are two common ways to get the coarse-grid matrix, A_H, either by directly discretizing on the coarse grid (known as rediscretization or the discretization coarse grid approximation, DCGA), or the Galerkin coarse grid approximation (GCGA) where $A_H = I_H^h A_h I_h^H$. Note that GCGA is sometimes advantageous, as A_H can be computed without knowledge of the underlying discretization (as is done in algebraic multigrid (AMG) [46, 119, 129]), but also that both are equivalent in some applications (as they often will be in this thesis. The choice of pre/post-relaxation algorithms usually depends on the matrix A_h (or, equivalently, on the underlying PDE), and should be chosen carefully as using the wrong relaxation can destroy the efficiency of the algorithm. We point out that local Fourier analysis (LFA) can help choose proper components of multigrid methods, and refer to [67, 80] for details.

For symmetric and positive-definite matrices, simple relaxation schemes such as the (weighted) Jacobi, Gauss-Seidel, and Richardson iterations can be used. For the two-dimensional problem in Example 2, an efficient solution scheme is to use the preconditioned conjugate gradient method with a multigrid preconditioner and Jacobi relaxation. Rather than use of Jacobi relaxation, we use two steps of Chebyshev iterations in the multigrid cycle on each level. Table 2.1 shows the number of iterations required for convergence, defined as reducing the Euclidean norm of the residual by a relative factor of 10^{-8} or until its value is below 10^{-8}. Dirichlet boundary conditions are enforced and a right-hand side function is chosen so that the exact solution is given by $u_{ex} = \sin(2\pi x)\cos(3\pi y)$, enabling us to also check that the final approximation is suitably accurate. We note that the number iterations to convergence recorded in Table 2.1 shows no degradation with either decreasing mesh size, h, or increasing polynomial order, k.

Simple relaxation schemes such as those applied in Table 2.1 do not work for saddle-point problems, such as those described in Equation (2.18), simply because

Table 2.1: Number of iterations to converge with preconditioned conjugate gradient and multigrid preconditioner with Jacobi relaxation scheme for Poisson problem in Example 2 for $u \in CG_k(\Omega, \tau_h)$

Order \ h^{-1}	2^6	2^7	2^8	2^9
$k = 1$	7	7	7	7
$k = 2$	7	7	7	7
$k = 3$	8	8	8	8

the main diagonal of the discretization matrix has zero entries and, therefore, when considering matrix splittings of $A = D - L - U$, where D is a diagonal matrix, L is a strict lower-triangular matrix, and U is a strict upper-triangular matrix, the matrices D and $D - L$ are singular. Since this will be the case for many of the linear systems in the chapters to follow, we now discuss how to choose relaxation for saddle-point systems.

2.5 Monolithic-multigrid with Vanka relaxation

Block preconditioners have been extensively studied to approximate the solution of saddle-point problems [29, 59]. Here, however, we are interested in monolithic multigrid preconditioners; in contrast to block preconditioners, which define solvers for approximations to the diagonal blocks of a linear system and/or appropriate Schur complements, monolithic multigrid methods are directly applied to the coupled systems. As mentioned above, standard point-wise relaxation schemes, such as Jacobi and Gauss-Seidel cannot be applied to such systems. Thus, several alternative relaxation schemes have been developed, including Braess-Sarazin schemes [36], Uzawa relaxation [99], distributed relaxation methods [37], and Vanka relaxation which will be of interest here. Vanka relaxation was first proposed by Vanka [130] as a relaxation scheme for nonlinear multigrid for the marker-and-cell finite-difference discretization of the Navier Stokes problem. The key to Vanka relaxation is the use of a "patch-based" relaxation, or overlapping Schwarz iteration, where the problem degrees of freedom are separated into a set of patches, and relaxation is performed by restricting the problem residual to each patch, computing a correction over that patch, then accumulating these corrections globally.

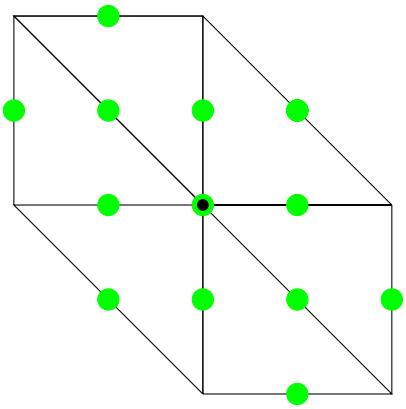

Figure 2.1: Vanka-exclusive patch for the Taylor-Hood ($[CG_2]^2 \times CG_1$) discretization of the Stokes equations. A green disc represents a velocity vector (2 degrees of freedom), while the black disc represents a single pressure degree of freedom.

MacLachlan and Oosterlee [98] observed that Vanka relaxation is naturally generalized to other saddle-point problems, arising from other discretizations and other PDEs. In that framework, the key observation for saddle-point systems is that each patch should contain all degrees of freedom in the linear system that are connected to a single Lagrange multiplier degree of freedom. A standard example of this is for the Taylor-Hood discretization of the Stokes problem [67, 98], where the velocity, $\vec{u}$ is discretized using vector $CG_2(\Omega, \tau_h)$ elements and the pressure is discretized using $CG_1(\Omega, \tau_h)$ elements. In this case, the degrees of freedom for pressure are located at the vertices of the mesh; Figure 2.1 shows the subdomain construction around vertices for this discretization.

Table 2.2 shows the efficiency of flexible GMRES with a monolithic-multigrid preconditioner using Vanka relaxation for the Stokes problem. As choosing relaxation parameters is more complicated in the saddle-point setting, we use two GMRES iterations preconditioned by the Vanka iteration as the pre- and post-relaxation scheme on each level. In later chapters, we primarily use the "vertex star" relaxation scheme [12, 68], which is similar to Vanka, but with patches constructed around each vertex, i, taking all degrees of freedom at vertex i itself, and on edges, faces, and elements directly adjacent to vertex i. In the notation of [68], Vanka relaxation is realized as the (partial) closure of these sets, relative to the grid topology.

Table 2.2: Number of iterations of FGMRES, preconditioned by monolithic multigrid with Vanka relaxation, for convergence of the Stokes problem, with $(\vec{v}_h, p_h) \in [CG_2(\Omega, \tau_h)]^2 \times CG_1(\Omega, \tau_h)$.

Order \ h^{-1}	2^6	2^7	2^8	2^9
$[CG_2]^2 \times CG_1$	12	12	13	13

Chapter 3

Mixed Finite-element methods for H^2 elliptic problems

Abstract[1]

Fourth-order differential equations play an important role in many applications in science and engineering. In this paper, we present a three-field mixed finite-element formulation for fourth-order problems, with a focus on the effective treatment of the different boundary conditions that arise naturally in a variational formulation. Our formulation is based on introducing the gradient of the solution as an explicit variable, constrained using a Lagrange multiplier. The essential boundary conditions are enforced weakly, using Nitsche's method where required. As a result, the problem is rewritten as a saddle-point system, requiring analysis of the resulting finite-element discretization and the construction of optimal linear solvers. Here, we discuss the analysis of the well-posedness and accuracy of the finite-element formulation. Moreover, we develop monolithic multigrid solvers for the resulting linear systems. Two and three-dimensional numerical results are presented to demonstrate the accuracy of the discretization and efficiency of the multigrid solvers proposed.

[1]This work is under revision as "Mixed finite-element methods for H^2 elliptic problems", by Patrick E. Farrell, Abdalaziz Hamdan, and Scott P. MacLachlan, for Computers & Mathematics with Applications, 2022.

3.1 Introduction

Fourth-order differential operators often appear in mathematical models of thin films and plates [62, 104, 112], and pose significant challenges in numerical simulation in comparison to equations governed by more familiar second-order operators. A motivating example arises from modeling equilibrium states of smectic A liquid crystals (LCs), which correspond to minimizers of a given energy functional. For example, the Pevnyi, Selinger, and Sluckin energy functional for smectic A liquid crystals is given by [112]:

$$E(u,\vec{\nu}) = \int_\Omega \frac{a}{2}u^2 + \frac{b}{3}u^3 + \frac{c}{4}u^4 + B\left[\nabla\nabla u + q^2\vec{\nu}\otimes\vec{\nu}\right]^2 + \frac{K}{2}|\nabla\vec{\nu}|^2, \tag{3.1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ is a bounded Lipschitz domain, a, b, q, B, and K are positive real-valued constants determined by the experiment and material under consideration, $\vec{\nu} : \Omega \to \mathbb{R}^d$ is a unit vector field called the director, and $u : \Omega \to \mathbb{R}$ is the smectic order parameter representing the density variation of the LC. This energy is to be minimized subject to the constraint that $\vec{\nu}\cdot\vec{\nu} = 1$ pointwise almost everywhere. When enforcing this constraint with a Lagrange multiplier, the Euler–Lagrange equations for (3.1) lead to a coupled system of PDEs, with a fourth-order operator applied to u, a second-order operator acting on $\vec{\nu}$, and an algebraic constraint.

Motivated by such examples, several families of finite-element methods have been developed to approximate solutions of PDEs with fourth-order terms. In this work, we consider the minimization of a simplified form of the energy (3.1) with suitable boundary conditions, given in variational form as

$$\min_{v\in H^2(\Omega)} \frac{1}{2}\int_\Omega (\Delta v)^2 + c_0\nabla v\cdot\nabla v + c_1 v^2 - \int_\Omega fv, \tag{3.2}$$

with nonnegative constants c_0 and c_1. While the variational formulation in (3.1) is written in terms of the Hessian operator, here we consider the classical fourth-order biharmonic (Laplacian squared) and will consider the Hessian problem (with appropriate boundary conditions) in the following Chapters. Sufficiently smooth extremizers of (3.2) must satisfy its Euler–Lagrange equations, which yield a fourth-order problem,

$$\Delta^2 u - c_0\Delta u + c_1 u = f. \tag{3.3}$$

We consider three-field mixed formulations for this fourth-order problem, with a particular focus on the treatment of the boundary conditions that arise naturally from the transition from the variational to strong forms. These formulations introduce the gradient of the solution as an explicit variable constrained using a Lagrange multiplier. Our approach is general in the sense that we are able to use elements of order k, $k+1$, and $k+1$ for the solution, its gradient, and the Lagrange multiplier respectively, where k can be as large as the smoothness of the solution allows. The existence and uniqueness proofs are not complicated. A drawback here is that our formulation provides suboptimal convergence for some boundary conditions, as discussed below.

If $c_0 = c_1 = 0$, then (3.2) represents the classical biharmonic equation. Many different types of finite-element methods have been considered in this context. Conforming methods, in which the finite-dimensional space is a subspace of the Sobolev space $H^2(\Omega)$, rely on the use of complicated basis functions. These require a high number of degrees of freedom per element, especially in three dimensions. Moreover, the elements are typically not affine equivalent; i.e. the basis functions cannot be mapped to each element using a reference element in the standard way, and more complicated approaches are needed [52, 90, 92]. In order to avoid the use of such C^1 elements, other types of finite elements can be used, leading to nonconforming methods in which the finite-element space is not a subspace of $H^2(\Omega)$, such as Morley and cubic Hermite elements [41, 52, 131]. These elements are also complex to implement and require analysis of the consistency error, which sometimes implies suboptimal convergence when the consistency error is larger than the interpolation error of the element [132].

C^0 interior penalty (C0IP) methods can also be used for fourth-order problems, where the continuity of the function derivatives are weakly enforced using stabilization terms on interior edges [18, 40, 42]. Brenner, Sung, and Zhang [42] solved the problem $\Delta^2 u - \nabla \cdot (\beta(x)\nabla u) = f$, where $\beta(x)$ is a nonnegative C^1 function. Their approach is to find $u \in CG_k(\Omega, \tau_h)$, $k \geq 2$, that satisfies the system

$$a_h(u, \phi) + b_h(u, \phi) + \gamma c_h(u, \phi) = \langle f, \phi \rangle, \ \forall \phi \in CG_k(\Omega, \tau_h), \tag{3.4}$$

where $\gamma > 0$ is a penalty parameter, $CG_k(\Omega, \tau_h)$ is the space of continuous Lagrange

elements of degree k on a triangulation τ_h of the domain Ω, and

$$a_h(u,\phi) = \sum_{T\in\tau_h}\int_T \left(\nabla\nabla u : \nabla\nabla\phi + \beta(x)\nabla u\cdot\nabla\phi\right), \tag{3.5}$$

$$b_h(u,\phi) = \sum_{e\in\epsilon_h}\int_e \left\{\frac{\partial^2 u}{\partial^2 n}\right\}\left[\frac{\partial\phi}{\partial n}\right] + \sum_{e\in\epsilon_h}\int_e \left\{\frac{\partial^2\phi}{\partial^2 n}\right\}\left[\frac{\partial u}{\partial n}\right], \tag{3.6}$$

$$c_h(u,\phi) = \sum_{e\in\epsilon_h}\frac{1}{|e|}\int_e \left[\frac{\partial u}{\partial n}\right]\left[\frac{\partial\phi}{\partial n}\right], \tag{3.7}$$

where $\left\{\frac{\partial u}{\partial n}\right\}$ and $\left[\frac{\partial\phi}{\partial n}\right]$ denote the standard average and jump on each edge. Here, τ_h is the set of cells in a mesh and ϵ_h is the set of edges. While C0IP methods have advantages, such as enabling the use of simple Lagrange elements and the ability to use arbitrarily high-order elements [42], they also have some disadvantages. The weak forms are more complicated than those used for classical conforming and nonconforming methods. Moreover, the need for the penalty parameter is also a drawback, as it is sometimes not trivial to decide how large this parameter must be to achieve stability, especially as parameters in the PDE are varied [108]. Similarly, discontinuous Galerkin approaches can also be applied to this problem [19], augmenting the forms in (3.4) to account for basis functions that do not enforce C^0 continuity across elements. These share the disadvantages of C0IP methods, while requiring more degrees of freedom than C^0 approaches.

Another attractive option to avoid using H^2-conforming methods is mixed finite-element methods, in which the gradient or the Laplacian of the solution are approximated in addition to the solution itself [24,25,27,51,52,97,102]. A natural classification of such mixed finite-element methods is based on how many functions (fields) are directly approximated. Given clamped boundary conditions, where both u and ∇u are prescribed on the boundary, two functions are approximated in [51, 52, 102], both u and either its gradient or its Laplacian. In [52], the biharmonic problem is rewritten as a coupled system of Poisson equations, in which the unknown and its Laplacian are both directly approximated. In [102], the 2D biharmonic problem is approximated by minimizing

$$J(u,\vec{v}) = \frac{1}{2}\|\nabla\vec{v}\|_0^2 + \frac{1}{2\epsilon}\|\rho_0(\vec{v}-\nabla u)\|_0^2 - \langle f,u\rangle, \quad \text{for } 0<\epsilon\le ch^2,$$

where ρ_0 is the orthogonal projection from $[L^2(\Omega)]^2$ to the space of piecewise constant

functions, and the functions u and $\vec{v}$ are approximated using bilinear elements on a rectangular mesh. An error analysis of this method requires the solution to be at least in $H^{4.73}(\Omega)$ [87]. A similar approach solves the d-dimensional biharmonic problem, replacing the L^2 projection onto piecewise constant functions with that onto the space of multilinear vector-valued functions whose i^{th} component is independent of x_i [51]. This approach requires less regularity on the solution, $u \in H^4(\Omega)$, than was required in [102]. These approaches only treat clamped boundary conditions.

A second class of mixed finite-element methods is that of four-field formulations, in which $u,\ \nabla u,\ \nabla^2 u,$ and $\nabla \cdot (\nabla^2 u)$ are directly approximated. In [97], a mixed formulation approximating these fields and its stability in $H_0^1(\Omega) \times [H_0^1(\Omega)]^2 \times \boldsymbol{L}^2_{\mathrm{sym}}(\Omega) \times H^{-1}(\mathrm{div}, \Omega)$ is discussed for $\Omega \subset \mathbb{R}^2$, where $\boldsymbol{L}^2_{\mathrm{sym}}(\Omega)$ is the space of 2×2 symmetric tensors with components in $L^2(\Omega)$, and $H^{-1}(\mathrm{div}; \Omega)$ is the dual space of $H_0(\mathrm{rot}, \Omega) = \{\vec{\psi} \in [L^2(\Omega)]^2 \mid \mathrm{rot}\, \vec{\psi} \in L^2(\Omega),\ \vec{\psi} \cdot \vec{t} = 0 \text{ on } \partial\Omega\}$, where $\vec{t}$ is the unit tangent vector to $\partial\Omega$. A similar approach with different function spaces is given in [27]. This approach, focused on the discrete level, finds $(u_h, \vec{q}_h,\ \bar{z}_h,\ \vec{\sigma}_h) \in DG_k(\Omega, \tau_h) \times [DG_k(\Omega, \tau_h)]^2 \times \boldsymbol{RT}_{k+1}(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h) \subset L^2(\Omega) \times [L^2(\Omega)]^2 \times \boldsymbol{H}(\mathrm{div}, \Omega) \times H(\mathrm{div}; \Omega)$, where $u_h, \vec{q}_h,\ \bar{z}_h,$ and $\vec{\sigma}_h$ are approximations of $u,\ \nabla u,\ \nabla^2 u,$ and $\nabla \cdot (\nabla^2 u)$ respectively, and $\bar{y} \in \boldsymbol{H}(\mathrm{div}; \Omega))$ means that each row of the tensor $\bar{y}$ belongs to $H(\mathrm{div}; \Omega)$. Here $DG_k(\Omega, \tau_h)$ and $RT_k(\Omega, \tau_h)$ denote the discontinuous Lagrange and Raviart–Thomas approximation spaces of order k on mesh τ_h, respectively, with $\boldsymbol{RT}_k(\Omega, \tau_h)$ denoting tensor-valued functions with rows in $RT_k(\Omega, \tau_h)$. However, these four-field formulations lead to discretizations with large numbers of degrees of freedom, posing difficulties in the development of efficient linear solvers.

The third class of mixed finite-element methods is that of three-field formulations [24]. The unknowns here are the function, its gradient, and a Lagrange multiplier. Assuming again homogeneous clamped boundary conditions, these lead to finding the saddle-point $(u, \vec{v}, \vec{\alpha}) \in H_0^1(\Omega) \times H_0(\mathrm{div}; \Omega) \times M$ of the Lagrangian functional

$$\mathcal{L}\big((u, \vec{v}), \vec{\alpha}\big) = \frac{1}{2}\|\nabla \cdot \vec{v}\|_0^2 + \int_\Omega \vec{\alpha} \cdot (\vec{v} - \nabla u) - \int_\Omega f u \tag{3.8}$$

where $M = \{\vec{\alpha} \in H_0(\mathrm{div}; \Omega) \mid \nabla \cdot \vec{\alpha} \in H^{-1}(\Omega)\}$. Here, $H_0(\mathrm{div}; \Omega) := \{\vec{v} \in H(\mathrm{div}; \Omega) \mid \vec{v} \cdot \vec{n} = 0 \text{ on } \partial\Omega\}$. At the discrete level, the method in [24] finds $(u_h, \vec{v}_h,\ \vec{\alpha}_{2h}) \in CG_1(\Omega, \tau_h) \times RT_1(\Omega, \tau_h) \times DG_0(\Omega, \tau_{2h})$, where the Lagrange multiplier $\vec{\alpha}_{2h}$ is constructed in τ_{2h} to guarantee well-posedness at the discrete level and to achieve an

optimal error estimate. Again, this approach only treats clamped boundary conditions, and requires the use of different meshes in the discretization. Moreover, it is not mentioned if the discretization can be generalized to higher orders. Here, we propose a similar three-field formulation, but treating the generalized problem in (3.3) with more general boundary conditions, and using different discretization spaces of arbitrarily high degree. Unlike conforming methods, our approach works effectively in both two and three dimensions. Finally, we point out that some methods merge mixed and nonconforming methods [50, 82, 95, 113] using the HHJ elements. These, however, are restricted to 2D problems.

Strongly imposing essential boundary conditions with some finite-element basis functions is difficult [90]. In addition, it can, sometimes, negatively affect properties of the finite-element method, such as its stability and accuracy [86,88]. Weakly imposing the boundary conditions via a penalty method [16,17] may help. An attractive family of penalty methods are the Nitsche-type methods [109] for which optimal convergence can be achieved. Applications of Nitsche's method to second-order PDEs can be found in [70, 86, 88]. Moreover, Nitsche-type penalty methods have been used to impose essential boundary conditions for some discretizations of the biharmonic and other fourth-order problems [28, 60, 90]. While we are able to impose a variety of boundary conditions directly in our variational formulation, we utilize Nitsche-type penalty methods for a particular case where strong enforcement of the boundary conditions leads to problems establishing inf-sup stability of the discretization.

At the discrete level, the resulting linear system of our three-field formulation is a saddle point system [29], of the form

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U \\ \alpha \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}$$

where U represents discrete degrees of freedom associated with both u and $\vec{v} = \nabla u$, while α represents discrete degrees of freedom associated with $\vec{\alpha}$, leading to matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$ and the zero matrix $0 \in \mathbb{R}^{m\times m}$. In our formulation, A will be symmetric and positive semi-definite. This kind of problem appears in many areas of computational science and engineering [29]. For discretized PDEs, the condition number of such systems usually grows like h^{-k} for $k > 0$, resulting in increasingly ill-conditioned systems as the mesh size, h, goes to zero. This growth of the condition

number leads to slow convergence of unpreconditioned Krylov methods. Therefore, we employ preconditioning in order to develop a mesh-independent algorithm to solve these systems. Two common families of preconditioners are block factorization [59,69, 105] and monolithic multigrid preconditioners [4, 6, 7, 130]. In this work, we propose an effective monolithic multigrid solver for the arising saddle-point systems [6, 7, 68].

This chapter is organized as follows. In Section 3.2, a brief summary is given of the Sobolev and finite-element spaces employed. The weak forms, uniqueness of solutions at the continuum and discrete levels, and an error analysis are presented in Sections 3.3 and 3.4. The monolithic multigrid preconditioner and the details of the linear solver are presented in Section 3.5. Finally, numerical experiments showing the accuracy of the finite-element method and the effectiveness of the linear solver are given in Section 3.6.

3.2 Background

Throughout this paper, we consider $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ to be a bounded, Lipschitz, and connected domain. On a simplex $T \in \tau_h$, all degrees of freedom of the discontinuous Lagrange $DG_k(\Omega, \tau_h)$ element are considered to be internal; i.e., no continuity is imposed by these elements [92]. In contrast, the continuous Lagrange $CG_k(\Omega, \tau_h)$ elements possess full C^0 continuity across element edges. Here, we primarily make use of $DG_k(\Omega, \tau_h)$ approximations of functions in $L^2(\Omega)$. We also consider the Raviart-Thomas $RT_k(\Omega, \tau_h)$ element, which is $H(\text{div})$-conforming, where the normal component is continuous across element faces, and $RT^{\Gamma}_{k+1}(\Omega, \tau_h) = \{\vec{v} \in RT_{k+1}(\Omega, \tau_h) | \vec{v} \cdot \vec{n} = 0 \text{ on } \Gamma \subset \partial\Omega\}$. A standard approximation result for these elements is stated next.

Theorem 7. *[32, 35, 92] Let $I^k_h : H^{k+1}(\Omega) \to DG_k(\Omega, \tau_h)$, $\mathit{\Pi}^k_h : [H^{k+1}(\Omega)]^d \to RT_k(\Omega, \tau_h)$, and $L^k_h : [H^{k+1}(\Omega)]^d \to [CG_k(\Omega, \tau_h)]^d$ be the finite-element interpolation operators. Then there exist constants $\tilde{c}$, c, and $\hat{c}$, such that for any $u \in H^{k+1}(\Omega)$ and $\vec{v} \in \left[H^{k+1}(\Omega)\right]^d$,*

$$\|u - I^k_h u\|_0 \leq \tilde{c} h^{k+1} |u|_{k+1}, \quad \forall k \geq 0, \tag{3.9}$$

$$\|\vec{v} - \mathit{\Pi}^k_h \vec{v}\|_{\text{div}} \leq c h^k \left(|\vec{v}|_k + |\vec{v}|_{k+1}\right), \quad \forall k > 0, \tag{3.10}$$

$$\|\vec{v} - L^k_h \vec{v}\|_1 \leq \hat{c} h^k |\vec{v}|_{k+1}, \quad \forall k > 0. \tag{3.11}$$

An important property of our discretization is that it benefits from the usual mimetic relationships between $RT_{k+1}(\Omega, \tau_h)$ and $DG_k(\Omega, \tau_h)$, summarized in the following results.

Lemma 2. *[12, 15] Assume Ω is simply-connected, and is convex if $d = 3$. Then the Helmholtz decomposition of $RT_{k+1}(\Omega, \tau_h)$ is*

$$RT_{k+1}(\Omega, \tau_h) = \Big(\nabla \times V_h\Big) \oplus \Big(\operatorname{grad}_h DG_k(\Omega, \tau_h)\Big), \tag{3.12}$$

where $\operatorname{grad}_h : DG_k(\Omega, \tau_h) \to RT_{k+1}(\Omega, \tau_h)$ is the discrete gradient operator, defined by

$$\int_\Omega \operatorname{grad}_h u \cdot \vec{v} = -\int_\Omega u \nabla \cdot \vec{v}, \quad \forall \vec{v} \in RT_{k+1}(\Omega, \tau_h).$$

For $d = 2$, $\nabla\times = \begin{bmatrix} -\frac{\partial}{\partial y} \\ \frac{\partial}{\partial x} \end{bmatrix}$ and $V_h = CG_{k+1}(\Omega, \tau_h)$, while $V_h = N^1_{k+1}(\Omega, \tau_h)$ for $d = 3$, where $N^1_{k+1}(\Omega, \tau_h)$ is the Nédélec element of the first kind of order $k + 1$.

Remark 3.2.1. The main idea of relating the spaces in Lemma 2 is that the following sequences are exact in 2D and 3D respectively.

$$0 \to CG_{k+1}(\Omega, \tau_h) \xrightarrow{\nabla\times} RT_{k+1}(\Omega, \tau_h) \xrightarrow{\nabla\cdot} DG_k(\Omega, \tau_h) \to 0,$$

and

$$0 \to CG_{k+1}(\Omega, \tau_h) \xrightarrow{\nabla} \mathrm{N}^1_{k+1}(\Omega, \tau_h) \xrightarrow{\nabla\times} RT_{k+1}(\Omega, \tau_h) \xrightarrow{\nabla\cdot} DG_k(\Omega, \tau_h) \to 0.$$

Remark 3.2.2. [32, 92] $\forall \vec{v} \in RT_{k+1}(\Omega, \tau_h)$, we have $\nabla \cdot \vec{v} \in DG_k(\Omega, \tau_h)$.

While we largely make use of the standard Sobolev norms, we will also use the "strengthened" norm,

$$\|\vec{v}\|^2_{\mathrm{div},\Gamma} = \|\vec{v}\|^2_{\mathrm{div}} + h\|\nabla \cdot \vec{v}\|^2_{0,\Gamma} + \frac{1}{h}\|\vec{v} \cdot \vec{n}\|^2_{0,\Gamma} \tag{3.13}$$

where $\Gamma \subset \partial\Omega$ (to be specified later), and

$$\|\vec{v} \cdot \vec{n}\|^2_{0,\Gamma} = \int_\Gamma |\vec{v} \cdot \vec{n}|^2, \quad \|\nabla \cdot \vec{v}\|^2_{0,\Gamma} = \int_\Gamma |\nabla \cdot \vec{v}|^2.$$

For these norms, the inverse trace inequality below is a useful result.

Theorem 8. *[117, 134] Let $T \in \tau_h$ be a d-simplex of $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. Then, for all $u \in DG_k(T)$,*

$$\|u\|^2_{0,\partial T} \leq \frac{(k+1)(k+d)\ \mathrm{Vol}(\partial T)}{d\ \mathrm{Vol}(T)} \|u\|^2_{0,T}, \tag{3.14}$$

where $\|u\|^2_{0,\partial T}$ is defined as $\|u\|^2_{0,\partial T} = \int_{\partial T} u^2$, and $\mathrm{Vol}(\cdot)$ is the Lebesgue measure.

Corollary 2. *Consider a triangulation τ_h of the domain $\Omega \subset \mathbb{R}^d$, and let $\partial\tau_h := \{T \in \tau_h \mid \partial T \cap \partial\Omega \neq \varnothing\}$. Then,*

$$\forall u_h \in DG_k(\Omega, \tau_h), \|u_h\|^2_{0,\partial\Omega} \leq \gamma_1(k, \tau_h)\|u_h\|^2_0$$
$$\forall \vec{v}_h \in RT_{k+1}(\Omega, \tau_h), \|\vec{v}_h \cdot \vec{n}\|^2_{0,\partial\Omega} \leq \gamma_1(k+1, \tau_h)\|\vec{v}_h\|^2_0,$$

where

$$\gamma_1(k, \tau_h) = \max_{T \in \partial\tau_h} \frac{(k+1)(k+d)\ \mathrm{Vol}(\partial T)}{d\ \mathrm{Vol}(T)}. \tag{3.15}$$

While the ratio between $\mathrm{Vol}(\partial T)$ and $\mathrm{Vol}(T)$ can be arbitrarily large, $\gamma_1(k, \tau_h)$ is readily bounded when we consider quasiuniform families of meshes [41, Definition 4.4.13], where $\mathrm{Vol}(\partial T)$ of each d-simplex T is bounded above by $\mathcal{O}(h^{d-1})$ and $\mathrm{Vol}(T)$ is bounded below by $\mathcal{O}(h^d)$. This naturally leads to an approximation property for the trace norm. These results will be useful in the analysis of the Nitsche boundary integrals. We note that Lemma 4 adds an important restriction, that Ω be polygonal or polyhedral, in order for the error estimate given there to hold. This is required only in Theorem 11 below; for the remainder of the analysis in Section 3.4, we only require that Ω is a bounded, Lipschitz, and connected domain.

Corollary 3. *Let $\{\tau_h\}$, $0 < h \leq 1$ be a family of quasiuniform meshes of the domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$. Then, there exists $C_\Omega > 0$ such that for any τ_h in the family,*

$$\forall u_h \in DG_k(\Omega, \tau_h), \|u_h\|^2_{0,\partial\Omega} \leq \frac{\gamma_1(k)}{h}\|u_h\|^2_0,$$
$$\forall \vec{v}_h \in RT_{k+1}(\Omega, \tau_h), \|\vec{v}_h \cdot \vec{n}\|^2_{0,\partial\Omega} \leq \frac{\gamma_1(k+1)}{h}\|\vec{v}_h\|^2_0,$$

where

$$\gamma_1(k) = C_\Omega(k+1)(k+d) \geq h \max_{T \in \partial\tau_h} \frac{(k+1)(k+d)\ \mathrm{Vol}(\partial T)}{d\ \mathrm{Vol}(T)}, \tag{3.16}$$

for all τ_h. The constant C_Ω is determined by the dimension, d, and the quasiuniformity parameter for the family.

Lemma 3. *Let $\{\tau_h\}$, $0 < h \leq 1$ be a family of quasiuniform meshes of the domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, and $\vec{v} \in \left[H^{k+2}(\Omega)\right]^d$. Then, there exists a constant, m_1, such that*

$$\left\| \left(\vec{v} - \Pi_h^{k+1}\vec{v}\right) \cdot \vec{n} \right\|_{0,\partial\Omega} \leq m_1 h^{k+1/2} |\vec{v}|_{k+2},$$

where Π_h^{k+1} is the natural Raviart-Thomas interpolation operator.

Proof. Applying the triangle inequality yields

$$\left\| \left(\vec{v} - \Pi_h^{k+1}\vec{v}\right) \cdot \vec{n} \right\|_{0,\partial\Omega} \leq \| \left(\vec{v} - L_h^{k+1}\vec{v}\right) \cdot \vec{n}\|_{0,\partial\Omega} + \| \left(L_h^{k+1}\vec{v} - \Pi_h^{k+1}\vec{v}\right) \cdot \vec{n}\|_{0,\partial\Omega},$$

where L_h^{k+1} is the $CG_k(\Omega, \tau_h)$ interpolation operator. Note that the vector-valued function $\vec{v} - L_h^{k+1}\vec{v} \in [H^1(\Omega)]^d$, and therefore, we use the trace theorem [77, Theorem 1.5.1.10],

$$\| \left(\vec{v} - L_h^{k+1}\vec{v}\right) \cdot \vec{n}\|_{0,\partial\Omega} \leq K \|\vec{v} - L_h^{k+1}\vec{v}\|_1,$$

where K is a positive constant independent of h. Also, $L_h^{k+1}\vec{v} - \Pi_h^{k+1}\vec{v} \in RT_{k+2}(\tau_h, \Omega)$. Applying Theorem 7 and Corollary 3,

$$\begin{aligned}
\left\| \left(\vec{v} - \Pi_h^{k+1}\vec{v}\right) \cdot \vec{n} \right\|_{0,\partial\Omega} \leq & K \|\vec{v} - L_h^{k+1}\vec{v}\|_1 + \|L_h^{k+1}\vec{v} - \Pi_h^{k+1}\vec{v}\|_{0,\partial\Omega} \\
\leq & K\hat{c}h^{k+1}|\vec{v}|_{k+2} + \sqrt{\frac{\gamma_1(k+2)}{h}} \|L_h^{k+1}\vec{v} - \Pi_h^{k+1}\vec{v}\|_0 \\
\leq & K\hat{c}h^{k+1}|\vec{v}|_{k+2} + \sqrt{\frac{\gamma_1(k+2)}{h}} \left(\|L_h^{k+1}\vec{v} - \vec{v}\|_0 + \|\vec{v} - \Pi_h^{k+1}\vec{v}\|_0\right) \\
\leq & K\hat{c}h^{k+1}|\vec{v}|_{k+2} + \sqrt{\frac{\gamma_1(k+2)}{h}} (c + \hat{c}) h^{k+1} |\vec{v}|_{k+2}
\end{aligned}$$

where the constants γ_1, c, and $\hat{c}$ are defined in Theorem 7 and Corollary 3. The choice $m_1 = K\hat{c} + (c + \hat{c})\sqrt{\gamma_1(k+2)}$ completes the proof. □

Lemma 4. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ be a bounded polygonal or polyhedral domain with Lipschitz boundary, $\{\tau_h\}$, $0 < h \leq 1$ be a family of quasiuniform meshes of Ω, and $\vec{v} \in [H^{k+2}(\Omega)]^d$, with $1 \leq k \leq 2^d$. Then, there exists a positive constant, m_2, such that*

$$\|\nabla \cdot (\vec{v} - \Pi_h^{k+1}\vec{v})\|_{\partial\Omega} \leq m_2 h^k |\vec{v}|_{k+2}.$$

where Π_h^{k+1} *is the Raviart-Thomas interpolation operator defined in Theorem 7.*

Proof. Define the finite-element spaces $[ARG_{2^d+1}(\Omega,\tau_h)]^d \subset [H^2(\Omega)]^d$, where $\mathrm{ARG}_5(\Omega,\tau_h)$ is the well-known Argyris elements for $d=2$ [52], and $ARG_9(\Omega,\tau_h)$ is the finite-element space with continuously differentiable functions of three variables defined in [141]. Standard interpolation results give an operator, $\pi_h : [H^{k+2}(\Omega)]^d \to [ARG_{2^d+1}(\Omega,\tau_h)]^d$ such that

$$\|\vec{v}-\pi_h\vec{v}\|_1 \le m_3 h^{k+1}|\vec{v}|_{k+2}, \quad \|\vec{v}-\pi_h\vec{v}\|_2 \le m_4 h^k|\vec{v}|_{k+2}. \tag{3.17}$$

Using the triangle inequality yields

$$\|\nabla\cdot(\vec{v}-\Pi_h^{k+1}\vec{v})\|_{\partial\Omega} \le \|\nabla\cdot(\vec{v}-\pi_h\vec{v})\|_{\partial\Omega} + \|\nabla\cdot(\pi_h\vec{v}-\Pi_h^{k+1}\vec{v})\|_{\partial\Omega}.$$

As $\nabla\cdot(\vec{v}-\pi_h\vec{v}) \in H^1(\Omega)$, we apply [77, Theorem 1.5.1.10]. That is,

$$\|\nabla\cdot(\vec{v}-\pi_h\vec{v})\|_{\partial\Omega} \le m_5\|\nabla\cdot(\vec{v}-\pi_h\vec{v})\,\|_1 \le m_5\|\vec{v}-\pi_h\vec{v}\|_2 \le m_6 h^k|\vec{v}|_{k+2}. \tag{3.18}$$

On the other hand, $\nabla\cdot(\pi_h\vec{v}-\Pi_h^{k+1}\vec{v}) \in DG_{2^d}(\Omega,\tau_h)$, and we therefore apply Corollary 3 to yield

$$\begin{aligned}
\|\nabla\cdot(\pi_h\vec{v}-\Pi_h^{k+1}\vec{v})\|_{\partial\Omega} &\le \frac{\sqrt{\gamma_1(2^d)}}{\sqrt{h}}\|\nabla\cdot(\pi_h\vec{v}-\Pi_h^{k+1}\vec{v})\|_0 \\
&\le \frac{\sqrt{\gamma_1(2^d)}}{\sqrt{h}}\left(\|\nabla\cdot(\pi_h\vec{v}-\vec{v})\|_0 + \|\nabla\cdot(\vec{v}-\Pi_h^{k+1}\vec{v})\|_0\right) \\
&\le \frac{m_7\sqrt{\gamma_1(2^d)}}{\sqrt{h}}h^{k+1}|v|_{k+2}.
\end{aligned} \tag{3.19}$$

Finally, combining (3.17)–(3.19) leads to the desired estimate. □

Remark 3.2.3. Lemma 4 can be generalized for any $k \ge 1$ by using higher-order continuously differentiable elements as intermediate elements. We refer to [92, 122] for higher-order Argyis-like elements in 2D, and [141] for the 3D case.

3.3 Continuum Analysis

Consider the fourth-order problem (3.3) with suitable boundary conditions (discussed below),

$$\Delta^2 u - c_0 \Delta u + c_1 u = f \quad \text{in } \Omega, \tag{3.20}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ is a bounded, Lipschitz, and connected domain with outward pointing normal $\vec{n}$, and c_0 and c_1 are nonnegative constants. Define $V = \{v \in H^1(\Omega) \mid \Delta v \in L^2(\Omega)\}$ with dual space V^*, and assume that $f \in V^*$. Multiplying by a test function, $v \in V$, integration by parts yields

$$\int_\Omega \Delta u \Delta v + c_0 \nabla u \cdot \nabla v + c_1 uv + \int_{\partial\Omega} v(\nabla \Delta u - c_0 \nabla u) \cdot \vec{n} - \int_{\partial\Omega} \Delta u \nabla v \cdot \vec{n} = \int_\Omega fv. \tag{3.21}$$

Since (3.20) is a fourth-order problem, we require two boundary conditions on any segment of $\partial\Omega$. Here, we focus on the boundary conditions that arise from the integration by parts (3.21):

$$u = 0, \quad \Delta u = 0, \quad \text{on } \Gamma_0, \tag{3.22}$$

$$u = 0, \quad \frac{\partial u}{\partial n} = 0, \quad \text{on } \Gamma_1, \tag{3.23}$$

$$\frac{\partial(\Delta u - c_0 u)}{\partial n} = 0, \quad \Delta u = 0, \quad \text{on } \Gamma_2, \tag{3.24}$$

$$\frac{\partial(\Delta u - c_0 u)}{\partial n} = 0, \quad \frac{\partial u}{\partial n} = 0, \quad \text{on } \Gamma_3. \tag{3.25}$$

Note that we consider homogeneous boundary conditions here, but the results hold true for nonhomogeneous boundary conditions if the traces of these quantities are smooth enough on $\partial\Omega$, using standard techniques (cf. [74]) to transform the inhomogeneous boundary conditions (3.22)-(3.25) into homogeneous ones. We note that the commonly-considered case of clamped boundary conditions corresponds to Γ_1 in this classification. In contrast, Γ_0 differs from the classical "simply supported" plate boundary conditions, although is equivalent since the tangential derivative of u along Γ_0 is equal to zero [41, Section 5.9]. Furthermore, writing $\partial\Omega = \Phi \cup (\cup_i \Gamma_i)$, where each Γ_i is open, and $\Phi = \cup_{i \neq j} \left(\bar{\Gamma}_i \cap \bar{\Gamma}_j\right)$, we assume that Φ provides a piecewise C^1 dissection of $\partial\Omega$ [75, Definition 2.2]. Roughly speaking, this requires that Φ is the union of closed curves that are piecewise C^1 [81]. Under suitable assumptions on

c_0 and c_1, we can prove a variety of results on the well-posedness of the resulting variational problems in the standard Hilbert space setting.

Lemma 5. *Equip V with the inner product*

$$(u, v)_V = \int_\Omega uv + \nabla u \cdot \nabla v + \Delta u \Delta v.$$

The normed space $(V, \|.\|_V)$ is a Hilbert space.

Defining $V_0 = \{v \in V \mid v = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \text{ and } \frac{\partial v}{\partial n} = 0 \text{ on } \Gamma_1 \cup \Gamma_3\}$, we first consider the H^2-conforming weak form of (3.20), requiring $u \in V_0$ such that

$$a(u, v) = \int_\Omega fv, \quad \forall v \in V_0, \tag{3.26}$$

where the bilinear form a is defined as

$$a(u, v) = \int_\Omega \Delta u \Delta v + c_0 \nabla u \cdot \nabla v + c_1 uv. \tag{3.27}$$

Using standard tools, it is straightforward to show that the weak form in (3.26) is well-posed (i.e., that $a(u, v)$ is coercive and continuous on V_0) when $c_0, c_1 > 0$, for any choice of boundary conditions. This can be extended to cover the case of $c_0 = 0$ if $\Gamma_2 = \emptyset$, using the remaining boundary conditions to show that there is a constant, C, such that $\|\nabla u\|_0^2 \leq C\left(\|u\|_0^2 + \|\Delta u\|_0^2\right)$. If $c_1 = 0$ for $c_0 > 0$, then well-posedness requires that $\Gamma_0 \cup \Gamma_1 \neq \emptyset$, in order to be able to apply the standard Poincaré inequality. If both $c_0 = c_1 = 0$, then both $\Gamma_0 \cup \Gamma_1 \neq \emptyset$ and $\Gamma_2 = \emptyset$ are required to show well-posedness.

Remark 3.3.1. Assume that $\Gamma_0 \cup \Gamma_1 \neq \emptyset$ and Γ_2 is empty. While the constants c_0 and c_1 in (3.20) are assumed to be nonnegative throughout this manuscript, we point out that well-posedness of (3.26) can be proved for negative values of c_0 and c_1, with $|c_0| + |c_1|$ sufficiently small and depending on the Poincaré inequality constant ρ, where $\|u\|_1^2 \leq \rho \|\nabla u\|_0^2$.

Remark 3.3.2. When $\partial\Omega = \Gamma_2$ (the analogous case to full Neumann boundary conditions), if $c_0 = 0$, then $a(u, v)$ is not coercive on $V_0 = V$. We illustrate this by considering the harmonic function $v = e^{-kx}\cos(ky)$, for which $a(v, v) = c_1\|v\|_0^2 = \mathcal{O}(k^{-2})$. On the other hand, $\|v\|_V^2 = \mathcal{O}(k^{-2}) + \mathcal{O}(1)$. Thus, as k gets larger, the

implied bound on the coercivity constant goes to zero. Thus, in what follows, c_0 is restricted to be positive if $\Gamma_2 \subseteq \partial\Omega$.

We now turn our attention to the mixed formulation at the continuum level. Letting $\vec{v} = \nabla u$ and $\vec{\alpha} = \nabla\nabla\cdot\vec{v} - c_0\vec{v}$, (3.20) is equivalent to the following system of first- and second-order PDEs.

$$\nabla\cdot\vec{\alpha} + c_1 u = f, \tag{3.28}$$

$$\vec{\alpha} - \nabla\nabla\cdot\vec{v} + c_0\vec{v} = 0, \tag{3.29}$$

$$\vec{v} - \nabla u = 0. \tag{3.30}$$

Considering the relevant spaces and applying the boundary conditions given in (3.22)-(3.25), the weak form of (3.28)-(3.30) is to find the triple

$$(u, \vec{v}, \vec{\alpha}) \in L^2(\Omega) \times H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega) \times H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega) \text{ such that}$$

$$\int_\Omega \nabla\cdot\vec{\alpha}\phi + c_1 u\phi = \int_\Omega f\phi, \tag{3.31}$$

$$\int_\Omega \vec{\alpha}\cdot\vec{\psi} + \nabla\cdot\vec{v}\nabla\cdot\vec{\psi} + c_0\int_\Omega \vec{v}\cdot\vec{\psi} = 0, \tag{3.32}$$

$$\int_\Omega \vec{\beta}\cdot\vec{v} + \int_\Omega u\nabla\cdot\vec{\beta} = 0, \tag{3.33}$$

$\forall(\phi, \vec{\psi}, \vec{\beta}) \in L^2(\Omega) \times H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega) \times H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)$, where, for $\Gamma \subset \partial\Omega$,

$$H_0^\Gamma(\mathrm{div};\Omega) = \{\vec{v} \in H(\mathrm{div};\Omega) \mid \vec{v}\cdot\vec{n} = 0 \text{ on } \Gamma\}.$$

We note that this formulation strongly imposes Dirichlet boundary conditions on $\vec{v}$ and α, but weakly imposes those on u and Δu.

