# Computational and Analytical Aspects of Energy Minimisation Problems in Cholesteric, Ferronematic and Smectic Liquid Crystals


Jingmin Xia

Keble College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2021

This thesis is dedicated to
my beloved family
for their boundless love and company

# Acknowledgements

Going abroad for a PhD was not an easy decision but rather a gamble. Anything could go wrong at any time: I may not like the topic to work on, or I may not meet a nice supervisor to get along with, or I may not be lucky enough to even finish it. However, as I gradually recall these four years that I committed to this journey, it becomes more and more clear in my mind that I have been blessed all the time. This thesis would have been impossible without my supervisor, collaborators, friends, family, and all those people who have shaped and affected me in my life.

I came to Oxford to partake in a four-year PDE CDT (Centre for Doctoral Training) program, with the first year taking classes and working on two mini-projects with different potential supervisors. I took the second mini-project under the supervision of Patrick Farrell and learnt many new skills through this progress. I was completely new to Firedrake and Defcon (two Python-based libraries related to finite elements and computing multiple solutions). My work would have been impossible without Patrick's lasting patience and patience of my ignorance, and most importantly, his excellent guidance. I greatly appreciate his strong support, especially during the last year when everything went online due to the covid-19 pandemic.

I also worked with Saullo Castro for my second mini-project on aircraft stiffeners and I made my first journey to Delft University of Technology because of this collaboration. I still remember my excitement of seeing the aircraft components I had analysed being manufactured in the lab, for use in actual experiments. It made me realise that our virtual simulations connect deeply to the real world.

I feel very lucky to have joined Patrick to attend the 2019 ICERM workshop on liquid crystals in Brown University, where I met some of my collaborators: Apala Majumdar (who invited me), James Dalby, Tim Atherton and Scott MacLachlan, whom I want to thank for their generous time and brilliant ideas shared with me. I also met Anca Andrei in that workshop with whom I visited Tufts and Boston shortly after and went to the Cheesecake Factory (an absolutely wonderful suggestion). I had no idea at that time this friendship would lead to us co-organising the mini-symposium MS66 in the SIAM Conference on Material Sciences (2021), for gave the very enjoyable experience of inviting some renowned researchers working on liquid crystals.

I also want to express my special thanks to the PDE CDT directors (Gui-Qiang Chen and Endre Süli) who provide their generous support for the students. The students in the PDE CDT (and the mathematical institute) have made every working day more enjoyable. In particular, thanks to the company of Timo Sprekeler, Fabian Laakmann, Christoph Höppke and Yikun Qiao, I was kept spiritually well with a rather relaxed and happy mind to commit myself to work. There was a difficult time during the lockdown: everyone had to stay at home and avoid going out. However, the company and amazing cooking of Yuan Zhou reduced the stress and loneliness of that time. The first part of my thesis relates to work done by Florian Wechsung, from whom I received some insightful suggestions that led to our joint paper.

The pandemic has changed everyone's original plans in an inevitable and unpredictable way. It definitely changed mine: I spent my last year of my studies in China. During this special year of completely online research, it was not only the regular catch-up with Patrick and our collaborators that keep me on track, but also the help from some people that comfort my life outside work. They include Zuohong Xu, Yang Zhang, Ping Liu, and Zhouran Zhang (with Lianlian).

The last gratitude goes to my beloved family: their encouragement and kindness nurture my heart silently and warmly. I want to give a special thanks to my dad, who always listens to my endless talks and guides me to think positively, thereby keeping me sane.

Of course, there are many other people whom I have not mentioned, yet whose support remains important. I am even more grateful as I finish this acknowledgement, since without all those people that I met and knew, I would not be who I am today.

# Abstract

Liquid crystals are widely used in display devices and their indispensable applications have driven more than a century of scientific investigations. They are of great interest in physics, for their striking defect structures, e.g., defect walls and focal conics in smectics; and in mathematics, for the questions arising in their modelling and analysis. Two successful mathematical theories are the Oseen–Frank (vector-based) and Landau–de Gennes (tensor-based) theories for nematics. In the former, the order parameter is simple but a nonlinear constraint must be enforced in the optimisation. The latter theory becomes more appealing in characterising complex defects, as it supports defects (e.g., half charge defects) that Oseen–Frank does not. However, when it comes to the phenomenological modelling of other phases of liquid crystals such as smectics, mathematical theories have not been extensively studied. This thesis takes a step forward in understanding several modelling and implementation issues related to three phases of liquid crystals: cholesterics, ferronematics and smectics.

In the first part of this thesis, we propose an augmented Lagrangian-type preconditioner to construct efficient solvers for Oseen–Frank problems arising in cholesterics. We analyse two advantages of the augmented Lagrangian formulation: (i) it helps in controlling the Schur complement matrix, enabling the development of block preconditioners; (ii) it improves the discrete enforcement of the unit-length constraint of the director. Since the augmentation makes the director block harder to solve, we develop a robust multigrid algorithm for the augmented block. The resulting preconditioner is an efficient and robust approach for solving director-based models of liquid crystals.

The second part is devoted to investigating defect structures (e.g., jumps of the director and magnetisation vector) in ferronematics, through numerical bifurcation analysis. Novel bifurcations of the ferronematic problem of interest are studied to give a more complete picture of solution landscapes as the parameter space varies. The reported numerical results validate the corresponding theoretical analysis of Dalby & Majumdar [Dal+21], and show us the potential of the Landau–de Gennes theory in characterising complicated defects.

Convinced by the successful application of the Landau–de Gennes model in ferronematics, we move to developing effective models of smectic-A liquid crystals

in the last part of this thesis. We propose a new continuum model, solving for a real-valued smectic order parameter for the density variation and a tensor-valued nematic order parameter for the director orientation. This expands on an idea mentioned by Ball & Bedford [BB15]. The model is challenging to discretise due to the high regularity of the density variation; to address this, a continuous interior penalty discretisation is employed. Numerical analysis and experiments are performed to confirm the effectiveness of the proposed model and discretisation. The model numerically captures important defect structures in focal conic domains and oily streaks for the first time.

# Contents

# 1 Introduction

## Contents



## 1.1 Background

Liquid crystals (LC), first discovered by Reinitzer in 1888 [Rei88], are materials that can exist in an intermediate mesophase between isotropic (i.e., the physical properties are uniform in all directions) liquids and solid crystals. That is to say, LC can flow like liquids while also possessing long-range orientational order. Based on different ordering symmetries, Friedel [Fri22] proposed to classify them into three broad categories: *nematic*, *smectic* and *cholesteric*, as shown in Figure 1.1. In the following, we briefly summarise the characteristics of these phases (refer to [Ste04; Cha92] for further details).

**Nematic phase** This is the simplest and most extensively studied form of LC phase where the molecules are not layered but tend to point in the same direction. The molecules are free to move (rotate or slide) in this phase

and align approximately parallel to each other, thus giving a long-range orientational ordering.

**Smectic phase** The molecules have similar orientation and point in the same direction as the molecules in nematic LC do but they also tend to line up into layers. Depending on the angles formed between the molecular axes and the planes of molecules, there are a number of different smectic phases. Figure 1.1 depicts the simplest smectic structure, the so-called smectic-A phase.

**Cholesteric phase** This is also known as chiral nematic phase and is characterised by the molecules being aligned and stacked in a helical pattern, with each layer rotated at an angle to the ones above and below it. It has a fixed pitch in its helical structure and is the last phase before the substance becomes a crystal or solid by decreasing the temperature.

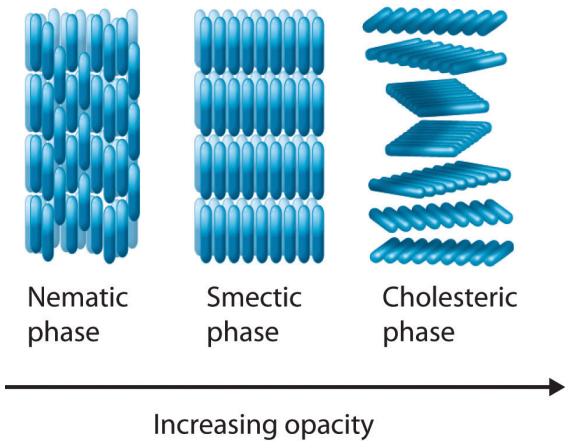



**Figure 1.1:** Three types of molecular orientations in LC. As the temperature is increased, the material goes from solid or crystal through the cholesteric, smectic, nematic and liquid phases. Source: [AE11, Figure 11.26, Section 8]. For example, 8CB melts from crystal at 22°C to the smectic phase, then transitions to the nematic phase at 34°C and becomes a conventional liquid above 42°C [Sci18].

Since the orientational properties of LC can be manipulated by imposing electric fields, they are often used to control light and have formed the basis of several important technologies in the area of electric display devices, e.g., digital screens. This has substantially increased interest in the scientific study of liquid crystals.

Some examples of thorough overviews on LC modelling and its history can be found in [Bal17; Ste04; Cha92]. More relevant to this thesis, there are two main continuum theories for modelling nematic LC: the Oseen–Frank and Landau–de Gennes theories, differing in the order parameters they use to describe the system. They both postulate a free energy, the minimisation of which gives the state of the LC. We include the detailed introduction of each theory later in the relevant part of this thesis.

The working flow of this thesis is as follows. We start with the director-based Oseen–Frank model for cholesteric liquid crystals in Part I. The presence of the unit-length constraint on the director in this model stimulates the need for an efficient and robust solver for the saddle point systems arising in finite element discretisations of the equations. This is inspired by the work [BO06; FMW19] for enforcing the divergence-free constraint in the stationary Navier–Stokes equations by applying the discrete augmented Lagrangian formulation. We propose an augmented-Lagrangian-type preconditioner and derive some robustness estimates in this part.

With this first experience of the Oseen–Frank model, its limitations in characterising more complicated defects (such as half-charge defects) become apparent, since it does not respect the head-to-tail symmetry of LC molecules. To explore and understand the typical defect structures, e.g., oily streaks and focal conics (see Figures 1.2 and 1.3) in smectics, we begin considering the Landau–de Gennes model employing a $\mathbf{Q}$-tensor as the state variable. As a step in this direction, Part II explores the effectiveness of the $\mathbf{Q}$-tensor model in characterising defects by considering a model problem of ferronematics, where magnetic nanoparticles (MNPs) are suspended in a nematic LC and thus induce a spontaneous magnetisation response without any external magnetic fields. In this part, we study the solution landscapes of the ferronematic problem for different parameter spaces and focus on the numerical validations of some theoretical analyses proven by Dalby & Majumdar in [Dal+21].

This substantial success in observing some interesting defect structures in ferronematics stimulates our interest in investigating more sophisticated defects in smectics

**Figure 1.2:** Schematic illustration of flattened hemicylinders (left) and curvature wall (right) in smectic-A thin films. Source: [Mic+04, Fig. 9 & Fig. 16].

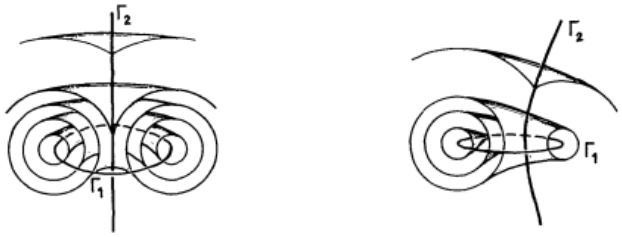



**Figure 1.3:** Schematic representation of toroidal focal conic domains (left) and focal conic domains (right) in smectic-A thin films. Here, $\Gamma_1$ and $\Gamma_2$ are two singularities resulting from keeping equidistant layer spacing. We can notice that the ellipse collapses to a circle and hyperbola into a straight line in the toroidal case. Source: [WK75, Fig. 1].

and thereby leads to our work in Part III. We propose a new mathematical model for smectic-A LC in this last part, which successfully captures typical structures of oily streaks and focal conic domains. We believe it can be applied to many other smectic scenarios that require an effective and efficient mathematical model.

Following this working flow, we divide the remainder of this thesis into three parts regarding different applications in liquid crystals, i.e., cholesterics, ferronematics and smectics, and close with some conclusions and potential directions for future work. Each part expands upon a relevant publication, as detailed below.

- Part I: *Xia, Farrell and Wechsung* (2021) [XFW21], published in *BIT Numerical Mathematics*.

- Part II: *Dalby, Farrell, Majumdar and Xia* (2021) [Dal+21], in review in *SIAM Journal on Applied Mathematics*.

- Part III: *Xia, MacLachlan, Atherton and Farrell* (2021) [Xia+21], published in *Physical Review Letters*.

## 1.2 Some common notation

$d$ spatial dimension, $d \in \{1, 2, 3\}$

$\Omega$ open, bounded $d$-dimensional domain with Lipschitz boundary $\partial\Omega$

$x, y, z$ coordinates of domain $\Omega$

$C$ generic constant that may not be the same constant for each appearance

$h$ mesh size

$\mathcal{T}_h$ mesh of $\Omega$

$T$ element of $\mathcal{T}_h$

$\mathcal{E}_I$ set of all interior edges/faces of the mesh $\mathcal{T}$

$\mathcal{E}_B$ set of all boundary edges/faces of the mesh $\mathcal{T}$

$\mathcal{E}$ set of all edges/faces of the mesh $\mathcal{T}$; $\mathcal{E} = \mathcal{E}_I \cup \mathcal{E}_B$

$S_0$ space of symmetric, traceless $d \times d$ matrices

$\mathcal{S}^{d-1}$ ($d \in \{2, 3\}$) surface of the unit ball in $\mathbb{R}^d$ centered at the origin

$\mathbb{M}^{d\times d}$ space of all $d \times d$ matrices

$\mathbf{I}_d$ identity matrix in $\mathbb{M}^{d\times d}$, $\mathbf{I}$ general identity matrix

$\mathcal{A}$ admissible space of a minimisation problem

$\mathbb{P}^{\Bbbk}$ piecewise continuous polynomials of degree $\Bbbk \geq 0$ on a simplicial mesh (intervals, triangles and tetrahedra)

$\mathbb{Q}^{\Bbbk}$ piecewise continuous polynomials of degree $\Bbbk \geq 0$ on a mesh of quadrilaterals or hexahedra

$\nu$ outward unit normal to the boundary $\partial\Omega$

$H^{\Bbbk}$ Sobolev space of square-integrable functionals with square-integrable weak derivatives up to $\Bbbk$ order with standard $H^{\Bbbk}$-norm $\|\cdot\|_{\Bbbk}$ on $\Omega$

$\mathbf{H}^{\Bbbk}$ vector-valued version of $H^{\Bbbk}$

$H_b^{\Bbbk}$, $\mathbf{H}_b^{\Bbbk}$ Sobolev spaces $H^{\Bbbk}$, $\mathbf{H}^{\Bbbk}$ with an addition of the trace

$\|\cdot\|_0$, $\|\cdot\|_\infty$ standard $L^2$- and $L^\infty$-norm on $\Omega$

$(\cdot, \cdot)_0$ inner product in $L^2(\Omega)$

$\Delta = \nabla \cdot \nabla$ Laplace operator

$\mathcal{D}^2$ Hessian operator

In order to avoid the proliferation of constants throughout this thesis, we use the notation $a \lesssim b$ (respectively, $b \gtrsim a$) to represent the relation $a \leq Cb$ (respectively, $b \geq Ca$) for some generic constant $C$ independent of the mesh.

## 1.3 Common solver details

Since the problems to be solved in this thesis are all nonlinear, we always use Newton's method with $L^2$ linesearch [Bru+15, Algorithm 2] as the outer nonlinear solver. The solver is implemented in the Firedrake [Rat+17] library, which relies on PETSc [Bal+18] for solving the linear systems resulted from linearising the nonlinear problem.

In addition, for those problems (e.g., in Parts II and III) where we are interested in multiple solutions or bifurcation diagrams, we use the so-called *deflation* technique as described in [FBF15] to compute multiple solutions. This technique is implemented in the Defcon library [Far17].

Further details of each solver used for the numerical experiments in Chapters 4, 6 and 9 will be specified later. For reproducibility, the exact versions of the implementation codes used have been archived on Zenodo; the appropriate archived code will be cited in the corresponding chapter.

# Part I

# Cholesteric Liquid Crystals

---

**This work expands upon *Xia, Farrell and Wechsung (2021)* [XFW21].**

---

# 2

# A mathematical model of cholesterics

## Contents



As mentioned in the previous chapter, one of the commonly used mathematical models for nematic and cholesteric liquid crystals is the Oseen–Frank theory [Ose33; Fra58], which takes a unit-length vector field as the state variable. We therefore introduce the form of the Oseen–Frank model that we will subsequently consider.

## 2.1 The Oseen–Frank model

Let $\Omega \subset \mathbb{R}^d, d \in \{2, 3\}$ be an open, bounded domain with Lipschitz boundary $\partial\Omega$. We triangulate the domain $\Omega$ and denote the mesh by $\mathcal{T}_h$ with each element represented by $\mathcal{T}$ and $h$ is the mesh size. Denote $\mathbf{H}^1_b(\Omega) = \{\mathbf{v} \in H^1(\Omega, \mathbb{R}^3) : \mathbf{v}|_{\partial\Omega} = \mathbf{n}_b\}$ for a given vector field $\mathbf{n}_b \in H^{1/2}(\partial\Omega, \mathcal{S}^2)$ with $\mathbf{H}^1_0$ given by zero trace $\mathbf{n}_b = \mathbf{0} \notin \mathcal{S}^2$. Assume that the (nematic or) cholesteric LC occupying the domain $\Omega$

is equipped with a rigid anchoring (Dirichlet) boundary condition $\mathbf{n}|_{\partial\Omega} = \mathbf{n}_b$[1]. The Oseen–Frank model [Fra58] considers the following minimisation problem:

$$\begin{aligned} &\min_{\mathbf{n}\in\mathbf{H}^1_b(\Omega)} \mathcal{J}^{OF}(\mathbf{n}) = \int_\Omega W^{OF}(\mathbf{n}), \\ &\text{subject to } \mathbf{n}\cdot\mathbf{n} = 1 \text{ a.e.}, \end{aligned} \tag{2.1.0.1}$$

where the Frank energy density $W^{OF}(\mathbf{n})$ is of the form

$$\begin{aligned} W^{OF}(\mathbf{n}) = \frac{K_1}{2}(\nabla\cdot\mathbf{n})^2 + \frac{K_2}{2}(\mathbf{n}\cdot(\nabla\times\mathbf{n}) + q_0)^2 + \frac{K_3}{2}|\mathbf{n}\times(\nabla\times\mathbf{n})|^2 \\ + \frac{K_2+K_4}{2}[\mathrm{tr}((\nabla\mathbf{n})^2) - (\nabla\cdot\mathbf{n})^2], \end{aligned} \tag{2.1.0.2}$$

with $\mathrm{tr}(\cdot)$ the trace of a matrix, $K_i \in \mathbb{R}$ $(i = 1, 2, 3, 4)$ elastic constants (called *Frank constants*) and $q_0 \geq 0$ the preferred pitch for the cholesteric. $K_1$, $K_2$, $K_3$, and $K_4$ are referred to as the splay, twist, bend, and saddle-splay constants, respectively. Note here $\nabla\mathbf{n}$ is matrix-valued and $(\nabla\mathbf{n})^2$ denotes the matrix multiplication of the matrix $\nabla\mathbf{n}$ and itself.

If $K_1 = K_2 = K_3 = K_c > 0$ and $K_4 = 0$, the energy density (2.1.0.2) reduces to the so-called *equal-constant* approximation, with energy density

$$W^{OF}(\mathbf{n}) = \frac{K_c}{2}\left[|\nabla\mathbf{n}|^2 + 2q_0\mathbf{n}\cdot(\nabla\times\mathbf{n}) + q_0^2\right],$$

which is a useful simplification to help us gain qualitative insight into more complex situations.

**Remark 2.1.** *When $q_0 = 0$, the energy density (2.1.0.2) corresponds to the nematic case. Furthermore, when combined with the equal-constant approximation, (2.1.0.2) reduces to*

$$W^{OF}(\mathbf{n}) = \frac{K_c}{2}|\nabla\mathbf{n}|^2.$$

*With this free energy density, the solution to the minimisation problem (2.1.0.1) is unique and is known as the harmonic map from a two- or three-dimensional compact manifold to $\mathcal{S}^2$ [Lin89]. Some fast numerical algorithms for this equal-constant approximation case have been proposed and tested in [HTW09].*

[1] The following theory also applies with mixed periodic and Dirichlet boundary conditions [Adl+15b; Bed14], which we shall use in some numerical examples in Chapter 4.

Using the fact that

$$\mathrm{tr}((\nabla \mathbf{n})^2) - (\nabla \cdot \mathbf{n})^2 = \nabla \cdot ((\mathbf{n} \cdot \nabla)\mathbf{n} - (\nabla \cdot \mathbf{n})\mathbf{n}),$$

the last term (the *saddle-splay* term or the *null Lagrangian*) in (2.1.0.2) can be dropped as its integral reduces to a surface integral, which is essentially a constant if applying Dirichlet boundary conditions to the model, via the divergence theorem. For mixed periodic and Dirichlet boundary conditions considered in Section 4.2.1, we can verify directly that this saddle-splay energy vanishes. Hence, for simplicity, it suffices to consider the following Frank energy density

$$W^{OF}(\mathbf{n}) = \frac{K_1}{2}(\nabla \cdot \mathbf{n})^2 + \frac{K_2}{2}(\mathbf{n} \cdot (\nabla \times \mathbf{n}) + q_0)^2 + \frac{K_3}{2}|\mathbf{n} \times (\nabla \times \mathbf{n})|^2. \quad (2.1.0.3)$$

In this chapter, we use a more compact form of the free energy (2.1.0.1) as in [Adl+15b; Adl+16] by introducing a symmetric dimensionless tensor

$$\mathbf{Z} = \kappa \mathbf{n} \otimes \mathbf{n} + (\mathbf{I}_3 - \mathbf{n} \otimes \mathbf{n}) = \mathbf{I}_3 + (\kappa - 1)\mathbf{n} \otimes \mathbf{n},$$

where $\kappa = K_2/K_3$. By the classical equality

$$|\nabla \times \mathbf{n}|^2 = (\mathbf{n} \cdot (\nabla \times \mathbf{n}))^2 + |\mathbf{n} \times (\nabla \times \mathbf{n})|^2, \quad (2.1.0.4)$$

the original energy functional $\mathcal{J}^{OF}(\mathbf{n})$ can be rewritten as

$$\begin{aligned}\mathcal{J}^{OF}(\mathbf{n}) = \frac{1}{2}\big(&K_1(\nabla \cdot \mathbf{n}, \nabla \cdot \mathbf{n})_0 + K_3(\mathbf{Z}\nabla \times \mathbf{n}, \nabla \times \mathbf{n})_0 \\ &+ 2K_2 q_0(\mathbf{n}, \nabla \times \mathbf{n})_0 + K_2(q_0, q_0)_0\big).\end{aligned} \quad (2.1.0.5)$$

It can be observed that the auxiliary tensor $\mathbf{Z}$ contributes to the nonlinearity of $\mathcal{J}^{OF}(\mathbf{n})$ in (2.1.0.5).

**Remark 2.2.** *There is another widely used simplification of the energy density (2.1.0.2), where $q_0 = 0$ and $K_2 = K_3 = K_1 + K_p$, $K_4 = -K_p$ [GLP03; LR07]. In this case, (2.1.0.2) becomes*

$$W^{OF}(\mathbf{n}) = \frac{1}{2}[K_1|\nabla \mathbf{n}|^2 + K_p|\nabla \times \mathbf{n}|^2],$$

*and it is expected that as $K_p \to \infty$, the asymptotic behavior of minimisers provides a description of the phase transition process of LC from the nematic phase to the smectic-A phase [GLP03; LR07; LT14].*

Furthermore, it is proven in [Adl+15b, Section 2.3] that $\mathbf{Z}$ is uniformly (with respect to $\mathbf{x} \in \Omega$) symmetric positive definite (USPD) as long as sufficient control is maintained on $|\mathbf{n}|^2$. This property of $\mathbf{Z}$ plays an essential role in proving the well-posedness of the saddle-point problem in the nematic case. We restate the result of $\mathbf{Z}$ being USPD in the following, as it is important later:

**Lemma 2.1.** *[Adl+15b, Section 2.3] Assume $\alpha \leq |\mathbf{n}|^2 \leq \beta \ \forall \mathbf{x} \in \Omega$ with $0 < \alpha \leq 1 \leq \beta$. If $\kappa > 1$, then $\mathbf{Z}$ is USPD on $\Omega$; for $0 < \kappa < 1$, then $\mathbf{Z}$ is USPD on $\Omega$ if $\beta < \frac{1}{1-\kappa}$.*

**Remark 2.3.** *Notice that the regularity of $\mathbf{n} \in \mathbf{H}^1(\Omega)$ is enough for the functional $\mathcal{J}^{OF}(\mathbf{n})$ of (2.1.0.5) to be well defined. In fact, $\mathbf{n} \in \mathbf{H}^1(\Omega)$ implies $\nabla \cdot \mathbf{n} \in L^2(\Omega)$ and $\nabla \times \mathbf{n} \in \mathbf{L}^2(\Omega)$. By (2.1.0.4), we have $\mathbf{n} \cdot (\nabla \times \mathbf{n}) \in L^2(\Omega)$. This ensures that the term $(q_0, \mathbf{n} \cdot (\nabla \times \mathbf{n}))_0$ in (2.1.0.5) is defined. Furthermore, Lemma 2.1 gives the boundedness of $\mathbf{Z}$, which guarantees the $L^2$-regularity of the term $\mathbf{Z}\nabla \times \mathbf{n}$ in (2.1.0.5).*

Naturally, the values of elastic constants and the cholesteric pitch will be an important factor in determining the minimisers. In particular, the free energy density should be bounded from below so to ensure the existence of minimisers. With an addition of arbitrary constant, we thus need additional assumptions on the parameters to satisfy non-negativity of the energy density, i.e.,

$$W^{OF}(\mathbf{n}) \geq 0 \quad \forall \mathbf{n} \in \mathbf{H}^1_b(\Omega).$$

This gives rise to Ericksen's inequalities (see [Bal17; Bed14] and references therein):

$$\begin{aligned} &K_1, K_2, K_3 \geq 0, K_2 + K_4 = 0 && \text{if } q_0 \neq 0, \\ &2K_1 \geq K_2 + K_4, K_2 \geq |K_4|, K_3 \geq 0 && \text{if } q_0 = 0. \end{aligned}$$

**Remark 2.4.** *We have included the inequalities with regard to constant $K_4$ here for generality, though they are not necessary in our work as we have eliminated the $K_4$-related term in the free energy. In this part, we will simply consider $K_i > 0$ $(i = 1, 2, 3)$ to avoid any technical issues.*

For the minimisation problem (2.1.0.1) arising in (nematic or cholesteric) liquid crystals, it has been proven in [Lin89, Theorem 2.1] that there exists a solution.

**Theorem 2.2.** *[Lin89, Theorem 2.1] Let $\Omega$ be a bounded Lipschitz domain and assume the Dirichlet boundary data $\mathbf{n}_b \in H^{1/2}(\partial\Omega, \mathcal{S}^2)$. If $K_1, K_2, K_3 > 0$, then there exists an $\mathbf{n} \in H^1_b(\Omega, \mathcal{S}^2) := \{\mathbf{n} \in H^1(\Omega, \mathcal{S}^2) : \mathbf{n} = \mathbf{n}_b \text{ on } \partial\Omega\}$ such that*

$$\mathcal{J}^{OF}(\mathbf{n}) = \inf_{\mathbf{u} \in H^1_b(\Omega, \mathcal{S}^2)} \mathcal{J}^{OF}(\mathbf{u}).$$

The main difficulty in numerically solving the Oseen–Frank model (2.1.0.1) is the enforcement of the unit-length constraint. There are several existing approaches to handling constraints, e.g., projection [LT14], Lagrange multipliers, and penalty methods [NW99, Section 12.3 & 17].

The projection method is numerically simple but the value of the energy functional may go up and down dramatically after each projection, making it difficult to control in the optimisation procedure [LT14]. A Lagrange multiplier is often used to replace constrained optimisation problems with unconstrained ones, but an important disadvantage of this approach is that it introduces another unknown (i.e., the Lagrange multiplier) and leads to a saddle-point structure which can be difficult to solve [BGL05]. On the other hand, the penalty method has the favorable property that the resulting system has an energy decay property [LR07] which may result in an easier theoretical and numerical study of the solution. However, the penalty parameter has to be very large for the accuracy of approximating the constraints, leading to an ill-conditioned system. Some works based on either projection or pure penalty methods for nematic phases can be found in [GLP03; LR07; GL89] and the references therein.

Fortunately, it is possible to amend the ill-conditioning effects with large penalty parameters that are inherent in the pure penalty method by combining it with a Lagrange multiplier. This is the *augmented Lagrangian* algorithm [FG83]. This strategy combines the advantages of both schemes: the penalty parameter can be relatively small due to the presence of the Lagrange multiplier, and the Schur complement of the saddle-point system is easier to solve due to the presence of

the penalty term [GLP03; GL89; Ols02; BO06; FMW19]. Since the concept of the Schur complement is closely related to this part of the thesis, we briefly summarise the approach of Schur complement reduction here. Consider a saddle-point system (that is, it has both positive and negative eigenvalues) of form

$$\mathbf{D}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} := \begin{bmatrix}\mathbf{A}_1 & \mathbf{B}_1^\top\\ \mathbf{B}_2 & \mathbf{C}_1\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathbf{c}\\ \mathbf{d}\end{bmatrix}. \tag{2.1.0.6}$$

Assuming that both $\mathbf{A}_1$ and $\mathbf{D}$ are nonsingular, it implies that $\mathbf{S}_1 = \mathbf{C}_1 - \mathbf{B}_2\mathbf{A}_1^{-1}\mathbf{B}_1^\top$ is also nonsingular [BGL05]. Here, $\mathbf{S}_1$ is the so-called *Schur complement*. Block Gaussian elimination then reduces the system (2.1.0.6) to

$$\begin{bmatrix}\mathbf{A}_1 & \mathbf{B}_1^\top\\ \mathbf{0} & \mathbf{S}_1\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} = \begin{bmatrix}\mathbf{c}\\ \mathbf{d} - \mathbf{B}_2\mathbf{A}_1^{-1}\mathbf{c}\end{bmatrix}. \tag{2.1.0.7}$$

If it is possible to solve linear systems involving $\mathbf{A}_1$ and $\mathbf{S}_1$, we can solve the coupled linear system.

In what follows, we first consider the method of Lagrange multipliers and then add the augmented Lagrangian term to control the Schur complement of the system.

## 2.2 Lagrange multiplier and Newton linearisation

By introducing the Lagrange multiplier $\lambda \in L^2(\Omega)$, the associated Lagrangian of the minimisation problem (2.1.0.1) is then defined as

$$\mathcal{L}(\mathbf{n}, \lambda) = \mathcal{J}^{OF}(\mathbf{n}) + (\lambda, \mathbf{n}\cdot\mathbf{n} - 1)_0\,, \tag{2.2.0.1}$$

and its first-order optimality conditions are: find $(\mathbf{n}, \lambda) \in \mathbf{H}_b^1(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned}
\mathcal{L}_{\mathbf{n}}[\mathbf{v}] &= \mathcal{J}_{\mathbf{n}}^{OF}[\mathbf{v}] + (\lambda, 2\mathbf{n}\cdot\mathbf{v})_0\\
&= K_1\left(\nabla\cdot\mathbf{n}, \nabla\cdot\mathbf{v}\right)_0 + K_3\left(\mathbf{Z}\nabla\times\mathbf{n}, \nabla\times\mathbf{v}\right)_0\\
&\quad + (K_2 - K_3)\left(\mathbf{n}\cdot\nabla\times\mathbf{n}, \mathbf{v}\cdot\nabla\times\mathbf{n}\right)_0\\
&\quad + K_2 q_0\left(\mathbf{v}, \nabla\times\mathbf{n}\right)_0 + K_2 q_0\left(\mathbf{n}, \nabla\times\mathbf{v}\right)_0 + (\lambda, 2\mathbf{n}\cdot\mathbf{v})_0\\
&= 0 \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega),\\
\mathcal{L}_\lambda[\mu] &= (\mu, \mathbf{n}\cdot\mathbf{n} - 1)_0 = 0 \quad \forall \mu \in L^2(\Omega).
\end{aligned} \tag{2.2.0.2}$$

As (2.2.0.2) is nonlinear, Newton linearisation is employed. Let $\mathbf{n}_j$ and $\lambda_j$ be the current approximations for $\mathbf{n}$ and $\lambda$, respectively, and denote the corresponding updates to these approximations as $\delta\mathbf{n} = \mathbf{n}_{j+1} - \mathbf{n}_j$ and $\delta\lambda = \lambda_{j+1} - \lambda_j$. Then the Newton iteration at $(\mathbf{n}_j, \lambda_j)$ in block form is given by: find $(\delta\mathbf{n}, \delta\lambda) \in \mathbf{H}^1_0(\Omega) \times L^2(\Omega)$ such that

$$\begin{bmatrix} \mathcal{L}_{\mathbf{nn}} & \mathcal{L}_{\mathbf{n}\lambda} \\ \mathcal{L}_{\lambda\mathbf{n}} & 0 \end{bmatrix} \begin{bmatrix} \delta\mathbf{n} \\ \delta\lambda \end{bmatrix} = - \begin{bmatrix} \mathcal{L}_{\mathbf{n}} \\ \mathcal{L}_{\lambda} \end{bmatrix}, \tag{2.2.0.3}$$

where

$$\begin{aligned} \mathcal{L}_{\mathbf{nn}}[\mathbf{v}, \delta\mathbf{n}] &= J_{\mathbf{nn}}[\mathbf{v}, \delta\mathbf{n}] + (\lambda_j, 2\delta\mathbf{n} \cdot \mathbf{v})_0 \\ &= K_1 \left(\nabla \cdot \delta\mathbf{n}, \nabla \cdot \mathbf{v}\right)_0 + K_3 \left(\mathbf{Z}(\mathbf{n}_j)\nabla \times \delta\mathbf{n}, \nabla \times \mathbf{v}\right)_0 \\ &\quad + (K_2 - K_3)\Big( (\delta\mathbf{n} \cdot \nabla \times \mathbf{n}_j, \mathbf{n}_j \cdot \nabla \times \mathbf{v})_0 + (\mathbf{n}_j \cdot \nabla \times \mathbf{n}_j, \delta\mathbf{n} \cdot \nabla \times \mathbf{v})_0 \\ &\quad + (\mathbf{v} \cdot \nabla \times \mathbf{n}_j, \mathbf{n}_j \cdot \nabla \times \delta\mathbf{n})_0 + (\mathbf{n}_j \cdot \nabla \times \mathbf{n}_j, \mathbf{v} \cdot \nabla \times \delta\mathbf{n})_0 \\ &\quad + (\delta\mathbf{n} \cdot \nabla \times \mathbf{n}_j, \mathbf{v} \cdot \nabla \times \mathbf{n}_j)_0 \Big) \\ &\quad + K_2 q_0 \left(\mathbf{v}, \nabla \times \delta\mathbf{n}\right)_0 + K_2 q_0 \left(\delta\mathbf{n}, \nabla \times \mathbf{v}\right)_0 + (\lambda_j, 2\delta\mathbf{n} \cdot \mathbf{v})_0 , \end{aligned} \tag{2.2.0.4}$$

and

$$\begin{aligned} \mathcal{L}_{\mathbf{n}\lambda}[\mathbf{v}, \delta\lambda] &= (\delta\lambda, 2\mathbf{n}_j \cdot \mathbf{v})_0 , \\ \mathcal{L}_{\lambda\mathbf{n}}[\mu, \delta\mathbf{n}] &= (\mu, 2\mathbf{n}_j \cdot \delta\mathbf{n})_0 . \end{aligned}$$

Since $\mathcal{L}(\mathbf{n}, \lambda)$ is linear in $\lambda$, $\mathcal{L}_{\lambda\lambda} = 0$. This results in (2.2.0.3) being a saddle-point problem.

With a suitable spatial discretisation (we only consider conforming finite elements throughout this part of the thesis, i.e., the finite dimensional space $V_h \subset \mathbf{H}^1_0(\Omega)$ that the finite element approximation $\mathbf{n}_h$ of $\mathbf{n}$ belongs to, and the finite dimensional space $Q_h \subset L^2(\Omega)$ that the approximation $\lambda_h$ of $\lambda$ belongs to), a symmetric saddle-point system must be solved at each Newton iteration:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} U \\ X \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{2.2.0.5}$$

where $U$ and $X$ represent the coefficient vectors of $\delta\mathbf{n}$ and $\delta\lambda$ in terms of the basis functions of $V_h$ and $Q_h$, respectively.

We can accordingly write the discrete variational problem as: find $\delta\mathbf{n}_h \in V_h$ and $\delta\lambda_h \in Q_h$ such that

$$\begin{aligned} \mathfrak{a}(\delta\mathbf{n}_h, \mathbf{v}_h) + \mathfrak{b}(\mathbf{v}_h, \delta\lambda_h) &= \mathfrak{f}(\mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_h, \\ \mathfrak{b}(\delta\mathbf{n}_h, \mu_h) &= \mathfrak{g}(\mu_h) \quad \forall \mu_h \in Q_h, \end{aligned} \tag{2.2.0.6}$$

where $\mathfrak{a}(\cdot,\cdot)$ and $\mathfrak{b}(\cdot,\cdot)$ are bilinear forms given by

$$\begin{aligned} \mathfrak{a}(\mathbf{u},\mathbf{v}) =& K_1 \left(\nabla\cdot\mathbf{u}, \nabla\cdot\mathbf{v}\right)_0 + K_3 \left(\mathbf{Z}(\mathbf{n}_j)\nabla\times\mathbf{u}, \nabla\times\mathbf{v}\right)_0 \\ &+ (K_2 - K_3)\Big( \left(\mathbf{u}\cdot\nabla\times\mathbf{n}_j, \mathbf{n}_j\cdot\nabla\times\mathbf{v}\right)_0 + \left(\mathbf{n}_j\cdot\nabla\times\mathbf{n}_j, \mathbf{u}\cdot\nabla\times\mathbf{v}\right)_0 \\ &+ \left(\mathbf{v}\cdot\nabla\times\mathbf{n}_j, \mathbf{n}_j\cdot\nabla\times\mathbf{u}\right)_0 + \left(\mathbf{n}_j\cdot\nabla\times\mathbf{n}_j, \mathbf{v}\cdot\nabla\times\mathbf{u}\right)_0 \\ &+ \left(\mathbf{u}\cdot\nabla\times\mathbf{n}_j, \mathbf{v}\cdot\nabla\times\mathbf{n}_j\right)_0 \Big) \\ &+ K_2 q_0 \left(\mathbf{v}, \nabla\times\mathbf{u}\right)_0 + K_2 q_0 \left(\mathbf{u}, \nabla\times\mathbf{v}\right)_0 + \left(\lambda_j, 2\mathbf{u}\cdot\mathbf{v}\right)_0, \end{aligned}$$

and

$$\mathfrak{b}(\mathbf{v}, p) = \left(p, 2\mathbf{n}_j\cdot\mathbf{v}\right)_0,$$

and $\mathfrak{f}$ and $\mathfrak{g}$ are linear functionals given by

$$\begin{aligned} \mathfrak{f}(\mathbf{v}) =& -\Big( K_1 \left(\nabla\cdot\mathbf{n}_j, \nabla\cdot\mathbf{v}\right)_0 + K_3 \left(Z(\mathbf{n}_j)\nabla\times\mathbf{n}_j, \nabla\times\mathbf{v}\right)_0 \\ &+ (K_2 - K_3)\left(\mathbf{n}_j\cdot\nabla\times\mathbf{n}_j, \mathbf{v}\cdot\nabla\cdot\mathbf{n}_j\right)_0 \\ &+ K_2 q_0 \left(\mathbf{v}, \nabla\times\mathbf{n}_j\right)_0 + K_2 q_0 \left(\mathbf{n}_j, \nabla\times\mathbf{v}\right)_0 \\ &+ \left(\lambda_j, 2\mathbf{n}_j\cdot\mathbf{v}\right)_0 \Big), \end{aligned}$$

and

$$\mathfrak{g}(\mu) = -\left(\mu, \mathbf{n}_j\cdot\mathbf{n}_j - 1\right)_0.$$

**Remark 2.5.** *The well-posedness of the continuous and discretised Newton system (with the $([\mathbb{Q}_{\Bbbk}]^d \oplus \mathbb{B}_F)$-$\mathbb{Q}_0$ finite element pair, $\Bbbk \geq 1$) for a generalised nematic LC problem is discussed in [Adl+15b], where $\mathbb{B}_F := \{\mathbf{v} \in [\mathcal{C}_c(\Omega)]^d : \mathbf{v}|_T = a_T b_T \mathbf{n}_j | T \ \forall T \in \mathcal{T}_h\}$ denotes the bubble space. Here, $\mathcal{C}_c(\Omega)$ includes compactly supported continuous functions, $b_T$ represents biquadratic bubble function that vanishes on $\partial T \in \mathcal{T}_h$ and satisfies*

$$\begin{cases} \int_T b_T = 1 & \forall T \in \mathcal{T}_h, \\ b_T(x) > 0 & \forall x \in T, \end{cases}$$

*and $a_T$ is a constant associated with $b_T$. Moreover, the authors of [Adl+16] considered the pure penalty approach for nematic LC and obtained a well-posedness result of the penalised Newton iteration through similar techniques. We will follow these analysis strategies in this section.*

It is straightforward to deduce the well-posedness of the discrete Newton iteration (2.2.0.6) for cholesteric problems under some proper assumptions on the problem-dependent constants. In fact, two additional $q_0$-related terms in $\mathcal{L}_{\mathbf{nn}}$ from (2.2.0.4) compared to the nematic energy density from [Adl+15b] are simply $L^2$ inner products, which can be easily bounded above using the Cauchy–Schwarz and triangle inequalities. We start with the assumptions and subsequently prove some necessary ingredients, e.g., the coercivity and boundedness of $\mathfrak{a}(\cdot,\cdot)$ and the discrete inf-sup condition for $\mathfrak{b}(\cdot,\cdot)$, of the well-posedness result.

**Assumption 2.3.** *Assume that there exist constants $0 < \alpha \leq 1 \leq \beta$ such that $\alpha \leq |\mathbf{n}_j|^2 \leq \beta$. For $0 < \kappa < 1$, assume further that $\beta < \frac{1}{1-\kappa}$. By Lemma 2.1, $\mathbf{Z}(\mathbf{n}_j)$ remains USPD with lower bound $\Lambda_l$ and upper bound $\Lambda_u$, i.e.,*

$$\Lambda_l \leq \frac{\mathbf{x}^\top \mathbf{Z}(\mathbf{n}_j)\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \leq \Lambda_u \quad \forall \mathbf{x} \in \mathbb{R}^d \backslash \{\mathbf{0}\}.$$

**Lemma 2.4.** *(Continuous coercivity) With Assumption 2.3, we assume further that the current Lagrange multiplier approximation $\lambda_j$ is pointwise non-negative almost everywhere. Let $K_1 > K_2 q_0 C_4$ and $K_3 \Lambda_l > K_2 q_0 (C_4 + 1)$ with $C_4$ to be defined. Then there exists an $\alpha_0 > 0$ such that*

$$\alpha_0 \|\mathbf{v}\|_1^2 \leq \mathfrak{a}(\mathbf{v}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega). \tag{2.2.0.7}$$

*Moreover, when $\kappa = 1$, i.e., $K_2 = K_3$, if $K_1 > K_2 q_0 C_4$ and $1 > q_0(C_4 + 1)$, then the coercivity result (2.2.0.7) also holds.*

**Remark 2.6.** *One may wonder how realistic that $\lambda_j$ can be pointwise non-negative almost everywhere during each nonlinear iteration. However, we do not observe any ill-posed problems during our numerical experiments that are illustrated in Chapter 4.*

*Proof.* With the lower bound $\Lambda_l$ of $\mathbf{Z}$, we compute:

$$\begin{aligned}\mathfrak{a}(\mathbf{v},\mathbf{v}) &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_3\Lambda_l\|\nabla\times\mathbf{v}\|_0^2 + 2K_2q_0\,(\mathbf{v},\nabla\times\mathbf{v})_0 + 2\,(\lambda_j,\mathbf{v}\cdot\mathbf{v})_0\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_3\Lambda_l\|\nabla\times\mathbf{v}\|_0^2 - 2K_2q_0|\,(\mathbf{v},\nabla\times\mathbf{v})_0\,|\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_3\Lambda_l\|\nabla\times\mathbf{v}\|_0^2 - 2K_2q_0\|\mathbf{v}\|_0\|\nabla\times\mathbf{v}\|_0\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_3\Lambda_l\|\nabla\times\mathbf{v}\|_0^2 - K_2q_0(\|\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2),\end{aligned}$$

where the first inequality comes from the assumption that $\lambda_j$ is non-negative pointwise and the last two inequalities are derived by Cauchy–Schwarz and Hölder inequalities, respectively.

By [GR11, Remark 2.7], for a bounded Lipschitz domain, there exists $C_1 > 0$ such that

$$\|\nabla\mathbf{v}\|_0^2 \leq C_1(\|\nabla\cdot\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2),$$

for all $\mathbf{v}\in\mathbf{H}_0(\mathrm{div},\Omega)\cap\mathbf{H}_0(\mathrm{curl},\Omega)$[2]. Here, we denote

$$\begin{aligned}\mathbf{H}_0(\mathrm{div},\Omega) &= \{\mathbf{v}\in\mathbf{L}^2(\Omega) : \nabla\cdot\mathbf{v}\in L^2(\Omega), \nu\cdot\mathbf{v} = 0 \text{ on } \partial\Omega\},\\ \mathbf{H}_0(\mathrm{curl},\Omega) &= \{\mathbf{v}\in\mathbf{L}^2(\Omega) : \nabla\times\mathbf{v}\in\mathbf{L}^2(\Omega), \nu\times\mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\}.\end{aligned}$$

Then using the classical Poincaré inequality, $\|\mathbf{v}\|_0^2 \leq C_3\|\nabla\mathbf{v}\|_0^2$ for all $\mathbf{v}\in\mathbf{H}_0^1(\Omega)$, and defining $C_4 = C_1C_3 > 0$, we have

$$\|\mathbf{v}\|_0^2 \leq C_4(\|\nabla\cdot\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2).$$

Furthermore, there exists $C_2 = C_4 + C_1 > 0$ such that

$$\|\mathbf{v}\|_1^2 \leq C_2(\|\nabla\cdot\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2).$$

It follows that

$$\begin{aligned}\mathfrak{a}(\mathbf{v},\mathbf{v}) &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_3\Lambda_l\|\nabla\times\mathbf{v}\|_0^2 - K_2q_0\left[C_4\left(\|\nabla\cdot\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2\right) - \|\nabla\times\mathbf{v}\|_0^2\right]\\ &= (K_1 - K_2q_0C_4)\|\nabla\cdot\mathbf{v}\|_0^2 + (K_3\Lambda_l - K_2q_0C_4 - K_2q_0)\|\nabla\times\mathbf{v}\|_0^2.\end{aligned}$$

Choosing $C_5 = \min\{K_1 - K_2q_0C_4, K_3\Lambda_l - K_2q_0C_4 - K_2q_0\} > 0$ (the positivity follows from the assumptions) and $\alpha_0 = C_5/C_2$, we find that the coercivity (2.2.0.7) holds.

[2] In fact, $\mathbf{H}_0^1(\Omega) = \mathbf{H}_0(\mathrm{div},\Omega)\cap\mathbf{H}_0(\mathrm{curl},\Omega)$ holds for any bounded Lipschitz domain $\Omega$ [GR11, Lemma 2.5].

In particular, when $\kappa = 1$ (i.e., $K_2 = K_3$), we have $\mathbf{Z} = \mathbf{I}_3$ and thus $\Lambda_l = 1$. Then, the bilinear form becomes

$$\begin{aligned}\mathfrak{a}(\mathbf{v},\mathbf{v}) &= K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_2\|\nabla\times\mathbf{v}\|_0^2 + 2K_2q_0\,(\mathbf{v},\nabla\times\mathbf{v})_0 + 2\,(\lambda_j,\mathbf{v}\cdot\mathbf{v})_0\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_2\|\nabla\times\mathbf{v}\|_0^2 - 2K_2q_0|\,(\mathbf{v},\nabla\times\mathbf{v})_0\,|\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_2\|\nabla\times\mathbf{v}\|_0^2 - 2K_2q_0\|\mathbf{v}\|_0\|\nabla\times\mathbf{v}\|_0\\ &\geq K_1\|\nabla\cdot\mathbf{v}\|_0^2 + K_2\|\nabla\times\mathbf{v}\|_0^2 - K_2q_0(\|\mathbf{v}\|_0^2 + \|\nabla\times\mathbf{v}\|_0^2).\end{aligned}$$

By choosing $C_6 = \min\{K_1 - K_2q_0C_4, K_2(1 - q_0C_4 - q_0)\} > 0$ (the positivity comes from the assumptions) and $\alpha_0 = C_6/C_2$, we obtain the desired coercivity

$$\mathfrak{a}(\mathbf{v},\mathbf{v}) \geq \alpha_0\|\mathbf{v}\|_1^2 \quad \forall\mathbf{v}\in\mathbf{H}_0^1(\Omega),$$

as stated in (2.2.0.7). □

So far, the coercivity of the bilinear form $\mathfrak{a}(\cdot,\cdot)$ has been shown for all functions in $\mathbf{H}_0^1(\Omega)$. Discrete coercivity follows if a conforming finite element for the director space is chosen.

The boundedness of the bilinear form $\mathfrak{a}(\cdot,\cdot)$ and the right-hand side functionals $\mathfrak{f}(\cdot)$ and $\mathfrak{g}(\cdot)$ can be obtained directly by following the proofs in [Adl+15b]. Hence, we omit the details here.

It remains to consider the discrete inf-sup condition of the bilinear form $\mathfrak{b}(\cdot,\cdot)$ for a finite element pair $V_h$-$Q_h$, i.e., whether there exists a constant $C$ such that

$$\sup_{\mathbf{u}_h\in V_h\setminus\{0\}} \frac{\mathfrak{b}(\mathbf{u}_h,\mu_h)}{\|\mathbf{u}_h\|} \geq C\|\mu_h\| \quad \forall\mu_h\in Q_h.$$

The continuous inf-sup condition was shown in [Eme15, Appendix B] and [HTW09, Theorem 3.1]. However, the discrete inf-sup condition is not inherited from the continuous problem. Some previous works have succeeded in obtaining a discrete inf-sup condition for some specific discretisations. A discrete inf-sup condition was proven for the $([\mathbb{Q}_k]^d \oplus \mathbb{B}_F)$-$\mathbb{Q}_0$ element on quadrilaterals in [Eme15, Lemma 2.5.14] and [Adl+15b, Lemma 3.12]. The discrete inf-sup condition for the $[\mathbb{P}_1]^2$-$\mathbb{P}_1$ discretisation is shown in [HTW09, Theorem 4.5], where the analysis is only valid for the two-dimensional case due to the use of some special inverse inequalities. It

is straightforward to deduce that an enrichment of $V_h$ still guarantees the stability of the discretisation, and thus $[\mathbb{P}_2]^2$-$\mathbb{P}_1$ is inf-sup stable under the same conditions. In three dimensions, there is not yet a discussion about the inf-sup stability of the finite element pair $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ for the bilinear form $\mathfrak{b}(\cdot,\cdot)$, however, we can observe that it is inf-sup stable at least in our numerical experiments in Chapter 4.

We now consider the matrix form of the saddle-point system (2.2.0.5) after discretisation. The coercivity of the bilinear form $\mathfrak{a}(\cdot,\cdot)$ implies the invertibility of the coefficient matrix $\mathbf{A}$ and the discrete inf-sup condition indicates that $\mathbf{B}$ has full row rank. We use the full block factorisation preconditioner

$$\mathcal{Q}^{-1} = \begin{bmatrix} \mathbf{I} & -\tilde{\mathbf{A}}^{-1}\mathbf{B}^\top \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}^{-1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}\tilde{\mathbf{A}}^{-1} & \mathbf{I} \end{bmatrix}$$

with approximate inner solves $\tilde{\mathbf{A}}^{-1}$ and $\tilde{\mathbf{S}}^{-1}$ for the director block and the Schur complement $\mathbf{S} = -\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$, respectively, for solving the saddle-point problem (2.2.0.5). With exact inner solves, this is an exact inverse. With this strategy, solving the original saddle-point problem (2.2.0.5) reduces to solving two smaller linear systems involving $\mathbf{A}$ and $\mathbf{S}$. Even though $\mathbf{A}$ is sparse, its inverse is generally dense, making it impractical to store $\mathbf{S}$ explicitly. In this situation, developing a fast solver for $\mathbf{A}$ is tractable while approximating $\mathbf{S}$ becomes difficult. We will return to this issue in Section 2.3.2 and Chapter 3.

## 2.3 Augmented Lagrangian form

In the previous section, we have considered the method of Lagrange multipliers to enforce the unit-length constraint. We now introduce one of the most famous and successful algorithms, as described in many text books, e.g., [FG83; NW99], for solving constrained optimisation problems: the augmented Lagrangian method. It can be interpreted as the combination of the pure penalty method and the method of Lagrange multipliers. The AL procedure is to transform the constrained minimisation problem into an unconstrained one by introducing a Lagrange multiplier $\lambda \in L^2(\Omega)$ and adding a term (to its Lagrangian) that penalises the constraint. Instead of solving the constrained problem, we seek the equilibrium

of the unconstrained minimisation problem. In this section, we utilise the AL stabilisation strategy and accordingly modify the discrete Newton-linearised system to control the Schur complement.

### 2.3.1 Penalising the constraint

Consider penalising the continuous form of the nonlinear constraint $\mathbf{n} \cdot \mathbf{n} = 1$ in the AL algorithm, then we obtain the associated Lagrangian

$$\tilde{\mathcal{L}}(\mathbf{n}, \lambda) = \mathcal{L}(\mathbf{n}, \lambda) + \frac{\gamma}{2} \left(\mathbf{n} \cdot \mathbf{n} - 1, \mathbf{n} \cdot \mathbf{n} - 1\right)_0 \tag{2.3.1.1}$$

with penalty parameter $\gamma \geq 0$. The weak form of the associated first-order optimality conditions is to find $(\mathbf{n}, \lambda) \in \mathbf{H}^1_b(\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} \tilde{\mathcal{L}}_{\mathbf{n}}[\mathbf{v}] &= \mathcal{L}_{\mathbf{n}}[\mathbf{v}] + 2\gamma \left(\mathbf{n} \cdot \mathbf{n} - 1, \mathbf{n} \cdot \mathbf{v}\right)_0 = 0 \qquad && \forall \mathbf{v} \in \mathbf{H}^1_0(\Omega), \\ \tilde{\mathcal{L}}_{\lambda}[\mu] &= \mathcal{L}_{\lambda}[\mu] = (\mu, \mathbf{n} \cdot \mathbf{n} - 1)_0 = 0 && \forall \mu \in L^2(\Omega). \end{aligned}$$

The Newton linearisation at a given approximation $(\mathbf{n}_j, \lambda_j)$ yields a system of the form:

$$\begin{bmatrix} \tilde{\mathcal{L}}_{\mathbf{nn}} & \mathcal{L}_{\mathbf{n}\lambda} \\ \mathcal{L}_{\lambda\mathbf{n}} & 0 \end{bmatrix} \begin{bmatrix} \delta\mathbf{n} \\ \delta\lambda \end{bmatrix} = - \begin{bmatrix} \tilde{\mathcal{L}}_{\mathbf{n}} \\ \mathcal{L}_{\lambda} \end{bmatrix}.$$

Thus, we have to solve the augmented discrete variational problem:

$$\begin{aligned} \mathfrak{a}^c(\delta\mathbf{n}_h, \mathbf{v}_h) + \mathfrak{b}(\mathbf{v}_h, \delta\lambda_h) &= \mathfrak{f}^c(\mathbf{v}_h) \qquad && \forall \mathbf{v}_h \in V_h, \\ \mathfrak{b}(\delta\mathbf{n}_h, \mu_h) &= \mathfrak{g}(\mu_h) && \forall \mu_h \in Q_h, \end{aligned} \tag{2.3.1.2}$$

where

$$\mathfrak{a}^c(\mathbf{u}, \mathbf{v}) = \mathfrak{a}(\mathbf{u}, \mathbf{v}) + 4\gamma \left(\mathbf{n}_j \cdot \mathbf{u}, \mathbf{n}_j \cdot \mathbf{v}\right)_0 + 2\gamma \left(\mathbf{n}_j \cdot \mathbf{n}_j - 1, \mathbf{u} \cdot \mathbf{v}\right)_0,$$

and

$$\mathfrak{f}^c(\mathbf{v}) = \mathfrak{f}(\mathbf{v}) - 2\gamma \left(\mathbf{n}_j \cdot \mathbf{n}_j - 1, \mathbf{n}_j \cdot \mathbf{v}\right)_0.$$

Comparing (2.3.1.2) to the original system (2.2.0.6), only the bilinear form $\mathfrak{a}(\cdot, \cdot)$ and the right-hand side functional $\mathfrak{f}(\cdot)$ have changed. The boundedness of $\mathfrak{f}^c(\cdot)$ follows straightforwardly via the Cauchy–Schwarz inequality. As for the coercivity of $\mathfrak{a}^c(\cdot, \cdot)$, an additional assumption on the penalty parameter $\gamma$ is needed.

**Lemma 2.5.** *(Continuous coercivity) Let $\alpha_0 > 0$ be the coercivity constant of $\mathfrak{a}(\cdot,\cdot)$. If $\alpha_0 > 2\gamma|\alpha - 1|$ with $0 < \alpha \leq 1 \leq \beta$ satisfying $\alpha \leq |\mathbf{n}_j|^2 \leq \beta$, there exists a $\beta_0 > 0$ such that*

$$\mathfrak{a}^c(\mathbf{v},\mathbf{v}) \geq \beta_0 \|\mathbf{v}\|_1^2 \quad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega).$$

*Proof.* Note that

$$\begin{aligned}\mathfrak{a}^c(\mathbf{v},\mathbf{v}) &= \mathfrak{a}(\mathbf{v},\mathbf{v}) + 4\gamma\|\mathbf{n}_j\cdot\mathbf{v}\|_0^2 + 2\gamma\left(\mathbf{n}_j\cdot\mathbf{n}_j - 1, \mathbf{v}\cdot\mathbf{v}\right)_0 \\ &\geq \mathfrak{a}(\mathbf{v},\mathbf{v}) + 2\gamma\left(\mathbf{n}_j\cdot\mathbf{n}_j - 1, \mathbf{v}\cdot\mathbf{v}\right)_0.\end{aligned}$$

By the assumption that $\mathfrak{a}(\mathbf{v},\mathbf{v}) \geq \alpha_0\|\mathbf{v}\|_1^2$ for some $\alpha_0 > 0$, we have

$$\mathfrak{a}^c(\mathbf{v},\mathbf{v}) \geq \alpha_0\|\mathbf{v}\|_1^2 + 2\gamma\left(\mathbf{n}_j\cdot\mathbf{n}_j - 1, \mathbf{v}\cdot\mathbf{v}\right)_0.$$

Moreover, since $\mathbf{n}_j\cdot\mathbf{n}_j \geq \alpha$ and $\alpha - 1 \leq 0$, we get

$$2\gamma\left(\mathbf{n}_j\cdot\mathbf{n}_j - 1, \mathbf{v}\cdot\mathbf{v}\right)_0 \geq 2\gamma(\alpha-1)\|\mathbf{v}\|_0^2 \geq 2\gamma(\alpha-1)\|\mathbf{v}\|_1^2.$$

Thus, by taking $\beta_0 = \alpha_0 - 2\gamma|\alpha - 1| > 0$, we obtain the desired coercivity property. □

The condition $\alpha_0 > 2\gamma|\alpha - 1|$ in Lemma 2.5 indicates a limit on the value of $\gamma$ to ensure the solvability of the augmented system (2.3.1.2). However, it is desirable to use large values of $\gamma$ to achieve better control of the Schur complement as we shall see in Chapter 4. We therefore choose to employ a Picard iteration to solve the nonlinear problem, omitting the term $2\gamma\left(\mathbf{n}_j\cdot\mathbf{n}_j - 1, \mathbf{v}\cdot\mathbf{v}\right)_0$ from the linearised equations. This yields the linearised problem: find $(\delta\mathbf{n}_h, \delta\lambda_h) \in V_h \times Q_h$ such that

$$\begin{aligned}\mathfrak{a}^m(\delta\mathbf{n}_h, \mathbf{v}_h) + \mathfrak{b}(\mathbf{v}_h, \delta\lambda_h) &= \mathfrak{f}^c(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h, \\ \mathfrak{b}(\delta\mathbf{n}_h, \mu_h) &= \mathfrak{g}(\mu_h) \qquad \forall \mu_h \in Q_h,\end{aligned} \tag{2.3.1.3}$$

with the modified bilinear form

$$\mathfrak{a}^m(\mathbf{u},\mathbf{v}) = \mathfrak{a}(\mathbf{u},\mathbf{v}) + 4\gamma\left(\mathbf{n}_j\cdot\mathbf{u}, \mathbf{n}_j\cdot\mathbf{v}\right)_0 \tag{2.3.1.4}$$

to be solved at each nonlinear iteration. This ensures that the $(1,1)$-block is coercive with a coercivity constant independent of $\gamma$. Moreover, in contrast to the situation with the Navier–Stokes equations, numerical experiments indicate that the use

of this Picard iteration requires *fewer* nonlinear iterations to converge to a given tolerance than using the full Newton linearisation (see Section 4.2.1).

The corresponding matrix form of the variational problem (2.3.1.3) becomes

$$\begin{bmatrix} \mathbf{A} + \gamma \mathbf{A}_* & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} U \\ X \end{bmatrix} = \begin{bmatrix} \mathbf{f} + \gamma \mathbf{l} \\ \mathbf{g} \end{bmatrix}, \tag{2.3.1.5}$$

where $\mathbf{A}_*$ is the assembly of $4\left(\mathbf{n}_j \cdot \mathbf{u}, \mathbf{n}_j \cdot \mathbf{v}\right)_0$ and $\mathbf{l}$ denotes the assembly of $-2\left(\mathbf{n}_j \cdot \mathbf{n}_j - 1, \mathbf{n}_j \cdot \mathbf{v}\right)_0$. Note that compared to the non-augmented version (2.2.0.5), the $(1,1)$ block in (2.3.1.5) has an additional semi-definite term $\gamma \mathbf{A}_*$ with a large coefficient $\gamma$. Its sparsity pattern remains unchanged. We will construct a robust multigrid method to solve this top-left block in Chapter 3.

Since the unit-length constraint is enforced exactly in (2.3.1.1), the continuous solutions to minimising both (2.3.1.1) and (2.2.0.1) are the same. However, the unit-length constraint is not enforced exactly in our finite element discretisation, and hence this AL stabilisation does change the computed discrete solution.

**Remark 2.7.** *When utilising the augmented Lagrangian strategy, one can apply it before discretisation or afterwards. In this part of work we consider the continuous penalisation, as it improves the enforcement of the nonlinear constraint, as shown later in Section 2.3.3. This is different to the approach considered in [BO06; FMW19] for the stationary Navier–Stokes equations, where the discrete AL stabilisation was used to yield a system that has the same discrete solution but a better Schur complement.*

### 2.3.2 Approximation to the Schur complement

The Schur complement of the augmented director block in (2.3.1.5) is given by

$$\mathbf{S}_\gamma = -\mathbf{B}\mathbf{A}_\gamma^{-1}\mathbf{B}^\top =: -\mathbf{B}(\mathbf{A} + \gamma \mathbf{A}_*)^{-1}\mathbf{B}^\top.$$

We now proceed to analyse this Schur complement by following similar techniques to those of [HR12, §4]. We will show that $\mathbf{A}_*$ is equal to the matrix arising from the *discrete* AL stabilisation (which controls the Schur complement) plus a perturbation that vanishes as the mesh is refined.

Let $\Pi_{Q_h} : L^2(\Omega) \to Q_h$ ($Q_h$ is a finite dimensional approximation space of $L^2(\Omega)$) be the orthogonal $L^2$ projection operator, i.e., there holds for $p \in L^2(Omega)$ that

$$(p - \Pi_{Q_h} p, q)_0 = 0 \quad \forall q \in Q_h.$$

We define the fluctuation operator $\mathfrak{F} \coloneqq \mathcal{I} - \Pi_{Q_h}$ where $\mathcal{I} : L^2(\Omega) \to L^2(\Omega)$ is the identity mapping. Therefore, one has

$$(\mathfrak{F}(p), q)_0 = 0 \quad \forall q \in Q_h.$$

For $\mathbf{u}_h, \mathbf{v}_h \in V_h$, one can split the term $4\,(\mathbf{n}_j \cdot \mathbf{u}_h, \mathbf{n}_j \cdot \mathbf{v})_0$ into the following terms using the properties of $\mathfrak{F}$ and $\Pi_{Q_h}$:

$$\begin{aligned}
4\,(\mathbf{n}_j \cdot \mathbf{u}, \mathbf{n}_j \cdot \mathbf{v})_0 &= (\Pi_{Q_h}(2\mathbf{n}_j \cdot \mathbf{n}), 2\mathbf{n}_j \cdot \mathbf{v})_0 + (\mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{u}), 2\mathbf{n}_j \cdot \mathbf{v})_0 \\
&= (\Pi_{Q_h}(2\mathbf{n}_j \cdot \mathbf{n}), (\Pi_{Q_h} + \mathfrak{F})(2\mathbf{n}_j \cdot \mathbf{v}))_0 + (\mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{u}), (\Pi_{Q_h} + \mathfrak{F})(2\mathbf{n}_j \cdot \mathbf{v}))_0 \\
&= (\Pi_{Q_h}(2\mathbf{n}_j \cdot \mathbf{u}), \Pi_{Q_h}(2\mathbf{n}_j \cdot \mathbf{v}))_0 + (\mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{u}), \mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{v}))_0 \,.
\end{aligned}$$

Note here that the assembly of the first term is $\mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B}$, where $\mathbf{M}_\lambda$ is the mass matrix associated with the finite element space $Q_h$ for the multiplier. This can then be readily used with the Sherman–Morrison–Woodbury formula to derive an approximation of the Schur complement. Moreover, the second term $(\mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{u}), \mathfrak{F}(2\mathbf{n}_j \cdot \mathbf{v}))_0$ in fact characterises the difference between $\mathbf{A}_*$ and $\mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B}$, since the assembly of $4\,(\mathbf{n}_j \cdot \mathbf{u}, \mathbf{n}_j \cdot \mathbf{v})_0$ is $\mathbf{A}_*$. The next result (see Theorem 2.6) shows that such difference vanishes as the mesh size $h \to 0$ and thus, in this limit, the tractable term $\mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B}$ dominates $\mathbf{A}_*$.

**Theorem 2.6.** *Let $(\delta\mathbf{n}_h, \delta\lambda_h) \in V_h \times Q_h$ be the solution of the augmented discrete system (2.3.1.3) with corresponding degrees of freedom $(U, X) \in \mathbb{R}^n \times \mathbb{R}^m$. Assume that $\|\delta\mathbf{n}_h\|_1$ is bounded as $h \to 0$. Then, for the Newton linearisation at a given approximation $(\mathbf{n}_j, \lambda_j)$ satisfying $\alpha \le |\mathbf{n}_j|^2 \le \beta$ with $0 < \alpha \le 1 \le \beta$ and $|\nabla \mathbf{n}_j|$ bounded pointwise a.e., we have*

$$\left\| \left( \mathbf{A}_* - \mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B} \right) U \right\|_{\mathbb{R}^n} \lesssim h^{1+\frac{d}{2}} \|\delta\mathbf{n}_h\|_1,$$

*where $\|\cdot\|_{\mathbb{R}^n}$ denotes the Euclidean norm.*

*Proof.* Assuming $\mathbf{v}_h \in V_h$ and using the basis representations in $V_h = \operatorname{span}\{\varphi_i\}$ for $\delta\mathbf{n}_h$ and $\mathbf{v}_h$:

$$\delta\mathbf{n}_h = \sum_{i=1}^{n} U_i\varphi_i, \quad \mathbf{v}_h = \sum_{i=1}^{n} Y_i\varphi_i,$$

we obtain

$$\begin{aligned}
\left\|\left(\mathbf{A}_* - \mathbf{B}^\top\mathbf{M}_\lambda^{-1}\mathbf{B}\right)U\right\|_{\mathbb{R}^n} &= \sup_{\|Y\|_{\mathbb{R}^n}=1} Y^\top\left(\mathbf{A}_* - \mathbf{B}^\top\mathbf{M}_\lambda^{-1}\mathbf{B}\right)U \\
&= \sup_{\substack{\mathbf{v}_h=\sum_{i=1}^n Y_i\varphi_i \\ \|Y\|_{\mathbb{R}^n}=1}} \left(\mathfrak{F}(2\mathbf{n}_j\cdot\delta\mathbf{n}_h), \mathfrak{F}(2\mathbf{n}_j\cdot\mathbf{v}_h)\right)_0 \\
&\le \sup_{\substack{\mathbf{v}_h=\sum_{i=1}^n Y_i\varphi_i \\ \|Y\|_{\mathbb{R}^n}=1}} \|\mathfrak{F}(2\mathbf{n}_j\cdot\delta\mathbf{n}_h)\|_0\|\mathfrak{F}(2\mathbf{n}_j\cdot\mathbf{v}_h)\|_0 \\
&\le \underbrace{\|\mathfrak{F}\|}_{G_1} \underbrace{\sup_{\substack{\mathbf{v}_h=\sum_{i=1}^n Y_i\varphi_i \\ \|Y\|_{\mathbb{R}^n}=1}} \|2\mathbf{n}_j\cdot\mathbf{v}_h\|_0}_{G_2} \underbrace{\|\mathfrak{F}(2\mathbf{n}_j\cdot\delta\mathbf{n}_h)\|_0}_{G_3}
\end{aligned}$$

by applying the Cauchy–Schwarz inequality.

One readily sees that $G_1 \le C_1$ for a certain constant $C_1$ from the continuity of $\mathfrak{F}$. Furthermore, we write

$$G_2 = \sup_{\mathbf{v}_h=\sum_{i=1}^n Y_i\varphi_i} \frac{\|2\mathbf{n}_j\cdot\mathbf{v}_h\|_0}{\|Y\|_{\mathbb{R}^n}}.$$

Note that [KA00, Theorem 3.43] as used in [HR12] gives the relation between the discrete vector $Y$ and its associated continuous function $\mathbf{v}_h$:

$$\|Y\|_{\mathbb{R}^n} \ge C_r h^{-\frac{d}{2}}\|\mathbf{v}_h\|_0,$$

for some $C_r > 0$. Then with the fact that $\mathbf{n}_j$ is bounded we have

$$G_2 \le \sup_{\mathbf{v}_h} \frac{\|2\mathbf{n}_j\cdot\mathbf{v}_h\|_0}{C_r h^{-\frac{d}{2}}\|\mathbf{v}_h\|_0} \le C_2 h^{\frac{d}{2}}.$$

Moreover, [Clé75, Theorem 1] implies

$$\|\mathfrak{F}(p)\|_0 = \|p - \Pi_{Q_h}p\|_0 \le C_4 h\|p\|_1 \quad \forall p \in H^1(\Omega),$$

and we can deduce the following $L^2$-projection error estimate

$$G_3 = \|\mathfrak{F}(2\mathbf{n}_j\cdot\delta\mathbf{n}_h)\|_0 \le C_4 h\|2\mathbf{n}_j\cdot\delta\mathbf{n}_h\|_1 \le C_3 h\|\delta\mathbf{n}_h\|_1.$$

Note here we have used the pointwise boundedness of $\mathbf{n}_j, \nabla \mathbf{n}_j$ a.e. and the fact that $\delta \mathbf{n}_h \in V_h \subset H^1(\Omega)$.

Combining these estimates regarding $G_1, G_2, G_3$, we find

$$\left\| \left( \mathbf{A}_* - \mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B} \right) U \right\|_{\mathbb{R}^n} \lesssim h^{1+\frac{d}{2}} \|\delta \mathbf{n}_h\|_1 \to 0 \quad \text{as } h \to 0.$$

The proof is complete. □

This result suggests the use of the algebraic approximation

$$\mathbf{S}_\gamma \approx -\mathbf{B} \left( \mathbf{A} + \gamma \mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B} \right)^{-1} \mathbf{B}^\top. \tag{2.3.2.1}$$

The reason for doing so is that we can straightforwardly calculate the inverse (note the solver requires the action of $\mathbf{S}_\gamma^{-1}$, i.e., solving linear systems involving $\mathbf{S}_\gamma$) of this approximation (2.3.2.1) by the Sherman–Morrison–Woodbury formula as shown in the following Lemma 2.7.

**Lemma 2.7.** *The Schur complement approximation satisfies*

$$\mathbf{S}_\gamma^{-1} = \mathbf{S}^{-1} - \gamma \mathbf{M}_\lambda^{-1}. \tag{2.3.2.2}$$

*Proof.* Recalling the Sherman–Morrison–Woodbury formula [Hag89]: for matrices $\mathbf{E}$, $\mathbf{U}_1$, $\mathbf{P}$ and $\mathbf{U}_2$ where $\mathbf{E}$, $\mathbf{P}$ and $\mathbf{P}^{-1} + \mathbf{U}_2 \mathbf{E}^{-1} \mathbf{U}_1$ are invertible, it holds that

$$(\mathbf{E} + \mathbf{U}_1 \mathbf{P} \mathbf{U}_2)^{-1} = \mathbf{E}^{-1} - \mathbf{E}^{-1} \mathbf{U}_1 \left( \mathbf{P}^{-1} + \mathbf{U}_2 \mathbf{E}^{-1} \mathbf{U}_1 \right)^{-1} \mathbf{U}_2 \mathbf{E}^{-1}. \tag{2.3.2.3}$$

We now apply this formula twice to obtain

$$\begin{aligned}
\mathbf{S}_\gamma^{-1} &= \left( -\mathbf{B} \left( \mathbf{A} + \gamma \mathbf{B}^\top \mathbf{M}_\lambda^{-1} \mathbf{B} \right)^{-1} \mathbf{B}^\top \right)^{-1} \\
&= -\left( \mathbf{B} \left( \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B}^\top \left( \frac{1}{\gamma} \mathbf{M}_\lambda + \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top \right)^{-1} \mathbf{B} \mathbf{A}^{-1} \right) \mathbf{B}^\top \right)^{-1} \quad \text{by (2.3.2.3)}, \\
&= -\left( \underbrace{\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top}_{-\mathbf{S}} - \underbrace{\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top}_{-\mathbf{S}} \left( \frac{1}{\gamma} \mathbf{M}_\lambda + \underbrace{\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top}_{-\mathbf{S}} \right)^{-1} \underbrace{\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top}_{-\mathbf{S}} \right)^{-1} \\
&= \left( \mathbf{S} + \mathbf{S} \left( \frac{1}{\gamma} \mathbf{M}_\lambda - \mathbf{S} \right)^{-1} \mathbf{S} \right)^{-1}
\end{aligned}$$

$$= \mathbf{S}^{-1} - \mathbf{S}^{-1}\mathbf{S}\left(\frac{1}{\gamma}\mathbf{M}_\lambda - \mathbf{S} + \mathbf{S}\mathbf{S}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}\mathbf{S}^{-1} \qquad \text{by (2.3.2.3)},$$
$$= \mathbf{S}^{-1} - \gamma\mathbf{M}_\lambda^{-1}.$$

This completes the proof. □

Induced from the above result (2.3.2.2) for the inverse of the Schur complement approximation, a simple and effective approach for large $\gamma$ is to employ the approximation

$$\mathbf{S}_\gamma^{-1} \approx -\gamma\mathbf{M}_\lambda^{-1}. \tag{2.3.2.4}$$

On the infinite-dimensional level, the effect of the augmented Lagrangian term is to make $-\gamma^{-1}\mathcal{I}$ ($\mathcal{I}$ the identity operator on the multiplier space) an effective approximation for the Schur complement [PT74, Lemma 3]. When discretised, this indicates that the weighted multiplier mass matrix $-\gamma^{-1}\mathbf{M}_\lambda$ will be an effective approximation for $\mathbf{S}_\gamma$, with the approximation improving as $\gamma \to \infty$.

In fact, the approximation of the inverse of the discretely augmented Schur complement (2.3.2.4) can be improved further by combining $-\gamma\mathbf{M}_\lambda^{-1}$ with a good approximation of the unaugmented Schur complement $\mathbf{S}$ [HVK18]. Given an approximation $\tilde{\mathbf{S}}$ of $\mathbf{S}$, we employ

$$\mathbf{S}_\gamma^{-1} \approx \tilde{\mathbf{S}}_\gamma^{-1} = \tilde{\mathbf{S}}^{-1} - \gamma\mathbf{M}_\lambda^{-1}. \tag{2.3.2.5}$$

It is therefore of interest to consider the Schur complement of the unaugmented problem. In the context of the Stokes equations, the Schur complement is spectrally equivalent to the viscosity-weighted pressure mass matrix [SW94; WS91; ESW14]. Following similar techniques, an approximation can be obtained by proving that $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ is spectrally equivalent to $\mathbf{M}_\lambda$ for the equal-constant nematic case. This gives us good insight into the choice of $\tilde{\mathbf{S}}^{-1}$.

**Theorem 2.8.** *Assume that the finite dimensional spaces $V_h \subset \mathbf{H}_0^1(\Omega)$ and $Q_h \subset L^2(\Omega)$ are inf-sup stable. For equal-constant nematic LC problems without augmented Lagrangian stabilisation, the matrix $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\top$ arising from the Newton-linearised*

*system is spectrally equivalent to the multiplier mass matrix* $\mathbf{M}_\lambda$*, under the same assumptions as in Lemma 2.4.*

*Proof.* For the equal-constant model with Dirichlet boundary conditions $\mathbf{n} = \mathbf{n}_b \in H^{1/2}(\partial\Omega, \mathcal{S}^2)$, its corresponding Lagrangian is

$$\mathcal{L}(\mathbf{n}, \lambda) = \frac{K_c}{2} (\nabla\mathbf{n}, \nabla\mathbf{n})_0 + (\lambda, \mathbf{n}\cdot\mathbf{n} - 1)_0 .$$

After Newton linearisation and due to the inf-sup stability of the finite element pair $V_h$-$Q_h$, the discrete variational problem is to find $\delta\mathbf{n}_h \in V_h$, $\delta\lambda_h \in Q_h$ satisfying

$$\begin{aligned} K_c (\nabla\delta\mathbf{n}_h, \nabla\mathbf{v}_h)_0 + 2(\lambda_j, \delta\mathbf{n}_h\cdot\mathbf{v}_h)_0 + 2(\delta\lambda_h, \mathbf{n}_j\cdot\mathbf{v}_h)_0 & \\ = -K_c (\nabla\mathbf{n}_j\cdot\nabla\mathbf{v}_h)_0 - 2(\lambda_j, \mathbf{n}_j\cdot\mathbf{v}_h)_0 \quad & \forall\mathbf{v}_h \in V_h, \\ 2(\mu_h, \mathbf{n}_j\cdot\delta\mathbf{n}_h)_0 = -(\mu_h, \mathbf{n}_j\cdot\mathbf{n}_j - 1)_0 \quad & \forall\mu_h \in Q_h, \end{aligned}$$

where $\mathbf{n}_j$ and $\lambda_j$ represent the current approximations to $\mathbf{n}$ and $\lambda$, respectively. This can be rewritten in block matrix form as

$$\mathcal{R}\begin{bmatrix} U \\ X \end{bmatrix} := \begin{bmatrix} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{bmatrix}\begin{bmatrix} U \\ X \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix},$$

where as before $U \in \mathbb{R}^n$ and $X \in \mathbb{R}^m$ are the unknown coefficients of the discrete director update and the discrete Lagrange multiplier update with respect to the basis functions in $V_h$ and $Q_h$, and $\mathbf{A}$ denotes the symmetric form $K_c(\nabla\delta\mathbf{n}_h, \nabla\mathbf{v}_h)_0 + 2(\lambda_j, \delta\mathbf{n}_h\cdot\mathbf{v}_h)_0$. The coercivity property of the bilinear form from Lemma 2.4 ensures that $\mathbf{A}$ is positive definite.

The coefficient matrix $\mathcal{R}$ is symmetric and indefinite (resulting in $\mathcal{R}$ possessing both positive and negative eigenvalues). Moreover, $\mathcal{R}$ is non-singular if and only if $\mathbf{B}$ has full row rank, which can be deduced from the discrete inf-sup condition.

Denote

$$\begin{aligned} \|\mathbf{u}_h\|_{lc}^2 &= K_c (\nabla\mathbf{u}_h, \nabla\mathbf{u}_h)_0 + (\lambda_j, 2\mathbf{u}_h\cdot\mathbf{u}_h)_0 , \\ \|\mu_h\|_0^2 &= (\mu_h, \mu_h)_0 . \end{aligned}$$

Notice that the validity of the first norm follows from the assumed pointwise non-negativity of $\lambda_j$.

For a stable mixed finite element, from the inf-sup condition, there exists a positive constant $C$ independent of the mesh size $h$ such that

$$\sup_{\mathbf{u}_h \in V_h \backslash \{0\}} \frac{(\mu_h, 2\mathbf{n}_j \cdot \mathbf{u}_h)_0}{\|\mathbf{u}_h\|_{lc}} \geq C\|\mu_h\|_0 \quad \forall \mu_h \in Q_h,$$

leading to its matrix form

$$\max_{U \in \mathbb{R}^n \backslash \{0\}} \frac{X^\top \mathbf{B} U}{[U^\top \mathbf{A} U]^{1/2}} \geq C[X^\top \mathbf{M}_\lambda X]^{1/2} \quad \forall X \in \mathbb{R}^m.$$

Thus, we have

$$\begin{aligned} C[X^\top \mathbf{M}_\lambda X]^{1/2} &\leq \max_{U \in \mathbb{R}^n \backslash \{0\}} \frac{X^\top \mathbf{B} U}{[U^\top \mathbf{A} U]^{1/2}} \\ &= \max_{\mathbf{z} = A^{1/2} U \neq 0} \frac{X^\top \mathbf{B} \mathbf{A}^{-1/2} \mathbf{z}}{[\mathbf{z}^\top \mathbf{z}]^{1/2}} \\ &= (X^\top \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top X)^{1/2} \quad \forall X \in \mathbb{R}^m, \end{aligned}$$

where the maximum is attained at $\mathbf{z} = (X^\top \mathbf{B} \mathbf{A}^{-1/2})^\top$. It yields

$$C^2 \frac{X^\top \mathbf{M}_\lambda X}{X^\top X} \leq \frac{X^\top \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top X}{X^\top X} \quad \forall X \in \mathbb{R}^m \backslash \{0\}. \tag{2.3.2.6}$$

Regardless of the stability of the finite element pair, we can deduce from the boundedness of $\mathfrak{b}(\cdot, \cdot)$ that there exists a positive constant $C_1$ such that

$$X^\top \mathbf{B} U \leq C_1 [X^\top \mathbf{M}_\lambda X]^{1/2} [U^\top \mathbf{A} U]^{1/2} \quad \forall U \in \mathbb{R}^n, \forall X \in \mathbb{R}^m.$$

Hence,

$$\begin{aligned} C_1[X^\top \mathbf{M}_\lambda X]^{1/2} &\geq \max_{U \in \mathbb{R}^n \backslash \{0\}} \frac{X^\top \mathbf{B} U}{[U^\top \mathbf{A} U]^{1/2}} \\ &= \max_{\mathbf{z} = \mathbf{A}^{1/2} U \neq 0} \frac{X^\top \mathbf{B} \mathbf{A}^{-1/2} \mathbf{z}}{[\mathbf{z}^\top \mathbf{z}]^{1/2}} \\ &= (X^\top \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top X)^{1/2} \quad \forall X \in \mathbb{R}^m, \end{aligned}$$

where again the maximum is attained at $\mathbf{z} = (X^\top \mathbf{B} \mathbf{A}^{-1/2})^\top$. This gives rise to

$$\frac{X^\top \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top X}{X^\top \mathbf{M}_\lambda X} \leq C_1^2 \quad \forall X \in \mathbb{R}^m \backslash \{0\}. \tag{2.3.2.7}$$

Therefore for inf-sup stable finite element pairs, we have by (2.3.2.6) and (2.3.2.7)

$$C^2 \leq \frac{X^\top \mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top X}{X^\top \mathbf{M}_\lambda X} \leq C_1^2 \quad \forall X \in \mathbb{R}^m \backslash \{0\}.$$

This indicates that $\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^\top$ is spectrally equivalent to $\mathbf{M}_\lambda$. □

**Remark 2.8.** *It follows from Theorem 2.8 that $\gamma = 0$ should show mesh-independence (i.e., the average number of Flexible GMRES (abbreviated as FGMRES in the following; this allows for the use of different preconditioner in each iteration step) iterations per nonlinear iteration does not deteriorate as one refines the mesh) in the case of equal-constant nematic LC. This can be observed in subsequent numerical experiments reported in Table 4.6 (see the column where $\gamma = 0$). One should also notice that such mesh-independence for $\gamma = 0$ is also shown in Table 4.2 for the non-equal-constant case, suggesting it has use outside the context of augmented Lagrangian methods also.*

Combining Theorem 2.8 with (2.3.2.5), our final approximation for $\mathbf{S}_\gamma^{-1}$ is given by

$$\mathbf{S}_\gamma^{-1} \approx \tilde{\mathbf{S}}_\gamma^{-1} = -(1+\gamma)\mathbf{M}_\lambda^{-1}. \tag{2.3.2.8}$$

### 2.3.3 Improvement of the constraint

We have now observed that the continuous AL form introduced in Section 2.3.2 can help control the Schur complement. Another contribution of this AL stabilisation is that it improves the discrete enforcement of the constraint as we increase the value of the penalty parameter $\gamma$. An example of improving the linear divergence-free constraint in the Stokes system can be found in [Joh+17, Section 5.1]. In this section, we will use a similar strategy to show the improvement of the discrete constraint as $\gamma$ increases.

We restrict ourselves to the equal-constant case with *constant* Dirichlet boundary conditions. That is to say, we consider the Oseen–Frank model with Dirichlet boundary condition $\mathbf{n}|_{\partial\Omega} = \mathbf{n}_b$, where $\mathbf{n}_b$ is a nonzero constant vector satisfying $|\mathbf{n}_b| = 1$.

**Remark 2.9.** *One may wonder whether the solution under this assumption of boundary conditions is the constant boundary data itself, i.e., $\mathbf{n}_h = \mathbf{n}_b$ in the domain $\Omega$, and thus the unit-length constraint is actually satisfied exactly. In fact, $\mathbf{n}_h = \mathbf{n}_b$ is indeed an equilibrium of the energy minimisation problem (2.1.0.1), however, it is not the only one and not in general the one with the lowest energy value. An example supports this fact can be seen in [Eme+18], where $\mathbf{n} = (0, 0, 1)$ is strongly*

*enforced on the boundary while many computed solutions are not* $(0,0,1)$ *everywhere in the domain* $\Omega$.

We use the $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ finite element pair in this section, so both the director $\mathbf{n}$ and the Lagrange multiplier $\lambda$ are approximated by continuous piecewise-linear polynomials. For this section, we denote finite element spaces for the director and the Lagrange multiplier by $V_{h,b} \coloneqq V_h \cap \mathbf{H}_b^1(\Omega)$ and $Q_h \subset L^2(\Omega)$, respectively, and denote $V_{h,0} = V_h \cap \mathbf{H}_0^1(\Omega)$.

We restate the associated nonlinear discrete variational problem as follows: find $(\mathbf{n}_h, \lambda_h) \in V_{h,b} \times Q_h$ such that

$$\begin{aligned} &K_c\left(\nabla \mathbf{n}_h, \nabla \mathbf{v}_h\right)_0 + K_c q_0\left(\mathbf{v}_h, \nabla \times \mathbf{n}_h\right)_0 + K_c q_0\left(\mathbf{n}_h, \nabla \times \mathbf{v}_h\right)_0 \\ &\quad + 2\left(\lambda_h, \mathbf{n}_h \cdot \mathbf{v}_h\right)_0 + 2\gamma\left(\mathbf{n}_h \cdot \mathbf{n}_h - 1, \mathbf{n}_h \cdot \mathbf{v}_h\right)_0 = 0 \quad \forall \mathbf{v}_h \in V_{h,0}, \end{aligned} \tag{2.3.3.1a}$$

$$\left(\mu_h, \mathbf{n}_h \cdot \mathbf{n}_h - 1\right)_0 = 0 \quad \forall \mu_h \in Q_h. \tag{2.3.3.1b}$$

Take the test function $\mathbf{v}_h = \mathbf{n}_h - \mathbf{n}_b \in V_{h,0}$ in (2.3.3.1a) to obtain

$$\begin{aligned} &K_c\|\nabla \mathbf{n}_h\|_0^2 + 2K_c q_0\left(\mathbf{n}_h, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, \mathbf{n}_h \cdot \mathbf{n}_h\right)_0 + 2\gamma\left(\mathbf{n}_h \cdot \mathbf{n}_h - 1, \mathbf{n}_h \cdot \mathbf{n}_h\right)_0 \\ &\quad = K_c q_0\left(\mathbf{n}_b, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, \mathbf{n}_h \cdot \mathbf{n}_b\right)_0 + 2\gamma\left(\mathbf{n}_h \cdot \mathbf{n}_h - 1, \mathbf{n}_h \cdot \mathbf{n}_b\right)_0. \end{aligned} \tag{2.3.3.2}$$

Note that in this step we have used the fact that since $\mathbf{n}_b$ is a constant vector, its derivative is zero.

As (2.3.3.1b) is valid for arbitrary $\mu_h \in Q_h$ and one can easily verify that $\mathbf{n}_h \cdot \mathbf{n}_b \in Q_h$, we have

$$\left(\mathbf{n}_h \cdot \mathbf{n}_b, \mathbf{n}_h \cdot \mathbf{n}_h - 1\right)_0 = 0.$$

Then taking $\mu_h = 1$ and $\mu_h = \lambda_h$ leads to

$$\left(1, \mathbf{n}_h \cdot \mathbf{n}_h - 1\right)_0 = 0 \text{ and } \left(\lambda_h, \mathbf{n}_h \cdot \mathbf{n}_h - 1\right)_0 = 0,$$

respectively. Thus, (2.3.3.2) collapses to

$$\begin{aligned} &K_c\|\nabla \mathbf{n}_h\|_0^2 + 2K_c q_0\left(\mathbf{n}_h, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, 1\right)_0 + 2\gamma\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0^2 \\ &\quad = K_c q_0\left(\mathbf{n}_b, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, \mathbf{n}_h \cdot \mathbf{n}_b\right)_0. \end{aligned} \tag{2.3.3.3}$$

By the Cauchy–Schwarz and Hölder inequalities, we observe an upper bound for the right-hand side of (2.3.3.3):

$$\begin{aligned} K_c q_0 \left(\mathbf{n}_b, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, \mathbf{n}_h \cdot \mathbf{n}_b\right)_0 &\leq K_c q_0 \|\nabla \times \mathbf{n}_h\|_0 + 2\|\lambda_h\|_0 \|\mathbf{n}_h\|_0 \\ &\leq \frac{K_c q_0}{2} + \frac{K_c q_0}{2}\|\nabla \times \mathbf{n}_h\|_0^2 + \|\lambda_h\|_0^2 + \|\mathbf{n}_h\|_0^2. \end{aligned} \tag{2.3.3.4}$$

Meanwhile, the left-hand side of (2.3.3.3) can be bounded from below:

$$\begin{aligned} &K_c\|\nabla \mathbf{n}_h\|_0^2 + 2K_c q_0 \left(\mathbf{n}_h, \nabla \times \mathbf{n}_h\right)_0 + 2\left(\lambda_h, 1\right)_0 + 2\gamma\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0^2 \\ &\quad \geq K_c\|\nabla \mathbf{n}_h\|_0^2 - 2K_c q_0 |\left(\mathbf{n}_h, \nabla \times \mathbf{n}_h\right)_0| - 2|\left(\lambda_h, 1\right)_0| + 2\gamma\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0^2 \\ &\quad \geq K_c\|\nabla \mathbf{n}_h\|_0^2 - K_c q_0\|\mathbf{n}_h\|_0^2 - K_c q_0\|\nabla \times \mathbf{n}_h\|_0^2 - \|\lambda_h\|_0^2 - |\Omega| + 2\gamma\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0^2, \end{aligned} \tag{2.3.3.5}$$

where $|\Omega|$ denotes the measure of the domain $\Omega$.

Hence, by combining (2.3.3.4) and (2.3.3.5), we have

$$\begin{aligned} K_c\|\nabla \mathbf{n}_h\|_0^2 - (K_c q_0 + 1)\|\mathbf{n}_h\|_0^2 - \frac{3}{2}K_c q_0\|\nabla \times \mathbf{n}_h\|_0^2 \\ - \|\lambda_h\|_0^2 + 2\gamma\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0^2 \leq \frac{K_c q_0}{2} + |\Omega|. \end{aligned} \tag{2.3.3.6}$$

Note that the right-hand side of (2.3.3.6) is a fixed constant independent of $\gamma$ and those negative terms on the left-hand side actually depends on $\gamma$ since both $\mathbf{n}_h$ and $\lambda_h$ depends on $\gamma$. Therefore, taking $\gamma$ larger value does not directly force the constraint approximation error $\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0$ to become smaller. That is to say, (2.3.3.6) does not imply that $\|\mathbf{n}_h \cdot \mathbf{n}_h - 1\|_0 \leq \mathcal{O}(\gamma^{-1/2})$. However, this improvement of the discrete constraint as $\gamma$ increases can be observed in our numerical experiments illustrated in Chapter 4.

**Remark 2.10.** *The technique shown in this section can be extended in a similar way to the multi-constant case; we omit the details here for brevity.*

## 2.4 Summary

In this chapter, we considered the Oseen–Frank model of cholesteric LC, which demands a unit-length constraint be enforced. We then applied the continuous augmented Lagrangian form for constraint penalisation and illustrated its two major

effects: the improvement of the discrete constraint $\mathbf{n}_h \cdot \mathbf{n}_h = 1$, and a better control on the Schur complement using a weighted mass matrix approximation. However, this results in a more complicated top-left block to be solved, which we will tackle by means of a robust multigrid method in the next chapter.

# 3

# A robust multigrid algorithm for the augmented director block

## Contents



As discussed in the previous chapter, the addition of the augmented Lagrangian term gives a better approximation to the Schur complement (we will see this in Tables 4.2 and 4.6). However, the tradeoff is that it complicates the solution of the top-left block $\mathbf{A}_\gamma$, as it adds a semi-definite term with a large coefficient. We demonstrate this effect in Table 3.1 where we apply the block preconditioner with the Schur complement approximation $\tilde{\mathbf{S}}_\gamma^{-1}$ as given by (2.3.2.8) and $\mathbf{A}_\gamma = \mathbf{A} + \gamma \mathbf{A}_*$ solved approximately with one V-cycle of standard geometric multigrid with Jacobi relaxation. Table 3.1 shows that the solver is neither $\gamma$- or $h$-robust. Thus, for the augmented Lagrangian strategy to be successful, we require a parameter-independent solver for the top-left block.

Fortunately, a rich literature is available to guide the development of multigrid solvers for nearly singular systems with the presence of a semi-definite term; see for

| | | $\gamma$ | | | | | |
|---|---|---|---|---|---|---|---|
| #refs | #dofs | $10^1$ | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 5,340 | 33.75(4) | 14.80(5) | 6.20(5) | 4.38(8) | 7.18(11) | 32.53(19) |
| 2 | 21,080 | 75.00(5) | 31.80(5) | 11.60(5) | 4.86(7) | 5.83(12) | 16.53(15) |
| 3 | 83,760 | >100 | 57.60(5) | 24.60(5) | 10.17(6) | 46.75(8) | >100 |
| 4 | 333,920 | >100 | >100 | 90.80(5) | 19.67(6) | >100 | >100 |

**Table 3.1:** The average number of FGMRES iterations per Newton iteration (total number of Newton iterations) for a nematic LC problem in a square slab. See the detailed problem description in Chapter 4.

instance [Sch99a; Sch99b; Lee+07]. Particularly, Schöberl's seminal paper [Sch99a] on the construction of parameter-robust multigrid schemes lists two requirements that must be satisfied for the top-left solve to be robust. The first requirement is a parameter-robust relaxation method; this is achieved by developing a space decomposition that stably captures the kernel of the semi-definite terms. The second requirement is a parameter-robust prolongation operator, i.e., one whose continuity constant is independent of the parameters. This is achieved by (approximately) mapping kernel functions on coarse grids to kernel functions on fine grids. We separately discuss both of these requirements below.

In this chapter, we will construct a parameter-robust multigrid algorithm based on these works [Sch99a; Sch99b; Lee+07]. Some extensions of the analysis from Schöberl's work [Sch99b] are given for the LC case. Then in order to verify the two aforementioned requirements of constructing the robust multigrid algorithm, a detailed example using the point-block Jacobi or star relaxation and the natural prolongation is illustrated for two-dimensional cholesteric problems.

For ease of notation, we consider the two-grid method applied to the equal-constant nematic case, and use subscripts $h$ and $H$ to distinguish fine and coarse mesh levels respectively. That is to say, $V_H$ represents the coarse-grid function space and we denote the associated operator $A_{H,\gamma} : V_H \to V_H^*$

$$(A_{H,\gamma}\mathbf{u}_H, \mathbf{v}_H)_0 \coloneqq \mathfrak{a}^m(\mathbf{u}_H, \mathbf{v}_H)$$

with approximations $\mathbf{u}_H, \mathbf{v}_H$ on $V_H$. The analysis in this chapter can be extended to more complicated cases, e.g., with non-equal constants and more than two levels of grids.

For the domain $\Omega$, we consider a non-overlapping triangulation $\mathcal{T}_H$, i.e.,

$$\cup_{T\in\mathcal{T}_H} T = \bar{\Omega} \text{ and } \operatorname{int}(T_i) \cap \operatorname{int}(T_j) = \emptyset \quad \forall T_i \neq T_j,\ T_i, T_j \in \mathcal{T}_H.$$

The fine grid $\mathcal{T}_h$ with $h = H/2$ is obtained by a regular refinement of the simplices in $\mathcal{T}_H$. In what follows we consider both the $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ and $[\mathbb{P}_2]^d$-$\mathbb{P}_1$ discretisations.

## 3.1 Relaxation

After applying the AL method introduced in Section 2.3.1, the discrete linear variational form corresponding to the top-left block $\mathbf{A}_\gamma = \mathbf{A} + \gamma \mathbf{A}_*$ is given by

$$\mathfrak{a}^m(\mathbf{u}_h, \mathbf{v}_h) = K_c \left(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h\right)_0 + 2\left(\lambda_j, \mathbf{u}_h \cdot \mathbf{v}_h\right)_0 + 4\gamma \left(\mathbf{n}_j \cdot \mathbf{u}_h, \mathbf{n}_j \cdot \mathbf{v}_h\right)_0, \quad (3.1.0.1)$$

with $\mathbf{u}_h \in V_h \subset \mathbf{H}_0^1(\Omega)$ being the trial function and $\mathbf{v}_h \in V_h$ the test function. Note that $\mathbf{n}_j$ and $\lambda_j$ are the current approximations to the director $\mathbf{n}$ and the Lagrange multiplier $\lambda$, respectively, in the Newton iteration. The first two terms of $\mathfrak{a}^m$ are symmetric and coercive because of the running assumption of uniform non-negativity of $\lambda_j$. The kernel of the semi-definite term involving $\gamma$ is

$$\mathcal{N}_h = \{\mathbf{u}_h \in V_h : \mathbf{n}_j \cdot \mathbf{u}_h = 0 \text{ a.e.}\}. \quad (3.1.0.2)$$

In the case of $\gamma$ being very large, the variational problem involving (3.1.0.1) is nearly singular and common relaxation methods like Jacobi and Gauss–Seidel will not yield effective multigrid cycles, as we explain below.

Relaxation schemes can be devised in a generic way by considering *space decompositions*

$$V_h = \sum_{i=1}^{M} V_i, \quad (3.1.0.3)$$

where the sum of vector spaces on the right is not necessarily a direct sum [Xu92]. For example, if $V_h = \operatorname{span}(\varphi_1, \ldots, \varphi_M)$, Jacobi and Gauss–Seidel iterations are induced by the space decomposition

$$V_i = \operatorname{span}(\varphi_i), \tag{3.1.0.4}$$

where the updates are performed additively for Jacobi and multiplicatively for Gauss–Seidel. One of the key insights of [Sch99a; Lee+07] was that the key requirement for parameter-robustness when applied to nearly singular problems is that the space decomposition must satisfy the *kernel-capturing property*

$$\mathcal{N}_h = \sum_{i=1}^{M} (V_i \cap \mathcal{N}_h), \tag{3.1.0.5}$$

that is, any kernel function can be written as a sum of kernel functions drawn from the subspaces. In particular, each subspace $V_i$ must be rich enough to support kernel functions; in our context, this is not satisfied by the choice (3.1.0.4), accounting for its poor behaviour shown in Table 3.1 as $\gamma \to \infty$.