This weak form is equivalent to the saddle-point problem of finding $(u, \vec{v}, \vec{\alpha}) \in L^2(\Omega) \times H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega) \times H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)$ such that

$$a\big((u,\vec{v}),(\phi,\vec{\psi})\big) + b\big((\phi,\vec{\psi}),\vec{\alpha}\big) = F(\phi), \tag{3.34}$$

$$b((u,\vec{v}),\vec{\beta}) = 0, \tag{3.35}$$

$\forall(\phi, \vec{\psi}, \vec{\beta}) \in L^2(\Omega) \times H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega) \times H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)$, where the linear and bilinear

forms $a,\ b,$ and F are given by

$$a\big((u,\vec{v}),(\phi,\vec{\psi})\big) = c_0\int_\Omega \vec{v}\cdot\vec{\psi} + \int_\Omega \nabla\cdot\vec{v}\,\nabla\cdot\vec{\psi} + c_1\int_\Omega u\phi, \tag{3.36}$$

$$b\big((u,\vec{v}),\vec{\beta}\big) = \int_\Omega \vec{\beta}\cdot\vec{v} + u\nabla\cdot\vec{\beta}, \tag{3.37}$$

$$F(\phi) = \int_\Omega f\phi. \tag{3.38}$$

As noted above, the boundary conditions imposed can have significant effects on the well-posedness of the problem. In particular, we now show that the mixed-formulation is well-posed under combinations of assumptions on $c_0,\ c_1$, and the boundary conditions.

Theorem 9. *Let $\partial\Omega = \Gamma_0\cup\Gamma_3$. The saddle-point problem* (3.34)-(3.35) *has a unique solution for any $c_0 \geq 0$ and $c_1 > 0$, and for $c_1 \geq 0$ if Γ_0 is nonempty.*

Proof. We verify the standard Brezzi conditions for well-posedness [32]. Continuity of a, b, and F in the product norm on $L^2(\Omega)\times H(\mathrm{div};\Omega)\times H(\mathrm{div};\Omega)$ is straightforward.

We next show that the bilinear form $a\big((u,\vec{v}),(\phi,\vec{\psi})\big)$ is coercive on the set

$$\eta = \{(u,\vec{v})\in L^2(\Omega)\times H_0^{\Gamma_3}(\mathrm{div};\Omega)\mid b\big((u,\vec{v}),\vec{\alpha}\big)=0,\ \ \forall\vec{\alpha}\in H_0^{\Gamma_3}(\mathrm{div};\Omega)\}.$$

Since the boundary conditions for $\vec{v}$ and $\vec{\alpha}$ are identical on $\Gamma_0\cup\Gamma_3$, the kernel condition implies that $b\big((u,\vec{v}),\vec{v}\big)=0$ for any $(u,\vec{v})$ in η, which implies that $\|\vec{v}\|_0^2 = -\int_\Omega u\nabla\cdot\vec{v} \leq \frac{1}{2}\left(\|u\|_0^2+\|\nabla\cdot\vec{v}\|_0^2\right)$. Then

$$\begin{aligned} a\big((u,\vec{v}),(u,\vec{v})\big) &= c_0\|\vec{v}\|_0^2 + \frac{1}{3}\big(\|\nabla\cdot\vec{v}\|_0^2 + c_1\|u\|_0^2\big) + \frac{2}{3}\big(\|\nabla\cdot\vec{v}\|_0^2 + c_1\|u\|_0^2\big) \\ &\geq c_0\|\vec{v}\|_0^2 + \frac{2\min\{1,c_1\}}{3}\|\vec{v}\|_0^2 + \frac{2\min\{1,c_1\}}{3}\big(\|\nabla\cdot\vec{v}\|_0^2+\|u\|_0^2\big) \\ &\geq \frac{2\min\{1,c_1\}}{3}\left(\|\vec{v}\|_{\mathrm{div}}^2+\|u\|_0^2\right), \end{aligned}$$

where $\|\vec{v}\|_{\mathrm{div}}^2 = \|\vec{v}\|_0^2+\|\nabla\cdot\vec{v}\|_0^2$.

If Γ_0 is nonempty and $c_1=0$, then for a given $(u,\vec{v})$, we choose $\vec{\alpha}=\mu\vec{v}+\vec{\alpha}_m$, where μ is a positive constant to be specified below, and $\vec{\alpha}_m$ is the solution of the

standard mixed Poisson problem,

$$\int_\Omega \delta \nabla \cdot \vec{\alpha}_m = \int_\Omega u\delta, \quad \forall \delta \in L^2(\Omega), \tag{3.39}$$

$$\int_\Omega \vec{\alpha}_m \cdot \vec{\beta} + \phi \nabla \cdot \vec{\beta} = 0, \quad \forall \vec{\beta} \in H_0^{\Gamma_3}(\Omega; \mathrm{div}), \tag{3.40}$$

which is well-posed with $\|\vec{\alpha}_m\|_{\mathrm{div}}^2 + \|\phi\|_0^2 \leq \Lambda \|u\|_0^2$, where Λ is a positive constant that depends on the coercivity and continuity constants and the inf-sup conditions for the mixed Poisson problem [32, 106]. Moreover, the choice of $\delta = u$ in Equation (3.39) implies that $\|u\|_0^2 = \int_\Omega u \nabla \cdot \vec{\alpha}_m$. Thus, for every $(u, \vec{v}) \in \eta$, we have

$$b((u, \vec{v}), \vec{\alpha}) = \mu \|\vec{v}\|_0^2 + \|u\|_0^2 + \int_\Omega \vec{\alpha}_m \cdot \vec{v} + \mu \int_\Omega u \nabla \cdot \vec{v} = 0.$$

Rearranging terms and using the Cauchy-Schwarz and Young's inequalities, we get

$$\frac{\mu}{2}\left(\frac{2\mu}{k_1}\|u\|_0^2 + \frac{k_1}{2\mu}\|\nabla \cdot \vec{v}\|_0^2\right) + \frac{1}{2}\left(\frac{2}{k_2}\|u\|_0^2 + \frac{k_2\Lambda}{2}\|\vec{v}\|_0^2\right) \geq \mu\|\vec{v}\|_0^2 + \|u\|_0^2$$

for arbitrary $k_1 > 0, k_2 > 0$, which can be further rearranged to yield

$$\frac{k_1}{4}\|\nabla \cdot \vec{v}\|_0^2 \geq \left(\mu - \frac{k_2\Lambda}{4}\right)\|\vec{v}\|_0^2 + \left(1 - \frac{\mu^2}{k_1} - \frac{1}{k_2}\right)\|u\|_0^2.$$

Choosing sufficiently large constants k_1 and μ and sufficiently small k_2 results in the coercivity condition that $a\left((u, \vec{v}), (u, \vec{v})\right) = \|\nabla \cdot \vec{v}\|_0^2 \geq K\left(\|\vec{v}\|_{\mathrm{div}}^2 + \|u\|_0^2\right)$, for some constant $K > 0$.

Finally, we establish the necessary inf-sup condition, that

$$\sup_{(u,\vec{v}) \in L^2(\Omega) \times H_0^{\Gamma_3}(\mathrm{div};\Omega)} \frac{b\big((u, \vec{v}), \vec{\alpha}\big)}{\sqrt{\|u\|_0^2 + \|\vec{v}\|_{\mathrm{div}}^2}} \geq \frac{1}{\sqrt{2}}\|\vec{\alpha}\|_{\mathrm{div}}, \ \forall \vec{\alpha} \in H_0^{\Gamma_3}(\mathrm{div}; \Omega)$$

The choice $u = \nabla \cdot \vec{\alpha}$, $\vec{v} = \vec{\alpha}$ completes the proof, noting this is compatible with $\partial\Omega = \Gamma_0 \cup \Gamma_3$, since $u \in L^2(\Omega)$, without an essential boundary condition strongly imposed on it. □

Corollary 4. *Let $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$. The saddle-point problem* (3.34)-(3.35) *has a unique solution for any $c_0 > 0$ and $c_1 > 0$.*

Proof. Under these assumptions, the bilinear form a is coercive for $(u,\vec{v}) \in L^2(\Omega) \times H_0^{\Gamma_3}(\mathrm{div};\Omega)$ since

$$a\left((u,\vec{v}),(u,\vec{v})\right) = c_0\|\vec{v}\|_0^2 + \|\nabla\cdot\vec{v}\|_0^2 + c_1\|u\|_0^2 \geq \min\left\{1,c_0,c_1\right\}\left(\|\vec{v}\|_{\mathrm{div}}^2 + \|u\|_0^2\right).$$

Moreover, the inf-sup condition,

$$\sup_{(u,\vec{v})\in L^2(\Omega)\times H_0^{\Gamma_3}(\mathrm{div};\Omega)} \frac{b\big((u,\vec{v}),\vec{\alpha}\big)}{\sqrt{\|u\|_0^2+\|\vec{v}\|_{\mathrm{div}}^2}} \geq \frac{1}{\sqrt{2}}\|\vec{\alpha}\|_{\mathrm{div}},\ \forall\vec{\alpha}\in H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega),$$

is readily shown by choosing $u=\nabla\cdot\vec{\alpha},\ \vec{v}=\vec{\alpha}$, noting that this is allowable because $\vec{\alpha}\in H_0^{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)\subset H_0^{\Gamma_3}(\mathrm{div};\Omega)$, and $\nabla\cdot\vec{\alpha}\in L^2(\Omega)$. □

Solving (3.34)–(3.35) when essential boundary conditions on $\vec{v}$ are strongly imposed while $\vec{\alpha}$ is free on the boundary, i.e. $\partial\Omega=\Gamma_1$, leads to difficulties in proving the inf-sup condition. This difficulty can easily be understood from the proof of the inf-sup condition in Theorem 9, in which we take $\vec{v}=\vec{\alpha}$ to provide a concrete bound on the supremum. In this setting, we are able to prove uniqueness of the solution to the continuum mixed form of the problem under suitable regularity assumptions.

Corollary 5. *Let $\partial\Omega=\Gamma_0\cup\Gamma_1\cup\Gamma_3$ and $\bar{u}$ solve* (3.26)*. The pair $(u,\vec{v})$ that solves the saddle-point problem* (3.34)-(3.35) *is unique for any $c_0\geq 0$ and $c_1\geq 0$ with $(u,\vec{v})=(\bar{u},\nabla\bar{u})$. Moreover, $\vec{\alpha}$ is unique if $\bar{u}\in H^t(\Omega)$, $t\geq 4$.*

Proof. As in the proof of Theorem 9, the bilinear form a is coercive on the set

$$\eta=\{(u,\vec{v})\in L^2(\Omega)\times H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega)\mid b\big((u,\vec{v}),\vec{\alpha}\big)=0,\quad\forall\vec{\alpha}\in H_0^{\Gamma_3}(\mathrm{div};\Omega)\}.$$

Therefore, the pair $(u,\vec{v})$ is uniquely determined [45, Remark 1.1]. As $(\bar{u},\nabla u)$ solves (3.35) for every $\vec{\beta}$, uniqueness of $(u,\vec{v})$ implies that $(u,\vec{v})=(\bar{u},\nabla\bar{u})$. If, additionally, $u=\bar{u}\in H^t(\Omega)$, $t\geq 4$, then we choose $(\phi,\vec{\psi})=(0,Q\left(\vec{\alpha}-\nabla\Delta u+c_0\nabla u\right))$ in (3.34), for any $Q\in C^{t-3}(\Omega)\cap H_0^1(\Omega)$ that is positive in Ω. Note that $\vec{\psi}\in H_0^{\Gamma_1\cup\Gamma_3}(\mathrm{div};\Omega)$ since $u\in H^t(\Omega)$ for $t\geq 4$. Integration by parts on (3.34) then yields

$$\int_\Omega Q\left(\vec{\alpha}-\nabla\Delta u+c_0\nabla u\right)\cdot\left(\vec{\alpha}-\nabla\Delta u+c_0\nabla u\right)=0.$$

As $Q\left(\vec{\alpha}-\nabla\Delta u+c_0\nabla u\right)\cdot\left(\vec{\alpha}-\nabla\Delta u+c_0\nabla u\right)$ is non-negative in Ω, this implies that

$\vec{\alpha} = \nabla\Delta u - c_0\nabla u.$ □

Remark 3.3.3. In the case $d = 2$, we can write $\partial\Omega = \left\{\cup_{i=1}^{M_1}\Gamma^i\right\} \cup \left\{\cup_{i=1}^{M_2}\tilde{\Gamma}^i\right\}$, where $\Gamma^i = (x, a_i x + b_i)$ for $x_i^0 < x < x_i^1$, and $\tilde{\Gamma}^i = (c_i y + d_i, y)$ for $y_i^0 < y < y_i^1$, with M_1 and M_2 positive integers. The function Q in Corollary (5) can be chosen as

$$Q = \Pi_{i=1}^{M_1}\left(y - a_i x - b_i\right)^2 \Pi_{i=1}^{M_2}\left(x - c_i y - d_i\right)^2,$$

with $Q \in C^{t-3}(\Omega) \cap H_0^1(\Omega)$ and Q positive in the interior of Ω. Similarly, Q can be constructed when $d = 3$ by writing the boundary faces of Ω in sets that can be parametrized as planes in each pair of two Cartesian coordinates.

3.4 Discrete Analysis

For what follows, we consider a conforming discretization of the mixed form, with

$$(u_h, \vec{v}_h, \vec{\alpha}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega, \tau_h),$$

for $k \geq 0$, where $RT_{k+1}^{\Gamma}(\Omega, \tau_h) = \{\vec{v}_h \in RT_{k+1}(\Omega, \tau_h) \mid \vec{v}_h \cdot \vec{n} = 0 \text{ on } \Gamma\}$, noting that $DG_k(\Omega, \tau_h) \subset L^2(\Omega)$ and $RT_{k+1}^{\Gamma}(\Omega, \tau_h) \subset H_0^{\Gamma}(\mathrm{div}; \Omega)$. We examine the problem of finding $(u, \vec{v}, \vec{\alpha}) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega, \tau_h)$ such that

$$a\big((u_h, \vec{v}_h), (\phi_h, \vec{\psi}_h)\big) + b\big((\phi_h, \vec{\psi}_h), \vec{\alpha}_h\big) = F\big(\phi_h\big), \tag{3.41}$$

$$b\big((u_h, \vec{v}_h), \vec{\beta}_h\big) = 0, \tag{3.42}$$

$\forall(\phi_h, \vec{\psi}_h, \vec{\beta}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega, \tau_h)$, where a, b and F are defined in (3.36)-(3.38). Boundary conditions on Γ_1 will be enforced with Nitsche's method. As in the continuum case, taking $\partial\Omega = \Gamma_0 \cup \Gamma_3$ is the easiest case to consider.

Corollary 6. *Let $\partial\Omega = \Gamma_0 \cup \Gamma_3$, $c_0 \geq 0, c_1 > 0$, and $c_1 \geq 0$ if Γ_0 is nonempty. Let $\{\tau_h\}$, $0 < h \leq 1$ be a quasiuniform family of triangular meshes of Ω. Then problem* (3.41)-(3.42) *has a unique solution for any τ_h in the family.*

Proof. Coercivity of the bilinear form $a\big((u_h, \vec{v}_h), (\phi, \vec{\psi}_h)\big)$ on the set

$$\{(u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h) \mid b\big((u_h, \vec{v}_h), \vec{\alpha}_h\big) = 0, \ \ \forall\vec{\alpha} \in RT_{k+1}^{\Gamma_3}(\Omega, \tau_h)\},$$

and the inf-sup condition of the form

$$\sup_{(u_h,\vec{v}_h)\in DG_k(\Omega,\tau_h)\times RT^{\Gamma_3}_{k+1}(\Omega,\tau_h)} \frac{b\big((u_h,\vec{v}_h),\vec{\alpha}_h\big)}{\sqrt{\|u_h\|_0^2+\|\vec{v}_h\|_{\text{div}}^2}} \geq \frac{1}{\sqrt{2}}\|\vec{\alpha}_h\|_{\text{div}},\ \forall \vec{\alpha}_h \in RT^{\Gamma_3}_{k+1}(\Omega,\tau_h),$$

can be proven exactly as in the continuum. Note that this is compatible with the finite-element spaces. For example, when $c_1 > 0$, we can choose $\vec{\alpha}_h = \vec{v}_h \in RT^{\Gamma_3}_{k+1}(\tau_h)$ in the kernel condition within the coercivity proof, and $\vec{v}_h = \vec{\alpha}_h \in RT^{\Gamma_3}_{k+1}(\tau_h)$ and $u_h = \nabla\cdot\vec{\alpha}_h$ in the proof of the inf-sup condition. Such a u_h is in $DG_k(\Omega,\tau_h)$ by Remark 3.2.2. □

Corollary 7. *Let the assumptions of Corollary 6 hold, and let*

$$(u,\vec{v},\vec{\alpha}) \in H^{k+1}(\Omega)\times[H^{k+2}(\Omega)]^d\times[H^{k+2}(\Omega)]^d$$

be the solution of (3.34)-(3.35)*. Let*

$$(u_h,\vec{v}_h,\vec{\alpha}_h) \in DG_k(\Omega,\tau_h)\times RT^{\Gamma_3}_{k+1}(\Omega,\tau_h)\times RT^{\Gamma_3}_{k+1}(\Omega,\tau_h)$$

be the solution of (3.41)-(3.42)*. Then, there exist constants m_1 and m_2 such that*

$$\|(u,\vec{v})-(u_h,\vec{v}_h)\|_{0,\text{div}} \leq m_1 h^{k+1}\left(|u|^2_{k+1}+|\vec{v}|^2_{k+1}+|\vec{v}|^2_{k+2}+|\vec{\alpha}|^2_{k+1}+|\vec{\alpha}|^2_{k+2}\right)^{1/2}, \tag{3.43}$$

$$\|\vec{\alpha}-\vec{\alpha}_h\|_{\text{div}} \leq m_2 h^{k+1}\left(|u|^2_{k+1}+|\vec{v}|^2_{k+1}+|\vec{v}|^2_{k+2}+|\vec{\alpha}|^2_{k+1}+|\vec{\alpha}|^2_{k+2}\right)^{1/2}. \tag{3.44}$$

Proof. Because this is a conforming discretization, standard approximation theory for mixed finite elements (e.g. [32]) yields optimal approximation results in the product norm, $\|\,(u_h,\vec{v}_h)\,\|^2_{0,\text{div}} = \|u_h\|_0^2+\|\vec{v}_h\|^2_{\text{div}}$. □

Corollary 8. *Problem* (3.41)-(3.42)*, with $\partial\Omega = \Gamma_0\cup\Gamma_2\cup\Gamma_3$, has a unique solution for $c_0, c_1 > 0$. If, further,*

$$(u,\vec{v},\vec{\alpha}) \in H^{k+1}(\Omega)\times[H^{k+2}(\Omega)]^d\times[H^{k+2}(\Omega)]^d$$

is the solution of (3.34)-(3.35) *and*

$$(u_h, \vec{v}_h, \vec{\alpha}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h)$$

is the solution of (3.41)-(3.42)*, then the error bounds in* (3.43)–(3.44) *also hold for this case.*

Proof. The proof follows exactly as those of Corollaries 4 and 7. The bilinear form a is coercive for every pair $(u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega, \tau_h)$. The finite-element approximation spaces allow the choice $u_h = \nabla \cdot \vec{\alpha}_h$ and $\vec{v}_h = \vec{\alpha}_h$ for the inf-sup condition. As this is a conforming discretization, standard theory yields optimal approximation results. □

As in the continuum case, solving (3.41)–(3.42) when essential boundary conditions on $\vec{v}$ are strongly imposed while $\vec{\alpha}$ is free on the boundary leads to difficulties in proving the inf-sup condition. When $\partial\Omega = \Gamma_1$, we cannot follow the proof technique used in Theorem 9 and Corollary 4, since $\vec{v}$ must satisfy the prescribed BC while $\vec{\alpha}$ is free on the boundary. In this case, the inf-sup condition has the form of finding $\tilde{c} > 0$ such that

$$I = \sup_{(u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_1}(\Omega, \tau_h)} \frac{\int_\Omega \vec{\alpha}_h \cdot \vec{v}_h + \int_\Omega u \nabla \cdot \vec{\alpha}_h}{\sqrt{\|u_h\|_0^2 + \|\vec{v}_h\|_{\text{div}}^2}} > \tilde{c} \|\vec{\alpha}_h\|_{\text{div}}, \ \forall \vec{\alpha}_h \in RT_{k+1}(\Omega, \tau_h).$$

To understand the challenge, we consider a simply-connected domain Ω. Then the two-dimensional discrete Helmholtz decomposition from Lemma 2,

$$RT_{k+1}(\Omega, \tau_h) = \Big(\nabla \times CG_{k+1}(\Omega, \tau_h) \Big) \oplus \Big(\operatorname{grad}_h DG_k(\Omega, \tau_h) \Big).$$

For any $\vec{\alpha}_h = \operatorname{grad}_h z$ where $z \in DG_k(\Omega, \tau_h)$, then the choice $\vec{v}_h = 0$ and $u_h = \nabla \cdot \vec{\alpha}_h - z$ satisfies the inf-sup condition, as

$$\begin{aligned} I &\geq \sup_{u_h \neq 0 \in DG_k(\Omega, \tau_h)} \frac{\int_\Omega u_h \nabla \cdot \vec{\alpha}_h}{\|u_h\|_0} \geq \frac{\|\nabla \cdot \vec{\alpha}_h\|_0^2 - \int_\Omega z \nabla \cdot (\operatorname{grad}_h z)}{\|\nabla \cdot \vec{\alpha}_h\|_0 + \|z\|_0} \\ &\geq \frac{\|\nabla \cdot \vec{\alpha}_h\|_0^2 + \|\operatorname{grad}_h z\|_0^2}{\|\nabla \cdot \vec{\alpha}_h\|_0 + c\|\operatorname{grad}_h z\|_0} \geq \tilde{c} \left(\|\nabla \cdot \vec{\alpha}_h\|_0 + \|\vec{\alpha}_h\|_0 \right), \end{aligned}$$

where the discrete Poincaré inequality [14], $\|z\|_0 \leq c\|\operatorname{grad}_h z\|_0$, is used here. In contrast, for $\vec{\alpha}_h \in \nabla \times CG_{k+1}(\Omega, \tau_h)$, we cannot establish a uniform inf-sup condition. As a simple example, take $\Omega = (0,1)^2$, with $\partial\Omega = \Gamma_1$, and consider $k = 0$, so $\vec{\alpha}_h \in RT_1(\Omega, \tau_h)$. Consider a mesh such that one triangle has vertices $(0,0)$, $(h,0)$, and $(0,h)$. Take $\vec{\alpha}_h$ to be nonzero only in this triangle, with value

$$\vec{\alpha}_h = \nabla \times \left(1 - \frac{x+y}{h}\right) = \frac{1}{h}\begin{bmatrix}-1\\1\end{bmatrix}. \tag{3.45}$$

Clearly, $\nabla \cdot \vec{\alpha}_h = 0$. Moreover, for any choice of $\vec{v}_h \in RT_1(\Omega, \tau_h)$ with $\vec{v}_h \cdot \vec{n} = 0$ on $\partial\Omega$ and $u_h \in DG_0(\Omega, \tau_h)$, we have $\int_\Omega \vec{\alpha}_h \cdot \vec{v}_h + \int_\Omega u \nabla \cdot \vec{\alpha}_h = 0$, which results in a zero inf-sup constant. Numerical experiments (not reported here) suggest that restricting the mesh so that no element has 3 vertices on the boundary yields an $\mathcal{O}(h)$ inf-sup constant.

We next show that weakly implementing the essential boundary condition on $\vec{v}$ yields a discretization with $\mathcal{O}(1)$ continuity and coercivity constants and $\mathcal{O}(h)$ inf-sup constant, without any mesh restrictions but in a modified norm. This is slightly disadvantageous, because the error estimate loses some convergence due to both the sub-optimal inf-sup constant and the error analysis in the modified norm; however, we view the lack of restrictions on the mesh construction to be preferable to possible further results in the direction considered above. To weakly impose the boundary condition, we will make use of a Nitsche-type penalty method. These approaches are based on adding three terms to the weak form, which are commonly denoted as the consistency, stability, and symmetry terms [86, 109]. Consider the case where $\partial\Omega = \Gamma_0 \cup \Gamma_1 \cup \Gamma_3$, and modify the bilinear form $a\big((u,\vec{v}),(\phi,\vec{\psi})\big)$ from (3.36), to be

$$\hat{a}\big((u_h,\vec{v}_h),(\phi_h,\vec{\psi}_h)\big) + b\big((\phi_h,\vec{\psi}_h),\vec{\alpha}_h\big) = F(\phi_h), \tag{3.46}$$

$$b\big((u_h,\vec{v}_h),\vec{\beta}_h\big) = 0, \tag{3.47}$$

$\forall (\phi_h, \vec{\psi}_h, \vec{\beta}_h) \in DG_k(\Omega,\tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h)$, where,

$$\begin{aligned}\hat{a}\big((u_h,\vec{v}_h),(\phi_h,\vec{\psi}_h)\big) =& a\big((u_h,\vec{v}_h),(\phi_h,\vec{\psi}_h)\big) - \int_{\Gamma_1} \nabla\cdot\vec{v}_h\,\vec{\psi}_h\cdot\vec{n} \\ &- \int_{\Gamma_1} \nabla\cdot\vec{\psi}_h\,\vec{v}_h\cdot\vec{n} + \frac{\lambda}{h}\int_{\Gamma_1} \vec{v}_h\cdot\vec{n}\,\vec{\psi}_h\cdot\vec{n}.\end{aligned} \tag{3.48}$$

Here, we impose the condition that $\vec{v}_h \cdot \vec{n} = 0$ on Γ_1 directly by adding $\frac{\lambda}{h}\int_{\Gamma_1} \vec{v}_h \cdot \vec{n}\, \vec{\psi}_h \cdot \vec{n}$ to a as defined in (3.36) for penalty parameter $\lambda > 0$. Consistent with the boundary condition, we add $-\int_{\Gamma_1} \nabla \cdot \vec{\psi}_h\, \vec{v}_h \cdot \vec{n}$ to the bilinear form. For symmetry, we add the term $-\int_{\Gamma_1} \nabla \cdot \vec{v}_h\, \vec{\psi}_h \cdot \vec{n}$ to the bilinear form. With these Nitsche terms, we now prove well-posedness of the weak form, but using a modified norm for $\vec{v}_h$, defined in (3.13), which we recall here:

$$\|\vec{v}_h\|^2_{\mathrm{div},\Gamma_1} = \|\vec{v}_h\|^2_{\mathrm{div}} + h\|\nabla \cdot \vec{v}_h\|^2_{0,\Gamma_1} + \frac{1}{h}\|\vec{v}_h \cdot \vec{n}\|^2_{0,\Gamma_1}.$$

Theorem 10. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, with $\partial\Omega = \Gamma_0 \cup \Gamma_1 \cup \Gamma_3$. Let $\{\tau_h\}$, $0 < h \leq 1$, be a quasiuniform family of meshes of Ω, and let $\lambda > 0$ be given. For sufficiently large λ, the weak form in (3.46)–(3.47) has a unique solution for $c_0 \geq 0$, $c_1 > 0$, and for $c_1 = 0$ if $\Gamma_0 \cup \Gamma_1$ is nonempty.*

Proof. As above, the existence and uniqueness of solutions follows from standard theory. We first show that the bilinear form $\hat{a}\big((u_h, \vec{v}_h), (\phi_h, \vec{\psi}_h)\big)$ is coercive for any pair $(u_h, \vec{v}_h) \in \eta_h$, where

$$\begin{aligned}\eta_h = \{&(u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times RT^{\Gamma_3}_{k+1}(\Omega, \tau_h) | \\ &b\big((u_h, \vec{v}_h), \vec{\alpha}_h\big) = 0, \quad \forall \vec{\alpha}_h \in RT^{\Gamma_3}_{k+1}(\Omega, \tau_h)\}.\end{aligned}$$

Since the strongly imposed boundary conditions for $\vec{v}$ and $\vec{\alpha}$ are identical, the kernel condition implies that $b\big((u_h, \vec{v}_h), \vec{v}_h\big) = 0$ for any pair $(u_h, \vec{v}_h) \in \eta_h$. Thus, any pair $(u_h, \vec{v}_h) \in \eta_h$ should satisfy $\|\vec{v}_h\|_0^2 = -\int_\Omega u_h \nabla \cdot \vec{v}_h \leq \frac{1}{2}\left(\|u_h\|_0^2 + \|\vec{v}_h\|_0^2\right)$. Employing the Cauchy-Schwarz inequality and letting $\gamma = \gamma_1(k)$ from Corollary 3, we then have

$$\begin{aligned}\hat{a}\big((u_h, \vec{v}_h), (u_h, \vec{v}_h)\big) =& c_0\|\vec{v}_h\|_0^2 + \|\nabla \cdot \vec{v}_h\|_0^2 + c_1\|u_h\|_0^2 - 2\int_{\Gamma_1} \nabla \cdot \vec{v}_h\, \vec{v}_h \cdot \vec{n} + \frac{\lambda}{h}\|\vec{v}_h \cdot \vec{n}\|^2_{0,\Gamma_1} \\ \geq& \|\nabla \cdot \vec{v}_h\|_0^2 + c_1\|u_h\|_0^2 - \frac{h}{3\gamma}\|\nabla \cdot \vec{v}_h\|^2_{0,\Gamma_1} + \frac{\lambda - 3\gamma}{h}\|\vec{v} \cdot \vec{n}\|^2_{0,\Gamma_1} \\ \geq& \frac{1}{3}\|\nabla \cdot \vec{v}_h\|_0^2 + c_1\|u_h\|_0^2 + \frac{h}{3\gamma}\|\nabla \cdot \vec{v}_h\|^2_{0,\Gamma_1} + \frac{\lambda - 3\gamma}{h}\|\vec{v} \cdot \vec{n}\|^2_{0,\Gamma_1} \\ \geq& \frac{2\min\{1/3, c_1\}}{3}\|\vec{v}_h\|_0^2 + \frac{2\min\{1/3, c_1\}}{3}\left(\|\nabla \cdot \vec{v}_h\|_0^2 + \|u\|_0^2\right) \\ &+ \frac{h}{3\gamma}\|\nabla \cdot \vec{v}_h\|^2_{0,\Gamma_1} + \frac{\lambda - 3\gamma}{h}\|\vec{v}_h \cdot \vec{n}\|^2_{0,\Gamma_1}. \end{aligned} \tag{3.49}$$

Clearly, any choice of $\lambda > 3\gamma$ completes the proof. When $\Gamma_0 \cup \Gamma_1$ is nonempty and $c_0 = 0$, the coercivity proof is similar to the one in Theorem 9.

Continuity of $\hat{a}$, b, and F, can be established using Cauchy-Schwarz inequality. The resulting inequalities are that

$$\hat{a}\left((u_h, \vec{v}_h), \left(\phi_h, \vec{\psi}_h\right)\right) \leq \|\hat{a}\| \|(u_h, \vec{v}_h)\|_{0,\mathrm{div},\Gamma_1} \|(\phi_h, \vec{\psi}_h)\|_{0,\mathrm{div},\Gamma_1},$$

and

$$b\left((u_h, \vec{v}_h), \vec{\alpha}_h\right) \leq \|\left(u_h, \vec{v}_h\right)\|_{0,\mathrm{div},\Gamma_1} \|\vec{\alpha}_h\|_{\mathrm{div}},$$

where $\|\hat{a}\| = 3 + c_0 + c_1 + \lambda$, and $\|\left(u_h, \vec{v}_h\right)\|^2_{0,\mathrm{div},\Gamma_1} = \|u_h\|_0^2 + \|\vec{v}_h\|^2_{\mathrm{div},\Gamma_1}$. Thus, the continuity constant of the bilinear form $\hat{a}$ is $\mathcal{O}(1)$.

Finally, we consider the inf-sup condition, that $\exists \theta > 0$ such that

$$\begin{aligned} I = & \sup_{(u_h,\vec{v}_h) \in DG_k(\Omega,\tau_h) \times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h)} \frac{\int_\Omega \vec{\alpha}_h \cdot \vec{v}_h + \int_\Omega u_h \nabla \cdot \vec{\alpha}_h}{\sqrt{\|\vec{v}_h\|^2_{\mathrm{div},\Gamma_1} + \|u_h\|_0^2}} \\ & \geq \theta \|\vec{\alpha}_h\|_{\mathrm{div}}, \ \forall \vec{\alpha}_h \in RT_{k+1}^{\Gamma_3}(\Omega, \tau_h). \end{aligned}$$

The choice $\vec{v}_h = \vec{\alpha}_h$ and $u_h = \nabla \cdot \vec{\alpha}_h$ and the inverse trace inequality from Corollary 3 implies that

$$\begin{aligned} I &\geq \frac{\|\vec{\alpha}_h\|^2_{\mathrm{div}}}{\sqrt{\|\vec{\alpha}_h\|^2_{\mathrm{div}} + h\|\nabla \cdot \vec{\alpha}_h\|^2_{0,\Gamma_1} + \frac{1}{h}\|\vec{\alpha}_h \cdot \vec{n}\|^2_{0,\Gamma_1} + \|\nabla \cdot \vec{\alpha}_h\|_0^2}} \\ &\geq \frac{\|\vec{\alpha}_h\|^2_{\mathrm{div}}}{\sqrt{\left(1 + \frac{\gamma_1(k+1)}{h^2}\right)\|\vec{\alpha}_h\|_0^2 + (2 + \gamma_1(k))\|\nabla \cdot \vec{\alpha}_h\|_0^2}} \\ &\geq \frac{1}{\sqrt{1 + \frac{\gamma_1(k+1)}{h^2}}}\|\vec{\alpha}_h\|_{\mathrm{div}} = \frac{h}{\sqrt{h^2 + \gamma_1(k+1)}}\|\vec{\alpha}_h\|_{\mathrm{div}} \geq \frac{h}{\sqrt{1 + \gamma_1(k+1)}}\|\vec{\alpha}_h\|_{\mathrm{div}}. \end{aligned}$$

□

While the above result establishes the existence and uniqueness of discrete solutions, we note that the inf-sup constant is $\mathcal{O}(h)$, owing to the contribution from the boundary terms in the formulation.

Theorem 11. *Assume $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$ is a bounded polygonal or polyhedral*

domain with Lipschitz boundary. Let the assumptions of Theorem 10 be satisfied, and assume that $(u, \vec{v}, \vec{\alpha}) \in H^{k+5}(\Omega) \times [H^{k+4}(\Omega)]^d \times [H^{k+2}(\Omega)]^d$ *is the solution of* (3.34)–(3.35). *Let* $(u_h, \vec{v}_h, \vec{\alpha}_h)$ *be the unique solution of Problem* (3.46)-(3.47). *Then, there exists positive constants* M_1 *and* M_2 *such that*

$$\|(u, \vec{v}) - (u_h, \vec{v}_h)\|_{0,\mathrm{div},\Gamma_1} \leq M_1 h^k \left(|u|_{k+1}^2 + |\vec{v}|_{k+1}^2 + |\vec{v}|_{k+2}^2 + |\vec{\alpha}|_{k+1}^2 + |\vec{\alpha}|_{k+2}^2 \right)^{1/2},$$

$$\|\vec{\alpha} - \vec{\alpha}_h\|_{\mathrm{div}} \leq M_2 h^{k-1} \left(|u|_{k+1}^2 + |\vec{v}|_{k+1}^2 + |\vec{v}|_{k+2}^2 + |\vec{\alpha}|_{k+1}^2 + |\vec{\alpha}|_{k+2}^2 \right)^{1/2}.$$

Proof. To prove the error estimates, we proceed in a similar way to [45, Section II.2.1], with modifications to account for the use of Nitsche-type penalty methods to weakly impose the boundary conditions on $\vec{v}_h$.

Coercivity of $\hat{a}$ over η_h, where η_h is the set defined in the proof of Theorem 10, leads to the fact that for all $(\phi_h, \vec{\psi}_h) \in \eta_h$, we have

$$\begin{aligned} C\|(\phi_h, \vec{\psi}_h) - (u_h, \vec{v}_h)\|_{0,\mathrm{div},\Gamma_1}^2 &\leq \hat{a}\big((\phi_h - u_h, \vec{\psi}_h - \vec{v}_h), (\phi_h - u_h, \vec{\psi}_h - \vec{v}_h)\big) \\ &= \hat{a}\big((\phi_h - u, \vec{\psi}_h - \vec{v}), (\phi_h - u_h, \vec{\psi}_h - \vec{v}_h)\big) \\ &\quad + \hat{a}\big((u - u_h, \vec{v} - \vec{v}_h), (\phi_h - u_h, \vec{\psi}_h - \vec{v}_h)\big). \end{aligned}$$

Here, C is the coercivity constant, which is $\mathcal{O}(1)$. Note, first, that $\hat{a}\big((\phi_h - u, \vec{\psi}_h - \vec{v}), (\phi_h - u_h, \vec{\psi}_h - \vec{v}_h)\big)$ is continuous in its arguments, with $\mathcal{O}(1)$ continuity constant, so the first term is readily bounded. We next establish that we have enough regularity on the solution $(u, \vec{v}, \vec{\alpha})$ of the continuum problem to show that

$$\hat{a}\big((u - u_h, \vec{v} - \vec{v}_h), (\phi_h - u_h, \vec{\psi}_h - \vec{v}_h)\big) = b\left((\phi_h - u_h, \vec{\psi}_h - \vec{v}_h), \vec{\alpha}_h - \vec{\alpha}\right), \tag{3.50}$$

where we note that the regularity is required to both integrate by parts and enforce relationships between the components of the continuum solution below. To establish

(3.50), note first (from the definition of $\hat{a}$) that

$$\begin{aligned}\hat{a}\big((u,\vec{v}),(\phi_h-u_h,\vec{\psi}_h-\vec{v}_h)\big) &= \int_\Omega \nabla\cdot\vec{v}\,\nabla\cdot(\vec{\psi}_h-\vec{v}_h) + c_0\int_\Omega \vec{v}\cdot(\vec{\psi}_h-\vec{v}_h)\\ &\quad + c_1\int_\Omega u(\phi_h-u_h) - \int_{\Gamma_1}\nabla\cdot\vec{v}\,(\vec{\psi}_h-\vec{v}_h)\cdot\vec{n}\\ &= -\int_\Omega(\nabla\nabla\cdot\vec{v}-c_0\vec{v})\cdot(\vec{\psi}_h-\vec{v}_h) + c_1\int_\Omega u(\phi_h-u_h)\\ &= -\int_\Omega\vec{\alpha}\cdot(\vec{\psi}_h-\vec{v}_h) + c_1\int_\Omega u(\phi_h-u_h),\end{aligned}$$

where we invoke Corollary 5 to ensure that $\vec{v}=\nabla u$, $\vec{\alpha}=\nabla\nabla\cdot\vec{v}-c_0\vec{v}$, and the boundary conditions on Γ_0 and Γ_3 ensure the other boundary integrals from integration by parts vanish. Next, note that $(u,\vec{v})$ satisfies (3.34); taking $\vec{\psi}=\vec{0}$ and $\phi=\phi_h-u_h\in L^2(\Omega)$ in (3.34) gives

$$c_1\int_\Omega u(\phi_h-u_h)+\int_\Omega(\phi_h-u_h)\nabla\cdot\vec{\alpha}=F(\phi_h-u_h).$$

Combining these, we have that

$$\begin{aligned}\hat{a}\big((u,\vec{v}),(\phi_h-u_h,\vec{\psi}_h-\vec{v}_h)\big) &= -\int_\Omega\vec{\alpha}\cdot(\vec{\psi}_h-\vec{v}_h)+F(\phi_h-u_h)-\int_\Omega(\phi_h-u_h)\,\nabla\cdot\vec{\alpha}\\ &= F(\phi_h-u_h)-b\left((\phi_h-u_h,\vec{\psi}_h-\vec{v}_h),\vec{\alpha}\right).\end{aligned}$$

On the other hand, we have, from (3.46), $\hat{a}\big((u_h,\vec{v}_h),(\phi_h-u_h,\vec{\psi}_h-\vec{v}_h)\big)=F(\phi_h-u_h)-b\left((\phi_h-u_h,\vec{\psi}_h-\vec{v}_h),\vec{\alpha}_h\right)$. Taking these together establishes (3.50).

Now, for any $\vec{\beta}_h\in RT^{\Gamma_3}_{k+1}(\Omega,\tau_h)$,

$$\begin{aligned}b\left((\phi_h-u_h,\vec{\psi}_h-\vec{v}_h),\vec{\alpha}-\vec{\alpha}_h\right) &= b\left((\phi_h-u_h,\vec{\psi}_h-\vec{v}_h),\vec{\alpha}-\vec{\beta}_h\right)\\ &\le \|(\phi_h-u_h,\vec{v}_h-\vec{\psi}_h)\|_{0,\mathrm{div},\Gamma_1}\|\vec{\alpha}-\vec{\beta}_h\|_{\mathrm{div}}\end{aligned}$$

since $(u_h-\phi_h,\vec{v}_h-\vec{\psi}_h)\in\eta_h$ and by the continuity of b. Thus,

$$C\|(\phi_h,\vec{\psi}_h)-(u_h,\vec{v}_h)\|_{0,\mathrm{div},\Gamma_1}\le\|\hat{a}\|\|(\phi_h,\vec{\psi}_h)-(u,\vec{v})\|_{0,\mathrm{div},\Gamma_1}+\|\vec{\alpha}-\vec{\beta}_h\|_{\mathrm{div}}.$$

We choose $(\phi_h,\vec{\psi}_h)\in\eta_h$ to be the solution of the mixed Poisson problem posed on $DG_k(\Omega,\tau_h)\times RT^{\Gamma_3}_{k+1}(\Omega,\tau_h)$ with source term $-\Delta u$. Stability of this mixed formulation

leads to the estimate

$$\begin{aligned}\inf_{(\phi_h,\vec{\psi}_h)\in\eta_h}\|(u,\vec{v})-(\phi_h,\vec{\psi}_h)\|_{0,\mathrm{div}} &\leq \inf_{(\phi_h,\vec{\psi}_h)\in DG_k(\Omega,\tau_h)\times RT_{k+1}^{\Gamma_3}}\|(u,\vec{v})-(\phi_h,\vec{\psi}_h)\|_{0,\mathrm{div}}\\ &\leq \|(u,\vec{v})-(I_h^k u,\Pi_h^k\vec{v})\|_{0,\mathrm{div}}\\ &\leq r_1 h^{k+1}\left(|u|_{k+1}^2+|\vec{v}|_{k+1}^2+|\vec{v}|_{k+2}^2\right)^{\frac{1}{2}},\end{aligned}$$

for constant r_1. Note, however, that our analysis is posed in the stronger norm, $\|(\cdot,\cdot)\|_{0,\mathrm{div},\Gamma_1}$. To analyse the error in this norm, we have

$$\begin{aligned}\inf_{(\phi_h,\vec{\psi}_h)\in\eta_h}\|(u,\vec{v})-(\phi_h,\vec{\psi}_h)\|_{0,\mathrm{div},\Gamma_1} &\leq \|(u,\vec{v})-(I_h^k u,\Pi_h^k\vec{v})\|_{0,\mathrm{div}}\\ &+\sqrt{h}\|\nabla\cdot(\vec{v}-\Pi_h^k\vec{v})\|_{0,\Gamma_1}+\frac{1}{\sqrt{h}}\|\vec{v}-\Pi_h^k\vec{v}\|_{0,\Gamma_1}\\ &\leq r_2\left(h^{k+1}+h^{k+1/2}+h^k\right)\left(|u|_{k+1}^2+|\vec{v}|_k^2+|\vec{v}|_{k+2}^2\right)^{1/2},\end{aligned}\tag{3.51}$$

where r_2 is a positive constant independent of h, and the terms above are bounded using Theorem 7 and Lemmas 3 and 4, resulting in degraded convergence rates. Finally, using the triangle inequality, (3.51), and Theorem 7 leads to the $\mathcal{O}(h^k)$ estimate on $(u,\vec{v})$. To find the error estimate for $\vec{\alpha}$, we first use the triangle inequality, writing

$$\|\vec{\alpha}-\vec{\alpha}_h\|_{\mathrm{div}}\leq\|\vec{\alpha}-\vec{\beta}_h\|_{\mathrm{div}}+\|\vec{\beta}_h-\vec{\alpha}_h\|_{\mathrm{div}}.\tag{3.52}$$

Choosing $\vec{\beta}_h$ to be the interpolant of $\vec{\alpha}$ gives an error bound for the first term that is $\mathcal{O}(h^{k+1})$, as in Theorem 7. We then use the discrete inf-sup condition to bound the

second term, writing $\gamma = \sqrt{h^2\left(1+\gamma_1(k+1)\right)+\gamma_1(k+1)}$,

$$\begin{aligned}
\|\vec{\beta}_h - \vec{\alpha}_h\|_{\text{div}} &\leq \frac{\gamma}{h} \sup_{(\phi_h,\vec{\psi}_h)\in DG_k(\Omega,\tau_h)\times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h)} \frac{b\left((\phi_h,\vec{\psi}_h),\vec{\beta}_h-\vec{\alpha}_h\right)}{\|(\phi_h,\vec{\psi}_h)\|_{0,\text{div},\Gamma_1}} \\
&= \frac{\gamma}{h} \sup_{(\phi_h,\vec{\psi}_h)\in DG_k(\Omega,\tau_h)\times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h)} \frac{b\left((\phi_h,\vec{\psi}_h),\vec{\alpha}-\vec{\alpha}_h\right)+b\left((\phi_h,\vec{\psi}_h),\vec{\beta}_h-\vec{\alpha}\right)}{\|(\phi_h,\vec{\psi}_h)\|_{0,\text{div},\Gamma_1}} \\
&= \frac{\gamma}{h} \sup_{(\phi_h,\vec{\psi}_h)\in DG_k(\Omega,\tau_h)\times RT_{k+1}^{\Gamma_3}(\Omega,\tau_h)} \frac{\hat{a}\left((u-u_h,\vec{v}-\vec{v}_h),(\phi_h,\vec{\psi}_h)\right)+b\left((\phi_h,\vec{\psi}_h),\vec{\alpha}-\vec{\beta}_h\right)}{\|(\phi_h,\vec{\psi}_h)\|_{0,\text{div},\Gamma_1}} \\
&\leq \frac{\gamma\|\hat{a}\|}{h}\|(u,\vec{v})-(u_h,\vec{v}_h)\|_{0,\text{div},\Gamma_1} + \frac{\gamma}{h}\|\vec{\alpha}-\vec{\beta}_h\|_{\text{div}},
\end{aligned}$$

where we use the continuity of $\hat{a}$ and b in the final inequality. The error estimate for $(u,\vec{v})$ above implies that the convergence rate of $\vec{\alpha}$ is $\mathcal{O}(h^{k-1})$. □

3.5 Monolithic multigrid preconditioner

We now consider the development of effective linear solvers for the resulting discretized systems. We first consider the case where $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$ with constants $c_0, c_1 > 0$; however, the same arguments allow the case where $c_0 = 0$ if Γ_2 is empty. The discretizations above lead to block-structured linear systems that can be written as

$$\begin{bmatrix} A_{11} & 0 & B_1^T \\ 0 & A_{22} & B_2^T \\ B_1 & B_2 & 0 \end{bmatrix} \begin{bmatrix} u \\ \vec{v} \\ \vec{\alpha} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ g \end{bmatrix}, \tag{3.53}$$

where $[u,\vec{v},\vec{\alpha}]^T$ now refers to the vector of coefficients of the finite element basis functions. For the weak form in (3.34)-(3.35), A_{11} is a mass matrix representing the discrete version of the $L^2(\Omega)$ inner product on $DG_k(\Omega,\tau_h)$ weighted by c_1, A_{22} is the discrete version of the $H(\text{div};\Omega)$ inner product on $RT_{k+1}(\Omega,\tau_h)$ with weight c_0 on the $L^2(\Omega)$ term, B_1 is the weak gradient operator, and B_2 is the $L^2(\Omega)$ inner product on $RT_{k+1}(\Omega,\tau_h)$.