In the mesh triangulation $\mathcal{T}_h$, we denote the *star* of a vertex $v_i$ as the patch of elements sharing $v_i$, i.e.,

$$\operatorname{star}(v_i) \coloneqq \bigcup_{T \in \mathcal{T}_h : v_i \in T} T.$$

This induces an associated space decomposition, called the *star patch*, by

$$V_i \coloneqq \{\mathbf{u}_h \in V_h : \operatorname{supp}(\mathbf{u}_h) \subset \operatorname{star}(v_i)\}.$$

This is illustrated in Figure 3.1 (left). We call the induced relaxation method a *star iteration*. In effect, each subspace solve solves for the degrees of freedom in the interior of the patch of cells, with homogeneous Dirichlet conditions on the boundary of the patch. Given a vertex or edge midpoint $v_i$, we denote the *point-block* patch $V_i$ as the span of the basis functions associated with degrees of freedom that evaluate a function at $v_i$ (see Figure 3.1, middle). The induced relaxation method solves for all colocated degrees of freedom simultaneously. These two space decompositions coincide for the $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ discretisation (see Figure 3.1, right).

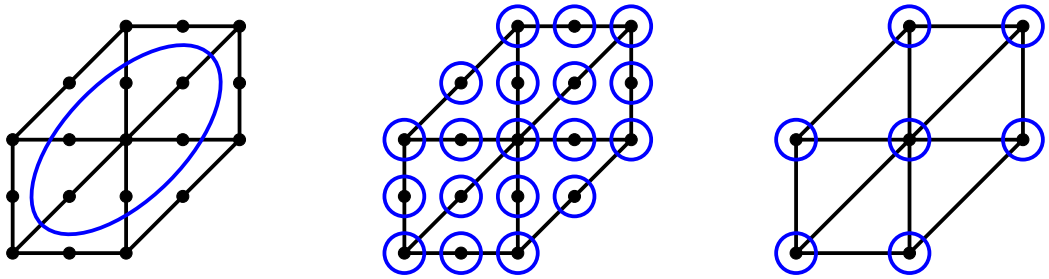


**Figure 3.1:** Illustrations of the star patch of the center vertex (left) and the point-block patch (middle) for the finite element pair $[\mathbb{P}_2]^2$-$\mathbb{P}_1$. Note that these two patches are the same for $[\mathbb{P}_1]^2$-$\mathbb{P}_1$ discretisation (right). Here, black dots represent the degrees of freedom, and the blue lines gather degrees of freedom solved for simultaneously in the relaxation.

We now briefly explain why these two decompositions approximately satisfy the kernel-capturing condition (3.1.0.5) for the finite element pair $[\mathbb{P}_1]^d$-$\mathbb{P}_1$. First, we define an approximate kernel

$$\tilde{\mathcal{N}}_h = \{\mathbf{u}_h \in V_h : \mathbf{n}_j \cdot \mathbf{u}_h = 0 \text{ on each vertex}\}. \tag{3.1.0.6}$$

Since $\mathbf{n}_j$ is the current approximation to the director $\mathbf{n}$, we have $\mathbf{n}_j \in V_h = \sum_i V_i$. We are therefore able to express $\mathbf{n}_j$ as $\mathbf{n}_j = \sum_i \mathbf{n}_j^i$, where $\mathbf{n}_j^i \in V_i$ describes the function at the vertex $v_i$. Similarly, we split $\mathbf{u}_h$ into $\mathbf{u}_h = \sum_i \mathbf{u}_h^i$ with $\mathbf{u}_h^i \in V_i$. For each vertex $v_i$, the requirement $\mathbf{u}_h \in \tilde{\mathcal{N}}_h$ yields

$$\mathbf{n}_j^i \cdot \mathbf{u}_h^i = 0 \quad \forall i. \tag{3.1.0.7}$$

The definition of $V_i$ ensures that $\mathbf{u}_h^i$ and $\mathbf{n}_j^i$ are only supported on the interior of the star of $v_i$. We deduce that on each vertex

$$\mathbf{n}_j^k \cdot \mathbf{u}_h^i = 0 \quad \forall i \neq k,$$

which yields $\sum_k \mathbf{n}_j^k \cdot \mathbf{u}_h^i = \mathbf{n}_j \cdot \mathbf{u}_h^i = 0$. Hence, $\mathbf{u}_h^i \in \tilde{\mathcal{N}}_h \forall i$ and we obtain the kernel-capturing condition (3.1.0.5) for the approximate kernel $\tilde{\mathcal{N}}_h$.

For the $[\mathbb{P}_2]^d$-$\mathbb{P}_1$ finite element pair, the satisfaction of the kernel-capturing property for the approximate kernel follows along similar lines. For the point-block patch, (3.1.0.7) still holds. The star patch uses larger subspaces, each one including multiple point-block patches, but it can be easily verified that (3.1.0.7) is still fulfilled.

### 3.1.1 Robustness analysis of the approximate kernel

While we are not able to prove the kernel capturing property for the exact kernel (3.1.0.2), we can still obtain the spectral inequalities

$$c_1 D_{h,\gamma} \leq A_{h,\gamma} \leq c_2 D_{h,\gamma}, \tag{3.1.1.1}$$

when using the approximate kernel (3.1.0.6). Here, $D_{h,\gamma}$ is the preconditioner to be specified later for the operator $A_{h,\gamma}$ and $C \leq D$ represents $\|\mathbf{u}\|_C \leq \|\mathbf{u}\|_D$ for all $\mathbf{u}$. We prove that $c_1$ depends on $\gamma$, but the dependence can be well controlled so that the preconditioner is not badly affected by varying $\gamma$, while $c_2$ is always independent of $\gamma$. For simplicity, we prove the case for the equal-constant nematic case with the $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ discretisation; extensions to the non-equal-constant cholesteric case and to the $[\mathbb{P}_2]^d$-$\mathbb{P}_1$ discretisation are possible.

We define the operator associated to $\mathfrak{a}^m$, $A_{h,\gamma} : V_h \to V_h^*$, by

$$(A_{h,\gamma}\mathbf{u}_h, \mathbf{v}_h)_0 \coloneqq \mathfrak{a}^m(\mathbf{u}_h, \mathbf{v}_h).$$

For the space decomposition $V_h = \sum_i V_i$, we denote the lifting operator (the natural inclusion) by $I_i : V_i \to V_h$ and choose the Galerkin subspace operator $A_i : V_i \to V_i$ to satisfy

$$(A_i\mathbf{u}_i, \mathbf{v}_i)_0 \coloneqq (A_{h,\gamma} I_i\mathbf{u}_i, I_i\mathbf{v}_i)_0 \quad \forall \mathbf{u}_i, \mathbf{v}_i \in V_i.$$

This implies that $A_i = I_i^* A_{h,\gamma} I_i$.

The additive Schwarz preconditioner $D_{h,\gamma}$ for a problem $A_{h,\gamma} w_h = d_h$ associated with the space decomposition (3.1.0.3) is defined by the action of its inverse [Xu92]:

$$w_h = D_{h,\gamma}^{-1} d_h$$

given by

$$w_h = \sum_{i=1}^{M} I_i w_i,$$

with $w_i \in V_i$ being the unique solution of

$$(A_i w_i, v_i)_0 = (d_h, I_i v_i)_0 \quad \forall v_i \in V_i.$$

Hence, we can rewrite the preconditioning operator $D_{h,\gamma}^{-1}$ in operator form as

$$D_{h,\gamma}^{-1} = \sum_{i=1}^{M} I_i A_i^{-1} I_i^*.$$

We now state for completeness a classical result in the analysis of additive Schwarz preconditioners, see e.g. [Sch99b, Theorem 3.1] and the references therein.

**Theorem 3.1.** *Define the splitting norm for $\mathbf{u}_h \in V_h$ as*

$$|||\mathbf{u}_h|||^2 \coloneqq \inf_{\substack{\mathbf{u}_h = \sum_i I_i \mathbf{u}_i \\ \mathbf{u}_i \in V_i}} \sum_{i=1}^{M} \|\mathbf{u}_i\|_{A_i}^2.$$

*This splitting norm is equal to the norm $\|\mathbf{u}_h\|_{D_{h,\gamma}} \coloneqq (D_{h,\gamma}\mathbf{u}_h, \mathbf{u}_h)_0^{1/2}$ generated by the additive Schwarz preconditioner, i.e. it holds that*

$$|||\mathbf{u}_h|||^2 = \|\mathbf{u}_h\|_{D_{h,\gamma}}^2 \quad \forall \mathbf{u}_h \in V_h.$$

To build intuition, let us examine why Jacobi relaxation defined by the space decomposition (3.1.0.4) is not robust as $\gamma \to \infty$. With (3.1.0.4), the decomposition $\mathbf{u}_h = \sum_i^M \mathbf{u}_i, \mathbf{u}_i \in V_i$ is unique. It yields that

$$\begin{aligned} \|\mathbf{u}_h\|_{D_{h,\gamma}}^2 = |||\mathbf{u}_h|||^2 &= \sum_i (A_i \mathbf{u}_i, \mathbf{u}_i)_0 = \sum_i (A_{h,\gamma} \mathbf{u}_i, \mathbf{u}_i)_0 \\ &\lesssim (1+\gamma) \sum_i \|\mathbf{u}_i\|_1^2 \lesssim \frac{1+\gamma}{h^2} \sum_i \|\mathbf{u}_i\|_0^2 \lesssim \frac{1+\gamma}{h^2} \|\mathbf{u}_h\|_0^2 \\ &\lesssim \frac{1+\gamma}{h^2} \|\mathbf{u}_h\|_{A_{h,\gamma}}^2. \end{aligned} \tag{3.1.1.2}$$

Note that the bound in (3.1.1.2) is parameter-dependent and deteriorates as $\gamma \to \infty$ or $h \to 0$.

In order to deduce the robustness result for our approximate kernel (3.1.0.6), we first derive the following lemma.

**Lemma 3.2.** *Let $\mathbf{u}_0 = \sum_i^M \mathbf{u}_0^i \in \tilde{\mathcal{N}}_h$ and assume $\mathbf{n}_j \in [\mathbb{P}_1]^d$. Then it holds that*

$$\sum_i^M \|\mathbf{u}_0^i \cdot \mathbf{n}_j\|_{L^2(\Omega)}^2 \lesssim h^2 \|\mathcal{D}\mathbf{n}_j\|_{L^\infty(\Omega)}^2 \|\mathbf{u}_0\|_{L^2(\Omega)}^2,$$

*where $\mathcal{D}\mathbf{n}_j$ denotes the Jacobian matrix of $\mathbf{n}_j$.*

*Proof.* Consider the vertex $v_i$ on the boundary of an element $T$. As $\mathbf{n}_j \in [\mathbb{P}_1]^d$, we have

$$(\mathbf{u}_0^i \cdot \mathbf{n}_j)(\mathbf{x}) = \mathbf{u}_0^i(\mathbf{x}) \cdot \mathbf{n}_j(v_i) + \mathbf{u}_0^i(\mathbf{x}) \cdot [\mathcal{D}\mathbf{n}_j(v_i)(\mathbf{x} - v_i)] \quad \forall \mathbf{x} \in T.$$

Note that $\mathbf{u}_0^i \cdot \mathbf{n}_j$ vanishes at the vertex $v_i$ as $\mathbf{u}_0 \in \tilde{\mathcal{N}}_h$. Moreover, we know $\mathbf{u}_0^i(\mathbf{x})/\|\mathbf{u}_0^i(\mathbf{x})\|$ is constant on the interior of the patch around $v_i$, and $\mathbf{u}_0^i(\mathbf{x})$ is zero on the boundary of the patch, since we can write $\mathbf{u}_0^i(\mathbf{x}) = \mathbf{u}_0(v_i)\psi_i(\mathbf{x})$ with $\psi_i$ denoting the scalar piecewise linear basis function (vanishing outside the patch) associated with $v_i$. Therefore, we can deduce $\mathbf{u}_0^i(\mathbf{x}) \cdot \mathbf{n}_j(v_i) = 0$ on $T$. In addition, we have $\|\mathbf{x} - v_i\| \lesssim h$ on the element $T$. We thus conclude that

$$\|\mathbf{u}_0^i \cdot \mathbf{n}_j\|_{L^2(T)} \lesssim h\|\mathcal{D}\mathbf{n}_j\|_{L^\infty(T)}\|\mathbf{u}_0^i\|_{L^2(T)}.$$

From this we are able to show that for both the star and point-block patches around $v_i$,

$$\begin{aligned}\sum_i \|\mathbf{u}_0^i \cdot \mathbf{n}_j\|^2_{L^2(\text{patch}(v_i))} &\lesssim \sum_i h^2\|\mathcal{D}\mathbf{n}_j\|^2_{L^\infty(\text{patch}(v_i))}\|\mathbf{u}_0^i\|^2_{L^2(\text{patch}(v_i))} \\ &\lesssim h^2\|\mathcal{D}\mathbf{n}_j\|^2_{L^\infty(\Omega)} \sum_i \|\mathbf{u}_0^i\|^2_{L^2(\Omega)} \\ &\lesssim h^2\|\mathcal{D}\mathbf{n}_j\|^2_{L^\infty(\Omega)}\|\mathbf{u}_0\|^2_{L^2(\Omega)}.\end{aligned}$$

Therefore, with the local support of $\mathbf{u}_0^i$ we have

$$\sum_i \|\mathbf{u}_0^i \cdot \mathbf{n}_j\|^2_{L^2(\Omega)} = \sum_i \|\mathbf{u}_0^i \cdot \mathbf{n}_j\|^2_{L^2(\text{patch}(v_i))} \lesssim h^2\|\mathcal{D}\mathbf{n}_j\|^2_{L^\infty(\Omega)}\|\mathbf{u}_0\|^2_{L^2(\Omega)}.$$

□

We now derive the general form of the spectral bounds in (3.1.1.1). This follows a similar approach to [Sch99b, Theorem 4.1], but with a different assumption on the splitting approximation, to allow for a dependence on $\gamma$. Given a space decomposition $V_h = \sum_i^M V_i$, we define its *overlap* $N_O$ as

$$N_O \coloneqq \max_{1 \le i \le M} \sum_{j=1}^{M} g_{ij},$$

where

$$g_{ij} = \begin{cases} 1 & \text{if } \exists \mathbf{v}_i \in V_i, \mathbf{v}_j \in V_j : |\text{supp}(\mathbf{v}_i) \cap \text{supp}(\mathbf{v}_j)| > 0, \\ 0 & \text{otherwise} \end{cases}$$

measures the interaction between each subspace.

**Theorem 3.3.** *Let $\{V_i\}$ be a subspace decomposition of $V_h$ with overlap $N_O$. Assume that the finite element pair $V_h$-$Q_h$ for $(u, \lambda)$ is inf-sup stable for the mixed problem*

$$\begin{aligned}\mathfrak{B}((\mathbf{u},\lambda);(\mathbf{v},\mu)) &:= K_c\left(\nabla\mathbf{u},\nabla\mathbf{v}\right)_0 + 2\left(\lambda,\mathbf{n}_j\cdot\mathbf{v}\right)_0 + 2\left(\mu,\mathbf{n}_j\cdot\mathbf{u}\right)_0\\ &= \mathcal{F}(\mathbf{v},\mu)\quad \forall(\mathbf{v},\mu)\in V_h\times Q_h,\end{aligned}$$

*where $\mathcal{F}$ is a known functional. Furthermore, assume that the function $\mathbf{u}_h \in V_h$ and the kernel function $\mathbf{u}_0 \in \mathcal{N}_h$ can be split locally with estimates depending on the mesh size $h$ and possibly on $\gamma$ if the kernel-capturing property is not satisfied:*

$$\begin{aligned}\inf_{\substack{\mathbf{u}_h=\sum_i \mathbf{u}_h^i\\ \mathbf{u}_h^i\in V_i}} \sum_i \|\mathbf{u}_h^i\|_1^2 &\le c_1(h)\|\mathbf{u}_h\|_0^2,\\ \inf_{\substack{\mathbf{u}_0=\sum_i \mathbf{u}_0^i\\ \mathbf{u}_0^i\in V_i}} \sum_i \|\mathbf{u}_0^i\|_{A_{h,\gamma}}^2 &\le \left(c_2(h)+c_3(h,\gamma)\right)\|\mathbf{u}_0\|_0^2.\end{aligned}$$

*Then the additive Schwarz preconditioner $D_{h,\gamma}$ built on the decomposition $\{V_i\}$ satisfies*

$$\left(c_1(h)+c_2(h)+c_3(h,\gamma)\right)^{-1} D_{h,\gamma} \le A_{h,\gamma} \le N_O D_{h,\gamma}, \tag{3.1.1.3}$$

*with constants $c_1$ and $c_2$ independent of $\gamma$.*

*Proof.* The upper bound can be directly given by [Sch99b, Lemma 3.2] independent of the form of partial differential equations.

For the lower bound, choose $\mathbf{u}_h \in V_h$ and split it into $\mathbf{u}_h = \mathbf{u}_0 + \mathbf{u}_1$, by solving

$$\mathfrak{B}((\mathbf{u}_1,\lambda_1),(\mathbf{v}_h,\mu_h)) = 2\left(\mu_h,\mathbf{n}_j\cdot\mathbf{u}_h\right)_0 \quad \forall(\mathbf{v}_h,\mu_h)\in V_h\times Q_h. \tag{3.1.1.4}$$

Testing with $\mathbf{v}_h = 0$ in (3.1.1.4), we obtain that

$$\left(\mu_h,\mathbf{n}_j\cdot\mathbf{u}_1\right)_0 = \left(\mu_h,\mathbf{n}_j\cdot\mathbf{u}_h\right)_0 \quad \forall\mu_h\in Q_h.$$

Furthermore, since the current approximation $\mathbf{n}_j$ is well-controlled as from Assumption 2.3, $\mathbf{n}_j\cdot\mathbf{u}$ belongs to $L^2(\Omega)$. Hence, $\mathbf{n}_j\cdot\mathbf{u}_0 = 0$ a.e., that is to say $\mathbf{u}_0 \in \mathcal{N}_h$.

By stability of the finite element pair $V_h$-$Q_h$, we have

$$\begin{aligned}\|\mathbf{u}_1\|_1 &\lesssim \sup_{\substack{\mathbf{v}_h\in V_h\\ \mu_h\in Q_h}} \frac{\mathfrak{B}((\mathbf{u}_1,\lambda_1),(\mathbf{v}_h,\mu_h))}{\|(\mathbf{v}_h,\mu_h)\|}\\ &\lesssim \sup_{\substack{\mathbf{v}_h\in V_h\\ \mu_h\in Q_h}} \frac{\|\mathbf{n}_j\cdot\mathbf{u}_h\|_0\|\mu_h\|_0}{\|(\mathbf{v}_h,\mu_h)\|}\\ &\leq \|\mathbf{n}_j\cdot\mathbf{u}_h\|_0.\end{aligned}$$

It follows that

$$\|\mathbf{u}_1\|_1 \lesssim \|\mathbf{u}_h\|_0$$

by the boundedness of $\mathbf{n}_j$ and

$$\|\mathbf{u}_1\|_1 \lesssim \gamma^{-1/2}\|\mathbf{u}_h\|_{A_{h,\gamma}}$$

by the form of the operator $A_{h,\gamma}$, respectively. Using $\mathbf{u}_0 = \mathbf{u}_h - \mathbf{u}_1$, we have in addition that

$$\|\mathbf{u}_0\|_1 \lesssim \|\mathbf{u}_h\|_1.$$

We now calculate

$$\begin{aligned}\|\mathbf{u}_h\|^2_{D_{h,\gamma}} &= |||\mathbf{u}_h|||^2\\ &\leq \inf_{\substack{\mathbf{u}_1=\sum_i \mathbf{u}_1^i\\ \mathbf{u}_1^i\in V_i}} \sum_i \|\mathbf{u}_1^i\|^2_{A_{h,\gamma}} + \inf_{\substack{\mathbf{u}_0=\sum_i \mathbf{u}_0^i\\ \mathbf{u}_0^i\in V_i}} \sum_i \|\mathbf{u}_0^i\|^2_{A_{h,\gamma}}\\ &\lesssim (1+\gamma)\inf_{\substack{\mathbf{u}_1=\sum_i \mathbf{u}_1^i\\ \mathbf{u}_1^i\in V_i}} \sum_i \|\mathbf{u}_1^i\|_1^2 + (c_2(h)+c_3(h,\gamma))\|\mathbf{u}_0\|_0^2\\ &\lesssim (1+\gamma)c_1(h)\|\mathbf{u}_1\|_0^2 + (c_2(h)+c_3(h,\gamma))\|\mathbf{u}_0\|_1^2\\ &\lesssim (1+\gamma)c_1(h)\|\mathbf{u}_1\|_1^2 + (c_2(h)+c_3(h,\gamma))\|\mathbf{u}_h\|_1^2\\ &\lesssim (c_1(h)+c_2(h)+c_3(h,\gamma))\|\mathbf{u}_h\|^2_{A_{h,\gamma}},\end{aligned} \tag{3.1.1.5}$$

completing the proof of the spectral estimates (3.1.1.3). □

**Remark 3.1.** *Note that in Theorem 3.3, if the kernel-capturing property (3.1.0.5) is satisfied, then $c_3$ will be zero. Hence, we will instead get a parameter-independent result.*

**Corollary 3.4.** *In Theorem 3.3, if we take $V_h$-$Q_h$ to be constructed by the $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ element, it holds that*

$$\left(c_1(h) + c_2(h) + \gamma h^2 \|\mathcal{D}\mathbf{n}_j\|_\infty^2\right)^{-1} D_{h,\gamma} \leq A_{h,\gamma} \leq N_O D_{h,\gamma},$$

*with constants $c_1(h)$, $c_2(h) \sim \mathcal{O}(h^{-2})$.*

*Proof.* We follow the main argument of Theorem 3.3. We have only proven the kernel-capturing property for the approximate kernel (3.1.0.6) rather than (3.1.0.2), and need to account for this in the estimates. From Lemma 3.2 and the definition of $A_{h,\gamma}$ we have that

$$c_3(h,\gamma) = \gamma h^2 \|\mathcal{D}\mathbf{n}_j\|_\infty^2.$$

With the choice of $V_h = [\mathbb{P}_1]^d$, we will use the so-called *inverse inequality* (its proof can be found in any finite element book, e.g., [Cia78]) which states that

$$\|\mathbf{v}_h\|_1 \lesssim h^{-1}\|\mathbf{v}_h\|_0 \quad \forall \mathbf{v}_h \in V_h.$$

Therefore, it is straightforward to obtain that $c_1$ and $c_2$ are actually $\mathcal{O}(h^{-2})$. Notice here we have also used the form of $\|\cdot\|_{A_{h,\gamma}}$ in estimating $c_2(h)$.

Finally, substituting the form of $c_3$ in (3.1.1.5), we derive

$$\|\mathbf{u}_h\|_{D_{h,\gamma}}^2 \lesssim \left(c_1(h) + c_2(h) + \gamma h^2 \|\mathcal{D}\mathbf{n}_j\|_\infty^2\right) \|\mathbf{u}_h\|_{A_{h,\gamma}}^2,$$

with constants $c_1(h)$, $c_2(h) \sim \mathcal{O}(h^{-2})$. □

The above Corollary 3.4 implies that we cannot entirely get rid of parameter $\gamma$ in the spectral estimates if the kernel-capturing property for the kernel (3.1.0.2) is not satisfied and instead we get an additional factor of $\gamma h^2 \|\mathcal{D}\mathbf{n}_j\|_\infty^2$. However, this $\gamma$-dependence can be well controlled and does not impinge on the effectiveness of our smoother; the dependence improves as the mesh becomes finer or as $\mathbf{n}_j$ becomes smoother.

## 3.2 Prolongation

To construct a parameter-robust multigrid method, the prolongation operator is also required to be continuous (in the energy norm associated with the PDE) with the continuity constant independent of the penalty parameter $\gamma$ [Sch99b, Theorem 4.2]. In the context of the Oseen, Navier–Stokes, and linear elasticity equations, the prolongation operator was modified in order to guarantee that the continuity constant is $\gamma$-independent [Sch99b; BO06; FMW19]. However, in our experiments with the Oseen–Frank system, we observe robust convergence with respect to $\gamma$, even when using the (cheaper) standard prolongation. This can be seen in Tables 4.7 and 4.8 of Chapter 4, for example. Hence, we will use the standard prolongation with no modification in this part of work.

**Remark 3.2.** *Since both discretisations* $[\mathbb{P}_1]^d$-$\mathbb{P}_1$ *and* $[\mathbb{P}_2]^d$-$\mathbb{P}_1$ *are nested, i.e.,* $V_H \subset V_h$*, the standard prolongation is actually a continuous (in the* $H^1$*-norm) natural inclusion.*

## 3.3 Summary

In this chapter, we discussed constructing a robust multigrid algorithm for solving the augmented top-left block in the derived saddle point system (2.3.1.5) of LC problems. Two essential ingredients for the guarantee of robustness were examined: a relaxation that captures the kernel of the augmentation term and a prolongation operator that possesses a parameter-independent continuity constant. We will present some numerical results to verify the effectiveness of our constructed AL preconditioner in the next chapter.

# 4

# Numerical experiments for cholesterics

## Contents



## 4.1 Algorithm details

In the following numerical experiments, we use the $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ element pair and use flexible GMRES [Saa93] as the outermost linear solver, since GMRES [SS86] is applied in the multigrid relaxation. An absolute tolerance of $10^{-8}$ was used for the nonlinear solver, except for the convergence rate tests in Figure 4.4, which used $10^{-10}$. A relative tolerance of $10^{-4}$ was used for the inner linear solver. We use the full block factorisation preconditioner

$$\mathcal{Q}^{-1} = \begin{bmatrix} \mathbf{I} & -\tilde{\mathbf{A}}_\gamma^{-1}\mathbf{B}^\top \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{A}}_\gamma^{-1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_\gamma^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}\tilde{\mathbf{A}}_\gamma^{-1} & \mathbf{I} \end{bmatrix},$$

where $\mathbf{I}$ is the identity matrix and $\tilde{\mathbf{A}}_\gamma^{-1}$ represents solving the top-left block $\mathbf{A}_\gamma$ inexactly by our specialised multigrid algorithm described in the previous chapter

and the Schur complement approximation $\tilde{\mathbf{S}}_\gamma^{-1}$ is given by (2.3.2.8). The multiplier mass matrix inverse $\mathbf{M}_\lambda^{-1}$ is solved using Cholesky factorisation.

For $\tilde{\mathbf{A}}_\gamma^{-1}$, we perform a multigrid V-cycle, where the problem on the coarsest grid is solved exactly by Cholesky decomposition. On each finer level, as relaxation we perform 3 GMRES iterations preconditioned by the additive star (denoted as ALMG-STAR) iteration or additive point-block Jacobi (denoted as ALMG-PBJ) iteration. In order to achieve convergence results independent of the number of cores used in parallel, we only report iteration counts using additive relaxation, although multiplicative ones generally give better convergence. The star and Vanka relaxation methods are implemented using the PCPATCH preconditioner recently included in PETSc [Far+21].

**Code availability.** For reproducibility, both the solver code [Xia20] and the exact version of Firedrake [Fir20] used to produce the numerical results of this chapter have been archived on Zenodo. An installation of Firedrake with components matching those used in this chapter can be obtained by following the instructions at `https://www.firedrakeproject.org/download.html` with

```
python3 firedrake-install --doi 10.5281/zenodo.4249051
```

## 4.2 Numerical results

We denote #refs and #dofs as the number of mesh refinements and degrees of freedom, respectively, in the following experiments. The test problems in this section assume that the domain represents a uniform slab in the $xy$-plane, i.e., $\mathbf{n}$ may have a nonzero $z$-component but $\frac{\partial \mathbf{n}}{\partial z} = \mathbf{0}$. Hence, though the domain is in two dimensions, we use the Cartesian representation of the director $\mathbf{n} = (n_x, n_y, n_z)$ throughout this chapter.

### 4.2.1 Periodic boundary condition in a square slab

Following the nematic benchmarks in [Adl+16, Section 5.1], we consider a generalised twist equilibrium configuration in a square $\Omega = [0, 1] \times [0, 1]$, which has an analytical

solution [Ste04]. We will investigate the robustness of the solver when applied to unequal Frank constants and nonzero cholesteric pitch.

We impose periodic boundary conditions in the $x$-direction and Dirichlet boundary conditions in the $y$-direction, with values

$$\mathbf{n} = [\cos\vartheta_0, 0, -\sin\vartheta_0]^\top \qquad \text{on} \quad y = 0,$$
$$\mathbf{n} = [\cos\vartheta_0, 0, \sin\vartheta_0]^\top \qquad \text{on} \quad y = 1,$$

where $\vartheta_0 = \pi/8$.

We first consider parameter values $K_1 = 1.0$, $K_2 = 1.2$, $K_3 = 1.0$, $q_0 = 0$ when solving the minimisation problem (2.1.0.1). The exact solution is given by

$$\mathbf{n} = [\cos(\vartheta_0(2y-1)), 0, \sin(\vartheta_0(2y-1))]^\top,$$

with true free energy $2K_2\vartheta_0^2 \approx 0.37011$. An example of the pure twist configuration is illustrated in Figure 4.1.

We use an initial guess of $\mathbf{n}_0 = [1, 0, 0]^\top$ in the Newton iteration and a $10 \times 10$ mesh of triangles of negative slope as the coarse grid.

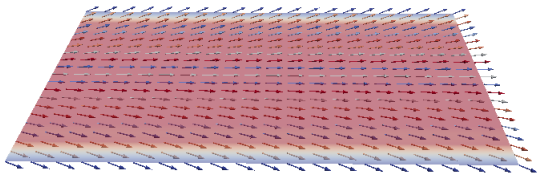


**Figure 4.1:** A sample solution of the twist configuration. Colours represent the magnitude of directors.

We first compare in Table 4.1 the nonlinear convergence of the Newton linearisation (2.3.1.2) against that of the Picard iteration (2.3.1.3) we propose. For these experiments we use the augmented Lagrangian preconditioner with ideal inner solvers (denoted as ALLU), i.e. where the top-left block is solved exactly by LU factorisation. The Picard iteration requires substantially fewer nonlinear iterations for large $\gamma$. We expect that this relates to the degradation of the coercivity estimate given in Lemma 2.5. Similar results were obtained on other test cases and we adopt the Picard iteration henceforth.

| | #refs | #dofs | $\gamma$ | | | |
|---|---|---|---|---|---|---|
| | | | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| Newton | 1 | 5,340 | 2.20 (5) | 1.14 (7) | 1.00 (10) | 1.00 (19) |
| | 2 | 21,080 | 3.20 (5) | 1.14 (7) | 1.00 (12) | 1.00 (15) |
| | 3 | 83,760 | 3.83 (6) | 1.57 (7) | 1.11 (9) | 1.00 (14) |
| | 4 | 333,920 | 4.67 (6) | 2.14 (7) | 1.00 (7) | 1.00 (11) |
| | 5 | 1,333,440 | 5.17 (6) | 2.43 (7) | 1.57 (7) | 1.00 (10) |
| Picard | 1 | 5,340 | 2.00 (5) | 1.20 (5) | 1.14 (7) | 1.11 (9) |
| | 2 | 21,080 | 3.00 (5) | 1.40 (5) | 1.17 (6) | 1.12 (8) |
| | 3 | 83,760 | 3.83 (6) | 2.00 (5) | 1.17 (6) | 1.14 (7) |
| | 4 | 333,920 | 4.67 (6) | 2.29 (7) | 1.14 (7) | 1.17 (6) |
| | 5 | 1,333,440 | 5.17 (6) | 2.57 (7) | 1.50 (8) | 1.17 (6) |

**Table 4.1:** A comparison of the nonlinear convergence of the Newton linearisation (2.3.1.2) and the Picard iteration (2.3.1.3) using ideal inner solvers for a nematic LC problem in a square slab. The table shows the average number of outer FGMRES iterations per nonlinear iteration and the total nonlinear iterations in brackets.

To see the efficiency of the Schur complement approximation (2.3.2.8) we used in Section 2.3.2, we give the number of Krylov iterations for ALLU in Table 4.2. It can be observed that as $\gamma$ increases, the preconditioner becomes a better approximation to the real Jacobian inverse and the preconditioner is mesh-independent.

**Remark 4.1.** *It can be noted from Table 4.2 that ALLU seems to give a rather reasonable solver for $\gamma = 0$ (and thus with no penalisation of the unit-length constraint). One may wonder whether the example illustrated in this subsection is a good one for testing the application of augmented Lagrangian methods. Indeed, this is a simpler case but with a known exact solution and it is intended for showing the efficiency and convergence rate of our proposed AL preconditioner. More complicated cases will be given later.*

The performance of ALMG-STAR (utilising the augmented Lagrangian preconditioner with star patch as the relaxation in the multigrid algorithm) and ALMG-PBJ (utilising the augmented Lagrangian preconditioner with the point-block Jacobian relaxation in the multigrid algorithm) are illustrated in Tables 4.3 and 4.4, respectively, where both mesh-independence for $\gamma = 10^6$ and $\gamma$-robustness are observed.

| | | $\gamma$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| #refs | #dofs | 0 | 1 | 10 | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 5,340 | 10.40 | 9.20 | 8.00 | 5.40 | 2.00 | 1.20 | 1.14 | 1.11 |
| 2 | 21,080 | 14.20 | 13.20 | 9.20 | 5.80 | 3.00 | 1.40 | 1.17 | 1.12 |
| 3 | 83,760 | 4.75 | 4.75 | 6.75 | 6.40 | 3.83 | 2.00 | 1.17 | 1.14 |
| 4 | 333,920 | 5.50 | 4.50 | 7.25 | 7.20 | 4.67 | 2.29 | 1.14 | 1.17 |
| 5 | 1,333,440 | 5.25 | 3.75 | 5.75 | 7.00 | 5.17 | 2.57 | 1.50 | 1.17 |

**Table 4.2:** ALLU: The average number of FGMRES iterations per nonlinear iteration for a nematic LC problem in a square slab using $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ discretisation. Note here the last four columns are excerpted from Table 4.1 using the Picard iteration.

| | | $\gamma$ | | | |
|---|---|---|---|---|---|
| #refs | #dofs | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 5,340 | 2.60 (5) | 2.40 (5) | 2.29 (7) | 2.29 (7) |
| 2 | 21,080 | 4.20 (5) | 2.20 (5) | 2.50 (6) | 3.29 (7) |
| 3 | 83,760 | 8.00 (5) | 3.00 (5) | 2.33 (6) | 3.33 (6) |
| 4 | 333,920 | 11.60 (5) | 5.17 (6) | 2.17 (6) | 2.29 (7) |
| 5 | 1,333,440 | 15.20 (5) | 8.43 (7) | 3.14 (7) | 1.78 (9) |

**Table 4.3:** ALMG-STAR: the average number of FGMRES iterations per nonlinear iteration (total Newton iterations) for the nematic LC problem in a square slab.

| | | $\gamma$ | | | |
|---|---|---|---|---|---|
| #refs | #dofs | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 5,340 | 3.20 (5) | 2.60 (5) | 3.00 (6) | 3.57 (7) |
| 2 | 21,080 | 5.60 (5) | 2.60 (5) | 2.83 (6) | 3.71 (7) |
| 3 | 83,760 | 10.00 (5) | 3.80 (5) | 2.80 (5) | 3.00 (6) |
| 4 | 333,920 | 15.40 (5) | 7.00 (5) | 2.50 (6) | 2.83 (6) |
| 5 | 1,333,440 | >100 | 11.83 (6) | 5.00 (5) | 2.83 (6) |

**Table 4.4:** ALMG-PBJ: the average number of FGMRES iterations per nonlinear iteration (total Newton iterations) for the nematic LC problem in a square slab.

We also test the robustness of ALMG-STAR and ALMG-PBJ on other problem parameters, e.g., the twist elastic constant $K_2 > 0$ and the cholesteric pitch $q_0$. To this end, we continue $K_2 \in [0.2, 8]$ and $q_0 \in [0, 8]$ with step 0.1. We fix $\gamma = 10^6$, since it gives the best performance in Tables 4.3 and 4.4. The numerical results of ALMG-STAR and ALMG-PBJ in $K_2$- and $q_0$-continuation are shown in Figures 4.2 and 4.3, respectively. Clearly, a stable number of linear iterations is shown for

both continuation experiments.

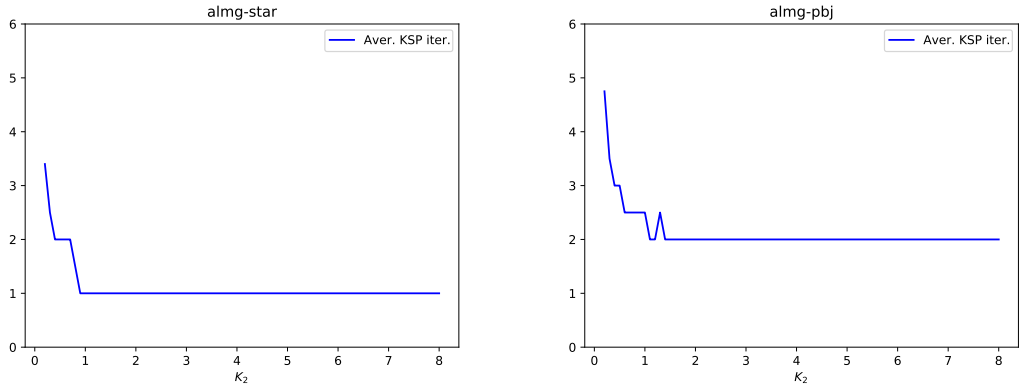



**Figure 4.2:** Average number of FGMRES iterations per nonlinear iteration when continuing in $K_2$ for the LC problem in a square slab.

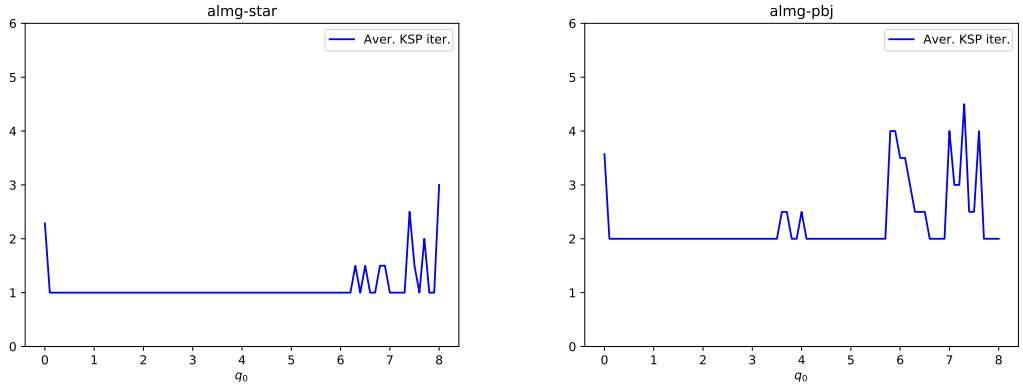



**Figure 4.3:** Average number of FGMRES iterations per nonlinear iteration when continuing in $q_0$ for the LC problem in a square slab.

To examine the convergence order of the discretisation as a function of $\gamma$, we apply the ALMG-PBJ solver for $\gamma = 10^4, 10^5$ and $10^6$. Note that the convergence result does not rely on the solver used. Figure 4.4 shows the $L^2$- and $H^1$-error between the computed director and the known analytical solution. We observe third order convergence of the director in the $L^2$ norm and second order convergence in the $H^1$ norm for all values of $\gamma$ considered.

To investigate the computational efficiency of the AL approach, we compare our proposed AL-based solvers (ALMG-PBJ and ALMG-STAR) with a monolithic multigrid preconditioner using Vanka relaxation [Adl+15a; Van86] on each level

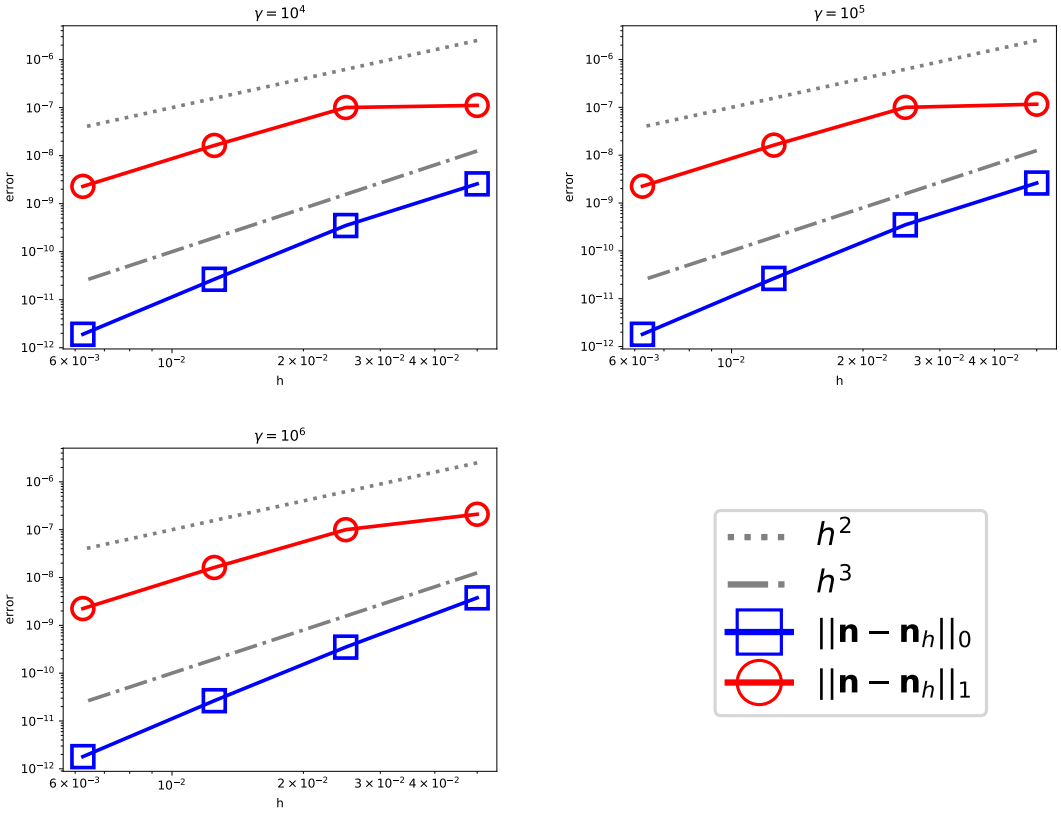



**Figure 4.4:** The convergence of the computed director as the mesh is refined for the nematic LC problem in a square slab.

(denoted as MGVANKA) in Table 4.5. Essentially, MGVANKA applies multigrid to the coupled director-multiplier problem, with an additive Schwarz relaxation organised around gathering all director dofs coupled to a given multiplier dof. All results are computed in serial. In our experiments, these two AL-based solvers outperform MGVANKA even for small problems of about five thousand dofs. In particular, ALMG-PBJ is the fastest method considered and is approximately five times faster than MGVANKA for a problem with about five million dofs. We also notice that ALMG-STAR is slower than ALMG-PBJ, which is caused by the size of the star patch being larger than that of the point-block patch, requiring more work in the multigrid relaxation.

| Computing time (in minutes) | | | | | | |
|---|---|---|---|---|---|---|
| #refs | 1 | 2 | 3 | 4 | 5 | 6 |
| #dofs | 5,340 | 21,080 | 83,760 | 333,920 | 1,333,440 | 5,329,280 |
| ALMG-PBJ | 0.02 | 0.04 | 0.09 | 0.32 | 1.17 | 5.53 |
| ALMG-STAR | 0.02 | 0.07 | 0.23 | 0.79 | 2.95 | 12.86 |
| MGVANKA | 0.04 | 0.15 | 0.38 | 1.44 | 5.91 | 25.09 |

**Table 4.5:** The computing time of ALMG-PBJ, ALMG-STAR and MGVANKA as a function of mesh refinement for the nematic LC problem in a square slab.

### 4.2.2 Equal-constant nematic case in an ellipse

Consider an ellipse of aspect ratio 3/2 with strong anchoring boundary condition $\mathbf{n} = [0, 0, 1]^\top$ imposed on the entire boundary. We consider the equal-constant nematic case $K_1 = K_2 = K_3 = 1$, $q_0 = 0$ in the minimisation problem (2.1.0.1) to verify the theoretical results presented in previous sections with corresponding discretisations. We use the initial guess $\mathbf{n}_0 = [0, 0, 0.8]^\top$ in the nonlinear iteration. The coarsest triangulation, generated in Gmsh [GR09], is illustrated in Figure 4.5.

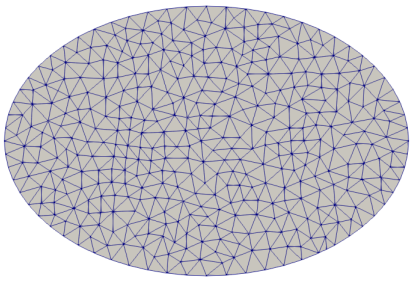


**Figure 4.5:** The coarse mesh of the ellipse.

To verify our theoretical results about the improvement of the discrete enforcement of the constraint in Section 2.3.3, we vary the penalty parameter $\gamma$, use one refinement for the fine mesh, and employ the $[\mathbb{P}_1]^3$-$\mathbb{P}_1$ element. The data is plotted in Figure 4.6. The $L^2$-norm $\|\mathbf{n} \cdot \mathbf{n} - 1\|_0$ of the residual of the constraint decreases as $\gamma$ grows, and scales like $\mathcal{O}(\gamma^{-1/2})$ as expected.

The efficiency of the Schur complement approximation of Section 2.3.2 for the $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ element can be observed in Table 4.6.

Tables 4.7 and 4.8 demonstrate the robustness of ALMG-STAR and ALMG-PBJ with respect to $\gamma$ and mesh refinement for the $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ element. It can be seen

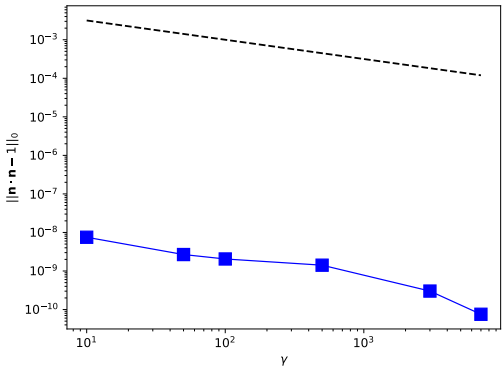



**Figure 4.6:** Comparison of the computed constraint $\|\mathbf{n}\cdot\mathbf{n}-1\|_0$ and the reference line $\mathcal{O}(\gamma^{-1/2})$ using the $[\mathbb{P}_1]^3$-$\mathbb{P}_1$ finite element pair for equal-constant nematic LC problems in an ellipse.

| | | $\gamma$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| #refs | #dofs | 0 | 1 | 10 | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 19,933 | 29.20 | 25.60 | 16.40 | 5.20 | 2.60 | 1.60 | 1.33 | 1.14 |
| 2 | 78,810 | 32.50 | 26.00 | 14.00 | 6.80 | 3.40 | 1.80 | 1.33 | 1.17 |
| 3 | 313,408 | 12.50 | 15.50 | 16.25 | 7.60 | 4.20 | 2.20 | 1.33 | 1.17 |
| 4 | 1,249,980 | 11.00 | 12.25 | 14.75 | 8.40 | 4.80 | 2.60 | 1.40 | 1.17 |
| 5 | 4,992,628 | 12.33 | 13.33 | 11.75 | 8.00 | 5.20 | 3.00 | 1.50 | 1.14 |

**Table 4.6:** ALLU: The average number of FGMRES iterations per nonlinear iteration for an equal-constant nematic problem in an ellipse using $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ discretisation.

that both solvers are robust with respect to the penalty parameter $\gamma$, and with respect to the mesh size $h$ for $\gamma = 10^6$. The number of nonlinear iterations and the number of FGMRES iterations per nonlinear step remain stable.

| | | $\gamma$ | | | |
|---|---|---|---|---|---|
| #refs | #dofs | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 19,933 | 2.60 (5) | 1.60 (5) | 1.80 (5) | 1.67 (6) |
| 2 | 78,810 | 4.40 (5) | 1.80 (5) | 1.60 (5) | 1.50 (6) |
| 3 | 313,408 | 6.80 (5) | 3.20 (5) | 1.50 (6) | 1.50 (6) |
| 4 | 1,249,980 | 10.00 (5) | 4.67 (6) | 1.80 (5) | 1.50 (6) |
| 5 | 4,992,628 | 14.40 (5) | 7.50 (6) | 4.20 (5) | 1.33 (6) |

**Table 4.7:** ALMG-STAR: the average number of FGMRES iterations per nonlinear iteration (total nonlinear iterations) for equal-constant nematic problem in an ellipse using $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ discretisation.

| | | $\gamma$ | | | |
|---|---|---|---|---|---|
| #refs | #dofs | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
| 1 | 19,933 | 3.80 (5) | 2.60 (5) | 2.60 (5) | 2.80 (5) |
| 2 | 78,810 | 6.80 (5) | 3.20 (5) | 2.60 (5) | 2.60 (5) |
| 3 | 313,408 | 9.00 (5) | 5.00 (5) | 2.60 (5) | 2.60 (5) |
| 4 | 1,249,980 | 14.80 (5) | 8.20 (5) | 3.80 (5) | 2.40 (5) |
| 5 | 4,992,628 | 19.00 (5) | 11.60 (5) | 6.80 (5) | 2.50 (6) |

**Table 4.8:** ALMG-PBJ: the average number of FGMRES iterations per Newton iteration (total Newton iterations) for equal-constant nematic problem in an ellipse using $[\mathbb{P}_2]^3$-$\mathbb{P}_1$ discretisation.

## 4.3 Summary

In this chapter, we presented numerical results of our proposed AL preconditioner for two examples of LC problems in two dimensions (an ellipse and a square slab). We demonstrated the effectiveness and robustness (regarding to problem-related parameters, the elastic constant $K_2$ and the cholesteric pitch $q_0$, and the mesh size $h$) of the preconditioner. We also tested the efficiency of preconditioners with star and point block patches and gave the numerical verification of the improvement of the constraint proven in Section 2.3.3.

This part of the thesis (from Chapter 2 to Chapter 4) resolves the difficulty of solving a unit-length constrained minimisation problem of the Oseen–Frank model for LC by applying augmented Lagrangian methods. It provides a viable approach to construct efficient and robust solvers for liquid crystal problems involving Oseen–Frank models, although the complexity can rapidly increase when it comes to more sophisticated phases requiring the coupling with other order parameters, e.g., in ferronematics and smectics. In the remainder of this thesis, we consider another modelling theory which avoids the imposition of unit-length constraint for a vector field, and instead turn to the so-called **Q**-tensor theory.

# Part II

# Ferronematic Liquid Crystals

---

**This work is derived from *Dalby, Farrell, Majumdar and Xia (2021)* [Dal+21].**

---

# 5

# A mathematical model of ferronematics

## Contents



In the previous part, we considered the Oseen–Frank model for cholesteric and nematic liquid crystals. This model uses a vector-valued order parameter and only applies to uniaxial LC, i.e., where only one direction of molecular alignment is preferred. In fact, the Oseen–Frank formulation is known to be limited, in the sense that it can only account for point defects, but not the more complicated line or surface defects that are observed experimentally [MZ10]. One can simply check this by observing that the Oseen–Frank free energy $\mathcal{J}^{OF}(\mathbf{n})$ (2.1.0.1), with equal Frank constants and zero cholesteric pitch $q_0$, blows up for the line defect $\mathbf{n} = [x, y, 0]^\top/\sqrt{x^2+y^2}$, while the energy functional is well-defined for the point defect $\mathbf{n} = [x, y, z]^\top/\sqrt{x^2+y^2+z^2}$. Another potential drawback of this theory is its inability of representing half-charge defects, due to the presence of director discontinuities in these defects [Bal17], which cannot be characterised by a continuous vector field. For example, around a $\pm 1/2$ defect where $\mathbf{n}$ rotates by $\pm\pi$ degrees, a

discontinuity line (i.e., *branch cut* in [BZ08]) where $\mathbf{n}$ reverses sign must exist.

Hence, to better characterise the defect structure, particularly in more complex liquid crystal phases and applications, we instead use a more complete phenomenological description for LC: the Landau–de Gennes (LdG) theory [Gen69; Gen74], which can account for both uniaxial and biaxial (having more than one preferred direction of molecular alignment) phases. The LdG theory is widely used in the modelling of phase transitions in liquid crystals [BCT07; Gen69] and we thus adopt it for ferronematics and smectics in the remainder of this thesis. In this part, we consider the case of ferronematics and we first give an introduction on some details of the LdG theory to prepare ourselves for modelling ferronematics.

## 5.1 The Landau–de Gennes model

In this framework, the state of nematic LC is modelled by a symmetric, traceless tensor field $\mathbf{Q} : \Omega \to S_0$, known as the tensor order parameter. Here, $S_0$ denotes the set of all symmetric, traceless $d \times d$ matrices. We consider a three-dimensional domain $\Omega \subset \mathbb{R}^3$ (i.e., $d = 3$) filled with liquid crystal as an example in this subsection; the two dimensional case is analogous. The eigenvectors $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$ of $\mathbf{Q} \in S_0$ are the directions of the preferred molecular orientations and their associated eigenvalues $\lambda_1, \lambda_2$ and $\lambda_3$ represent the degree of order along each corresponding direction [MN14].

We say that liquid crystals are (a) *isotropic* if $\mathbf{Q}$ has three equal eigenvalues, i.e., $\lambda_1 = \lambda_2 = \lambda_3$ (and hence, $\mathbf{Q} = 0$). They are (b) *uniaxial* when $\mathbf{Q}$ has two equal nonzero eigenvalues (say, $2|\lambda_1| = 2|\lambda_2| = |\lambda_3|$, thus $\lambda_3$ is the major eigenvalue). Such uniaxial $\mathbf{Q}$-tensors can be written in the special form

$$\mathbf{Q} = s\left(\mathbf{n} \otimes \mathbf{n} - \frac{\mathbf{I}_3}{3}\right), \quad s : \Omega \to \mathbb{R}, \quad \mathbf{n} : \Omega \to \mathcal{S}^2,$$

where $s = \frac{3}{2}\lambda_3$. Finally, they are (c) *biaxial* when $\mathbf{Q}$ has three distinct eigenvalues. A biaxial $\mathbf{Q}$-tensor can always be represented by

$$\mathbf{Q} = s\left(\mathbf{n} \otimes \mathbf{n} - \frac{1}{3}\mathbf{I}_3\right) + t\left(\mathbf{r} \otimes \mathbf{r} - \frac{1}{3}\mathbf{I}_3\right), s, t : \Omega \to \mathbb{R}, \mathbf{n}, \mathbf{r} : \Omega \to \mathcal{S}^2. \quad (5.1.0.1)$$

The Landau–de Gennes energy for nematic LC is of the form

$$\mathcal{J}^{LdG}(\mathbf{Q}) = \int_\Omega \left\{ f_n^e(\nabla \mathbf{Q}) + f_n^b(\mathbf{Q}) \right\},$$

where $f_n^e$ and $f_n^b$ correspond to the nematic *elastic* and *bulk* energy densities to be defined in the following. Note that the minimisation problem with functional $\mathcal{J}^{LdG}$ is unconstrained, as opposed to the constrained minimisation problem (2.1.0.1) in the Oseen–Frank theory.

The elastic part consists of three independent quadratic terms with respect to the first partial derivatives of components of $\mathbf{Q}$. Specifically, we take the form

$$f_n^e(\nabla \mathbf{Q}) = \frac{1}{2}\{K_a \mathbf{Q}_{ij,n}\mathbf{Q}_{ij,n} + K_b \mathbf{Q}_{ij,j}\mathbf{Q}_{in,n} + K_c \mathbf{Q}_{ij,n}\mathbf{Q}_{in,j}\}, \tag{5.1.0.2}$$

where $K_a$, $K_b$ and $K_c$ are *elastic* constants depending on the material. Here, we adopted the Einstein summation convention for repeated indices.

The bulk energy density is typically a truncated expansion in the scalar invariants of $\mathbf{Q}$ and accounts for bulk effects. One commonly used form [Gen74; MN14] is

$$f_n^b(\mathbf{Q}) = \frac{l_a}{2}\mathrm{tr}\left(\mathbf{Q}^2\right) - \frac{l_b}{3}\mathrm{tr}\left(\mathbf{Q}^3\right) + \frac{l_c}{4}\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right)^2. \tag{5.1.0.3}$$

Here, $l_b, l_c > 0$ are material-dependent *bulk* constants, independent of temperature, whereas $l_a < 0$ depends on the temperature.

Taking $K_a = K_{LdG}$, $K_b = K_c = 0$ in (5.1.0.2), we obtain the *one-constant* form of the LdG energy for nematic LC:

$$\mathcal{J}^{LdG}(\mathbf{Q}) = \int_\Omega \left\{ \frac{K_{LdG}}{2}|\nabla \mathbf{Q}|^2 + \frac{l_a}{2}\mathrm{tr}\left(\mathbf{Q}^2\right) - \frac{l_b}{3}\mathrm{tr}\left(\mathbf{Q}^3\right) + \frac{l_c}{4}\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right)^2 \right\}, \tag{5.1.0.4}$$

which will be employed in several places, e.g., (5.2.0.2) in ferronematics and our new proposed smectic model (7.3.1.2).

## 5.2 Full model of ferronematics

To start with the first application of the LdG theory in this thesis, we now briefly introduce ferronematic materials and their modelling.

Nematic LC are anisotropic materials that can respond to applied external fields and are thus suitable for a wide range of electro-optic devices, especially liquid crystal displays. One immediate example is the twisted nematic display [DS11, Technical Box 10.1] where the display is switched on and off by activating or deactivating an electric field applied to the nematic LC. In fact, this response relies on the dielectric anisotropy of nematics, that is to say, the directional response to external electric fields [LS12]. In contrast, when exposed to magnetic fields, their responses are much weaker (perhaps seven orders of magnitude smaller) than that of electric fields [Ste04]. Consequently, nemato-magnetic coupling effect has not been extensively exploited for nematic applications, e.g., sensors, displays, microfluidics etc. One pioneering work dating back to 1970 by Brochard & de Gennes [BG70] found that a suspension of magnetic nanoparticles (MNPs) in a nematic phase can induce a spontaneous magnetisation in the absence of an external magnetic field, and substantially enhance the nemato-magnetic response. They referred to this new class of materials as *ferronematics*, possessing the useful feature that the nematic and magnetic order parameters are strongly coupled. Subsequently, there were some notable theoretical contributions regarding ferronematic modelling made by Burylov & Raikher [BR95] and Calderer et al. [Cal+14], where continuum models were discussed and analysed. Meanwhile, some experiments about ferronematics were also realised by Rauly, Cladis and Burger [RCB70], and more recently by Mertelj et al. [Mer+13]. Due to their special responses in the absence of any external magnetic fields, ferronematics may find potential use in magneto-optic devices.

In this chapter, we study a dilute suspension of MNPs in a three-dimensional nematic-filled channel, $\tilde{\Omega} = [-L, L] \times [-D, D] \times [0, G]$, where $L \gg D$ is the length of the channel, $D$ is the width and $G$ is the height. Since $L \gg D$, it is reasonable to assume that molecules are uniform along the length and across the height of the channel, and there are no boundary constraints imposed at the two ends $x = \pm L$. Thus, we can restrict ourselves to a one-dimensional geometry: $\bar{\Omega} = [-D, D]$. We then rescale this domain to $\Omega = [-1, 1]$ for simplicity, similarly to [Bis+19].

The suspended MNPs generate a spontaneous magnetisation (in the absence of any external magnetic fields) by means of the nemato-magnetic coupling. In this system, there are two order parameters: (i) a nematic tensor parameter $\mathbf{Q} : \Omega \to S_0$ (symmetric, traceless $2 \times 2$ matrices), indicating the preferred molecular alignment of the director in the nematic host and (ii) a vector-valued magnetic order parameter $\mathbf{M} : \Omega \to \mathbb{R}^2$, $\mathbf{M} = (M_1, M_2)^\top$, generated by the suspended MNPs.

In the uniaxial case, as discussed above the nematic order parameter $\mathbf{Q}$ can be written as

$$\mathbf{Q} = s(2\mathbf{n} \otimes \mathbf{n} - \mathbf{I}_2), \tag{5.2.0.1}$$

where $s$ is a scalar order parameter and $\mathbf{n}$ is the nematic director. Here, $s$ can be interpreted as a measure of the degree of the orientational order for director $\mathbf{n}$, so that the nodal set of $s$ (i.e., where $s = 0$) indicates the presence of nematic defects (where an orientation is not well-defined). We denote the two independent components of $\mathbf{Q}$ by $Q_{11}$ and $Q_{12}$ such that

$$Q_{11} = s\cos 2\vartheta, \quad Q_{12} = s\sin 2\vartheta,$$

where $\mathbf{n} = (\cos\vartheta, \sin\vartheta)$ and $\vartheta$ denotes the angle between $\mathbf{n}$ and the horizontal axis. To avoid writing $\mathbf{Q}$ in the matrix form $\left[\begin{smallmatrix} Q_{11} & Q_{12} \\ Q_{12} & -Q_{11} \end{smallmatrix}\right]$, we henceforth label $\mathbf{Q}$ in terms of its two independent components $(Q_{11}, Q_{12})$, when this causes no confusions. Consequently, we use the vector norm $|\mathbf{Q}| = \sqrt{Q_{11}^2 + Q_{12}^2}$, as opposed to the usual matrix norm. The conventional definition of the vector norm is adopted for the magnetisation vector $\mathbf{M}$, that is to say, $|\mathbf{M}| = \sqrt{M_1^2 + M_2^2}$.

By following the methods in [Mer+13; Bis+19], we use the total rescaled and dimensionless ferronematic energy of the form

$$\begin{aligned}\mathcal{J}^{fer}(Q_{11}, Q_{12}, M_1, M_2) \coloneqq \int_\Omega \Bigg\{ & \frac{k_1}{2}\left[\left(\frac{\mathrm{d}Q_{11}}{\mathrm{d}y}\right)^2 + \left(\frac{\mathrm{d}Q_{12}}{\mathrm{d}y}\right)^2\right] + \left(Q_{11}^2 + Q_{12}^2 - 1\right)^2 \\ & + \frac{\xi k_2}{2}\left[\left(\frac{\mathrm{d}M_1}{\mathrm{d}y}\right)^2 + \left(\frac{\mathrm{d}M_2}{\mathrm{d}y}\right)^2\right] + \frac{\xi}{4}\left(M_1^2 + M_2^2 - 1\right)^2 \\ & - cQ_{11}\left(M_1^2 - M_2^2\right) - 2cQ_{12}M_1M_2 \Bigg\}\, \mathrm{d}y,\end{aligned} \tag{5.2.0.2}$$

and the associated minimisation problem is

$$\min \mathcal{J}^{fer}(Q_{11}, Q_{12}, M_1, M_2). \tag{5.2.0.3}$$

Here, $k_1 > 0$ and $k_2 > 0$ are scaled elastic constants (in practice, $k_1 > k_2$ since the nematic effect dominates in ferronematics), $\xi > 0$ is a parameter that weighs the relative strength of the nematic and magnetic energies, and $c$ is a coupling parameter. Since we consider a dilute suspension of MNPs, there are only "small" interactions between MNPs while the nemato-magnetic interactions are taken into account through the coupling energy term. Therefore, we can see that the magnetic energy part is not dominating and it is reasonable that $\xi \leq 1$ [Cal+14].

The ferronematic free energy is a sum of three energetic contributions: a LdG-type nematic energy of $\mathbf{Q}$, a magnetisation energy of $\mathbf{M}$, and a coupling energy between $\mathbf{Q}$ and $\mathbf{M}$. Substituting the uniaxial expression (5.2.0.1) into the coupling energy, we observe that

$$-cQ_{11}\left(M_1^2 - M_2^2\right) - 2cQ_{12}M_1M_2 \propto -c\left(\mathbf{n}\cdot\mathbf{M}\right)^2.$$

In this part of work, we only focus on positive coupling ($c > 0$) so that the coupling energy favours co-alignment between the nematic director $\mathbf{n}$ and magnetic vector $\mathbf{M}$.

Furthermore, we consider imposing Dirichlet boundary conditions for both $\mathbf{Q}$ and $\mathbf{M}$ on the ends $y = \pm 1$:

$$Q_{11}(-1) = M_1(-1) = 1, \tag{5.2.0.4a}$$

$$Q_{12}(-1) = Q_{12}(1) = M_2(-1) = M_2(1) = 0, \tag{5.2.0.4b}$$

$$Q_{11}(1) = M_1(1) = -1. \tag{5.2.0.4c}$$

Here, the boundary conditions for $\mathbf{Q}$ correspond to $\mathbf{n} = (1, 0)$ on $y = -1$ and $\mathbf{n} = (0, 1)$ on $y = 1$, that is to say, we are essentially enforcing planar boundary conditions for $\mathbf{Q}$ at $y = -1$ and homeotropic boundary conditions at the other end $y = +1$. Meanwhile, the boundary conditions for $\mathbf{M}$ describe a $\pi$-rotation of magnetic orientation between the bounding plates $y = \pm 1$. Then, the admissible space of the minimisation problem (5.2.0.3) is given by

$$\mathcal{A}_f = \Big\{ \mathbf{Q} \in H^1(\Omega, S_0), \mathbf{M} \in H^1\left(\Omega, \mathbb{R}^2\right), \\ \mathbf{Q} \text{ and } \mathbf{M} \text{ satisfy the boundary conditions (5.2.0.4)} \Big\}. \quad (5.2.0.5)$$

We are interested in the local or global energy minimisers $(\mathbf{Q}, \mathbf{M})$, being stable and potentially observable, of the ferronematic free energy (5.2.0.2) in the admissible space $\mathcal{A}_f$. In fact, they are *classical* solutions (which can be verified by elliptic regularity, suitable Sobolev embeddings and bootstrapping arguments) of the associated Euler–Lagrange equations

$$k_1 \frac{\mathrm{d}^2 Q_{11}}{\mathrm{d}y^2} = 4Q_{11}(Q_{11}^2 + Q_{12}^2 - 1) - c\left(M_1^2 - M_2^2\right), \quad (5.2.0.6a)$$

$$k_1 \frac{\mathrm{d}^2 Q_{12}}{\mathrm{d}y^2} = 4Q_{12}(Q_{11}^2 + Q_{12}^2 - 1) - 2cM_1M_2, \quad (5.2.0.6b)$$

$$\xi k_2 \frac{\mathrm{d}^2 M_1}{\mathrm{d}y^2} = \xi M_1\left(M_1^2 + M_2^2 - 1\right) - 2cQ_{11}M_1 - 2cQ_{12}M_2, \quad (5.2.0.6c)$$

$$\xi k_2 \frac{\mathrm{d}^2 M_2}{\mathrm{d}y^2} = \xi M_2\left(M_1^2 + M_2^2 - 1\right) + 2cQ_{11}M_2 - 2cQ_{12}M_1. \quad (5.2.0.6d)$$

**Remark 5.1.** *For simplicity and brevity, we take $k_1 = k_2 = k$ and $\xi = 1$ hereafter. One can tackle the cases of $k_1 \neq k_2$ and $\xi \neq 1$ using similar mathematical methods.*

An immediate question arises regarding the existence and uniqueness of minimisers of the problem (5.2.0.3) in the admissible space $\mathcal{A}_f$. The existence result is proven in [Dal+21] via the direct method of the calculus of variations. Uniqueness holds for sufficiently large $k$. We quote the theorem below for self-containment.

**Theorem 5.1.** *[Dal+21] (Uniqueness of minimisers for sufficiently large $k$) For a fixed $c$ and for $k$ sufficiently large, there exists a unique critical point (and hence global minimiser) of the ferronematic free energy (5.2.0.2) in the admissible space (5.2.0.5).*

Moreover, a maximum principle for the solutions $(Q_{11}, Q_{12}, M_1, M_2)$ of the system (5.2.0.6a)-(5.2.0.6d) is obtained in [Dal+21] and we include this result in the following so that we can numerically verify it later in Chapter 6.

**Theorem 5.2.** *[Dal+21] (Maximum principle) There exists an $L^\infty$ bound for the solutions $(Q_{11}, Q_{12}, M_1, M_2)$ of the system (5.2.0.6a)-(5.2.0.6d) subject to the boundary conditions (5.2.0.4). Specifically,*

$$Q_{11}^2(y) + Q_{12}^2(y) \leq (\rho^*)^2, \;\; M_1^2(y) + M_2^2(y) \leq 1 + 2c\rho^* \quad \forall y \in [-1, 1], \tag{5.2.0.7}$$

*where $\rho^*$ is given by*

$$\rho^* = \left( \frac{c}{8} + \sqrt{\frac{c^2}{64} - \frac{1}{27}\left(1 + \frac{c^2}{2}\right)^3} \right)^{\frac{1}{3}} + \left( \frac{c}{8} - \sqrt{\frac{c^2}{64} - \frac{1}{27}\left(1 + \frac{c^2}{2}\right)^3} \right)^{\frac{1}{3}}. \tag{5.2.0.8}$$

**Remark 5.2.** *We will verify the $L^\infty$ bound (5.2.0.7) numerically for each solution in Chapter 6.*

With the uniqueness and maximum principle results at hand, we can notice that in the $k \to \infty$ limit, it is theoretically expected to have only one minimiser of the ferronematic free energy (5.2.0.2) and there is a $k$-independent $L^\infty$ bound (given by (5.2.0.7)) for $\mathbf{Q}, \mathbf{M}$. Moreover, in this limit, one can easily see that the Euler–Lagrange equations (5.2.0.6a)-(5.2.0.6d) reduce to the Laplace equations

$$\begin{aligned} \frac{\mathrm{d}^2 Q_{11}}{\mathrm{d}y^2} &= 0, \quad \frac{\mathrm{d}^2 Q_{12}}{\mathrm{d}y^2} = 0, \\ \frac{\mathrm{d}^2 M_1}{\mathrm{d}y^2} &= 0, \quad \frac{\mathrm{d}^2 M_2}{\mathrm{d}y^2} = 0, \end{aligned} \tag{5.2.0.9}$$

subject to the boundary conditions (5.2.0.4). This Laplace system then admits a unique solution:

$$(\mathbf{Q}^\infty, \mathbf{M}^\infty) = (Q_{11}^\infty, Q_{12}^\infty, M_1^\infty, M_2^\infty) = (-y, 0, -y, 0), \tag{5.2.0.10}$$

where $Q_{12}, M_2$ are zero-valued and $Q_{11}, M_1$ are linear profiles. The solution (5.2.0.10) is also referred to as an *order reconstruction* solution, with only two degrees of freedom $(Q_{11}, M_1)$ reduced from the full four degrees of freedom $(Q_{11}, Q_{12}, M_1, M_2)$. We will discuss this reduced system further in Section 5.3. The convergence result regarding the limit regime $k \to \infty$ is proven in [Dal+21] using the method of sub- and super-solutions and we quote the theorem below, which is to be numerically validated as well in Chapter 6.

**Theorem 5.3.** *[Dal+21] (Convergence result of $k \to \infty$) Assume $k$ is sufficiently large so that the uniqueness result Theorem 5.1 holds. Let $(\mathbf{Q}^k, \mathbf{M}^k)$ be the unique solution of the Euler–Lagrange equations (5.2.0.6a)-(5.2.0.6d) in the admissible space (5.2.0.5), subject to the boundary conditions (5.2.0.4). Then $(\mathbf{Q}^k, \mathbf{M}^k)$ converge to $(\mathbf{Q}^\infty, \mathbf{M}^\infty)$ as $k \to \infty$ with the following estimates:*

$$\forall j = 1, 2, \quad \|Q_{1j}^k - Q_{1j}^\infty\|_\infty \leq \alpha_1 k^{-1}, \ \|M_j^k - M_j^\infty\|_\infty \leq \alpha_2 l^{-1},$$

*for positive constants $\alpha_1, \alpha_2$ independent of $k$.*

**Remark 5.3.** *It implies that when $k$ is sufficiently large, there is only one unique minimiser of the form (5.2.0.10) which gives a linear order reconstruction profile.*

The case of $k \to 0$ is more complicated due to the non-uniqueness of solutions and in fact its convergence information requires more delicate $\Gamma$-convergence analysis. However, some preliminary properties about the limiting profile for $k \to 0$ can be obtained by examining the so-called *bulk minimisers* that minimise the bulk energy (i.e., eliminating all gradient terms in the ferroenematic full energy (5.2.0.2)):

$$\begin{aligned} F_b(Q_{11}, Q_{12}, M_1, M_2) \coloneqq &\left(Q_{11}^2 + Q_{12}^2 - 1\right)^2 + \frac{1}{4}\left(M_1^2 + M_2^2 - 1\right)^2 \\ &- cQ_{11}\left(M_1^2 - M_2^2\right) - 2cQ_{12}M_1M_2. \end{aligned} \tag{5.2.0.11}$$

Substituting the parametrisation

$$\begin{aligned} Q_{11} = \rho\cos(\theta), Q_{12} = \rho\sin(\theta), \\ M_1 = \sigma\cos(\phi), M_2 = \sigma\sin(\phi), \end{aligned} \tag{5.2.0.12}$$

into (5.2.0.11), we can deduce that the minimisers of $F_b$ belong to the set

$$\begin{aligned} \mathcal{M}_{\min} \coloneqq \{(Q_{11}, Q_{12}, M_1, M_2) = &(\rho^*\cos(\theta), \rho^*\sin(\theta), \\ &\sqrt{1 + 2c\rho^*}\cos(\phi), \sqrt{1 + 2c\rho^*}\sin(\phi)) : \\ &\theta = 2\phi + 2z\pi, \text{ for } z \in \mathbb{Z}\}, \end{aligned}$$

where $\rho^*$ is given by (5.2.0.8). Thus, we can define the limiting minimisers for $k \to 0$ as

$$\begin{aligned} \mathbf{Q}^f(c, y) &= \rho^*(\cos(2\phi(y)), \sin(2\phi(y))), \\ \mathbf{M}^f(c, y) &= \sqrt{1 + 2c\rho^*}\,(\cos(\phi(y)), \sin(\phi(y)))\,, \end{aligned} \tag{5.2.0.13}$$

where there are two choices of $\phi$ due to the imposed boundary conditions for $\mathbf{M}^f$:

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}y^2} = 0, \tag{5.2.0.14a}$$

$$\phi(-1) = 0, \phi(1) = \pi \quad \text{or} \quad \phi(-1) = 0, \phi(1) = -\pi, \tag{5.2.0.14b}$$

$$\theta - 2\phi = 2z\pi. \tag{5.2.0.14c}$$

**Remark 5.4.** *It is obvious from the definition (5.2.0.13) of the limiting minimisers for $k \to 0$ that neither $\mathbf{Q}$ nor $\mathbf{M}$ vanishes since $\rho^*$ is nonzero.*

Therefore, we expect that the energy minimisers $(\mathbf{Q}^f, \mathbf{M}^f)$ of the full energy (5.2.0.2) should converge to one of the defined limiting minimisers in (5.2.0.13) almost everywhere as $k \to 0$. The exception happens close to the boundary end points $y = \pm 1$ (due to the incompatible boundary conditions with the limiting minimisers) or at interior points that are associated with jumps in $(2\phi - \theta)$ (since $(2\phi - \theta)$ is only constrained to be an even multiple of $2\pi$ in the $k \to 0$ limit). The numerical verification of this hypothesis is illustrated in Section 6.2.

## 5.3 Reduced model: order reconstruction

The previous section concerns the full ferronematic problem (5.2.0.6a)-(5.2.0.6d) with four degrees of freedom $(\mathbf{Q}, \mathbf{M}) = (Q_{11}, Q_{12}, M_1, M_2)$, i.e., four scalar unknowns. One can observe that profiles with $Q_{12} = M_2 = 0$ can always contribute to a branch of solutions of the Euler–Lagrange equations (5.2.0.6a)-(5.2.0.6d). We refer to these solutions with only two degrees of freedom, $(\mathbf{Q}, \mathbf{M}) = (Q_{11}, 0, M_1, 0)$ as *order reconstruction* (OR) solutions. This leads to the following reduced functional, denoted as the *OR* energy, from the full energy (5.2.0.2):

$$\begin{aligned} E(Q_{11}, M_1) \coloneqq \int_{-1}^{1} \Bigg\{ & \frac{k}{2}\left(\frac{\mathrm{d}Q_{11}}{\mathrm{d}y}\right)^2 + \frac{k}{2}\left(\frac{\mathrm{d}M_1}{\mathrm{d}y}\right)^2 + (Q_{11}^2 - 1)^2 \\ & + \frac{1}{4}\left(M_1^2 - 1\right)^2 - cQ_{11}M_1^2 \Bigg\} \, \mathrm{d}y, \end{aligned} \tag{5.3.0.1}$$

subject to the boundary conditions

$$\begin{aligned} Q_{11}(-1) = M_1(-1) = 1, \\ Q_{11}(1) = M_1(1) = -1, \end{aligned} \tag{5.3.0.2}$$

in the admissible space

$$\mathcal{A}'_f = \left\{ Q_{11}, M_1 \in H^1(\Omega, \mathbb{R}), Q_{11} \text{ and } M_1 \text{ satisfy the boundary conditions (5.3.0.2)} \right\}. \tag{5.3.0.3}$$

Consequently, OR solutions are classical solutions of the following coupled ordinary differential equations,

$$\begin{aligned} k_1 \frac{\mathrm{d}^2 Q_{11}}{\mathrm{d}y^2} &= 4Q_{11}(Q_{11}^2 - 1) - cM_1^2, \\ k_2 \frac{\mathrm{d}^2 M_1}{\mathrm{d}y^2} &= M_1(M_1^2 - 1) - 2cQ_{11}M_1. \end{aligned} \tag{5.3.0.4}$$

**Remark 5.5.** *The reason why we are interested in the OR solutions is not only due to a reduction of unknowns that benefits our subsequent analysis, but also due to one of their special solutions, the so-called* domain wall *(i.e.,* $\mathbf{Q} = \mathbf{M} = \mathbf{0}$*) profiles that separate* polydomains*, i.e., distinctly ordered domains. A nematic (resp. magnetic) domain wall is a point* $y = y^* \in (-1, 1)$ *such that* $\mathbf{Q}(y^*) = (Q_{11}(y^*), Q_{12}(y^*)) = 0$ *(resp.* $\mathbf{M}(y^*) = 0$*).*

One can note from our applied inhomogeneous boundary conditions (5.3.0.2) for $Q_{11}$ (resp. $M_1$) that there must exist an interior point, $y^* \in (-1, 1)$ such that $Q_{11}(y^*) = 0$ (resp. $M_1(y^*) = 0$) since $Q_{11}(-1) = M_1(-1) = 1$ and $Q_{11}(1) = M_1(1) = -1$. That is to say, we expect to see nematic and magnetic interior domain walls for the solutions $(\mathbf{Q}, \mathbf{M})$. Moreover, these domain walls can occur at different points (which we shall demonstrate in Chapter 6). In fact, using the parameterisation

$$\begin{aligned} Q_{11} &= \rho\cos(\theta), Q_{12} = \rho\sin(\theta), \\ M_1 &= \sigma\cos(\phi), M_2 = \sigma\sin(\phi), \end{aligned} \tag{5.3.0.5}$$

we can notice that $Q_{12} = M_2 = 0$ implies $\theta = \mathrm{z}_1\pi$ and $\phi = \mathrm{z}_2\pi$ for some integers $\mathrm{z}_1, \mathrm{z}_2$. Furthermore, due to the imposed inhomogeneous boundary conditions, there is necessarily a domain wall in $\mathbf{Q}$ such that $\theta = 2\mathrm{z}_1\pi$ on one side of the domain wall containing the end point $y = -1$, and $\theta = (2\mathrm{z}_2 + 1)\pi$ (for some integers $\mathrm{z}_1, \mathrm{z}_2$) on the other side of the domain wall containing the end point $y = 1$; analogously, there is a domain wall in $\mathbf{M}$ that separates two polydomains, with $\phi = 2\mathrm{z}_1\pi$ and $\phi = (2\mathrm{z}_2 + 1)\pi$ for some integers $\mathrm{z}_1$ and $\mathrm{z}_2$ respectively.

Similarly, there are some qualitative results regarding the existence, uniqueness, maximum principle and instability of the OR solutions, proven in detail by Dalby & Majumdar [Dal+21]. We again quote the following result so that we can numerically verify it in Chapter 6.

**Theorem 5.4.** *[Dal+21] (Uniqueness and maximum principle) For sufficiently large $k$ and a fixed positive $c$, the OR solution $(\mathbf{Q}^{OR}, \mathbf{M}^{OR}) := (Q_{11}^*, 0, M_1^*, 0)$ is the unique critical point, and hence, global minimiser of the energy (5.2.0.2), as in Theorem 5.1. Moreover, we have the $L^\infty$ bound*

$$|Q_{11}(y)| \leq \rho^*, \ M_1^2(y) \leq 1 + 2c\rho^* \quad \forall y \in [-1, 1], \tag{5.3.0.6}$$

*where $\rho^*$ is given by (5.2.0.8).*

It follows from Theorem 5.4 that the OR solution is the global minimiser for sufficiently large $k$ and there is an $L^\infty$ bound for $Q_{11}, M_1$. However, the OR solution loses its stability as $k$ decreases, similarly to the study of the pure nematic case (i.e., $c = 0$) in [Lam14; CMS19]. We include the result below for self containment.

**Theorem 5.5.** *[Dal+21] (Instability of the OR solution) For sufficiently small $k$ and a fixed positive $c$, the OR energy minimiser, $(\mathbf{Q}^{OR}, \mathbf{M}^{OR})$, is an unstable critical point of the full energy (5.2.0.2), in the full admissible space (5.2.0.5).*

The convergence result for $k \to 0$ limiting regime is given by Dalby & Majumdar in [Dal+21] using $\Gamma$-convergence methods by directly following [WCM19, Proposition 4.1]. More precisely, when $k$ is very small, the minimisers to the OR energy (5.3.0.1) is closely related to the OR bulk minimisers:

$$\mathbf{p}^* = (Q_{11}, M_1) = \left(\rho^*, \sqrt{1 + 2c\rho^*}\right), \quad \text{or } \mathbf{p}^{**}(Q_{11}, M_1) = \left(\rho^*, -\sqrt{1 + 2c\rho^*}\right). \tag{5.3.0.7}$$

**Remark 5.6.** *Note that these profiles in (5.3.0.7) are not compatible with the boundary conditions (5.3.0.2). Thus, there are necessarily boundary layers close to $y = \pm 1$ in the OR energy minimisers as $k \to 0$.*

We do not include a detailed description of the convergence results as $k \to 0$, however, we can numerically demonstrate that the OR energy minimiser converges to $\mathbf{p}^*$ almost everywhere as it has the least transition costs. To see this, we need to define the non-negative OR bulk energy:

$$\tilde{f}(Q_{11}, M_1) := \left(Q_{11}^2 - 1\right)^2 + \frac{1}{4}\left(M_1^2 - 1\right)^2 - cQ_{11}M_1^2 - \beta(c) \geq 0, \tag{5.3.0.8}$$

where the $c$-dependent constant $\beta(c)$ is the minimum value of the OR bulk potential.

Following [Bra06] and [WCM19], we let $\mathbf{p} = (Q_{11}, M_1)$ and define the following metric $\omega$ (which is in fact the geodesic distance associated with the Riemannian metric $\tilde{f}^{1/2}$ [WCM19]) in the $\mathbf{p}$-plane, for any two points $\mathbf{p}_0, \mathbf{p}_1 \in \mathbb{R}^2$:

$$\begin{aligned}\omega\left(\mathbf{p}_0, \mathbf{p}_1\right) = \inf\Bigg\{ \int_{-1}^{1} \tilde{f}^{1/2}\left(\mathbf{p}(t)\right) \left|\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t}\right| \,\mathrm{d}t : \mathbf{p}(t) \in C^1\left([-1,1]; \mathbb{R}^2\right), \\ \mathbf{p}(-1) = \mathbf{p}_0, \mathbf{p}(1) = \mathbf{p}_1 \Bigg\}.\end{aligned} \tag{5.3.0.9}$$

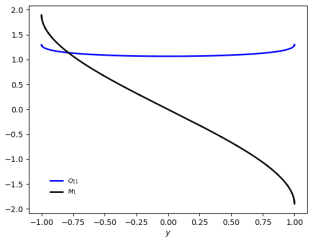
**Remark 5.7.** *It is obvious to see that this metric is degenerate (i.e., zero-valued) as* $\tilde{f}(\mathbf{p}) = 0$ *for* $\mathbf{p} = \mathbf{p}^* = (\rho^*, \sqrt{1 + 2c\rho^*})$ *and* $\mathbf{p} = \mathbf{p}^{**} = (\rho^*, -\sqrt{1 + 2c\rho^*})$. *Despite such degeneracy, the infimum in (5.3.0.9) can be attained for arbitrary* $\mathbf{p}_0$ *and* $\mathbf{p}_1$ *(see [Bra06, Lemma 9] and [WCM19]).*

In fact, the metric $\omega(\mathbf{p}_1, \mathbf{p}_2)$ accounts for the transition costs between the profiles $\mathbf{p}_1$ and $\mathbf{p}_2$. Thus, one can deduce the energetically preferable minimisers, say $\mathbf{p}^k = (Q_{11}^k, M_1^k)$ by minimising the total transition costs that is a sum of $\omega(\mathbf{p}^k, \mathbf{p}_b(1))$, $\omega(\mathbf{p}^k, \mathbf{p}_b(-1))$. Here, $\mathbf{p}_b(1) = (-1, -1)$ and $\mathbf{p}_b(-1) = (1, 1)$ denote the boundary profiles of $(Q_{11}, M_1)$.

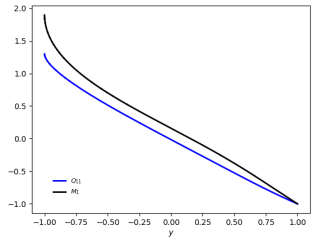
According to [WCM19, Section 5.1], the distance $\omega(\cdot, \cdot)$ can be calculated alternatively by

$$\omega(\mathbf{p}^*, \mathbf{p}^{**}), \omega(\mathbf{p}^*, \mathbf{p}_b(1)), \omega(\mathbf{p}^{**}, \mathbf{p}_b(-1)), \omega(\mathbf{p}^*, \mathbf{p}_b(-1)), \omega(\mathbf{p}^{**}, \mathbf{p}_b(1)). \tag{5.3.0.10}$$

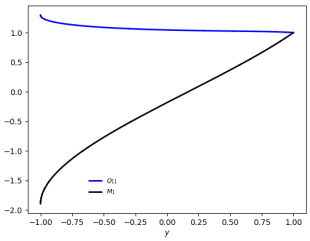
$$\begin{aligned}\omega\left(\mathbf{p}_0, \mathbf{p}_1\right) = \inf\Bigg\{ \left(\int_{-1}^{1} \tilde{f}\left(\mathbf{p}(t)\right) \left|\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t}\right|^2 \,\mathrm{d}t\right)^{1/2} : \mathbf{p}(t) \in C^1\left([-1,1]; \mathbb{R}^2\right), \\ \mathbf{p}(-1) = \mathbf{p}_0, \mathbf{p}(1) = \mathbf{p}_1 \Bigg\}.\end{aligned}$$



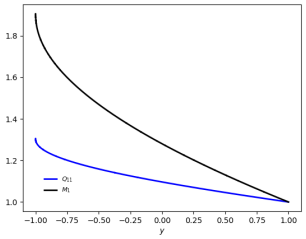
$\omega(\mathbf{p}^*, \mathbf{p}^{**}) \approx 3.008.$



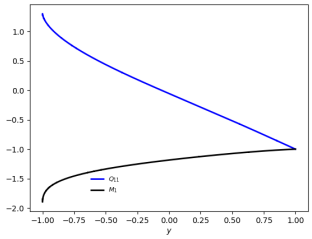
$\omega(\mathbf{p}^*, \mathbf{p}_b(1)) \approx 3.967.$



$\omega(\mathbf{p}^{**}, \mathbf{p}_b(-1)) \approx 2.577.$



$\omega(\mathbf{p}^*, \mathbf{p}_b(-1)) \approx 0.455.$



$\omega(\mathbf{p}^{**}, \mathbf{p}_b(1)) \approx 2.591.$

**Figure 5.1:** The profiles of $\mathbf{p}$ and their corresponding transition costs in (5.3.0.10).

The profiles of $\mathbf{p} = (Q_{11}, M_1)$ for each case in (5.3.0.10) are shown in Figure 5.1 which indicates that

$$\omega(\mathbf{p}^*, \mathbf{p}_b(-1)) < \omega(\mathbf{p}^{**}, \mathbf{p}_b(-1)) < \omega(\mathbf{p}^{**}, \mathbf{p}_b(1)) < \omega(\mathbf{p}^*, \mathbf{p}^{**}) < \omega(\mathbf{p}^*, \mathbf{p}_b(1)).$$

Using the computed values of those transition costs, it is clear that the OR energy minimiser converges to $\mathbf{p}^*$ almost everywhere, except close to the boundary end points $y = \pm 1$. Moreover, no interior jumps are expected in the OR minimiser.

## 5.4 Summary

In this chapter, we used the LdG $\mathbf{Q}$-tensor theory to investigate the solution structure of a ferronematic problem. We introduced both the full and reduced models of ferronematics and quoted some theoretical results proven in [Dal+21]. Our aim in the next chapter is twofold: first, verify these theoretical results computationally, and second, provide more information on the solution landscapes via numerical experiments.

# 6

# Numerical verifications for ferronematics

## Contents



In this chapter, we perform numerical experiments to validate the theoretical results proven in [Dal+21] and quoted in the previous chapter, and understand the interplay between the elastic constant $k$ and the coupling parameter $c$ for the solution landscapes. For simplicity, we fix the scaling $\xi = 1$ throughout this chapter.

For the visualisation, we plot the (headless) director $\mathbf{n}$ with rods and the normalised magnetisation vector field $\mathbf{m} = \frac{\mathbf{M}}{|\mathbf{M}|}$ with arrows.

## 6.1 Solver details

The nonlinear solver is deemed to have converged when the Euclidean norm of the residual falls below $10^{-8}$, or reduces from its initial value by a factor of $10^{-6}$, whichever comes first. For the inner solver, the linearised systems are solved using the sparse LU factorisation library MUMPS [ADL00]. We partition the whole

interval $[-1, 1]$ into 1000 equi-distant subintervals and numerically approximate the solutions using $\mathbb{P}^1$ finite elements.

**Code availability.** For reproducibility and more details of the implementation, we have archived the solver code [Xia21b] and the exact version of Firedrake [Fir21a] used to produce the numerical results of this work. An installation of Firedrake with components matching those used in this chapter can be obtained by following the instructions at https://www.firedrakeproject.org/download.html with

```
python3 firedrake-install --doi 10.5281/zenodo.4449535
```

Defcon version #aaa4ef should then be installed, as described in https://bitbucket.org/pefarrell/defcon/.

## 6.2 Solutions of the full problem

In this section, we focus on the full problem (5.2.0.6a)-(5.2.0.6d) with four scalar-valued solution variables $(Q_{11}, Q_{12}, M_1, M_2)$. We only present the result with small $k_1 = k_2 = k = 0.01$ (while varying the coupling $c$) here, since Theorem 5.4 implies that the OR solution branch remains as the unique minimiser of the full problem for a sufficiently large $k$ and the OR solution will be reported later in Section 6.3. In fact, we shall see the uniqueness of solution for large $k$ in the next section.

We first take the coupling parameter $c = 1$ and present four examples of stable stationary profiles $(Q_{11}, Q_{12}, M_1, M_2)$ in Figure 6.1. One can compare the $L^\infty$ bound (5.2.0.7) with the computed values of vector norms $|\mathbf{Q}| = \sqrt{Q_{11}^2 + Q_{12}^2}$ and $|\mathbf{M}| = \sqrt{M_1^2 + M_2^2}$, and note that the pointwise maximum principle given by Theorem 5.2 is respected. By Remark 5.4, we expect that there is no interior domain wall with $|\mathbf{Q}| = |\mathbf{M}| = 0$, for small $k$, which is indeed noticeable from the presented solution profiles. Moreover, we can see that Solutions 1, 2 and 3 in Figure 6.1 only have boundary layers with constant $|\mathbf{Q}|, |\mathbf{M}|$-profiles in the interior domain, whereas Solution 4 has an interior non-zero local minimum (thus an interior jump) in $|\mathbf{Q}|$ and $|\mathbf{M}|$. In addition, Solutions 1 and 2 only differ in their orientational $\mathbf{m}$-patterns (more precisely, possessing opposite signs of $M_2$) and they are the energy minimisers

having the same energy value, while Solutions 3 and 4 are non-minimising stable critical points of the full energy (5.2.0.2).

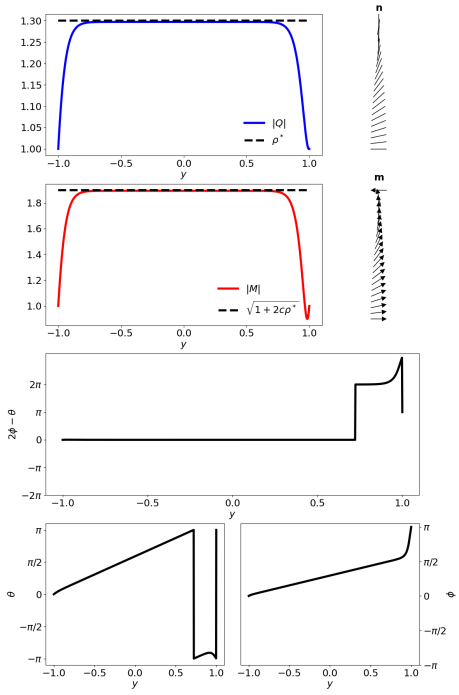
Moreover, we compute the values of orientational angles $\theta$ and $\phi$, defined as

$$\theta = \arctan\left(\frac{Q_{12}}{Q_{11}}\right), \ \phi = \arctan\left(\frac{M_2}{M_1}\right) \tag{6.2.0.1}$$

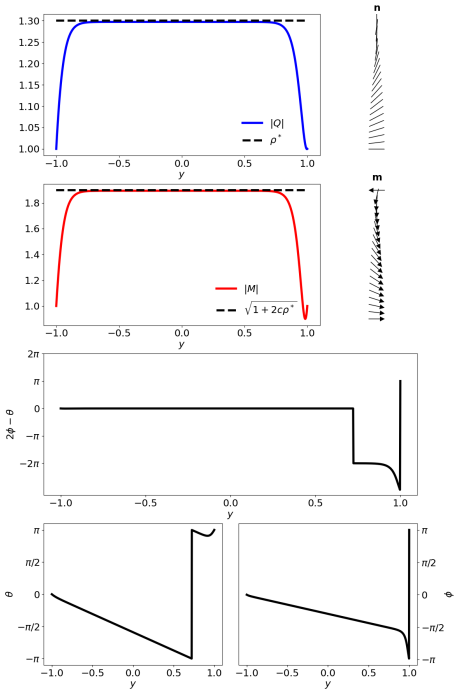
for each numerical solution profile $(Q_{11}, Q_{12}, M_1, M_2)$, so to verify the relation (5.2.0.14), in particular the constraint (5.2.0.14c). It can be seen from Figure 6.1 that $|\mathbf{Q}| \to \rho^*, |\mathbf{M}| \to 1 + 2c\rho^*$ for the energy minimisers (Solutions 1 and 2), whereas $(2\phi - \theta)$ tends to be an even multiple of $\pi$ almost everywhere, except close to the end point $y = 1$. Furthermore, we plot the separate values of $\theta$ and $\phi$ to demonstrate the linear behaviour consistent with (5.2.0.14) for $\phi$ and thus $\theta$ as $k \to 0$. This linearity of $\theta$ and $\phi$ can be seen in Figure 6.1 except around the local minima and boundary layers.

Now, we repeat the simulations for $c = 5$. Two stable stationary profiles are illustrated in Figure 6.2. Again, we observe that $|\mathbf{Q}| \to \rho*$ and $|\mathbf{M}|^2 \to 1 + 2c\rho^*$ almost everywhere, as expected from the maximum principle Theorem 5.2. Here, Solution 2 has lower energy than Solution 1, since Solution 1 has more jumps in $|\mathbf{Q}|$ and $|\mathbf{M}|$ than Solution 2. Further, $(2\phi - \theta)$ is an even multiple of $\pi$ almost everywhere, with the jumps being associated with the jumps in $|\mathbf{Q}|$ and $|\mathbf{M}|$, thus verifying the constraint (5.2.0.14c). Additionally, we plot $\phi$ and $\theta$ in Figure 6.2, and observe almost linear profiles of $\theta$ and $\phi$, except around the local minima and the boundary layers.

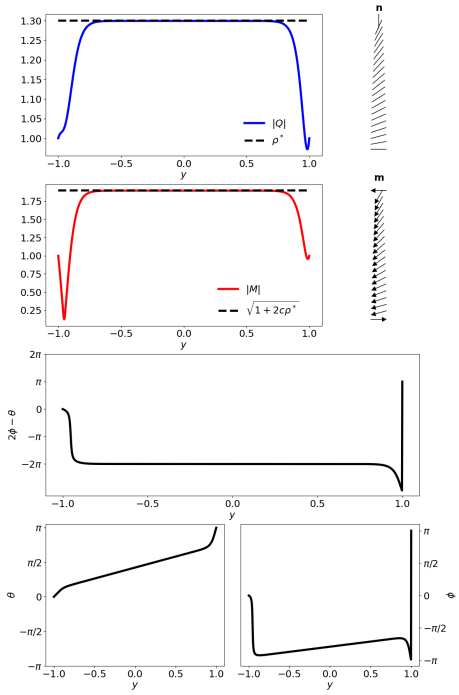
To summarise, the numerical experiments in this section and the theoretical heuristics in (5.2.0.14) suggest that there are at least two energy minimisers, characterised by $(\rho_1, \sigma_1, \theta_1, \phi_1)$ and $(\rho_2, \sigma_2, \theta_2, \phi_2)$ of the full ferronematic energy (5.2.0.2) in the $k \to 0$ limit, such that $\rho_1, \rho_2 \to \rho^*$, $\sigma_1^2, \sigma_2^2 \to 1 + 2c\rho^*$ almost everywhere away from the boundary plates $y = \pm 1$. Moreover, it holds that $\theta_2 = -\theta_1$, $\phi_2 = -\phi_1$ and $2\phi_{1,2} - \theta_{1,2}$ an even multiple of $\pi$ except near $y = 1$ or close to some local jumps of $\mathbf{Q}$ and $\mathbf{M}$. The two energy minimisers only differ in the sense of rotation, in $\mathbf{n}$ and $\mathbf{m}$, between $y = -1$ and $y = 1$.



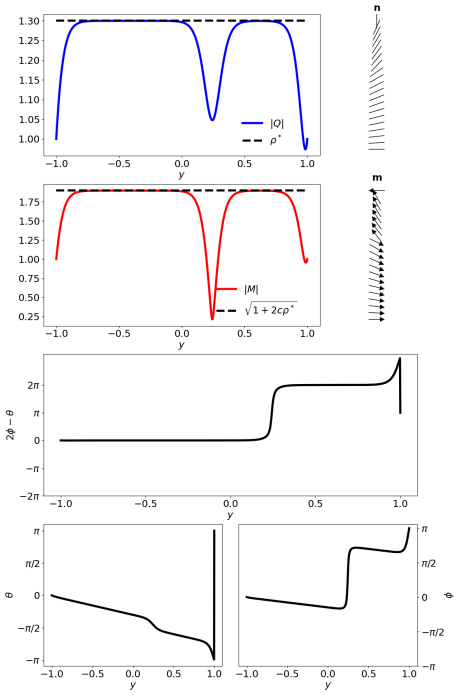

(Solution 1; stable)




(Solution 2; stable)

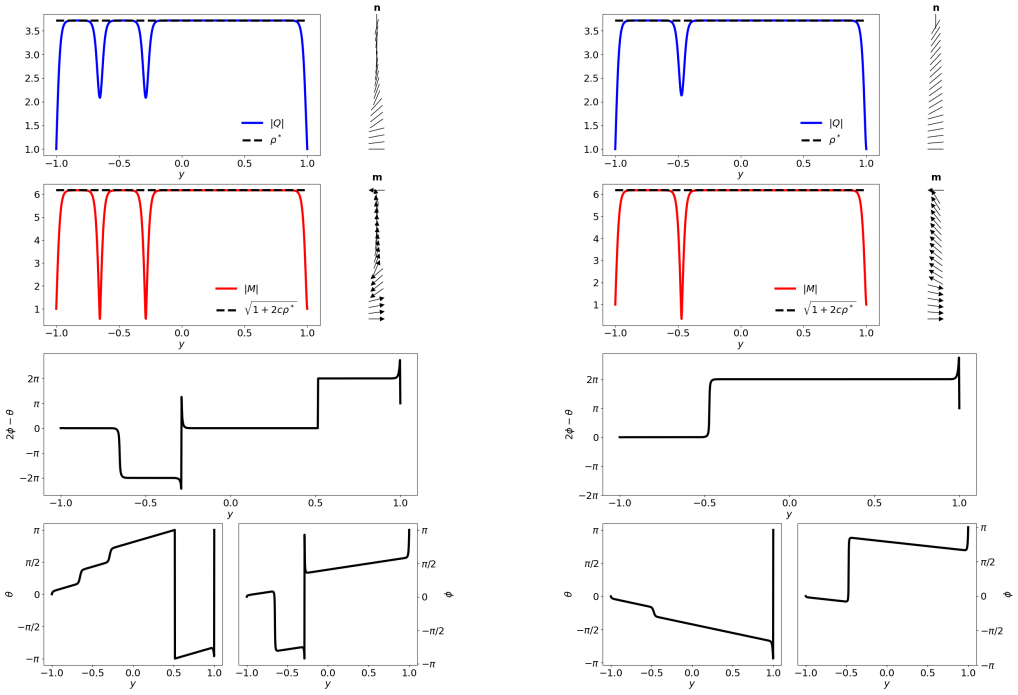



(Solution 3; stable)




(Solution 4; stable)

**Figure 6.1:** Four stable stationary profiles, $(Q_{11}, Q_{12}, M_1, M_2)$, of (5.2.0.2) with $k = 0.01$ and $c = \xi = 1$, along with plots of $(2\phi - \theta)$, $\theta$, and $\phi$ to verify the relation (5.2.0.14). Solutions 1 and 2 have the lowest full energy value (5.2.0.2).