In order to efficiently solve such linear systems, we consider preconditioned Krylov subspace methods. Two families of preconditioners are popular for such block-structured

problems. Block factorization methods [59, 69] approximate Gaussian elimination applied to the blocks of the discretization matrix, writing

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} = \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix},$$

where $S = -(C + BA^{-1}B^T)$ is the Schur complement of A, assuming A is invertible. Natural block preconditioners are of block diagonal, $\mathbf{P}_d$, and block triangular, $\mathbf{P}_t$ form, given as

$$\mathbf{P}_d = \begin{bmatrix} A & 0 \\ 0 & \hat{S} \end{bmatrix}, \quad \mathbf{P}_t = \begin{bmatrix} A & B^T \\ 0 & -\hat{S} \end{bmatrix},$$

where $\hat{S}$ is some approximation of S. The quality of these preconditioners naturally depends on the approximation $\hat{S} \approx S$, and their efficiency depends on the availability of effective fast solvers for the linear systems involving A and $\hat{S}$. Preliminary experiments in this direction revealed some difficulties approximating the Schur complement in the presence of the Nitsche terms that would require further investigation. Therefore, we focus on the development of efficient monolithic multigrid preconditioners [7, 105] in this setting.

We use standard multigrid V-cycles with a direct solve on the coarsest level (taken in the examples to be the mesh with $h = 1/4$ for problems on unit-length domains), and factor-2 coarsening between all grids. These cycles are employed as preconditioners for FGMRES [118]. We use standard interpolation operators, partitioned based on the discretized fields, of the form

$$P = \begin{bmatrix} I_h^k & & \\ & \Pi_h^{k+1} & \\ & & \Pi_h^{k+1} \end{bmatrix},$$

where the blocks I_h^k and Π_h^{k+1} are the natural finite-element interpolation operators for the $DG_k(\Omega, \tau_h)$ and $RT_{k+1}(\Omega, \tau_h)$ spaces, respectively. Coarse-grid operators are formed by rediscretization, which is equivalent to a Galerkin coarse-grid operator for constant $c_0, c_1 \in \mathbb{R}$.

The main challenge with monolithic multigrid methods is to develop an effective relaxation method. In this work, as relaxation we make use of an additive overlapping Schwarz relaxation, which can be considered as a variant of the family of Vanka

relaxation schemes originally proposed in [130] to solve the saddle-point systems that arise from the marker-and-cell (MAC) finite-difference discretization of the Navier-Stokes equations. Vanka relaxation methods encompass a variety of overlapping multiplicative or additive Schwarz methods applied to saddle-point problems, in which the subdomains are chosen so that the corresponding subsystems are also saddle-point systems. Vanka-type relaxation has been used extensively for finite-element discretizations, such as the discretizations arising from the Stokes equations [98], magnetohydrodynamics [6], and liquid crystals [4]. Recently, a general-purpose implementation of patch-based relaxation schemes, including Vanka relaxation, was provided in [68], which we employ via the finite-element discretization package Firedrake [93, 115].

Like other Schwarz methods, Vanka relaxation can be understood algebraically. Denoting the set of all degrees of freedom in the problem by Q, we partition Q into s overlapping subdomains or patches, $Q = \cup_{i=1}^{s} Q_i$, and consider the stationary additive iteration with updates given by

$$x \leftarrow x + \sum_{i=1}^{s} R_i^T \mathcal{A}_{ii}^{-1} R_i (b - \mathcal{A}x),$$

where $\mathcal{A}x = b$ represents the linear system to be solved, R_i is the injection operator from a global vector, x, to a local vector, x_i, on Q_i (with $R_i x = x_i$), and $\mathcal{A}_{ii} = R_i \mathcal{A} R_i^T$ is the restriction of the global system $\mathcal{A}$ to the degrees of freedom in Q_i. While inexact solution of the subdomain problems is relevant when the cardinality of Q_i is large, we consider small subdomain sizes, where direct solution remains practical. We construct the patches, $\{Q_i\}$, topologically, as the so-called star patch around each vertex [68] in the mesh, taking all degrees of freedom on vertex i, on edges adjacent to vertex i, and on all cells adjacent to vertex i to form Q_i. Figure 3.1 shows the subdomain construction around a typical vertex for the cases of discretization using DG_0 and RT_1 elements (left) and using DG_1 and RT_2 elements (right), noting that the RT_k degrees of freedom for both $\vec{v}$ and $\vec{\alpha}$ are included in the patch (and are collocated on the mesh).

There is no guarantee that an unweighted stationary relaxation iteration with the additive Vanka relaxation method would lead to a convergent iteration scheme; however, determining optimal relaxation parameters is difficult. Thus, rather than use the stationary iteration given above, we use two steps of GMRES preconditioned by the Schwarz method as the (pre- and post-) relaxation in the multigrid cycle on

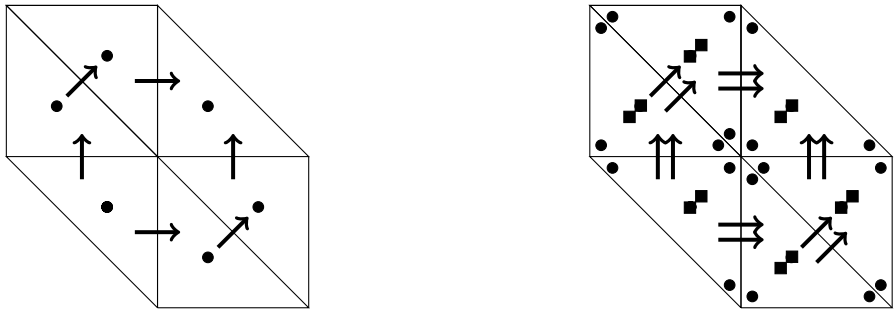

Figure 3.1: Star patches for $DG_0 - RT_1$(left) and $DG_1 - RT_2$(right) discretizations. Filled discs denote DG degrees of freedom, while arrows and filled squares denote edge and interior RT degrees of freedom, respectively.

each level. As noted in Section 3.6, GMRES performs well in giving the relaxation property that we need despite of the lack of analysis that confirms it. We point out that LFA could be used to compute damping parameters to ensure both convergence of the (damped) stationary iteration and some sort of smoothing property, and it can outperform GMRES if the parameters are well-chosen. We, however, leave LFA for future work as it is complicated for this class of problems.

For the case where $\partial\Omega = \Gamma_0 \cup \Gamma_1$ with $c_0 = c_1 = 0$, a modification of the above solver framework is needed. Note that, in this case, while the linear system is well-posed by Theorem 10, the modified weak form in (3.46)-(3.47) has the same structure as (3.53) except that A_{11} becomes the zero matrix and Nitsche boundary terms appear in A_{22}. The approach above performs poorly in this case, as might be expected, particularly with the zero block for A_{11}. To overcome this, we adopt the idea of preconditioning the resulting discretization matrix using an auxiliary operator that corresponds to the discretization of another PDE, related to the inner products in which the PDE is analyzed [91, 103]. Here, since the biharmonic operator is equivalent to the norm in Lemma 5, we add a scaling factor times the $L^2(\Omega)$ inner product in the (1,1) block of the auxiliary operator. That is, the PDE that corresponds to the auxiliary operator is $\Delta^2 u + \chi u$, where χ is a positive constant. Preliminary experiments (reported below in Table 3.4) indicate that choosing the scaling factor χ to be $\mathcal{O}(h^{-1})$ gives better performance than $\mathcal{O}(1)$ values.

3.6 Numerical experiments

In this section, we present numerical experiments to measure finite-element convergence rates and demonstrate the performance of the proposed monolithic multigrid method, stopping when either the residual norm or its relative reduction is less than 10^{-8}. These numerical results were calculated using the finite-element discretization package Firedrake [115], which offers close integration with PETSc for the linear solvers [20, 93]. The relaxation scheme is implemented using the PCPATCH framework [68]. All numerical experiments were run on a workstation with dual 8-core Intel Xeon 1.7 GHz CPUs and 384 GB of RAM. For reproducibility, the codes used to generate the numerical results, as well as the major Firedrake components needed, have been archived on Zenodo [140]. To measure solution quality, we make use of the method of manufactured solutions, prescribing forcing terms and boundary data to exactly match those of a known solution, u_{ex}. Taking uniform meshes, as described below, with representative mesh size h, we define u_h to be the finite-element solution on the mesh, and define the approximation error $e_h = u_{ex} - u_h$. With this, we can define the relative error in the $L^2(\Omega)$ norm on mesh h as $R_e(h) = \frac{\|e_h\|_0}{\|u_{ex}\|_0}$. As needed, we extend these definitions to other quantities, such as the $L^2(\Omega)$ error in $\vec{v}$ and $\vec{\alpha}$, the error in $\vec{v}$ and $\vec{\alpha}$ in the $H(\mathrm{div})$ (semi-)norm, and the error in any boundary terms included in the norms used above.

3.6.1 2D Experiments

In $2D$, we consider experiments on the unit square using uniform "right" triangular meshes (Figure 3.2, left) and on the L-shaped domain with vertices

$$(0,0),\ (0,1),\ \left(\frac{1}{2},1\right),\ \left(\frac{1}{2},\ \frac{1}{2}\right),\ \left(1,\frac{1}{2}\right),\ \text{and } (1,0)$$

using uniform "crossed" triangular meshes (Figure 3.2, right). We consider the smooth exact solution $u_{1ex} = \sin(2\pi x)\cos(3\pi y)$ and an exact solution that is in $H^4(\Omega)$, but not $H^p(\Omega)$ for any integer $p > 4$, given by $u_{2ex} = \left(\sin(2\pi x) + x^{\frac{9}{2}}\right)\left(\cos(3\pi y) + y^{\frac{17}{4}}\right)$.

The discretizations proposed here have larger numbers of degrees of freedom and nonzeros in their matrices than are typically encountered with Lagrange elements for second-order problems. We therefore record the matrix dimensions, N, and number

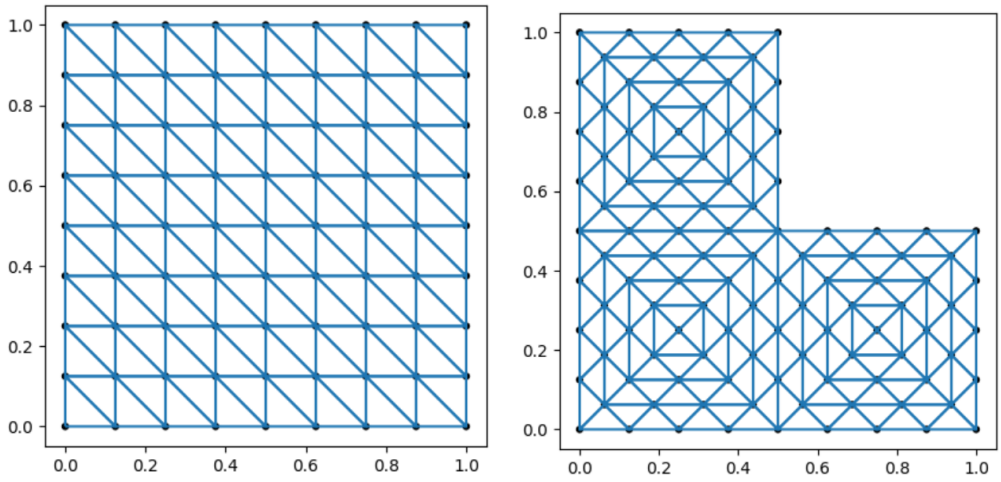

Figure 3.2: Left: unit square domain with uniform right triangular mesh ($h = \frac{1}{8}$). Right: L-shaped domain with uniform crossed triangular mesh ($h = \frac{1}{8}$).

Table 3.1: Dimension, N, and the number of nonzeros, nnz, in the system matrix for $u \in DG_k(\Omega, \tau_h)$, $\vec{v} \in RT_{k+1}(\Omega, \tau_h)$, $\vec{\alpha} \in RT_{k+1}(\Omega, \tau_h)$ on uniform meshes of the unit square domain in 2D.

	$k=0$		$k=1$		$k=2$	
$1/h$	N	nnz	N	nnz	N	nnz
2^6	33,024	352,768	107,008	2,762,752	221,952	10,179,072
2^7	131,584	1,410,048	427,008	11,046,912	886,272	40,707,072
2^8	525,312	5,638,144	1,705,984	44,179,456	3,542,016	162,809,856
2^9	2,099,200	22,548,480	6,819,840	176,701,440	14,161,920	651,202,560

of nonzeros, nnz, in Table 3.1 for the discretizations on the unit square, for several levels of refinement, h, and orders of the discretization, k. In all figures, we use blue, red, and green lines to present results for $k = 0, 1, 2$, respectively, with filled discs denoting the measured (u, v) error in the $L^2 \times H(\text{div})$ norm, and squares denoting the error in α measured in the $H(\text{div})$ norm. The values of m approximate slopes of the lines and, therefore, the convergence rates.

We present two-dimensional numerical experiments in two parts. First, in Subsection 3.6.1, we investigate the finite-element convergence and provide a comparison between direct solvers and the multigrid-preconditioned FGMRES algorithm for H^2 elliptic problems with $c_0 \geq 0$ and $c_1 > 0$. Then, in Subsection 3.6.1, we focus on the classical biharmonic problem, i.e., $c_0 = c_1 = 0$, where we investigate discretization errors in one case that requires the use of the Nitsche penalty method and a second case

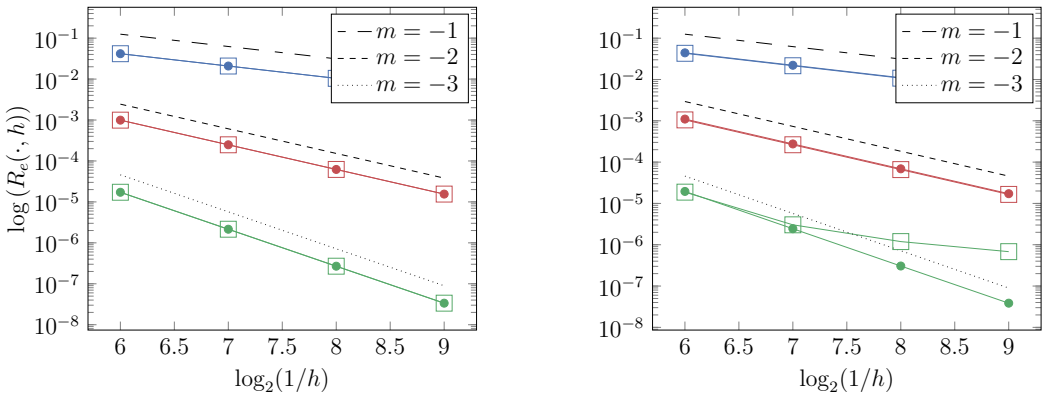

Figure 3.3: Relative approximation errors and rate of convergence for the unit square domain with $c_0 = 0$, $c_1 = 1$, $\partial\Omega = \Gamma_0 \cup \Gamma_3$ and $(u, \vec{v}, \vec{\alpha}) \in DG_k(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h)$, $k = 0, 1, 2$. Blue, red, and green lines present results for $k = 0, 1, 2$, respectively. Left: smooth solution $u_{ex} = u_{1ex}$. Right: rough solution $u_{ex} = u_{2ex}$.

that does not. Finally, we present multigrid-preconditioned FGMRES iteration counts for the biharmonic problem with clamped boundary conditions, which is challenging due to the inclusion of the Nitsche boundary terms.

2D experiments with positive c_1

We consider the unit square domain and plot $\log(R_e(\cdot, h))$ against $\log_2(1/h)$, so that the slopes of the lines represent the experimentally measured convergence rates for $(u, \vec{v}, \vec{\alpha}) \in DG_k(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h)$, for $k = \{0, 1, 2\}$. Let $\partial\Omega = \Gamma_N \cup \Gamma_S \cup \Gamma_E \cup \Gamma_W$, meaning the North, South, East, and West faces of the square. Figure 3.3 presents results for the problem with $c_0 = 0$ and $c_1 = 1$ with boundary $\partial\Omega = \Gamma_0 \cup \Gamma_3$, where $\Gamma_0 = \Gamma_E \cup \Gamma_W$, and $\Gamma_3 = \Gamma_N \cup \Gamma_S$. Figure 3.4 presents results when $c_0 = 2$, $c_1 = 4$, and $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$ with $\Gamma_0 = \Gamma_E \cup \Gamma_W$, $\Gamma_2 = \Gamma_S$, and $\Gamma_3 = \Gamma_N$. Since we omit Γ_1 from these examples, there is no need to use the Nitsche boundary terms considered in that case. We note that we see optimal convergence for all k with $(u, \vec{v})$ in the $L^2 \times H(\text{div})$ norm and $\vec{\alpha}$ in the $H(\text{div})$ norm for the smooth exact solution, u_{1ex}, on the left of these figures. Considering the $H^4(\Omega)$ solution, u_{2ex}, on the right, we see optimal convergence for small k, but degraded performance for $k = 2$, where the lack of smoothness in α is reflected in the numerical results. These results are consistent with the analysis in Corollaries 7 and 8, although we note that the $H^4(\Omega)$ case outperforms the expected convergence from the analysis.

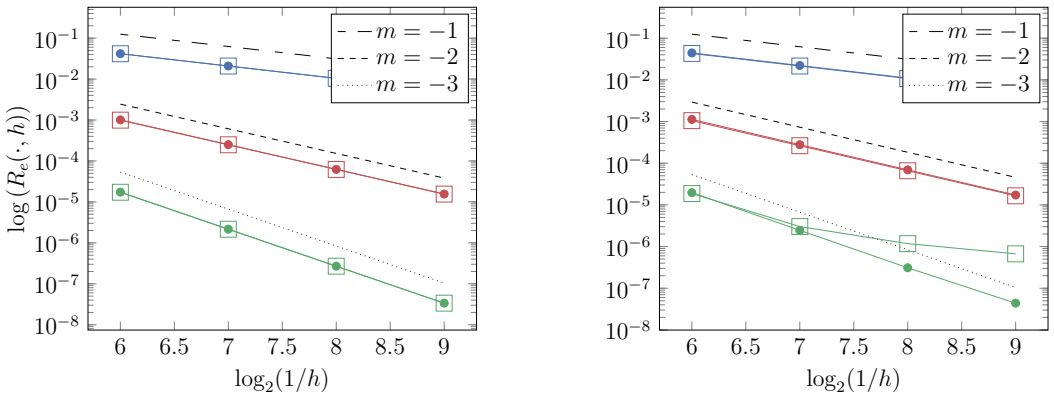

Figure 3.4: Relative approximation errors and rates of convergence for the unit square domain with $c_0 = 2$, $c_1 = 4$, $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$ and $(u, \vec{v}, \vec{\alpha}) \in DG_k(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h)$, $k = 0, 1, 2$. Blue, red, and green lines present results for $k = 0, 1, 2$, respectively. Left: smooth solution $u_{ex} = u_{1ex}$. Right: rough solution $u_{ex} = u_{2ex}$.

To demonstrate the effectiveness of the monolithic multigrid preconditioner, Table 3.2 presents iteration counts and CPU times to solution for both the multigrid-preconditioned FGMRES iterations and the use of a direct solver (MUMPS [11], via the PETSc interface) for the unit square domain with $(u, \vec{v}, \vec{\alpha}) \in DG_2(\Omega, \tau_h) \times RT_3(\Omega, \tau_h) \times RT_3(\Omega, \tau_h)$ and $\partial\Omega = \Gamma_0 \cup \Gamma_3$. We note that the iteration counts for monolithic multigrid-preconditioned FGMRES are consistent through all runs and mesh sizes, and that the scaling of wall-clock time for this approach is $\mathcal{O}(N)$ or better throughout. While the direct solver is slightly faster for small mesh sizes, we see worse than $\mathcal{O}(N)$ scaling for the wall-clock time with MUMPS at larger mesh sizes, showing the expected behaviour. Moreover, as we vary the number of processors over which we parallelize the computation, we see that, for sufficiently large problems, we have good strong parallel scalability with the monolithic multigrid solver, although MUMPS is always faster than our multigrid implementation for this two-dimensional problem. Table 3.3 presents the case of $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$, $(u, \vec{v}, \vec{\alpha}) \in DG_1(\Omega, \tau_h) \times RT_2(\Omega, \tau_h) \times RT_2(\Omega, \tau_h)$. As we increase number of processors from 4 to 16, we see better performance with the multigrid solver than the direct solver. Again, we have good strong parallel scalability with the monolithic multigrid solver, showing 3.98x speedup for the 1024^2 mesh, while the direct solver (MUMPS) shows only 1.54x speedup.

Table 3.2: Wall-clock time (in seconds) and iterations to convergence with varying numbers of processors, p, for monolithic multigrid and a direct solver (MUMPS) for the unit square domain with $c_0 = 0$, $c_1 = 1$, $\partial\Omega = \Gamma_0 \cup \Gamma_3$ and $(u, \vec{v}, \vec{\alpha}) \in DG_2(\Omega, \tau_h) \times RT_3(\Omega, \tau_h) \times RT_3(\Omega, \tau_h)$.

h^{-1}	Monolithic			MUMPS	
	Iterations	Time ($p = 1$)	Time ($p = 4$)	Time ($p = 1$)	Time ($p = 4$)
2^6	5	27.99	11.63	5.44	2.82
2^7	5	102.18	35.35	21.45	11.40
2^8	5	405.00	127.43	93.27	46.84
2^9	5	1621.38	569.88	438.91	218.88

Table 3.3: Wall-clock time (in seconds) and iterations to convergence with varying numbers of processors, p, for monolithic multigrid and a direct solver (MUMPS) for the unit square domain with $c_0 = 2$, $c_1 = 4$, $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$ and $(u, \vec{v}, \vec{\alpha}) \in DG_1(\Omega, \tau_h) \times RT_2(\Omega, \tau_h) \times RT_2(\Omega, \tau_h)$.

h^{-1}	Monolithic			MUMPS	
	Iterations	Time ($p = 4$)	Time ($p = 16$)	Time ($p = 4$)	Time ($p = 16$)
2^7	5	13.91	7.15	4.97	3.29
2^8	5	44.31	16.08	20.72	13.30
2^9	5	171.90	50.39	93.93	62.33
2^{10}	5	843.30	211.65	446.23	289.33

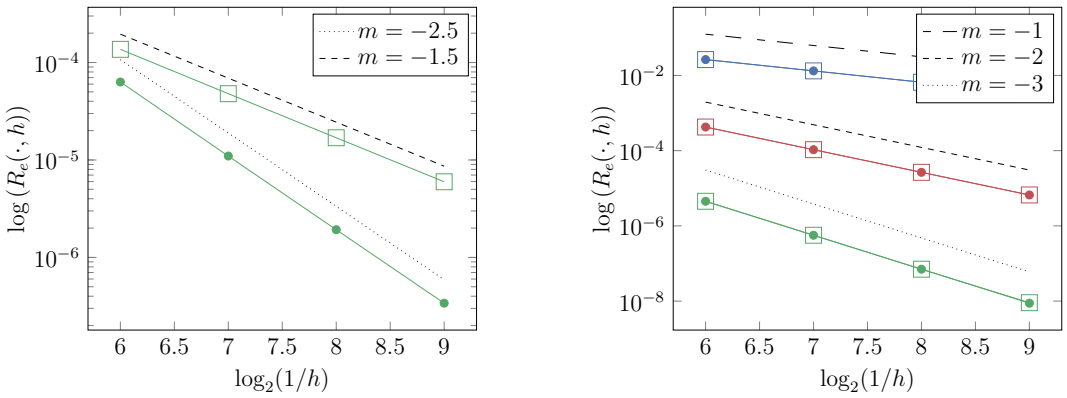

Figure 3.5: Relative approximation errors and rate of convergence for the biharmonic problem in the unit square domain and $\partial\Omega = \Gamma_0 \cup \Gamma_1 \cup \Gamma_3$ (left), and the L-shaped domain with $\partial\Omega = \Gamma_0$ (right). Blue, red, and green lines present results for $k = 0, 1, 2$, respectively.

The 2D biharmonic

We next consider the classical biharmonic problem with the exact solution u_{1ex}. Figure 3.5 (left) shows results for the case where $\partial\Omega = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$, and we take $(u, \vec{v}, \vec{\alpha}) \in DG_2(\Omega, \tau_h) \times RT_3(\Omega, \tau_h) \times RT_3(\Omega, \tau_h)$. For this choice ($k = 2$), Corollary 11 gives an expected convergence rate of $\mathcal{O}(h^2)$ for $(u, \vec{v})$ and of $\mathcal{O}(h)$ for $\vec{\alpha}$. In contrast with that result, we observe almost $\mathcal{O}(h^{5/2})$ for $(u, \vec{v})$ in the modified $L^2 \times H(\text{div})$ norm and $\mathcal{O}(h^{3/2})$ convergence for $\vec{\alpha}$. Figure 3.5 (right) considers the L-shaped domain with $\partial\Omega = \Gamma_0$ and $k = 0, 1, 2$. Here, we observe optimal convergence rates. We note that this is a test using a manufactured solution and not a generic smooth forcing function, so the lack of full regularity from the domain is not expected to degrade the expected convergence rates.

Remark 3.6.1. For the right-triangular meshes given in Figure 3.2 (Left), $\gamma_1(2) \approx 41$ and $\gamma_1(3) \approx 62$. Therefore, we choose $\lambda = 125,$ and 210 respectively for the numerical results given in Figure 3.5(Left) and Table 3.4. Note that, for both cases, $\lambda > 3\gamma_1$, which is necessary to satisfy the coercivity condition (3.49) for the bilinear form $\hat{a}$.

Finally, we consider the classical biharmonic operator with clamped boundary conditions, i.e., $c_0 = c_1 = 0$ and $\partial\Omega = \Gamma_1$ and $(u, \vec{v}, \vec{\alpha}) \in DG_3(\Omega, \tau_h) \times RT_4(\Omega, \tau_h) \times RT_4(\Omega, \tau_h)$. Table 3.4 shows the effectiveness of the monolithic multigrid solver with an $\mathcal{O}(h^{-1})$ weight on the auxiliary operator. Dashes in the table mean that more than 100 iterations were required to converge when the residual norm or its relative reduction is less than 10^{-14}. We note that, due to a technical limitation in PCPATCH

Table 3.4: Number of iterations to converge with different weights on the auxiliary operator. Here $(u, \vec{v}, \vec{\alpha}) \in DG_3(\Omega, \tau_h) \times RT_4(\Omega, \tau_h) \times RT_4(\Omega, \tau_h)$. A dash means that convergence was not achieved in 100 iterations.

h^{-1} \ weight	1	10	20	40	80	h^{-1}
2^6	25	13	13	12	12	12
2^7	-	18	14	12	12	12
2^8	-	43	24	15	13	11
2^9	-	-	-	-	90	11

(where Nitsche boundary terms cannot be treated), these results use an alternate implementation of the star relaxation scheme that is less efficient than PCPATCH. Consequently, we do not report timings for these experiments, as they are not comparable to the timings reported elsewhere in this paper.

3.6.2 $3D$ experiments

Here, we consider a test case on the unit cube, with right-hand side and boundary conditions chosen so that the exact solution is $u_{ex} = \sin(2\pi x)\cos(3\pi y)\sinh(\pi z)$. Finite-element convergence is demonstrated in Figure 3.6 for $k \in \{0, 1, 2\}$ with $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$, with Γ_0 corresponding to $z = 0$ and $z = 1$, Γ_2 corresponding to $y = 0$ and $y = 1$, and Γ_3 corresponding to $x = 0$ and $x = 1$, showing convergence consistent with the analysis of Corollary 8. Table 3.5 details the performance of the monolithic multigrid-preconditioned FGMRES solver for $k = 0$, compared with a standard direct solver (MUMPS). We see excellent performance of the monolithic multigrid method, with iteration counts that are independent of problem size and CPU time scaling linearly with problem size, and decreasing with parallelization for sufficiently large problems. In contrast, we see the expected rapid growth of required CPU times for MUMPS, and suboptimal parallel scaling, showing the utility and power of the monolithic multigrid approach.

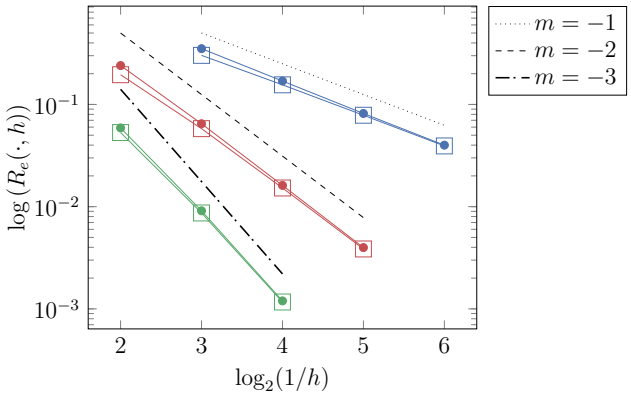

Figure 3.6: Relative approximation errors and rates of convergence for the unit cube domain $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$, $c_0 = 4$ and $c_1 = 2$. Blue, red, and green lines present results for $k = 0, 1, 2$, respectively.

Table 3.5: Wall-clock time (in seconds) and iterations to convergence with varying numbers of processors, p, for monolithic multigrid and a direct solver (MUMPS) for the unit cube domain with $c_0 = 4$, $c_1 = 2$, $\partial\Omega = \Gamma_0 \cup \Gamma_2 \cup \Gamma_3$ and $(u, \vec{v}, \vec{\alpha}) \in DG_0(\Omega, \tau_h) \times RT_1(\Omega, \tau_h) \times RT_1(\Omega, \tau_h)$.

h^{-1}	Monolithic			MUMPS	
	Iterations	Time ($p = 1$)	Time ($p = 8$)	Time ($p = 1$)	Time ($p = 8$)
2^3	9	5.55	2.80	1.51	0.69
2^4	9	31.10	7.30	11.05	3.75
2^5	9	229.65	38.31	154.33	44.07
2^6	9	1847.65	280.42	5173.58	1167.10

3.7 Conclusion

We consider the mixed finite-element approximation of solutions to H^2-elliptic fourth-order problems, achieved by the transformation of the fourth-order equation into a system of PDEs. We find that under natural assumptions on the coefficients of the problem, three combinations of boundary conditions lead to optimal finite-element convergence. For the fourth case of boundary conditions ("clamped" boundary conditions, on the solution and its normal derivative), suboptimal rates of convergence are expected and observed when implemented using Nitsche's method. While the approach is applicable in both two and three dimensions, we note that it is particularly attractive in 3D, where the cost of conforming methods is prohibitive. It remains an open question whether or not it is possible to employ alternative approaches (such as adapting the Nitsche boundary conditions, or the use of alternative penalty approaches) to regain optimal finite-element convergence for the boundary conditions where suboptimal convergence is proven and observed here. We additionally propose a monolithic multigrid algorithm with optimal scaling for the resulting discrete linear systems. For three-dimensional problems, this approach yields a preconditioned FGMRES iteration that dramatically outperforms state-of-the-art direct solvers.

Chapter 4

Finite-element discretization of the smectic density equation

Abstract[1]

The density variation of smectic A liquid crystals is modelled by a fourth-order PDE, which exhibits two complications over the biharmonic or other typical H^2-elliptic fourth-order problems. First, the equation involves a "Hessian-squared" (div-div-grad-grad) operator, rather than a biharmonic (div-grad-div-grad) operator. Secondly, while positive-definite, the equation has a "wrong-sign" shift, making it somewhat more akin to a Helmholtz operator, with lowest-energy modes arising from certain plane waves, than an elliptic one. In this paper, we analyze and compare three finite-element formulations for such PDEs, based on H^2-conforming elements, the C^0 interior penalty method, and a mixed finite-element formulation that explicitly introduces approximations to the gradient of the solution and a Lagrange multiplier. The conforming method is simple but is impractical to apply in three dimensions; the interior-penalty method works well in two and three dimensions but has lower-order convergence and (in preliminary experiments) seems difficult to precondition; the mixed method uses more degrees of freedom, but is amenable to monolithic multigrid preconditioning. Numerical results verify the finite-element convergence for all discretizations, and illustrate the trade-offs between the three schemes.

[1]This work is to be submitted as "Finite-element discretization of the smectic density equation" by Patrick E. Farrell, Abdalaziz Hamdan, and Scott P. MacLachlan.

4.1 Introduction

Recent years have seen significant and successful effort in developing numerical models of various liquid crystalline materials [4, 5, 9, 23, 53, 100, 112, 114]. In these models, equilibrium states of liquid crystals usually correspond to minimizers of a given energy functional, which can be directly discretized using finite-element (or other) variational techniques. Smectic A liquid crystals are characterized by their natural propensity to form layers with periodic variation in the density of the liquid crystal aligned with the orientation of the molecules. While some models make use of a complex order parameter as a model of the energy of liquid crystals [55], several recent papers have proposed models based on a real-valued density variation [22, 112, 138]. For example, Pevnyi et al. [112] propose a model

$$E(u,\vec{\nu}) = \int_\Omega \frac{a}{2}u^2 + \frac{b}{3}u^3 + \frac{c}{4}u^4 + B\left|\nabla\nabla u + q^2\vec{\nu}\otimes\vec{\nu}u\right|^2 + \frac{K}{2}|\nabla\vec{\nu}|^2, \tag{4.1}$$

where $\Omega \subset \mathbb{R}^d$, for $d \in \{2,3\}$, $u : \Omega \to \mathbb{R}$ represents the variation in the density of the liquid crystal from its average density, $\vec{\nu}$ is the unit-length director of the liquid crystal (the local axis of average molecular alignment), and a, b, c, q, K, and B are real-valued constants determined by the liquid crystal under consideration. Of these, the smectic wavenumber, q, is notable because it prescribes a preferred wavelength for the solution of $2\pi/q$. Here, and in what follows, we use $|\boldsymbol{T}|^2 = \boldsymbol{T} : \boldsymbol{T}$ to denote the Frobenius norm squared of a tensor $\boldsymbol{T}$ (of any rank), defined as the sum of squares of the entries in $\boldsymbol{T}$ at a given point in Ω.

It is well-known that representing the orientation of the liquid crystal with a vector-valued director cannot represent certain defects of the liquid crystal [21]. In [138], Xia et al. adapted (4.1) to make use of a tensor-valued order parameter in place of the director field, proposing

$$E(u,\boldsymbol{Q}) = \int_\Omega \frac{a}{2}u^2+\frac{b}{3}u^3+\frac{c}{4}u^4+B\left|\nabla\nabla u + q^2\left(\boldsymbol{Q}+\frac{I_d}{d}\right)u\right|^2+\frac{K}{2}|\nabla\boldsymbol{Q}|^2+f_n(\boldsymbol{Q}), \tag{4.2}$$

where $\boldsymbol{Q}$ is the tensor-valued order parameter, I_d is the $d \times d$ identity matrix, and $f_n(\boldsymbol{Q}) = -l\,\mathrm{tr}(\boldsymbol{Q}^2) + l\left(\mathrm{tr}(\boldsymbol{Q}^2)\right)^2$ for $d = 2$ and $f_n(\boldsymbol{Q}) = -l\,\mathrm{tr}(\boldsymbol{Q}^2) - \frac{l}{3}\mathrm{tr}(\boldsymbol{Q}^3) + \frac{l}{2}\left(\mathrm{tr}(\boldsymbol{Q}^2)\right)^2$ in three dimensions. Here, the penalty parameter, l, and the functions $f_n(\boldsymbol{Q})$ are chosen so that the minimizer of $\int_\Omega f_n(\boldsymbol{Q})$ is of the form $\boldsymbol{Q} = \vec{\nu}\otimes\vec{\nu} - \frac{I_d}{d}$, and

are included in the energy to weakly enforce the rank-one condition implied by (4.1). A related model was proposed by Ball & Bedford [22]. While there remain many open questions about the physical values of the constants a, b, c, q, K, and B, an important feature of these models is the energetic competition between the term encouraging alignment of the orientation and density variation, scaled by B, and the deformation of the director field, scaled by K. The Euler–Lagrange equations for either of these functionals naturally lead to a coupled system of PDEs, with a fourth-order operator acting on u and a second-order operator acting on $\vec{\nu}$ or $\boldsymbol{Q}$. While the discretization of the vector or tensor Laplacian is relatively standard, the fourth-order PDE involving u, which we refer to as the "smectic density equation," is of a type not previously studied in the literature.

Motivated by such examples, we consider minimization of a simplification of Equation (4.2) with suitable boundary conditions, given in variational form as

$$\min_{v \in H^2(\Omega)} \frac{B}{2} \int_\Omega \left| \nabla\nabla v + q^2 \boldsymbol{T} v \right|^2 + \frac{m}{2} \int_\Omega v^2 - \int_\Omega f v, \tag{4.3}$$

where, motivated by the above, $\boldsymbol{T}$ is a given bounded $d \times d$ tensor, with $|\boldsymbol{T}|^2 = \boldsymbol{T} : \boldsymbol{T} \leq \mu_1$ for $\mathcal{O}(1)$ constant μ_1, while m and q are $\mathcal{O}(1)$ positive constants, and $0 < B \leq 1$. We assume $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, to be a bounded simply connected polytopal domain with Lipschitz boundary. Note that if $b = c = 0$, and $\boldsymbol{Q}$ is fixed in (4.2), then Energies (4.2) and (4.3) are the same, but the source f is enforced in (4.3). Sufficiently smooth extremizers of this energy must satisfy its Euler–Lagrange equations, which yield the fourth-order smectic density equation,

$$B\nabla \cdot \nabla \cdot \left(\nabla\nabla u + q^2 \boldsymbol{T} u \right) + Bq^2 \boldsymbol{T} : \nabla\nabla u + (Bq^4 \boldsymbol{T} : \boldsymbol{T} + m) u = f. \tag{4.4}$$

We consider finite-element formulations for this fourth-order problem, with a particular focus on the treatment of the boundary conditions that arise naturally from the transition from the variational to strong forms. Two formulations are based on discretizing (4.3), using either H^2-conforming or C^0 interior penalty (C0IP) methods. The third formulation is based on mixed finite-element principles, introducing the gradient of the solution as an explicit variable constrained using a Lagrange multiplier, and leverages standard discretizations for the Stokes problem in order to achieve inf-sup stability. Both the C0IP and mixed approaches are quite general, in the sense

that we achieve high-order convergence when using high-order elements if the solution is sufficiently smooth. However, these formulations provide slightly suboptimal convergence, as shown later. Complications to achieving optimal convergence come from the use of Nitsche's method to weakly enforce essential boundary conditions, weakly imposing C^1 continuity in the C0IP approach, and the spaces chosen to achieve inf-sup stability in the mixed approach.

Finite-element methods for fourth-order H^2-elliptic problems have been extensively studied. These include conforming methods, such as the use of Argyris elements, nonconforming methods [41, 52, 131], C^0 interior penalty methods (C0IP, which are also nonconforming) [18,40,42,127], and mixed-finite element methods, including two-field [51, 52, 102], three-field [24, 66], and four-field discretizations [27, 50, 97]. While conforming methods are attractive in two dimensions, the natural extension of Argyris elements to $\mathbb{R}^3$ requires the use of ninth-order polynomials on each element [141], which is prohibitively expensive in comparison to low-order methods. While a C0IP method was used in [138] and is analyzed herein, preliminary experiments showed that it is difficult to develop effective preconditioners for this discretization, motivating the consideration of alternate approaches. Thus, we also propose a mixed finite-element discretization of (4.3) that does not require growth in polynomial order in three dimensions, and which we expect to be more amenable to the development of effective preconditioners, similar to those in [66].

An additional challenge in considering the models of smectic LCs in [22, 112, 138] is that of proper treatment of the boundary conditions. In particular, these models typically include only natural BCs on the density variation, u, and do not strongly impose Dirichlet boundary conditions, such as the "clamped" boundary conditions that are commonly considered for the biharmonic problem. Indeed, the case of clamped boundary conditions, where the Hessian and Laplacian weak forms of a fourth-order operator are equivalent, has been extensively studied [27, 42, 50, 51, 97, 102]. A central question in this work is how to treat the more general forms of boundary conditions that arise when moving from the variational form in (4.3) to the strong form in (4.4), summarized in (4.13)-(4.16) below. To our knowledge, existing results in the literature treat the cases of clamped boundary conditions, Cahn-Hilliard boundary conditions (a special case of those in (4.16) when $q = 0$) [40], and simply supported boundary conditions (a special case of those in (4.13)) [39], but not the case of full Neumann boundary conditions. Given the simply supported boundary conditions, existence of

minimizers of (4.2) when $q \geq 0$ as well as error estimates for its discretization using C^0 interior penalty methods when $q = 0$ were obtained in [137]. Here, we prove well-posedness of (4.3) and provide error estimates for its discretization using Argyris elements, C0IP, and a mixed finite-element method.

This paper is organized as follows. A brief summary of the tools needed for the finite-element analysis is presented in Section 4.2. The continuum analysis, including the weak forms and uniqueness theory, is presented in Section 4.3. In Section 4.4, we present the conforming, C0IP, and mixed finite-element methods, and analysis of both well-posedness and error estimates for these methods. Finally, numerical experiments to compare the different finite-element methods are presented in Section 4.5.

4.2 Preliminaries

We recall the standard Sobolev spaces

$$\begin{aligned} H^1_\Gamma(\Omega) &= \left\{u \in H^1(\Omega) \middle|\ u = 0 \text{ on } \Gamma\right\}, \\ H_\Gamma(\mathrm{div};\Omega) &= \left\{\vec{v} \in H(\mathrm{div};\Omega) \middle|\ \vec{v}\cdot\vec{n} = 0 \text{ on } \Gamma\right\}, \end{aligned}$$

where $\Gamma \subset \partial\Omega$ and $\vec{n}$ is the outward unit normal to Γ. Let $\{\tau_h\}$ be a quasiuniform family of triangulations of Ω, and let $CG_k(\Omega,\tau_h)$, $DG_k(\Omega,\tau_h)$, and $RT_k(\Omega,\tau_h)$ be the standard continuous Lagrange, discontinuous Lagrange, and Raviart-Thomas approximation spaces of degree k, respectively, on mesh τ_h. We also define $CG^\Gamma_k(\Omega,\tau_h) = CG_k(\Omega,\tau_h) \cap H^1_\Gamma(\Omega)$ and $RT^\Gamma_k(\Omega,\tau_h) = RT_k(\Omega,\tau_h) \cap H_\Gamma(\mathrm{div};\Omega)$.