(Solution 1; stable) (Solution 2; stable)

**Figure 6.2:** Two examples of stable stationary profiles $(Q_{11}, Q_{12}, M_1, M_2)$ of the full energy (5.2.0.2) with $k = 0.01$, $c = 5$ and $\xi = 1$, along with plots of $(2\phi - \theta)$, $\theta$, and $\phi$ to verify the relation (5.2.0.14). Solution 2 has lower energy than Solution 1.

## 6.3 Solutions of the reduced problem

By the definition of the OR solution in Section 5.3, we know it is fully characterised by two degrees of freedom $(Q_{11}, M_1)$ of the boundary-value problem (5.3.0.4) while $Q_{12} = M_2 = 0$ always holds. We now numerically investigate the limiting behaviours of the OR solution for $k \to 0$ and $k \to \infty$ illustrated in Section 5.3.

As $k \to \infty$, recall Theorem 5.1 to deduce that the OR solution branch is approximately given by $\left(\mathbf{Q}^{OR}, \mathbf{M}^{OR}\right) \approx (-y, 0, -y, 0)$, for a fixed $c$, and that $\left(\mathbf{Q}^{OR}, \mathbf{M}^{OR}\right)$ is the unique minimiser of both the OR energy (5.3.0.1) and the full energy (5.2.0.2). In Figure 6.3, we plot the OR solution of (5.3.0.4) for $c = 1$ and $k = 10$. The profile is indeed linear, and we do not numerically obtain any other solutions, supporting the uniqueness result in this regime. We notice that the OR solution vanishes at the channel centre $y = 0$, i.e., $Q_{11}(0) = M_1(0) = 0$, and thus both the nematic and magnetic domain walls coincide at $y = 0$. Therefore, the normalised magnetisation vector $\mathbf{m}$ and director $\mathbf{n}$ have a jump discontinuity at

$y = 0$. In fact, $\mathbf{m}$ jumps from $\mathbf{m} = (1, 0)$ for $y < 0$ to $\mathbf{m} = (-1, 0)$ for $y > 0$, while $\mathbf{n}$ jumps from $\mathbf{n} = (1, 0)$ (modulo a sign) for $y < 0$ to $\mathbf{n} = (0, 1)$ (modulo a sign) for $y > 0$. Hence, the nematic and magnetic domain walls at $y = 0$ separate two distinct polydomains in $\mathbf{n}$ and $\mathbf{m}$, respectively. We also plot the pointwise $L^\infty$ bound (5.3.0.6) as blue solid lines in Figure 6.3, and as expected, this bound is indeed respected.

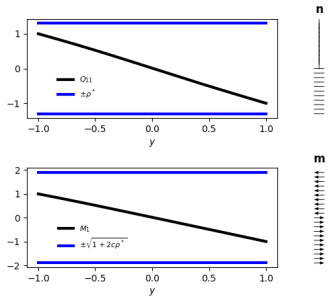



**Figure 6.3:** The only (stable) solution of (5.3.0.1) for $c = \xi = 1$, and $k = 10$.

As $k \to 0$ with fixed positive $c$, the OR solution is not unique anymore and we expect to see that $(Q_{11}, M_1^2) \to (\rho^*, 1 + 2c\rho^*)$ uniformly everywhere away from the edges $y = \pm 1$, for the minimiser of the OR energy (5.3.0.1). Of course, all OR solutions are unstable critical points of the full energy (5.2.0.2) in the $k \to 0$ limit, as shown in Theorem 5.5. We now numerically corroborate these theoretical results with fixed $k = 0.01$ and $\xi = 1$.

In Figure 6.4, we present four example solutions by taking $c = 1$. In fact, they are all unstable critical points of the full energy (5.2.0.2) whilst being stable critical points of the OR energy (5.3.0.1) (in the sense that the Hessian of second variation of the OR energy about these critical points has only positive eigenvalues). Consistent with the discussion of the convergence regime for $k \to 0$ in Section 5.3, these solution profiles $(Q_{11}, M_1)$ have a domain wall in $\mathbf{Q}$ near the end point $y = 1$, where $Q_{11}$ jumps from $Q_{11} = \rho^* > 1$ to the boundary value $Q_{11})(1) = -1$. Analogously, we can see that all solution profiles illustrated in Figure 6.4 have a boundary layer close the other end point $y = -1$, within which $Q_{11}$ jumps from $Q_{11}(-1) = 1$ to $Q_{11} = \rho^* > 1$. However, we should note that this boundary layer does not contain a domain wall with $Q_{11} = 0$. An additional observation is the presence of interior

transition layers in $M_1$ (near the center $y = 0$) in Solutions 3 and 4 of Figure 6.4. The $L^\infty$ bounds (5.3.0.6) (blue solid line) for $|Q_{11}|$ and $|M_1|$ are also satisfied.

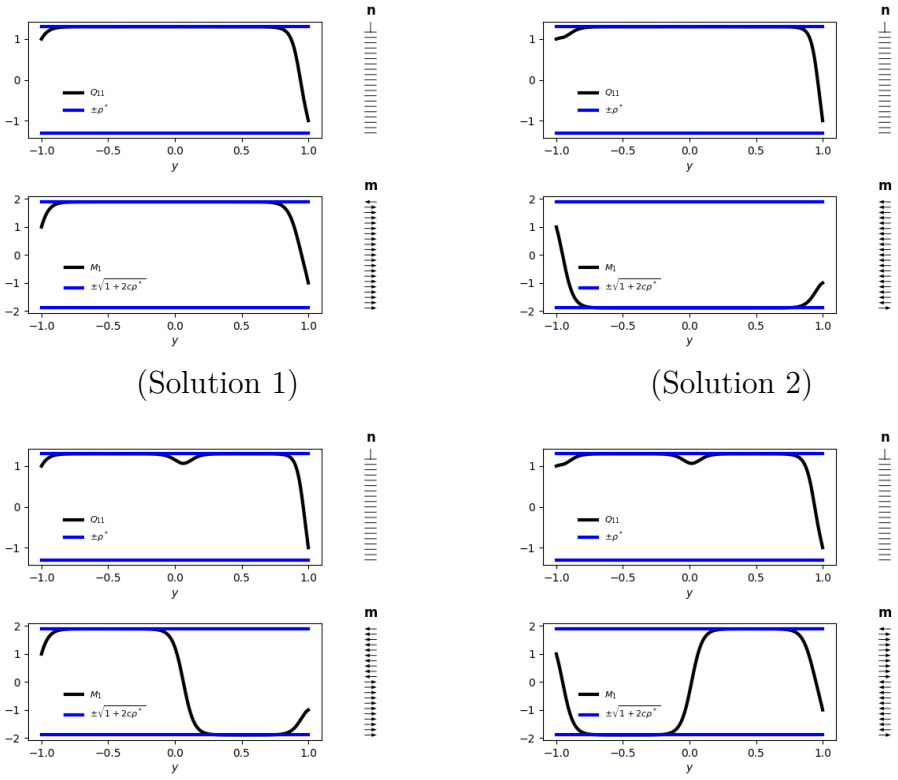



**Figure 6.4:** Four OR solution profiles with $c = \xi = 1$ and $k = 0.01$. Solution 1 is the OR energy minimiser. (5.3.0.1).

In Figure 6.5, we plot the stable stationary profiles of the OR energy (5.3.0.1) for a larger value $c = 5$, whereas they are unstable critical points of the full energy (5.2.0.2). Indeed, each of the solutions in Figure 6.5 has one unstable eigendirection, in the context of the full energy (5.2.0.2). The two profiles in Figure 6.5, have boundary layers near $y = \pm 1$, and essentially differ in the sign of $M_1$ in the interior; $Q_{11}$ only vanishes near $y = 1$, so that we have a nematic domain wall close to the end point $y = 1$. On the other hand, $M_1$ can vanish either near $y = -1$ or near $y = 1$, so that the corresponding magnetic domain wall can occur near either boundary. We also note that Solution 2 in Figure 6.5 is the OR energy minimiser which indeed converges to $\mathbf{p}^*$ almost everywhere except close to the boundary plates $y = \pm 1$. This verifies the heuristics explained by computing the transition costs in Figure 5.1.

Additionally, we present two more solution examples with interior transition layers for $M_1$ in Figure 6.6 with $c = 5$, where single and multiple interior transition layers in $M_1$ are observed. They are also stable critical points of the OR energy (5.3.0.1), and unstable critical points of the full energy (5.2.0.2). The transition layers in $M_1$ necessarily contain a magnetic domain wall with $M_1 = 0$, and these interior magnetic domain walls are not accompanied by associated nematic domain walls. Moreover, solutions with interior transition layers have higher OR energy (5.3.0.1) than solutions without interior transition layers in Figure 6.5, since each transition layer requires an energetic cost of $\omega(p^*, p^{**})$. Again, the $L^\infty$ bound (5.3.0.6) is satisfied for the solutions illustrated in Figure 6.6.

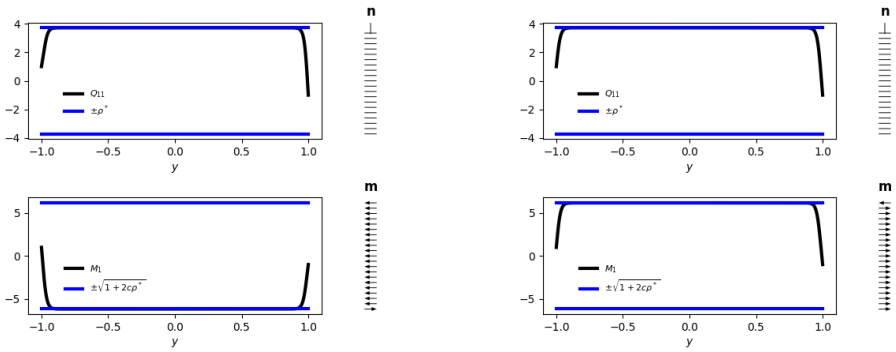



**Figure 6.5:** Two stable OR critical points of (5.3.0.1), for $c = 5$, $\xi = 1$ and $k = 0.01$. The right profile has lower OR energy than the left profile and the solutions in Figure 6.6.

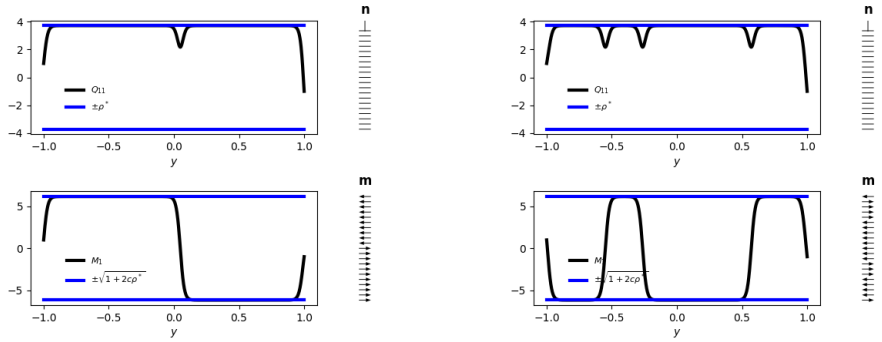



**Figure 6.6:** Two stable OR solutions with single (left) and multiple (right) interior transition layers for $c = 5$, $\xi = 1$ and $k = 0.01$. The left profile has lower OR energy than the right profile.

All above numerical experiments show that the domain walls in the OR energy minimisers migrate from the channel centre to the channel boundaries at $y = \pm 1$, as $k$ decreases. Therefore, we can manipulate the location and multiplicity of nematic and magnetic domain walls in the OR solutions by varying $k$.

## 6.4 Asymptotics checking for $k \to \infty$

We then theoretically and numerically illustrate the asymptotic behaviour as $k \to \infty$ in this section, to investigate the convergence to the unique OR minimiser $(\mathbf{Q}^{OR}, \mathbf{M}^{OR}) = (-y, 0, -y, 0)$ in this limit regime.

As $k \to \infty$, we can compute useful asymptotic expansions of the OR solution branch for large $k$ and small $c$, by setting $k = \frac{1}{c}$ in the Euler–Lagrange equations (5.2.0.6a)-(5.2.0.6d) and expanding around $(\mathbf{Q}^{\infty}, \mathbf{M}^{\infty})$ as shown below:

$$Q_{11}(y) = -y + cf_2(y) + c^2 f_3(y) + \mathcal{O}(c^3), \quad M_1(y) = -y + cf_2^*(y) + c^2 f_3^*(y) + \mathcal{O}(c^3).$$

Substituting the above into (5.2.0.6a) and (5.2.0.6d) (with $k = \frac{1}{c}$) yields

$$\frac{d^2 f_1}{dy^2} + c\frac{d^2 f_2}{dy^2} + c^2\frac{d^2 f_3}{dy^2} = 4c\left(f_1^3 - f_1\right) + c^2\left(12 f_1^2 f_2 - 4f_2 - (f_1^*)^2\right) + \mathcal{O}(c^3) \quad (6.4.0.1a)$$

$$\frac{d^2 f_1^*}{dy^2} + c\frac{d^2 f_2^*}{dy^2} + c^2\frac{d^2 f_3^*}{dy^2} = c\left((f_1^*)^3 - f_1^*\right) + c^2\left(3(f_1^*)^2 f_2^* - f_2^* - \frac{2}{\xi} f_1 f_1^*\right) + \mathcal{O}(c^3). \quad (6.4.0.1b)$$

By equating powers of $c$, we solve the computed second order ordinary differential equations for $f_2, f_3, f_2^*, f_3^*$, subject to the boundary conditions $f_2(-1) = f_2(1) = f_3(-1) = f_3(1) = 0$ and $f_2^*(-1) = f_2^*(1) = f_3^*(-1) = f_3^*(1) = 0$. This gives

$$c^0 : \frac{d^2 f_1}{dy^2} = 0 \Rightarrow f_1(y) = -y$$

$$c^1 : \frac{d^2 f_2}{dy^2} = 4(f_1^2 - 1)f_1 \Rightarrow f_2(y) = -\frac{1}{5}y^5 + \frac{2}{3}y^3 - \frac{7}{15}y$$

$$c^2 : \frac{d^2 f_3}{dy^2} = 4(3f_1^2 - 1)f_2 - (f_1^*)^2 \Rightarrow f_3(y) = p(y),$$

and

$$c^0 : \frac{d^2(f_1^*)}{dy^2} = 0 \Rightarrow f_1^*(y) = -y,$$
$$c^1 : \frac{d^2(f_2^*)}{dy^2} = ((f_1^*)^2 - 1)f_1^* \Rightarrow f_2^*(y) = -\frac{1}{20}y^5 + \frac{1}{6}y^3 - \frac{7}{60}y,$$
$$c^2 : \frac{d^2(f_3^*)}{dy^2} = 3(f_1^*)^2 f_2^* - f_2^* - \frac{2}{\xi}f_1 f_1^* \Rightarrow f_3(y) = q(y).$$

Here,

$$p(y) = -\frac{1}{30}y^9 + \frac{22}{105}y^7 - \frac{31}{75}y^5 - \frac{1}{12}y^4 + \frac{14}{45}y^3 - \frac{233}{3150}y + \frac{1}{12},$$

and

$$q(y) = -\frac{1}{480}y^9 + \frac{11}{840}y^7 - \frac{31}{1200}y^5 - \frac{1}{6}y^4 + \frac{7}{360}y^3 - \frac{233}{50400}y + \frac{1}{6}.$$

Thus, the expansions for $Q_{11}$ and $M_1$ are

$$Q_{11}(y) = -y + c\left(-\frac{1}{5}y^5 + \frac{2}{3}y^3 - \frac{7}{15}y\right) + c^2 p(y) + \mathcal{O}(c^3), \tag{6.4.0.2}$$

and

$$M_1(y) = -y + c\left(-\frac{1}{20}y^5 + \frac{1}{6}y^3 - \frac{7}{60}y\right) + c^2 q(y) + \mathcal{O}(c^3), \tag{6.4.0.3}$$

for $k = \frac{1}{c}$ and $k \gg 1$.

We now check the validity of these expansions, (6.4.0.2) and (6.4.0.3), numerically. To this end, we compare $(\cdot)^{num}$ and $(\cdot)^{asymp}$ in the $L^\infty$-norm, where $(\cdot)^{num}$ is the numerical solution and $(\cdot)^{asymp}$ corresponds to the asymptotic expansion, depending on the truncation of the expansions in (6.4.0.2) and (6.4.0.3). For instance, a first order truncation (with respect to $c$) yields

$$Q_{11}^{asymp} = -y + c\left(-\frac{1}{5}y^5 + \frac{2}{3}y^3 - \frac{7}{15}y\right),$$
$$M_1^{asymp} = -y + c\left(-\frac{1}{20}y^5 + \frac{1}{6}y^3 - \frac{7}{60}y\right).$$

The left-hand column of Figure 6.7 shows a first order convergence by truncating the expansions up to $\mathcal{O}(c^0)$, whilst a first order truncation leads to a second order convergence as shown in the middle column of Figure 6.7 and finally, in the right-hand column, a truncation up to $\mathcal{O}(c^2)$ demonstrates a third order convergence with respect to $c$, for both $Q_{11}$ and $M_1$.

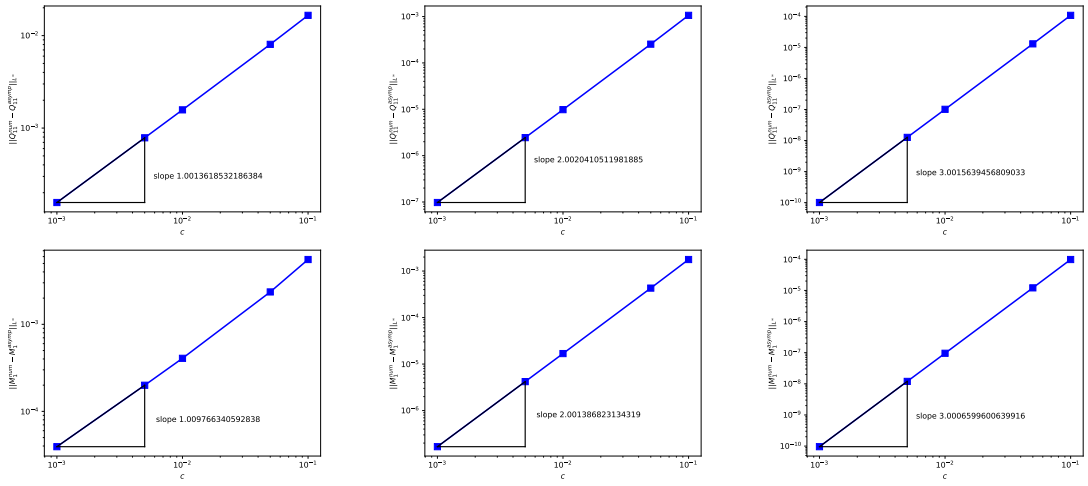



**Figure 6.7:** Log-log plots of $\|Q_{11}^{num} - Q_{11}^{asymp}\|_\infty$ (top row) and $\|M_1^{num} - M_1^{asymp}\|_\infty$ (bottom row). Left: truncating asymptotic expansions (6.4.0.2) and (6.4.0.3) at $c^0$. Middle: truncating asymptotic expansions at $c^1$. Right: truncating asymptotic expansions at $c^2$.

## 6.5 Bifurcation diagrams

The proceeding sections examine the behaviour of the solution profiles for certain specific choices of parameters. One can obtain further information about the solutions to the Euler–Lagrange equations (5.2.0.6a)-(5.2.0.6d) by continuing the parameter and plotting bifurcation diagrams for the parameter space of interest. We thus perform numerical experiments as we continue the coupling parameter $c$ or the elastic constant $k$.

The first experiment regarding varying $c$ is illustated in Figure 6.8. Here, we choose $k_1 = k_2 = k \in [0.2, 3.0]$ with fixed step size 0.01 and $c = 1$. It can be seen that there is only one stable OR solution for $k \in [1.25, 3.0]$, being the energy minimiser of the full energy (5.2.0.2). For $k \approx 1.25$, there is a pitchfork bifurcation consisting of two stable branches and one unstable OR branch (see Figure 6.9 for an illustration of these three solutions at $k = 1$). In fact, the two stable solutions (Solutions 1 and 3 in Figure 6.9) differ by the sign of $Q_{12}$ and $M_2$, i.e., for every solution branch, $(Q_{11}, Q_{12}, M_1, M_2)$, there exists another solution branch with $(Q_{11}, -Q_{12}, M_1, -M_2)$. The stable solution branches correspond to a smooth rotation in $\mathbf{n}$, between the two end points $y = \pm 1$ and are actually the global energy minimisers for $k \leq 1.25$.

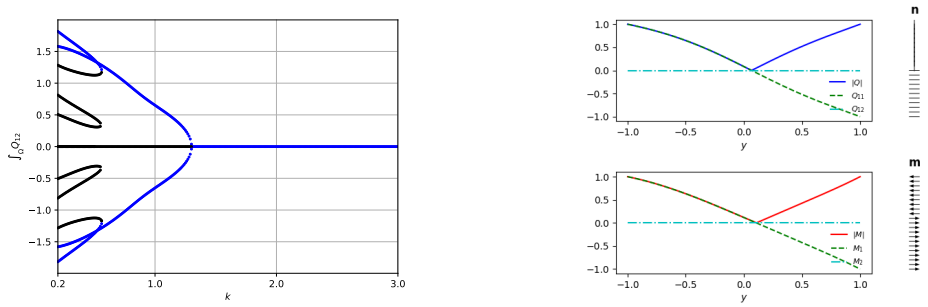
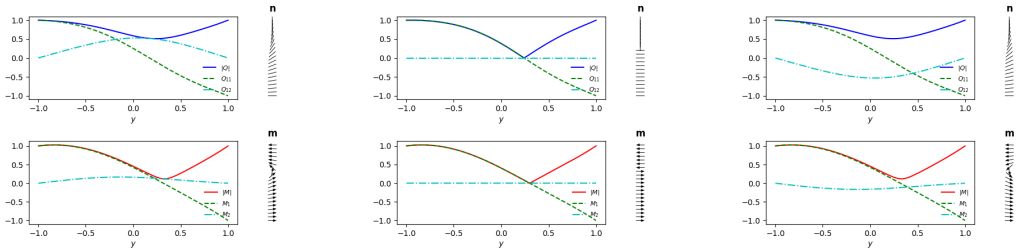


**Figure 6.8:** Left: the bifurcation diagram of continuing $k_1 = k_2 = k \in [0.2, 3.0]$ with fixed $c = \xi = 1$; here, black represents unstable solutions while blue indicates stable solutions. Right: the stable solution for $k = 2$.

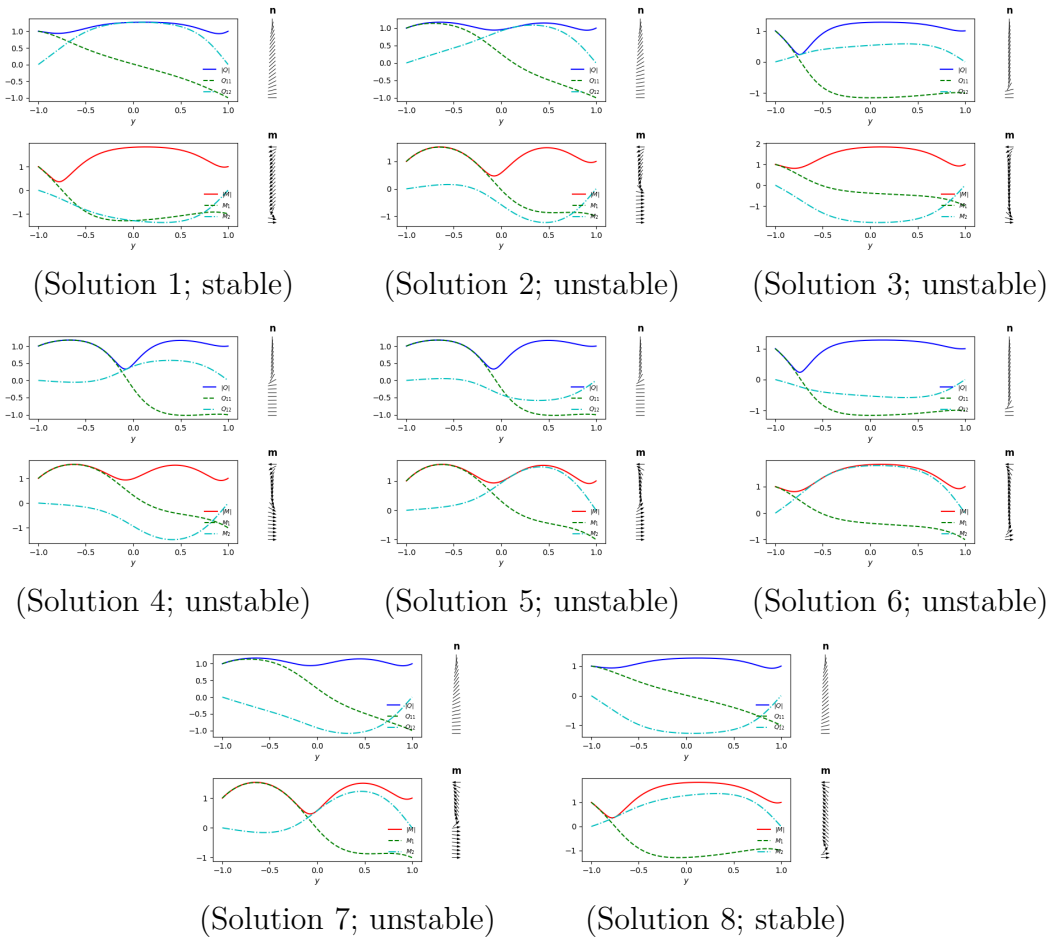


(Solution 1; stable) (Solution 2; unstable) (Solution 3; stable)

**Figure 6.9:** Three solutions for $k = 1$ in Figure 6.8. Solutions 1 and 3 are global energy minimisers.

As $k$ becomes smaller, more (stable or unstable) solutions are found. More specifically, there are four disconnected bifurcations appearing around $k = 0.55$, giving two further stable solutions, which are also local energy minimisers (see Solutions 1 and 8 in Figure 6.10 for an illustration) for $k \in [0.2, 0.55]$. Again, they only differ by the sign of $Q_{12}$ and $M_2$. In Figure 6.10, we plot eight newly found solution profiles, along with their stabilities. The stable solutions typically correspond to a smooth **n**-profiles with minimal rotation (minimal topological degree consistent with the boundary conditions), while the stable normalised magnetisation profiles **m** are also smooth, except for a thin interval of large rotation in **m** localised near the end points $y = \pm 1$. Meanwhile, it can be seen that the unstable solution pairs, i.e., Solutions 2 & 7, Solutions 3 & 6 and Solutions 4 & 5 also differ by the sign of $Q_{12}$ and $M_2$. Interestingly, all profiles with interior jumps in **n** and **m** are unstable.

We next investigate the loss of stability of the OR solution branch for a larger value of $c$, i.e., we numerically compute a bifurcation diagram in Figure 6.11, for



(Solution 1; stable) (Solution 2; unstable) (Solution 3; unstable)

(Solution 4; unstable) (Solution 5; unstable) (Solution 6; unstable)

(Solution 7; unstable) (Solution 8; stable)

**Figure 6.10:** Eight new solutions for $k = 0.2$ in Figure 6.8. Solutions 1 and 8 are global energy minimisers.

the solutions of (5.2.0.6a)-(5.2.0.6d), by continuing $k \in [3, 5]$ with a step size of 0.015, and fixed $c = 5$. One stable OR solution is shown in Figure 6.11 and it loses stability at the pitchfork bifurcation point $k \approx 4.46$, leading to two new stable branches (see illustrations in Figure 6.12 for $k = 4.43$). We observe that they only differ in the signs of $Q_{12}$ and $M_2$ and in fact are energy minimisers for $k \leq 4.43$. Thus, the qualitative features of the bifurcation diagram are unchanged by increasing $c$ but the OR solution branch loses stability for $k < k^*(c)$, where $k^*(c)$ is an increasing function of $c$. Hence, as $c$ increases, OR solutions are increasingly difficult to find owing to their shrinking window of stability.

**Remark 6.1.** *One may wonder about the appearance of the two folds in the bifurcation diagram depicted in Figure 6.11. They do not represent the same*

*solution at the intersection points. Instead, they are just overlapping points in this plot of* $\int_\Omega Q_{12}$ *against* $k$*. Choosing a different functional may yield a bifurcation diagram without these intersection points.*

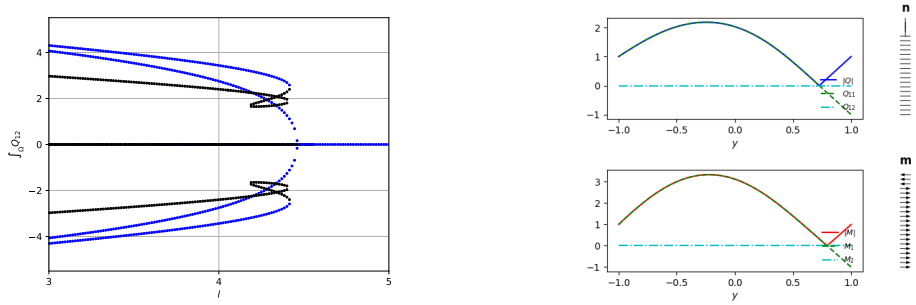


**Figure 6.11:** Left: the bifurcation diagram with fixed $c = 5$ and $\xi = 1$; here, black labels unstable solutions while blue labels stable solutions. Right: one stable OR solution for $k = 4.45$.

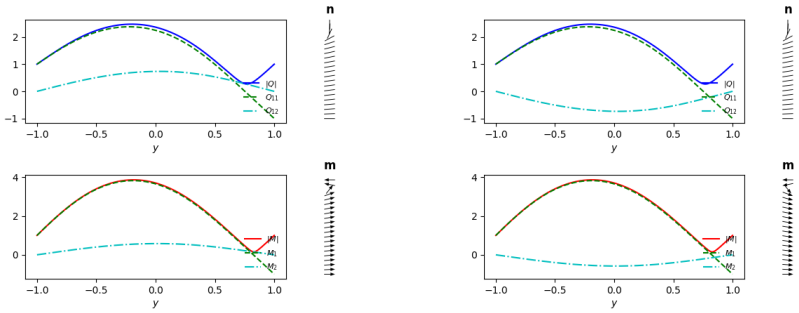


**Figure 6.12:** Two new stable solutions at $k = 4.43$ in Figure 6.11.

## 6.6 Summary

In this chapter, we performed several numerical experiments that validate the theoretical analysis derived in [Dal+21]. These include providing more complete solution landscapes of the ferronematic problem, stability analysis, and showing multiple patterns of domain walls in the interior. We demonstrated the strength of **Q**-tensor theory for characterising defects (i.e., domain walls in director **n** and normalised magnetisation **m**) in one-dimensional ferronematics. We will further consider more complicated defect structures (in higher dimensions) in the next part of this thesis.

# Part III

# Smectic Liquid Crystals

---

**This work expands upon *Xia, MacLachlan, Atherton, and Farrell (2021)* [Xia+21].**

---

# 7

# A mathematical model of smectics

## Contents



In the proceeding part, we have considered the application of the **Q**-tensor theory in ferronematics, which can possess multiple domain walls (i.e., where the nematic tensor **Q** or magnetic order parameter **M** vanishes) separating polydomains. In this last bulk of this thesis, we study and model more complicated defect structures that exist in smectics, more precisely, in the smectic-A phase.

Smectic liquid crystals are layered mesophases that have a periodic modulation of the mass density along one spatial direction. Roughly speaking, they can be thought of as one-dimensional solids along the direction of periodicity and two-dimensional fluids along the other two remaining directions. Due to their periodic structures, smectic liquid crystals have drawn extensive research attention and are directly related to some applications in photonic band-gap materials, metamaterials, and templates for guided particle self-assembly [ZL08].

Two common phases of smectic liquid crystals are the smectic-A and smectic-C phases (see Figure 7.1 for an illustration). In smectic-A phases, the director is parallel to the normal of the smectic layers while smectic-C phases allow the director to freely rotate around the normal, and thus present a tilted angle between the director and the layer normals. In order to characterise the periodic property of the density in smectic phases, de Gennes first proposed to use a complex-valued variable as the smectic order parameter, based on an analogue to superfluids in superconductors [Gen72]. This theory (abbreviated as the $dG$ theory) for modelling smectics has been a popular tool for investigating defect structures in smectic phases, e.g., [SK07; OU06] and for modelling smectic liquid crystal fluids [E97].

In this chapter, we first review the classical dG model for smectic-A liquid crystals and then a more recent model by Pevnyi, Selinger and Sluckin [PSS14] using a real-valued smectic order parameter. Next, we propose a new model inheriting the advantages of the real-valued smectic model, which can also represent half charge defects by adopting a $\mathbf{Q}$-tensor as the nematic order parameter.

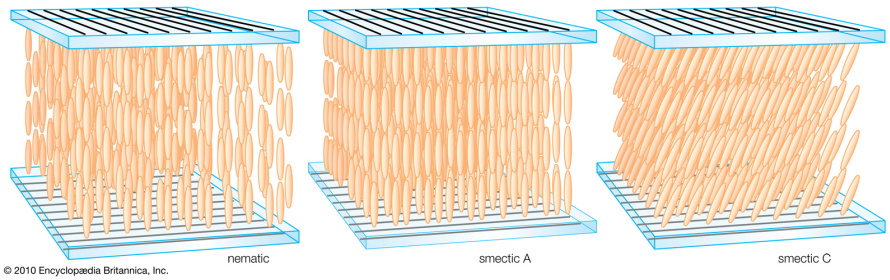



**Figure 7.1:** Graphical illustrations of nematic, smectic-A and smectic-C phases. The top and bottom substrate plates are polarisers with perpendicular alignment directions. This type of polarisers is for example used in twisted nematic display [DS11, Technical Box 10.1]. Picture is taken from [Wal20].

## 7.1 The de Gennes model

According to de Gennes' theory [Gen72; Gen74], one can model smectic liquid crystals based on a complex-valued order parameter $\psi : \Omega \to \mathbb{C}$, which describes the magnitude $|\psi|$ and the phase $\nabla\psi$ of smectic layer ordering, and a real vector-valued nematic order parameter $\mathbf{n}$ satisfying the unit-length constraint $|\mathbf{n}|$=1. Furthermore,

the phase $\nabla\psi$ indicates the position of the layers. There is a strong analogy between the derivation of de Gennes's formulation for smectics and that of superconductors, as discussed in [Gen72; HL74].

More precisely, de Gennes proposed the free energy of smectic-A LC to be

$$\mathcal{J}^{dG}(\mathbf{n},\psi)=\int_{\Omega}\left(F_S(\mathbf{n},\psi)+W^{OF}(\mathbf{n})\right), \tag{7.1.0.1}$$

where $\Omega\subset\mathbb{R}^d$ ($d\in\{2,3\}$) is the region occupied by liquid crystals, $W^{OF}$ denotes the nematic Oseen–Frank energy density of the form (2.1.0.3), and $F_S$ represents the smectic energy density given by

$$F_S(\mathbf{n},\psi)=|\nabla\psi-\mathrm{i}q\mathbf{n}\psi|^2+\varsigma|\psi|^2+\frac{\varpi}{2}|\psi|^4. \tag{7.1.0.2}$$

Here, $\mathrm{i}=\sqrt{-1}$, $q$ represents the length of the favoured wave-vector, $\varpi>0$ a fixed number and $\varsigma=\varsigma_0(T_m-T_{ns})$ the discrepancy between the material temperature $T_m$ and nematic-smectic transition temperature $T_{ns}$, with $\varsigma_0>0$. Since we are focusing on the smectic phase where $T_m$ is normally below the transition temperature $T_{ns}$, it holds that $\varsigma<0$.

It is obvious to see that when $\psi=0$, (7.1.0.1) reduces to the nematic phase; and $\psi\neq0$ corresponds to the smectic phase. Furthermore, one can note there is no odd power of the amplitude $|\psi|$ in the smectic energy density (7.1.0.2). This is because a change in sign, $\psi\to-\psi$, corresponds to a uniform translation of the smectic layers by one smectic layer and it should cost no additional energy to do so [LS91].

**Remark 7.1.** *We have some comments regarding the derivation of dG model (7.1.0.1) for smectic liquid crystals: (a) the coefficients in (7.1.0.1) are phenomenological and their relations to molecular properties are not revealed [LS91]; (b) the smectic order parameter $\psi$ is assumed to vary spatially on a length scale larger than the layer thickness $\tau$; (c) the free energy (7.1.0.1) only includes independent fluctuations (i.e., the $W^{OF}$ density term and the $\varsigma,\varpi$ term in $F_S$ are dependent only on $\mathbf{n}$ and $\psi$ respectively) in the quantities $\psi$ and $\mathbf{n}$; (d) no orientational order parameter (e.g., tensor order parameter $\mathbf{Q}$) has been involved. Linhananta and Sullivan [LS91] have presented a modified dG energy to overcome the above limitations by means of molecular density functional theories.*

It is important to understand what the coupling term $|\nabla\psi - \mathrm{i}q\mathbf{n}\psi|^2$ describes. To this end, we can express the smectic order parameter by

$$\psi(\mathbf{x}) = \varrho(\mathbf{x})e^{\mathrm{i}\iota(\mathbf{x})}, \quad \varrho : \Omega \to \mathbb{R}, \quad \iota : \Omega \to \mathbb{R}, \tag{7.1.0.3}$$

where $\varrho(\mathbf{x}) = |\psi(\mathbf{x})|$ denotes the mass density of the smectic layers at a point $\mathbf{x} \in \Omega$ and $\iota$ parametrises the layers so that $\nabla\iota$ indicates the direction of the layer normal. Substituting the above expression into the coupling term, we obtain

$$|\nabla\psi - \mathrm{i}q\mathbf{n}\psi|^2 = |\nabla\varrho|^2 + \varrho^2|\nabla\iota - q\mathbf{n}|^2,$$

and the smectic energy density $F_S$ becomes

$$F_S(\mathbf{n}, \varrho, \iota) = |\nabla\varrho|^2 + \varrho^2|\nabla\iota - q\mathbf{n}|^2 + \varsigma|\varrho|^2 + \frac{\varpi}{2}|\varrho|^4. \tag{7.1.0.4}$$

Consequently, as we perform minimisation over $F_S$, we are actually penalising the nematic-smectic coupling constraint $\nabla\iota = q\mathbf{n}$. This illustrates how smectic layers align with nematic directors $\mathbf{n}$, that is to say, the smectic layer normals should be parallel to the director.

**Remark 7.2.** *If* $\mathbf{n}$ *is a gradient (i.e.,* $q\mathbf{n} = \nabla\iota$ *which can be derived from penalising the coupling term* $\varrho^2|\nabla\iota - q\mathbf{n}|^2$ *in (7.1.0.4)), then the twist-effect* $\mathbf{n} \cdot (\nabla \times \mathbf{n})$ *in* $W^{OF}(\mathbf{n})$ *is zero. This is known as the incompatibility between smectic order and twist (see e.g., [CP00, Section 1.6] and [SK07]).*

Moreover, the molecular mass density is defined as

$$\varrho_m(\mathbf{x}) = \varrho_0 + \frac{1}{2}\left(\psi(\mathbf{x}) + \psi^*(\mathbf{x})\right) = \varrho_0 + \varrho(\mathbf{x})\cos\iota(\mathbf{x}) = \varrho_0 + |\psi(\mathbf{x})|\cos\iota(\mathbf{x}), \tag{7.1.0.5}$$

where $\varrho_0$ is the average density and $\psi^*$ represents the complex conjugate of $\psi$. Hence, $|\psi(\mathbf{x})|\cos\iota(\mathbf{x})$ gives the real-valued density variation between the molecular mass density and the average density. The derivation of the model by Pevnyi, Selinger and Sluckin [PSS14] to be introduced in the next section in fact utilises such a real variable as the smectic order parameter, as we shall now see.

## 7.2 The Pevnyi–Selinger–Sluckin model

As discussed in [Bed14; PSS14], the classical dG model (7.1.0.1) using the complex order parameter $\psi$ gives rise to a direct difficulty: $\mathrm{Im}(\psi)$ does not relate to anything physical. A resulting branch-cut due to the presence of this issue is schematically illustrated in Figure 7.2 with a $+1/2$-charge disclination. This situation is similar to the case of representing the $+1/2$-charge defect by the vector-valued director $\mathbf{n}$, where the head-to-tail symmetry of molecules is not respected and thus a branch-cut occurs when $\mathbf{n}$ changes to $-\mathbf{n}$. To avoid the use of a complex variable, Pevnyi, Selinger and Sluckin [PSS14] proposed a new model (abbreviated henceforth as the *PSS* model) adopting the director $\mathbf{n} : \Omega \to \mathbb{R}^d$ and the density variation $u : \Omega \to \mathbb{R}$ from the average density as state variables.

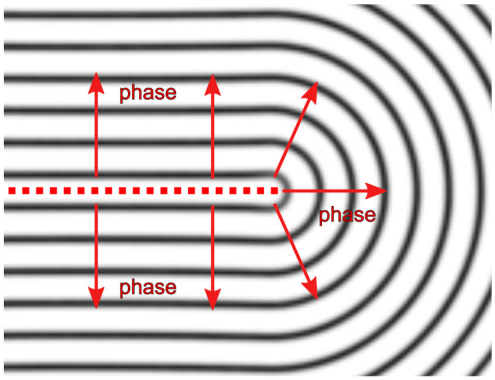



**Figure 7.2:** Illustration of the branch-cut (red dotted line) resulted from the non-physical imaginary part of $\psi$ in a $+1/2$-charge disclination. Credit: [PSS14, Fig. 1]

The form of the PSS free energy is given by

$$\mathcal{J}^{PSS}(u,\mathbf{n}) = \int_\Omega \left( f_s(u) + \frac{K}{2}|\nabla \mathbf{n}|^2 + B\left|\mathcal{D}^2 u + q^2\,(\mathbf{n}\otimes\mathbf{n})\,u\right|^2 \right), \tag{7.2.0.1}$$

where the smectic bulk energy density is given by

$$f_s(u) = \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4.$$

Here $a_1, a_2, a_3, B, K$ and $q$ are some known real parameters. Moreover, the unit length constraint $\mathbf{n}\cdot\mathbf{n} = 1$ for the director must be enforced. In order to keep $f_s$ bounded from below, we need to choose $a_3 > 0$, and to possess nonzero (i.e., $u \neq 0$) minimisers of $f_s$ (thus not pure nematic minimisers), we should choose $a_1 < 0$.

**Remark 7.3.** *One can notice that a cubic term of $u$ is added to $f_s$ in (7.2.0.1) when comparing it with the dG model (7.1.0.1). This is allowed because we should not expect symmetry between positive density variation $u > 0$ and negative density variation $u < 0$.*

The derivation of the PSS model comes from the density functional theory (based on a molecular statistical description) analogous to early work in [LS91; PS91]; however, a detailed explanation (in particular, about how the model parameters are related to some physically measurable constants) is not given in [PSS14]. In fact, the idea of [LS91] is to divide the total free energy into local and nonlocal parts. The local energy includes an isotropic term, modelled by the standard quartic order Landau–Ginzburg free energy with regard to the smectic density variable, and an anisotropic term of **Q**-tensor to characterise nematic LC. For the nonlocal part, they adopt the typical form of two-body contributions to the free energy occurring in mean-field density functional theories, which gives rise to a fourth order term similar to the coupling $B$-term in (7.2.0.1).

For a better understanding of the PSS model, particularly how the coupling term in (7.2.0.1) relates to the physical constraint of smectics, we give our interpretation in the following. As described in [PSS14] and illustrated in (7.1.0.5), the density variation $u$ can be related to the complex order parameter $\psi$ in the dG model by the expression

$$u = \Re\psi = |\psi| \cos(\iota)$$

with $\Re$ denotes the real part of a complex number. Note that the amplitude $|\psi|$ of the density modulation does not vary spatially as it refers to the largest mass density and we can actually see this fact in the numerical results in Chapter 9. From what we have discussed in Section 7.1, minimising $|\nabla\psi - \mathrm{i}q\mathbf{n}\psi|^2$ in fact promotes the relation $\nabla\iota = q\mathbf{n}$. Subsequently combining with (7.1.0.3), one can expect the following expression of $\psi$,

$$\psi(\mathbf{x}) = |\psi| e^{\mathrm{i}q\mathbf{n}\cdot\mathbf{x}}.$$

Therefore, we obtain the corresponding form of the density variation $u$ as follows:

$$u(\mathbf{x}) = |\psi| \cos(q\mathbf{n} \cdot \mathbf{x}). \tag{7.2.0.2}$$

We then calculate

$$\mathcal{D}^2 u = \mathcal{D}(\mathcal{D}u) = \mathcal{D}(-|\psi| \sin(q\mathbf{n} \cdot \mathbf{x}) q\mathbf{n}) = -q^2(\mathbf{n} \otimes \mathbf{n})u,$$

and it follows that

$$\mathcal{D}^2 u + q^2(\mathbf{n} \otimes \mathbf{n})u = 0.$$

Hence, one can interpret minimising the coupling term $\left|\mathcal{D}^2 u + q^2 (\mathbf{n} \otimes \mathbf{n}) u\right|^2$ as respecting the periodicity of the smectic density, i.e., $u = |\psi| \cos(q\mathbf{n} \cdot \mathbf{x})$.

The PSS model helps investigate defect structures appearing in the smectic-A phase in a more physically reasonable way without using the complex order parameter $\psi$ in the classical dG model (7.1.0.1). There are some numerical examples of smectic layers respecting different topological defects illustrated in the work [PSS14]. However, by solving the PSS model as described using $\mathbf{n} \in H^1(\Omega, \mathcal{S}^{d-1})$, we cannot reproduce the experiments of half-charge defects that are shown in [PSS14]. This is due to the presence of a discontinuity in the director $\mathbf{n}$ in these defects, which cannot be characterised by a continuous vector field [Bal17]. As a matter of fact, in private communication, the authors of [PSS14] have commented that they actually implemented their model with the tensor product $\mathbf{n} \otimes \mathbf{n}$, thus enforcing the unit length constraint of director $\mathbf{n}$ implicitly through introducing the tensor $\mathbf{n} \otimes \mathbf{n}$. This allows them to represent half-charge defects [Bal17], but numerically enforcing that the order parameter is a line field of the form $\mathbf{n} \otimes \mathbf{n}$ in minimisation is difficult [BNW20].

## 7.3 Our proposed model

A new mathematical model that incorporates both a tensor field and a real-valued density variation field could be useful in representing smectic liquid crystals with complex defect structures. In fact, the idea of combining a $\mathbf{Q}$-tensor variable and a

real-valued density variable to model smectic LC has been previously discussed in [LS91; MZ15; Han+15]. However, these works are all molecular-based microscopic models which are difficult to implement due to their natural complexity in relating statistical parameters to physically realistic experimental results. It is an open problem to combine both the microscopic and macroscopic sides for modelling smectic liquid crystals, as discussed by Ball & Bedford [BB15]. Moreover, these authors [BB15; Bed14] have noticed the necessity of combining the nematic order parameter $\mathbf{Q}$ and the real-valued smectic order parameter to characterise defects and thus modified the PSS model by replacing $\mathbf{n} \otimes \mathbf{n}$ by $(\mathbf{Q}/s + \mathbf{I}_3/3)$ arising from the uniaxial expression of $\mathbf{Q}$-tensor:

$$\mathcal{J}^{BB}(u, \mathbf{Q}) = \int_\Omega \frac{K}{2}|\nabla \mathbf{Q}|^2 + B\left|\mathcal{D}^2 u + q^2\left(\frac{\mathbf{Q}}{s} + \frac{\mathbf{I}_3}{3}\right)u\right|^2 + \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4, \tag{7.3.0.1}$$

with $u \in H^2(\Omega, \mathbb{R})$ and $\mathbf{Q} \in SBV(\Omega, S_0)$ where $SBV$ denotes special functions of bounded variation. A preliminary result of existence of minimisers for their modified model is also briefly included. Nevertheless, the possibility of characterising defects existing in smectic liquid crystals and the implementation of their model has not been investigated or realised. One can readily notice the numerical singularities caused by the denominator $s$ whenever it is near zero (which is likely to happen around defects). To avoid the aforementioned issue of the denominator $s$, we assume that the scalar order parameter $s$ is a fixed constant, which can be determined by the form of the additional nematic bulk energy (we will discuss this point in detail later) arising from the phenomenological LdG model of nematics.

### 7.3.1 A unified framework

In this part, we further assume that $\Omega$ is convex as such convexity is needed for the regularity result (see Theorem 8.1).

Considering that smectic-A liquid crystals are optically uniaxial [Gen73; Gen74], we can express the $\mathbf{Q}$ tensor in a uniaxial form: $\mathbf{Q} = s\left(\mathbf{n} \otimes \mathbf{n} - \frac{\mathbf{I}_d}{d}\right)$, where the

director $\mathbf{n}$ is the corresponding eigenvector of $\mathbf{Q}$ with the major eigenvalue, say, $\lambda_{eig}$. One can readily check that $s$ and $\lambda_{eig}$ satisfy the relation

$$\begin{aligned} s &= 2\lambda_{eig} \quad \text{for } d = 2, \\ s &= \frac{3}{2}\lambda_{eig} \quad \text{for } d = 3. \end{aligned}$$

Moreover, the symmetric traceless $\mathbf{Q}$-tensor has two degrees of freedom $(Q_{11}, Q_{12})$ in two dimensions or five degrees of freedom $(Q_{11}, Q_{12}, Q_{13}, Q_{22}, Q_{23})$ in three dimensions. Thus, it can be expressed in the form of

$$\mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & -Q_{11} \end{bmatrix} \quad \text{or} \quad \mathbf{Q} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12} & Q_{22} & Q_{23} \\ Q_{13} & Q_{23} & -(Q_{11}+Q_{22}) \end{bmatrix}. \tag{7.3.1.1}$$

In particular, we note that $\mathrm{tr}(\mathbf{Q}^3) = 0$ for $d = 2$ which can be easily checked via computations using (7.3.1.1).

We now propose the following $\mathbf{Q}$-tensor model that incorporates the dG theory for smectic LC and LdG model for nematics while also keeping the density variable $u$ to be real-valued, as discussed in [PSS14]:

$$\mathcal{J}(u, \mathbf{Q}) = \int_\Omega \left( f_s(u) + B \left| \mathcal{D}^2 u + q^2 \left( \mathbf{Q} + \frac{\mathbf{I}_d}{d} \right) u \right|^2 + f_n(\mathbf{Q}, \nabla \mathbf{Q}) \right), \tag{7.3.1.2}$$

where

$$f_s(u) := \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 \tag{7.3.1.3}$$

and

$$\begin{aligned} f_n(\mathbf{Q}, \nabla \mathbf{Q}) &= f_n^e(\nabla \mathbf{Q}) + f_n^b(\mathbf{Q}) \\ &:= \frac{K}{2}|\nabla \mathbf{Q}|^2 + \begin{cases} \left( -l\left(\mathrm{tr}(\mathbf{Q}^2)\right) + l\left(\mathrm{tr}(\mathbf{Q}^2)\right)^2 \right), & \text{if } d = 2, \\ \left( -\frac{l}{2}\left(\mathrm{tr}(\mathbf{Q}^2)\right) - \frac{l}{3}\left(\mathrm{tr}(\mathbf{Q}^3)\right) + \frac{l}{2}\left(\mathrm{tr}(\mathbf{Q}^2)\right)^2 \right), & \text{if } d = 3. \end{cases} \end{aligned} \tag{7.3.1.4}$$

Here, $K$ is the nematic elastic constant, $l$ represents the nematic bulk parameter that can depend on temperature and $a_1, a_2, a_3, B, q$ are inherited from the PSS model. We refer to the *decoupled* case when $q = 0$.

**Remark 7.4.** *We can observe some differences between our proposed model (7.3.1.2) and the Ball–Bedford model (7.3.0.1): (a) we have taken the scalar order parameter*

*$s$ to be a fixed constant (in fact, $s = 1$), which is weakly preferred due to the addition of the nematic bulk term $f_n^b(\mathbf{Q})$; (b) we give a unified framework applicable to both two and three dimensions.*

One may notice that the term $f_n$ arises in the classical LdG model (5.1.0.4) for nematic LC. Furthermore, it is known that the global minimiser of the bulk energy $f_n^b$ is a uniaxial $\mathbf{Q}$ tensor with scalar order parameter $s = 1$ (one can check this by some calculations or using [MZ10, Proposition 15] as quoted below for self containment). Adding the bulk energy terms helps in deciding the scalar order parameter $s$, and therefore we can adjust the coefficients in the bulk energy density so to promote $s = 1$ almost everywhere.

**Proposition 7.1.** *[MZ10, Proposition 15] Assume that $l_a, l_b, l_c$ are positive parameters and consider the bulk energy in the following form*

$$f_n^b(\mathbf{Q}) = -\frac{l_a}{2}\left(tr(\mathbf{Q}^2)\right) - \frac{l_b}{3}\left(tr(\mathbf{Q}^3)\right) + \frac{l_c}{4}\left(tr(\mathbf{Q}^2)\right)^2. \tag{7.3.1.5}$$

*Then its minimiser is a uniaxial tensor of the form*

$$\mathbf{Q} = s_+\left(\mathbf{n}\otimes\mathbf{n} - \frac{\mathbf{I}_3}{3}\right),$$

*where*

$$s_+ = \frac{l_b + \sqrt{l_b^2 + 24 l_a l_c}}{4 l_c}.$$

### 7.3.2 Existence of minimisers

We have proposed a unified functional (7.3.1.2) for both two- and three-dimensional cases to be minimised on some admissible set. An immediate question is whether minimisers exist.

We define the admissible space $\mathcal{A}^s$ of our proposed functional $\mathcal{J}$ as

$$\mathcal{A}^s = \Big\{ u \in H^2(\Omega,\mathbb{R}),\ \mathbf{Q} \in H^1(\Omega, S_0) : \\ \mathbf{Q} = s\left(\mathbf{n}\otimes\mathbf{n} - \frac{\mathbf{I}_d}{d}\right) \text{ for some } s \in [0,1], \mathbf{Q} = \mathbf{Q}_b \text{ on } \partial\Omega \Big\}, \tag{7.3.2.1}$$

with $\mathbf{n} \in H^1(\Omega, \mathcal{S}^{d-1})$ and the Dirichlet boundary data $\mathbf{Q}_b \in H^{1/2}(\partial\Omega, S_0)$. For simplicity, we only consider Dirichlet boundary conditions for $\mathbf{Q}$ in this section, but other types of boundary conditions (e.g., a mixture of the Dirichlet and natural boundary conditions as used in Chapter 9) can be taken.

Notice that $f_n(\mathbf{Q}, \nabla\mathbf{Q})$ is the classical LdG model for nematic LC. It is a known result from Davis & Gartland [DG98, Corollary 4.4] that there exists a minimiser of the functional $\int_\Omega f_n$ on $\mathbf{Q} \in H^1(\Omega, S_0)$ in three dimensions. Furthermore, Bedford [Bed14, Theorem 5.18] has given an existence result of the Ball–Bedford model (7.3.0.1) for $\mathbf{Q} \in SBV(\Omega, S_0)$ and $u \in H^2(\Omega, \mathbb{R})$, also in three dimensions. Motivated by these two results, we can give the existence result of minimising our proposed free energy (7.3.1.2) via the direct method of calculus of variations (see e.g., [Gia83, Section 3, Chapter 1]) in the admissible space $\mathcal{A}^s$.

**Theorem 7.2.** *(Existence of minimisers) Let $\mathcal{J}$ be of the form (7.3.1.2) with positive parameters $a_3$, $B$, $q$, $K$, $l$. Then there exists a solution pair $(u^*, \mathbf{Q}^*)$ that minimises $\mathcal{J}$ over the admissible set $\mathcal{A}^s$.*

*Proof.* Note that both the smectic density $f_s$ and the nematic bulk density $f_n^b$ are bounded from below as $a_3, l > 0$. Thus, $\mathcal{J}$ is also bounded from below and we can choose a minimising sequence $\{(u_j, \mathbf{Q}_j)\}$, i.e.,

$$\begin{gathered}(u_j, \mathbf{Q}_j) \in \mathcal{A}^s, \quad \mathbf{Q}_j - \tilde{\mathbf{Q}} \in H_0^1(\Omega, S_0), \\ \mathcal{J}(u_j, \mathbf{Q}_j) \overset{j\to\infty}{\longrightarrow} \inf\{\mathcal{J}(u, \mathbf{Q}) : (u, \mathbf{Q}) \in \mathcal{A}^s,\ \mathbf{Q} - \tilde{\mathbf{Q}} \in H_0^1(\Omega, S_0)\} < \infty.\end{gathered} \tag{7.3.2.2}$$

Here, we define $\tilde{\mathbf{Q}} \in H^1(\Omega, S_0)$ to be the extended function with trace $\mathbf{Q}_b$. We tackle the three terms in (7.3.1.2) separately in the following.

First, for the nematic energy term $\int_\Omega f_n(\mathbf{Q}, \nabla\mathbf{Q})$, we can follow the proof of [DG98, Theorem 4.3] to obtain that $f_n(\mathbf{Q}_j, \nabla\mathbf{Q}_j)$ is coercive in $H^1(\Omega, S_0)$ in the sense that $f_n$ grows unbounded as $\|\mathbf{Q}\|_1 \to \infty$, and thus the minimising sequence $\{\mathbf{Q}_j\}$ must be bounded. Since $H^1(\Omega)$ is a reflexive Banach Space, we have a subsequence (also denoted as $\{\mathbf{Q}_j\}$) that weakly converges to $\mathbf{Q}^* \in H^1(\Omega, S_0)$ such that $\mathbf{Q}^* - \tilde{\mathbf{Q}} \in H_0^1(\Omega)$, and from the Rellich–Kondrachov theorem it follows that

$$\begin{aligned}\mathbf{Q}_j &\to \mathbf{Q}^* && \text{in } L^2(\Omega), \\ \nabla\mathbf{Q}_j &\rightharpoonup \nabla\mathbf{Q}^* && \text{in } L^2(\Omega).\end{aligned}$$

The weakly lower semi-continuity of the nematic energy density $f_n$ in (7.3.1.4) is guaranteed by [DG98, Lemma 4.2], therefore,

$$\liminf_{j\to\infty}\int_\Omega f_n(\mathbf{Q}_j,\nabla\mathbf{Q}_j)\geq\int_\Omega f_n(\mathbf{Q}^*,\nabla\mathbf{Q}^*). \tag{7.3.2.3}$$

Next, for the smectic bulk term $\int_\Omega f_s(u)$, we can follow the proof in [Bed14, Theorem 5.19] with further details. By (7.3.2.2), we have

$$\sup_j\int_\Omega\left(\left|\mathcal{D}^2u_j\right|^2+|u_j|^2\right)<\infty,$$

which implies an upper bound for $\nabla u_j$ using [Bed14, Equation (5.42)]:

$$\int_\Omega|\nabla v|^2\leq C\left(\int_\Omega\left|\mathcal{D}^2v\right|^2+v^2\right)\quad\forall v\in H^2(\Omega,\mathbb{R}).$$

Hence, $\{u_j\}$ is bounded in $H^2(\Omega)$ and thus there is a subsequence (also denoted as $\{u_j\}$) such that

$$u_j\rightharpoonup u^*\quad\text{in } H^2(\Omega).$$

Moreover, one can readily check that $\|u^*\|_\infty<\infty$ by the Sobolev embedding of $H^2(\Omega)$ into the Hölder spaces $\mathcal{C}^{\mathfrak{t},\varkappa_0}(\Omega)$ ($\mathfrak{t}+\varkappa_0=1$ for $d=2$ and $\mathfrak{t}+\varkappa_0=1/2$ for $d=3$) and the boundedness of domain $\Omega$. Again, by the Rellich–Kondrachov theorem, we have

$$\begin{aligned}u_j&\to u^*&&\text{in } L^2(\Omega),\\ \mathcal{D}^2u_j&\rightharpoonup\mathcal{D}^2u^*&&\text{in } L^2(\Omega).\end{aligned}$$

Noting that $f_s$ is bounded from below for all $u\in H^2(\Omega)$, then there holds that

$$\liminf_{j\to\infty}\int_\Omega f_s(u_j)\geq\int_\Omega f_s(u^*). \tag{7.3.2.4}$$

Now, we consider the nematic-smectic coupling term in (7.3.1.2). Since the admissible space $\mathcal{A}^s$ admits uniaxial tensors, we calculate

$$\begin{aligned}|\mathbf{Q}_j|^2&=\left|s_j\left(\mathbf{n}_j\otimes\mathbf{n}_j-\frac{\mathbf{I}_d}{d}\right)\right|^2\\&=|s_j|^2\left(|\mathbf{n}_j\otimes\mathbf{n}_j|^2+\left|\frac{\mathbf{I}_d}{d}\right|^2-\frac{2}{d}\mathbf{n}_j\otimes\mathbf{n}_j\colon\mathbf{I}_d\right)\\&=|s_j|^2\left(1+\frac{1}{d}-\frac{2}{d}\right)\\&=|s_j|^2\left(1-\frac{1}{d}\right)\\&<|s_j|^2,\end{aligned}$$

implying that $|\mathbf{Q}_j|^2$ is always bounded in $\Omega$. By this boundedness and the fact that $\|u^*\|_\infty < \infty$, we can deduce

$$\begin{aligned}\int_\Omega |u_j\mathbf{Q}_j - u^*\mathbf{Q}^*|^2 &= \int_\Omega |(u_j - u^*)\mathbf{Q}_j + u^*(\mathbf{Q}_j - \mathbf{Q}^*)|^2 \\ &\le 2\int_\Omega \left(|u_j - u^*|^2|\mathbf{Q}_j|^2 + |u^*|^2|\mathbf{Q}_j - \mathbf{Q}^*|^2\right) \\ &\to 0 \quad \text{as } u_j \to u^*, \mathbf{Q}_j \to \mathbf{Q}^* \text{ in } L^2(\Omega).\end{aligned}$$

Hence, $u_j\mathbf{Q}_j \to u^*\mathbf{Q}^*$ in $L^2(\Omega)$, and further,

$$\begin{aligned} u_j\left(\mathbf{Q}_j + \frac{\mathbf{I}_d}{d}\right) &\to u^*\left(\mathbf{Q}^* + \frac{\mathbf{I}_d}{d}\right) && \text{in } L^2(\Omega), \\ u_j\left(\mathbf{Q}_j + \frac{\mathbf{I}_d}{d}\right) : \mathcal{D}^2 u_j &\rightharpoonup u^*\left(\mathbf{Q}^* + \frac{\mathbf{I}_d}{d}\right) : \mathcal{D}^2 u^* && \text{in } L^1(\Omega). \end{aligned}$$

Therefore, we have

$$\begin{aligned}&\liminf_{j\to\infty} \int_\Omega \left|\mathcal{D}^2 u_j + q^2\left(\mathbf{Q}_j + \frac{\mathbf{I}_d}{d}\right)u_j\right|^2 \\ &= \liminf_{j\to\infty} \int_\Omega \left(\left|\mathcal{D}^2 u_j\right|^2 + 2q^2 u_j\left(\mathbf{Q}_j + \frac{\mathbf{I}_d}{d}\right) : \mathcal{D}^2 u_j + q^4\left|u_j\left(\mathbf{Q}_j + \frac{\mathbf{I}_d}{d}\right)\right|^2\right) \\ &\ge \int_\Omega \left(\left|\mathcal{D}^2 u^*\right|^2 + 2q^2 u^*\left(\mathbf{Q}^* + \frac{\mathbf{I}_d}{d}\right) : \mathcal{D}^2 u^* + q^4\left|u^*\left(\mathbf{Q}^* + \frac{\mathbf{I}_d}{d}\right)\right|^2\right) \\ &= \int_\Omega \left|\mathcal{D}^2 u^* + q^2\left(\mathbf{Q}^* + \frac{\mathbf{I}_d}{d}\right)u^*\right|^2. \end{aligned} \tag{7.3.2.5}$$

Finally, we only need to check that $\mathbf{Q}^*$ is uniaxial, i.e., $\mathbf{Q}^* = s^*\left(\mathbf{n}^* \otimes \mathbf{n}^* - \frac{\mathbf{I}_d}{d}\right)$ for certain $s^*$ and $\mathbf{n}^*$. This is indeed guaranteed by the $L^2$ convergence of $\mathbf{Q}_j$ and the compactness of the unit sphere $\mathbf{n}_j$ lies. Hence, we can conclude that $\mathcal{J}(u^*, \mathbf{Q}^*)$ achieves its minimum in the admissible space $\mathcal{A}^s$ by combining (7.3.2.3), (7.3.2.4) and (7.3.2.5). □

## 7.4 Summary

In this chapter, we reviewed three models for smectic-A LC: the classical dG model, a more recent model by Pevnyi, Selinger and Sluckin and the Ball–Bedford model. Through discussing their potential issues, it motivated us to propose a new model, incorporating the nematic tensor order parameter $\mathbf{Q}$ and a real smectic order parameter $u$, to characterise the complex defect structures existing in smectic liquid crystals. We then gave an existence result for the proposed model.

# 8

# Finite element discretisation

## Contents



It is implied from Theorem 7.2 that there exist minimisers of the free energy functional (7.3.1.2). One might then naturally ask how those solutions behave. We therefore consider the discretisation of the problem in this chapter. For simplicity, we only consider the decoupled case, i.e., $q = 0$ where two separate problems are to be solved: a second order PDE for the nematic tensor order parameter $\mathbf{Q}$ and a fourth order PDE for the smectic density variation $u$. With the derived a priori error estimates at hand, we then choose a suitable finite element pair for $(\mathbf{Q}, u)$, to be used in the implementations of some realistic scenarios as illustrated in Chapter 9. We verify the expected convergence behaviour via the manufactured method of solutions for both $q = 0$ and $q > 0$.

## 8.1 A priori analysis for $q = 0$

In the decoupled case, we are to solve two independent minimisation problems: one for the tensor field $\mathbf{Q}$,

$$\min_{\mathbf{Q}\in H^1_b(\Omega,S_0)} \quad \mathcal{J}_1(\mathbf{Q}) = \int_\Omega \left(f_n(\mathbf{Q}, \nabla\mathbf{Q})\right),$$

and the other for the density variation $u$:

$$\min_{u\in H^2(\Omega,\mathbb{R})} \quad \mathcal{J}_2(u) = \int_\Omega \left(B\left|\mathcal{D}^2 u\right|^2 + f_s(u)\right).$$

One can derive the following strong forms of their equilibrium equations using integration by parts (and assuming that $u \in H^4(\Omega)$),

$$(\mathcal{P}1) \quad \begin{cases} d = 2 \Rightarrow -K\Delta\mathbf{Q} + 2l\left(2|\mathbf{Q}|^2 - 1\right)\mathbf{Q} = 0 & \text{in } \Omega, \\ d = 3 \Rightarrow -K\Delta\mathbf{Q} + l\left(-\mathbf{Q} - |\mathbf{Q}|^2 + 2|\mathbf{Q}|^2\mathbf{Q}\right) = 0 & \text{in } \Omega, \\ \mathbf{Q} = \mathbf{Q}_b & \text{on } \partial\Omega, \end{cases} \tag{8.1.0.1}$$

and

$$\begin{cases} 2B\left(\mathcal{D}^2 : \mathcal{D}^2\right)u + a_1 u + a_2 u^2 + a_3 u^3 = 0 & \text{in } \Omega, \\ S^0_{bc}(u; v) = 0 \ \forall v \in H^2(\Omega) & \text{on } \partial\Omega, \end{cases} \tag{8.1.0.2}$$

with the natural boundary data given by

$$S^0_{bc}(u; v) \coloneqq \int_{\partial\Omega} \left\{\nu \cdot \left(\mathcal{D}^2 u \cdot \nabla v\right) - \left(\left(\nabla \cdot \mathcal{D}^2 u\right) \cdot \nu\right) v\right\}.$$

Note that we are not enforcing any essential boundary conditions for the real variable $u$ in (8.1.0.2). This can be insufficient to guarantee the uniqueness of solutions, thus leading to ill-posed problems. In fact, both $u$ and $-u$ are admissible solutions if $a_2 = 0$. Moreover, one expect the solution $u$ of the smectic-A model (7.3.1.2) to be a cosine function that describes the periodicity as illustrated in [PSS14, Eq. (5)] due to the alignment between smectic layer normals and directors. Therefore, the lack of essential boundary conditions may result in multiple solutions with shifted phases.

To facilitate our analysis, we assume that the fourth order problem is imposed with a Dirichlet boundary condition $u = u_b$ on $\partial\Omega$ and a natural boundary condition

regarding the second derivative of $u$. That is to say, we consider the following minimisation problem for $u$:

$$\min_{u\in H^2\cap H_b^1(\Omega,\mathbb{R})} \quad \mathcal{J}_2(u) = \int_\Omega \left( B\left|\mathcal{D}^2 u\right|^2 + f_s(u)\right),$$

which corresponds to a strong form

$$(\mathcal{P}2) \quad \begin{cases} 2B\left(\mathcal{D}^2 : \mathcal{D}^2\right) u + a_1 u + a_2 u^2 + a_3 u^3 = 0 & \text{in } \Omega, \\ u = u_b & \text{on } \partial\Omega, \\ \mathcal{D}^2 u \cdot \nu = \mathcal{D}^2 u_b \cdot \nu & \text{on } \partial\Omega, \end{cases} \tag{8.1.0.3}$$

where $H_b^1(\Omega,\mathbb{R}) \coloneqq \{v \in H^1(\Omega,\mathbb{R}) : v = u_b \text{ on } \partial\Omega\}$.

**Remark 8.1.** *The uniqueness result of the problem (8.1.0.3) is still not guaranteed, though we have imposed additional boundary conditions. This can be resulted from the presence of the nonlinear term.*

**Remark 8.2.** *In the coupled case that we implement in Chapter 9, the boundary conditions on $u$ are somewhat different. No essential boundary conditions are enforced, and some second derivative terms arise in the natural boundary condition. See (A.0.0.2) for details.*

Essentially, $(\mathcal{P}1)$ is a second order semi-linear PDE while $(\mathcal{P}2)$ yields a fourth order semi-linear PDE. To be more specific, both PDEs possess cubic nonlinearities. We now consider these two problems separately.

### 8.1.1 A priori error estimates for $(\mathcal{P}1)$

Note that problem $(\mathcal{P}1)$ is a special form of the classical Landau–de Gennes model of nematic liquid crystals. Finite element analysis for a more general form using conforming discretisations has been studied in [DG98] with homogeneous Dirichlet boundary data and in [Dav94] with inhomogeneous Dirichlet and natural boundary conditions. More specifically, Davis and Gartland [DG98] gave an abstract nonlinear finite element convergence analysis where an optimal $H^1$ error bound is proved on convex domains with piecewise linear polynomial approximations. However, the $L^2$ error bound is not derived. Quite recently, Maity, Majumdar

and Nataraj [MMN20] analysed the discontinuous Galerkin finite element methods (dGFEM) for a two-dimensional reduced Landau–de Gennes free energy, where optimal a priori error estimates in the $L^2$-norm with exact solutions being in $H^2$ and piecewise linear polynomial approximations are achieved. Their representations of the nonlinear variational form and approaches of deriving error estimates are different from those of Davis and Gartland. We follow similar techniques from [MMN20] in this subsection for concreteness.

We use the common continuous Lagrange elements for the problem $(\mathcal{P}1)$. For simplicity, we only illustrate the analysis in two dimensions for the model problem (7.3.1.2); the three dimensional case has an additional quadratic term $|\mathbf{Q}|^2$ in the strong form which can be tackled similarly. Since $(\mathcal{P}1)$ arises in the classical LdG model for nematic LC, we can quote some existing results (e.g., regularity, convergence rate in the $H^1$ norm).

**Theorem 8.1.** *[DG98, Theorem 6.3] (Regularity) Let $\Omega$ be an open, bounded, Lipschitz and convex domain. If the Dirichlet data $\mathbf{Q}_b \in H^{1/2}(\partial\Omega, S_0)$, then any solution of $(\mathcal{P}1)$ belongs to $H^2(\Omega, S_0)$.*

**Remark 8.3.** *One may wonder that the $H^2$-regularity of $\mathbf{Q}$ possibly excludes the appearance of singularities, e.g., the half-charge defects in nematics. Indeed, we do not consider the case with singularities in the analysis throughout this part of work.*

Suppose $\mathbf{Q}_h \in \mathbf{V}_h$ is the approximate solution (of the discrete problem (8.1.1.2) introduced later) by finite element methods on a finite dimensional space $\mathbf{V}_h \subset H_b^1(\Omega, S_0)$. For simplicity, we restrict ourselves in the case that $\mathbf{V}_h$ consists of piecewise linear polynomials. An a priori estimate in the $H^1$ norm has been shown in [Dav94, Theorem 2.3.3] and [DG98, Theorem 7.3] and we include it here for self-containment.

**Theorem 8.2.** *[DG98, Theorem 7.3] ($H^1$ error estimate for $\mathbf{Q}$) Let $\Omega$ be an open, bounded, polygonal and convex domain. If $\mathbf{Q} \in H^2 \cap H_b^1(\Omega, S_0)$ and $\mathbf{Q}_h \in \mathbf{V}_h$ represents an approximated solution to $\mathbf{Q}$, it holds that*

$$\|\mathbf{Q} - \mathbf{Q}_h\|_1 \lesssim h\|\mathbf{Q}\|_2. \tag{8.1.1.1}$$

**Remark 8.4.** *Theorems 8.1 and 8.2 hold for both* $\Omega \subset \mathbb{R}^2$ *and* $\Omega \subset \mathbb{R}^3$.

Following the same representation of the nonlinear variational form as in [MMN20], we introduce the continuous weak formulation of $(\mathcal{P}1)$: find $\mathbf{Q} \in H_b^1(\Omega, S_0)$ such that

$$\mathcal{N}^n(\mathbf{Q})\mathbf{P} \coloneqq A^n(\mathbf{Q},\mathbf{P}) + B^n(\mathbf{Q},\mathbf{Q},\mathbf{Q},\mathbf{P}) + C^n(\mathbf{Q},\mathbf{P}) = 0 \quad \forall \mathbf{P} \in \mathbf{H}_0^1(\Omega), \tag{8.1.1.2}$$

where the bilinear forms are

$$A^n(\mathbf{Q},\mathbf{P}) \coloneqq K \int_\Omega \nabla\mathbf{Q} \vdots \nabla\mathbf{P},$$
$$C^n(\mathbf{Q},\mathbf{P}) \coloneqq -2l \int_\Omega \mathbf{Q} : \mathbf{P},$$

and the nonlinear operator is given by

$$B^n(\Psi,\Phi,\Theta,\Xi) \coloneqq \frac{4l}{3} \int_\Omega \left((\Psi : \Phi)(\Theta : \Xi) + 2(\Psi : \Theta)(\Phi : \Xi)\right). \tag{8.1.1.3}$$

Since (8.1.1.2) is nonlinear, we need to approximate the solution of its linearised version, i.e., find $\Theta \in \mathbf{H}_0^1(\Omega)$ such that

$$\langle \mathcal{D}\mathcal{N}^n(\mathbf{Q})\Theta, \Phi\rangle \coloneqq A^n(\Theta,\Phi) + 3B^n(\mathbf{Q},\mathbf{Q},\Theta,\Phi) + C^n(\Theta,\Phi) = -\mathcal{N}^n(\mathbf{Q})\Phi \quad \forall \Phi \in \mathbf{H}_0^1(\Omega), \tag{8.1.1.4}$$

where $\langle\cdot,\cdot\rangle$ represents the dual pairing between $\mathbf{H}^{-1}(\Omega)$ and $\mathbf{H}_0^1(\Omega)$. We use continuous Lagrange elements and the finite dimensional approximation space $\mathbf{V}_h \subset \mathbf{H}^1(\Omega)$, thus, the discrete bilinear form inherits from (8.1.1.4).

**Remark 8.5.** *We only consider the approximation of a regular or non-singular solution* $\mathbf{Q}$ *of (8.1.1.2). This means that the Implicit Function Theorem can be applied in the Banach space* $\mathbf{H}^1(\Omega)$ *and it is equivalent to the following continuous inf-sup condition [MMN20, Equation (2.8)]:*

$$0 < \beta_Q \coloneqq \inf_{\substack{\Theta\in\mathbf{H}^1(\Omega)\\ \|\Theta\|_1=1}} \sup_{\substack{\Phi\in\mathbf{H}^1(\Omega)\\ \|\Phi\|_1=1}} \langle \mathcal{D}\mathcal{N}^n(\mathbf{Q})\Theta,\Phi\rangle = \inf_{\substack{\Phi\in\mathbf{H}^1(\Omega)\\ \|\Phi\|_1=1}} \sup_{\substack{\Theta\in\mathbf{H}^1(\Omega)\\ \|\Theta\|_1=1}} \langle \mathcal{D}\mathcal{N}^n(\mathbf{Q})\Theta,\Phi\rangle. \tag{8.1.1.5}$$

To deduce the $L^2$ error estimate of regular solutions one can use the Aubin–Nitsche duality argument; however due to the nonlinearity, it is nontrivial to derive the dual problem. To this end, we consider the following linear dual problem to the primary nonlinear problem (8.1.0.1): find $\mathbf{N} \in \mathbf{H}_0^1(\Omega)$ such that

$$\begin{cases} -K\Delta\mathbf{N} + 4l|\mathbf{Q}|^2\mathbf{N} + 8l(\mathbf{Q}:\mathbf{N})\mathbf{Q} - 2l\mathbf{N} = \mathbf{G} & \text{in } \Omega, \\ \mathbf{N} = \mathbf{0} & \text{on } \partial\Omega, \end{cases} \tag{8.1.1.6}$$

for a given $\mathbf{G} \in \mathbf{L}^2(\Omega)$ (we will see the choice of $\mathbf{G}$ in the proof of Theorem 8.8). Here, $\mathbf{Q} \in \mathbf{H}_b^1(\Omega)$. Furthermore, one can obtain the weak form of (8.1.1.6): find $\mathbf{N} \in \mathbf{H}_0^1(\Omega)$ such that

$$\langle \mathcal{D}\mathcal{N}^n(\mathbf{Q})\mathbf{N}, \Phi\rangle = A^n(\mathbf{N},\Phi) + 3B^n(\mathbf{Q},\mathbf{Q},\mathbf{N},\Phi) + C^n(\mathbf{N},\Phi) = (\mathbf{G},\Phi)_0. \tag{8.1.1.7}$$

The technique follows [MMN20], where their proofs based on dGFEM with the broken Sobolev space

$$\mathbf{H}_0^1(\mathcal{T}_h) = \left\{\mathbf{v} \in \mathbf{L}^2(\Omega) : \mathbf{v}|_T \in \mathbf{H}^1(T)\ \forall T \in \mathcal{T}_h, \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\right\},$$

are derived with the mesh-dependent norm

$$\|\mathbf{v}\|_{dG}^2 = \sum_{T\in\mathcal{T}_h}\int_T |\nabla\mathbf{v}|^2 + \sum_{e\in\mathcal{E}}\int_e \frac{\sigma_m}{h_e}[\![\mathbf{v}]\!]^2.$$

Here, $\sigma_m > 0$ is the penalty parameter. Moreover, for any interior edge $e \in \mathcal{E}_I$ shared by two cells $T_-$ and $T_+$, we define the jump $[\![\mathbf{v}]\!]$ by $[\![\mathbf{v}]\!] = \mathbf{v}_- \cdot \nu_- + \mathbf{v}_+ \cdot \nu_+$ with $\nu_-, \nu_+$ representing the restriction of outward normals in $T_-, T_+$ respectively. On the boundary edge/face $e \in \mathcal{E}_B$, we define $[\![\mathbf{v}]\!] = \mathbf{v} \cdot \nu$.

One can easily check that for a continuous approximation $\mathbf{v}_h \in \mathbf{H}_0^1(\Omega)$, it holds that $[\![\mathbf{v}_h]\!] = 0$ and the $\|\cdot\|_{dG}$-norm is in fact the $\mathbf{H}^1$ semi-norm in the Sobolev space $\mathbf{H}^1(\Omega)$ and equivalent to the $\|\cdot\|_1$-norm in the Sobolev space $\mathbf{H}_0^1(\Omega)$ by the Poincaré inequality. Hence, it is straightforward to derive similar results for the $\|\cdot\|_1$-norm as in [MMN20]. We give some auxiliary results about the operators $A^n(\cdot,\cdot)$, $B^n(\cdot,\cdot,\cdot,\cdot)$ and $C^n(\cdot,\cdot)$.

**Lemma 8.3.** *(Boundedness and coercivity of $A^n(\cdot,\cdot)$) For $\Theta, \Phi \in \mathbf{H}_0^1(\Omega)$, there holds*

$$A^n(\Theta, \Phi) \lesssim \|\Theta\|_1 \|\Phi\|_1,$$

*and*

$$\|\Theta\|_1^2 \lesssim A^n(\Theta, \Theta) \quad \forall \Theta \in \mathbf{H}_0^1(\Omega).$$

*Proof.* An application of the Cauchy–Schwarz inequality yields the boundedness result while the coercivity follows from the Poincaré inequality. □

**Lemma 8.4.** *(Boundedness of $B^n(\cdot,\cdot,\cdot,\cdot)$, $C^n(\cdot,\cdot)$) For $\Psi, \Phi, \Theta, \Xi \in \mathbf{H}^1(\Omega)$, there holds*

$$B^n(\Psi, \Phi, \Theta, \Xi) \lesssim \|\Psi\|_1 \|\Phi\|_1 \|\Theta\|_1 \|\Xi\|_1, \quad C^n(\Psi, \Phi) \lesssim \|\Psi\|_1 \|\Phi\|_1, \tag{8.1.1.8}$$

*and for $\Psi, \Phi \in \mathbf{H}^2(\Omega)$, $\Theta, \Xi \in \mathbf{H}^1(\Omega)$,*

$$B^n(\Psi, \Phi, \Theta, \Xi) \lesssim \|\Psi\|_2 \|\Phi\|_2 \|\Theta\|_1 \|\Xi\|_1. \tag{8.1.1.9}$$

*Proof.* For $\Psi, \Phi, \Theta, \Xi \in \mathbf{H}^1(\Omega)$, we use Hölder's inequality and the embedding result $\mathbf{H}^1(\Omega) \hookrightarrow \mathbf{L}^4(\Omega)$ to obtain

$$B^n(\Psi, \Phi, \Theta, \Xi) \lesssim \|\Psi\|_{L^4} \|\Phi\|_{L^4} \|\Theta\|_{L^4} \|\Xi\|_{L^4} \lesssim \|\Psi\|_1 \|\Phi\|_1 \|\Theta\|_1 \|\Xi\|_1.$$

The proof of (8.1.1.9) follows analogously to that of (8.1.1.8) with the use of the embedding result $\mathbf{H}^2(\Omega) \hookrightarrow \mathbf{L}^\infty(\Omega)$ and Cauchy–Schwarz inequality:

$$B^n(\Psi, \Phi, \Theta, \Xi) \lesssim \|\Psi\|_\infty \|\Phi\|_\infty \|\Theta\|_0 \|\Xi\|_0 \lesssim \|\Psi\|_2 \|\Phi\|_2 \|\Theta\|_1 \|\Xi\|_1.$$

This completes the proof. □

We also quote interpolation estimates that will be frequently used.

**Lemma 8.5.** *[BS08a] (Interpolation estimates) For $\mathbf{v} \in \mathbf{H}^2(\Omega)$ there exists $I_h\mathbf{v} \in \mathbf{V}_h$ such that*

$$\|\mathbf{v} - I_h\mathbf{v}\|_0 \lesssim h^2 \|\mathbf{v}\|_2,$$

$$\|\mathbf{v} - I_h\mathbf{v}\|_1 \lesssim h \|\mathbf{v}\|_2.$$

*Here, $I_h : \mathbf{H}^2 \to \mathbf{V}_h$ is the interpolation operator.*

To derive the $L^2$ a priori error estimates, we need two more auxiliary results.

**Lemma 8.6.** *For* $\mathbf{Q} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_b^1(\Omega)$, $\mathbf{N} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega)$ *and* $I_h\mathbf{Q} \in \mathbf{V}_h \subset \mathbf{H}_b^1(\Omega)$, *it holds that*

$$A^n(I_h\mathbf{Q} - \mathbf{Q}, \mathbf{N}) \lesssim h^2 \|\mathbf{Q}\|_2 \|\mathbf{N}\|_2.$$

*Proof.* By the definition of the bilinear form $A^n(\cdot,\cdot)$, integration by parts (note that $(I_h\mathbf{Q} - \mathbf{Q})|_{\partial\Omega} = 0$), Cauchy–Schwarz inequality and interpolation estimates from Lemma 8.5, we have

$$\begin{aligned} A^n(I_h\mathbf{Q} - \mathbf{Q}, \mathbf{N}) &= \int_\Omega K\nabla(I_h\mathbf{Q} - \mathbf{Q}) \cdot \nabla\mathbf{N} \\ &= -\int_\Omega K(I_h\mathbf{Q} - \mathbf{Q}) \cdot \Delta\mathbf{N} \\ &\lesssim \|I_h\mathbf{Q} - \mathbf{Q}\|_0 \|\mathbf{N}\|_2 \\ &\lesssim h^2 \|\mathbf{Q}\|_2 \|\mathbf{N}\|_2. \end{aligned}$$

This completes the proof. □

We then show that the $H^2$-norm of the dual solution is bounded by the source term $\mathbf{G} \in \mathbf{L}^2(\Omega)$.

**Lemma 8.7.** *(Boundedness of the dual solution in the $H^2$-norm) The solution* $\mathbf{N}$ *to the weak form (8.1.1.7) of the dual linear problem belongs to* $\mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega)$ *and it holds that*

$$\|\mathbf{N}\|_2 \lesssim \|\mathbf{G}\|_0. \tag{8.1.1.10}$$

*Proof.* We use the inf-sup condition (8.1.1.5) for the linear operator $\langle \mathcal{DN}^n(\mathbf{Q})\cdot, \cdot\rangle$, the weak formulation (8.1.1.7) and Cauchy–Schwarz inequality to obtain

$$\beta_Q \|\mathbf{N}\|_1 \leq \sup_{\substack{\Phi \in \mathbf{H}_0^1 \\ \|\Phi\|_1 = 1}} \langle \mathcal{DN}^n(\mathbf{Q})\mathbf{N}, \Phi\rangle = \sup_{\substack{\Phi \in \mathbf{H}_0^1 \\ \|\Phi\|_1 = 1}} (\mathbf{G}, \Phi)_0 \leq \|\mathbf{G}\|_0. \tag{8.1.1.11}$$

By the form of (8.1.1.7) and boundedness of $B^n(\mathbf{Q}, \mathbf{Q}, \cdot, \cdot)$ and $C^n(\cdot,\cdot)$, we have

$$\|K\Delta\mathbf{N}\|_0 = \| -3B^n(\mathbf{Q}, \mathbf{Q}, \cdot, \mathbf{N}) - C^n(\cdot, \mathbf{N}) + (\mathbf{G}, \cdot)_0\|_0 \lesssim \|\mathbf{N}\|_1 + \|\mathbf{G}\|_0. \tag{8.1.1.12}$$

Note that the linear dual problem (8.1.1.6) includes a Laplace operator. Using the elliptic regularity result on a domain with polygonal boundary (see e.g., [Gri85,

Theorem 4.3.1.4]) for Laplace operators, we deduce that $\mathbf{N} \in \mathbf{H}^2(\Omega)$. Combining Equations (8.1.1.11) and (8.1.1.12) and the fact that $\|\Delta \cdot\|_0$ is indeed a norm in $\mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega)$, we can get (8.1.1.10). □

Finally, we are ready to deduce the optimal $L^2$ error estimate.

**Theorem 8.8.** *($L^2$ error estimate) Let $\mathbf{Q}$ be a regular solution of the nonlinear weak problem (8.1.1.2). For sufficiently small mesh size $h$, there exists a unique approximate solution $\mathbf{Q}_h$ of the discrete problem (having the same weak formulation as (8.1.1.2)) such that*

$$\|\mathbf{Q} - \mathbf{Q}_h\|_0 \lesssim h^2 \left(2 + \left(3 + 2h + 2h^2\right) \|\mathbf{Q}\|_2^2\right) \|\mathbf{Q}\|_2. \tag{8.1.1.13}$$

*Proof.* We take $\mathbf{G} = I_h\mathbf{Q} - \mathbf{Q}_h$ in the linear dual problem (8.1.1.6), multiply (8.1.1.6) by $I_h\mathbf{Q} - \mathbf{Q}_h$ and integrate by parts to obtain the weak formulation

$$\langle \mathcal{DN}^n(\mathbf{Q})(I_h\mathbf{Q} - \mathbf{Q}_h), \mathbf{N}\rangle = \|I_h\mathbf{Q} - \mathbf{Q}_h\|_0^2.$$

Here, $\langle \mathcal{DN}^n(\mathbf{Q})(I_h\mathbf{Q} - \mathbf{Q}_h), \mathbf{N}\rangle = A^n(I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N}) + 3B^n(\mathbf{Q}, \mathbf{Q}, I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N}) + C^n(I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N})$. Since both $\mathbf{Q}$ and its approximation $\mathbf{Q}_h$ satisfy the weak formulation (8.1.1.2), we know

$$\mathcal{N}^n(\mathbf{Q})I_h\mathbf{N} = 0 \quad \text{and} \quad \mathcal{N}^n(\mathbf{Q}_h)I_h\mathbf{N} = 0.$$

By the definitions of the nonlinear operator $\mathcal{N}^n(\mathbf{Q})\cdot$ and bilinear form $\langle \mathcal{DN}^n(\mathbf{Q})\cdot, \cdot\rangle$, we calculate

$$\begin{aligned}
\|I_h\mathbf{Q} - \mathbf{Q}_h\|_0^2 &= \langle \mathcal{DN}^n(\mathbf{Q})(I_h\mathbf{Q} - \mathbf{Q}_h), \mathbf{N}\rangle + \mathcal{N}^n(\mathbf{Q}_h)I_h\mathbf{N} - \mathcal{N}^n(\mathbf{Q})I_h\mathbf{N} \\
&= A^n(I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N}) + 3B^n(\mathbf{Q}, \mathbf{Q}, I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N}) + C^n(I_h\mathbf{Q} - \mathbf{Q}_h, \mathbf{N}) \\
&\quad + A^n(\mathbf{Q}_h, I_h\mathbf{N}) + B^n(\mathbf{Q}_h, \mathbf{Q}_h, \mathbf{Q}_h, I_h\mathbf{N}) + C^n(\mathbf{Q}_h, I_h\mathbf{N}) \\
&\quad - A^n(\mathbf{Q}, I_h\mathbf{N}) - B^n(\mathbf{Q}, \mathbf{Q}, \mathbf{Q}, I_h\mathbf{N}) - C^n(\mathbf{Q}, I_h\mathbf{N}) \\
&= \underbrace{A^n(I_h\mathbf{Q} - \mathbf{Q}, \mathbf{N}) + A^n(\mathbf{Q} - \mathbf{Q}_h, \mathbf{N} - I_h\mathbf{N})}_{U_1} \\
&\quad + \underbrace{C^n(I_h\mathbf{Q} - \mathbf{Q}, \mathbf{N}) + C^n(\mathbf{Q} - \mathbf{Q}_h, \mathbf{N} - I_h\mathbf{N})}_{U_2}
\end{aligned}$$

$$
\begin{aligned}
&+ \underbrace{3B^n(\mathbf{Q},\mathbf{Q},I_h\mathbf{Q}-\mathbf{Q}_h,\mathbf{N}-I_h\mathbf{N})+3B^n(\mathbf{Q},\mathbf{Q},I_h\mathbf{Q}-\mathbf{Q},I_h\mathbf{N})}_{U_3}\\
&+ \underbrace{B^n(\mathbf{Q}_h,\mathbf{Q}_h,\mathbf{Q}_h,I_h\mathbf{N})-3B^n(\mathbf{Q},\mathbf{Q},\mathbf{Q}_h,I_h\mathbf{N})+2B^n(\mathbf{Q},\mathbf{Q},\mathbf{Q},I_h\mathbf{N})}_{U_4}\\
&=: U_1+U_2+U_3+U_4. \qquad (8.1.1.14)
\end{aligned}
$$

We now use the previous auxiliary results to bound $U_1,\cdots,U_4$ separately, yielding

$$
\begin{aligned}
U_1 &= A^n(I_h\mathbf{Q}-\mathbf{Q},\mathbf{N})+A^n(\mathbf{Q}-\mathbf{Q}_h,\mathbf{N}-I_h\mathbf{N})\\
&\lesssim h^2\|\mathbf{Q}\|_2\|\mathbf{N}\|_2+\|\mathbf{Q}-\mathbf{Q}_h\|_1\|\mathbf{N}-I_h\mathbf{N}\|_1, && \text{by Lemma 8.6 and Lemma 8.3,}\\
&\lesssim h^2\|\mathbf{Q}\|_2\|\mathbf{N}\|_2, && \text{by (8.1.1.1) and Lemma 8.5,}
\end{aligned}
\qquad (8.1.1.15)
$$

$$
\begin{aligned}
U_2 &= C^n(I_h\mathbf{Q}-\mathbf{Q},\mathbf{N})+C^n(\mathbf{Q}-\mathbf{Q}_h,\mathbf{N}-I_h\mathbf{N})\\
&\lesssim \|I_h\mathbf{Q}-\mathbf{Q}\|_0\|\mathbf{N}\|_0+\|\mathbf{Q}-\mathbf{Q}_h\|_1\|\mathbf{N}-I_h\mathbf{N}\|_1, && \text{by CS and (8.1.1.8),}\\
&\lesssim h^2\|\mathbf{Q}\|_2(\|\mathbf{N}\|_0+\|\mathbf{N}\|_2), && \text{by Lemma 8.5 and (8.1.1.1),}\\
&\lesssim h^2\|\mathbf{Q}\|_2\|\mathbf{N}\|_2, && \text{by } \|\mathbf{N}\|_0\le\|\mathbf{N}\|_2,
\end{aligned}
\qquad (8.1.1.16)
$$

and

$$
\begin{aligned}
U_3 &= 3B^n(\mathbf{Q},\mathbf{Q},I_h\mathbf{Q}-\mathbf{Q}_h,\mathbf{N}-I_h\mathbf{N})+3B^n(\mathbf{Q},\mathbf{Q},I_h\mathbf{Q}-\mathbf{Q},I_h\mathbf{N})\\
&\lesssim \|\mathbf{Q}\|_2^2\|I_h\mathbf{Q}-\mathbf{Q}_h\|_1\|\mathbf{N}-I_h\mathbf{N}\|_1+\|\mathbf{Q}\|_2^2\|I_h\mathbf{Q}-\mathbf{Q}\|_0\|I_h\mathbf{N}\|_0, && \text{by (8.1.1.9) and CS,}\\
&\lesssim h\|\mathbf{Q}\|_2^2\|I_h\mathbf{Q}-\mathbf{Q}_h\|_1\|\mathbf{N}\|_2+h^2\|\mathbf{Q}\|_2^3\|I_h\mathbf{N}\|_0, && \text{by Lemma 8.5.}
\end{aligned}
\qquad (8.1.1.17)
$$

Here, CS abbreviates for the Cauchy–Schwarz inequality. Note that by triangle inequality, Lemma 8.5 and (8.1.1.1), it holds that

$$
\begin{aligned}
&\|I_h\mathbf{Q}-\mathbf{Q}_h\|_1\le\|I_h\mathbf{Q}-\mathbf{Q}\|_1+\|\mathbf{Q}-\mathbf{Q}_h\|_1\lesssim h\|\mathbf{Q}\|_2,\\
&\|I_h\mathbf{N}\|_0\le\|I_h\mathbf{N}-\mathbf{N}\|_0+\|\mathbf{N}\|_0\lesssim(1+h^2)\|\mathbf{N}\|_2,\\
&\|I_h\mathbf{N}\|_1\le\|I_h\mathbf{N}-\mathbf{N}\|_1+\|\mathbf{N}\|_1\lesssim(1+h)\|\mathbf{N}\|_2. \qquad (8.1.1.18)
\end{aligned}
$$

Therefore, we further estimate $U_3$ in (8.1.1.17) to obtain

$$
U_3\lesssim h^2(2+h^2)\|\mathbf{Q}\|_2^3\|\mathbf{N}\|_2. \qquad (8.1.1.19)
$$

It remains to bound the $U_4$ term in (8.1.1.14). Let $\mathbf{E} = \mathbf{Q}_h - \mathbf{Q}$. We use the definition (8.1.1.3) of $B^n(\cdot,\cdot)$ and manipulate terms as follows:

$$\begin{aligned}
U_4 &= B^n(\mathbf{Q}_h,\mathbf{Q}_h,\mathbf{Q}_h,I_h\mathbf{N}) - 3B^n(\mathbf{Q},\mathbf{Q},\mathbf{Q}_h,I_h\mathbf{N}) + 2B^n(\mathbf{Q},\mathbf{Q},\mathbf{Q},I_h\mathbf{N}) \\
&= 4l\int_\Omega |\mathbf{Q}_h|^2(\mathbf{Q}_h\colon I_h\mathbf{N}) - 4l\int_\Omega \left(|\mathbf{Q}|^2(\mathbf{Q}_h\colon I_h\mathbf{N}) + 2(\mathbf{Q}\colon\mathbf{Q}_h)(\mathbf{Q}\colon I_h\mathbf{N})\right) \\
&\quad + 8l\int_\Omega |\mathbf{Q}|^2(\mathbf{Q}\colon I_h\mathbf{N}) \\
&= 4l\int_\Omega \left(|\mathbf{Q}_h|^2 - |\mathbf{Q}|^2\right)(\mathbf{Q}_h\colon I_h\mathbf{N}) - 8l\int_\Omega \mathbf{Q}\colon(\mathbf{Q}_h-\mathbf{Q})(\mathbf{Q}\colon I_h\mathbf{N}) \\
&= 4l\int_\Omega \mathbf{E}\colon(\mathbf{E}+2\mathbf{Q})(\mathbf{E}+\mathbf{Q})\colon I_h\mathbf{N} - 8l\int_\Omega(\mathbf{E}\colon\mathbf{Q})(\mathbf{Q}\colon I_h\mathbf{N}) \\
&= 4l\int_\Omega(\mathbf{E}\colon\mathbf{E})(\mathbf{E}\colon I_h\mathbf{N}) + 4l\int_\Omega(\mathbf{E}\colon\mathbf{E})(\mathbf{Q}\colon I_h\mathbf{N}) + 8l\int_\Omega(\mathbf{E}\colon\mathbf{Q})(\mathbf{E}\colon I_h\mathbf{N}).
\end{aligned}$$

Using the Hölder's inequality and the embedding result $\mathbf{H}^1(\Omega)\hookrightarrow\mathbf{L}^4(\Omega)$, we can bound $U_4$ further to obtain

$$\begin{aligned}
U_4 &\lesssim \|\mathbf{E}\|_1^3\|I_h\mathbf{N}\|_1 + \|\mathbf{E}\|_1^2\|\mathbf{Q}\|_1\|I_h\mathbf{N}\|_1 + \|\mathbf{E}\|_1^2\|\mathbf{Q}\|_1\|I_h\mathbf{N}\|_1 \\
&= \|\mathbf{E}\|_1^2(\|\mathbf{E}\|_1 + 2\|\mathbf{Q}\|_1)\|I_h\mathbf{N}\|_1 \\
&\lesssim (1+h)\|\mathbf{E}\|_1^2(\|\mathbf{E}\|_1 + \|\mathbf{Q}\|_1)\|\mathbf{N}\|_2, && \text{by (8.1.1.18),} \\
&\lesssim h^2(1+h)\|\mathbf{Q}\|_2^2(h\|\mathbf{Q}\|_2 + \|\mathbf{Q}\|_2)\|\mathbf{N}\|_2, && \text{by Lemma 8.5,} \\
&= h^2(1+h)^2\|\mathbf{Q}\|_2^3\|\mathbf{N}\|_2. && (8.1.1.20)
\end{aligned}$$

Combining the estimates (8.1.1.15), (8.1.1.16), (8.1.1.19) and (8.1.1.20) and applying Lemma 8.7 yields

$$\begin{aligned}
\|I_h\mathbf{Q} - \mathbf{Q}_h\|_0^2 &\lesssim h^2\left(2 + \left(2 + h^2 + (1+h)^2\right)\|\mathbf{Q}\|_2^2\right)\|\mathbf{Q}\|_2\|\mathbf{N}\|_2 \\
&\lesssim h^2\left(2 + \left(3 + 2h + 2h^2\right)\|\mathbf{Q}\|_2^2\right)\|\mathbf{Q}\|_2 \underbrace{\|\mathbf{G}\|_0}_{=\|I_h\mathbf{Q}-\mathbf{Q}_h\|_0},
\end{aligned}$$

implying that

$$\|I_h\mathbf{Q} - \mathbf{Q}_h\|_0 \lesssim h^2\left(2 + \left(3 + 2h + 2h^2\right)\|\mathbf{Q}\|_2^2\right)\|\mathbf{Q}\|_2. \tag{8.1.1.21}$$

By the triangle inequality and Lemma 8.5, we have

$$\|\mathbf{Q} - \mathbf{Q}_h\|_0 \le \|\mathbf{Q} - I_h\mathbf{Q}\|_0 + \|I_h\mathbf{Q} - \mathbf{Q}_h\|_0 \lesssim h^2\left(2 + \left(3 + 2h + 2h^2\right)\|\mathbf{Q}\|_2^2\right)\|\mathbf{Q}\|_2.$$

This completes the proof. □

**Remark 8.6.** *One can follow [MMN20] to obtain optimal error estimates in both norms* $\|\cdot\|_1$ *and* $\|\cdot\|_0$ *for higher degrees* ($\geq 2$) *of approximating polynomials. We omit further details since* $\|\cdot\|_1$ *is actually equivalent to the norm* $\|\cdot\|_{dG}$ *(in [MMN20]) in the* $\mathbf{H}_0^1(\Omega)$ *space and the technique can be directly applied to our case here.*

In this subsection, we have obtained the optimal a priori error estimates of the regular solution $\mathbf{Q}$ in the $L^2$-norm (see Theorem 8.8) and in the $H^1$-norm (see Theorem 8.2). We will verify this in our subsequent numerical experiments in Section 8.2 with different choices of approximations.