Remark 4.2.1. In what follows, we use C to represent a generic positive constant that can depend on the domain, shape regularity of the triangulation, τ_h, and the polynomial degree k of the finite-element space, but not on the mesh parameter, h, nor the smectic wavenumber, q, and may be different in different instances. Where needed, we will use $\left\{C_i\right\}$ to denote different arbitrary constants in the same expression.

To prove well-posedness of the original continuum problem, we make use of a standard estimate of $\|u\|^2_{0,\partial\Omega} = \int_{\partial\Omega} u^2 ds$.

Theorem 12. *[77, Theorem 1.5.1.10] There exists a constant $C > 0$, such that*

$$\|u\|_{0,\partial\Omega}^2 \leq C \left[\epsilon^{\frac{1}{2}} \|\nabla u\|_0^2 + \frac{1}{\epsilon^{\frac{1}{2}}} \|u\|_0^2 \right], \tag{4.5}$$

for all $u \in H^1(\Omega)$ and $\epsilon \in (0, 1)$.

As is typical in the analysis of finite-element methods, our well-posedness results will rely on a Poincaré inequality. In order to treat a more general set of boundary conditions, we state a more general form of the standard inequality.

Lemma 6. *(Poincaré Inequality [38, 44, 107, 135]) If $u \in H^j(\Omega),\ j \in \{1, 2\}$, then*

$$\|u\|_{j-1}^2 \leq C \left(|u|_j^2 + \xi^2(u) \right),$$

where ξ is any seminorm on $H^j(\Omega)$ with the properties that

- *There exists a constant $C > 0$ such that, for all $u \in H^j(\Omega)$, we have*

$$\xi(u) \leq C \|u\|_j.$$

- *If a is a polynomial of degree less than j (i.e., a constant function if $j = 1$ or linear function if $j = 2$), $\xi(a) = 0$ if and only if $a = 0$.*

In particular, for $j = 1$ and any functions ψ_1 and ψ_2 that are square integrable on Ω and $\partial\Omega$, respectively, with $\int_\Omega \psi_1 \neq 0$, and $\int_{\partial\Omega} \psi_2 \neq 0$, then the Poincaré inequality above holds for either seminorm $\xi_1(u)$ or $\xi_2(u)$ [38], defined as

$$\xi_1(u) = \left| \int_\Omega \psi_1 u \right|, \quad \xi_2(u) = \left| \int_{\partial\Omega} \psi_2 u \right|.$$

Note that the choices of $\psi_1 = 1$ on Ω, and $\psi_2 = 1$ on $\partial\Omega$ lead to the classical Poincaré inequalities that are most commonly used. For $j = 2$, we will use $\xi(u) = \|u\|_0$, which satisfies the conditions above.

A useful function in our analysis is the solution of $-\Delta S = 1$ with homogeneous Dirichlet boundary conditions. We next recall some properties of this function and its discrete approximation.

Lemma 7. *The weak solution of $-\Delta S = 1$, with homogeneous Dirichlet boundary conditions, has positive mean and bounded H^1 norm. In addition, the discrete solution, $S_h \in CG_1^{\partial\Omega}(\Omega, \tau_h)$, of $\langle \nabla S_h, \nabla v\rangle = \langle 1, v\rangle$ for all $v \in CG_1^{\partial\Omega}(\Omega, \tau_h)$ has the same properties.*

Proof. The solutions, S and S_h , satisfy $\|\nabla S\|_0^2 = \int_\Omega S$ and $\|\nabla S_h\|_0^2 = \int_\Omega S_h$. Therefore, S and S_h have positive means. In addition, the H^1-regularity results that $\|S\|_1 \leq C\|1\|_0 = C\sqrt{\text{Area}(\Omega)}$ and $\|S_h\|_1 \leq C\sqrt{\text{Area}(\Omega)}$ also hold [32]. □

Our well-posedness proofs also rely on the standard Helmholtz decomposition in 2D, which we first state for the continuum case. We use the standard definition of the curl of a scalar function in 2D, as $\nabla \times p = \begin{bmatrix} p_y \\ -p_x \end{bmatrix}$.

Lemma 8. *(2D continuous Helmholtz decomposition [14, 74, 120]) Let $\partial\Omega = \Gamma_a \cup \Gamma_b$, where Γ_a and Γ_b are disjoint. For $\vec{\alpha} \in H_{\Gamma_a}(\text{div}; \Omega)$, the following Helmholtz decomposition holds*

$$\vec{\alpha} = \nabla\phi + \nabla \times p, \tag{4.6}$$

where $p \in H^1_{\Gamma_a}(\Omega)$, and $\phi \in H^1_{\Gamma_b}(\Omega)$ is the solution of $\int_\Omega \nabla\phi \cdot \nabla\chi = -\int_\Omega \nabla \cdot \vec{\alpha}\chi,\ \forall\chi \in H^1_{\Gamma_b}(\Omega)$. Furthermore, p is a zero-mean function if $\Gamma_a = \emptyset$, and ϕ is a zero-mean function if $\Gamma_b = \emptyset$. This decomposition is orthogonal in the L^2 and $H(\text{div})$ norms.

Remark 4.2.2. Assuming, in addition to the assumptions of Lemma 8 that $\vec{\alpha} \in [H^{t+2}(\Omega)]^2 \cap H_{\Gamma_a}(\text{div}; \Omega)$, $\nabla \cdot \vec{\alpha} \in H_0^{t+1}(\Omega)$, for $t \geq 0$, $\phi \in H^1_{\Gamma_b}(\Omega)$ is the solution of the mixed boundary value problem of the form $\int_\Omega \nabla\phi \cdot \nabla\chi = -\int_\Omega \nabla \cdot \vec{\alpha}\chi,\ \forall\chi \in H^1_{\Gamma_b}(\Omega)$, and the partition into Γ_a and Γ_b satisfies the conditions in [77, Theorem 5.1.1.5] for there to be no admissible "singular solutions" to the variational problem. Then, $\phi \in H^{t+3}(\Omega) \cap H^1_{\Gamma_b}(\Omega)$, and p is at least in $\varrho = \{p \in H^{t+2}(\Omega) \cap H^1_{\Gamma_a} \mid \nabla \times p \in [H^{t+2}(\Omega)]^2\}$.

To see that this regularity can be achieved, we next state a regularity result for solution of the Poisson problem with mixed boundary conditions on the unit square.

Lemma 9. *Let u be the solution of*

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega, \\ u &= 0, && \text{on } \mathscr{D}, \\ \frac{\partial u}{\partial \vec{n}} &= 0, && \text{on } \mathscr{N}, \end{aligned}$$

where $f \in H_0^t(\Omega)$, $t > 0$, *and* $\Omega = (0,1)^2$, $\partial\Omega = \Gamma_N \cup \Gamma_S \cup \Gamma_E \cup \Gamma_W$, *where* Γ_N, Γ_S, Γ_E, *and* Γ_W *are the North, South, East, and West faces of the square, respectively, and* $\mathscr{D}$, $\mathscr{N}$ *are the faces on which we impose Dirichlet and Neumann boundary conditions, respectively. Then, if any of the following conditions hold,* $u \in H^{t+2}(\Omega)$,

- $\mathscr{D} = \partial\Omega$ *or* $\mathscr{N} = \partial\Omega$ *, and* t *is even,*
- $\mathscr{D} = \Gamma_E \cup \Gamma_W$ *and* $\mathscr{N} = \Gamma_N \cup \Gamma_S$*, and* t *is odd, or*
- $\mathscr{D} = \Gamma_N \cup \Gamma_S$ *and* $\mathscr{N} = \Gamma_E \cup \Gamma_W$*, and* t *is odd.*

Proof. This result is a direct consequence of [77, Theorem 5.1.1.5]. Note that, in the notation of [77], we have $\Phi_j = 0$ if edge $\Gamma_j \in \mathscr{N}$, and $\Phi_j = \frac{\pi}{2}$ if edge $\Gamma_j \in \mathscr{D}$. Then, if j and $j+1$ denote adjacent edges, we require that $\frac{\Phi_j - \Phi_{j+1} + (t+1)\frac{\pi}{2}}{\pi}$ is not an integer to achieve the stated regularity result, which is guaranteed for the cases given above. In addition, for each pair of adjacent edges, $2\frac{\Phi_j - \Phi_{j+1} + m\pi}{\pi}$ is an integer for any integer m, which precludes any singular solutions. These conditions are sufficient to guarantee that $u \in H^{t+2}(\Omega)$. □

We also make use of the discrete analogue of Lemma 8, stated next.

Lemma 10. *[14, 15] Under the same assumptions as Lemma 8, the Helmholtz decomposition of* $RT_{k+1}^{\Gamma_a}(\Omega, \tau_h)$ *is*

$$RT_{k+1}(\Omega, \tau_h) = \left(\nabla_h^{\Gamma_a} DG_k(\Omega, \tau_h)\right) \oplus \left(\nabla \times CG_{k+1}^{\Gamma_a}(\Omega, \tau_h)\right), \tag{4.7}$$

where $\nabla_h^{\Gamma_a}$ *is the discrete gradient operator,* $\nabla_h^{\Gamma_a} : DG_k(\Omega, \tau_h) \to RT_{k+1}^{\Gamma_a}(\Omega, \tau_h)$*, such that*

$$\int_\Omega \nabla_h^{\Gamma_a} u \cdot \vec{v} = -\int_\Omega u \nabla \cdot \vec{v}, \quad \forall \vec{v} \in RT_{k+1}^{\Gamma_a}(\Omega, \tau_h). \tag{4.8}$$

This decomposition is orthogonal in the L^2 *and* $H(\mathrm{div})$ *norms.*

Finally, we note that the Helmholtz decomposition allows us to define an alternative norm on $H(\mathrm{div}; \Omega)$, which will be useful in the later analysis.

Remark 4.2.3. Let $\vec{\alpha}_1$ and $\vec{\alpha}_2 \in H(\mathrm{div}; \Omega)$ with $\vec{\alpha}_1 = \nabla\phi_1 + \nabla \times p_1$ and $\vec{\alpha}_2 = \nabla\phi_2 + \nabla \times p_2$, computed as in Lemma 8. The following defines an inner product on

$H(\mathrm{div};\Omega)$:

$$(\vec{\alpha}_1,\vec{\alpha}_2)_{\mathrm{Div}} = q^{-4}\left(\int_\Omega p_1 p_2 + \int_\Omega \nabla\phi_1\cdot\nabla\phi_2 + \int_\Omega \nabla\cdot\vec{\alpha}_1\nabla\cdot\vec{\alpha}_2\right), \tag{4.9}$$

where q is a positive constant (which will be taken as the q in (4.3)).

4.3 Continuum Analysis

Consider the fourth-order problem,

$$B\nabla\cdot\nabla\cdot\left(\nabla\nabla u + q^2\boldsymbol{T}u\right) + Bq^2\nabla\nabla u:\boldsymbol{T} + (Bq^4\boldsymbol{T}:\boldsymbol{T}+m)u = f, \tag{4.10}$$

where $\boldsymbol{T}$ is a $d\times d$ tensor with $|\boldsymbol{T}|^2 = \boldsymbol{T}:\boldsymbol{T}\le\mu_1$ for some constant μ_1 almost everywhere in $\bar{\Omega}$, m and q are $\mathcal{O}(1)$ positive constants, and $0<B\le 1$. Let $u\in H^2(\Omega)$ satisfy (4.10); given $\phi\in H^2(\Omega)$, a test function, integration by parts gives

$$\int_\Omega f\phi = a(u,\phi) + B\int_{\partial\Omega}\phi\nabla\cdot(\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} - B\int_{\partial\Omega}\nabla\phi\cdot(\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} \tag{4.11}$$

where the bilinear form, a, is given by

$$\begin{aligned} a(u,\phi) &= B\int_\Omega \nabla\nabla u:\nabla\nabla\phi + Bq^2\int_\Omega \boldsymbol{T}u:\nabla\nabla\phi + Bq^2\int_\Omega \nabla\nabla u:\boldsymbol{T}\phi \\ &+ \int_\Omega (Bq^4\boldsymbol{T}:\boldsymbol{T}+m)u\phi. \end{aligned} \tag{4.12}$$

From (4.11), we identify that two boundary conditions are required on any segment of $\partial\Omega$, and that certain boundary conditions on u arise naturally from the variational formulation. Consequently, we write $\partial\Omega = \Gamma_0\cup\Gamma_1\cup\Gamma_2\cup\Gamma_3$ with $\Gamma_i\cap\Gamma_j=\emptyset$ for $i\ne j$, and specify

$$u = 0, \quad (\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} = \vec{0}, \quad \text{on } \Gamma_0, \tag{4.13}$$

$$u = 0, \quad \nabla u = \vec{0}, \quad \text{on } \Gamma_1, \tag{4.14}$$

$$\nabla\cdot(\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} = 0, \quad (\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} = \vec{0}, \quad \text{on } \Gamma_2, \tag{4.15}$$

$$\nabla\cdot(\nabla\nabla u + q^2\boldsymbol{T}u)\cdot\vec{n} = 0, \quad \nabla u = \vec{0}, \quad \text{on } \Gamma_3. \tag{4.16}$$

As typical, we consider homogeneous boundary conditions here, so that the boundary integrals in (4.11) vanish, but the results hold true for inhomogeneous boundary conditions if the traces of these quantities are smooth enough on $\partial\Omega$. Consequently, we define

$$\mathscr{V} = \left\{ u \in H^2(\Omega) : u = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \text{ and } \nabla u = \vec{0} \text{ on } \Gamma_1 \cup \Gamma_3 \right\}.$$

Because of the dependence on q in the bilinear form, we analyze the problem in a q-dependent norm on $\mathscr{V}$, given by

$$\|u\|_{2,q}^2 = q^{-4}\|\nabla\nabla u\|_0^2 + q^{-4}\|\nabla u\|_0^2 + \|u\|_0^2. \tag{4.17}$$

Assumptions. In all results that follow, we assume $q \geq 1$ and that there exists an $\mathcal{O}(1)$ constant, s, such that $sq^{-4} \leq B \leq 1$, as this is the case of interest.

Theorem 13. *Given $f \in L^2(\Omega)$, the variational problem to find $u \in \mathscr{V}$ such that*

$$a(u,\phi) = \int_\Omega f\phi \text{ for all } \phi \in \mathscr{V} \tag{4.18}$$

is well-posed.

Proof. By the Lax–Milgram Theorem, this variational problem has a unique solution if the bilinear form a is coercive and continuous on $\mathscr{V}$ in the $\|\cdot\|_{2,q}$ norm, and the associated linear form is continuous. The assumption that $f \in L^2(\Omega)$ is sufficient to guarantee that the linear form is continuous.

To prove continuity of a, we have

$$\begin{aligned}
a(u,\phi) &= B\int_\Omega \nabla\nabla u : \nabla\nabla\phi + Bq^2\int_\Omega \nabla\nabla u : \boldsymbol{T}\phi + Bq^2\int_\Omega \nabla\nabla\phi : \boldsymbol{T}u \\
&+ \int_\Omega (Bq^4\boldsymbol{T} : \boldsymbol{T} + m)u\phi \\
&\leq B\|\nabla\nabla u\|_0\|\nabla\nabla\phi\|_0 + Bq^2\left(\|\nabla\nabla u\|_0\|\boldsymbol{T}\phi\|_0\right) + Bq^2\left(\|\nabla\nabla\phi\|_0\|\boldsymbol{T}u\|_0\right) \\
&+ (Bq^4\mu_1 + m)\|u\|_0\|\phi\|_0 \\
&\leq Bq^4\|u\|_{2,q}\|\phi\|_{2,q} + 2B\sqrt{\mu_1}q^4\|u\|_{2,q}\|\phi\|_{2,q} + (Bq^4\mu_1 + m)\|u\|_{2,q}\|\phi\|_{2,q},
\end{aligned}$$

where we have used the fact that $\|\nabla\nabla u\|_0 \leq q^2\|u\|_{2,q}$. This gives a q-dependent continuity bound, with $a(u,\phi) \leq CBq^4\|u\|_{2,q}\|\phi\|_{2,q}$, where $C = 1 + 2\sqrt{\mu_1} + \mu_1 + m/s$.

To prove coercivity, we first observe that, for any $0 < C_1 < 1$,

$$\begin{aligned} a(u,u) &= B\|\nabla\nabla u\|_0^2 + 2Bq^2\int_\Omega \nabla\nabla u : \boldsymbol{T}u + \int_\Omega (Bq^4\boldsymbol{T}:\boldsymbol{T} + m)u^2 \\ &\geq B(1-C_1)\|\nabla\nabla u\|_0^2 + Bq^4\left(1-\frac{1}{C_1}\right)\|\boldsymbol{T}u\|_0^2 + m\|u\|_0^2 \\ &\geq B\left(1-C_1\right)\|\nabla\nabla u\|_0^2 + \left(m + Bq^4\mu_1\left(1-\frac{1}{C_1}\right)\right)\|u\|_0^2. \end{aligned} \tag{4.19}$$

Note that since $0 < C_1 < 1$, $1 - \frac{1}{C_1} < 0$, so we get a lower bound for the expression on the second line by using the upper bound $\|\boldsymbol{T}u\|_0^2 \leq \mu_1\|u\|_0^2$. Let $C_1 = \frac{B\mu_1 q^4}{B\mu_1 q^4 + C_2}$ for any constant $0 < C_2 < m$. Then, $1 - C_1 = \frac{C_2}{Bq^4\mu_1 + C_2}$, and $1 - \frac{1}{C_1} = -\frac{C_2}{Bq^4\mu_1}$, giving

$$a(u,v) \geq B\left(\frac{C_2}{B\mu_1 q^4 + C_2}\right)\|\nabla\nabla u\|_0^2 + (m - C_2)\|u\|_0^2. \tag{4.20}$$

Now, since $sq^{-4} \leq B \leq 1$, $B\left(\frac{C_2}{B\mu_1 q^4 + C_2}\right) = \left(\frac{C_2}{\mu_1 q^4 + \frac{C_2}{B}}\right)$ is an $\mathcal{O}(q^{-4})$ constant. That is, there exists a constant C_3 such that

$$a(u,u) \geq C_3\left(q^{-4}\|\nabla\nabla u\|_0^2 + \|u\|_0^2\right). \tag{4.21}$$

We then use Lemma 6, with $j = 2$, and $\Phi(u) = \|u\|_0$. That is, there exists a constant C_4 such that $\|\nabla u\|_0^2 \leq C_4\left(\|\nabla\nabla u\|_0^2 + \|u\|_0^2\right)$, giving

$$\begin{aligned} a(u,u) &\geq C_3 q^{-4}\left(\|\nabla\nabla u\|_0^2 + q^4\|u\|_0^2\right) \\ &\geq \frac{C_3 q^{-4}}{2}\left(\|\nabla\nabla u\|_0^2 + \|u\|_0^2\right) + \frac{C_3}{2}\left(q^{-4}\|\nabla\nabla u\|_0^2 + \|u\|_0^2\right) \\ &\geq \left(\frac{C_3}{2C_4}q^{-4}\|\nabla u\|_0^2 + \frac{C_3}{2}\left(q^{-4}\|\nabla\nabla u\|_0^2 + \|u\|_0^2\right)\right), \end{aligned}$$

As a result, $a(u,u) \geq \left(\min\{\frac{C_3}{2C_4}, \frac{C_3}{2}\}\right)\|u\|_{2,q}^2$, giving an $\mathcal{O}(1)$ coercivity constant.

□

The mixed finite-element formulation presented below relies on a reformulation of (4.10) to a lower-order system. We introduce $\vec{v} = \nabla u$ and $\vec{\alpha} = B\nabla\cdot(\nabla v + q^2\boldsymbol{T}u)$.

Then, Equation (4.10) is equivalent to the system of equations

$$\begin{aligned} \nabla \cdot \vec{\alpha} + Bq^2 \nabla \vec{v} : \boldsymbol{T} + (Bq^4 \boldsymbol{T} : \boldsymbol{T} + m)u &= f, && (4.22) \\ \vec{\alpha} - B\Delta \vec{v} - Bq^2 \nabla \cdot (\boldsymbol{T}u) &= 0, && (4.23) \\ (\vec{v} - \nabla u) &= 0. && (4.24) \end{aligned}$$

To convert this system to weak form, we multiply (4.22), (4.23), and (4.24) by the test functions $\phi, \vec{\psi}$, and $\vec{\beta}$, respectively, and integrate by parts. This yields the weak form of finding $(u, \vec{v}, \vec{\alpha}) \in L^2(\Omega) \times V \times H_{\Gamma_2 \cup \Gamma_3}(\text{div}; \Omega)$ such that

$$\begin{aligned} \int_\Omega \nabla \cdot \vec{\alpha}\phi + Bq^2 \nabla \vec{v} : \boldsymbol{T}\phi + (Bq^4 \boldsymbol{T} : \boldsymbol{T} + m)u\phi &= \int_\Omega f\phi, \ \forall \phi \in L^2(\Omega), && (4.25) \\ \int_\Omega \vec{\alpha} \cdot \vec{\psi} + \left(B\nabla \vec{v} + Bq^2 \boldsymbol{T}u\right) : \nabla \vec{\psi} &= 0, \ \forall \vec{\psi} \in V, && (4.26) \\ \int_\Omega \vec{\beta} \cdot \vec{v} + \int_\Omega u \nabla \cdot \vec{\beta} &= 0, \ \forall \vec{\beta} \in H_{\Gamma_2 \cup \Gamma_3}(\text{div}; \Omega), && (4.27) \end{aligned}$$

where $V = \left\{ \vec{v} \in [H^1_{\Gamma_1 \cup \Gamma_3}(\Omega)]^d, \ \vec{v} \times \vec{n} = 0 \text{ on } \Gamma_0 \right\}$. Note that since $u = 0$ on Γ_0, we are free to require the boundary condition $\vec{v} \times \vec{n} = 0$ on Γ_0, which will be needed below. Equivalently, the system (4.25)-(4.27) can be written as a saddle-point system, to find $(u, \vec{v}, \vec{\alpha}) \in L^2(\Omega) \times V \times H_{\Gamma_2 \cup \Gamma_3}(\text{div}; \Omega)$ such that

$$\begin{aligned} \mathcal{A}\left((u, \vec{v}), (\phi, \vec{\psi})\right) + b\left(\vec{\alpha}, (\phi, \vec{\psi})\right) &= F(\phi), \quad \forall (\phi, \vec{\psi}) \in L^2(\Omega) \times V && (4.28) \\ b\left(\vec{\beta}, (u, \vec{v})\right) &= 0, \quad \forall \vec{\beta} \in H_{\Gamma_2 \cup \Gamma_3}(\text{div}; \Omega) && (4.29) \end{aligned}$$

where

$$\begin{aligned} \mathcal{A}\left((u, \vec{v}), (\phi, \vec{\psi})\right) &= B \int_\Omega \nabla \vec{v} : \nabla \vec{\psi} + Bq^2 \int_\Omega \nabla \vec{v} : \boldsymbol{T}\phi + Bq^2 \int_\Omega \nabla \vec{\psi} : \boldsymbol{T}u \\ &+ \int_\Omega (Bq^4 \boldsymbol{T} : \boldsymbol{T} + m)u\phi, && (4.30) \\ b\left(\vec{\alpha}, (u, \vec{v})\right) &= \int_\Omega \vec{\alpha} \cdot \vec{v} + \int_\Omega u \nabla \cdot \vec{\alpha}, && (4.31) \\ F(\phi) &= \int_\Omega f\phi. && (4.32) \end{aligned}$$

Here, we see that while α is defined in terms of u above, it serves as a Lagrange multiplier in the saddle-point form, weakly enforcing that $\vec{v} = \nabla u$.

In the next theorem, we prove well-posedness of (4.28)-(4.29) in the two-dimensional case. Here, we make use of a similarly weighted product norm for $(u, \vec{v})$, defined as

$$\|(u, \vec{v})\|_{0,q,1}^2 = \|u\|_0^2 + q^{-4}\|\vec{v}\|_1^2. \tag{4.33}$$

Theorem 14. *Assume that $\partial\Omega \neq \Gamma_3$. Given $f \in L^2(\Omega)$, (4.28)-(4.29) is well-posed using the product norm defined in (4.33) for $(u, \vec{v})$ and the H_{Div} norm induced by (4.9) for the Helmholtz decomposition, $\vec{\alpha} = \nabla\phi + \nabla \times p$, given in Lemma 8.*

Proof. By Brezzi's theory, proving well-posedness requires establishing the coercivity and continuity of $\mathcal{A}$, an inf-sup condition on b, and the continuity of b. As above, $F(\phi)$ is clearly continuous.

We first consider the two continuity bounds. The continuity bound for $\mathcal{A}$ gives an $\mathcal{O}(Bq^4)$ continuity constant, established similarly to the proof in Theorem 13. The Helmholtz decomposition in Lemma 8 and integration by parts are needed to show continuity of b. Let $\vec{\alpha} \in H_{\Gamma_2\cup\Gamma_3}(\text{div}; \Omega)$, writing $\vec{\alpha} = \nabla\phi + \nabla \times p$, where $\phi \in H^1_{\Gamma_0\cup\Gamma_1}(\Omega)$ and $p \in H^1_{\Gamma_2\cup\Gamma_3}(\Omega)$. Then,

$$\begin{aligned}
b\left(\vec{\alpha}, (u, \vec{v})\right) &= \int_\Omega \vec{\alpha}\cdot\vec{v} + \int_\Omega u\nabla\cdot\vec{\alpha} = \int_\Omega (\nabla\phi + \nabla\times p)\cdot\vec{v} + \int_\Omega u\nabla\cdot\vec{\alpha} \\
&= \int_\Omega \nabla\phi\cdot\vec{v} + \int_\Omega p\nabla\times\vec{v} + \int_\Omega u\nabla\cdot\vec{\alpha} \\
&\leq q^4\left(q^{-2}\|\nabla\phi\|_0\right)\left(q^{-2}\|\vec{v}\|_0\right) + q^4\left(q^{-2}\|p\|_0\right)\left(q^{-2}\|\nabla\times\vec{v}\|_0\right) \\
&+ q^2\|u\|_0\left(q^{-2}\|\nabla\cdot\vec{\alpha}\|_0\right) \\
&\leq q^4\left(1 + \sqrt{2} + q^{-2}\right)\|(u, \vec{v})\|_{0,q,1}\|\vec{\alpha}\|_{\text{Div}},
\end{aligned}$$

where we have used the fact that $\|\nabla\times\vec{v}\|_0 \leq \sqrt{2}\|\nabla\vec{v}\|_0$. We note that the "extra" boundary condition imposed on $\vec{v} \in V$ is needed here to ensure the boundary integral from integration-by-parts is identically zero.

We now show that the bilinear form, $\mathcal{A}$, is coercive on

$$\Lambda = \{(u, v) \in L^2(\Omega) \times H^1_{\Gamma_1\cup\Gamma_3}(\Omega) \ : \ b\left(\vec{\alpha}, (u, \vec{v})\right) = 0, \ \forall\vec{\alpha} \in H_{\Gamma_2\cup\Gamma_3}(\text{div}; \Omega)\}.$$

Similarly to the proof of Inequality (4.21), there exists $C_1 > 0$ such that

$$\mathcal{A}\big((u,\vec{v}),(u,\vec{v})\big) \geq C_1 \left(q^{-4}\|\nabla\vec{v}\|_0^2 + \|u\|_0^2\right) \quad \forall\ (u,\vec{v}) \in \Lambda. \tag{4.34}$$

If $(u,\vec{v}) \in \Lambda$, then $b(\beta_i,(u,\vec{v})) = 0$ for the particular choices $\beta_1 = [S,0]^\top$ and $\beta_2 = [0,S]^\top$, where S is the function defined in Lemma 7. In other words, $\int_\Omega \vec{\beta}_i \cdot \vec{v} + \int_\Omega u\nabla\cdot\vec{\beta}_i = 0$ for $i = 1,2$, which gives

$$\left|\int_\Omega S v_1\right| = \left|\int_\Omega S_x u\right| \leq \|S_x\|_0\|u\|_0 \leq C_2\|u\|_0, \quad \text{and}$$

$$\left|\int_\Omega S v_2\right| = \left|\int_\Omega S_y u\right| \leq \|S_y\|_0\|u\|_0 \leq C_2\|u\|_0.$$

As S has a positive mean, we can apply the Poincaré inequality in the form given in Lemma 6, which leads to the fact

$$\|\vec{v}\|_0^2 \leq C_3\left(\|u\|_0^2 + \|\nabla\vec{v}\|_0^2\right). \tag{4.35}$$

Combining this with Inequality (4.34) then gives

$$\mathcal{A}\left((u,\vec{v}),(u,\vec{v})\right) \geq C_4\|(u,\vec{v})\|_{0,q,1}.$$

That is, coercivity of $\mathcal{A}$ holds with an $\mathcal{O}(1)$ constant.

Next, we prove the required inf-sup condition, of the form

$$I = \sup_{(u,\vec{v})\in L^2\times V} \frac{\int_\Omega \vec{\alpha}\cdot\vec{v} + \int_\Omega u\nabla\cdot\vec{\alpha}}{\sqrt{\|u\|_0^2 + q^{-4}\|\vec{v}\|_1^2}} \geq Cq^2\|\vec{\alpha}\|_{\mathrm{Div}}, \quad \forall\vec{\alpha} \in H_{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega).$$

In our mixed formulation, the boundary condition on u is weakly enforced. Therefore, given $\vec{\alpha} = \nabla\eta + \nabla\times p$, where $p \in H^1_{\Gamma_2\cup\Gamma_3}(\Omega)$ and $\eta \in H^1_{\Gamma_0\cup\Gamma_1}$, we may choose $u = C_1\left(\nabla\cdot\vec{\alpha} - \eta\right)$, for a positive constant, C_1, to be specified later. This choice of u might not be zero on $\Gamma_0\cup\Gamma_1$. Note that

$$\|u\|_0 = C_1\|\nabla\cdot\vec{\alpha} - \eta\|_0 \leq C_1\left(\|\nabla\cdot\vec{\alpha}\|_0 + \|\eta\|_0\right), \text{ and } \int_\Omega u\nabla\cdot\vec{\alpha} = C_1\left(\|\nabla\cdot\vec{\alpha}\|_0^2 + \|\nabla\eta\|_0^2\right).$$

If Γ_3 is a proper subset of $\partial\Omega$, then the inf-sup condition of [30, Lemma 2.7] establishes

that for all $p \in H^1_{\Gamma_2 \cup \Gamma_3}(\Omega)$, there exists $\vec{\psi} \in [H^1_{\Gamma_1 \cup \Gamma_2}(\Omega)]^2$ such that $\int_\Omega p \nabla \cdot \vec{\psi} \geq \|p\|_0^2$, and $\|\vec{\psi}\|_1^2 \leq C_2 \|\vec{p}\|_0^2$. Note that

- If $\partial\Omega = \Gamma_0 \cup \Gamma_1$, then p is a zero-mean function, and such a $\vec{\psi}$ exists by the standard inf-sup condition for the Stokes problem with Dirichlet boundary conditions.
- Otherwise (so long as $\partial\Omega \neq \Gamma_3$), such a $\vec{\psi} \in V$ exists with $\vec{\psi} = \vec{0}$ on $\Gamma_0 \cup \Gamma_1 \cup \Gamma_3$ and $\vec{\psi} \times \vec{n} = 0$ on Γ_2, following [30, Lemma 2.7].

To establish the inf-sup condition needed here, we then choose $\vec{v} = [\psi_2, -\psi_1]^T$ which, by construction, belongs to V, giving $\nabla \cdot \vec{\psi} = \nabla \times \vec{v}$ and $\|\vec{\psi}\|_1^2 = \|\vec{v}\|_1^2$. With this,

$$
\begin{aligned}
I \geq & \sup_{(u,\vec{v}) \in L^2 \times V} \frac{\int_\Omega \nabla\eta \cdot \vec{v} + \int_\Omega p \nabla \times \vec{v} + \int_\Omega u \nabla \cdot \vec{\alpha}}{\sqrt{\|u\|_0^2 + q^{-4}\|\vec{v}\|_1^2}} \\
\geq & \sup_{(u,\vec{v}) \in L^2 \times V} \frac{\int_\Omega \nabla\eta \cdot \vec{v} + \int_\Omega p \nabla \times \vec{v} + \int_\Omega u \nabla \cdot \vec{\alpha}}{\sqrt{\|u\|_0^2 + \|\vec{v}\|_1^2}} \\
\geq & \frac{\|p\|_0^2 + C_1\|\nabla \cdot \vec{\alpha}\|_0^2 + C_1\|\nabla\eta\|_0^2 - \frac{C_3}{2}\|\nabla\eta\|_0^2 - \frac{1}{2C_3}\|\vec{v}\|_0^2}{\sqrt{C_2\|p\|_0^2 + 2C_1^2\|\eta\|_0^2 + 2C_1^2\|\nabla \cdot \vec{\alpha}\|_0^2}} \\
& \geq \frac{\left(1 - \frac{C_2}{2C_3}\right)\|p\|_0^2 + C_1\|\nabla \cdot \vec{\alpha}\|_0^2 + \left(C_1 - \frac{C_3}{2}\right)\|\nabla\eta\|_0^2}{\sqrt{C_2\|p\|_0^2 + 2C_1^2\|\eta\|_0^2 + 2C_1^2\|\nabla \cdot \vec{\alpha}\|_0^2}},
\end{aligned}
$$

where we use the facts that $\int_\Omega p \nabla \times \vec{v} \geq \|p\|_0^2$, $\|\vec{v}\|_1^2 \leq C_2\|p\|_0^2$, and $\left|\int_\Omega \nabla\eta \cdot \vec{v}\right| \leq \frac{C_3}{2}\|\nabla\eta\|_0^2 + \frac{1}{2C_3}\|\vec{v}\|_0^2$. Choose $C_1 > \frac{C_3}{2} > \frac{C_2}{4}$. Note that $\|\eta\|_0^2 \leq C_4\|\nabla\eta\|_0^2$ for some positive C_4 by the Poincaré inequality. Thus,

$$
\begin{aligned}
I \quad \geq \quad & \frac{\left(1 - \frac{C_2}{2C_3}\right)\|p\|_0^2 + C_1\|\nabla \cdot \vec{\alpha}\|_0^2 + \left(C_1 - \frac{C_3}{2}\right)\|\nabla\eta\|_0^2}{\sqrt{C_2\|p\|_0^2 + 2C_1^2C_4\|\nabla\eta\|_0^2 + 2C_1^2\|\nabla \cdot \vec{\alpha}\|_0^2}} \\
\geq \quad & C\left(\|p\|_0 + \|\nabla\eta\|_0 + \|\nabla \cdot \vec{\alpha}\|_0\right) \geq Cq^2\|\vec{\alpha}\|_{\mathrm{Div}}.
\end{aligned}
$$

As a result, the inf-sup condition holds true and the three-field formulation is well-posed. □

Remark 4.3.1. The dependence on q of the continuity, coercivity, and inf-sup constants in the proofs above can lead to pessimistic error bounds for the finite-element methods developed below. While it is tempting to try and prove convergence using

other weighted H^2 norms or $L^2 \times H^1$ product norms, we are unaware of a simple weighting of the terms in such a norm that leads to $\mathcal{O}(1)$ continuity and coercivity constants in the weighted norms. We note that a common use case of these results will be when Bq^4 is an $\mathcal{O}(1)$ constant, in which case it is only the inf-sup constant and continuity bounds for b that are suboptimal.

Corollary 9. *Suppose $\Omega \subset \mathbb{R}^2$ has $\partial\Omega = \left\{\cup_{i=1}^{N_1}\Gamma^i\right\} \cup \left\{\cup_{i=1}^{N_2}\bar{\Gamma}^i\right\}$, for $\Gamma^i = (x, a_i x + b_i)$, and $\bar{\Gamma}^i = (c_i y + d_i, y)$ where N_1 and N_2 are positive integers. Let u and $(\bar{u}, \vec{v}, \vec{\alpha})$ be the solutions of Problems (4.18) and (4.28)-(4.29), respectively. Further, assume u and $\bar{u}$ are in $H^t(\Omega)$, $t \geq 4$, and $\boldsymbol{T} \in \boldsymbol{C}^{t-2}(\Omega)$, where $\boldsymbol{C}^m(\Omega)$ is the space of 2×2 tensors with each component in $C^m(\Omega)$. Then u and $(\bar{u}, \vec{v}, \vec{\alpha})$ are equivalent in the sense that $u = \bar{u}$, $\vec{v} = \nabla u$, and $\vec{\alpha} = \nabla \cdot (\nabla\nabla u + q^2 \boldsymbol{T} u)$.*

Proof. First, let $u \in H^t(\Omega)$, $t \geq 4$, be the solution of Problem (4.18). By direct calculation, $(u, \vec{v}, \vec{\alpha})$ is a solution to Problem (4.28)-(4.29), which is unique by Theorem 14. Note that $u \in H^t(\Omega)$ is sufficient to guarantee that $\vec{v} \in \left(H^1(\Omega)\right)^2$ and $\vec{\alpha} \in H(\mathrm{div}; \Omega)$.

Conversely, let $(\bar{u}, \vec{v}, \vec{\alpha})$ be the solution of (4.28)-(4.29) with $\bar{u} \in H^t(\Omega)$. Let $D = \Pi_{i=1}^{N_1} \left(y - a_i x - b_i\right)^2 \Pi_{i=1}^{N_2} \left(x - c_i y - d_i\right)^2$; $D \in C^\infty(\Omega) \cap H_0^1(\Omega)$ is positive in the interior of Ω. Choosing $(\phi, \vec{\psi}) = (0, DB\nabla \cdot (\nabla\nabla\bar{u} + q^2 \boldsymbol{T}\bar{u}))$ and $\vec{\beta} = D\left(\vec{v} - \nabla\bar{u}\right)$ in (4.28)–(4.29) and integrating by parts imply that $\vec{v} = \nabla\bar{u}$ and $\vec{\alpha} = B\nabla \cdot (\nabla\nabla\bar{u} + q^2 \boldsymbol{T}\bar{u})$:

$$\int_\Omega \vec{\beta} \cdot \left(\vec{v} - \nabla\bar{u}\right) = \int_\Omega D\left(\vec{v} - \nabla\bar{u}\right) \cdot \left(\vec{v} - \nabla\bar{u}\right) = 0.$$

Note that D is sufficiently smooth so that

$$\vec{\beta} = D\left(\vec{v} - \nabla\bar{u}\right) \in H_0(\mathrm{div}; \Omega) \subset H_{\Gamma_2 \cup \Gamma_3}(\mathrm{div}; \Omega).$$

As $D\left(\vec{v} - \nabla u\right) \cdot \left(\vec{v} - \nabla u\right)$ is non-negative over Ω, this implies that $\vec{v} = \nabla\bar{u}$. Similarly, one can prove that the value of $\vec{\alpha} = \nabla \cdot (\nabla\nabla\bar{u} + q^2 \boldsymbol{T}\bar{u})$. As above, the regularity of $\bar{u}$ is necessary to ensure that $\vec{v}$ and $\vec{\alpha}$ (as well as $\vec{\psi}$ and $\vec{\beta}$ defined above) have the regularity to satisfy these equations. With this value for $\vec{\alpha}$, taking $\vec{\psi} = \vec{0}$ and $\phi \in \mathscr{V} \subset L^2(\Omega)$ in (4.28) leads to the fact that $\bar{u}$ is a solution of Problem (4.18). Thus, $\bar{u} = u$ by the uniqueness of the solution of Problem (4.18). $\square$

4.4 Discrete Analysis

We now consider three different discretizations of the smectic density equation. First, in Section 4.4.1, we consider an H^2-conforming discretization based on Argyris elements for $\Omega \subset \mathbb{R}^2$. The method offers optimal convergence bounds, but its analogue for $\Omega \subset \mathbb{R}^3$ is very difficult to implement. Thus, in Section 4.4.2, we consider a C^0-interior-penalty method, which allows for the use of continuous Lagrange elements of any order $k \geq 2$, in both two and three dimensions. Finally, in Section 4.4.3, we develop a mixed finite-element formulation similar to that proposed in [66], that also offers some advantages, especially in the construction of preconditioners.

4.4.1 Conforming Methods

We first consider the case of $\Omega \subset \mathbb{R}^2$ with full Neumann boundary conditions, where $\partial\Omega = \Gamma_2$. While several choices of conforming elements are possible, we focus on Argyris elements, $\mathrm{ARG}_5(\Omega, \tau_h)$, which arise from choosing a basis for the 21 degrees of freedom for a fifth-order polynomial space, $CG_5(\Omega, T)$, on each triangle, T, in such a way as to ensure that the resulting space is H^2-conforming [92]. The weak form is to find $u_h \in \mathrm{ARG}_5(\Omega, \tau_h)$ such that [52]

$$a(u_h, \phi_h) = F(\phi_h), \quad \forall \phi_h \in \mathrm{ARG}_5(\Omega, \tau_h), \tag{4.36}$$

where a is defined in (4.12) and $F(\phi_h)$ is defined in (4.32).

Corollary 10. *Let $f \in L^2(\Omega)$, and let $\{\tau_h\}$ be a family of quasiuniform meshes of Ω. Problem* (4.36) *is well-posed for $\partial\Omega = \Gamma_2$. Moreover, if $hq \leq 1$ and $u \in H^t(\Omega)$ for $3 \leq t \leq 6$ is the solution of Problem* (4.12)*, then*

$$\|u - u_h\|_{2,q} \leq CBq^2 h^{t-2} |u|_t. \tag{4.37}$$

Proof. For $\partial\Omega = \Gamma_2$, the bilinear form $a(u_h, \phi_h)$ is symmetric, continuous and coercive and the linear form F is continuous, as shown above. Since $u_h, \phi_h \in \mathrm{ARG}_5(\Omega, \tau_h) \subset H^2(\Omega)$, this is a conforming discretization and is well-posed following Theorem 13 and the Lax–Milgram theorem. Finally, Céa's lemma and standard bounds on the Argyris

interpolation operator [35] lead to the estimate in (4.37), as

$$\begin{aligned}\|u-u_h\|_{2,q} &\leq CBq^4 \inf_{v_h \in \mathrm{ARG}_5(\Omega,\tau_h)} \left(q^{-2}\|\nabla\nabla(u-v_h)\|_0 + q^{-2}\|\nabla(u-v_h)\|_0 + \|u-v_h\|_0\right) \\ &\leq CBq^4 \left(q^{-2}h^{t-2} + q^{-2}h^{t-1} + h^t\right)|u|_t \leq CBq^2 h^{t-2}|u|_t.\end{aligned}$$

□

Remark 4.4.1. We are naturally interested in how the error estimate above depends on q. From the coercivity and continuity constants of Theorem 13, which scale as $\mathcal{O}(1)$ and $\mathcal{O}(Bq^4)$, respectively, we see that the quasioptimality constant scales like $\mathcal{O}(Bq^4)$. When $Bq^4 = \mathcal{O}(1)$, as can be the case (see, e.g., [112]), this gives an error bound that $\|u-u_h\|_{2,q} \leq (q^{-2}h^{t-2})|u|_t$, which is optimal in h. For larger values of B, we retain optimality in h, but see some degradation in q, as might be expected. Moreover, for the case of expected solutions to (4.1) that behave like $e^{iq\vec{\nu}\cdot\vec{x}}$ (showing similar behaviour to the observed solutions of the generalized models in [112, 138]), we have $|u|_t \sim \mathcal{O}(q^t)$. Considering the case of a strong solution with $u \in H^6(\Omega)$, this gives an error estimate that scales like $\mathcal{O}(Bq^8)$, but with an L^2 error estimate of h^4. Again, the value of B strongly influences the impact of this scaling: when $Bq^4 = \mathcal{O}(1)$, then this necessitates choosing a mesh, τ_h, such that $hq < 1$, which is not an unreasonable requirement when q is, itself, an $\mathcal{O}(1)$ constant. If, however, $B = \mathcal{O}(1)$, the requirement on τ_h is stricter, needing $h^4q^8 < 1$ in order to guarantee convergence in the large q limit. While we are not interested in prohibitively large values of q (as in [112, 138], we consider $q \sim 40$), this recalls standard results in the literature on numerical approximation of solutions to the Helmholtz equation and the *pollution effect* that leads to similar restrictions [26, 84].

Remark 4.4.2. The above result naturally extends to domains $\Omega \subset \mathbb{R}^3$ with three-dimensional $H^2(\Omega)$ conforming elements [141].