### 8.1.2 A priori error estimates for $(\mathcal{P}2)$

Since the PDE (8.1.0.3) for the density variation $u$ is a fourth order problem, a conforming discretisation requires a finite dimensional subspace of the Sobolev space $H^2(\Omega)$, which necessitates the use of $\mathcal{C}^1$-continuous elements. The construction of these elements is quite involved, particularly in three dimensions; without a special mesh structure, the lowest-degree conforming elements are the Argyris [AFS68] and Zhang [Zha09] elements, of degree 5 and 9 in two and three dimensions respectively. One approach to avoid this is to use mixed formulations by solving two second order systems, and we refer to [Sch78; CHH00] for instance. However, this substantially increases the size of the linear systems to be solved. Alternatively, one can directly tackle the fourth-order problem with non-conforming elements, that do not satisfy the $\mathcal{C}^1$-requirement. For instance, the so-called *continuous/discontinuous Galerkin* (C/DG) methods and $\mathcal{C}^0$ *interior penalty* methods ($\mathcal{C}^0$-IP) are analysed in [Eng+02; BS05], combining concepts from the theory of continuous and discontinuous Galerkin methods. Essentially, these methods use $\mathcal{C}^0$-conforming elements and penalise inter-element jumps in first derivatives to weakly enforce $\mathcal{C}^1$-continuity. This has the advantages of both convenience and efficiency: the weak form is simple, with only minor modifications from a conforming method, and fewer degrees of freedom are used than with a fully discontinuous Galerkin method.

We thus adopt the idea of $\mathcal{C}^0$-IP methods to solve the nonlinear fourth order problem ($\mathcal{P}2$). Specifically, we use the usual $\mathcal{C}^0$-continuous Lagrange elements and penalise jumps of the gradient across facets. In what follows, we derive some a priori error estimates for the fourth order problem ($\mathcal{P}2$) with the strong form derived in (8.1.0.3).

For simplicity, we only consider the cubic nonlinearity (i.e., $a_2 = 0$) in this analysis. The quadratic term can be tackled similarly. We therefore analyse the following strong form,

$$\begin{cases} 2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2u)) + a_1u + a_3u^3 = 0 & \text{in } \Omega, \\ u = u_b & \text{on } \partial\Omega, \\ \mathcal{D}^2u\cdot\nu = \mathcal{D}^2u_b\cdot\nu & \text{on } \partial\Omega. \end{cases} \tag{8.1.2.1}$$

The corresponding continuous weak form is defined as: find $u \in H^2(\Omega)\cap H^1_b(\Omega)$ such that

$$\mathcal{N}^s(u)v \coloneqq A^s(u,v) + B^s(u,u,u,v) + C^s(u,v) = L^s(v) \quad \forall v \in H^2(\Omega)\cap H^1_0(\Omega), \tag{8.1.2.2}$$

where for $v, w \in H^2(\Omega)$,

$$\begin{aligned} A^s(v,w) &= 2B\int_\Omega \mathcal{D}^2v : \mathcal{D}^2w, \\ C^s(v,w) &= a_1\int_\Omega vw, \\ L^s(v) &\coloneqq 2B\int_{\partial\Omega}\left(\mathcal{D}^2u_b\cdot\nabla v\right)\cdot\nu, \end{aligned}$$

and for $\mu, \zeta, \eta, \xi \in H^2(\Omega)$,

$$B^s(\mu,\zeta,\eta,\xi) = a_3\int_\Omega \mu\zeta\eta\xi.$$

Since (8.1.2.2) is nonlinear, we derive its linearisation: find $v \in H^2(\Omega)\cap H^1_0(\Omega)$ such that

$$\langle \mathcal{D}\mathcal{N}^s(u)v, w\rangle_{H^2} \coloneqq A^s(v,w) + 3B^s(u,u,v,w) + C^s(v,w) = L^s(w) \quad \forall w \in H^2(\Omega)\cap H^1_0(\Omega), \tag{8.1.2.3}$$

where $\langle\cdot,\cdot\rangle_{H^2}$ represents the dual pairing between $(H^2(\Omega)\cap H^1_0(\Omega))^*$ and $H^2(\Omega)\cap H^1_0(\Omega)$.

It is straightforward to derive the coercivity and boundedness of the bilinear operator $A^s(\cdot,\cdot)$ with the semi-norm $|\cdot|_2$ (in fact, this is indeed a norm in $H^2(\Omega)\cap H_0^1(\Omega)$).

**Lemma 8.9.** *For $v, w \in H^2(\Omega)\cap H_0^1(\Omega)$, there holds*

$$A^s(v,w) \lesssim |v|_2|w|_2 \quad \textit{and} \quad A^s(v,v) \gtrsim |v|_2^2.$$

Define the broken Sobolev space by

$$H^2(\mathcal{T}_h) \coloneqq \{v \in H^1(\Omega) : v|_T \in H^2(T)\ \forall T \in \mathcal{T}_h\},$$

equipped with the broken norm $\|v\|_{2,\mathcal{T}_h}^2 = \sum_{T\in\mathcal{T}_h}\|v\|_{2,T}^2$.

We take the nonconforming but still continuous approximation $u_h$ for the solution $u$ of (8.1.2.2), that is to say, $u_h \in W_{h,b} \subset H^2(\mathcal{T}_h)\cap H_b^1(\Omega)$ with some related definitions for deg $\geq 2$

$$\begin{aligned}
W_h &\coloneqq \{v \in H^2(\mathcal{T}_h)\cap H^1(\Omega) : v \in \mathbb{Q}_{\deg}(T)\ \forall T \in \mathcal{T}_h\},\\
W_{h,0} &\coloneqq \{v \in H^2(\mathcal{T}_h)\cap H^1(\Omega) : v = 0 \text{ on } \partial\Omega, v \in \mathbb{Q}_{\deg}(T)\ \forall T \in \mathcal{T}_h\},\\
W_{h,b} &\coloneqq \{v \in H^2(\mathcal{T}_h)\cap H^1(\Omega) : v = u_b \text{ on } \partial\Omega, v \in \mathbb{Q}_{\deg}(T)\ \forall T \in \mathcal{T}_h\}.
\end{aligned}$$

Following the derivation of $\mathcal{C}^0$-IP formulation similar to [Bre11, Section 3], we introduce the discrete nonlinear weak form: find $u_h \in W_{h,b}$ such that

$$\mathcal{N}_h^s(u_h)v_h \coloneqq A_h^s(u_h,v_h)+P_h^s(u_h,v_h)+B_h^s(u_h,u_h,u_h,v_h)+C_h^s(u_h,v_h) = L^s(v_h) \quad \forall v_h \in W_{h,0}, \tag{8.1.2.4}$$

where for all $u, v, \mu, \zeta, \eta, \xi \in W_h$,

$$A_h^s(u,v) \coloneqq 2B\bigg(\sum_{T\in\mathcal{T}_h}\int_T \mathcal{D}^2u\colon\mathcal{D}^2v - \sum_{e\in\mathcal{E}_I}\int_e\left\{\!\!\left\{\frac{\partial^2 u}{\partial\nu^2}\right\}\!\!\right\}[\![\nabla v]\!] - \sum_{e\in\mathcal{E}_I}\int_e\left\{\!\!\left\{\frac{\partial^2 v}{\partial\nu^2}\right\}\!\!\right\}[\![\nabla u]\!]\bigg),$$

$$C_h^s(u,v) = C^s(u,v) = a_1\int_\Omega uv,$$

$$B_h^s(\mu,\zeta,\eta,\xi) = B^s(\mu,\zeta,\eta,\xi) = a_3\int_\Omega \mu\zeta\eta\xi,$$

and

$$P_h^s(u,v) \coloneqq \sum_{e\in\mathcal{E}_I}\frac{2B\epsilon}{h_e^3}\int_e [\![\nabla u]\!][\![\nabla v]\!]. \tag{8.1.2.5}$$

Here, $\epsilon$ is the penalty parameter (to be specified in the implementations later), the average $\left\{\left\{\frac{\partial^2 u}{\partial \nu^2}\right\}\right\}$ of the second derivatives of $u$ along tangential directions across $e$ is defined as

$$\left\{\left\{\frac{\partial^2 u}{\partial \nu^2}\right\}\right\} = \frac{1}{2}\left(\frac{\partial^2 u_+}{\partial \nu^2}\bigg|_e + \frac{\partial^2 u_-}{\partial \nu^2}\bigg|_e\right),$$

with $\nu$ denoting the outward normal. In fact, the operator $P_h^s$ penalises the first derivatives across the interior facet since the function in $H^1(\Omega)$ is not necessarily continuously differentiable.

**Remark 8.7.** *The nonlinear problems (8.1.2.2) and (8.1.2.4) are equivalent for the solution $u$ of the strong form (8.1.0.3) since the jump term $[\![\nabla u]\!]$ vanishes for $u \in H^2(\Omega)$, however they are* not *equivalent for $u_h \in W_{h,b} \subset H^1(\Omega)$.*

The linearised version of the discrete nonlinear problem (8.1.2.4) yields the following discrete linear weak form: seek $v_h \in W_{h,0}$ such that

$$\langle \mathcal{DN}_h^s(u_h)v_h, w_h\rangle = L^s(w_h) \quad \forall w_h \in W_{h,0}, \tag{8.1.2.6}$$

where

$$\langle \mathcal{DN}_h^s(u_h)v_h, w_h\rangle := A_h^s(v_h, w_h) + P_h^s(v_h, w_h) + 3B_h^s(u_h, u_h, v_h, w_h) + C_h^s(v_h, w_h). \tag{8.1.2.7}$$

We also define the mesh-dependent $H^2$-like semi-norm for $v \in W_h$,

$$|||v|||_h^2 := \sum_{T\in\mathcal{T}_h} |v|^2_{H^2(T)} + \sum_{e\in\mathcal{E}_I}\int_e \frac{1}{h_e^3}|[\![\nabla v]\!]|^2. \tag{8.1.2.8}$$

Note that $|||\cdot|||_h$ is indeed a norm on $W_{h,0}$. This norm will be used in the well-posedness and convergence analysis below.

We first give an immediate result about the consistency of the discrete form (8.1.2.4).

**Theorem 8.10.** *(Consistency) Assuming that $u \in H^4(\Omega)$. The solution $u$ of the continuous weak form (8.1.2.2) solves the discrete weak problem (8.1.2.4).*

*Proof.* Multiplying the fourth order term $2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2u))$ in (8.1.2.1) with $v \in W_{h,0}$ and using piecewise integration by parts with the boundary condition specified in (8.1.2.1) for $u$, one can obtain

$$2B\sum_{T\in\mathcal{E}_h}\int_T \nabla\cdot(\nabla\cdot(\mathcal{D}^2u))v = 2B\sum_{T\in\mathcal{E}_h}\int_T \mathcal{D}^2u\colon\mathcal{D}^2v - 2B\sum_{e\in\mathcal{E}_I}\int_e \left\{\!\!\left\{\frac{\partial^2u}{\partial\nu^2}\right\}\!\!\right\}[\![\nabla v]\!]. \tag{8.1.2.9}$$

Since $u \in H^4(\Omega)$ implies $\nabla u$ is continuous on the whole domain $\Omega$, the jump term $[\![\nabla u]\!]$ then becomes zero and we can thus symmetrise and penalise the form (8.1.2.9). This leads to the presence of $A_h^s(u,v) + P_h^s(u,v)$. The remaining terms involving $B_h^s$ and $C_h^s$ are straightforward as one takes the test function $v \in W_{h,0}$. Therefore, $u$ satisfies (8.1.2.4). □

### 8.1.2.1 Elliptic regularity

Essentially, the strong form (8.1.2.1) is similar to the model problem given as [Bre11, Example 2] of form

$$\begin{aligned} \Delta^2u &= f \quad \text{in } \Omega,\\ u = \Delta u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{8.1.2.10}$$

**Remark 8.8.** *The boundary condition for the second derivative of $u$ in (8.1.2.10) is different from what we have imposed in (8.1.2.1). We just want to comment about the regularity of the problem (8.1.2.1) by extending the results for (8.1.2.10).*

Noticing that

$$\left(\mathcal{D}^2\colon\mathcal{D}^2\right)u = \left[\left(\partial_x^2\right)^2 + \left(\partial_y^2\right)^2 + 2\left(\partial_{xy}^2\right)^2\right]u = \Delta^2u,$$

it is natural to extend the classical elliptic regularity result [BB80] for the biharmonic operator $\Delta^2$ to the case of the bi-Hessian operator $\mathcal{D}^2\colon\mathcal{D}^2$. In general, the weak solution of (8.1.2.10) in a bounded polygonal domain $\Omega$ belongs to $H^{2+\varkappa}(\Omega)$ for some elliptic regularity index $\varkappa \in (0,2]$. More specifically, by [BB80, Theorem 2], we know that if each interior angle is smaller than $\pi/2$, then for $f \in H^{-1}(\Omega)$ there holds

$$\|u\|_{H^3(\Omega)} \lesssim \|f\|_{H^{-1}(\Omega)}.$$

In addition, if the domain $\Omega$ is smooth, the weak solutions even belong to $H^4(\Omega)$ by classical elliptic regularity results and thus we take this as an assumption throughout the analysis for simplicity.

Hence, we assume the solution $u$ of the strong form (8.1.2.1) is sufficiently regular in what follows and only consider to approximate such regular or non-singular solutions of the continuous weak form (8.1.2.2). Moreover, to facilitate the following analysis, we further assume that $u$ is an isolated solution, i.e., within a sufficiently small ball $\{v \in H^2(\Omega) \cap H_0^1(\Omega) : |v-u|_2 \leq r_b\}$ with radius $r_b$, there is only one solution $u$ satisfying (8.1.2.1). These assumptions then imply that the linearised operator $\langle \mathcal{DN}^s(u)\cdot,\cdot\rangle_{H^2}$ satisfies the following inf-sup condition:

$$0 < \beta_u = \inf_{\substack{v\in H^2(\Omega)\cap H_0^1(\Omega)\\ |v|_2=1}} \sup_{\substack{w\in H^2(\Omega)\cap H_0^1(\Omega)\\ |w|_2=1}} \langle \mathcal{DN}^s(u)v, w\rangle_{H^2} = \inf_{\substack{w\in H^2(\Omega)\cap H_0^1(\Omega)\\ |w|_2=1}} \sup_{\substack{v\in H^2(\Omega)\cap H_0^1(\Omega)\\ |v|_2=1}} \langle \mathcal{DN}^s(u)v, w\rangle_{H^2} \tag{8.1.2.11}$$

### 8.1.2.2 Well-posedness of the discrete form

Recalling [Bre11, Eq. (3.20)] that for $u, v \in W_{h,0}$,

$$\sum_{e\in\mathcal{E}_I} \left| \int_e \left\{\!\!\left\{ \frac{\partial^2 w}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla v]\!] \right| \lesssim \left( \sum_{T\in\mathcal{T}_h} \int_T \mathcal{D}^2 w \colon \mathcal{D}^2 w \right)^{1/2} \left( \sum_{e\in\mathcal{E}_I} \frac{1}{h_e} \int_e ([\![\nabla v]\!])^2 \right)^{1/2},$$

we can immediately obtain

$$\sum_{e\in\mathcal{E}_I} \left| \int_e \left\{\!\!\left\{ \frac{\partial^2 w}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla v]\!] \right| \lesssim \left( \sum_{T\in\mathcal{T}_h} \int_T \mathcal{D}^2 w \colon \mathcal{D}^2 w \right)^{1/2} \left( \sum_{e\in\mathcal{E}_I} \frac{1}{h_e^3} \int_e ([\![\nabla v]\!])^2 \right)^{1/2}, \tag{8.1.2.12}$$

since the edge or facet size $h_e < 1$. With the estimate (8.1.2.12) at hand, we can apply the Cauchy–Schwarz inequality and use the definition (8.1.2.8) of $|||\cdot|||_h$ to obtain the boundedness of $A_h^s(\cdot,\cdot)$ and $P_h^s(\cdot,\cdot)$. That is to say, for $u, v \in W_{h,0}$, there holds

$$|A_h^s(u,v)| \lesssim |||u|||_h |||v|||_h,$$
$$|P_h^s(u,v)| \lesssim |||u|||_h |||v|||_h.$$

We omit the details of their proofs here and only illustrate the boundedness result for $B_h^s(\cdot,\cdot,\cdot,\cdot)$ and $C_h^s(\cdot,\cdot)$ below.

**Lemma 8.11.** *(Boundedness of $B_h^s(\cdot,\cdot,\cdot,\cdot)$ and $C_h^s(\cdot,\cdot)$) For $u,v,w,p \in W_{h,0}$, we have*

$$\begin{aligned} |B_h^s(u,v,w,p)| &\lesssim |||u|||_h |||v|||_h |||w|||_h |||p|||_h, \\ |C_h^s(u,v)| &\lesssim |||u|||_h |||v|||_h. \end{aligned} \tag{8.1.2.13}$$

*For $u,v \in H^2(\Omega)$, $w,p \in W_h$,*

$$|B_h^s(u,v,w,p)| \lesssim \|u\|_2 \|v\|_2 |||w|||_h |||p|||_h. \tag{8.1.2.14}$$

*Proof.* By Hölder's inequality, Sobolev embedding $H^1(\Omega) \hookrightarrow L^4(\Omega)$, and the fact that the $H^1$ semi-norm $|\cdot|_1$ is a norm in $H_0^1(\Omega)$, we deduce

$$\begin{aligned} |B_h^s(u,v,w,p)| &\lesssim \|u\|_{L^4} \|v\|_{L^4} \|w\|_{L^4} \|p\|_{L^4} \\ &\lesssim |u|_1 |v|_1 |w|_1 |p|_1. \end{aligned}$$

It then follows from a Poincaré inequality [Bre03] [BS05, Eq. (4.22)] for piecewise $H^1$ functions that

$$\sum_{T\in\mathcal{T}_h} |v|_{1,T}^2 \lesssim \sum_{T\in\mathcal{T}_h} |v|_{2,T}^2 + \sum_{e\in\mathcal{E}_I} \frac{1}{h_e^3} \|[\![\nabla v]\!]\|_{0,e}^2 = |||v|||_h^2 \quad \forall v \in W_{h,0}. \tag{8.1.2.15}$$

Thus, we obtain

$$|B_h^s(u,v,w,p)| \lesssim |||u|||_h |||v|||_h |||w|||_h |||p|||_h.$$

The boundedness of $C_h^s(\cdot,\cdot)$ follows similarly by Cauchy–Schwarz inequality, the Sobolev embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ and the use of (8.1.2.15).

The proof of (8.1.2.14) is analogous to that of (8.1.2.13) with a use of the embedding result $H^2(\Omega) \hookrightarrow L^\infty(\Omega)$ and the Cauchy–Schwarz inequality. □

We give the coercivity result for the bilinear form $(A_h^s(\cdot,\cdot) + P_h^s(\cdot,\cdot))$.

**Lemma 8.12.** *(Coercivity of $A_h^s + P_h^s$) For a sufficiently large penalty parameter $\epsilon$, there holds*

$$|||v_h|||_h^2 \lesssim A_h^s(v_h,v_h) + P_h^s(v_h,v_h) \quad \forall v_h \in W_{h,0}. \tag{8.1.2.16}$$

*Proof.* By (8.1.2.12) and the inequality of geometric and arithmetic means, we deduce for $v \in W_h$,

$$\begin{aligned} A_h^s(v,v) + P_h^s(v,v) &\geq 2B \sum_{T\in\mathcal{T}_h} |v|^2_{H^2(T)} - 2BC \left( \sum_{T\in\mathcal{T}_h} |v|^2_{2,T} \right)^{1/2} \left( \sum_{e\in\mathcal{E}_I} \frac{1}{h_e^3} \|[\![\nabla v]\!]\|^2_{0,e} \right)^{1/2} \\ &\quad + 2B \left( \sum_{e\in\mathcal{E}_I} \int_e \frac{\epsilon}{h_e^3} |[\![\nabla v]\!]|^2 \right) \\ &\geq 2B \left[ \frac{1}{2} \sum_{T\in\mathcal{T}_h} |v|^2_{H^2(T)} + \left( \epsilon - \frac{C^2}{2} \right) \sum_{e\in\mathcal{E}_I} \frac{1}{h_e^3} \|[\![\nabla v]\!]\|^2_{0,e} \right] \\ &\geq B |\!|\!| v |\!|\!|^2_h, \end{aligned}$$

provided the penalty parameter $\epsilon$ is sufficiently large with the generic constant $C$ from (8.1.2.12). □

An important question about the well-posedness is the coercivity of the bilinear operator $\langle \mathcal{DN}_h^s(u_h)\cdot,\cdot\rangle$. Due to the presence of $B_h^s$ and $C_h^s$ terms in $\langle \mathcal{DN}_h^s(u_h)\cdot,\cdot\rangle$, it is not trivial to derive its coercivity. Instead, we discuss the weak coercivity of the bilinear form $\langle \mathcal{DN}_h^s(u)\cdot,\cdot\rangle$ defined as

$$\langle \mathcal{DN}_h^s(u)v_h, w_h\rangle := A_h^s(v_h,w_h) + P_h^s(v_h,w_h) + 3B_h^s(u,u,v_h,w_h) + C_h^s(v_h,w_h) \quad \forall v_h, w_h \in W_h. \tag{8.1.2.17}$$

We first give a useful lemma illustrating some estimates related to the enrichment operator $E_h : W_h \to W_C \subset H^2(\Omega)$ with $W_C$ being the Hsieh–Clough–Tocher macro finite element space. The degrees of freedom of $w \in W_C$ include: (i) the values of the derivatives of $w$ up to order 1 at the interior vertices and (ii) the values of the normal derivative of $w$ at the midpoints of the interior edges/facets in $\mathcal{E}_I$. The following lemma is adapted to our notations and definition of $|\!|\!|\cdot|\!|\!|_h$ using the result [Bre11, Lemma 1].

**Lemma 8.13.** *[Bre11, Lemma 1] For $v_h \in W_{h,0}$, there holds that*

$$\begin{aligned} &\sum_{T\in\mathcal{T}_h} \left( h^{-4} \|v_h - E_h v_h\|^2_{L^2(T)} + h^{-2} |v_h - E_h v_h|^2_{H^1(T)} + |v_h - E_h v_h|^2_{H^2(T)} \right) \\ &\quad \lesssim \sum_{e\in\mathcal{E}_I} \frac{1}{h_e^3} \|[\![\nabla v_h]\!]\|^2_{L^2(e)} \lesssim |\!|\!| v_h |\!|\!|^2_h. \end{aligned} \tag{8.1.2.18}$$

We can obtain the discrete inf-sup condition for the discrete bilinear operator $\langle \mathcal{DN}_h^s(u)\cdot,\cdot\rangle$.

**Theorem 8.14.** *(Weak coercivity of $\langle \mathcal{DN}_h^s(u)\cdot,\cdot\rangle$) Let $u$ be a regular isolated solution of the nonlinear continuous weak form (8.1.2.4). For a sufficiently large $\epsilon$ and a sufficiently small mesh size $h$, the following discrete inf-sup condition holds on a smooth domain $\Omega$ with a positive constant $\beta_c > 0$:*

$$0 < \beta_c \leq \inf_{\substack{v\in W_h \\ |||v_h|||_h=1}} \sup_{\substack{w\in W_h \\ |||w_h|||_h=1}} \langle \mathcal{DN}_h^s(u)v_h, w_h\rangle. \tag{8.1.2.19}$$

*Proof.* For $v \in H^2(\Omega)\cap H_0^1(\Omega)$, it follows from the boundedness result of $B_h^s, C_h^s$ that $B_h^s(u,u,v,\cdot)$, $B^s(u,u,v,\cdot)$, $C_h^s(v,\cdot)$ and $C^s(v,\cdot) \in L^2(\Omega)$. Furthermore, since $A^s(\cdot,\cdot)$ is bounded and coercive as given by Lemma 8.9, for a given $v_h \in W_h$ with $|||v_h|||_h = 1$, there exists $\xi$ and $\eta \in H^4(\Omega)\cap H_0^1(\Omega)$ that solve the linear systems:

$$A^s(\xi, w) = 3B_h^s(u,u,v_h,w) + C_h^s(v_h,w) \quad \forall w \in H^2(\Omega)\cap H_0^1(\Omega), \tag{8.1.2.20a}$$

$$A^s(\eta, w) = 3B^s(u,u,E_hv_h,w) + C^s(E_hv_h,w) \quad \forall w \in H^2(\Omega)\cap H_0^1(\Omega). \tag{8.1.2.20b}$$

It then follows from the standard elliptic regularity result that $\|\eta\|_4 \lesssim C_{BC}$ with constant $C_{BC}$ depending on $\|u\|_2$.

Subtracting (8.1.2.20a) from (8.1.2.20b), then taking $w = \eta - \xi$ and using the coercivity of $A^s(\cdot,\cdot)$ and boundedness of $B_h^s, C_h^s$, we obtain

$$\begin{aligned} |\eta-\xi|_2 &\lesssim \left(3\|u\|_2^2 + 1\right)\|E_hv_h - v_h\|_0 \\ &\lesssim h^2 \underbrace{|||v_h|||_h}_{=1} \qquad \text{by Lemma 8.13.} \end{aligned} \tag{8.1.2.21}$$

Here, we have used the fact that $B_h^s$ and $C_h^s$ are in fact equivalent to $B^s$ and $C^s$, respectively, by their definitions. Since $u$ is a regular isolated solution of (8.1.2.2), it yields by (8.1.2.11) that there exists $w \in H^2(\Omega)\cap H_0^1(\Omega)$ with $|w|_2 = 1$ such that

$$\begin{aligned} |E_hv_h|_2 &\lesssim \langle \mathcal{DN}^s(u)E_hv_h, w\rangle_{H^2} \\ &= A^s(E_hv_h,w) + 3B^s(u,u,E_hv_h,w) + C^s(E_hv_h,w) \\ &= A^s(E_hv_h+\eta, w) \qquad \text{by (8.1.2.20b),} \end{aligned}$$

$$
\begin{aligned}
&\lesssim |E_h v_h + \eta|_2 \underbrace{|w|_2}_{=1} && \text{by Lemma 8.9,}\\
&= \|\!|E_h v_h + \eta|\!\|_h && \text{since } E_h v_h + \eta \in H^2,\\
&\le \|\!|E_h v_h - v_h|\!\|_h + \|\!|v_h + I_h\xi|\!\|_h + \|\!|I_h\xi - \xi|\!\|_h + \underbrace{\|\!|\xi - \eta|\!\|_h}_{=|\xi-\eta|_2} && \text{by triangle inequality.}
\end{aligned}
\tag{8.1.2.22}
$$

Note that $[\![\nabla\xi]\!] = 0$ on $\mathcal{E}_I$ since $\xi \in H^4(\Omega)$. We can thus calculate

$$
\begin{aligned}
\|\!|E_h v_h - v_h|\!\|_h^2 &\lesssim \sum_{e\in\mathcal{E}_I} \int_e \frac{1}{h_e^3} |[\![\nabla v_h]\!]|^2 && \text{by Lemma 8.13,}\\
&\lesssim \sum_{e\in\mathcal{E}_I} \int_e \frac{1}{h_e^3} |[\![\nabla(v_h + \xi)]\!]|^2\\
&\le \|\!|v_h + \xi|\!\|_h^2.
\end{aligned}
$$

Further, by the triangle inequality, we get

$$
\|\!|E_h v_h - v_h|\!\|_h \lesssim \|\!|v_h + \xi|\!\|_h \le \|\!|v_h + I_h\xi|\!\|_h + \|\!|\xi - I_h\xi|\!\|_h. \tag{8.1.2.23}
$$

Since $v_h + I_h\xi \in W_h$, it follows from the coercivity result (8.1.2.16) that there exists $w_h \in W_h$ with $\|\!|w|\!\|_h = 1$ such that

$$
\begin{aligned}
\|\!|v_h + I_h\xi|\!\|_h &\lesssim A_h^s(v_h + I_h\xi, w_h) + P_h^s(v_h + I_h\xi, w_h)\\
&= \langle \mathcal{DN}_h^s(u)v_h, w_h\rangle - 3B_h^s(u, u, v_h, w_h) - C_h^s(v_h, w_h)\\
&\quad + A_h^s(I_h\xi - \xi, w_h) + P_h^s(I_h\xi - \xi, w_h) + A_h^s(\xi, w_h) + P_h^s(\xi, w_h)\\
&= \langle \mathcal{DN}_h^s(u)v_h, w_h\rangle + 3B_h^s(u, u, v_h, E_h w_h - w_h) + C_h^s(v_h, E_h w_h - w_h)\\
&\quad + A_h^s(I_h\xi - \xi, w_h) + P_h^s(I_h\xi - \xi, w_h)\\
&\quad + A_h^s(\xi, w_h - E_h w_h) + P_h^s(\xi, w_h - E_h w_h),
\end{aligned}
\tag{8.1.2.24}
$$

where in the last equality we have used the fact that

$$
3B_h^s(u, u, v_h, E_h w_h) + C_h^s(v_h, E_h w_h) = A^s(\xi, E_h w_h) = A_h^s(\xi, E_h w_h) + P^s(\xi, E_h w_h)
$$

because of (8.1.2.20a) and $[\![\nabla\xi]\!] = [\![\nabla E_h w_h]\!] = 0$.

Using the boundedness result Lemma 8.11 and the enrichment estimates Lemma 8.13, we obtain

$$
3B_h^s(u, u, v_h, E_h w_h - w_h) + C_h^s(v_h, E_h w_h - w_h) \lesssim \underbrace{\|v_h\|_0}_{\lesssim |v_h|_1 \lesssim \|\!|v_h|\!\|_h = 1} \underbrace{\|E_h w_h - w_h\|_0}_{\lesssim h^2 \|\!|w_h|\!\|_h = h^2}.
\tag{8.1.2.25}
$$

By the boundedness of the bilinear form $A_h^s + P_h^s$ and standard interpolation estimates, we have

$$\begin{aligned} A_h^s(I_h\xi - \xi, w_h) + P_h^s(I_h\xi - \xi, w_h) &\lesssim |||I_h\xi - \xi|||_h \underbrace{|||w_h|||_h}_{=1} \\ &\lesssim h^{\min\{\deg -1, 2\}} \|\xi\|_4, \end{aligned} \tag{8.1.2.26}$$

where $\deg \geq 2$ denotes the degree of the approximating polynomials. Moreover, by the enrichment estimate Lemma 8.13 and the fact that $[\![\nabla \xi]\!] = [\![\nabla(E_h w_h)]\!] = 0$, there holds

$$\begin{aligned} &A_h^s(\xi, w_h - E_h w_h) + P_h^s(\xi, w_h - E_h w_h) \\ &\quad = 2B \sum_{T\in\mathcal{T}_h} \int_T \mathcal{D}^2\xi : \mathcal{D}^2(w_h - E_h w_h) - 2B \sum_{e\in\mathcal{E}_I} \int_e \left\{\!\!\left\{ \frac{\partial^2 \xi}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla(w_h - E_h w_h)]\!] \\ &\quad = 2B \sum_{T\in\mathcal{T}_h} \nabla \cdot \Big( \nabla \cdot (\mathcal{D}^2\xi) \Big) (w_h - E_h w_h) && \text{by (8.1.2.9)}, \\ &\quad \lesssim \|\xi\|_4 \|w_h - E_h w_h\|_0 \\ &\quad \lesssim h^2 \|\xi\|_4 && \text{by Lemma 8.13.} \end{aligned} \tag{8.1.2.27}$$

Combine Equations (8.1.2.25) to (8.1.2.27) in (8.1.2.24) to obtain

$$|||E_h v_h - v_h|||_h \lesssim \langle \mathcal{D}\mathcal{N}_h^s(u) v_h, w_h \rangle + h^2 + h^{\min\{\deg -1, 2\}}. \tag{8.1.2.28}$$

Substituting (8.1.2.28) into (8.1.2.23) and using standard interpolation estimates yield that

$$|||E_h v_h - v_h|||_h \lesssim \langle \mathcal{D}\mathcal{N}_h^s(u) v_h, w_h \rangle + h^2 + h^{\min\{\deg -1, 2\}}. \tag{8.1.2.29}$$

A use of Equations (8.1.2.28) and (8.1.2.29), standard interpolation estimates and (8.1.2.21) in (8.1.2.22) leads to

$$|E_h v_h|_2 \lesssim \langle \mathcal{D}\mathcal{N}_h^s(u) v_h, w_h \rangle + h^2 + h^{\min\{\deg -1, 2\}}.$$

Then, by the triangle inequality, we have

$$\begin{aligned} 1 = |||v_h|||_h &\leq |||v_h - E_h v_h|||_h + \underbrace{|||E_h v_h|||_h}_{=|E_h v_h|_2} \\ &\leq C_t \left( \langle \mathcal{D}\mathcal{N}_h^s(u) v_h, w_h \rangle + h^2 + h^{\min\{\deg -1, 2\}} \right). \end{aligned}$$

Therefore, for the mesh size $h$ satisfying

$$h^2 + h^{\min\{\deg -1,2\}} < \frac{1}{2C_t},$$

the discrete inf-sup condition (8.1.2.19) holds for $\beta_c = \frac{1}{2C_t}$. □

Moreover, we can obtain the discrete inf-sup condition for the perturbed bilinear form $\langle \mathcal{DN}_h^s(I_h u)\cdot,\cdot\rangle$, i.e.,

$$\langle \mathcal{DN}_h^s(I_h u)v_h, w_h\rangle = A_h^s(v_h,w_h)+P_h^s(v_h,w_h)+3B_h^s(I_h u, I_h u, v_h, w_h)+C_h^s(v_h,w_h) \quad \forall v_h, w_h \in W_h. \tag{8.1.2.30}$$

**Theorem 8.15.** *(Weak coercivity of $\langle \mathcal{DN}_h^s(I_h u)\cdot,\cdot\rangle$) Let $u$ be a regular isolated solution of the nonlinear continuous weak form (8.1.2.4) and $I_h u$ the interpolation of $u$. For a sufficiently large $\epsilon$ and a sufficiently small mesh size $h$, the following discrete inf-sup condition holds:*

$$0 < \frac{\beta_c}{2} \le \inf_{\substack{v_h\in W_h \\ |||v_h|||_h=1}} \sup_{\substack{w_h\in W_h \\ |||w_h|||_h=1}} \langle \mathcal{DN}_h^s(I_h u)v_h, w_h\rangle. \tag{8.1.2.31}$$

*Proof.* Denote $\tilde{u} = u - I_h u$. By the definition (8.1.2.30) of the bilinear form $\langle \mathcal{DN}_h^s(I_h u)\cdot,\cdot\rangle$, we have

$$\langle \mathcal{DN}_h^s(I_h u)v_h, w_h\rangle = A_h^s(v_h,w_h)+P_h^s(v_h,w_h)+3B_h^s(u-\tilde{u}, u-\tilde{u}, v_h, w_h)+C_h^s(v_h,w_h).$$

It follows from the definition of $B_h^s$ and its boundedness result Lemma 8.11 that

$$\begin{aligned} B_h^s(u-\tilde{u},u-\tilde{u},v_h,w_h) &= B_h^s(u,u,v_h,w_h) + B_h^s(\tilde{u},\tilde{u},v_h,w_h) - 2B_h^s(u,\tilde{u},v_h,w_h) \\ &\ge B_h^s(u,u,v_h,w_h) + B_h^s(\tilde{u},\tilde{u},v_h,w_h) - 2C_1|||u|||_h|||\tilde{u}|||_h|||v_h|||_h|||w_h|||_h, \end{aligned}$$

where $C_1$ is the generic constant arising in the boundedness result Lemma 8.11 for $B_h^s(\cdot,\cdot,\cdot,\cdot)$. Therefore, we obtain that

$$\langle \mathcal{DN}_h^s(I_h u)v_h, w_h\rangle \ge \langle \mathcal{DN}_h^s(u)v_h, w_h\rangle+3B_h^s(\tilde{u},\tilde{u},v_h,w_h)-6C_1|||u|||_h|||\tilde{u}|||_h|||v_h|||_h|||w_h|||_h.$$

Now using the inf-sup condition Theorem 8.14 for the bilinear form $\langle \mathcal{DN}_h^s(u)\cdot,\cdot\rangle$, boundedness result Lemma 8.11 and interpolation estimates, we get

$$\begin{aligned}\sup_{\substack{|||w_h|||_h=1\\ w_h\in W_h}} \langle \mathcal{DN}_h^s(I_h u)v_h, w_h\rangle &\geq \sup_{\substack{|||w_h|||_h=1\\ w_h\in W_h}} \langle \mathcal{DN}_h^s(u)v_h, w_h\rangle - 3|B_h^s(\tilde{u},\tilde{u},v_h,w_h)| \\ &\quad - 6C_1 h^{\min\{\deg-1,\Bbbk_u-2\}}|||u|||_h|||v_h|||_h \\ &\geq \left(\beta_c - C_2 h^{\min\{\deg-1,\Bbbk_u-2\}}\right)|||v_h|||_h \\ &\geq \frac{\beta_c}{2}|||v_h|||_h,\end{aligned}$$

for a sufficiently small mesh size $h$ such that $h^{\min\{\deg-1,\Bbbk_u-2\}} < \frac{\beta_c}{2C_2}$. Here, $C_2$ depends on $C_1$ and $\|u\|_2$ and $\Bbbk_u > 2$ gives the regularity of $u$, i.e., $u \in H^{\Bbbk_u}(\Omega)$. Therefore, the inf-sup condition (8.1.2.31) holds. □

### 8.1.2.3 Convergence analysis

We proceed to the error analysis for the discrete nonlinear problem (8.1.2.4). Let

$$\mathcal{B}_\rho(I_h u) := \{v_h \in W_h : |||I_h u - v_h|||_h \leq \rho\},$$

where $I_h$ is the interpolation operator mapping from the infinite dimensional space $H^2(\mathcal{T}_h)\cap H^1(\Omega)$ to the finite dimensional space $W_h$. We define the nonlinear map $\mu_h : W_h \to W_h$ by

$$\langle \mathcal{DN}_h^s(I_h u_h)\mu_h(v_h), w_h\rangle = 3B_h^s(I_h u_h, I_h u_h, v_h, w_h) + L^s(w_h) - B_h^s(v_h,v_h,v_h,w_h). \tag{8.1.2.32}$$

Due to the weak coercivity property in Theorem 8.15, the nonlinear map $\mu_h$ is well-defined.

The existence and local uniqueness result of the discrete solution $u_h$ to the discrete nonlinear problem (8.1.2.4) will be proven via an application of Brouwer's fixed point theorem, which necessitates the use of two auxiliary lemmas illustrating that (i) $\mu_h$ maps from a ball to itself; and (ii) the map $\mu_h$ is contracting.

**Lemma 8.16.** *(Mapping from a ball to itself) Let $u$ be a regular isolated solution of the continuous nonlinear weak problem (8.1.2.2). For a sufficiently large $\epsilon$ and a sufficiently small mesh size $h$, there exists a positive constant $R(h) > 0$ such that:*

$$|||v_h - I_h u|||_h \leq R(h) \Rightarrow |||\mu_h(v_h) - I_h u|||_h \leq R(h) \quad \forall v_h \in W_{h,0}.$$

*Proof.* Note that the solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of (8.1.2.2) satisfies the discrete weak formulation (8.1.2.4) due to the consistency result Theorem 8.10, that is to say, there holds that

$$A_h^s(u, w_h) + P_h^s(u, w_h) + B_h^s(u, u, u, w_h) + C_h^s(u, w_h) = L^s(w_h) \quad \forall w_h \in W_{h,0}. \tag{8.1.2.33}$$

By the linearity of $\langle \mathcal{DN}_h^s(I_h u)\cdot, \cdot\rangle_{H^2}$, the definition (8.1.2.32) of the nonlinear map $\mu_h$ and formulation (8.1.2.33), we calculate

$$\begin{aligned}
&\langle \mathcal{DN}_h^s(I_h u)(I_h u - \mu_h(v_h)), w_h\rangle \\
&\quad = \langle \mathcal{DN}_h^s(I_h u) I_h u, w_h\rangle - \langle \mathcal{DN}_h^s(I_h u)\mu_h(v_h), w_h\rangle \\
&\quad = A_h^s(I_h u, w_h) + P_h^s(I_h u, w_h) + 3B_h^s(I_h u, I_h u, I_h u, w_h) + C_h^s(I_h u, w_h) \\
&\qquad - 3B_h^s(I_h u, I_h u, v_h, w_h) + B_h^s(v_h, v_h, v_h, w_h) - L^s(w_h) \\
&\quad = \underbrace{A_h^s(I_h u - u, w_h) + P_h^s(I_h u - u, w_h)}_{\mathfrak{N}_1} + \underbrace{C_h^s(I_h u - u, w_h)}_{\mathfrak{N}_2} \\
&\qquad + \underbrace{(B_h^s(I_h u, I_h u, I_h u, w_h) - B_h^s(u, u, u, w_h))}_{\mathfrak{N}_3} \\
&\qquad + \underbrace{(2B_h^s(I_h u, I_h u, I_h u, w_h) - 3B_h^s(I_h u, I_h u, v_h, w_h) + B_h^s(v_h, v_h, v_h, w_h))}_{\mathfrak{N}_4} \\
&\quad =: \mathfrak{N}_1 + \mathfrak{N}_2 + \mathfrak{N}_3 + \mathfrak{N}_4.
\end{aligned}$$

In what follows, we give the upper bounds for each $\mathfrak{N}_i, i = 1, 2, 3, 4$. A use of the boundedness of $A_h^s + P_h^s, C_h^s$ and the interpolation estimate [BS05, Eq. (5.3)] in the $|||\cdot|||$-norm, we obtain

$$\begin{aligned}
\mathfrak{N}_1 &\lesssim |||I_h u - u|||_h |||w_h|||_h \lesssim h^{\min\{\deg-1, \Bbbk_u - 2\}} |||w_h|||_h, \\
\mathfrak{N}_2 &\lesssim |||I_h u - u|||_h |||w_h|||_h \lesssim h^{\min\{\deg-1, \Bbbk_u - 2\}} |||w_h|||_h.
\end{aligned}$$

We rearrange terms in $\mathfrak{N}_3$ and use the boundedness result Lemma 8.11 and the interpolation result [BS05, Eq. (5.3)] to obtain

$$\begin{aligned}
\mathfrak{N}_3 &= B_h^s(I_h u, I_h u, I_h u, w_h) - B_h^s(u, u, u, w_h) \\
&= B_h^s(I_h u - u, I_h u - u, I_h u, w_h) + 2B_h^s(I_h u - u, I_h u - u, u, w_h) + 3B_h^s(u, u, I_h u - u, w_h) \\
&\lesssim \left(|||I_h u - u|||_h^2 |||I_h u|||_h + |||I_h u - u|||_h^2 |||u|||_h + \|u\|_2^2 \|I_h u - u\|_0\right) |||w_h|||_h \\
&\lesssim \left(h^{2\min\{\deg-1, \Bbbk_u - 2\}} + h^{\min\{\deg+1, \Bbbk_u\}}\right) |||w_h|||_h.
\end{aligned}$$

Let $e_I = v_h - I_h u$. We use the definition of $B_h^s(\cdot,\cdot,\cdot,\cdot)$ and use its boundedness to deduce that

$$\begin{aligned}
\mathfrak{N}_4 &= 2B_h^s(I_h u, I_h u, I_h u, w_h) - 3B_h^s(I_h u, I_h u, v_h, w_h) + B_h^s(v_h, v_h, v_h, w_h) \\
&= a_3 \int_\Omega \left\{ 2(I_h u)^3 w_h - 3(I_h u)^2 v_h w_h + v_h^3 w_h \right\} \\
&= a_3 \int_\Omega \left\{ \left( v_h^2 - (I_h u)^2 \right) v_h w_h + 2(I_h u)^2 (I_h u - v_h) w_h \right\} \\
&= a_3 \int_\Omega \left\{ e_I (e_I + 2I_h u)(e_I + I_h u) w_h - 2(I_h u)^2 e_I w_h \right\} \\
&= a_3 \int_\Omega \left\{ e_I \left( e_I^2 + 3e_I I_h u + 2(I_h u)^2 \right) w_h - 2(I_h u)^2 e_I w_h \right\} \\
&= a_3 \int_\Omega \left( e_I^3 + 3e_I^3 I_h u \right) w_h \\
&= B_h^s(e_I, e_I, e_I, w_h) + 3B_h^s(e_I, e_I, I_h u, w_h) \\
&\lesssim |||e_I|||_h^2 \left( |||e_I|||_h + |||I_h u|||_h \right) |||w_h|||_h.
\end{aligned}$$

Hence, we combine the above bounds for $\mathfrak{N}_i$, $i = 1, 2, 3, 4$ to have

$$\begin{aligned}
&\langle D\mathcal{N}_h^s(I_h u)(I_h u - \mu_h(v_h)), w_h \rangle \\
&\quad \lesssim \left( h^{\min\{\deg - 1, \Bbbk_u - 2\}} + h^{\min\{2\deg - 2, 2\Bbbk_u - 4, \deg + 1, \Bbbk_u\}} + |||e_I|||_h^2 \left( |||e_I|||_h + 1 \right) \right) |||w_h|||_h.
\end{aligned}$$

By the inf-sup condition (8.1.2.31) for the perturbed bilinear form, we further deduce that there exists a $w_h \in W_h$ with $|||w_h|||_h = 1$ such that

$$|||I_h u - \mu_h(v_h)|||_h \lesssim \langle D\mathcal{N}_h^s(I_h u)(I_h u - \mu_h(v_h)), w_h \rangle.$$

Since $|||e_I|||_h \leq R(h)$, we obtain

$$\begin{aligned}
|||I_h u - \mu_h(v_h)|||_h &\lesssim \left( h^{\min\{\deg - 1, \Bbbk_u - 2\}} + h^{\min\{2\deg - 2, 2\Bbbk_u - 4, \deg + 1, \Bbbk_u\}} + R(h)^2 \left( R(h) + 1 \right) \right) \\
&\leq \begin{cases} C_u \left( 2h^{\min\{\deg - 1, \Bbbk_u - 2\}} + R(h)^2 (1 + R(h)) \right) & \text{for } 2 \leq \deg \leq 3, \Bbbk_u \leq 4, \\ C_u \left( h^{\min\{\deg - 1, \Bbbk_u - 2\}} + h^{\min\{\deg + 1, 2\Bbbk_u - 4\}} + R(h)^2 (1 + R(h)) \right) & \text{for } \deg > 3, \Bbbk_u \leq 4. \end{cases}
\end{aligned}$$

Note that there are other cases when $\Bbbk_u > 4$ and we only focus on the case of $\Bbbk_u \leq 4$ here for brevity. Hence, the idea of the remainder of the proof is to choose an appropriate $R(h)$ so that $|||I_h u - \mu_h(v_h)|||_h \leq R(h)$. For simplicity of the calculation, we illustrate the case when $2 \leq \deg \leq 3$, $\Bbbk_u \leq 4$. To this end, we take $R(h) = 4C_u h^{\min\{\deg - 1, \Bbbk_u - 2\}}$ and choose $h$ satisfying

$$h^{2\min\{\deg - 1, \Bbbk_u - 2\}} \leq \frac{1}{32C_u} - \frac{1}{16}.$$

This yields

$$\begin{aligned}|||I_h u - \mu_h(v_h)|||_h &\leq 2C_u h^{\min\{\deg-1,\Bbbk_u-2\}}\left(1 + C_u R(h)^2 + C_u\right)\\ &= 2C_u h^{\min\{\deg-1,\Bbbk_u-2\}}\left(1 + 32C_u^3 h^{2\min\{\deg-1,\Bbbk_u-2\}} + 2C_u\right)\\ &\leq R(h).\end{aligned}$$

This completes the proof. □

**Lemma 8.17.** *(Contraction result) For a sufficiently large $\epsilon$, a sufficiently small mesh size $h$ and any $v_1, v_2 \in \mathcal{B}_{R(h)}(I_h u)$, there holds*

$$|||\mu_h(v_1) - \mu_h(v_2)|||_h \lesssim h^{\min\{\deg-1,\Bbbk_u-2\}}|||v_1 - v_2|||_h. \tag{8.1.2.34}$$

*Proof.* For $w_h \in W_h$, we use the definition (8.1.2.32) of the nonlinear map $\mu_h$, definition and linearity of $\langle \mathcal{DN}_h^s(I_h u)\cdot,\cdot\rangle$ to calculate

$$\begin{aligned}&\langle \mathcal{DN}_h^s(I_h u)(\mu_h(v_1) - \mu_h(v_2)), w_h\rangle\\ &\quad= 3B_h^s(I_h u, I_h u, v_1, w_h) - B_h^s(v_1, v_1, v_1, w_h)\\ &\qquad - 3B_h^s(I_h u, I_h u, v_2, w_h) + B_h^s(v_2, v_2, v_2, w_h)\\ &\quad= a_3\int_\Omega \left(3(I_h u)^2 v_1 w_h - v_1^3 w_h\right) - a_3\int_\Omega \left(3(I_h u)^2 v_2 w_h - v_2^3 w_h\right)\\ &\quad= a_3\int_\Omega \left(\left((I_h u)^2 - v_1^2\right)v_1 w_h + 2(I_h u)^2(v_1 - v_2)w_h - \left((I_h u)^2 - v_2^2\right)v_2 w_h\right)\\ &\quad= a_3\int_\Omega ((I_h u - v_1)(v_1 - I_h u)(v_1 - v_2)w_h + 2(I_h u - v_1)I_h u(v_1 - v_2)w_h\\ &\qquad\qquad + (I_h u - v_1)(I_h u + v_1)v_2 w_h)\\ &\qquad + 2a_3\int_\Omega (I_h u(v_1 - v_2)(I_h u - v_2)w_h + I_h u(v_1 - v_2)v_2 w_h)\\ &\qquad - a_3\int_\Omega (I_h u - v_2)(I_h u + v_2)v_2 w_h\\ &\quad= a_3\int_\Omega (I_h u - v_1)(v_1 - I_h u)(v_1 - v_2)w_h + 2a_3\int_\Omega (I_h u - v_1)I_h u(v_1 - v_2)w_h\\ &\qquad 2a_3\int_\Omega (I_h u - v_2)I_h u(v_1 - v_2)w_h\\ &\qquad + a_3\int_\Omega (v_1 - v_2)\left((I_h u - v_1) + (I_h u - v_2)\right)\left((v_2 - I_h u) + I_h u\right)w_h.\end{aligned}$$

Let $e_1 = I_h u - v_1$, $e_2 = I_h u - v_2$ and $e = v_1 - v_2$. We make some elementary manipulations and use the boundedness result of $B_h^s$ and the inequality of geometric

and arithmetic means to get

$$
\begin{aligned}
&\langle \mathcal{DN}_h^s(I_h u)(\mu_h(v_1)-\mu_h(v_2)), w_h\rangle \\
&\quad = a_3\int_\Omega(-e_1^2)ew_h + 2a_3\int_\Omega e_1(I_hu)ew_h + 2a_3\int_\Omega e_2(I_hu)ew_h \\
&\quad + a_3\int_\Omega\{ew_h(e_1I_hu + e_2I_hu - e_1e_2 - e_2^2)\} \\
&\quad \lesssim \left(|||e_1|||_h^2 + |||I_hu|||_h|||e_1|||_h + |||e_2|||_h|||I_hu|||_h + |||e_1|||_h|||e_2|||_h + |||e_2|||_h^2\right)|||e|||_h|||w_h|||_h \\
&\quad \lesssim \left(|||e_1|||_h^2 + |||e_2|||_h^2 + |||e_1|||_h + |||e_2|||_h\right)|||e|||_h|||w_h|||_h \\
&\quad \lesssim \left(R(h)^2 + R(h)\right)|||e|||_h|||w_h|||_h.
\end{aligned}
$$

By the inf-sup condition (8.1.2.31), we know that there exist $w_h \in W_h$ with $|||w_h|||_h = 1$ such that

$$
\frac{\beta_c}{2}|||\mu_h(v_1)-\mu_h(v_2)|||_h \lesssim \langle \mathcal{DN}_h^s(I_hu)(\mu_h(v_1)-\mu_h(v_2)), w_h\rangle.
$$

Therefore, we have

$$
|||\mu_h(v_1)-\mu_h(v_2)|||_h \lesssim R(h)(1+R(h))|||e|||_h.
$$

Note that $R(h)(1+R(h)) < 1$ for $R(h) < 1$. This completes the proof. □

The existence and local uniqueness of the discrete solution $u_h$ can now be obtained via the application of Brouwer's fixed point theorem [Kes89].