Strongly enforcing essential boundary conditions using Argyris elements is well-known to be difficult [90], although extensions of Corollary 10 would hold if we could do so. Instead, if $\Gamma_0 \cup \Gamma_1 \cup \Gamma_3 \neq \emptyset$, we enforce the essential boundary conditions weakly using Nitsche-type penalty methods. Then, the weak form is to find $u_h \in \mathrm{ARG}_5(\Omega, \tau_h)$ such that

$$A_h(u_h, \phi_h) = F(\phi_h), \quad \forall \phi_h \in \mathrm{ARG}_5(\Omega, \tau_h), \tag{4.38}$$

where

$$\begin{aligned}A_h(u_h,\phi_h) =& a(u_h,\phi_h) + B\int_{\Gamma_0\cup\Gamma_1}\phi_h\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\\ &-B\int_{\Gamma_0\cup\Gamma_1}u_h\nabla\cdot(\nabla\nabla\phi_h + q^2\boldsymbol{T}\phi_h)\cdot\vec{n} + \frac{1}{qh^3}\int_{\Gamma_0\cup\Gamma_1}u_h\phi_h\\ &+\frac{1}{q^3h}\int_{\Gamma_1\cup\Gamma_3}\nabla u_h\cdot\nabla\phi_h - B\int_{\Gamma_1\cup\Gamma_3}\nabla\phi_h\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\\ &+B\int_{\Gamma_1\cup\Gamma_3}\nabla u_h\cdot(\nabla\nabla\phi_h + q^2\boldsymbol{T}\phi_h)\cdot\vec{n}\end{aligned}$$

We prove coercivity and continuity of the bilinear form, A_h, in the strengthened $H^2(\Omega)$ norm, $|||\cdot|||_{2,q,h}$, defined as

$$\begin{aligned}|||u_h|||^2_{2,q,h} &= \|u_h\|^2_{2,q} + \frac{1}{qh^3}\|u_h\|^2_{0,\Gamma_0\cup\Gamma_1} + \frac{h^3}{q^7}\|\nabla\cdot\left(\nabla\nabla u_h + q^2\boldsymbol{T}u_h\right)\cdot\vec{n}\|^2_{0,\Gamma_0\cup\Gamma_1}\\ &\quad+\frac{1}{q^3h}\|\nabla u_h\|^2_{0,\Gamma_1\cup\Gamma_3} + \frac{h}{q^5}\|\left(\nabla\nabla u_h + q^2\boldsymbol{T}u_h\right)\cdot\vec{n}\|^2_{0,\Gamma_1\cup\Gamma_3}. \qquad (4.39)\end{aligned}$$

As shown below, the weights in the norm $|||\cdot|||_{2,q,h}$ allow us to prove optimal-in-q error estimates for solutions in the space $H^t(\Omega),\ t\geq 4$. Note that the choice of the Nitsche formulation in (4.38) results in non-symmetric linear systems to be solved; while we do not focus on effective linear solvers here, this nonsymmetry may be seen as a downside of this approach. However, we note that using a symmetric Nitsche formulation led to suboptimal error bounds in the analogous results to those that follow.

Theorem 15. *Let $f\in L^2(\Omega)$, and let $\{\tau_h\}$ be a family of quasiuniform meshes of Ω. Let $\boldsymbol{T}$ be given s.t. $|\boldsymbol{T}|^2 = \boldsymbol{T}:\boldsymbol{T}\leq\mu_1$ and $|\nabla\boldsymbol{T}|^2 = (\nabla\boldsymbol{T}):(\nabla\boldsymbol{T})\leq\mu_2$ pointwise on $\bar{\Omega}$. Then, there exist constants C_1 and C_2 such that for any $u_h,\phi_h\in ARG_5(\Omega,\tau_h)$,*

$$\begin{aligned}|A_h(u_h,\phi_h)| &\leq C_1Bq^4|||u_h|||_{2,q,h}|||\phi_h|||_{2,q,h},\\ A_h(u_h,u_h) &\geq C_2|||u_h|||^2_{2,q,h}.\end{aligned}$$

Moreover, Problem (4.38) *is well-posed over $ARG_5(\Omega,\tau_h)$.*

Proof. The continuity of $A_h(u_h,\phi_h)$ and $F(\phi_h)$ for $u_h,\phi_h\in \mathrm{ARG}_5(\Omega,\tau_h)$ follow directly from the Cauchy–Schwarz inequality applied termwise, making use of the leeway offered by the suboptimal continuity of $a(u,\phi)$ with respect to the $\|\cdot\|_{2,q}$ norm.

For example, we use the bound

$$
\begin{aligned}
B\int_{\Gamma_0\cup\Gamma_1} & \phi_h\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n} \\
&\leq B\left(\left(\frac{q^3}{h^3}\right)^{1/2}\|\phi_h\|_{0,\Gamma_0\cup\Gamma_1}\right)\left(\left(\frac{h^3}{q^3}\right)^{1/2}\|\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|_{0,\Gamma_0\cup\Gamma_1}\right) \\
&= Bq^4\left(\left(\frac{1}{qh^3}\right)^{1/2}\|\phi_h\|_{0,\Gamma_0\cup\Gamma_1}\right)\left(\left(\frac{h^3}{q^7}\right)^{1/2}\|\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|_{0,\Gamma_0\cup\Gamma_1}\right) \\
&\leq Bq^4|||\phi_h|||_{2,q,h}|||u_h|||_{2,q,h},
\end{aligned}
$$

with similar bounds for the other terms in $A_h(u_h,\phi_h)$ (recalling that $1\leq\frac{1}{s}Bq^4$ for $\mathcal{O}(1)$ constant, s). Therefore, we focus on the proof of coercivity.

Trace theorems [117, Section 2.1.3], standard inverse estimates [41, Theorem 4.5.11], and Theorem 12 can be used to prove the following inequalities that are needed to show coercivity of A_h. For $\phi_h\in\mathrm{ARG}_5(\Omega,\tau_h)$, we have

$$
\|\nabla\cdot\nabla\nabla\phi_h\cdot\vec{n}\|^2_{0,\Gamma_0\cup\Gamma_3}\leq\frac{C}{h}\|\nabla\cdot\nabla\nabla\phi_h\|^2_0, \tag{4.40}
$$

$$
\|\nabla\cdot\nabla\nabla\phi_h\|^2_0\leq C\|\nabla\nabla\nabla\phi_h\|^2_0\leq\frac{C}{h^2}\|\nabla\nabla\phi_h\|^2_0, \tag{4.41}
$$

$$
\begin{aligned}
\|\nabla\cdot(\nabla\nabla\phi_h + q^2\boldsymbol{T}\phi_h)\cdot\vec{n}\|^2_{0,\Gamma_0\cup\Gamma_1} &\leq C\left(\frac{1}{h^3}\|\nabla\nabla\phi_h\|^2_0 + q^4\left(\frac{\mu_2}{h}\|\phi_h\|^2_0+\frac{\mu_1}{h}\|\nabla\phi_h\|^2_0\right)\right) \\
&\leq C\left(\frac{1}{h^3}\|\nabla\nabla\phi_h\|^2_0 + q^4\left(\frac{\mu_2}{h}\|\phi_h\|^2_0+\frac{\mu_1}{h^3}\|\phi_h\|^2_0\right)\right) \\
&\leq \frac{C_3q^4}{h^3}\|\phi_h\|^2_{2,q}.
\end{aligned} \tag{4.42}
$$

In addition,

$$
\|(\nabla\nabla\phi_h+q^2\boldsymbol{T}\phi_h)\cdot\vec{n}\|^2_{0,\Gamma_1\cup\Gamma_3}\leq C\left(\frac{1}{h}\|\nabla\nabla\phi_h\|^2_0+\frac{q^4\mu_1}{h}\|\phi_h\|^2_0\right)\leq\frac{C_4q^4}{h}\|\phi_h\|^2_{2,q}, \tag{4.43}
$$

where the constants, C_3 and C_4, will be used below. From Theorem 13, we have that

$a(u_h, u_h) \geq C_5 \|u_h\|^2_{2,q}$, for $C_5 > 0$. Using this, we have

$$\begin{aligned}
A_h(u_h, u_h) &\geq C_5 \|u_h\|^2_{2,q} + \frac{1}{qh^3}\|u_h\|^2_{0,\Gamma_0\cup\Gamma_1} + \frac{1}{q^3 h}\|\nabla u_h\|^2_{\Gamma_1\cup\Gamma_3} \\
&\geq \frac{C_5}{3}\left(\|u_h\|^2_{2,q} + \frac{h^3}{C_3 q^4}\|\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|^2_{0,\Gamma_0\cup\Gamma_1}\right) \\
&+ \frac{C_5}{3}\left(\frac{h}{C_4 q^4}\|(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|^2_{0,\Gamma_1\cup\Gamma_3}\right) + \frac{1}{qh^3}\|u_h\|^2_{0,\Gamma_0\cup\Gamma_1} \\
&+ \frac{1}{q^3 h}\|\nabla u_h\|^2_{\Gamma_1\cup\Gamma_3} \\
&\geq \frac{C_5}{3}\|u_h\|^2_{2,q} + \frac{C_5 h^3}{3C_3 q^7}\|\nabla\cdot(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|^2_{0,\Gamma_0\cup\Gamma_1} \\
&+ \frac{C_5 h}{3C_4 q^5}\|(\nabla\nabla u_h + q^2\boldsymbol{T}u_h)\cdot\vec{n}\|^2_{0,\Gamma_1\cup\Gamma_3} + \frac{1}{qh^3}\|u_h\|^2_{0,\Gamma_0\cup\Gamma_1} \\
&+ \frac{1}{q^3 h}\|\nabla u_h\|^2_{\Gamma_1\cup\Gamma_3}.
\end{aligned}$$

That is, there exists a constant $C_2 = \min\{\frac{C_5}{3}, \frac{C_5}{3C_3}, \frac{C_5}{3C_4}, 1\}$ such that $A_h(u_h, u_h) \geq C_2 |||u_h|||^2_{2,q,h}$. □

Remark 4.4.3 (Galerkin orthogonality)**.** Let $u \in H^s(\Omega)$, $s \geq 4$ be the solution of (4.10). If $u_h \in \mathrm{ARG}_5(\Omega, \tau_h)$ is the solution of (4.38), then $A_h(u - u_h, \phi_h) = 0$, $\forall \phi_h \in \mathrm{ARG}_5(\Omega, \tau_h)$.

We carry out standard error analysis using the Galerkin orthogonality property, following the same approach as used for Poisson's equation in [128].

Lemma 11. *Let the assumptions of Theorems 12 and 15 hold. Given functions $v \in H^t(\Omega)$, $t \geq 4$ and $v_h \in ARG_5(\Omega, \tau_h)$, there exists a positive constant, C, such that*

$$\begin{aligned}
|||v - v_h|||_{2,q,h} &\leq \frac{C}{h^2}\left(\|v - v_h\|_0 + \frac{h}{q}\|v - v_h\|_1 + \frac{h^2}{q^2}\|v - v_h\|_2 + \frac{h^3}{q^3}\sum_{\tau\in\tau_h}\|v - v_h\|_{3,\tau}\right) \\
&+ \frac{C}{h^2}\left(\frac{h^4}{q^4}\sum_{\tau\in\tau_h}\|v - v_h\|_{4,\tau}\right). \qquad (4.44)
\end{aligned}$$

Proof. Define $r = v - v_h$, and note that $r \in H^4(\tau), \forall \tau \in \tau_h$. Apply Theorem 12 to

the boundary integrals in (4.39), with $\epsilon^{\frac{1}{2}} = \frac{h}{q} \in (0,1)$, yields

$$\begin{aligned}\frac{h^3}{q^7}\|\nabla\cdot(\nabla\nabla r + q^2\boldsymbol{T}r)\cdot\vec{n}\|^2_{0,\partial\tau\cap(\Gamma_0\cup\Gamma_1)} &\leq C\frac{h^3}{q^7}\|\nabla\nabla\nabla r\|^2_{0,\partial\tau\cap(\Gamma_0\cup\Gamma_1)}\\ &+ \mu_1\frac{h^3}{q^3}\|\nabla r\|^2_{0,\partial\tau\cap(\Gamma_0\cup\Gamma_1)} + \mu_2\frac{h^3}{q^3}\|r\|^2_{0,\partial\tau\cap(\Gamma_0\cup\Gamma_1)}\\ &\leq C\frac{h^3}{q^7}\left(\frac{h}{q}\|\nabla\nabla\nabla\nabla r\|^2_{0,\tau} + \frac{q}{h}\|\nabla\nabla\nabla r\|^2_{0,\tau}\right)\\ &+ C\frac{h^3}{q^3}\left(\frac{h}{q}\|\nabla\nabla r\|^2_{0,\tau} + \frac{q}{h}\|\nabla r\|^2_{0,\tau} + \frac{q}{h}\|r\|^2_{0,\tau}\right)\\ &\leq C\left(\frac{h^4}{q^8}\|\nabla\nabla\nabla\nabla r\|^2_{0,\tau} + \frac{h^2}{q^6}\|\nabla\nabla\nabla r\|^2_{0,\tau}\right)\\ &+ C\left(\frac{h^4}{q^4}\|\nabla\nabla r\|^2_{0,\tau} + \frac{h^2}{q^2}\|\nabla r\|^2_{0,\tau} + \frac{h^2}{q^2}\|r\|^2_{0,\tau}\right)\end{aligned} \tag{4.45}$$

$$\begin{aligned}\frac{h}{q^5}\|(\nabla\nabla r + q^2\boldsymbol{T}r)\cdot\vec{n}\|^2_{0,\partial\tau\cap(\Gamma_1\cup\Gamma_3)} &\leq C\frac{h}{q^5}\|\nabla\nabla r\|^2_{0,\partial\tau\cap(\Gamma_1\cup\Gamma_3)} + C\mu_1\frac{h}{q}\|r\|^2_{0,\partial\tau\cap(\Gamma_1\cup\Gamma_3)}\\ &\leq C\left(\frac{h^2}{q^6}\|\nabla\nabla\nabla r\|^2_{0,\tau} + \frac{1}{q^4}\|\nabla\nabla r\|^2_{0,\tau}\right)\\ &+ C\left(\frac{h^2}{q^2}\|\nabla r\|^2_{0,\tau} + \|r\|^2_{0,\tau}\right),\end{aligned} \tag{4.46}$$

$$\frac{1}{q^3h}\|\nabla r\|^2_{0,\partial\tau\cap(\Gamma_1\cup\Gamma_3)} \leq C\left(\frac{1}{q^4}\|\nabla\nabla r\|^2_{0,\tau} + \frac{1}{q^2h^2}\|\nabla r\|^2_{0,\tau}\right), \tag{4.47}$$

(4.48)

and

$$\frac{1}{qh^3}\|r\|^2_{0,\partial\tau\cap(\Gamma_0\cup\Gamma_1)} \leq C\left(\frac{1}{q^2h^2}\|\nabla r\|^2_{0,\tau} + \frac{1}{h^4}\|r\|^2_{0,\tau}\right). \tag{4.49}$$

Summing inequalities (4.45)-(4.49) over $\tau \in \tau_h$ and then combining these with the

fact that $r \in H^2(\Omega)$ leads to

$$|||v - v_h|||^2_{2,q,h} \leq \frac{C}{h^4}\left(\|v - v_h\|_0^2 + \frac{h^2}{q^2}\|v - v_h\|_1^2 + \frac{h^4}{q^4}\|v - v_h\|_2^2 + \frac{h^6}{q^6}\sum_{\tau\in\tau_h}\|v\|^2_{3,\tau}\right) + \frac{C}{h^4}\left(\frac{h^8}{q^8}\sum_{\tau\in\tau_h}\|v - v_h\|^2_{4,\tau}\right) \tag{4.50}$$

Taking the square root of both sides and using the fact that $\sqrt{\sum_{i=1}^4 x_i^2} \leq \sum_{i=1}^4 x_i,\ \forall x_i > 0$ completes the proof. □

Theorem 16. *Let the assumptions of Lemma 11 hold, let $u_h \in ARG_5(\Omega, \tau_h)$ be the solution of* (4.38)*, and let $u \in H^t(\Omega),\ 4 \leq t \leq 6$ be the solution of* (4.10)*. Then,*

$$|||u - u_h|||_{2,q,h} \leq CBq^4h^{t-2}|u|_t. \tag{4.51}$$

Proof. For any $v_h \in \mathrm{ARG}_5(\Omega, \tau_h)$, we have by the triangle inequality

$$|||u - u_h|||_{2,q,h} \leq |||u - v_h|||_{2,q,h} + |||u_h - v_h|||_{2,q,h}, \tag{4.52}$$

By the coercivity and continuity of A_h, and Remark 4.4.3,

$$\begin{aligned} |||u_h - v_h|||^2_{2,q,h} &\leq CA_h(u_h - v_h, u_h - v_h) = CA(u - v_h, u_h - v_h) && (4.53)\\ &\leq CBq^4|||u - v_h|||_{2,q,h}|||u_h - v_h|||_{2,q,h} && (4.54)\end{aligned}$$

Therefore, $|||u_h - v_h|||_{2,q,h} \leq CBq^4|||u - v_h|||_{2,q,h}$ and

$$|||u - u_h|||_{2,q,h} \leq CBq^4\left(\inf_{v_h\in\mathrm{ARG}_5(\Omega,\tau_h)}|||u - v_h|||_{2,q,h}\right). \tag{4.55}$$

Applying Lemma 11 and existing bounds on the Argyris interpolation operator [35] leads to the bound

$$\begin{aligned}|||u - u_h|||_{2,q,h} \leq& \frac{CBq^4}{h^2}\inf_{v_h\in\mathrm{ARG}_5(\Omega,\tau_h)}\left(\|u - v_h\|_0 + \frac{h}{q}|u - v_h|_1 + \frac{h^2}{q^2}|u - v_h|_2\right.\\ &\left.+\frac{h^3}{q^3}\sum_{\tau\in\tau_h}|u - v_h|_{3,\tau} + \frac{h^4}{q^4}\sum_{\tau\in\tau_h}|u - v_h|_{4,\tau}\right)\\ \leq& CBq^4h^{t-2}|u|_t, \quad \text{for } 4 \leq t \leq 6.\end{aligned}$$

□

Remark 4.4.4. Comparing this with the bound in Corollary 10, we see a slight degradation in the power of q, but no degradation in h. Thus, if $u \in H^6(\Omega)$ and behaves like $e^{iq\vec{n}\cdot\vec{x}}$, we now seek a mesh with $h^4q^6 < 1$ when $Bq^4 = \mathcal{O}(1)$, which is still reasonable when we expect $q \approx 40$ at its largest. Numerical results show, however, that even this estimate is pessimistic, and that a reasonable error tolerance can generally be achieved when $h = \mathcal{O}(1)$, i.e., independent of q.

Remark 4.4.5. The discrete solution, u_h, may fail to exactly satisfy the essential boundary conditions, with $u_h \neq 0$ on $\Gamma_0 \cup \Gamma_1$ and/or $\nabla u_h \neq 0$ on $\Gamma_1 \cup \Gamma_3$. Nevertheless, the error in these terms on the boundary also converges to zero, since we have the bounds

$$\|u_h\|_{0,\Gamma_0\cup\Gamma_1} = \|u - u_h\|_{0,\Gamma_0\cup\Gamma_1} \leq (qh^3)^{1/2}|||u - u_h|||_{2,q,h} \leq CBq^{9/2}h^{t-1/2}|u|_t, \tag{4.56}$$

$$\|\nabla u_h\|_{0,\Gamma_1\cup\Gamma_3} = \|\nabla(u - u_h)\|_{0,\Gamma_0\cup\Gamma_3} \leq (q^3h)^{1/2}|||u - u_h|||_{2,q,h} \leq CBq^{11/2}h^{t-3/2}|u|_t. \tag{4.57}$$

Remark 4.4.6. Preliminary results, not reported here, showed that the symmetric version of Nitsche's method resulted in the same dependence on h as above, but with worse dependence on q. In particular, we can also recover optimal-in-h convergence for the error in u measured in the L^2-norm in that setting, but with a dramatic increase in the power of q in the approximation results. Here, we can prove a slight improvement in the L^2-error estimate for u using arguments similar to [48, Proposition 5.3], but such estimates have little value, since they again trade worse dependence on q for better dependence on h. Whether such results can be improved (e.g., using a nonsymmetric penalty-free version of Nitsche's method, as in [48]) is left for future work.

4.4.2 C0IP methods

We next apply a C0IP method for the primal formulation (4.10), aiming to approximate the solution with a $H^1(\Omega)$-conforming function and to weakly enforce $H^2(\Omega)$-conformity. Such an interior penalty method for the biharmonic operator, with either homogeneous clamped boundary conditions or Cahn–Hilliard type boundary conditions with a vanishing corner DoF (to guarantee uniqueness of the solution), was

presented in [39]. As in that approach, we use Nitsche-type penalty methods to implement essential boundary conditions on the gradient, but strongly impose essential boundary conditions on the solution. The Nitsche term is, consequently, added to $\Gamma_1 \cup \Gamma_3$. In this case, the weak form is to find $u_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h),\ k \geq 2$, such that

$$\tilde{a}_h(u_h, \phi_h) = \int_\Omega f\phi_h, \quad \forall \phi_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h), \tag{4.58}$$

where

$$\begin{aligned}\tilde{a}(u_h, \phi_h) = \hat{a}(u_h, \phi_h) &- B \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla u_h + q^2 \boldsymbol{T} u_h \right) \cdot \vec{n} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial \phi_h}{\partial n} \right]\!\!\right] \\ &+ B \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla \phi_h + q^2 \boldsymbol{T} \phi_h \right) \cdot \vec{n} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right] \\ &+ \frac{1}{q^3 h} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right] \left[\!\!\left[\frac{\partial \phi_h}{\partial n} \right]\!\!\right],\end{aligned}$$

and

$$\begin{aligned}\hat{a}(u_h, \phi_h) = B \sum_{\tau \in \tau_h} &\left(\int_\tau \nabla\nabla u_h : \nabla\nabla \phi_h + Bq^2 \int_\tau \nabla\nabla u_h : \boldsymbol{T}\phi_h + Bq^2 \int_\tau \nabla\nabla \phi_h : \boldsymbol{T} u_h \right) \\ &+ \int_\Omega Bq^4 (\boldsymbol{T} : \boldsymbol{T} + m) u_h \phi_h.\end{aligned}$$

Here, $[\![\cdot]\!]$, $\{\!\{\cdot\}\!\}$ denote the standard jump and average functions defined in [18,39,42,43], and τ_h and ϵ_h are the sets of cells and edges (including the boundary) in the mesh, respectively. We define the following norm on $CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$,

$$\begin{aligned}|||u_h|||_h^2 = q^{-4} &\left(\sum_{\tau \in \tau_h} |u_h|_{2,\tau}^2 + \|\nabla u_h\|_0^2 \right) + \|u_h\|_0^2 \\ &+ \frac{h}{q^5} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla u_h + q^2 \boldsymbol{T} u_h \right) \cdot \vec{n} \right\}\!\!\right\}^2 \\ &+ \frac{1}{q^3 h} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right]^2\end{aligned} \tag{4.59}$$

The following inequalities are useful in proving the well-posedness of (4.58).

Lemma 12. *Let $\{\tau_h\}$ be a family of quasiuniform meshes of Ω and $\boldsymbol{T} : \boldsymbol{T} \leq \mu_1$*

pointwise on $\bar{\Omega}$.

- *The* H^2*-discrete Poincaré inequality is that there exists* $C_1 > 0$ *such that*

$$\|\nabla u_h\|_0^2 \leq C_1 \left(\|u_h\|_0^2 + \sum_{\tau \in \tau_h} |u_h|_{2,\tau}^2 + \frac{1}{h} \sum_{e \in \epsilon_h \setminus \partial\Omega} \int_e \left[\!\left[\frac{\partial u_h}{\partial n} \right]\!\right]^2 \right), \quad \forall \phi_h \in CG_k(\Omega, \tau_h). \tag{4.60}$$

- *There exists* $C_2 > 0$ *such that*

$$\sum_{e \in \epsilon_h} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla\phi_h + q^2 \boldsymbol{T} \phi_h \right) \cdot \vec{n} \right\}\!\!\right\}^2 \leq \frac{C_2}{h} \left(\left(\sum_{\tau \in \tau_h} |\phi_h|_{2,\tau}^2 \right) + q^4 \|\phi_h\|_0^2 \right) \tag{4.61}$$

Proof. From [44, Example 5.4], we have

$$\|u_h\|_1^2 \leq C \left(\sum_{\tau \in \tau_h} |u_h|_{2,\tau}^2 + \frac{1}{h} \sum_{e \in \epsilon_h \setminus \partial\Omega} \int_e \left[\!\left[\frac{\partial u_h}{\partial n} \right]\!\right]^2 + [\Phi(u_h)]^2 \right), \tag{4.62}$$

where Φ is a seminorm that satisfies Equations (1.2), (1.3), (2.15), and (3.3) in [44]. As $\Phi(u) = \|u\|_0$ satisfies these properties, Inequality (4.60) holds. While (1.2) and (1.3) can be trivially proved, (2.15) is shown in [44, Corollary 2.2], and (3.3) follows directly from [44, Inequality 3.2]. To prove Inequality (4.61), we first use [39, Inequality 3.20] to bound the term containing $\vec{n} \cdot (\nabla\nabla\phi_h) \cdot \vec{n}$. To bound the remaining term, we apply standard inverse trace inequalities to get

$$\sum_{e \in \epsilon_h} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(q^2 \boldsymbol{T} \phi_h \right) \cdot \vec{n} \right\}\!\!\right\}^2 \leq \frac{C q^4}{h} \|\phi_h\|_0^2. \tag{4.63}$$

Adding this to the right-hand side of [39, Inequality 3.20] completes the proof. □

These inequalities are enough to establish coercivity (and, thus, well-posedness) of the discrete problem in (4.58).

Theorem 17. *Let* $\{\tau_h\}$ *be a family of quasiuniform meshes of* Ω, $f \in L^2(\Omega)$, *and* $\boldsymbol{T} : \boldsymbol{T} \leq \mu_1$ *pointwise on* $\bar{\Omega}$. *Then, Problem* (4.58) *is well-posed.*

Proof. The bilinear form $\tilde{a}$ defined in (4.58) is continuous and coercive in the norm defined in (4.59). Proving continuity is straightforward using the Cauchy–Schwarz

inequality, yielding a continuity constant that is $\mathcal{O}(Bq^4)$, as in the conforming case. Coercivity of $\tilde{a}$ can be proven by combining the inequalities of Lemma 12 and Cauchy–Schwarz inequality. By direct substitution, we have

$$\begin{aligned}\tilde{a}(u_h,u_h) &= B\sum_{\tau\in\tau_h}\left(\int_\tau |u_h|^2_{2,\tau} + 2q^2\int_\tau \nabla\nabla u_h : \boldsymbol{T}u_h\right) + \int_\Omega (Bq^4\boldsymbol{T}:\boldsymbol{T}+m)u_h^2 \\ &+ \frac{1}{q^3h}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e \left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]^2.\end{aligned}$$

Similar to the proof of Theorem 13, we can show that there exists $C_3>0$ such that

$$\begin{aligned}B\sum_{\tau\in\tau_h}\left(\int_\tau |u_h|^2_{2,\tau} + 2q^2\int_\tau \nabla\nabla u_h : \boldsymbol{T}u_h\right) + \int_\Omega (Bq^4\boldsymbol{T}:\boldsymbol{T}+m)u_h^2 \\ \geq C_3\left(q^{-4}\sum_{\tau\in\tau_h}|u_h|^2_{2,\tau} + \|u_h\|_0^2\right).\end{aligned}$$

With this, we have

$$\begin{aligned}\tilde{a}(u_h,u_h) &\geq C_3\left(q^{-4}\sum_{\tau\in\tau_h}|u_h|^2_{2,\tau} + \|u_h\|_0^2\right) + \frac{1}{q^3h}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]^2 \\ &\geq \frac{C_3}{3}\left(q^{-4}\sum_{\tau\in\tau_h}|u_h|^2_{2,\tau} + \|u_h\|_0^2\right) + \frac{2}{3q^3h}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]^2 \\ &+ \frac{C_3h}{3C_2q^4}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left\{\!\!\left\{\vec{n}\cdot\left(\nabla\nabla\phi_h + q^2\boldsymbol{T}\phi_h\right)\cdot\vec{n}\right\}\!\!\right\}^2 \\ &+ \frac{\min\{C_3,1\}}{3}\left(q^{-4}\sum_{\tau\in\tau_h}|u_h|^2_{2,\tau} + \|u_h\|_0^2 + \frac{1}{q^3h}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]^2\right) \\ &\geq \frac{\min\{C_3,1\}}{3}\left(q^{-4}\sum_{\tau\in\tau_h}|u_h|^2_{2,\tau} + \frac{q^{-4}}{C_1}\|\nabla u_h\|_0^2 + \|u_h\|_0^2\right) \\ &+ \frac{C_3h}{3C_2q^5}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left\{\!\!\left\{\vec{n}\cdot\left(\nabla\nabla\phi_h + q^2\boldsymbol{T}\phi_h\right)\cdot\vec{n}\right\}\!\!\right\}^2 \\ &+ \frac{2}{3q^3h}\sum_{e\in\epsilon_h\setminus\Gamma_0\cup\Gamma_2}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]^2,\end{aligned}$$

where C_1 and C_2 are defined in Lemma 12. Thus, the coercivity constant is $\mathcal{O}(1)$. $\square$

Remark 4.4.7. Note that the bilinear form $\tilde{a}$ is also continuous with respect to the mesh-dependent norm for functions in $H^t(\Omega) \cap H^1_{\Gamma_0 \cup \Gamma_1}(\Omega),\ t \geq 4$, where the jump terms vanish over interior edges. That is, there exists a positive constant, C, such that

$$\tilde{a}(u, \phi) \leq CBq^4 |||u|||_h |||\phi|||_h,\ \forall u, \phi \in H^t(\Omega) \cap H^1_{\Gamma_0 \cup \Gamma_1}(\Omega). \tag{4.64}$$

Lemma 13. *Let $\{\tau_h\}$ be a family of quasiuniform meshes of Ω, $v \in H^t(\Omega) \cap H^1_{\Gamma_0 \cup \Gamma_1}(\Omega)$, $s \geq 4$, $\boldsymbol{T} : \boldsymbol{T} \leq \mu_1$ pointwise on $\bar{\Omega}$, and $v_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$. Then,*

$$\begin{aligned} |||v - v_h|||_h^2 \leq \frac{C}{h^4} \left(\|v - v_h\|_0^2 + \frac{h^2}{q^2} |v - v_h|_1^2 + \frac{h^4}{q^4} \sum_{\tau \in \tau_h} |v - v_h|_{2,\tau} + \frac{h^6}{q^6} \sum_{\tau \in \tau_h} |v - v_h|_{3,\tau}^2 \right) \\ + \frac{C}{h^4} \left(\frac{h^8}{q^8} \sum_{\tau \in \tau_h} |v - v_h|_{4,\tau}^2 \right) \end{aligned} \tag{4.65}$$

Proof. First note that $\left[\!\left[\frac{\partial v}{\partial n}\right]\!\right] = 0$ on the interior edges of τ_h. For $v \in H^t(\Omega) \cap H^1_{\Gamma_0 \cup \Gamma_1}(\Omega),\ t \geq 4$, we have that $r = v - v_h \in H^4(\tau),\ \forall \tau \in \tau_h$. We apply Theorem 12 to the boundary integrals in (4.59) with $\epsilon^{\frac{1}{2}} = \frac{h}{q} \in (0, 1)$, yielding

$$\frac{1}{q^3 h} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left[\!\left[\frac{\partial r}{\partial n} \right]\!\right]^2 \leq C \sum_{\tau \in \tau_h} \left(\frac{1}{q^2 h^2} \|\nabla r\|_{0,\tau}^2 + \frac{1}{q^4} |r|_{2,\tau}^2 \right) \tag{4.66}$$

and

$$\begin{aligned} \frac{h}{q^5} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla \nabla r + q^2 \boldsymbol{T} r \right) \cdot \vec{n} \right\}\!\!\right\}^2 &\leq C \sum_{\tau \in \tau_h} \left(\frac{1}{q^4} |r|_{2,\tau}^2 + \frac{h^2}{q^6} |r|_{3,\tau} \right) \\ &+ C \left(\|r\|_0^2 + \frac{h^2}{q^2} |\nabla r|_1^2 \right). \end{aligned} \tag{4.67}$$

As a result, Inequality (4.65) holds. □

Remark 4.4.8 (Galerkin orthogonality)**.** Let $u \in H^t(\Omega),\ t \geq 4$ be the solution of (4.10). If $u_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$ is the solution of (4.38), then $A_h(u - u_h, \phi_h) = 0$, $\forall \phi_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$.

Theorem 18. *Assume that the solution of* (4.10) *satisfies $u \in H^t(\Omega)$, for $t \geq 4$. If $u_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$ is the solution of* (4.58)*, then*

$$|||u - u_h|||_h \leq CBq^4 h^{\min\{t, k+1\} - 2} |u|_t.$$

Proof. We use the orthogonality property to prove the error estimates as in Theorem 16. Given $u \in H^t(\Omega),\ t \geq 4$, then we have the standard quasi-optimality result,

$$|||u - u_h|||_h \leq CBq^4 \inf_{v_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega,\tau_h)} |||u - v_h|||_h. \tag{4.68}$$

Now, we use Lemma 13 and the standard Lagrange interpolation error estimates [41, 52] to yield

$$\begin{aligned} |||u - u_h|||_h &\leq \frac{CBq^4}{h^2}\left(\|u - v_h\|_0 + \frac{h}{q}\|\nabla(u - v_h)\|_0 + \frac{h^2}{q^2}\sum_{\tau\in\tau_h}|u - v_h|_{2,\tau}\right) \\ &+ \frac{CBq^4}{h^2}\left(\frac{h^3}{q^3}\sum_{\tau\in\tau_h}|u - v_h|_{3,\tau} + \frac{h^4}{q^4}\sum_{\tau\in\tau_h}|u - v_h|_{4,\tau}\right) \\ &\leq \frac{CBq^4}{h^2}h^{\min\{t,k+1\}}|u|_t = CBq^4h^{\min\{t,k+1\}-2}|u|_t. \end{aligned}$$

□

4.4.3 Mixed finite elements

We now consider a mixed finite-element discretization of the systems reformulation given in (4.28)-(4.29). We consider a conforming discretization, with $u_h \in DG_k(\Omega,\tau_h) \subset L^2(\Omega)$ and $\vec{\alpha}_h \in RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega,\tau_h) \subset H_{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)$. In order to prove the required inf-sup condition on $b(\vec{\alpha}_h,(u_h,\vec{v}_h))$, the choice of space for $\vec{v}_h$ is based on generalized Taylor–Hood elements, writing $\vec{v}_h \in V_h$, where

$$V_h = \left\{\vec{\psi}_h \;\middle|\; \vec{\psi}_h \in [CG_{k+2}(\Omega,\tau_h)]^2 \cap V\right\},$$

with (as before) $V = \left\{\vec{v} \in [H^1_{\Gamma_1\cup\Gamma_3}(\Omega)]^2 \middle|\ \vec{v}\times\vec{n} = 0 \text{ on } \Gamma_0\right\}$.

Theorem 19. *Let the assumptions of Theorem 14 be satisfied, and let τ_h be a quasi-uniform family of triangular meshes of Ω. Let the bilinear forms $\mathcal{A}$ and b and linear form F be defined as in* (4.30)-(4.32)*. For sufficiently small h, the discrete saddle-point problem of finding $(u_h,\vec{v}_h,\vec{\alpha}_h) \in DG_k(\Omega,\tau_h)\times V_h \times RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega,\tau_h)$ such that*

$$\begin{aligned} \mathcal{A}\big((u_h,\vec{v}_h),(\phi_h,\vec{\psi}_h)\big) + b(\vec{\alpha}_h,(\phi_h,\vec{\psi}_h)) &= F(\phi_h), \quad \forall(\phi_h,\vec{\psi}_h) \in DG_k(\Omega,\tau_h)\times V_h \\ b\big(\vec{\beta}_h,(u_h,\vec{v}_h)\big) &= 0, \quad \forall\vec{\beta} \in RT_{k+1}^{\Gamma_2\cup\Gamma_3}(\Omega,\tau_h) \end{aligned} \tag{4.69}$$

is well-posed for $k \geq 1$.

Proof. We follow the standard theory (see, e.g., [32]), requiring continuity of $\mathcal{A}$ and b, coercivity of $\mathcal{A}$, and an inf-sup condition on b. Because we consider a conforming discretization, continuity of both $\mathcal{A}$ and b follow directly as in Theorem 14, with the same constants in the norms used there, where we use the discrete Helmholtz decomposition of $\vec{\alpha}_h = \nabla \times p_h + \nabla_h^{\Gamma_2 \cup \Gamma_3} \phi_h$ defined in Lemma 10 to prove continuity of b. Similarly, we consider coercivity of the bilinear form $\mathcal{A}\left((u_h, \vec{v}_h), (\phi_h, \vec{\psi}_h)\right)$ on the set

$$\Lambda_h = \left\{ (u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times V_h \,\middle|\, b\left(\vec{\alpha}_h, (u_h, \vec{v}_h)\right) = 0,\ \forall \vec{\alpha}_h \in RT_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h) \right\}.$$

This follows again as in the continuum case, using $\vec{\beta}_1 = \begin{bmatrix} S_h \\ 0 \end{bmatrix}$ and $\vec{\beta}_2 = \begin{bmatrix} 0 \\ S_h \end{bmatrix}$, with $S_h \in CG_1^{\partial\Omega}(\Omega, \tau_h)$ as defined in Lemma 7 in place of the continuum analogues in Theorem 14.

Finally, we establish the discrete inf-sup condition, that

$$I = \sup_{(u_h, \vec{v}_h) \in DG_k(\Omega, \tau_h) \times V_h} \frac{\int_\Omega \vec{\alpha_h} \cdot \vec{v}_h + \int_\Omega u_h \nabla \cdot \vec{\alpha}_h}{\|(u_h, \vec{v}_h)\|_{0,q,1}} \geq C q^2 \|\vec{\alpha}_h\|_{\mathrm{Div}}, \tag{4.70}$$

for some constant, C. By Lemma 10, for any $\vec{\alpha}_h \in RT_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h)$, there exists $p_h \in CG_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h)$ and $\eta_h \in DG_k(\Omega, \tau_h)$ such that

$$\vec{\alpha}_h = \nabla \times p_h + \nabla_h^{\Gamma_2 \cup \Gamma_3} \eta_h. \tag{4.71}$$

This gives the equivalent form to (4.70) of

$$I = \sup_{(u_h, \vec{v}_h) \in DG_k \times V_h} \frac{\int_\Omega \left(\nabla_h^{\Gamma_2 \cup \Gamma_3} \eta_h + \nabla \times p_h\right) \cdot \vec{v}_h + \int_\Omega u_h \nabla \cdot \vec{\alpha}_h}{\sqrt{\|u_h\|_0^2 + q^{-4} \|\vec{v}_h\|_1^2}} \geq C q^2 \|\vec{\alpha}_h\|_{\mathrm{Div}},$$

$\forall \vec{\alpha}_h \in RT_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h)$. We show this by choosing $u_h = C_1 (\nabla \cdot \vec{\alpha}_h - \eta_h)$. Let h be sufficiently small so that the inf-sup condition of [30, Lemma 3.5] holds. Then, for all $p_h \in CG_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h)$, there exists a vector $\vec{\psi}_h \in V_h$ such that $\int_\Omega p_h \nabla \cdot \vec{\psi}_h \geq \|p_h\|_0^2$, and $\|\vec{\psi}_h\|_1^2 \leq C_2 \|\vec{p}_h\|_0^2$. To establish the inf-sup condition needed here, we choose $\vec{v}_h = [\psi_{2,h}, -\psi_{1,h}]^T$ which also belongs to V_h, giving $\nabla \cdot \vec{\psi}_h = \nabla \times \vec{v}_h$ and $\|\vec{\psi}_h\|_1^2 = \|\vec{v}_h\|_1^2$. The remainder of the proof follows identically as in the continuum case. □

To measure the error estimates that arise from our three-field mixed formulation, we define the approximation errors,

$$E_{(u,\vec{v})} := \inf_{(\phi_h,\vec{\psi}_h)\in DG_k(\Omega,\tau_h)\times V} \|(u,\vec{v})-(\phi_h,\vec{\psi}_h)\|_{0,q,1}, \tag{4.72}$$

$$E_{\vec{\alpha}} := \inf_{\vec{\beta}_h\in RT_{k+1}(\Omega,\tau_h)} \|\vec{\alpha}-\vec{\beta}_h\|_{\mathrm{Div}}. \tag{4.73}$$

Corollary 11. *Let the assumptions of Theorem 19 be satisfied. Assume that $u \in H^{k+5}$ and $\boldsymbol{T} \in \boldsymbol{C}^{k+2}(\Omega)$, for $k \geq 1$, $(u,\vec{v},\vec{\alpha})$ is the unique solution of Problem* (4.28)-(4.29)*, and $(u_h,\vec{v}_h,\vec{\alpha}_h)$ is the solution of Problem* (4.69)*. Then,*

$$\|(u,\vec{v})-(u_h,\vec{v}_h)\|_{0,q,1} \leq C_1\left(\left(Bq^4+B^{1/2}q^4\right)E_{(u,\vec{v})}+q^4E_{\vec{\alpha}}\right), \tag{4.74}$$

$$\|\vec{\alpha}-\vec{\alpha}_h\|_{\mathrm{Div}} \leq C_2\left(\left(B^{3/2}q^4+Bq^4\right)E_{(u,\vec{v})}+B^{1/2}q^4E_{\vec{\alpha}}\right), \tag{4.75}$$

where C_1, C_2 are positive constants independent of h, B and q.

Proof. The standard error estimate, for example in [32, Theorem 5.2.2], leads to (4.74) and (4.75). Note that $\mathcal{A}$ and b in (4.28)-(4.29) are continuous with $\mathcal{O}(Bq^4)$ and $\mathcal{O}(q^4)$ continuity constants, respectively, the coercivity constant is $\mathcal{O}(1)$, and the inf-sup constant is $\mathcal{O}(q^2)$. □

In the next corollary, we bound the approximation errors $E_{(u,\vec{v})}$ and $E_{\vec{\alpha}}$ when $u \in H^{k+5}(\Omega)$ and $\boldsymbol{T} \in \boldsymbol{C}^{k+2}(\Omega)$. Note that, in this case, $\vec{v} = \nabla u \in \left[H^{k+4}(\Omega)\right]^2$ and $\vec{\alpha} = \nabla\cdot(\nabla\nabla u + q^2\boldsymbol{T}u) \in \left[H^{k+2}(\Omega)\right]^2$ by Corollary 9.

Corollary 12. *Let the assumptions of Corollary 11 be satisfied and write $\vec{\alpha} = \nabla\phi + \nabla\times p$. If, furthermore, $\phi \in H^{k+3}(\Omega)$ and $p \in \Phi$ (as defined in Remark 4.2.2), then*

$$E_{(u,\vec{v})} \leq Ch^{k+1}\left(|u|^2_{k+1}+\frac{1}{q^4}|\vec{v}|^2_{k+3}\right)^{1/2}, \tag{4.76}$$

$$E_{\vec{\alpha}} \leq \frac{C}{q^2}h^{k+1}\left(|\nabla\phi|^2_{k+1}+|\Delta\phi|^2_{k+1}+|p|^2_{k+2}\right)^{1/2}. \tag{4.77}$$

Proof. Inequality (4.76) holds using the classical continuous/discontinuous Lagrange interpolants [32]. To prove Inequality (4.77), we use the fact that $\vec{\alpha} \in [H^{k+2}(\Omega)]^2 \cap H_{\Gamma_2\cup\Gamma_3}(\mathrm{div};\Omega)$ and, therefore, the functions p and ϕ are in $H^{k+2}(\Omega)\cap H^1_{\Gamma_0\cup\Gamma_1}(\Omega)$ and $H^{k+3}(\Omega)\cap H^1_{\Gamma_2\cup\Gamma_3}(\Omega)$ respectively if the conditions of Remark 4.2.2 are satisfied. We

bound $\|\vec{\alpha} - \vec{\beta}_h\|_{\text{Div}}$ by writing $\vec{\beta}_h = \nabla_h^{\Gamma_2 \cup \Gamma_3} \phi_h + \nabla \times p_h$ and noting that

$$\|\vec{\alpha} - \vec{\beta}_h\|_{\text{Div}}^2 = q^{-4} \left(\|p - p_h\|_0^2 + \|\nabla \phi - \nabla_h^{\Gamma_2 \cup \Gamma_3} \phi_h\|_{\text{div}}^2 \right).$$

We choose $\left(\phi_h, \vec{\zeta}_h \right) \in DG_k(\Omega, \tau_h) \times RT_{k+1}^{\Gamma_2 \cup \Gamma_3}(\Omega, \tau_h)$ to be the solution of the mixed Poisson problem,

$$\int_\Omega \gamma_h \nabla \cdot \vec{\zeta}_h = \int_\Omega \Delta \phi \gamma_h, \quad \forall \gamma_h \in DG_k(\Omega, \tau_h),$$
$$\int_\Omega \vec{\zeta}_h \cdot \vec{\Upsilon}_h + \phi_h \nabla \cdot \vec{\Upsilon}_h = 0, \quad \forall \vec{\Upsilon}_h \in RT_{k+1}^{\Gamma_a}(\Omega, \tau_h).$$

The standard error estimate for $\vec{\zeta}_h$ is that

$$\|\vec{\zeta}_h - \nabla \phi\|_{\text{div}} \leq C h^{k+1} \left(|\nabla \phi|_{k+1} + |\Delta \phi|_{k+1} \right);$$

however, $\vec{\zeta}_h = \nabla_h^{\Gamma_2 \cup \Gamma_3} \phi_h$, giving

$$\|\nabla \phi - \nabla_h^{\Gamma_2 \cup \Gamma_3} \phi_h\|_{\text{div}} \leq C h^{k+1} \left(|\nabla \phi|_{k+1} + |\Delta \phi|_{k+1} \right). \tag{4.78}$$

Choosing p_h to be the interpolant of p in $CG_{k+1}^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h)$ gives

$$\|p - p_h\|_0 \leq C h^{k+2} |p|_{k+2}. \tag{4.79}$$

Adding Inequalities (4.78) and (4.79) leads to (4.77). □

While we can always compute the discrete Helmholtz decomposition of $\vec{\alpha}_h$, it is not always possible to compute the corresponding continuum Helmholtz decomposition of $\vec{\alpha}$, which would be needed to verify the above results by computing $\|\vec{\alpha} - \vec{\alpha}_h\|_{\text{Div}}$. Therefore, we use the $H(\text{div})$ norm in practice. We next show that the approximation error of $\vec{\alpha}$ in the $H(\text{div})$ norm can be bounded by that in the strengthened norm.

Corollary 13. *Let the assumptions of Corollary 12 be satisfied. Then,*

$$q^{-2} \|\vec{\alpha} - \vec{\alpha}_h\|_0 \;\; \leq \;\; C \left(h^{k+1} |p|_{k+2} + \frac{1}{h} \|\vec{\alpha} - \vec{\alpha}_h\|_{\text{Div}} \right), \tag{4.80}$$

where C is a positive constant independent of h.

Proof. Write $\vec{\alpha}_h = \nabla_h^{\Gamma_3}\phi_h + \nabla \times p_h$. Then we have

$$\begin{aligned} \|\vec{\alpha} - \vec{\alpha}_h\|_0 &\leq C\left(\|\nabla\phi - \nabla_h^{\Gamma_3}\phi_h\|_0 + \|\nabla \times p - \nabla \times p_h\|_0\right) \\ &\leq C\left(\|\vec{\alpha} - \vec{\alpha}_h\|_{\mathrm{Div}} + \|\nabla \times p - \nabla \times p_h\|_0\right). \end{aligned} \tag{4.81}$$

Thus, we only need to bound $\|\nabla \times p - \nabla \times p_h\|_0$. To do this, we note that

$$\|\nabla \times p - \nabla \times p_h\|_0 \leq \|\nabla \times p - \nabla \times z_h\|_0 + \|\nabla \times z_h - \nabla \times p_h\|,$$

where z_h is the interpolant of p in $CG_{k+1}(\Omega, \tau_h)$, for which $\|\nabla \times p - \nabla \times z_h\|_0 \leq Ch^{k+1}|p|_{k+2}$. Also, using standard arguments, we have that $\|\nabla \times z_h - \nabla \times p_h\|_0 \leq \frac{C}{h}\|z_h - p_h\|_0$. Thus,

$$\begin{aligned} \|\nabla \times p - \nabla \times p_h\|_0 &\leq C\left(h^{k+1}|p|_{k+2} + \frac{1}{h}\|z_h - p_h\|_0\right) \\ &\leq C\left(h^{k+1}|p|_{k+2} + \frac{1}{h}\|z_h - p\|_0 + \frac{1}{h}\|p - p_h\|_0\right) \\ &\leq C\left(h^{k+1}|p|_{k+2} + \frac{1}{h}\|p - p_h\|_0\right) \\ &\leq C\left(h^{k+1}|p|_{k+2} + \frac{1}{h}\|\vec{\alpha} - \vec{\alpha}_h\|_{\mathrm{Div}}\right). \end{aligned} \tag{4.82}$$

Combining Inequalities (4.81) and (4.82) leads to (4.80). □

Remark 4.4.9. Let the assumptions of Corollary 12 be satisfied, and let $W = \left(Bq^2 + B^{1/2}q^2\right)$, $Z_1 = \left(|u|_{k+1}^2 + \frac{1}{q^4}|\vec{v}|_{k+3}^2\right)^{1/2}$, and $Z_2 = \left(|\nabla\phi|_{k+1}^2 + |\Delta\phi|_{k+1}^2 + |p|_{k+2}^2\right)^{1/2}$. Then,

$$\begin{aligned} \|(u, \vec{v}) - (u_h, \vec{v}_h)\|_{0,q,1} &\leq C_1 q^2 h^{k+1}\left(WZ_1 + Z_2\right), \\ \frac{1}{q^2}\|\vec{\alpha} - \vec{\alpha}_h\|_0 \leq \frac{C}{h}E_{\vec{\alpha}} &\leq C_2 B^{1/2} q^2 h^k\left(WZ_1 + Z_2\right), \\ \frac{1}{q^2}\|\nabla \cdot \vec{\alpha} - \nabla \cdot \vec{\alpha}_h\|_0 \leq E_{\vec{\alpha}} &\leq C_3 B^{1/2} q^2 h^{k+1}\left(WZ_1 + Z_2\right). \end{aligned}$$

4.5 Numerical experiments

To verify the analyses of the three finite-element discretizations, we next present numerical experiments to measure convergence rates. The experiments were done using the finite-element package Firedrake [115], which offers close integration with PETSc for the linear solvers [20, 93]. All numerical experiments were run on a workstation with dual 8-core Intel Xeon 1.7 GHz CPUs and 384 GB of RAM. While development of efficient linear solvers for these discretizations is an important task, we consider only solution using the sparse direct solver, PaStiX [83], in all cases.

We use the method of manufactured solutions to estimate convergence rates, where we fix forcing terms and boundary data for the PDE to exactly match those for a known solution, u. In all experiments, we consider uniform triangular meshes of the unit square in two dimensions, generated by uniformly meshing the unit square into square elements with edge length $h = 1/N$, and then cutting each square into two triangles, from bottom left to top right. For the tests below, we write the boundary of the unit square as $\partial\Omega = \Gamma_N \cup \Gamma_S \cup \Gamma_E \cup \Gamma_W$, denoting the North, South, East, and West edges of the square, and fix $\Gamma_0 = \Gamma_S, \Gamma_1 = \Gamma_N, \Gamma_2 = \Gamma_E$, and $\Gamma_3 = \Gamma_W$. For mesh size h, we define u_h to be the finite-element solution on the mesh and the approximation error to be $E_h = u - u_h$. We can measure E_h in several ways, such as the absolute $L_2(\Omega)$-error, which we denote by $Abs_e(u_h, h) = \|E_h\|_0$. Similar definitions are used, as needed, for other quantities, such as the weighted $H^2(\Omega)$-error in u_h as given in Section 4.4.1, the weighted $L^2(\Omega) \times H^1(\Omega)$-error in $(u_h, \vec{v}_h)$ as given in Equation (4.33), and the weighted $L^2(\Omega)$-error and $H(\mathrm{div})$-seminorm errors in $\vec{\alpha}_h$.

We first consider an exact solution given by $u = \sin(q\vec{\nu} \cdot [x, y])$, with $q = 40$, $\boldsymbol{T} = \vec{\nu} \otimes \vec{\nu}$, $\vec{\nu} = \left[\frac{3}{5}, \frac{4}{5}\right]$, $m = 10$, and either $B = 1$ or q^{-4}. We plot $\log(Abs_e(\cdot))$ against $\log_2(1/h)$, so that slopes of the data plotted correspond to the experimental convergence rates. We approximate the slope, S, of each line using the last two points. In Figure 4.1, we show results for both the conforming method, with $u_h \in \mathrm{ARG}_5(\Omega, \tau_h)$ (in green), and the C0IP method with $u_h \in CG_k(\Omega, \tau_h)$, $k = 2, 3, 4$, (in blue, red, and orange, respectively). Triangles denote the absolute error in the L^2-norm and filled discs denote the weighted H^2-norm defined in (4.39) for Argyris elements and the norm defined in (4.59) for the C0IP method. We see that these results agree with the theoretical analysis in Sections 4.4.1 and 4.4.2. For $B = 1$ (shown at left), we note that the absolute errors in both the L^2-norm and the weighted H^2-norm are slightly

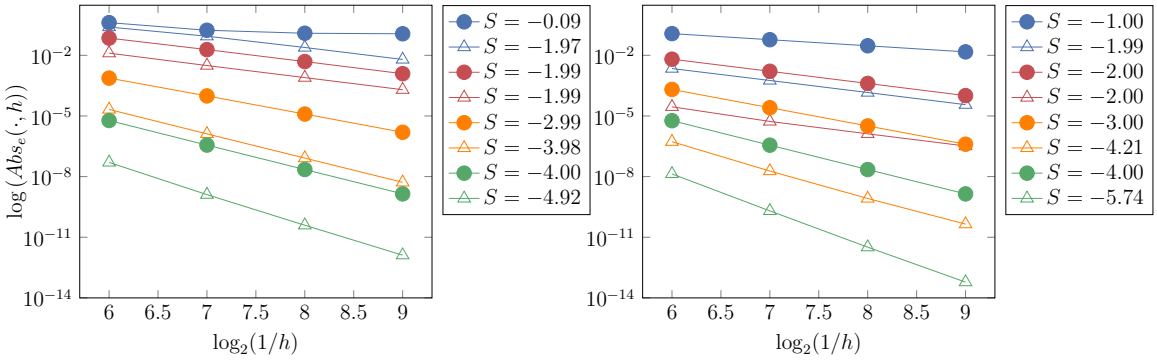

Figure 4.1: Absolute approximation errors and rate of convergence with $u \in \mathrm{ARG}_5(\Omega, \tau_h)$ (Green), and $u \in CG_k(\Omega, \tau_h)$, $k = 2, 3, 4$ with the C0IP formulation, where blue, red, and orange lines present results for $k = 2, 3, 4$, respectively. Triangles denote errors in the L^2 norm, while filled discs denote errors in the appropriately weighted H^2 norm. Left: $B = 1$. Right: $B = q^{-4}$.

larger than for the case $B = q^{-4}$ (shown at right), which is expected as the error estimates depend on Bq^4. Moreover, for the C0IP method with $u \in CG_2(\Omega, \tau_h)$, we see poor convergence that agrees with the fact that Bq^4h is quite large for values of q and h considered here. Finally, we see that, with Argyris elements, the convergence rate in the L^2-norm tends to be optimal for $B = q^{-4}$, but suboptimal when $B = 1$. The degraded convergence rates in the L^2-norm for both conforming and C0IP methods result from the use of nonsymmetric versions of Nitsche's method and C0IP, as expected.

Figure 4.2 presents results for the 3-field mixed finite-element discretization with $(u_h, \vec{v}_h, \vec{\alpha}_h) \in DG_k(\Omega, \tau_h) \times V_{k+2} \times RT_{k+1}(\Omega, \tau_h)$. Here, blue, red, and green lines present results for $k = 1, 2, 3$, while filled discs and squares denote the $L^2(\Omega)$ and weighted $H^1(\Omega)$ errors for u_h and $\vec{v}_h$, respectively, while triangles and diamonds denote the weighted $L^2(\Omega)$ and $H(\mathrm{div}; \Omega)$-seminorm errors for $\vec{\alpha}_h$. We see optimal convergence rates for u, degraded H^1 convergence for $\vec{v}$, which is expected because of the mismatch between the orders of DG_{k+1} and V_{k+2}, and optimal convergence rates for $\vec{\alpha}$. These results are consistent with Corollary 13. While the analysis of our mixed formulations in Remark 4.4.9 show that the convergence rates depend on B, we see that the error estimates for both $B = 1$ and $B = q^{-4}$ are almost the same for u and $\vec{v}$. The absolute errors in $\vec{\alpha}$ with $B = q^{-4}$ are equal to those with $B = 1$ multiplied by q^{-4}. The fact that the experimental error estimates of the mixed formulation are independent of B is a substantial advantage over the conforming and C0IP methods.

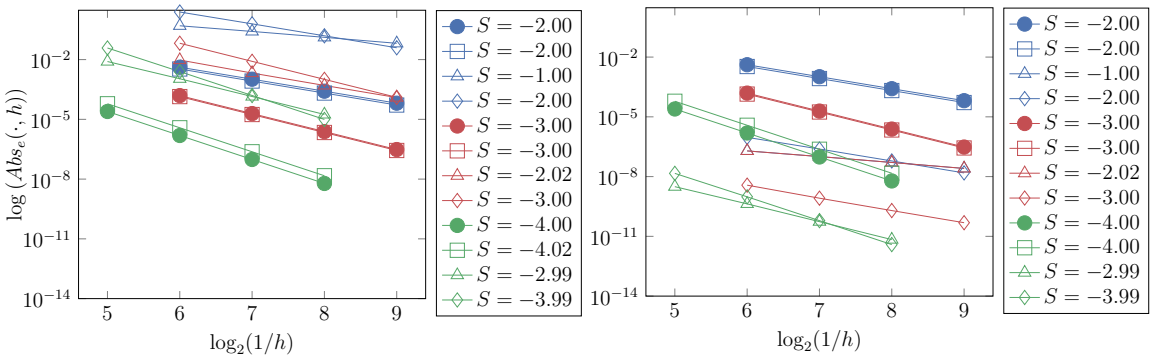

Figure 4.2: Absolute approximation errors and rates of convergence for the mixed formulation analyzed in Section 4.4.3, with $(u, \vec{v}, \vec{\alpha}) \in DG_k(\Omega, \tau_h) \times V_{k+2}(\Omega, \tau_h) \times RT_{k+1}(\Omega, \tau_h)$, with blue, red, and green lines presenting results for $k = 1, 2, 3$, respectively. Filled discs and squares denote the $L^2(\Omega)$ and weighted $H^1(\Omega)$ errors for u_h and $\vec{v}_h$, respectively, while triangles and diamonds denote the weighted $L^2(\Omega)$ and $H(\mathrm{div};\Omega)$-seminorm errors for $\vec{\alpha}_h$. Left: $B = 1$. Right: $B = q^{-4}$.

Note that for $k = 3$ and $h = 1/512$, we have so many DoFs that the direct solver fails, motivating future work on deriving efficient iterative solution algorithms.

An important question is whether the dependence on q in the theoretical results above is due to inefficient proof techniques, or is an actual dependence that is seen in the finite-element results. Figure 4.3 presents results for the same boundary conditions as above, considering two test solutions, $u = \sin(q(\frac{3x}{5} + \frac{4y}{5}))$ from above (left) and $u_1 = 100\sin(2\pi x + 3\pi y)\left(xy(1-x)(1-y)\right)^3$(right), with $h = 2^{-7}$. When $B = 1$ (solid lines), we see that the errors grow more slowly with q than predicted by the theoretical results. We note that these results are consistent with $\mathcal{O}(1)$ coercivity and continuity constants, scaling instead like the relevant semi-norms of u in the error bounds. To validate this hypothesis, we consider the case where u is independent of q, with $u = 100\sin(2\pi x + 3\pi y)\left(xy(1-x)(1-y)\right)^3$ (dashed lines). Here, we see no dependence on q in the error bounds for the Argyris or C0IP discretizations, and a moderate dependence on q in the H^2 norm error for the mixed method. This suggests that a finer analysis may improve the dependence of the error bounds on q.

4.6 Conclusions

We consider different finite-element techniques to discretize a fourth-order PDE describing the density variation of a smectic A liquid crystal. These models have two

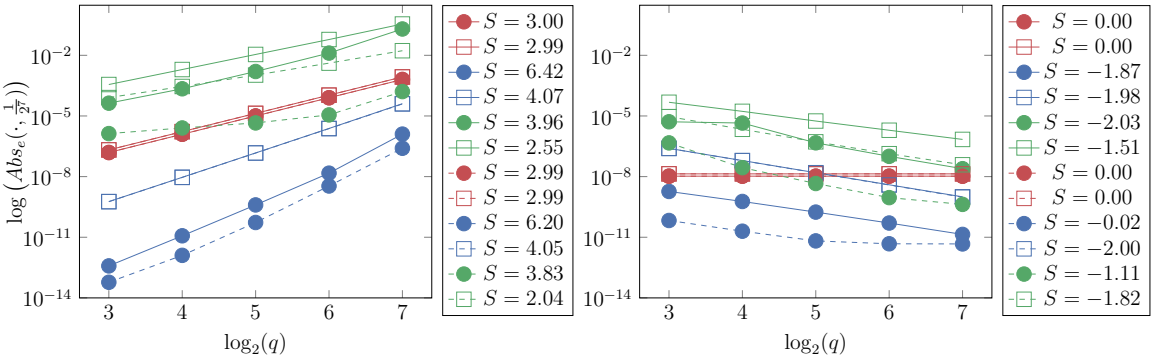

Figure 4.3: The L^2 (filled discs) and weighted H^2 (squares) absolute approximation errors at $1/h = 2^7$ and different values of q, for the conforming method, with $u \in \mathrm{ARG}_5(\Omega, \tau_h)$ (blue), the C0IP method with $u \in CG_3(\Omega, \tau_h)$ (green), and the mixed method with $(u, \vec{v}) \in DG_2(\Omega, \tau_h) \times V_4(\Omega, \tau_h)$ (red) for $B = 1$(solid lines) and $B = q^{-4}$(dashed lines). Left: $u = \sin(q\left(\frac{3x}{5} + \frac{4y}{5}\right))$. Right: $u = 100\sin(2\pi x + 3\pi y)\left(xy(1-x)(1-y)\right)^3$.

complications in comparison to classical biharmonic operators, as they are more akin to Helmholtz operators than elliptic ones, and involve Hessian-squared (div-div-grad-grad) operators rather than the classical biharmonic operator (div-grad-div-grad), with boundary conditions that preclude this potential simplification. We analyzed H^2-conforming, C0IP, and mixed finite-element methods.

In the H^2-conforming case, we use C^1 Argyris/Zhang elements. In practice, these elements can be expensive to work with, due to their high order (fifth-order piecewise polynomials in 2D and ninth-order in 3D), but they offer high-order approximation of smooth solutions as well. In this case, we implement essential boundary conditions using non-symmetric Nitsche-type penalty methods, which somewhat degrades the error estimates from the case where essential BCs are imposed strongly. C0IP methods have the advantage over H^2-conforming elements that there is greater flexibility in choosing the order of approximation, at the cost of more complicated weak forms, where C^1-conformity is weakly enforced by penalizing inter-element jumps in the first derivative. Our error estimates in this case match the dependence on q from the conforming case, but with an h-dependence in line with the lower polynomial order. Finally, we consider a three-field mixed finite-element formulation that explicitly introduces the gradient as an independent variable constrained using a Lagrange multiplier. The mixed formulation offers better robustness in B than the other schemes. Numerical results confirm the theoretical expectations.

The mixed formulation proposed here was motivated by the observation that design of optimal linear solvers for the C0IP formulation is not straightforward (with direct solvers used in [138]), coupled with the observed success of monolithic multigrid methods for a similar mixed discretization for the H^2-elliptic case of fourth-order operators in [66]. A natural step for future work is in extending these linear solvers to the mixed formulation proposed herein, in parallel to investigating effective linear solver strategies for the other discretizations. As this work is motivated by considering the more complex models in [138], coupling the smectic density to a director field or tensor-valued order parameter, the other natural direction for future research is to extend the analysis proposed herein to mixed formulations of the energy minimization problem associated with Equation (4.2).

Chapter 5

Efficient numerical simulation of smectic liquid crystals

Abstract[1]

Liquid crystalline materials are abundant in both the natural world (e.g., cholesterol and other molecules) and science and engineering practice (e.g., in liquid crystal displays). Because they possess properties that are intermediate between those of liquids and solid crystals, as well being electromagnetically active, there are a wide range of potential scientific and industrial uses for liquid crystals. However, their use in many contexts is held back by poor theoretical understanding of their mechanical properties. One approach to gaining such understanding is through the use of computer simulation and, in recent years, several families of finite-element methods have been proposed and developed to model various equilibrium states of different types of liquid crystals. Among common liquid crystal phases, smectic phases are distinguished by their "soap-like" properties, forming distinct layers at equilibrium. Until recently, there had been little success in developing finite-element simulation tools for smectic liquid crystals, primarily due to the complex nature of their governing free-energy functionals. In this paper, we discuss the challenges in developing such models, and build on recent work by Pevnyi, Selinger, and Sluckin [112] and by Xia et al. [138] to propose a new mixed finite-element formulation for one model of smectic A liquid

[1]This work to be submitted as "Efficient numerical simulation of smectic liquid crystals", by Patrick E. Farrell, Abdalaziz Hamdan, and Scott P. MacLachlan.

crystals. In particular, we demonstrate effective nonlinear and linear solvers for this formulation, combining nested iteration (grid continuation) and monolithic multigrid principles.

5.1 Introduction

The unusual properties of liquid crystals were first observed by the Austrian chemist Reinitzer in 1888 [116]. These substances display physical properties that are somehow "between" those expected of liquids and those seen in solid crystals, including the ability to maintain crystal-like molecular orientation while flowing like a liquid. Many of the interesting properties of liquid crystals can be tied to the development of symmetry-breaking structures and/or *defects* in the crystal structure [94], where non-smooth configurations are energetically preferable, particularly when constrained by a "mismatch" between the geometry of the domain and the inherent properties of the crystal under consideration (known as geometric frustration). Different materials achieve these properties in different ways, with some liquid crystals changing their behaviour with temperature (*thermotropic* liquid crystals) and some with chemical concentration (*lyotropic* liquid crystals). Among these types, there are further phases of the liquid crystals; in particular, thermotropic liquid crystals at high temperature behave as an isotropic liquid, then display a *nematic* phase as the temperature drops, then a *smectic* phase at lower temperatures, before acting as a convential crystal at sufficiently low temperatures.

The governing free-energy model for nematic liquid crystals behaves as a nonlinear and anisotropic second-order div-curl system, making it amenable to finite-element simulation with standard H^1 conforming spaces on convex domains [4,5,9,23,100,114]. Smectic liquid crystals, in contrast, exhibit much more complicated behaviour, including coupling between the liquid-crystal director field and a scalar order parameter related to the density variation of the liquid crystal, leading to free-energy functionals involving higher-order derivatives of the order parameter [1, 78, 112, 138]. The main focus of this paper is on the numerical modeling of equilibrium states of Smectic-A liquid crystals, which are characterized by their natural propensity to form layers with periodic variation in the density of the liquid crystal along lines orthogonal to

the orientation of the crystals. While some models make use of a complex order parameter as a model of the energy of liquid crystals [1, 55], several recent papers have proposed models based directly on the (real-valued) density variation [22, 112, 138]. For example, Pevnyi et al. [112] propose the energy functional

$$\mathscr{E}(u,\vec{\nu}) = \int_\Omega \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 + B\left|\nabla\nabla u + q^2\vec{\nu}\otimes\vec{\nu}u\right|^2 + \frac{K}{2}|\nabla\vec{\nu}|^2, \tag{5.1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, $u : \Omega \to \mathbb{R}$ represents the variation in the density of the liquid crystal from its average density, $\vec{\nu}$ is the unit-length director of the liquid crystal (the local axis of average molecular alignment), and a_1, a_2, a_3, q, K, and B are real valued constants determined by the liquid crystal under consideration. Of these, the smectic wavenumber, q, is notable because it prescribes a preferred wavelength for the solution of $2\pi/q$. Here, and in what follows, we use $|\boldsymbol{Q}|^2 = \boldsymbol{Q} : \boldsymbol{Q}$ to denote the Frobenius norm squared of tensor $\boldsymbol{Q}$ (of any rank), defined as the sum of squares of the entries in $\boldsymbol{Q}$ at a given point in Ω. While numerical experiments in [112] demonstrate that this model is capable of reproducing both the expected behaviour of Smectic-A liquid crystals and simulated results with complex-valued order parameters, the use of a vector-valued director degree of freedom limits the range of defects that can be represented, since $\vec{\nu}$ should be directionless (with no distinction between $\pm\vec{\nu}$ in the pointwise energy functional), but some natural structures cannot be represented with a continuous vector field, $\vec{\nu}$.

To overcome this limitation, Ball and Bedford [22] propose replacing the vector degrees of freedom for $\vec{\nu}$ with a tensor related to $\vec{\nu}\otimes\vec{\nu}$. In [138], Xia et al. adapted the Ball-and-Bedford model to achieve a more robust simulation framework, writing

$$\mathscr{J}_1(u,\boldsymbol{Q}) = \int_\Omega \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 + B\left|\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right|^2 + \frac{K}{2}|\nabla\boldsymbol{Q}|^2 + f_n(\boldsymbol{Q}), \tag{5.2}$$

where $\boldsymbol{Q}$ is a traceless tensor-valued order parameter, I_d is the identity matrix, and $f_n(\boldsymbol{Q}) = -l\,\mathrm{tr}(\boldsymbol{Q}^2) + l\left(\mathrm{tr}(Q^2)\right)^2$ for $d = 2$ and $f_n(\boldsymbol{Q}) = -\frac{l}{2}\mathrm{tr}(\boldsymbol{Q}^2) - \frac{l}{3}\mathrm{tr}(\boldsymbol{Q}^3) + \frac{l}{2}\left(\mathrm{tr}(\boldsymbol{Q}^2)\right)^2$ in three dimensions. Here, the functions $f_n(\boldsymbol{Q})$ and the penalty parameter, l, are chosen so that the minimizer of $\int_\Omega f_n(\boldsymbol{Q})$ is of the form $\boldsymbol{Q} = \vec{\nu}\otimes\vec{\nu} - \frac{I_d}{d}$, and are included in the energy to weakly enforce the rank-one condition implied by Pevnyi et al.'s model, without the potential singularity when including a scalar order parameter, as in [22] or the difficulty in representing certain expected defect structures

when directly discretizing (5.1). While there remain many open questions about the physical values of the constants a_1, a_2, a_3, q, K, and B, an important feature of the model is the energetic competition between the Hessian term, scaled by B, and the deformation of the director field, represented by either $\vec{\nu}$ or $\boldsymbol{Q}$. Compared to what has been done in Chapter 4, both u and the tensor $\boldsymbol{Q}$ are variables and, therefore, the Euler-Lagrange equations for either of these functionals naturally lead to a coupled system of PDEs, with a fourth-order operator acting on u and a second-order operator acting on $\vec{\nu}$ or $\boldsymbol{Q}$. We point out that the case of natural boundary conditions on the fourth-order operator and essential (Dirichlet) boundary condition on the second-order operator is the one of interest, precluding the possibility of making use of some discretization techniques for fourth-order problems that rely on clamped or simply supported boundary conditions. Several mixed finite-element techniques can be used to discretize (5.2), including conforming methods that require the use of H^2-conforming elements for u, and H^1-conforming elements for $\boldsymbol{Q}$. While this is possible using, for example, fifth-order Argyris elements on triangles in 2D, it is problematic in 3D, where the lowest-order conforming space on triangles is ninth-order [141]. This motivated the use of a C^0 interior-penalty (C0IP) approach in [138], where C^1-continuity is weakly enforced by penalizing inter-element jumps in the gradient of u, modifying the energy model (5.2) into

$$\hat{\mathscr{J}}_1(u, \boldsymbol{Q}) = \mathscr{J}_1(u, \boldsymbol{Q}) + \frac{1}{h_e^3} \sum_{e \in \epsilon_h} \int_e \left[\!\left[\frac{\partial u}{\partial n} \right]\!\right]^2, \tag{5.3}$$

where $[\![\cdot]\!]$ denotes the standard jump on each edge, and ϵ_h is the set of (interior) edges in the mesh. This technique enables the use of simple continuous (H^1-conforming) Lagrange elements, with only simple modification to the weak form required [42]. However, it is sometimes difficult to decide how large the penalty parameter on this term must be to achieve stability without harming convergence. Furthermore, developing fast solvers for the resulting systems is often difficult, and numerical experiments in [138] were limited to relatively coarse meshes, due to their reliance on direct solvers. Analysis of C0IP and conforming methods for (5.2) have been studied in [79, 137] with some simplifications. For the case of a fixed tensor, Q, and with $a_2 = a_3 = 0$, Argyris elements with a nonsymmetric version of Nitsche's method to impose Dirichlet BCs on u, as well as a nonsymmetric version of C0IP methods were discussed in [79]. The main motivation for using the nonsymmetric forms of these

discretizations is to get optimal-in-q convergence estimates, that yield optimal-in-h convergence rates in a weighted $H^2(\Omega)$ norm, but degraded h-convergence rates in the $L^2(\Omega)$ norm. A different simplification of (5.2) was considered in [137], assuming that $q = 0$ and only implementing the simply supported boundary conditions on u. A standard C0IP method was analysed with $\mathcal{O}(1/h^3)$ weights on the inter-element jumps in the first derivative of u. In addition, the weakly-over penalised symmetric interior penalty (WOPSIP) method, where the facet integrals arising from integration by parts are not included in the discrete forms, but higher weights appear on the inter-element jump of the first derivative, is shown numerically to be efficient. Numerical experments are also given in [137] for the coupled case($q \neq 0$).

An alternative approach to either conforming or C0IP discretizations for u is to introduce additional variables and use a mixed finite-element formulation for an augmented system [24, 25, 27, 51, 66, 79, 96, 97, 104]. In this paper, we adapt the mixed formulation from [79], which considered the minimization of $\mathscr{J}_1(u, \boldsymbol{Q}^*)$ for fixed $\boldsymbol{Q}^*$, by introducing an additional variable, $\vec{v}$, to represent ∇u, along with a Lagrange multiplier, $\vec{\alpha}$, to weakly enforce $\vec{v} = \nabla u$. The resulting modified smectic A energy functional is

$$\begin{aligned}\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q}) = \int_\Omega \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 + B\left|\nabla\vec{v} + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right|^2 \\ + \frac{K}{2}|\nabla\boldsymbol{Q}|^2 + f_n(\boldsymbol{Q}) + \int_\Omega \vec{\alpha}\cdot\vec{v} + u\nabla\cdot\vec{\alpha}. \qquad (5.4)\end{aligned}$$

In what follows, we show the expected relationship between minimizers of $\mathscr{J}_1(u, \boldsymbol{Q})$ and saddle points of $\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})$, as well as establishing that, under suitable "small data" assumptions, the Newton linearizations of $\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})$ are well-posed both in the continuum and when discretized appropriately. We also demonstrate that the resulting nonlinear systems are amenable to efficient solution using a Nested Iteration (NI) solver [10, 124] with Newton's method used to linearize on each grid, and monolithic multigrid used to solve the resulting linearizations.

The remainder of this paper is organized as follows. Section 5.2 presents background results on finite-element approximation needed for the later sections, as well as a review of existing theory for the smectic energy functional from [138]. The existence of minimizers of $\mathscr{J}_1$ and their equivalence to saddle points of $\mathscr{J}_2$ is presented

in Section 5.3. The linearization of $\mathscr{J}_2$ is considered in Section 5.4, establishing well-posedness (under typical assumptions) for both the continuum Hessian system and a particular mixed finite-element discretization. The details of the Nested Iteration-Newton-Krylov-Multigrid solver are presented in Section 5.5, followed by numerical results in Section 5.6. Concluding remarks and directions for future work are discussed in Section 5.7.

5.2 Background

5.2.1 Finite-element preliminaries

Throughtout this paper, we consider $\Omega \subset \mathbb{R}^d$, $d = \{2, 3\}$ to be an open, bounded and convex domain. For $d = 3$, we make the stricter assumption that Ω is a polyhedron. We take $\{\tau_h\}$, $0 < h < 1$, to be a quasiuniform family of triangulations of Ω. On a simplex, $T \in \tau_h$, we take $P_k(T)$ to be the space of multivariate polynomials of degree at most k. With this, the space of discontinuous Lagrange elements $DG_k(\Omega, \tau_h) \subset L^2(\Omega)$, $k \geq 0$ is defined as $DG(\Omega, \tau_h) = \{u_h \in L^2(\Omega),\ u_h \in P_k(T)\ \forall T \in \tau_h\}$. All degrees of freedom in this space are internal; i.e. functions belong to this space are piecewise continuous. In contrast, the continuous Lagrange elements, $CG_k(\Omega, \tau_h) \subset H^1(\Omega)$, $k \geq 1$, possess C^0 continuity across each element edges. The $H(\mathrm{div})$-conforming elements, $RT_k(\Omega, \tau_h)$, $k \geq 1$, where the normal components are continuous across element faces. In particular, for any $\vec{v}_h \in RT_k(\Omega, \tau_h)$, $\vec{v}_h|_T \in [P_{k-1}(T)]^d + P_{k-1}(T)\vec{x}$, $\forall T \in \tau_h$. We also consider Nédélec elements of the first kind, $\mathrm{N}_k(\Omega, \tau_h)$, $k \geq 1$ which are $H(\mathrm{curl})$-conforming elements, where the tangential component is continuous across element faces, and for any $\vec{v}_h \in \mathrm{N}_k(\Omega, \tau_h)$,$\vec{v}_h|_T \in [P_{k-1}(T)]^2 + \mathcal{S}_k(T)$, where $\mathcal{S}_k(T) = \left\{\vec{s} \in [P_k(T)]^d,\ \vec{s}(\vec{x}) \cdot \vec{x} = 0,\ \forall \vec{x} \in T\right\}$. Finally, we define the spaces $CG_k^\Gamma(\Omega, \tau_h)$ and $RT_k^\Gamma(\Omega, \tau_h)$ to be subspaces of $CG_k^\Gamma(\Omega, \tau_h)$ and $RT_k^\Gamma(\Omega, \tau_h)$, respectively, where

$$\begin{aligned} CG_k^\Gamma(\Omega, \tau_h) &= \{u_h \in CG_k(\Omega, \tau_h) |\ u_h = 0 \text{ on } \Gamma \subset \partial\Omega\}, \\ RT_k^\Gamma(\Omega, \tau_h) &= \{\vec{v}_h \in RT_k(\Omega, \tau_h) |\ \vec{v}_h \cdot \vec{n} = 0 \text{ on } \Gamma \subset \partial\Omega\}, \\ \mathrm{N}_k^\Gamma(\Omega, \tau_h) &= \{\vec{v}_h \in \mathrm{N}_k(\Omega, \tau_h) |\ \vec{v}_h \times \vec{n} = 0 \text{ on } \Gamma \subset \partial\Omega\}, \end{aligned}$$

where $\vec{n}$ is the outward unit normal to Γ. We recall standard approximation results for these spaces.

Theorem 20. *[32, 35, 92] Let $I^k_{1,h} : H^{k+1}(\Omega) \to DG_k(\Omega, \tau_h)$, $I^k_{2,h} : H^{k+1}(\Omega) \to CG_k(\Omega, \tau_h)$, $I^k_{3,h} : [H^{k+1}(\Omega)]^d \to RT_k(\Omega, \tau_h)$, and $I^k_{4,h} : [H^{k+1}(\Omega)]^d \to N_k(\Omega, \tau_h)$ be the finite-element interpolation operators. Then there exist a constant C, such that for any $u \in H^{k+1}(\Omega)$ and $\vec{v} \in \left[H^{k+1}(\Omega)\right]^d$,*

$$\begin{aligned}
\|u - I^k_{1,h}u\|_0 &\leq Ch^{k+1}|u|_{k+1}, \quad \forall k \geq 0, \\
\|u - I^k_{2,h}u\|_1 &\leq Ch^{k}|\vec{v}|_{k+1}, \quad \forall k > 0, \\
\|\vec{v} - I^k_{3,h}\vec{v}\|_{\text{div}} &\leq Ch^k \left(|\vec{v}|_k + |\vec{v}|_{k+1}\right), \quad \forall k > 0, \\
\|\vec{v} - I^k_{4,h}\vec{v}\|_{\text{curl}} &\leq Ch^k \left(|\vec{v}|_k + |\vec{v}|_{k+1}\right), \quad \forall k > 0,
\end{aligned}$$

Remark 5.2.1. In what follows, we use C to represent a generic positive constant that can depend on the domain, shape regularity of the triangulation, τ_h, and the polynomial degree of the finite-element space, but not on the mesh parameter, h, nor the smectic wavenumber, q, and may be different in different instances. Where needed, we will use $\left\{C_i\right\}$ to denote different arbitrary constants in the same expression.

Following [79], we make use of a stronger norm on $H_0(\text{div}, \Omega)$ induced by the Helmholtz decomposition. We shall use the spaces

$$\begin{aligned}
H^1_0(\Omega) &= \left\{u \in H^1(\Omega),\ u = 0 \text{ on } \partial\Omega\right\}, \\
H_0(\text{div}; \Omega) &= \left\{\vec{v} \in H(\text{div}; \Omega),\ \vec{v} \cdot \vec{n} = 0,\ \text{on } \partial\Omega\right\}, \\
H(\text{div}^0; \Omega) &= \left\{\vec{v} \in H(\text{div}; \Omega),\ \nabla \cdot \vec{v} = 0\right\}, \\
H_0(\text{curl}; \Omega) &= \left\{\vec{v} \in H(\text{curl}; \Omega),\ \vec{v} \times \vec{n} = \mathring{0},\ \text{on } \partial\Omega\right\},
\end{aligned}$$

where $\mathring{0} = 0$ in $2d$ and $\mathring{0}$ is the zero vector in $3d$.

Lemma 14. *(The Helmholtz decomposition [32, 74]) For $\vec{\alpha} \in H_0(div; \Omega)$, the following Helmholtz decomposition holds*

$$\vec{\alpha} = \nabla\phi + \nabla \times \mathring{p}, \tag{5.5}$$

where $\phi \in H^1(\Omega, \mathbb{R})$ is a zero-mean function, and $\mathring{p} = p \in H^1_0(\Omega, \mathbb{R})$ for $d = 2$, and $\mathring{p} = \vec{p} \in H_0(curl; \Omega) \cap H(div^0; \Omega)$ for $d = 3$. Furthermore, given $\vec{\alpha}_1$ and $\vec{\alpha}_2$ in $H(div; \Omega)$ with $\vec{\alpha}_1 = \nabla\phi_1 + \nabla \times \mathring{p}_1$, and $\vec{\alpha}_2 = \nabla\phi_2 + \nabla \times \mathring{p}_2$. The following defines an inner

product on $H(div;\Omega)$.

$$(\vec{\alpha}_1,\vec{\alpha}_2)_{Div} = q^{-4}\left(\int_\Omega \mathring{p}_1\mathring{p}_2 + \int_\Omega \nabla\phi_1\cdot\nabla\phi_2 + \int_\Omega \nabla\cdot\vec{\alpha}_1\nabla\cdot\vec{\alpha}_2\right), \tag{5.6}$$

where q *is a* $\mathcal{O}(1)$ *positive constant.*

We will also make use of this decomposition and norm for functions in $RT_{k+1}^{\partial\Omega}(\Omega,\tau_h)$.

Lemma 15. *[14, 15] The Helmholtz decomposition of* $RT_{k+1}^{\partial\Omega}(\Omega,\tau_h)$ *is*

$$RT_{k+1}(\Omega,\tau_h) = \left(\nabla_h^{\partial\Omega} DG_k(\Omega,\tau_h)\right) \oplus \left(\nabla\times V_h\right), \tag{5.7}$$

where $\nabla_h^{\Gamma_a}$ *is the discrete gradient operator,* $\nabla_h^{\Gamma_a} : DG_k(\Omega,\tau_h) \to RT_{k+1}^{\Gamma_a}(\Omega,\tau_h)$*, such that*

$$\int_\Omega \nabla_h^{\Gamma_a} u\cdot\vec{v} = -\int_\Omega u\nabla\cdot\vec{v}, \quad \forall\vec{v}\in RT_{k+1}^{\Gamma_a}(\Omega,\tau_h). \tag{5.8}$$

In two dimensions, we take $V_h = CG_{k+1}^{\partial\Omega}(\Omega,\tau_h)$*, while* $V_h = N_{k+1}^{\partial\Omega}(\Omega,\tau_h)$ *in three dimensions. This decomposition is orthogonal in the* L^2 *and* $H(div)$ *norms.*

5.2.2 Existing results

As proving well-posedness of the nonlinear systems arising from discretizing Problem (5.2) is either very complicated or requires strict restrictions on the constants $a_1,\ a_2,\ a_3,\ q,\ l,$ and B, different types of simplifications have been considered in the literature. In this section, we summarize existing results from [79, 137]. In [137], Xia and Farrell simplified (5.2) by assuming that $q = a_2 = 0$ and applying the simply-supported boundary conditions on u. In this case, minimizers of (5.2) should solve the two independent problems, with u determined by

$$2B\int_\Omega \nabla\nabla u : \nabla\nabla\phi + a_1 u\phi + a_3 u^3\phi = 0,\ \forall\phi\in H^2(\Omega)\cap H_0^1(\Omega). \tag{5.9}$$

A C0IP discretization on quadrilateral meshes in $\mathbb{R}^2$ and hexahedral meshes in $\mathbb{R}^3$ was proposed, with weak form to find $u\in CG_k^{\partial\Omega}(\Omega,\hat{\tau}_h)$ such that

$$A_h(u_h,\phi_h) + a_1\int_\Omega u_h\phi_h + a_3\int_\Omega u_h^3\phi_h = 0, \quad \forall\phi_h\in CG_k^{\partial\Omega}(\Omega,\hat{\tau}_h), \tag{5.10}$$

where

$$\begin{aligned} A_h(u_h,\phi_h) =& 2B\sum_{\tau\in\hat{\tau}_h}\nabla\nabla u_h:\nabla\nabla\phi_h - 2B\sum_{e\in\epsilon_h\backslash\partial\Omega}\int_e\left\{\!\!\left\{\vec{n}\cdot\left(\nabla\nabla u_h\right)\cdot\vec{n}\right\}\!\!\right\}\left[\!\left[\frac{\partial\phi_h}{\partial n}\right]\!\right] \\ & - 2B\sum_{e\in\epsilon_h\backslash\partial\Omega}\int_e\left\{\!\!\left\{\vec{n}\cdot\left(\nabla\nabla\phi_h\right)\cdot\vec{n}\right\}\!\!\right\}\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right] \\ & + \frac{2B\kappa}{h^3}\sum_{e\in\epsilon_h\backslash\partial\Omega}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]\left[\!\left[\frac{\partial\phi_h}{\partial n}\right]\!\right], \end{aligned}$$

where κ is a penalty parameter.

Theorem 21. *Let u be a regular isolated solution of Problem* (5.9)*, meaning there exists an $r>0$ such that there is only one solution within*

$$\left\{v\in H^2(\Omega)\cap H_0^1(\Omega),\ |u-v|_2\le r\right\},$$

then for sufficiently large κ, a_1, and a_3, there exists a unique solution $u_h\in CG_k^{\partial\Omega}(\Omega,\hat{\tau}_h)$ of the discrete problem (5.10) *within the ball*

$$\left\{v_h\in CG_k(\Omega,\hat{\tau}_h), |||I_{2,h}^k u - v_h|||_h\le R(h)\right\}.$$

Moreover, if $u\in H^p(\Omega)$, $p\ge 4$ then $\exists C>0$ such that

$$\|u-u_h\|_h\le Ch^{\min\{k-1,p-2\}}, \tag{5.11}$$

where $I_{2,h}^k$ is the continuous Lagrange interpolation operator defined in Theorem (20)*, and*

$$\|\phi_h\|_h = \sum_{\tau\in\hat{\tau}_h}|\phi_h|_{2,\tau}^2 + \frac{1}{h^3}\sum_{e\in\epsilon_h\backslash\partial\Omega}\int_e\left[\!\left[\frac{\partial\phi_h}{\partial n}\right]\!\right]^2. \tag{5.12}$$

Remark 5.2.2. Xia et al. in [137] experimentally validated a second discretization for (5.9) where

$$A_h(u_h,\phi_h) \;=\; 2B\sum_{\tau\in\hat{\tau}_h}\nabla\nabla u_h:\nabla\nabla\phi_h + \frac{2B\kappa}{h^3}\sum_{e\in\epsilon_h\backslash\partial\Omega}\int_e\left[\!\left[\frac{\partial u_h}{\partial n}\right]\!\right]\left[\!\left[\frac{\partial\phi_h}{\partial n}\right]\!\right].$$

In this case, A_h is missing the interior facet integrals arising from the integration by parts and its symmetrization. This discretization is inconsistent, as the solution u

of (5.9) fails to satisfy the discrete form. Numerical experiments in [137] show that optimal convergence rates can, nonetheless, be achieved if the derivative-jump term is suitable over-penalized. This is similar to the inconsistent discretization in (5.3) used to discretize (5.2) in [138].

When $q = 0$, the tensor-valued order parameter, $\boldsymbol{Q}$, is determined by

$$\begin{aligned} K\int_\Omega \nabla\boldsymbol{Q} : \nabla\boldsymbol{T} + 2l\left(2|\boldsymbol{Q}|^2 - 1\right)\boldsymbol{Q} : \boldsymbol{T} &= 0, \quad \forall \boldsymbol{T} \in H^1(\Omega), \text{ for } d = 2, \\ K\int_\Omega \nabla\boldsymbol{Q} : \nabla\boldsymbol{T} + l\left(-\boldsymbol{Q} - |\boldsymbol{Q}|^2 + |\boldsymbol{Q}|^2\boldsymbol{Q}\right) : \boldsymbol{T} &= 0, \quad \forall \boldsymbol{T} \in H^1(\Omega), \text{ for } d = 3. \end{aligned}$$

[137, Theorem 3.24] proves well-posedness and convergence results for the discretization of this problem using $H^1(\Omega)$-conforming continuous Lagrange elements.

In [79], a different type of simplification of (5.2) is applied. Here, Q is taken to be a given bounded tensor, $a_2 = a_3 = 0$, a forcing term f, and a mix of the boundary conditions (5.14)-(5.17) are applied. In this setting, it was shown that the minimizer is unique and should solve the problem

$$a(u, \phi) = \int_\Omega f\phi, \tag{5.13}$$

for all $\phi \in H_0^2(\Omega) = \{u \in H^2(\Omega),\ u = 0 \text{ on } \Gamma_0 \cup \Gamma_1 \text{ and } \nabla u = 0 \text{ on } \Gamma_1 \cup \Gamma_3\}$, and

$$\begin{aligned} a(u,\phi) &= B\int_\Omega \nabla\nabla u : \nabla\nabla\phi + Bq^2\int_\Omega \nabla\nabla u : \boldsymbol{Q}\phi + Bq^2\int_\Omega \nabla\nabla\phi : \boldsymbol{Q}u \\ &+ \int_\Omega (Bq^4\boldsymbol{Q} : \boldsymbol{Q} + a_1)u\phi. \end{aligned}$$

In this setting, $\partial\Omega$ is decomposed as $\partial\Omega = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ with $\Gamma_i \cap \Gamma_j = \emptyset$ for $i \neq j$, and the problem is posed with boundary conditions

$$u = 0, \quad (\nabla\nabla u + q^2\boldsymbol{Q}u)\cdot\vec{n} = \vec{0}, \quad \text{on } \Gamma_0, \tag{5.14}$$

$$u = 0, \quad \nabla u = \vec{0}, \quad \text{on } \Gamma_1, \tag{5.15}$$

$$\nabla\cdot(\nabla\nabla u + q^2\boldsymbol{Q}u)\cdot\vec{n} = 0, \quad (\nabla\nabla u + q^2\boldsymbol{Q}u)\cdot\vec{n} = \vec{0}, \quad \text{on } \Gamma_2, \tag{5.16}$$

$$\nabla\cdot(\nabla\nabla u + q^2\boldsymbol{Q}u)\cdot\vec{n} = \vec{0}, \quad \nabla u = \vec{0}, \quad \text{on } \Gamma_3. \tag{5.17}$$

A conforming finite-element discretization of (5.13) was shown to off optimal convergence; however, conforming discretizations of $H^2(\Omega)$ problems in three dimensions are prohibitively expensive, requiring ninth-order polynomials on tetrahedra. Two alternative formulations were provided in [79], a mixed formulation similar to that discussed below and a nonsymmetric C0IP method. The weak form in that case is to find $u_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h),\ k \geq 2$, such that

$$\tilde{a}_h(u_h, \phi_h) = \int_\Omega f\phi_h, \quad \forall \phi_h \in CG_k^{\Gamma_0 \cup \Gamma_1}(\Omega, \tau_h), \tag{5.18}$$

where

$$\begin{aligned}\tilde{a}(u_h, \phi_h) =& \hat{a}(u_h, \phi_h) - B \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla u_h + q^2 \boldsymbol{Q} u_h \right) \cdot \vec{n} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial \phi_h}{\partial n} \right]\!\!\right] \\ &+ B \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla \phi_h + q^2 \boldsymbol{Q} \phi_h \right) \cdot \vec{n} \right\}\!\!\right\} \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right] \\ &+ \frac{1}{q^3 h} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right] \left[\!\!\left[\frac{\partial \phi_h}{\partial n} \right]\!\!\right],\end{aligned}$$

and

$$\begin{aligned}\hat{a}(u_h, \phi_h) =& B \sum_{\tau \in \tau_h} \left(\int_\tau \nabla\nabla u_h : \nabla\nabla \phi_h + Bq^2 \int_\tau \nabla\nabla u_h : \boldsymbol{Q}\phi_h + Bq^2 \int_\tau \nabla\nabla \phi_h : \boldsymbol{Q} u_h \right) \\ &+ \int_\Omega B q^4 (\boldsymbol{Q} : \boldsymbol{Q} + a_1) u_h \phi_h.\end{aligned}$$

This problem is well-posed, and offers an optimal approximation in a similar weighted norm to that in (5.12).

Theorem 22. *Let $f \in L^2(\Omega)$ and $\boldsymbol{Q} : \boldsymbol{Q}$ be bounded pointwise on $\bar{\Omega}$. Then, Problem* (5.18) *is well-posed in the norm*

$$\begin{aligned}|||u_h|||_h^2 &= q^{-4} \left(\sum_{\tau \in \tau_h} |u_h|_{2,\tau}^2 + \|\nabla u_h\|_0^2 \right) + \|u_h\|_0^2 + \frac{1}{q^3 h} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left[\!\!\left[\frac{\partial u_h}{\partial n} \right]\!\!\right]^2 \\ &+ \frac{h}{q^5} \sum_{e \in \epsilon_h \setminus \Gamma_0 \cup \Gamma_2} \int_e \left\{\!\!\left\{ \vec{n} \cdot \left(\nabla\nabla u_h + q^2 \boldsymbol{T} u_h \right) \cdot \vec{n} \right\}\!\!\right\}^2.\end{aligned}$$

Moreover, if $u \in H^p(\Omega)$, $p \geq 4$, *is the unique solution of* (5.13), *then*

$$|||u - u_h|||_h \leq CBq^4 h^{\min\{p,k+1\}-2}|u|_p.$$

5.3 Equilibria of Energy Functionals

While the introduction of auxiliary variables to $\mathscr{J}_1(u, \boldsymbol{Q})$ in (5.2) to get $\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})$ in (5.4) is straightforward, it is natural to question what relation exists between minimizers of $\mathscr{J}_1$ and saddle points of $\mathscr{J}_2$. We consider this question in this section, but first state a slight generalization of [137, Theorem 2.1] where existence of a minimizer for $\mathscr{J}_1$ was proved over the space $\{(u, \boldsymbol{Q}) \in (H^2(\Omega, \mathbb{R}) \cap H_0^1(\Omega, \mathbb{R})) \times H_0^1(\Omega, S)\}$.

Corollary 14. *Given that the parameters* a_3, B, K, l, *and* q *are positive. The energy functional* $\mathscr{J}_1$ *has a minimizer over the admissible space*

$$\mathscr{A}_1 = \left\{(u, \boldsymbol{Q}) \in H^2(\Omega, \mathbb{R}) \times H_0^1(\Omega, S)\right\}, \tag{5.19}$$

where $H_0^1(\Omega, S)$ *is the space of* $d \times d$ *symmetric and traceless matrices with each component in* $H_0^1(\Omega, \mathbb{R})$, *giving* $\frac{d(d+1)}{2} - 1$ *degrees of freedom that can be represented as*

$$\boldsymbol{Q} = \begin{bmatrix} q_1 & q_2 \\ q_2 & -q_1 \end{bmatrix}, \quad \textit{in 2d, and } \boldsymbol{Q} = \begin{bmatrix} q_1 & q_3 & q_4 \\ q_3 & q_2 & q_5 \\ q_4 & q_5 & -(q_1+q_2) \end{bmatrix}, \quad \textit{in 3d.}$$

Proof. The proof is identical to [137, Theorem 2.1]. $\mathscr{J}_1$ is bounded from below as a_3 and l. $f_n(\boldsymbol{Q})$ is coercive in $H^1(\Omega, S)$ [137]. In addition, making use of the standard Poincaré inequality on $H^2(\Omega)$ [44], we have $\|\nabla u\|_0^2 \leq C\left(\|u\|_0^2 + \|\nabla\nabla u\|_0^2\right)$, $\forall u \in H^2(\Omega)$. □

Remark 5.3.1. The Euler-Lagrange equations for minimizers of $\mathscr{J}_1$ over the space

$\mathscr{A}_1$ are

$$\begin{aligned} 0 &= a_1 u + a_2 u^2 + a_3 u^3 + 2B\nabla\cdot\nabla\cdot\left[\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right] \\ &+ 2B\left[\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right] : \left[q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)\right], \\ \mathbf{0} &= 2B\left[\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{2}\right)u\right](q^2 u) + K\Delta\boldsymbol{Q} + \mathscr{F}_1, \end{aligned}$$

where

$$\mathscr{F}_1 = \begin{cases} -2l\boldsymbol{Q} + 4l\left|\boldsymbol{Q}\right|^2\boldsymbol{Q} & \text{for } d = 2, \\ -l\boldsymbol{Q} - l\boldsymbol{Q}^2 + 2l\left|\boldsymbol{Q}\right|^2\boldsymbol{Q} & \text{for } d = 3, \end{cases}$$

and $\mathbf{0} \in H_0^1(\Omega, S)$ is the d-dimensional zero tensor. In addition to the Dirichlet BCs on the tensor $\boldsymbol{Q}$, we consider the following natural BCs on u,

$$\nabla\cdot\left[\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right]\cdot\vec{n} = 0, \quad \text{and} \quad \left[\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right]\cdot\vec{n} = \vec{0}.$$

Now, we turn our attention to saddle points of $\mathscr{J}_2$ over the space

$$\mathscr{A}_2 = \left\{(u, \vec{v}, \boldsymbol{Q}, \vec{\alpha}) \in L^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S) \times H_0(\mathrm{div}; \Omega)\right\}.$$

The quadruple $(u, \vec{v}, \boldsymbol{Q}, \vec{\alpha})$ is a saddle point of $\mathscr{J}_2$ if it satisfies the first-order optimality conditions

$$\mathscr{J}_{2,u}[\phi] = \frac{\partial}{\partial u}\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})[\phi] = 0, \qquad \forall \phi \in L^2(\Omega, \mathbb{R}), \tag{5.20}$$

$$\mathscr{J}_{2,\vec{v}}[\vec{\psi}] = \frac{\partial}{\partial \vec{v}}\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})[\vec{\psi}] = 0, \qquad \forall \vec{\psi} \in H^1(\Omega, \mathbb{R}^d), \tag{5.21}$$

$$\mathscr{J}_{2,\boldsymbol{Q}}[\boldsymbol{T}] = \frac{\partial}{\partial \boldsymbol{Q}}\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})[\boldsymbol{T}] = 0, \qquad \forall \boldsymbol{T} \in H_0^1(\Omega, S). \tag{5.22}$$

$$\mathscr{J}_{2,\vec{\alpha}}[\vec{\beta}] = \frac{\partial}{\partial \vec{\alpha}}\mathscr{J}_2(u, \vec{v}, \vec{\alpha}, \boldsymbol{Q})[\vec{\beta}] = 0, \qquad \forall \vec{\beta} \in H_0(\mathrm{div}; \Omega), \tag{5.23}$$

The following variational system arises from these derivatives,

$$a\left((u,\vec{v},\boldsymbol{Q});(u,\vec{v},\boldsymbol{Q}),\left(\phi,\vec{\psi},\boldsymbol{T}\right)\right)+b\left(\vec{\alpha},\left(\phi,\vec{\psi},\boldsymbol{T}\right)\right) = 0,\forall\left(\phi,\vec{\psi},\boldsymbol{T}\right)\in V \tag{5.24}$$

$$b\left(\vec{\beta},(u,\vec{v},\boldsymbol{Q})\right) = 0, \qquad \forall\vec{\beta}\in H_0(\mathrm{div};\Omega) \tag{5.25}$$

,

where $V = L^2(\Omega,\mathbb{R})\times H^1(\Omega,\mathbb{R}^d)\times H_0^1(\Omega,S)$,

$$\begin{aligned}
a\left((u,\vec{v},\boldsymbol{Q});(u,\vec{v},\boldsymbol{Q}),\left(\phi,\vec{\psi},\boldsymbol{T}\right)\right) &= \int_\Omega \phi\left(a_1u+a_2u^2+a_3u^3\right)\\
&\quad + 2q^2B\int_\Omega\left[\nabla\vec{v}+q^2\left(\boldsymbol{Q}+\frac{I_d}{d}\right)u\right]:\left[\left(\boldsymbol{Q}+\frac{I_d}{d}\right)\phi\right]\\
&\quad + 2B\int_\Omega\left[\nabla\vec{v}+q^2\left(\boldsymbol{Q}+\frac{I_d}{d}\right)u\right]:\nabla\vec{\psi}\\
&\quad + 2q^2B\int_\Omega\left(\nabla\vec{v}+q^2\left(\boldsymbol{Q}+\frac{I_d}{d}\right)u\right):\boldsymbol{T}u\\
&\quad + K\int_\Omega\nabla\boldsymbol{Q}:\nabla\boldsymbol{T}+\int_\Omega\mathscr{F}_1:\boldsymbol{T},\\
b\left((u,\vec{v},\boldsymbol{Q}),\vec{\beta}\right) &= \int_\Omega\vec{\beta}\cdot\vec{v}+u\nabla\cdot\vec{\beta}.
\end{aligned}$$

In the next theorems, we show that minimizers of $\mathscr{J}_1$ and saddle points of $\mathscr{J}_2$ are equivalent.

Theorem 23. *Let $(u,\vec{v},\boldsymbol{Q},\vec{\alpha})\in\mathscr{A}_2$ be a saddle point of $\mathscr{J}_2$. Then, the pair $(u,\boldsymbol{Q})$ is also in $\mathscr{A}_1$ and is a minimizer of $\mathscr{J}_1$.*

Proof. Let $(u,\vec{v},\boldsymbol{Q},\vec{\alpha})\in\mathscr{A}_2$ be a saddle point of $\mathscr{J}_2$, so that

$$\int_\Omega\vec{\beta}\cdot\vec{v}+u\nabla\cdot\vec{\beta}=0,\quad\forall\vec{\beta}\in H_0(\mathrm{div};\Omega).$$

Let $g\in C_0^\infty(\Omega)$ be an arbitrary function and define $\vec{e}^{(i)}$ to be the d-dimensional canonical unit vector whose i^{th} entry is 1. Choosing $\vec{\beta}=g\vec{e}^{(i)}$ implies that $\vec{v}=\nabla u$ by the definition of the weak derivative [63]. Then, since $u\in L^2(\Omega)$ and $\nabla u\in H^1(\Omega,\mathbb{R}^d)$, it must be the case that $u\in H^2(\Omega)$. Moreover, since weak derivatives are unique, it

must be the case that

$$\begin{aligned}\mathrm{Ker}(b) &= \left\{ \left(\phi, \vec{\psi}, \boldsymbol{T}\right) \in V \,\middle|\, b\left(\vec{\beta}, \left(\phi, \vec{\psi}, \boldsymbol{T}\right)\right) = 0, \forall \vec{\beta} \in H_0^1(\mathrm{div}; \Omega) \right\} \\ &= \left\{ (\phi, \nabla\phi, \boldsymbol{T}) \in H^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S) \right\},\end{aligned}$$

where $V = L^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S)$. Equation (5.24) implies that

$$a\left((u, \nabla u, \boldsymbol{Q}); (u, \nabla u, \boldsymbol{Q}), (\phi, \nabla\phi, \boldsymbol{0})\right) = 0, \quad \forall \phi \in H^2(\Omega, \mathbb{R}) \tag{5.26}$$

$$a\left((u, \nabla u, \boldsymbol{Q}); (u, \nabla u, \boldsymbol{Q}), (0, 0, \boldsymbol{T})\right) = 0, \quad \forall \boldsymbol{T} \in H_0^1(\Omega, S), \tag{5.27}$$

which are the first-order optimality conditions of $\mathscr{J}_1$ over $\mathscr{A}_1$.Therefore, the pair $(u, \boldsymbol{Q})$ is a minimizer of $\mathscr{J}_1$. □

Theorem 24. *Let $(u, \boldsymbol{Q}) \in H^2(\Omega, \mathbb{R}) \times H_0^1(\Omega, S)$ be a minimizer of $\mathscr{J}_1$. Then, $(u, \nabla u, \boldsymbol{Q}, \vec{\alpha})$ is a saddle point of $\mathscr{J}_2$ for some $\vec{\alpha} \in H_0(div; \Omega)$.*

Proof. Let $(u, \boldsymbol{Q}) \in H^2(\Omega, \mathbb{R}) \times H_0^1(\Omega, S)$ be a minimizer of $\mathscr{J}_1$. Then, the first-order optimality conditions for $\mathscr{J}_1$ imply that

$$a\left((u, \nabla u, \boldsymbol{Q}); (u, \nabla u, \boldsymbol{Q}), (\phi, \nabla\phi, \boldsymbol{T})\right) = 0, \quad \forall (\phi, \nabla\phi, \boldsymbol{T}) \in \mathrm{Ker}(b)$$

Defining $\| (u, \vec{v}) \|_{0,q,1}^2 = \|u\|_0^2 + q^{-4}\|\vec{v}\|_1^2$, by [74, Theorem 1.4], if the inf-sup condition

$$I = \sup_{(u, \vec{v}, \boldsymbol{Q}) \in L^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S)} \frac{\int_\Omega \vec{\alpha} \cdot \vec{v} + \int_\Omega u \nabla \cdot \vec{\alpha}}{\sqrt{\| (u, \vec{v}) \|_{0,q,1} + \|\boldsymbol{Q}\|_1^2}} \geq C q^2 \|\vec{\alpha}\|_{\mathrm{Div}}, \tag{5.28}$$

is satisfied, then for every $(u, \nabla u, \boldsymbol{Q})$ that is a solution of (5.24) over $\mathrm{Ker}(b)$ there exists a unique $\vec{\alpha}$ such that $(u, \nabla u, \boldsymbol{Q}, \vec{\alpha})$ is a solution of (5.24)-(5.25). The choice $\boldsymbol{Q} = \boldsymbol{0}$ implies that the proof of the inf-sup condition for $d = 2$ is identical to the proof of the three-field formulation in [79].

We next prove the inf-sup condition (5.28) for $d = 3$. Given $\vec{\alpha} \in H_0(\mathrm{div}; \Omega)$, write $\vec{\alpha} = \nabla \times \vec{p} + \nabla\phi$, where $\phi \in H^1(\Omega, \mathbb{R})$ is a zero-mean function, and $\vec{p} \in H_0(\mathrm{curl}; \Omega) \cap H(\mathrm{div}^0; \Omega)$. Choose $\boldsymbol{Q} = \boldsymbol{0}$, and $u = c_1 (\nabla \cdot \vec{\alpha} - \phi)$, for a positive constant, c_1, to be chosen later. Note that

$$\|u\|_0 = c_1 \|\nabla \cdot \vec{\alpha} - \phi\|_0 \leq c_1 \left(\|\nabla \cdot \vec{\alpha}\|_0 + \|\phi\|_0\right), \text{ and } \int_\Omega u \nabla \cdot \vec{\alpha} = c_1 \left(\|\nabla \cdot \vec{\alpha}\|_0^2 + \|\nabla\phi\|_0^2\right),$$

with $\|\phi\|_0^2 \leq c_2\|\nabla\phi\|_0^2$ by the Poincaré inequality. By [73, Theorem 2.1], there exists $\vec{\psi} \in H^1(\Omega, \mathbb{R}^3)$ such that $\vec{p} = \nabla \times \vec{\psi}$, $\nabla \cdot \vec{\psi} = 0$, and $\vec{\psi} \cdot \vec{n} = 0$ on $\partial\Omega$. Furthermore, the following inequality holds [73, Theorem 2.3]

$$\|\vec{\psi}\|_1 \leq c_3\|\nabla \times \vec{\psi}\|_0^2 \leq c_3\|\vec{p}\|_0^2.$$

Taking $\vec{v} = \vec{\psi}$, we have

$$\begin{aligned} I &\geq \frac{\int_\Omega \left(\nabla \times \nabla \times \vec{\psi} + \nabla\phi\right) \cdot \vec{\psi} + c_1\|\nabla\phi\|_0^2 + c_1\|\nabla \cdot \vec{\alpha}\|_0^2}{\sqrt{c_3 q^{-4}\|\vec{p}\|_0^2 + c_1^2\|\nabla \cdot \vec{\alpha}\|_0^2 + c_1^2 c_2\|\nabla\phi\|_0^2}} \\ &\geq \frac{\|p\|_0^2 + \int_\Omega \nabla\phi \cdot \vec{\psi} + c_1\|\nabla\phi\|_0^2 + c_1\|\nabla \cdot \vec{\alpha}\|_0^2}{\sqrt{c_3 q^{-4}\|\vec{p}\|_0^2 + c_1^2\|\nabla \cdot \vec{\alpha}\|_0^2 + c_1^2 c_2\|\nabla\phi\|_0^2}}, \end{aligned}$$

where we use [32, Theorem 2.1.1] to establish

$$\int_\Omega \nabla \times \nabla \times \vec{\psi} \cdot \vec{\psi} = \|\nabla \times \vec{\psi}\|_0^2 + \int_{\partial\Omega} \vec{n} \times \left(\vec{\psi} \times \vec{n}\right) \cdot \left(\nabla \times \vec{\psi}\right) \times \vec{n} = \|\nabla \times \vec{\psi}\|_0^2 = \|\vec{p}\|_0^2,$$

as $\left(\nabla \times \vec{\psi}\right) \times \vec{n} = \vec{p} \times \vec{n} = 0$. We next bound $\int_\Omega \nabla\phi \cdot \vec{\psi}$,

$$\int_\Omega \nabla\phi \cdot \vec{\psi} \leq \frac{c_4}{2}\|\nabla\phi\|_0^2 + \frac{1}{2c_4}\|\vec{v}\|_0^2 \leq \frac{c_4}{2}\|\nabla\phi\|_0^2 + \frac{c_3}{2c_4}\|p\|_0^2.$$

Thus,

$$I \geq \frac{\left(1 - \frac{c_3}{2c_4}\right)\|p\|_0^2 + \left(c_1 - \frac{c_4}{2}\right)\|\nabla\phi\|_0^2 + c_1\|\nabla \cdot \vec{\alpha}\|_0^2}{\sqrt{c_3 q^{-4}\|\vec{p}\|_0^2 + c_1^2\|\nabla \cdot \vec{\alpha}\|_0^2 + c_1^2 c_2\|\nabla\phi\|_0^2}}.$$

If we choose $c_4 > \frac{c_3}{2}$ and $c_1 > \frac{c_4}{2}$. Then, there exits a constant c such that $I \geq cq^2\|\vec{\alpha}\|_{\text{Div}}$, where the $H(Div)$ norm of $\vec{\alpha}$ is defined in (5.6). □

Corollary 15. *If, in addition to the assumptions of Theorem 24, $(u, \boldsymbol{Q}) \in H^4(\Omega, \mathbb{R}) \times (H^2(\Omega, S) \cap H_0^1(\Omega, S))$ is a minimizer of $\mathscr{J}_1$ with $u\boldsymbol{Q} \in H^2(\Omega, S)$. Then, $(u, \vec{v}, \boldsymbol{Q}, \vec{\alpha})$ is a saddle point of $\mathscr{J}_2$ for*

$$\vec{v} = \nabla u, \tag{5.29}$$

$$\vec{\alpha} = 2B\nabla \cdot \left(\nabla\nabla u + q^2\left(\boldsymbol{Q} + \frac{I_d}{d}\right)u\right). \tag{5.30}$$

Proof. Multiplying Equations (5.29) and (5.30) by $\vec{\psi} \in H^1(\Omega, \mathbb{R}^d)$ and $\vec{\beta} \in H_0(\mathrm{div}; \Omega)$ respectively and integrating by parts leads to the fact that Equations (5.21) and (5.23) are satisfied. In addition, integrating Equations (5.26) and (5.27) by parts implies that

$$a_1 u + a_2 u^2 + a_3 u^3 + 2q^2 B \left(\nabla \vec{v} + q^2 \left(\boldsymbol{Q} + \frac{I_d}{d} \right) u \right) : \left(\boldsymbol{Q} + \frac{I_d}{d} \right) + \nabla \cdot \vec{\alpha} = 0 \quad (5.31)$$

$$2q^2 B u \left(\nabla \vec{v} + q^2 \left(\boldsymbol{Q} + \frac{I_d}{d} \right) u \right) - K \Delta \boldsymbol{Q} + \mathscr{F}_1 = 0. \quad (5.32)$$

Finally, multiplying Equations (5.31) and (5.32) by $\phi \in L^2(\Omega, \mathbb{R})$ and $\boldsymbol{T} \in H_0^1(\Omega, S)$ respectively, and integrating by parts implies that Equations (5.20) and (5.22) hold. □

5.4 Linearization

As the Euler-Lagrange equations for $\mathscr{J}_2$ are nonlinear, we use Newton's method to construct a sequence of approximations to equilibria points, where $\nabla \mathscr{J}_2 = \vec{0}$. Let $U = (u, \vec{v}, \boldsymbol{Q}, \vec{\alpha})$ be such an equilibrium point, and let $U_k = (u_k, \vec{v}_k, \boldsymbol{Q}_k, \vec{\alpha}_k)$ be an approximation to U that is, in some sense, close to U. Then, the first order Taylor expansion of $\nabla \mathscr{J}_2(U)$ is

$$\nabla \mathscr{J}_2(U) \approx \nabla \mathscr{J}_2(U_k) + H\left(\mathscr{J}_2(U_k)\right)(U - U_k),$$

where $H(\mathscr{J}_2)$ is the Hessian of $\mathscr{J}_2$. Rewriting this, we have

$$H\left(\mathscr{J}_2(U_k)\right)(U - U_k) \approx -\nabla \mathscr{J}_2(U_k), \quad (5.33)$$

Now, define $\delta u = u_{k+1} - u_k$, $\delta \vec{v} = \vec{v}_{k+1} - \vec{v}_k$, $\delta \boldsymbol{Q} = \boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k$, $\delta \vec{\alpha} = \vec{\alpha}_{k+1} - \vec{\alpha}_k$ and $\delta U = U_{k+1} - U_k = (\delta u, \delta \vec{v}, \delta \vec{\alpha}, \delta \boldsymbol{Q})$. Then, Newton's method finds a sequence of approximations to U by starting from some initial guess U_0 and successively solving

$$\begin{bmatrix} \mathscr{J}_{2,uu} & \mathscr{J}_{2,u\vec{v}} & \mathscr{J}_{2,u\boldsymbol{Q}} & \mathscr{J}_{2,u\vec{\alpha}} \\ \mathscr{J}_{2,\vec{v}u} & \mathscr{J}_{2,\vec{v}\vec{v}} & \mathscr{J}_{2,\vec{v}\boldsymbol{Q}} & \mathscr{J}_{2,\vec{v}\vec{\alpha}} \\ \mathscr{J}_{2,\boldsymbol{Q}u} & \mathscr{J}_{2,\boldsymbol{Q}\vec{v}} & \mathscr{J}_{2,\boldsymbol{Q}\boldsymbol{Q}} & \mathscr{J}_{2,\boldsymbol{Q}\vec{\alpha}} \\ \mathscr{J}_{2,\vec{\alpha}\vec{u}} & \mathscr{J}_{2,\vec{\alpha}\vec{v}} & \mathscr{J}_{2,\vec{\alpha}\boldsymbol{Q}} & \mathscr{J}_{2,\vec{\alpha}\vec{\alpha}} \end{bmatrix} \begin{bmatrix} \delta u \\ \delta \vec{v} \\ \delta \boldsymbol{Q} \\ \delta \vec{\alpha} \end{bmatrix} = -\nabla \mathscr{J}_2\left(U_k\right). \quad (5.34)$$

At each step, the Hessian is computed at $U_k = (u_k, \vec{v}_k, \boldsymbol{Q}_k, \vec{\alpha}_k)$. Here, the right-hand side is understood to be the variational terms given in (5.20) through (5.23), for every ϕ, $\vec{\psi}$, $\boldsymbol{T}$, $\vec{\beta}$, also evaluated at U_k, while the matrix-vector multiplication on the left denotes the directions in which the derivatives in the Hessian are taken [4]. For instance, $\mathscr{J}_{2,u\vec{v}}[\phi] \cdot \delta\vec{v} = \frac{\partial}{\partial \vec{v}} \left(\mathscr{J}_{2,u}(u_k, \vec{v}_k, \boldsymbol{Q}_k, \vec{\alpha}_k)[\phi] \right) [\delta\vec{v}]$. Using this formula, Components of the Hessian are given by

$$
\begin{aligned}
\mathscr{J}_{2,uu}[\phi] \cdot \delta u &= \int_\Omega \left(a_1 + 2a_2 u_k + 3a_3 u_k^2 + 2q^4 B \left| \boldsymbol{Q}_k + \frac{I_d}{d} \right|^2 \right) \phi \delta u, \\
\mathscr{J}_{2,u\vec{v}}[\phi] \cdot \delta\vec{v} &= 2q^2 B \int_\Omega \nabla\delta\vec{v} : \left(\boldsymbol{Q}_k + \frac{I_d}{d} \right) \phi, \\
\mathscr{J}_{2,u\boldsymbol{Q}}[\phi] \cdot \delta\boldsymbol{Q} &= 2q^2 B \int_\Omega \left(\nabla\vec{v}_k + 2q^2 \left(\boldsymbol{Q}_k + \frac{I_d}{d} \right) u_k \right) \phi \delta\boldsymbol{Q}, \\
\mathscr{J}_{2,u\vec{\alpha}}[\phi] \cdot \delta\vec{\alpha} &= \int_\Omega \phi \cdot \nabla\delta\vec{\alpha}, \\
\mathscr{J}_{2,\vec{v}\vec{v}}[\vec{\psi}] \cdot \delta\vec{v} &= 2B \int_\Omega \nabla\delta\vec{v} : \nabla\vec{\psi}, \quad \mathscr{J}_{2,\vec{v}\boldsymbol{Q}}[\vec{\psi}] \cdot \delta\boldsymbol{Q} = 2q^2 B \int_\Omega u_k \delta\boldsymbol{Q} : \nabla\vec{\psi}, \\
\mathscr{J}_{2,\boldsymbol{Q}\boldsymbol{Q}}[\boldsymbol{T}] \cdot \delta\boldsymbol{Q} &= 2q^2 B \int_\Omega q^2 u_k^2 \delta\boldsymbol{Q} : \boldsymbol{T} + K \nabla\delta\boldsymbol{Q} : \nabla\boldsymbol{T} + \int_\Omega \mathscr{F}_2 : \boldsymbol{T}, \\
\mathscr{J}_{2,\vec{v}\vec{\alpha}}[\vec{\psi}] \cdot \delta\vec{\alpha} &= \int_\Omega \delta\vec{\alpha} \cdot \vec{\psi},
\end{aligned}
$$

with $\mathscr{J}_{2,\boldsymbol{Q}\vec{\alpha}} = 0$ and $\mathscr{J}_{2,\vec{\alpha}\vec{\alpha}} = 0$. Here, we have

$$
\mathscr{F}_2 = \begin{cases} -2l\delta\boldsymbol{Q} + 8l \left(\boldsymbol{Q}_k : \delta\boldsymbol{Q} \right) \boldsymbol{Q}_k + 4l \left| \boldsymbol{Q}_k \right|^2 \delta\boldsymbol{Q} & \text{for } d = 2, \\ -l\delta\boldsymbol{Q} - 2l\boldsymbol{Q}_k\delta\boldsymbol{Q} + 4l \left(\boldsymbol{Q}_k : \delta\boldsymbol{Q} \right) \boldsymbol{Q}_k + 2l \left| \boldsymbol{Q}_k \right|^2 \delta\boldsymbol{Q} & \text{for } d = 3. \end{cases}
$$

As $\mathscr{J}_{2,\vec{\alpha}\vec{\alpha}} = 0$ and the Hessian matrix is symmetric, the system (5.34) can be rewritten in saddle-point form [29] to find $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in L^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S) \times H_0(\mathrm{div}; \Omega)$ such that,

$$
A\left((\delta u, \delta\vec{v}, \delta\boldsymbol{Q}), (\phi, \vec{\psi}, \boldsymbol{T}) \right) + b\left(\delta\vec{\alpha}, (\phi, \vec{\psi}, \boldsymbol{T}) \right) = F\left((\phi, \vec{\psi}, \boldsymbol{T}) \right) \quad \forall (\phi, \vec{\psi}, \boldsymbol{T}) \in V, \tag{5.35}
$$

$$
b\left(\vec{\beta}, (\delta u, \delta\vec{v}, \delta\boldsymbol{Q}) \right) = G\left(\vec{\beta} \right) \ \forall \vec{\beta} \in H_0(\mathrm{div}; \Omega), \tag{5.36}
$$

where $V = L^2(\Omega, \mathbb{R}) \times H^1(\Omega, \mathbb{R}^d) \times H_0^1(\Omega, S)$,

$$
\begin{aligned}
A\left((\delta u, \delta\vec{v}, \delta\boldsymbol{Q}), (\phi, \vec{\psi}, \boldsymbol{T})\right) = & \int_\Omega \left(a_1 + 2a_2 u_k + 3a_3 u_k^2 + 2q^4 B \left|\boldsymbol{Q}_k + \frac{I_d}{d}\right|^2\right) \phi\delta u \\
& + 2q^2 B \int_\Omega \nabla\delta\vec{v} : \left(\boldsymbol{Q}_k + \frac{I_d}{d}\right)\phi \\
& + 2q^2 B \int_\Omega \left(\nabla\vec{v}_k + 2q^2\left(\boldsymbol{Q}_k + \frac{I_d}{d}\right)u_k\right)\phi\delta\boldsymbol{Q} \\
& + 2q^2 B \int_\Omega \nabla\vec{\psi} : \left(\boldsymbol{Q}_k + \frac{I_d}{d}\right)\delta u + 2B\int_\Omega \nabla\delta\vec{v} : \nabla\vec{\psi} \\
& + 2q^2 B \int_\Omega u_k \delta\boldsymbol{Q} : \nabla\vec{\psi} \\
& + 2q^2 B \int_\Omega \left(\nabla\vec{v}_k + 2q^2\left(\boldsymbol{Q}_k + \frac{I_d}{d}\right)u_k\right) : \boldsymbol{T}\delta u \\
& + 2q^2 B \int_\Omega u_k \boldsymbol{T} : \nabla\delta\vec{v} \\
& + 2Bq^4 \int_\Omega u_k^2 \delta\boldsymbol{Q} : \boldsymbol{T} + K\int_\Omega \nabla\delta\boldsymbol{Q} : \nabla\boldsymbol{T} \\
& + \int_\Omega \mathscr{F}_2 : \boldsymbol{T}, \\
b\left(\vec{\beta}, (\delta u, \delta\vec{v}, \delta\boldsymbol{Q})\right) = & \int_\Omega \vec{\beta}\cdot\delta\vec{v} + \delta u \nabla\cdot\vec{\beta}, \\
F((\phi, \vec{\psi}, \boldsymbol{T})) = & - a\left((u_k, \vec{v}_k, \boldsymbol{Q}_k); (u_k, \vec{v}_k, \boldsymbol{Q}_k), \left(\phi, \vec{\psi}, \boldsymbol{T}\right)\right) \\
& - b\left(\vec{\alpha}_k, \left(\phi, \vec{\psi}, \boldsymbol{T}\right)\right), \\
G(\vec{\beta}) = & -b\left(\vec{\beta}, (u_k, \vec{v}_k, \boldsymbol{Q}_k)\right).
\end{aligned}
$$

Once the components $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha})$ are computed from solving (5.35)-(5.36), the current approximations $(u_k, \vec{v}_k, \boldsymbol{Q}_k, \vec{\alpha}_k)$ are updated, possibly with a step-length limited by a line-search or trust-region methodology.

Well-posedness of (5.35)-(5.36) can be proven following similar arguments to those used for just the density terms in [79], under assumptions on the problem parameters and that we linearize suitably close to a solution. We note, however, that the most natural assumption on problem parameters is to assume that pointwise values of $a_1 + 2a_2u_k + 3a_3u_k^2$ are suitably large and positive throughout the domain, and that this may not hold for "physical" values of the parameters a_1, a_2, a_3. Whether or not

the Newton linearizations are well-posed in a neighbourhood of a solution under more reasonable assumptions is a difficult question to answer, as the bilinear form A is expected to only be semi-definite, and not coercive; we leave investigation of this question for future work.

In the continuum, the necessary inf-sup condition was proven in [79] in two dimensions, while the proof in three-dimensions is given as the main part of the argument in the proof of Theorem 24. In two spatial dimensions, [79] also establishes a discrete inf-sup condition when taking $u_h \in DG_k(\Omega, \tau_h)$, $\vec{v}_h \in (CG_{k+2}(\Omega, \tau_h))^2$ and $\alpha_h \in RT^{\partial\Omega}_{k+1}(\Omega, \tau_h)$. In three spatial dimensions, this space for $\vec{v}_h$ does not appear to be rich enough to establish the inf-sup condition. Thus, we instead consider taking $u_h \in DG_k(\Omega, \tau_h)$, $\vec{v}_h \in (CG_{k+2}(\Omega, \tau_h) + B_{k+4}(\Omega, \tau_h))^3$ and $\alpha_h \in RT^{\partial\Omega}_{k+1}(\Omega, \tau_h)$, where $B_{k+4}(\Omega, \tau_h)$ is the standard "bubble" space of degree $k+4$ (including functions in $P_{k+4}(T)$ on each triangle $T \in \tau_h$, but having zero trace on ∂T). While we do not prove here that the inf-sup condition holds over this space, numerical results indicate no issues with stability. For notational simplicity, we write

$$\mathscr{V}_k(\Omega, \tau_h) = \begin{cases} (CG_{k+2}(\Omega, \tau_h))^2 & d = 2 \\ (CG_{k+2}(\Omega, \tau_h) + B_{k+4}(\Omega, \tau_h))^3 & d = 3 \end{cases}.$$

5.5 Nonlinear and linear solvers

At the core of our solver methodology is the use of Newton-Krylov-Multigrid methods. We use a standard Newton's method to solve our nonlinear systems, augmented with a secant line search using the ℓ^2-norm of the discretized nonlinear functional, $\nabla \mathscr{J}_2$; see [47], for example. In some instances, we damp the iteration by enforcing a maximum stepsize constraint that is less than 1, or using an initial step of less than 1 to compute the secant step. We typically use a stopping criterian also based on the ℓ^2 norm of the discretized nonlinear functional, requiring that either its absolute value be reduced below 10^{-8} or that it be reduced by the same factor times the initial value of nonlinear functional in the current nonlinear solver.

Even when using direct solvers, we find that many iterations of Newton's method can be needed for convergence, particularly when solving from poor initial guesses on fine computational grids. For this reason, we augment the Newton-Krylov-Multigrid

solution methodology with Nested Iteration (NI) [10, 124] (also known as *grid continuation*), where we first solve the nonlinear system to convergence on a coarse grid with a fixed initial guess, then interpolate this solution to use as the initial guess on a uniformly refined mesh, and repeat the procedure until we reach the desired finest-grid mesh for the simulation. While we could make use of variable solver tolerances on coarsest grids in our simulations, we find that the dominant time in our simulations is always the finest-grid solves (even when the coarsest-grid solves are notably inefficient), so do not pursue this here. In the numerical experiments in Section 5.6, we demonstrate that this NI-Newton-Krylov-Multigrid methodology vastly outperforms its finest-grid counterpart when not using the Nested Iteration methodology.

On the coarsest grids of our hierarchy, we use the sparse direct solver, MUMPS [11], as a direct solver for the Hessian system in (5.34) for each linearization. On finer-grids, however, we use preconditioned FGMRES [118, 119], with the ℓ^2 norm of the residual is below 10^{-8} in Tables 5.1 and 5.2 and the Eisenstat-Walker criteria in all other tables for determining linear solver stopping criteria [57] as a function of the convergence criteria and performance of the outer Newton iteration. We note that we use flexible GMRES for two reasons. First of all, as described below, we find we achieve the most robust performance when we use non-stationary relaxation schemes within our multigrid preconditioner. Secondly, even when using a stationary preconditioner, we find that the memory cost of extra vector storage needed for FGMRES is preferable to the computational cost of an extra application of the preconditioner needed for classical right-preconditioned GMRES for our problem.

The numerical results below are implemented using Firedrake [115] for the finite-element discretization and PETSc [20] for the nonlinear and linear solvers. This pairing is chosen because of the close integration between solvers and discretization in the two packages [93], particularly for the relaxation scheme used in the monolithic multigrid preconditioner described below, which is implemented using PCPatch [68].

5.5.1 Monolithic multigrid preconditioner

We now consider the development of effective linear solvers for the resulting discretized systems at each Newton step, for the Hessian system given in (5.34). Preliminary numerical results showed that directly applying a monolithic multigrid methodology to the linearizations given in (5.34) was somewhat unreliable, and that building a

preconditioner based on adding a multiple of the mass matrix for u, M_{uu}, to the $(1,1)$ block of the system is much more effective. We adopt this approach here, adding $100M_{uu}$ to the $(1,1)$ block and constructing a preconditioner based on this perturbed Hessian system.

We build the monolithic multigrid preconditioner using the same mesh hierarchy that we adopt for the nested iteration solver. That is, we have a fixed coarsest grid (on which we use MUMPS [11] as a direct solver), and each grid in the hiearchy is a uniform factor-two refinement of the next coarser grid. We use standard multigrid V- and W-cycles that, for nonlinear solves on any mesh in the nested iteration hierarchy iterate from that grid to the coarsest grid and back. We use standard finite-element interpolation operators on this hierarchy, partitioned based on the discretized fields, with matrix form

$$P = \begin{bmatrix} I_{1,h}^{k} & & & \\ & I_{2,h}^{k+1} & & \\ & & I_{3,h}^{k+1} & \\ & & & I_{4,h}^{k+1} \end{bmatrix},$$

where the blocks $I_{1,h}^{k}$, $I_{2,h}^{k}$, $I_{3,h}^{k}$, and $I_{4,h}^{k}$ are the natural finite-element interpolation operators for the $DG_k(\Omega,\tau_h)$, $\mathscr{V}_k(\Omega,\tau_h)$, $\boldsymbol{CG}_{k+1}(\Omega,\tau_h)$, and $RT_{k+1}(\Omega,\tau_h)$ spaces, respectively. Coarse-grid operators are formed by rediscretization.

As relaxation scheme, we make use of an additive overlapping Schwarz relaxation, which can be considered as a variant of the family of Vanka relaxation schemes originally proposed in [130] to solve the saddle-point systems that arise from the marker-and-cell (MAC) finite-difference discretization of the Navier-Stokes equations. Vanka relaxation methods encompass a variety of overlapping multiplicative or additive Schwarz methods applied to saddle-point problems, in which the subdomains are chosen so that the corresponding subsystems are also saddle-point systems. Vanka-type relaxation has been used extensively for finite-element discretizations [4, 6, 98]. Recently, a general-purpose implementation of patch-based relaxation schemes, including Vanka relaxation, was provided in [68], which we employ here.

Like other Schwarz methods, the relaxation used here can be understood algebraically. Denoting the set of all degrees of freedom in the problem by $\mathscr{L}$, we partition $\mathscr{L}$ into s overlapping subdomains or patches, $\mathscr{L} = \cup_{i=1}^{s}\mathscr{L}_i$, and consider the

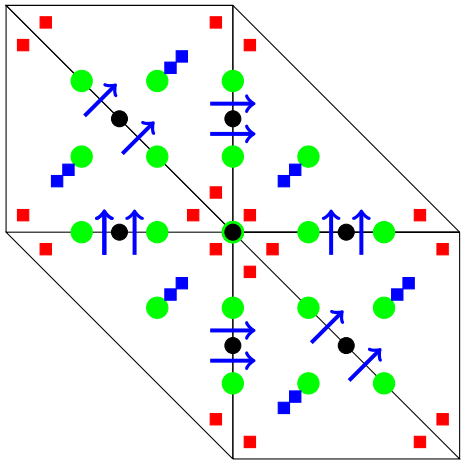

Figure 5.1: Star patches for $DG_1 - [CG_3]^2 - \boldsymbol{CG}_2 - RT_2$ discretizations. Red, green, black, and blue degrees of freedom denote DG_1, $[CG_3]^2$, $\boldsymbol{CG}_2$ and RT_2 degrees of freedom respectively.

stationary additive iteration with updates given by

$$x \leftarrow x + \sum_{i=1}^{s} R_i^T \mathcal{A}_{ii}^{-1} R_i (b - \mathcal{A}x),$$

where $\mathcal{A}x = b$ represents the Hessian linear system to be solved, R_i is the injection operator from a global vector, x, to a local vector, x_i, on $\mathscr{L}_i$ (with $R_i x = x_i$), and $\mathcal{A}_{ii} = R_i \mathcal{A} R_i^T$ is the restriction of the global system $\mathcal{A}$ to the degrees of freedom in $\mathscr{L}_i$. While inexact solution of the subdomain problems is relevant when the cardinality of $\mathscr{L}_i$ is large, we consider small subdomain sizes, where direct solution remains practical. We construct the patches, $\{\mathscr{L}_i\}$, topologically, as the so-called star patch around each vertex in the mesh (see [68]), taking all degrees of freedom at vertex i, on edges and faces adjacent to vertex i, and on all cells adjacent to vertex i to form $\mathscr{L}_i$. Figure 5.1 shows the subdomain construction around a typical vertex for $d = 2$ for the cases of discretization using $DG_1 \times [CG_3]^2 \times \boldsymbol{CG}_2 \times RT_2$ elements, where the red, green, black, and blue degrees of freedom denote DG_1, CG_3, $\boldsymbol{CG}_2$ and RT_2 degrees of freedom respectively, and each green/black circle represents 2 degrees of freedom (a vector). Rather than use the stationary iteration given above, we use three steps of GMRES preconditioned by the Schwarz method as the (pre- and post-) relaxation in the multigrid cycle on each level.

5.6 Numerical Results

In this section, we present numerical experiments to validate the mixed finite-element discretization that is used to discretize the energy in (5.4), and numerically demonstrate the efficiency of the NI solver with Newton's method to linearize on each grid, and monolithic multigrid to solve the resulting linearizations.

As a first example, we apply the NI-Newton-Krylov-multigrid scheme on the domain $\Omega = [0,1]^2$ with parameters $a_1 = -5,\ a_2 = 0,\ a_3 = 5,\ B = 10^{-5},\ K = 0.3,\ q = 40$, and $l = 30$. Dirichlet boundary conditions are imposed to match $\boldsymbol{Q} = \begin{bmatrix} \frac{x^2}{x^2+y^2+\epsilon} - \frac{1}{2} & 0 \\ 0 & \frac{y^2}{x^2+y^2+\epsilon} - \frac{1}{2} \end{bmatrix}$, where ϵ is a very small positive real number to avoid the singularity at $(0,0)$. We take the coarsest grid in the mesh hierarchy to be with $h = 1/32$, forming a uniform square mesh of size 32×32, then subdividing each square into two triangles by cutting from the top-left corner to the bottom-right corner. The initial guesses on the coarsest grid ($h = 1/32$) are taken to be $u = 1,\ \vec{v} = \vec{0},\ \vec{\alpha} = \vec{0}$ and $\boldsymbol{Q} = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$. For this experiment, we do not augment the linearizations by adding $100M_{uu}$ to the (1,1) block, showing that the solver can be effective without this augmentation (although we note that later results in this section are heavily dependent on this augmentation).

Table 5.1 presents Newton iteration counts, monolithic multigrid iteration counts (averaged over Newton iterations) using W(3,3) cycles, and wall-clock time to solution on each mesh (in minutes) for both the Newton-MG and Newton-LU solvers with varying numbers of processors, p, for a discretization with $(\delta u, \vec{v}, \delta\vec{\alpha}, \delta\boldsymbol{Q}) \in DG_2 \times [CG_4]^2 \times RT_3 \times \boldsymbol{CG}_3$. In these results, we see that the Newton-Krylov-MG solvers outperform Newton-LU when the mesh is fine enough for $p = 16$. In particular, when going from 4 to 16 processors, the parallel speedup for the Newton-Krylov-multigrid solver is 3.70x for the 512^2 mesh, while the speedup for Newton-LU is only 2.34x. Table 5.2 presents a comparison between Newton's method with nested iteration (using either MG or LU as the linear solver) and straight Newton-LU solvers on each mesh.

Preliminary results showed that the solver tested above was not efficient when changing the problem parameters and domain; therefore, for the remainder of this

Table 5.1: Newton iteration counts, averaged monolithic multigrid iteration counts using W(3,3) cycles, and wall-clock time on each mesh (in minutes) to convergence for NI-Newton-Krylov-MG and NI-Newton-LU solvers with varying numbers of processors, p, for $(\delta u, \delta\vec{v}, \delta\vec{\alpha}, \delta\boldsymbol{Q}) \in DG_2 \times [CG_4]^2 \times RT_3 \times \boldsymbol{CG}_3$.

h^{-1}	NI-Newton-Krylov-MG			NI-Newton-LU	
	Newton iter	MG iter	Time ($p = 4,\ 16$)	Newton iter	Time ($p = 4,\ 16$)
2^5	57	-	0.40, 0.28,	57	0.40, 0.28
2^6	2	8	0.52, 0.20	2	0.33, 0.13
2^7	2	5.5	2.01,0.66	2	1.62, 0.59
2^8	2	5	10.80, 2.79	2	8.55, 3.27
2^9	2	5	59.44, 16.06	2	49.12, 20.94

Table 5.2: Newton iteration counts, and total wall-clock time (in minutes) to convergence for Newton-LU solvers on each grid (with no nested iteration) using 16 processors, compared with the accumulated times for NI-Newton-LU and NI-Newton-Krylov-MG solvers, with $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_2 \times [CG_4]^2 \times \boldsymbol{CG}_3 \times RT_3$.

h^{-1}	Standard Newton-LU		NI-Newton-LU	NI-Newton-Krylov-MG[W(3,3)]
	Newton iter	Time	Time	Time
2^5	57	0.28	0.28	0.28
2^6	71	2.64	0.53	0.60
2^7	70	12.99	1.12	1.26
2^8	36	43.35	4.39	4.05
2^9	38	369.03	25.33	20.11

Table 5.3: Newton iteration counts, averaged monolithic multigrid V(3,3) iteration counts, and wall-clock time to convergence on each mesh (in minutes) for the NI-Newton-Krylov-MG and NI-Newton-LU solvers with varying numbers of processors, p, with the approximations $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_1 \times [CG_3] \times \boldsymbol{CG}_2 \times RT_2$. Results marked with a dash indicate where the solver was unsuccessful, due to memory requirements.

h^{-1}	NI-Newton-Krylov-MG			NI-Newton-LU	
	Newton iter	MG iter	Time ($p = 4,\ 16$)	Newton iter	Time ($p = 4,\ 16$)
2^5	50	-	1.81, 0.71	50	1.81, .71
2^6	4	3.00	1.11, 0.38	8	1.37, 0.54
2^7	4	3.50	4.79, 1.38	2	2.15, 0.94
2^8	4	2.75	18.47, 4.84	2	13.48, 5.19
2^9	4	3.5	98.67, 23.75	-	-

section, we use the solver described in Section 5.5, including the mass matrix augmentation to the (1,1) block. In the next example, we consider the so-called "oily streaks" scenario from [138], posed on the square domain, $\Omega = [-1, 1] \times [0, 2]$. Following [138], we use parameters $a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-5},\ K = 0.3,\ q = 30, l = 1$, with Dirichlet boundary conditions implemented on $\boldsymbol{Q}$ such that $\boldsymbol{Q} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}$ on $y = 0$, and $\boldsymbol{Q} = \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$ on $y = 2$, $x = -1$, and $x = 1$. The coarsest grid here is with $h = 1/32$, using a uniform 64×64 mesh, again with each square element cut into two triangles. As an initial guess on the coarsest grid, we take $u = \sin(\frac{q}{2}y),\ \vec{v} = \nabla u,\ \vec{\alpha} = 0$, and $\boldsymbol{Q} = \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$. Table 5.3 shows Newton iteration counts, averaged monolithic multigrid iteration counts using V(3,3) cycles, and wall-clock time on each mesh (in minutes) to convergence for the NI-Newton-MG and NI-Newton-LU solvers with varying numbers of processors, p. Here, the updates $(u, \vec{v}, \boldsymbol{Q}, \vec{\alpha})$ are approximated using $DG_1 \times [CG_3]^2 \times \boldsymbol{CG}_2 \times RT_2$. We note that here, and in later cases, the NI-Newton-LU algorithm fails on the finest grid, as the LU factorization requires more memory than is available on the workstation used in these tests.

For comparison, Table 5.4 shows the performance of the Newton-Krylov-multigrid method if the coarsest level is taken to be $h = 1/16$, with a 32×32 mesh, again using $p = 16$ processors. While it is common practice for elliptic problems to take a very coarse mesh, we see here that the indefinite shifts in (5.4) have an effect similar to

Table 5.4: Newton iteration counts, averaged V(3,3) monolithic multigrid iteration counts, and wall-clock time to convergence on each mesh (in minutes) for NI-Newton-MG solvers with $p = 16$ and coarsest level with $h = 1/16$, taking $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_1 \times [CG_3] \times \boldsymbol{CG}_2 \times RT_2$.

h^{-1}	Newton iter	MG iter	Time(p=16)
2^4	37	-	0.17
2^5	86	7.05	2.77
2^6	4	5.30	0.50
2^7	4	8.50	2.57
2^8	4	5.75	7.63
2^9	4	7.75	43.20

that observed for the Helmholtz equation [58], where taking a coarsest grid that fails to minimally resolve the natural wave behaviour in the system is counterproductive. In particular, what we notice in Table 5.4 is that we require many more Newton iterations to solve the problem with $h = 1/32$ when using the two-grid preconditioner than we did with the direct solver in Table 5.3, and that while the number of Newton steps is stable for finer grids, we require more multigrid iterations per Newton solve on the finer grids, leading to a substantially higher total time to solution. Thus, we emphasize that the grid resolution on the coarsest level of the hierarchy must be carefully chosen to optimize overall performance of the solvers. We note that, in these examples, we have used an approximation with $u \in DG_1$, in contrast to the use of $u \in DG_2$ in the first example (and in the example that follows) and, correspondingly, find that we can effectively use a coarser grid when we use a higher-order finite-element space for our solution.

In the third example, we again consider $\Omega = [-1, 1] \times [0, 2]$, now with $a_1 = -5,\ a_2 = 0, a_3 = 5,\ B = 10^{-5},\ k = 0.3,\ q = 40$, and $l = 30$. Dirichlet boundary on $\boldsymbol{Q}$ are imposed so that $\boldsymbol{Q} = \begin{bmatrix} \frac{x^2}{x^2+y^2+\epsilon} - \frac{1}{2} & \frac{xy}{\sqrt{x^2+y^2}+\epsilon} \\ \frac{xy}{x^2+y^2+\epsilon} & \frac{y^2}{x^2+y^2+\epsilon} - \frac{1}{2} \end{bmatrix}$ on all four edges of the square. For this example, we return to approximating the solution in $DG_2 \times [CG_4]^2 \times \boldsymbol{CG}_3 \times RT_3$, allowing a coarsest grid with $h = 1/16$ to be effectively used. As an initial guess on the coarsest grid, we take $u = \sin(5y),\ \vec{v} = \nabla u,\ \vec{\alpha} = 0$, and $\boldsymbol{Q} =$

Table 5.5: Newton iteration counts, averaged V(3,3) monolithic multigrid counts, and wall-clock time to convergence on each mesh (in minutes) for NI-Newton-Krylov-MG and NI-Newton-LU solvers with varying numbers of processors, p, taking $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_2 \times [CG_4]^2 \times \boldsymbol{CG}_3 \times RT_3$. Results marked with a dash indicate where the solver was unsuccessful, due to memory requirements.

h^{-1}	NI-Newton-Krylov-MG			NI-Newton-LU	
	Newton iter	MG iter	Time ($p = 4,\ 16$)	Newton iter	Time ($p = 4,\ 16$)
2^4	52	-	1.13, 0.50	52	1.13, 0.50
2^5	5	4.40	0.93, 0.32	22	1.98, 0.73
2^6	4	4.25	3.70, 1.17	3	1.67, 0.60
2^7	4	3.50	15.83, 4.42	2	5.56, 2.13
2^8	4	3.00	80.43, 23.01	-	-

$\begin{bmatrix} \frac{x^2}{x^2+y^2+\epsilon} - \frac{1}{2} & \frac{xy}{x^2+y^2+\epsilon} \\ \frac{xy}{x^2+y^2+\epsilon} & \frac{y^2}{x^2+y^2+\epsilon} - \frac{1}{2} \end{bmatrix}$. Table 5.5 compares the NI-Newton-LU and NI-Newton-Krylov-multigrid elapsed times and numbers of iterations to convergence using higher-order elements in comparison to Tables 5.3 and 5.4.

In the last example, we consider a three-dimensional problem on a unit cube domain, $\Omega = [0,1]^3$, with parameters $a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-3},\ K = 0.03,\ q = 30$, and $l = 1$. Dirichlet boundary conditions on $\boldsymbol{Q}$ are imposed, requiring $\boldsymbol{Q} = \begin{bmatrix} \frac{x^2}{x^2+y^2+\epsilon} & \frac{xy}{x^2+y^2+\epsilon} & 0 \\ \frac{xy}{x^2+y^2+\epsilon} & \frac{y^2}{x^2+y^2+\epsilon} & 0 \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}$ on the face $z = 0$, and $\boldsymbol{Q} = \begin{bmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} \end{bmatrix}$ on the face $z = 1$ with homogeneous Neumann boundary conditions on the other faces of the cube. We discretize the updates using the lowest-order elements suggested in Section 5.4, with $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_1 \times [CG_3 + B_5]^3 \times \boldsymbol{CG}_2 \times RT_1$, and take the coarsest level in the hierarchy to be at $h = 1/8$ (generated by taking a uniform $8 \times 8 \times 8$ hexahedral mesh of the unit cube, then cutting each hexahedron into 6 tetrahedra). The initial guess is taken to be $u = \sin(5z)$, and $\boldsymbol{Q} = \begin{bmatrix} \frac{x^2}{x^2+y^2+\epsilon} & \frac{xy}{x^2+y^2+\epsilon} & 0 \\ \frac{xy}{x^2+y^2+\epsilon} & \frac{y^2}{x^2+y^2+\epsilon} & 0 \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}$. A comparison between NI-Newton-Krylov-multigrid performance (using both V- and W-cycles) and NI-Newton-LU using $p = 16$ processors is presented in Table 5.6. We note that there is no difference between V- and W-cycles for a two-grid method, so only report differences in iterations and timings for the finest grid of this three-grid hierarchy.

Table 5.6: Newton iteration counts, averaged V(3,3) and W(3,3) monolithic multigrid counts, and wall-clock time to solution on each mesh (in minutes) for NI-Newton-Krylov-MG and NI-Newton-LU solvers with $p = 16$ processors. Here, the domain is $\Omega = [0,1]^3$, and we take $(\delta u, \delta\vec{v}, \delta\boldsymbol{Q}, \delta\vec{\alpha}) \in DG_1 \times [CG_3 + B_5]^3 \times \boldsymbol{CG}_2 \times RT_2$. Results marked with a dash indicate where the solver was unsuccessful, due to memory requirements.

h^{-1}	NI-Newton-Krylov-MG[V,W(3,3)]			NI-Newton-LU	
	Newton iter	MG iter(V, W)	Time (V, W)	Newton iter	Time ($p = 16$)
2^3	60	-	9.47	60	9.47
2^4	7	4.86	28.51	5	26.30
2^5	4	4.75, 3.50	133.76, 118.11	-	-

To give an indication of the solutions found, Figure 5.2 shows solutions for the three two-dimensional examples, visualizing only u. As expected, we see oscillatory solutions, whose structure is strongly determined by the behaviour of $\boldsymbol{Q}$, reflected particularly in the imposed boundary conditions.

5.7 Conclusions

Numerical simulation of liquid crystalline materials has been a focus of significant research in recent years, including recent advances in the simulation of smectic-A liquid crystals [22, 112, 138]. In this work, we build on the mathematical model introduced in [138], by introducing a mixed formulation using the gradient of the smectic order parameter explicitly and constraining its value using a Lagrange multiplier. We prove that, under some restrictions, equilibria of the two energies are equivalent, and we develop mixed finite-element methods to discretize our new formulation. Finally, we provide efficient nonlinear and linear solvers for the resulting nonlinear systems, using nested iteration to improve the efficiency of the nonlinear solution process, along with a Newton-Krylov-Multigrid solution strategy for each nonlinear problem. Numerical results demonstrate the efficiency of this approach.

In important step in future work is to combine the discretization and solver proposed here with deflation techniques [8, 61, 65], both for computing multiple solutions for fixed choices of the liquid crystal problem parameters and for computing bifurcation diagrams as the problem parameters and domain are varied, as was done in [138]. We also believe that the discretization and solver proposed herein could be extended

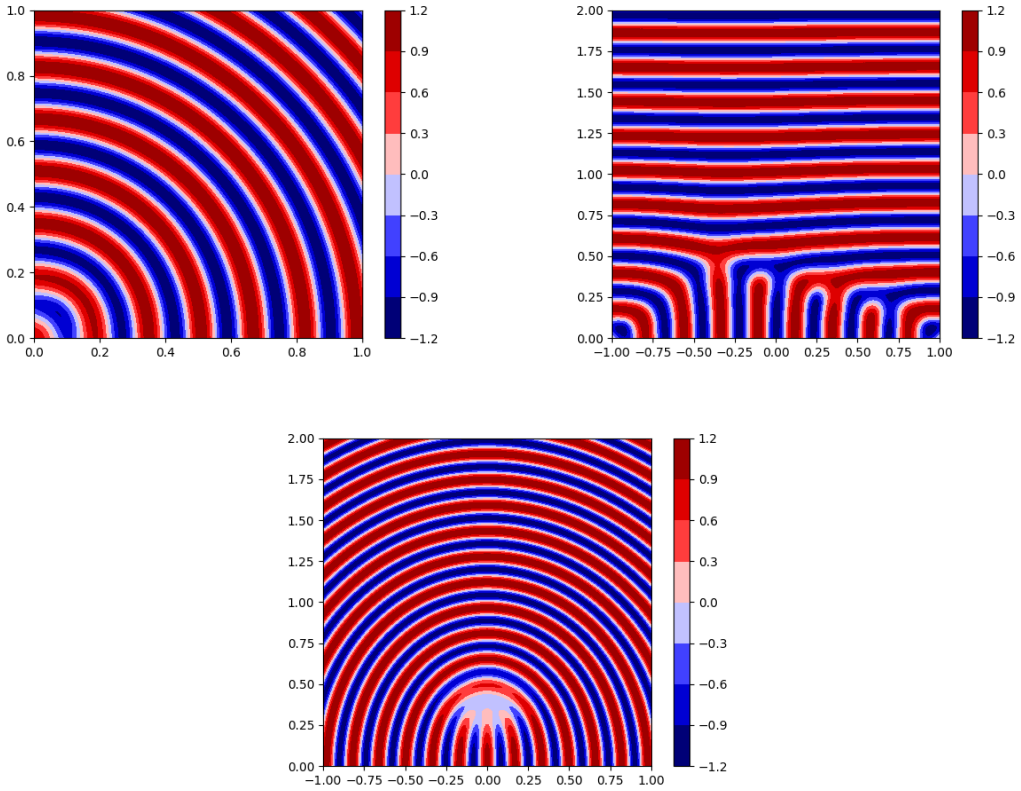

Figure 5.2: The variation in the density, u of the smectic A liquid crystals for the three two-dimensional examples (in order from top left) at $h = 1/128$.

for similar models of smectic-C liquid crystals, where only a single nonlinear term needs to be added to the energy in (5.2), as discussed in [136].

Chapter 6

Conclusions and future work

In this thesis, we develop, analyse, and apply finite-element methods for fourth-order PDEs similar to those that appear in the energy models of smectic-A liquid crystals. These include conforming methods, using Argyris elements with Nitsche-type penalty methods for essential boundary conditions, which are very expensive for three-dimensional problems because of the need to use very high-order elements. Therefore, we considered a nonsymmetric version of the C0IP method, to get an optimal convergence rate in terms of the physical problem parameter, q, and the mesh parameter, h. Preliminary computations showed that developing efficient preconditioners for the resulting systems is difficult. We address this by developing two new mixed finite-element formulations, based on introducing the gradient as an explicit variable constrained using a Lagrange multiplier. We prove error bounds for these methods and verify the analysis experimentally. Despite the fact that mixed finite-element formulations are sometimes not preferred because of the harder analysis of the resulting saddle-point problems compared to the conforming and C0IP methods and their high number of degrees of freedom, we find that they provide very nice properties, including the ability to easily implement all boundary conditions strongly, avoiding penalty terms that can worsen the condition numbers of the resulting systems. In addition, our mixed formulations offer the advantage of being able to construct efficient monolithic-multigrid preconditioners such as the ones given in Chapters 3 and 5. As shown in Chapter 5, we successfully use the mixed finite-element discretization and monolithic-multigrid preconditioner in an effective and efficient simulation tool for

smectic-A liquid crystals, building a nested iteration-Newton-Krylov-multigrid framework to efficiently solve the arising nonlinear systems.

There are several important directions for potential future research:

1. Testing the mixed finite-element formulations for less regular problems. Currently, we assume that the exact solution $u \in H^{k+5}(\Omega)$, where $k \geq 0$. However, there are cases where the solution to the biharmonic (and related equations) is only guaranteed to be in $H^{2+\gamma}(\Omega)$ for $0 < \gamma < 1/3$ (for L-shaped domains) [40]. Note that, for such a solution, we cannot interpret the Lagrange multiplier, $\vec{\alpha}$, as an approximation of the third derivative of u. It is not clear from the analysis in Chapter 3 whether the proposed method yields useful approximations with such minimal regularity of the solution to the continuum problem.

2. Generalizing the mixed finite-element formulations for general polyharmonic problems [72, Section 3.2]. Preliminary computational work showed that it is possible to generalize our mixed formulations to the sixth-order PDEs presented in [31], when the solution is sufficiently smooth. It is not clear how to extend them for less regular solutions.

3. Developing and analyzing alternative mixed finite-element discretizations for the full smectic-A model, including adapting our current Newton-Krylov-MG solvers for these discretizations. One such approach is to approximate the gradient, $\vec{v} = \nabla u$, using Raviart-Thomas or Brezzi-Douglas-Marini elements that are conforming for $H(div)$ but not H^1. As the full Hessian, $\nabla\vec{v} = \nabla\nabla u$ appears in the smectic-A model, we then must use element-wise gradients of $\vec{v}$ in our weak form and penalize inter-element jumps in the tangential derivatives (noting that the normal derivatives of any $H(div)$ function must be continuous, and that the RT and BDM elements are $H(div)$-conforming).

4. The methods proposed here should be able to be adapted to simulate the smectic-C models, where the continuum energy is similar to the one given in (1.3). One can distinguish the smectic-A and smectic-C phases by adding the term
$$\frac{e}{2}\int_\Omega |\boldsymbol{Q}\vec{v} \times \vec{v}|^2,$$
where $\vec{v} = \nabla u$ and e is a constant, to the smectic-A energy model [136]. In a

similar direction, it would be interesting to investigate adapting these models to consider phase change, such as the transition from smectic to nematic phases.

Bibliography

[1] N. M. Abukhdeir and A. D. Rey. Defect kinetics and dynamics of pattern coarsening in a two-dimensional smectic-a system. *New Journal of Physics*, 10(6):063025, jun 2008.

[2] D. J. Acheson. *Elementary fluid dynamics*. Oxford Applied Mathematics and Computing Science Series. The Clarendon Press, Oxford University Press, New York, 1990.

[3] J. Adler, H. D. Sterck, S. MacLachlan, and L. Olson. *Numerical Partial Differential Equations*. 2022. In preparation.

[4] J. H. Adler, T. J. Atherton, T. R. Benson, D. B. Emerson, and S. P. MacLachlan. Energy minimization for liquid crystal equilibrium with electric and flexoelectric effects. *SIAM J. Sci. Comput.*, 37(5):S157–S176, 2015.

[5] J. H. Adler, T. J. Atherton, D. B. Emerson, and S. P. MacLachlan. An energy-minimization finite-element approach for the Frank-Oseen model of nematic liquid crystals. *SIAM J. Numer. Anal.*, 53(5):2226–2254, 2015.

[6] J. H. Adler, T. Benson, E. C. Cyr, P. E. Farrell, S. MacLachlan, and R. Tuminaro. Monolithic multigrid for magnetohydrodynamics. *SIAM J. Sci. Comput.*, pages S70–S91, 2021.

[7] J. H. Adler, T. R. Benson, and S. P. MacLachlan. Preconditioning a mass-conserving discontinuous Galerkin discretization of the Stokes equations. *Numer. Linear Algebra Appl.*, 24(3):e2047, 23, 2017.

[8] J. H. Adler, D. B. Emerson, P. E. Farrell, and S. P. MacLachlan. A deflation technique for detecting multiple liquid crystal equilibrium states. *SIAM J. Sci. Comp.*, 39(1):B29–B52, 2017.

[9] J. H. Adler, D. B. Emerson, S. P. MacLachlan, and T. A. Manteuffel. Constrained optimization for liquid crystal equilibria. *SIAM J. Sci. Comput.*, 38(1):B50–B76, 2016.

[10] J. H. Adler, T. A. Manteuffel, S. F. McCormick, J. W. Ruge, and G. D. Sanders. Nested iteration and first-order system least squares for incompressible, resistive magnetohydrodynamics. *SIAM J. Sci. Comput.*, 32(3):1506–1526, 2010.

[11] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Anal. Appl.*, 23(1):15–41, 2001.

[12] D. Arnold, R. Falk, and R. Winther. Multigrid in $H(\mathrm{div})$ and $H(\mathrm{curl})$. *Numer. Math.*, 85(2):197–217, 2000.

[13] D. N. Arnold. Lecture notes on numerical analysis of partial differential equations. *University of Minnesota*, 2018.

[14] D. N. Arnold, R. S. Falk, and J. Gopalakrishnan. Mixed finite element approximation of the vector Laplacian with Dirichlet boundary conditions. *Math. Models Methods Appl. Sci.*, 22(9):1250024, 26, 2012.

[15] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning in $H(\mathrm{div})$ and applications. *Math. Comp.*, 66(219):957–984, 1997.

[16] J.-P. Aubin. *Approximation of elliptic boundary-value problems.* Wiley-Interscience [A division of John Wiley & Sons, Inc.], New York-London-Sydney, 1972. Pure and Applied Mathematics, Vol. XXVI.

[17] I. Babuška. The finite element method with penalty. *Math. Comp.*, 27:221–228, 1973.

[18] I. Babuška and M. Zlámal. Nonconforming elements in the finite element method with penalty. *SIAM J. Numer. Anal.*, 10:863–875, 1973.

[19] G. A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31(137):45–59, 1977.

[20] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkhout, W. Gropp, et al. PETSc users manual: Revision 3.10. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States), 2018.

[21] J. M. Ball. Mathematics and liquid crystals. *Mol. Cryst. Liq. Cryst.*, 647(1):1–27, 2017.

[22] J. M. Ball and S. J. Bedford. Discontinuous order parameters in liquid crystal theories. *Molecular Crystals and Liquid Crystals*, 612(1):1–23, 2015.

[23] J. M. Ball and A. Majumdar. Nematic liquid crystals: From maier-saupe to a continuum theory. *Molecular Crystals and Liquid Crystals*, 525(1):1–11, 2010.

[24] L. Banz, B. P. Lamichhane, and E. P. Stephan. A new three-field formulation of the biharmonic problem and its finite element discretization. *Numer. Methods Partial Differential Equations*, 33(1):199–217, 2017.

[25] L. Banz, J. Petsche, and A. Schröder. Two stabilized three-field formulations for the biharmonic problem. In *Chemnitz Fine Element Symposium*, pages 41–55. Springer, 2017.

[26] A. Bayliss, C. Goldstein, and E. Turkel. On accuracy conditions for the numerical computation of waves. *Journal of Computational Physics*, 59(3):396–404, 1985.

[27] E. M. Behrens and J. Guzmán. A mixed method for the biharmonic problem based on a system of first-order equations. *SIAM J. Numer. Anal.*, 49(2):789–817, 2011.

[28] J. Benzaken, J. A. Evans, S. F. McCormick, and R. Tamstorf. Nitsche's method for linear Kirchhoff–Love shells: Formulation, error analysis, and verification. *Comput. Methods Appl. Mech. Engrg.*, 374:113544, 2021.

[29] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.

[30] S. Bertoluzza, V. Chabannes, C. Prud'homme, and M. Szopos. Boundary conditions involving pressure for the Stokes problem and applications in computational hemodynamics. *Comput. Methods Appl. Mech. Engrg.*, 322:58–80, 2017.

[31] A. Bock and C. Cotter. A note on error analysis for a nonconforming discretisation of the tri-helmholtz equation with singular data, 2021.

[32] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2013.

[33] J. P. Borthagaray, R. H. Nochetto, and S. W. Walker. A structure-preserving FEM for the uniaxially constrained Q-tensor model of nematic liquid crystals. *Numer. Math.*, 145(4):837–881, 2020.

[34] J. P. Borthagaray and S. W. Walker. Chapter 5 - the q-tensor model with uniaxial constraint. In A. Bonito and R. H. Nochetto, editors, *Geometric Partial Differential Equations - Part II*, volume 22 of *Handbook of Numerical Analysis*, pages 313–382. Elsevier, 2021.

[35] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[36] D. Braess and R. Sarazin. An efficient smoother for the stokes problem. *Applied Numerical Mathematics*, 23(1):3–19, 1997. Multilevel Methods.

[37] A. Brandt and N. Dinar. Multigrid solutions to elliptic flow problems. In S. Parter, editor, *Numerical Methods for Partial Differential Equations*, pages 53–147. Academic Press, New York, 1979.

[38] S. C. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[39] S. C. Brenner. C^0 interior penalty methods. In *Frontiers in numerical analysis—Durham 2010*, volume 85 of *Lect. Notes Comput. Sci. Eng.*, pages 79–147. Springer, Heidelberg, 2012.

[40] S. C. Brenner, S. Gu, T. Gudi, and L.-y. Sung. A quadratic C^0 interior penalty method for linear fourth order boundary value problems with boundary conditions of the Cahn-Hilliard type. *SIAM J. Numer. Anal.*, 50(4):2088–2110, 2012.

[41] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[42] S. C. Brenner and L.-Y. Sung. C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. *J. Sci. Comput.*, 22/23:83–118, 2005.

[43] S. C. Brenner, L.-y. Sung, and Y. Zhang. C^0 interior penalty methods for an elliptic state-constrained optimal control problem with Neumann boundary condition. *J. Comput. Appl. Math.*, 350:212–232, 2019.

[44] S. C. Brenner, K. Wang, and J. Zhao. Poincaré-Friedrichs inequalities for piecewise H^2 functions. *Numer. Funct. Anal. Optim.*, 25(5-6):463–478, 2004.

[45] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[46] W. Briggs, V. Henson, and S. McCormick. *A Multigrid Tutorial, 2nd Edition*. 01 2000.

[47] P. R. Brune, M. G. Knepley, B. F. Smith, and X. Tu. Composing scalable nonlinear algebraic solvers. *SIAM Review*, 57(4):535–565, 2015.

[48] E. Burman. A penalty-free nonsymmetric nitsche-type method for the weak imposition of boundary conditions. *SIAM J. Numer. Anal.*, 50(4):1959–1981, 2012.

[49] S. Chandrasekhar. *Liquid Crystals*. Cambridge University Press, 2 edition, 1992.

[50] L. Chen and X. Huang. Decoupling of mixed methods based on generalized Helmholtz decompositions. *SIAM J. Numer. Anal.*, 56(5):2796–2825, 2018.

[51] X.-l. Cheng, W. Han, and H.-c. Huang. Some mixed finite element methods for biharmonic equation. *J. Comput. Appl. Math.*, 126(1-2):91–109, 2000.

[52] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[53] T. A. Davis and E. C. Gartland, Jr. Finite element analysis of the Landau-de Gennes minimization problem for liquid crystals. *SIAM J. Numer. Anal.*, 35(1):336–362, 1998.

[54] P. de and J. Prost. *The Physics of Liquid Crystal*, volume 2. 01 1993.

[55] P. G. de Gennes. An analogy between superconductors and smectic A. *Solid State Commun.*, 10:753–756, 1972.

[56] D. S. Dugdale. *Elements of Elasticity: The Commonwealth and International Library: Structures and Solid Body Mechanics Division*. Elsevier, 2014.

[57] S. C. Eisenstat and H. F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996.

[58] H. C. Elman, O. G. Ernst, and D. P. O'Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM J. Sci. Comput.*, 23(4):1291–1315, 2001.

[59] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005.

[60] A. Embar, J. Dolbow, and I. Harari. Imposing Dirichlet boundary conditions with Nitsche's method and spline-based finite elements. *Internat. J. Numer. Methods Engrg.*, 83(7):877–898, 2010.

[61] D. B. Emerson, P. E. Farrell, J. H. Adler, S. P. MacLachlan, and T. J. Atherton. Computing equilibrium states of cholesteric liquid crystals in elliptical channels with deflation algorithms. *Liquid Crystals*, 45(3):341–350, 2018.

[62] G. Engel, K. Garikipati, T. J. R. Hughes, M. G. Larson, L. Mazzei, and R. L. Taylor. Continuous/discontinuous finite element approximations of fourth-order elliptic problems in structural and continuum mechanics with applications to

thin beams and plates, and strain gradient elasticity. *Comput. Methods Appl. Mech. Engrg.*, 191(34):3669–3750, 2002.

[63] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[64] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[65] P. E. Farrell, Ã. Birkisson, and S. W. Funke. Deflation techniques for finding distinct solutions of nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 37(4):A2026–A2045, 2015.

[66] P. E. Farrell, A. Hamdan, and S. P. MacLachlan. A new mixed finite-element method for H^2 elliptic problems, 2021.

[67] P. E. Farrell, Y. He, and S. P. MacLachlan. A local Fourier analysis of additive Vanka relaxation for the Stokes equations. *Numer. Linear Algebra Appl.*, 28(3):Paper No. e2306, 28, 2021.

[68] P. E. Farrell, M. G. Knepley, F. Wechsung, and L. Mitchell. PCPATCH: software for the topological construction of multigrid relaxation methods. *CoRR*, abs/1912.08516, 2019.

[69] P. E. Farrell, L. Mitchell, and F. Wechsung. An Augmented Lagrangian Preconditioner for the 3D Stationary Incompressible Navier–Stokes Equations at High Reynolds Number. *SIAM J. Sci. Comput.*, 41(5):A3073–A3096, 2019.

[70] J. Freund and R. Stenberg. On weakly imposed boundary conditions for second order problems. *Proceedings of the Ninth International Conference on Finite Elements in Fluids*, pages 327–336, 01 1995.

[71] Friedel, G. Les états mésomorphes de la matière. *Ann. Phys.*, 9(18):273–474, 1922.

[72] D. Gallistl. Stable splitting of polyharmonic operators by generalized Stokes systems. *Math. Comp.*, 86(308):2555–2577, 2017.

[73] V. Girault. Incompressible finite element methods for Navier-Stokes equations with nonstandard boundary conditions in $\mathbf{R}^3$. *Math. Comp.*, 51(183):55–74, 1988.

[74] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[75] J. Gopalakrishnan and W. Qiu. Partial expansion of a Lipschitz domain and some applications. *Front. Math. China*, 7(2):249–272, 2012.

[76] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[77] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[78] F. Guillén-González and G. Tierra. Approximation of Smectic-A liquid crystals. *Computer Methods in Applied Mechanics and Engineering*, 290:342–361, 2015.

[79] A. Hamdan, P. Farrell, and S. MacLachlan. Finite-element discretization of the smectic density equation. 2022. In preparation.

[80] Y. He and S. P. MacLachlan. Local Fourier analysis of block-structured multigrid relaxation schemes for the Stokes equations. *Numer. Linear Algebra Appl.*, 25(3):e2147, 28, 2018.

[81] R. Hiptmair and C. Pechstein. A review of regular decompositions of vector fields: continuous, discrete, and structure-preserving. In *Spectral and high order methods for partial differential equations—ICOSAHOM 2018*, volume 134 of *Lect. Notes Comput. Sci. Eng.*, pages 45–60. Springer, Cham, [2020] ©2020.

[82] J. Huang, X. Huang, and Y. Xu. Convergence of an adaptive mixed finite element method for Kirchhoff plate bending problems. *SIAM J. Numer. Anal.*, 49(2):574–607, 2011.

[83] P. Hénon, P. Ramet, and J. Roman. PaStiX: a high-performance parallel direct solver for sparse symmetric positive definite systems. *Parallel Comput.*, 28(2):301–321, 2002.

[84] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM. *Computers and Mathematics with Applications*, 30(9):9–37, 1995.

[85] D. III, P. Keller, J. Naciri, R. Pink, H. Jeon, D. Shenoy, and B. Ratna. Liquid crystal elastomers with mechanical properties of a muscle. *Macromolecules*, 34, 07 2001.

[86] M. Ilyas and B. Lamichhane. A three-field formulation of the poisson problem with nitsche approach. *ANZIAM Journal*, 59, 11 2017.

[87] C. Johnson and J. Pitkäranta. Analysis of some mixed finite element methods related to reduced integration. *Math. Comp.*, 38(158):375–400, 1982.

[88] M. Juntunen and R. Stenberg. Nitsche's method for general boundary conditions. *Math. Comp.*, 78(267):1353–1374, 2009.

[89] Y. H. Kim, D. K. Yoon, H. S. Jeong, O. D. Lavrentovich, and H.-T. Jung. Smectic liquid crystal defects for self-assembling of building blocks and their lithographic applications. *Advanced Functional Materials*, 21(4):610–627, 2011.

[90] R. Kirby and L. Mitchell. Code generation for generally mapped finite elements. *ACM Transactions on Mathematical Software*, 45:1–23, 12 2019.

[91] R. C. Kirby. From functional analysis to iterative methods. *SIAM Rev.*, 52(2):269–293, 2010.

[92] R. C. Kirby, A. Logg, M. E. Rognes, and A. R. Terrel. Common and unusual finite elements. In *Automated Solution of Differential Equations by the Finite Element Method*, pages 95–119. Springer, 2012.

[93] R. C. Kirby and L. Mitchell. Solver composition across the PDE/linear algebra barrier. *SIAM Journal on Scientific Computing*, 40(1):C76–C98, 2018.

[94] M. Kleman. Defects in liquid crystals. *Reports on Progress in Physics*, 52(5):555–654, may 1989.

[95] W. Krendl, K. Rafetseder, and W. Zulehner. A decomposition result for biharmonic problems and the Hellan-Herrmann-Johnson method. *Electron. Trans. Numer. Anal.*, 45:257–282, 2016.

[96] B. P. Lamichhane. A stabilized mixed finite element method for the biharmonic equation based on biorthogonal systems. *J. Comput. Appl. Math.*, 235(17):5188–5197, 2011.

[97] Z. Li and S. Zhang. A stable mixed element method for the biharmonic equation with first-order function spaces. *Comput. Methods Appl. Math.*, 17(4):601–616, 2017.

[98] S. P. MacLachlan and C. W. Oosterlee. Local Fourier analysis for multigrid with overlapping smoothers applied to systems of PDEs. *Numer. Linear Algebra Appl.*, 18(4):751–774, 2011.

[99] J. F. Maitre, F. Musy, and P. Nignon. A fast solver for the Stokes equations using multigrid with a UZAWA smoother. In D. Braess, W. Hackbusch, and U. Trottenberg, editors, *Advances in Multi–Grid Methods*, volume 11 of *Notes on Numerical Fluid Mechanics*, pages 77–83, Braunschweig, 1984. Vieweg.

[100] A. Majumdar. Equilibrium order parameters of nematic liquid crystals in the Landau-de Gennes theory. *European J. Appl. Math.*, 21(2):181–203, 2010.

[101] A. Majumdar and A. Zarnescu. Landau-De Gennes theory of nematic liquid crystals: the Oseen-Frank limit and beyond. *Arch. Ration. Mech. Anal.*, 196(1):227–280, 2010.

[102] D. S. Malkus and T. J. Hughes. Mixed finite element methods, reduced and selective integration techniques: a unification of concepts. *Comput. Meth. Appl. Mech. Eng*, 15(1):63–81, 1978.

[103] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Lin. Alg. Appl.*, 18(1):1–40, 2011.

[104] P. Monk. A mixed finite element method for the biharmonic equation. *SIAM J. Numer. Anal.*, 24(4):737–749, 1987.

[105] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972, 2000.

[106] A. Natale, J. Shipton, and C. J. Cotter. Compatible finite element spaces for geophysical fluid dynamics. *Dynamics and Statistics of the Climate System*, 1(1), 11 2016. dzw005.

[107] J. Nečas. *Les méthodes directes en théorie des équations elliptiques*. Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague, 1967.

[108] T. D. Nguyen. Discontinious galerkin formulations for thin bending problems. 2008.

[109] J. Nitsche. Über ein variationsprinzip zur lösung von dirichlet-problemen bei verwendung von teilräumen, die keinen randbedingungen unterworfen sind. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 36(1):9–15, jul 1971.

[110] R. H. Nochetto, S. W. Walker, and W. Zhang. The ericksen model of liquid crystals with colloidal and electric effects. *Journal of Computational Physics*, 352:568–601, jan 2018.

[111] J. Peiró and S. Sherwin. Finite difference, finite element and finite volume methods for partial differential equations. In S. Yip, editor, *Handbook of Materials Modeling: Methods*, pages 2415–2446. Springer Netherlands, Dordrecht, 2005.

[112] M. Y. Pevnyi, J. V. Selinger, and T. J. Sluckin. Modeling smectic layers in confined geometries: Order parameter and defects. *Physical Review E*, 90(3):032507, 2014.

[113] K. Rafetseder and W. Zulehner. A decomposition result for Kirchhoff plate bending problems and a new discretization approach. *SIAM J. Numer. Anal.*, 56(3):1961–1986, 2018.

[114] A. Ramage and E. C. Gartland, Jr. A preconditioned nullspace method for liquid crystal director modeling. *SIAM J. Sci. Comput.*, 35(1):B226–B247, 2013.

[115] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. McRae, G.-T. Bercea, G. R. Markall, and P. H. Kelly. Firedrake: automating the finite element method by composing abstractions. *ACM Transactions on Mathematical Software (TOMS)*, 43(3):24, 2017.

[116] F. Reinitzer. Beiträge zur kenntniss des cholesterins. *Monatshefte für Chemie*, 9(1):421–441, 1888.

[117] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, volume 35 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.

[118] Y. Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM J. Sci. Comput.*, 14(2):461–469, 1993.

[119] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[120] M. Schedensack. *A class of mixed finite element methods based on the Helmholtz decomposition in computational mechanics*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2015.

[121] B. Seibold. Minimal positive stencils in meshfree finite difference methods for the poisson equation. *Computer Methods in Applied Mechanics and Engineering*, 198(3):592–601, 2008.

[122] P. Solin, K. Segeth, and I. Dolezel. *Higher-order finite element methods*. CRC Press, 2003.

[123] E. Sonnendrücker and A. Ratnani. Advanced finite element methods. *Lecture notes WS2016/17*, 2016.

[124] G. Starke. Gauss-Newton multilevel methods for least-squares finite element computations of variably saturated subsurface flow. *Computing*, 64:323–338, 2000.

[125] I. W. Stewart. *The static and dynamic continuum theory of liquid crystals: a mathematical introduction*. Crc Press, 2019.

[126] E. Süli. Lecture notes on finite element methods for partial differential equations. *Mathematical Institute, University of Oxford*, 2012.

[127] E. Süli and I. Mozolevski. *hp*-version interior penalty DGFEMs for the biharmonic equation. *Comput. Methods Appl. Mech. Engrg.*, 196(13-16):1851–1863, 2007.

[128] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1997.

[129] U. Trottenberg, C. W. Oosterlee, and A. Schuller. *Multigrid*. Elsevier, 2000.

[130] S. P. Vanka. Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. *J. Comput. Phys.*, 65(1):138–158, 1986.

[131] M. Wang and J. Xu. The Morley element for fourth order elliptic equations in any dimensions. *Numer. Math.*, 103(1):155–169, 2006.

[132] P. Wang, L. Jiang, and S. Chen. A nonconforming scheme for non-Fickian flow in porous media. *J. Inequal. Appl.*, pages Paper No. 142, 16, 2017.

[133] W. Wang, L. Zhang, and P. Zhang. Modeling and computation of liquid crystals, 2021.

[134] T. Warburton and J. S. Hesthaven. On the constants in *hp*-finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.*, 192(25):2765–2773, 2003.

[135] J. Wloka. *Partial Differential Equations*. Cambridge University Press, 1987.

[136] J. Xia. *Computational and analytical aspects of energy minimisation problems in cholesteric, ferronematic and smectic liquid crystals*. PhD thesis, University of Oxford, 2021.

[137] J. Xia and P. E. Farrell. Variational and numerical analysis of a **Q**-tensor model for smectic-A liquid crystals, 2021.

[138] J. Xia, S. MacLachlan, T. J. Atherton, and P. E. Farrell. Structural landscapes in geometrically frustrated smectics. *Phys. Rev. Lett.*, 126:177801, Apr 2021.

[139] H. Zeng. *Light Driven Microscopic Robot*. PhD thesis, 03 2015.

[140] Software used in 'A new mixed finite-element method for H^2 elliptic problems', dec 2021.

[141] S. Zhang. A family of 3D continuously differentiable finite elements on tetrahedral grids. *Appl. Numer. Math.*, 59(1):219–233, 2009.