**Theorem 8.18.** *(Convergence in $|||\cdot|||_h$-norm) Let $u$ be a regular isolated solution of the nonlinear problem (8.1.2.2). For a sufficiently large $\epsilon$ and a sufficiently small $h$, there exists a unique solution $u_h$ of the discrete nonlinear problem (8.1.2.4) within the local ball $\mathcal{B}_{R(h)}(I_hu)$. Furthermore, we have the following bound:*

$$
|||u-u_h|||_h \lesssim h^{\min\{\deg -1, \Bbbk_u - 2\}},
$$

*where $\deg \geq 1$ denotes the degree of the polynomial approximation and $\Bbbk_u \geq 2$ is the regularity index of $u$.*

*Proof.* A use of Lemma 8.16 yields that the nonlinear map $\mu_h$ maps a closed convex set $\mathcal{B}_{R(h)}(I_hu) \subset W_h$ to itself. Moreover it is a contracting map. Therefore, an application of the Brouwer fixed point theorem [Kes89] yields that $\mu_h$ has at least

one fixed point, say $u_h$, in this ball $\mathcal{B}_{R(h)}(I_h u)$. The uniqueness of the solution to (8.1.2.4) in that ball $\mathcal{B}_{R(h)}(I_h u)$ follows from the contraction result in Lemma 8.17. Meanwhile, we have by Lemma 8.16 that

$$|||u_h - I_h u|||_h \lesssim h^{\min\{\deg -1, \Bbbk_u - 2\}}. \tag{8.1.2.35}$$

The error estimate is then obtained straightforwardly using the triangle inequality

$$|||u - u_h|||_h \leq |||u - I_h u|||_h + |||I_h u - u_h|||_h,$$

combined with (8.1.2.35) and the interpolation estimate [BS05, Eq. (5.3)]. □

It is implied from Theorem 8.18 that optimal convergence rates have been shown in the mesh-dependent norm $|||\cdot|||_h$. We will see the numerical verifications of this in Section 8.2.

#### 8.1.2.4 Estimates in the $L^2$-norm

We derive an $L^2$ error estimate using a duality argument in this subsection. To this end, we consider the following linear dual problem to the primary nonlinear problem (8.1.0.3):

$$\begin{cases} 2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2\chi)) + a_1\chi + 3a_3u^2\chi = f_{dual} & \text{in } \Omega, \\ \chi = 0 & \text{on } \partial\Omega, \\ \nu\cdot\mathcal{D}^2\chi = \mathbf{0} & \text{on } \partial\Omega, \end{cases} \tag{8.1.2.36}$$

for $f_{dual} \in L^2(\Omega)$. For smooth domains $\Omega$, it can be deduced by a classical elliptic regularity result that $\chi \in H^4(\Omega)$. The corresponding weak form is derived: find $\chi \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$2B\int_\Omega \mathcal{D}^2\chi : \mathcal{D}^2 v + a_1\int_\Omega \chi v + 3a_3\int_\Omega u^2\chi v = \int_\Omega f_{dual} v \quad \forall v \in H^2(\Omega)\cap H_0^1(\Omega),$$

that is to say,

$$\langle \mathcal{DN}^s(u)\chi, v\rangle_{H^2} = \langle \mathcal{DN}_h^s(u)\chi, v\rangle = (f_{dual}, v)_0. \tag{8.1.2.37}$$

**Remark 8.9.** *The first equality in (8.1.2.37) holds since $u \in H^2(\Omega), \chi \in H^2(\Omega)$ and $v \in H^2(\Omega)$.*

We give two auxiliary results in the following.

**Lemma 8.19.** *For $u \in H^{\mathbb{k}_u}(\Omega)$, $\mathbb{k}_u > 2$, $\chi \in H^4(\Omega) \cap H_0^1(\Omega)$ and $I_h u \in W_{h,0} \subset H_0^1(\Omega)$, there holds that*

$$A_h^s(I_h u - u, \chi) + P_h^s(I_h u - u, \chi) \lesssim h^{\min\{\deg+1, \mathbb{k}_u\}} \|\chi\|_4.$$

*Proof.* Note that $[\![\nabla \chi]\!] = 0$ since $\chi \in H^4(\Omega)$ and $\chi = 0$ on $\partial\Omega$. We calculate

$$\begin{aligned}
&A^s(I_h u - u, \chi) + P_h^s(I_h u - u, \chi) \\
&\quad = \sum_{T \in \mathcal{E}_h} \int_T 2B\mathcal{D}^2(I_h u - u) \colon \mathcal{D}^2 \chi \\
&\qquad - 2B \sum_{e \in \mathcal{E}_I} \left\{\!\!\left\{ \frac{\partial^2 (I_h u - u)}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla \chi]\!] - 2B \sum_{e \in \mathcal{E}_I} \left\{\!\!\left\{ \frac{\partial^2 \chi}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla (I_h u - u)]\!] \\
&\qquad + \sum_{e \in \mathcal{E}_I} \frac{2B\epsilon}{h_e^3} \int_e [\![\nabla (I_h u - u)]\!] [\![\nabla \chi]\!] \\
&\quad = \sum_{T \in \mathcal{E}_h} \int_T 2B\mathcal{D}^2(I_h u - u) \colon \mathcal{D}^2 \chi - 2B \sum_{e \in \mathcal{E}_I} \left\{\!\!\left\{ \frac{\partial^2 \chi}{\partial \nu^2} \right\}\!\!\right\} [\![\nabla (I_h u - u)]\!] \\
&\quad = \sum_{T \in \mathcal{E}_h} \int_T 2B(I_h u - u) \nabla \cdot (\nabla \cdot (\mathcal{D}^2 \chi)) \\
&\quad \lesssim \|I_h u - u\|_0 \|\nabla \cdot (\nabla \cdot (\mathcal{D}^2 \chi))\|_0 \\
&\quad \lesssim h^{\min\{\deg+1, \mathbb{k}_u\}} \|\chi\|_4.
\end{aligned}$$

Here, the last, second last, third last steps follow from the standard interpolation estimates, the Cauchy–Schwarz inequality, and integration by parts twice, respectively. □

**Lemma 8.20.** *The solution $\chi$ of the linear dual problem (8.1.2.36) belongs to $H^4(\Omega)$ on a smooth domain $\Omega$ and it holds that*

$$\|\chi\|_4 \lesssim \|f_{dual}\|_0. \tag{8.1.2.38}$$

*Proof.* We can use the inf-sup condition (8.1.2.11) for the linear operator $\langle \mathcal{DN}^s(u)\cdot, \cdot \rangle$, the weak form (8.1.2.37) and the Cauchy–Schwarz inequality to obtain

$$|\chi|_2 \lesssim \sup_{\substack{w \in H^2 \cap H_0^1 \\ |w|_2 = 1}} \langle \mathcal{DN}^s(u)\chi, w \rangle_{H^2} = \sup_{\substack{w \in H^2 \cap H_0^1 \\ |w|_2 = 1}} (f_{dual}, w)_0 \lesssim \|f\|_0 \underbrace{\|w\|_0}_{\lesssim |w|_2 = 1}. \tag{8.1.2.39}$$

By the form of (8.1.2.37) and the boundedness of $B^s(u,u,\cdot,\cdot)$ and $C^s(\cdot,\cdot)$, we have

$$\begin{aligned}\|2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2\chi))\|_0 &= \|-3B^s(u,u,\chi,\cdot)-C^s(\chi,\cdot)+(f_{dual},\cdot)_0\|_0\\ &\lesssim \underbrace{\|\chi\|_0}_{\lesssim|\chi|_2}+\|f_{dual}\|_0\\ &\lesssim \|f_{dual}\|_0 \qquad \text{by (8.1.2.39).}\end{aligned} \tag{8.1.2.40}$$

Using a bootstrapping argument in elliptic regularity (see, e.g., [Eva10, Section 6.3]), we can deduce that $\chi\in H^4(\Omega)$ in a smooth domain $\Omega$. Moreover, it is implied from (8.1.2.40) that the regularity estimate (8.1.2.38) holds. □

We are ready to derive the $L^2$ a priori error estimates.

**Theorem 8.21.** *($L^2$ error estimate) Under the same conditions as in Theorem 8.18 and assuming further that* $\deg\geq 1, \Bbbk_u\geq 2$*, the discrete solution $u_h$ approximates $u$ such that*

$$\|u-u_h\|_0 \lesssim \begin{cases} h^{\min\{\deg+1,\Bbbk_u\}} & \text{for } \deg\geq 3,\\ h^{2\min\{\deg-1,\Bbbk_u-2\}} & \text{for } \deg=2.\end{cases}$$

*Proof.* Taking $f_{dual}=I_hu-u_h\in W_h\subset H^1(\Omega)\cap H^2(\mathcal{T}_h)$ in (8.1.2.36) and multiplying (8.1.2.36) by a test function $v_h=I_hu-u_h$ with integration by parts, we obtain

$$\langle\mathcal{DN}_h^s(u)\chi, I_hu-u_h\rangle = \|I_hu-u_h\|_0^2.$$

It follows from the fact that $u\in H^{\Bbbk_u}(\Omega)$, $\Bbbk_u\geq 2$, and the definition (8.1.2.2) of the nonlinear continuous weak form $\mathcal{N}^s(u)\cdot$ that

$$\begin{aligned}\|I_hu-u_h\|_0^2 &= \langle\mathcal{DN}_h^s(u)\chi, I_hu-u_h\rangle+\mathcal{N}_h^s(u_h)(I_h\chi)-\mathcal{N}_h^s(u)(I_h\chi)\\ &= A_h^s(\chi,I_hu-u_h)+P_h^s(\chi,I_hu-u_h)+C_h^s(\chi,I_hu-u_h)+3B_h^s(u,u,\chi,I_hu-u_h)\\ &\quad+A_h^s(u_h,I_h\chi)+P_h^s(u_h,I_h\chi)+C_h^s(u_h,I_h\chi)+B_h^s(u_h,u_h,u_h,I_h\chi)\\ &\quad-A_h^s(u,I_h\chi)-P_h^s(u,I_h\chi)-C_h^s(u,I_h\chi)-B_h^s(u,u,u,I_h\chi)\\ &= \underbrace{A_h^s(I_hu-u,\chi)+A_h^s(u-u_h,\chi-I_h\chi)+P_h^s(I_hu-u,\chi)+P_h^s(u-u_h,\chi-I_h\chi)}_{\mathfrak{U}_1}\\ &\quad+\underbrace{C_h^s(I_hu-u,\chi)+C_h^s(u-u_h,\chi-I_h\chi)}_{\mathfrak{U}_2}\\ &\quad+\underbrace{3B_h^s(u,u,I_hu-u_h,\chi-I_h\chi)+3B_h^s(u,u,I_hu-u,I_h\chi)}_{\mathfrak{U}_3}\end{aligned}$$

$$
\begin{aligned}
&\quad + \underbrace{B_h^s(u_h,u_h,u_h,I_h\chi) - 3B_h^s(u,u,u_h,I_h\chi) + 2B_h^s(u,u,u,I_h\chi)}_{\mathfrak{U}_4} \\
&=: \mathfrak{U}_1 + \mathfrak{U}_2 + \mathfrak{U}_3 + \mathfrak{U}_4.
\end{aligned}
$$

We then bound each $\mathfrak{U}_i$ separately using the boundedness results for $A_h^s, P_h^s, B_h^s$ and $C_h^s$ and standard interpolation estimates. This leads to

$$
\begin{aligned}
\mathfrak{U}_1 &\lesssim h^{\min\{\deg+1,\Bbbk_u\}}\|\chi\|_4 + \underbrace{|||u-u_h|||_h}_{\lesssim h^{\min\{\deg-1,\Bbbk_u-2\}}} \underbrace{|||\chi - I_h\chi|||_h}_{\lesssim h^2\|\chi\|_4} \qquad \text{by Theorem 8.18,} \\
&\lesssim h^{\min\{\deg+1,\Bbbk_u\}}\|\chi\|_4,
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{U}_2 &\lesssim \underbrace{\|I_hu-u\|_0}_{\lesssim h^{\min\{\deg+1,\Bbbk_u\}}} \underbrace{\|\chi\|_0}_{\leq\|\chi\|_4} + |||u-u_h|||_h|||\chi - I_h\chi|||_h \\
&\lesssim h^{\min\{\deg+1,\Bbbk_u\}}\|\chi\|_4,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathfrak{U}_3 &= 3B_h^s(u,u,I_hu-u_h,\chi-I_h\chi) + 3B_h^s(u,u,I_hu-u,I_h\chi) \\
&\lesssim \|u\|_2^2 \underbrace{|||I_hu-u_h|||_h}_{\lesssim h^{\min\{\deg-1,\Bbbk_u-2\}}} \underbrace{|||\chi-I_h\chi|||_h}_{\lesssim h^2\|\chi\|_4} + \|u\|_2^2 \underbrace{\|I_hu-u\|_0}_{\lesssim h^{\min\{\deg+1,\Bbbk_u\}}} \underbrace{\|I_h\chi\|_0}_{\lesssim\|\chi\|_4} \\
&\lesssim h^{\min\{\deg+1,\Bbbk_u\}}\|\chi\|_4.
\end{aligned}
$$

Setting $e_3 = u_h - u$ and estimating $\mathfrak{U}_4$ as in $\mathfrak{R}_4$ of Lemma 8.16 yield

$$
\begin{aligned}
\mathfrak{U}_4 &\lesssim |||e_3|||_h^2\,(|||e_3|||_h + |||u|||_h) \underbrace{|||I_h\chi|||_h}_{\lesssim\|\chi\|_2\leq\|\chi\|_4} \\
&\lesssim h^{2\min\{\deg-1,\Bbbk_u-2\}}(h^{\min\{\deg-1,\Bbbk_u-2\}}+1)\|\chi\|_4 \qquad \text{by Theorem 8.18.}
\end{aligned}
$$

Combining the above estimates for $\mathfrak{U}_i$ $(i = 1,2,3,4)$ and using the regularity estimate (8.1.2.38), we obtain

$$
\|I_hu-u_h\|_0^2 \lesssim \begin{cases} h^{\min\{\deg+1,\Bbbk_u\}} \underbrace{\|\chi\|_4}_{\lesssim\|I_hu-u_h\|_0} & \text{if } \deg\geq 3, \Bbbk_u\geq 4, \\ h^{2\min\{\deg-1,\Bbbk_u-2\}} \underbrace{\|\chi\|_4}_{\lesssim\|I_hu-u_h\|_0} & \text{if } \deg = 2, \Bbbk_h\leq 4. \end{cases}
$$

Using the triangle inequality and standard interpolation estimates, we get

$$
\|u-u_h\|_0 \leq \|u-I_hu\|_0 + \|I_hu-u_h\|_0
$$

$$\lesssim h^{\min\{\deg+1,\Bbbk_u\}} + \begin{cases} h^{\min\{\deg+1,\Bbbk_u\}} & \text{for } \deg \geq 3, \\ h^{2\min\{\deg-1,\Bbbk_u-2\}} & \text{for } \deg = 2, \end{cases}$$

$$\lesssim \begin{cases} h^{\min\{\deg+1,\Bbbk_u\}} & \text{for } \deg \geq 3, \\ h^{2\min\{\deg-1,\Bbbk_u-2\}} & \text{for } \deg = 2. \end{cases}$$

This complies the proof. □

Theorem 8.21 implies that for quadratic approximations to the sufficiently regular solution of (8.1.2.1), there is a sub-optimal convergence rate in the $L^2$-norm while for higher order ($\geq 3$) approximations, we expect optimal $L^2$ error rates. We shall see numerical verifications of this in the subsequent sections.

#### 8.1.2.5 The inconsistent discrete form

The above analysis considers the consistent weak formulation (8.1.2.4) in finite element discretisations. In practice, we adopt the inconsistent discrete weak form in the implementations in Chapter 9 because of its simplicity in the discrete weak form: find $u_h \in W_{h,b}$ such that

$$\tilde{\mathcal{N}}_h^s(u_h)v_h = \tilde{A}_h^s(u_h,v_h)+B_h^s(u_h,u_h,u_h,v_h)+C_h^s(u_h,v_h)+P_h^s(u_h,v_h) = 0 \quad \forall v_h \in W_{h,0}, \tag{8.1.2.41}$$

where

$$\tilde{A}_h^s(u,v) := 2B \sum_{T\in\mathcal{T}_h} \int_T \mathcal{D}^2 u \colon \mathcal{D}^2 v.$$

Note that the missing terms by comparing $\tilde{A}_h^s$ and $A_h^s$ are those interelement summations involving the average of the second tangential derivatives, arising from piecewise integration by parts and the symmetrisation. Due to the absence of those terms in $\tilde{A}_h^s$, one can immediately notice that the discrete weak formulation (8.1.2.41) is inconsistent in the sense that the solution $u$ of the strong form (8.1.2.1) does not satisfy the weak form (8.1.2.41), as opposed to Theorem 8.10.

Regardless of this inconsistency that complicates the convergence analysis, our choice of the discrete weak form (8.1.2.4) reduces the complexity of the implementation and in practice leads to a converging numerical scheme (though may not possess optimal convergence rates), as illustrated in Section 8.2. This

is not surprising; a similar idea has also been applied and introduced as *weakly over-penalised symmetric interior penalty* (WOPSIP) methods in [BS08b] for second order elliptic PDEs and in [BGS10] for biharmonic equations.

**Remark 8.10.** *The excessive size of the penalty parameter in the WOPSIP method could induce ill-conditioned linear systems. It is also discussed in [BS08b] how to design block preconditioners and analyse the conditioning of the linear systems. Moreover, in all of our numerical experiments in the next section, we do not observe any ill-conditioning effects.*

In our numerical examinations of the convergence rate for the inconsistent discrete weak form (8.1.2.41), we find that the inconsistency does not substantially alter the convergence rate proved for the consistent form. Thus, the inconsistent formulation (8.1.2.41) can be a viable choice in implementations.

## 8.2 Convergence tests

The proceeding section presents some a priori error estimates for the continuous Lagrange finite elements for both $\mathbf{Q}$ and $u$ in the decoupled case $q = 0$. We now test the convergence rate of the finite element approximations by the method of manufactured solutions (MMS) and experimentally investigate the coupled case $q \neq 0$ in two dimensions. To this end, we choose a nontrivial solution for each state variable and add an appropriate source term to the equilibrium equations (see Appendix A for its derivation), thus modifying the energy accordingly. Therefore, our chosen solution should solve the equilibrium equations exactly when we take a suitable initial guess and we can compute the numerical convergence order.

**Remark 8.11.** *Since this is purely a numerical verification exercise, the manufactured solution can be physically unrealistic. Moreover, we must specify a reasonable initial guess for Newton's iteration due to the nonlinearity of the problem. The initial guess throughout this section is taken to be* $\left(\frac{1}{2}(\textit{exact solution}) + 10^{-9}\right)$.

We take the following exact expressions for each state variable,

$$\begin{aligned} Q_{11}^e &= \left(\cos\left(\frac{\pi(2y-1)(2x-1)}{8}\right)\right)^2 - \frac{1}{2}, \\ Q_{12}^e &= \cos\left(\frac{\pi(2y-1)(2x-1)}{8}\right)\sin\left(\frac{\pi(2y-1)(2x-1)}{8}\right), \\ u^e &= 10\left((x-1)x(y-1)y\right)^3, \end{aligned} \tag{8.2.0.1}$$

and substitute them into the derived equilibrium equations (A.0.0.2) to obtain the source terms $\mathfrak{s}_1, \mathfrak{s}_2$ and $\mathfrak{s}_3$, yielding

$$\begin{aligned} \mathfrak{s}_1 \coloneqq\ & 4Bq^4(u^e)^2 q_1^e + 2Bq^2 u^e\left(\partial_x^2 u^e - \partial_y^2 u^e\right) - 2K\Delta Q_{11}^e \\ & - 4lQ_{11}^e + 16lq_1^e\left((Q_{11}^e)^2 + (q_2^e)^2\right), \\ \mathfrak{s}_2 \coloneqq\ & 4Bq^4(u^e)^2 q_2^e + 4Bq^2 u^e\left(\partial_x\partial_y u^e\right) - 2K\Delta Q_{12}^e \\ & - 4lQ_{12}^e + 16lq_2^e\left((Q_{11}^e)^2 + (Q_{12}^e)^2\right), \\ \mathfrak{s}_3 \coloneqq\ & a_1 u^e + a_2(u^e)^2 + a_3(u^e)^3 + 2B\Delta^2 u^e \\ & + Bq^4\left(4\left((Q_{11}^e)^2 + (Q_{12}^e)^2\right) + 1\right)u^e + 2Bq^2(t_1^e + t_2^e), \end{aligned}$$

with

$$\begin{aligned} t_1^e &\coloneqq (Q_{11}^e + 1/2)\partial_x^2 u^e + (-Q_{11}^e + 1/2)\partial_y^2 u^e + 2Q_{12}^e\partial_x\partial_y u^e, \\ t_2^e &\coloneqq \partial_x^2\left(u^e\left(Q_{11}^e + 1/2\right)\right) + \partial_y^2(u^e(-Q_{11}^e + 1/2)) + 2\partial_x\partial_y(u^e Q_{12}^e). \end{aligned}$$

We take $t_1$ and $t_2$ when replacing the exact expressions of $Q_{11}^e, Q_{12}^e, u^e$ by the unknowns $Q_{11}, Q_{12}, u$.

Therefore, in conducting the MMS, we are to solve the following governing equations

$$\begin{aligned} & 4Bq^4u^2Q_{11} + 2Bq^2u\left(\partial_x^2 u - \partial_y^2 u\right) - 2K\Delta Q_{11} - 4lQ_{11} + 16lQ_{11}\left(Q_{11}^2 + Q_{12}^2\right) = \mathfrak{s}_1, \\ & 4Bq^4u^2Q_{12} + 4Bq^2u\left(\partial_x\partial_y u\right) - 2K\Delta Q_{12} - 4lQ_{12} + 16lQ_{12}\left(Q_{11}^2 + Q_{12}^2\right) = \mathfrak{s}_2, \\ & a_1u + a_2u^2 + a_3u^3 + 2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2 u)) + Bq^4\left(4\left(Q_{11}^2 + Q_{12}^2\right) + 1\right)u + 2Bq^2(t_1 + t_2) = \mathfrak{s}_3, \end{aligned} \tag{8.2.0.2}$$

subject to Dirichlet boundary conditions for both $u$ and $Q$ and a natural boundary condition for $u$ arising from the manufactured solutions (8.2.0.1).

We partition the domain into $N \times N$ small squares with the uniform mesh size $h = \frac{1}{N}$ ($N = 6$, 12, 24, 48) and denote numerical solutions $u_h$, $Q_{11,h}$ and $Q_{12,h}$. The

numerical errors of $u$ and $\mathbf{Q}$ in the $\|\cdot\|_0$-, $\|\cdot\|_1$- and $|||\cdot|||_h$-norms are defined as

$$\|\mathbf{e}_u\|_0 = \|u^e - u_h\|_0, \quad \|\mathbf{e}_u\|_1 = \|u^e - u_h\|_1, \quad |||\mathbf{e}_u|||_h = |||u^e - u_h|||_h,$$
$$\|\mathbf{e}_\mathbf{Q}\|_0 = \|(Q^e_{11}, Q^e_{12}) - (Q_{11,h}, Q_{12,h})\|_0, \quad \|\mathbf{e}_\mathbf{Q}\|_1 = \|(Q^e_{11}, Q^e_{12}) - (Q_{11,h}, Q_{12,h})\|_1.$$

The convergence order is then calculated from the formula

$$\log_2\left(\frac{\text{error}_{h/2}}{\text{error}_h}\right).$$

Throughout this section, we use the parameter values

$$a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-5},\ K = 0.3 \text{ and } l = 30,$$

yielding a similar choice as in the simulations of oily streaks in Section 9.4.

### 8.2.1 Convergence rate for $q = 0$

In the case of $q = 0$, we essentially solve two independent nonlinear problems: one second order PDE for the tensor order parameter $\mathbf{Q}$ and a fourth order PDE for the density variation $u$. Therefore, we present the convergence results for $\mathbf{Q}$ and $u$ separately in this subsection to verify the a priori error estimates proven in Section 8.1.

For the tensor variable $\mathbf{Q}$, we expect both optimal $H^1$ and $L^2$ rates, as illustrated in Theorems 8.2 and 8.8. Table 8.1 presents the numerical convergence rate for the finite elements $[\mathbb{Q}_1]^2$, $[\mathbb{Q}_2]^2$ and $[\mathbb{Q}_3]^2$. It is clear to see that optimal $L^2$ and $H^1$ rates are shown with all choices of finite elements. More specifically, second order in $L^2$ and first order in $H^1$ are observed for the approximation $[\mathbb{Q}_1]^2$. This is consistent with the proven error estimates in Section 8.1.1.

Regarding the density variation $u$, we first present the convergence behaviour of the consistent discrete formulation (8.1.2.4) with penalty parameter $\epsilon = 1$, since we have proven the optimal error rate in the mesh-dependent norm $|||\cdot|||_h$. The errors and convergence orders are listed in Table 8.2. Optimal rates are observed in the $|||\cdot|||_h$-norm. Furthermore, optimal orders of convergence in the $L^2$-norm

| | $N = \frac{1}{h}$ | $\|\mathbf{e_Q}\|_0$ | rate | $\|\mathbf{e_Q}\|_1$ | rate |
|---|---|---|---|---|---|
| $[\mathbb{Q}_1]^2$ | 6 | 8.12e-04 | – | 3.78e-02 | – |
| | 12 | 2.02e-04 | 2.01 | 1.88e-02 | 1.01 |
| | 24 | 5.05e-05 | 2.00 | 9.39e-03 | 1.00 |
| | 48 | 1.26e-05 | 2.00 | 4.69e-03 | 1.00 |
| $[\mathbb{Q}_2]^2$ | 6 | 2.92e-05 | – | 1.11e-03 | – |
| | 12 | 3.90e-06 | 2.90 | 2.71e-04 | 2.04 |
| | 24 | 5.02e-07 | 2.96 | 6.72e-05 | 2.01 |
| | 48 | 6.36e-08 | 2.99 | 1.68e-05 | 2.00 |
| $[\mathbb{Q}_3]^2$ | 6 | 3.02e-07 | – | 2.25e-05 | – |
| | 12 | 2.17e-08 | 3.80 | 2.72e-06 | 3.05 |
| | 24 | 1.45e-09 | 3.90 | 3.34e-07 | 3.03 |
| | 48 | 9.33e-11 | 3.96 | 4.13e-08 | 3.01 |

**Table 8.1:** The convergence rate of $\mathbf{Q}$ with different degrees of polynomial approximation in the decoupled case $q = 0$.

are shown for approximating polynomials of degree greater than 2, while a sub-optimal rate in the $L^2$-norm is given for piecewise quadratic polynomials, exactly as expected. The theoretical a priori error estimates are indeed verified. Sub-optimal convergence rates for quadratic polynomials were also illustrated in the numerical results of [SM07]. We also tested the convergence with the penalty parameter $\epsilon = 5 \times 10^4$ and found that the discrete norms are very similar to Table 8.2. We therefore avoid repeating the details here.

We next give the error rates for the inconsistent discrete formulation (8.1.2.41) which is actually used in the applications in Chapter 9. Though the analysis is not given for such discretisation, we wish to demonstrate that it is still convergent. We illustrate the discrete norms and the computed convergence rates in Table 8.3 with the penalty parameter $\epsilon = 1$. It can be observed that only first order of convergence is obtained in the $H^2$-like norm $|||\cdot|||_h$ even with different approximating polynomials. Moreover, we notice by comparing Tables 8.2 and 8.3 that the convergence rate deteriorates slightly for polynomials of degree 3 (although not for degree 4). This, however, can be improved by choosing a larger penalty parameter, as shown in Table 8.4 with $\epsilon = 5 \times 10^4$, where optimal rates are shown for the discrete norms

| | $N = \frac{1}{h}$ | $\|\mathbf{e}_u\|_0$ | rate | $\|\mathbf{e}_u\|_1$ | rate | $\|\|\|\mathbf{e}_u\|\|\|_h$ | rate |
|---|---|---|---|---|---|---|---|
| $\mathbb{Q}_2$ | 6 | 1.17e-05 | – | 3.46e-04 | – | 1.36e-02 | – |
| | 12 | 2.60e-06 | 2.17 | 9.81e-05 | 1.82 | 7.25e-03 | 0.91 |
| | 24 | 6.37e-07 | 2.03 | 2.54e-05 | 1.95 | 3.54e-03 | 1.03 |
| | 48 | 1.82e-07 | 1.80 | 6.88e-06 | 1.88 | 1.76e-03 | 1.01 |
| $\mathbb{Q}_3$ | 6 | 4.73e-06 | – | 1.32e-04 | – | 4.98e-03 | – |
| | 12 | 3.32e-07 | 3.83 | 1.41e-05 | 3.23 | 9.96e-04 | 2.32 |
| | 24 | 2.12e-08 | 3.97 | 1.63e-06 | 3.12 | 2.46e-04 | 2.02 |
| | 48 | 1.32e-09 | 4.00 | 1.99e-07 | 3.03 | 6.14e-05 | 2.00 |
| $\mathbb{Q}_4$ | 6 | 2.01e-07 | – | 7.76e-06 | – | 3.94e-04 | – |
| | 12 | 5.40e-09 | 5.22 | 4.30e-07 | 4.17 | 4.88e-05 | 3.01 |
| | 24 | 1.68e-10 | 5.00 | 2.68e-08 | 4.00 | 6.11e-06 | 2.99 |
| | 48 | 5.27e-12 | 4.99 | 1.68e-09 | 3.99 | 7.64e-07 | 3.00 |

**Table 8.2:** Convergence rates using the consistent discrete formulation (8.1.2.4) with the penalty parameter $\epsilon = 1$ and different polynomial degrees.

$|||\cdot|||_h$, $\|\cdot\|_1$ and $\|\cdot\|_0$ for all polynomial degrees (except only sub-optimal in $\|\cdot\|_0$ when a piecewise quadratic polynomial is used as the approximation).

| | $N = \frac{1}{h}$ | $\|\mathbf{e}_u\|_0$ | rate | $\|\mathbf{e}_u\|_1$ | rate | $\|\|\|\mathbf{e}_u\|\|\|_h$ | rate |
|---|---|---|---|---|---|---|---|
| $\mathbb{Q}_2$ | 6 | 3.50e-06 | – | 1.06e-04 | – | 5.60e-03 | – |
| | 12 | 8.76e-08 | 5.32 | 5.41e-06 | 4.29 | 2.56e-03 | 1.13 |
| | 24 | 1.77e-08 | 2.31 | 7.47e-07 | 2.86 | 1.28e-03 | 0.99 |
| | 48 | 4.35e-09 | 2.02 | 1.24e-07 | 2.56 | 6.42e-04 | 1.00 |
| $\mathbb{Q}_3$ | 6 | 6.47e-06 | – | 1.86e-04 | – | 7.59e-03 | – |
| | 12 | 3.40e-07 | 4.25 | 1.73e-05 | 3.43 | 2.74e-03 | 1.47 |
| | 24 | 1.98e-08 | 4.10 | 2.03e-06 | 3.09 | 1.31e-03 | 1.07 |
| | 48 | 3.73e-09 | 2.39 | 2.63e-07 | 2.95 | 6.45e-04 | 1.02 |
| $\mathbb{Q}_4$ | 6 | 2.05e-07 | – | 7.85e-06 | – | 3.93e-04 | – |
| | 12 | 5.40e-09 | 5.24 | 4.31e-07 | 4.19 | 4.88e-05 | 3.01 |
| | 24 | 1.68e-10 | 5.00 | 2.68e-08 | 4.01 | 6.11e-06 | 3.00 |
| | 48 | 5.27e-12 | 5.00 | 1.67e-09 | 4.00 | 7.64e-07 | 3.00 |

**Table 8.3:** Convergence rates using the inconsistent discrete formulation (8.1.2.41) with the penalty parameter $\epsilon = 1$ and different polynomial degrees.

### 8.2.2 Convergence rate for $q \neq 0$

We next investigate the numerical convergence behaviour in the coupled case, i.e., $q \neq 0$, in this subsection. Its analysis is left for future work, but since it is the

| | $N = \frac{1}{h}$ | $\|\mathbf{e}_u\|_0$ | rate | $\|\mathbf{e}_u\|_1$ | rate | $\|\|\|\mathbf{e}_u\|\|\|_h$ | rate |
|---|---|---|---|---|---|---|---|
| $\mathbb{Q}_2$ | 6 | 1.17e-05 | – | 3.48e-04 | – | 1.36e-02 | – |
| | 12 | 2.62e-06 | 2.16 | 9.86e-05 | 1.82 | 7.26e-03 | 0.91 |
| | 24 | 6.38e-07 | 2.04 | 2.54e-05 | 1.96 | 3.54e-03 | 1.03 |
| | 48 | 1.82e-07 | 1.81 | 6.88e-06 | 1.88 | 1.76e-03 | 1.01 |
| $\mathbb{Q}_3$ | 6 | 4.80e-06 | – | 1.35e-04 | – | 4.92e-03 | – |
| | 12 | 3.35e-07 | 3.84 | 1.43e-05 | 3.23 | 9.86e-04 | 2.32 |
| | 24 | 2.14e-08 | 3.97 | 1.63e-06 | 3.13 | 2.45e-04 | 2.01 |
| | 48 | 1.33e-09 | 4.01 | 1.99e-07 | 3.04 | 6.13e-05 | 2.00 |
| $\mathbb{Q}_4$ | 6 | 2.05e-07 | – | 7.85e-06 | – | 3.93e-04 | – |
| | 12 | 5.40e-09 | 5.24 | 4.31e-07 | 4.19 | 4.88e-05 | 3.01 |
| | 24 | 1.68e-10 | 5.00 | 2.68e-08 | 4.01 | 6.11e-06 | 3.00 |
| | 48 | 5.27e-12 | 5.00 | 1.67e-09 | 4.00 | 7.64e-07 | 3.00 |

**Table 8.4:** Convergence rates using the inconsistent discrete formulation (8.1.2.41) with the penalty parameter $\epsilon = 5 \times 10^4$ and different polynomial degrees.

coupled case that is solved in practice it is important to assure ourselves that the discretisation is sensible. For brevity, we fix the model parameter $q = 30$.

We directly examine the inconsistent discretisation for $u$ with the penalty parameter $\epsilon = 5 \times 10^4$ in the coupled case where $q \neq 0$ and fixing the $[\mathbb{Q}_2]^2$-approximation for $\mathbf{Q}$. In some unreported preliminary experiments, we observed that varying the approximations for $u$ does not affect the convergence behaviour of $\mathbf{Q}$, that is to say, the error in $\mathbf{Q}$ depends mainly on the element used for $\mathbf{Q}$, but the polynomial that approximates $u$ should have at least the same degree as that for $\mathbf{Q}$. We thus give the convergence rates separately for $u$ and $\mathbf{Q}$ in Tables 8.5 and 8.6. It can be seen that $\mathbf{Q}$ retains optimal rates in both the $H^1$ and $L^2$ norms, and though there are some fluctuations of the order for $u$, it still possesses very similar convergence rates when compared with the decoupled case described in Table 8.4.

**Remark 8.12.** *We also tested the convergence with the consistent weak formulation for $u$ under the same numerical settings as in Tables 8.5 and 8.6. We found that in both cases they present very similar convergence behaviour and thus we skip the details here.*

| | $N = \frac{1}{h}$ | $\|\mathbf{e}_u\|_0$ | rate | $\|\mathbf{e}_u\|_1$ | rate | $\|\|\|\mathbf{e}_u\|\|\|_h$ | rate |
|---|---|---|---|---|---|---|---|
| $\mathbb{Q}_2$ | 6 | 1.21e-05 | – | 3.59e-04 | – | 1.37e-02 | – |
| | 12 | 3.98e-06 | 1.61 | 1.42e-04 | 1.34 | 8.30e-03 | 0.72 |
| | 24 | 1.57e-06 | 1.35 | 4.99e-05 | 1.51 | 3.89e-03 | 1.09 |
| | 48 | 2.58e-07 | 2.60 | 9.06e-06 | 2.46 | 1.78e-03 | 1.13 |
| $\mathbb{Q}_3$ | 6 | 7.36e-06 | – | 2.25e-04 | – | 9.10e-03 | – |
| | 12 | 4.13e-07 | 4.16 | 1.86e-05 | 3.60 | 1.11e-03 | 3.03 |
| | 24 | 4.23e-08 | 3.29 | 2.24e-06 | 3.05 | 2.53e-04 | 2.14 |
| | 48 | 3.01e-09 | 3.81 | 2.28e-07 | 3.29 | 6.15e-05 | 2.04 |

**Table 8.5:** Convergence orders for $u$ with $q = 30$ and the penalty parameter $\epsilon = 5 \times 10^4$ in the inconsistent discretisation (8.1.2.4) for $u$, while fixing $\mathbf{Q}$ with the approximation $[\mathbb{Q}_2]^2$.

| | $N = \frac{1}{h}$ | $\|\mathbf{e}_\mathbf{Q}\|_0$ | rate | $\|\mathbf{e}_\mathbf{Q}\|_1$ | rate |
|---|---|---|---|---|---|
| $[\mathbb{Q}_1]^2$ | 6 | 8.12e-04 | – | 3.78e-02 | – |
| | 12 | 2.02e-04 | 2.01 | 1.88e-02 | 1.01 |
| | 24 | 5.05e-05 | 2.00 | 9.39e-03 | 1.00 |
| | 48 | 1.26e-05 | 2.00 | 4.69e-03 | 1.00 |
| $[\mathbb{Q}_2]^2$ | 6 | 2.92e-05 | – | 1.11e-03 | – |
| | 12 | 3.90e-06 | 2.90 | 2.71e-04 | 2.04 |
| | 24 | 5.02e-07 | 2.96 | 6.72e-05 | 2.01 |
| | 48 | 6.37e-08 | 2.98 | 1.68e-05 | 2.00 |
| $[\mathbb{Q}_3]^2$ | 6 | 3.02e-07 | – | 2.25e-05 | – |
| | 12 | 2.17e-08 | 3.80 | 2.72e-06 | 3.05 |
| | 24 | 1.45e-09 | 3.90 | 3.34e-07 | 3.03 |
| | 48 | 9.32e-11 | 3.96 | 4.13e-08 | 3.01 |

**Table 8.6:** Convergence orders for $\mathbf{Q}$ with $q = 30$ when coupled with the inconsistent discretisation for $u$ employing the penalty parameter $\epsilon = 5 \times 10^4$, while fixing $u$ with the approximation $\mathbb{Q}_3$.

Since the error norms for the finite element pair $\mathbb{Q}_3 \times [\mathbb{Q}_2]^2$ for $(u, \mathbf{Q})$ are in a rather close level of magnitude with a reasonable computational cost, we choose this approximation in our subsequent numerical experiments in Chapter 9.

## 8.3 Summary

In this chapter, we derived some a priori error estimates related to our proposed model (7.3.1.2) for smectics and examined the convergence rates in two dimensions via the method of manufactured solutions. We focused the analysis on the decoupled

case for simplicity. Optimal rates in both $L^2$ and $H^1$ norms were shown and verified for the tensor $\mathbf{Q}$. Moreover, we proved optimal convergence rates for $u$ in the mesh-dependent norm $|||\cdot|||_h$ and the $L^2$ norm $\|\cdot\|_0$ (only suboptimal for piecewise quadratic polynomials). This was also illustrated in numerical experiments. By studying the convergence behaviour of different finite element choices, we noted that $\mathbb{Q}_3 \times [\mathbb{Q}_2]^2$ for $(u, \mathbf{Q})$ with the penalty parameter $\epsilon = 5 \times 10^4$ is a suitable choice to be applied to further scenarios where physically realistic defects need to be characterised. We will apply our model and discretisation to situations of physical interest in the next chapter.

# 9 Numerical experiments for smectics

## Contents



With the convergent finite element pair $\mathbb{Q}_3 \times [\mathbb{Q}_2]^2$ for $(u, \mathbf{Q})$ at hand, we now consider three scenarios of physical interest: the defect-free example from the work of Williams & Kléman [WK75], a focal conic domain simulation, and an oily streaks simulation. The first scenario is a simple example intended to examine the bending effect in smectics, while the latter two experiments depict two typical defects in smectics, thus elucidating the effectiveness of our proposed model.

For the choice of parameters, we mainly use the values suggested in Pevnyi et al. [PSS14], occasionally varying them based on physical intuition (e.g., choosing a larger wave number $q$ to achieve thinner layers, or a larger anchoring weight $w$ to more strongly enforce the boundary conditions). The new parameters that do not appear in the model of Pevnyi et al. (e.g., $l$ and $w$) were chosen via unreported initial numerical experiments.

## 9.1 Implementation details

As discussed in Section 8.2, we choose $\mathcal{C}^0$-continuous finite element pairs for $(u, \mathbf{Q})$ with the penalty parameter $\epsilon = 5 \times 10^4$ throughout this chapter. In two dimensions, we use quadrilateral meshes. Since we restrict $\mathbf{Q}$ to be a symmetric and traceless tensor, it has two independent components in two dimensions and we thus seek the components of $\mathbf{Q}$ in $[\mathbb{Q}_2]^2$ and $u$ in $\mathbb{Q}_3$. We utilise hexahedral meshes in three dimensions, and since $\mathbf{Q}$ has five independent components, we then seek its components in $[\mathbb{Q}_2]^5$, while retaining $u$ in $\mathbb{Q}_3$.

In the numerical experiments, the nonlinear solve is deemed to have converged when the Euclidean norm of the residual falls below $10^{-8}$, or reduces from its initial value by a factor of $10^{-8}$, whichever comes first. For the inner solves, the linearised systems are solved using the sparse LU factorisation library MUMPS [ADL00]. The mesh scale, $h_e$, employed in the $\mathcal{C}^0$ interior penalty approach is chosen to be the average of the diameters of the cells on either side of an edge.

To compute the stability of each solution profile, we calculate the inertia of the Hessian matrix of the energy functional with a Cholesky factorisation, implemented in MUMPS [ADL00]. If the Hessian matrix is positive semi-definite, we characterise the solution as stable, while any nonzero number of negative eigenvalues characterises an unstable solution [NW99]. Note that no zero eigenvalues of Hessians were observed in this chapter, i.e., the stable solutions all in fact had positive-definite Hessian matrices. For a handful of parameter values where deflation yields a solution of lowest energy that is unstable (i.e., does not find a candidate ground state), we then calculate the eigen-directions of negative curvature using the Krylov–Schur algorithm [Ste02] implemented in SLEPc [HRV05]. We then perturb the lowest-energy solution along its eigen-directions of negative curvature and employ the bounded Newton line search algorithm of TAO [Den+20] to converge to a stable solution of minimal energy.

We give further details for the configuration of each example in the remainder of this chapter.

**Code availability.** For reproducibility, both the solver code [Xia21c] and the exact version of Firedrake [Fir21b] used to produce the numerical results in this chapter have been archived on Zenodo. An installation of Firedrake with components matching those used here can be obtained by following the instructions at https://www.firedrakeproject.org/download.html with

```
python3 firedrake-install --doi 10.5281/zenodo.4441123
```

Defcon version #11e883c should then be installed, as described in https://bitbucket.org/pefarrell/defcon/.

## 9.2 Scenario I: defect free

This is a simple example proposed by the work of Williams and Kléman [WK75] to examine the bending effect in smectics. For a rectangle $\Omega = [-2, 2] \times [0, 2]$ with boundary labels

$$\Gamma_l = \{(x, y) : x = -2\}, \qquad \Gamma_r = \{(x, y) : x = 2\},$$
$$\Gamma_b = \{(x, y) : y = 0\}, \qquad \Gamma_t = \{(x, y) : y = 2\},$$

we strongly impose

$$\mathbf{Q} = \begin{bmatrix} (\cos\theta_0)^2 - \frac{1}{2} & -\cos\theta_0 \sin\theta_0 \\ -\cos\theta_0 \sin\theta_0 & (\sin\theta_0)^2 - \frac{1}{2} \end{bmatrix} \quad \text{on } \Gamma_b,$$
$$\mathbf{Q} = \begin{bmatrix} (\cos\theta_0)^2 - \frac{1}{2} & \cos\theta_0 \sin\theta_0 \\ \cos\theta_0 \sin\theta_0 & (\sin\theta_0)^2 - \frac{1}{2} \end{bmatrix} \quad \text{on } \Gamma_t,$$

and enforce periodic boundary conditions on the left and right boundaries, $\Gamma_l$ and $\Gamma_r$. The above Dirichlet data for $\mathbf{Q}$ is derived from imposing $\mathbf{n}_e = (\cos\theta_0, -\sin\theta_0)$ at the bottom boundary, $\Gamma_b$, and with $\mathbf{n}_e = (\cos\theta_0, \sin\theta_0)$ at the top boundary, $\Gamma_t$, for fixed $\theta_0 \in [0, \pi/2]$.

We discretise the domain $\Omega$ into $90 \times 30$ quadrilateral elements and take the following initial guesses for $u$ and $\mathbf{Q}$:

$$u = 1, \quad \mathbf{Q} = \mathbf{Q}_0, \tag{9.2.0.1}$$

where $\mathbf{Q}_0 = (\mathbf{n}_I \otimes \mathbf{n}_I - \frac{\mathbf{I}_2}{2})$ with

$$\mathbf{n}_I = \frac{1}{m_I}\begin{bmatrix} x\,(|x| - R) \\ (|x|)\,y \end{bmatrix},$$

and

$$m_I = |x|\sqrt{(R - |x|)^2 + y^2}.$$

Here, the initial guess for the $\mathbf{Q}$-tensor is computed from a simplified two-dimensional mathematical representation of a family of tori, and we have taken the major radius $R = 0.5$ in this implementation.

Furthermore, we specify the values of parameters in this experiment:

$$a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-5},\ K = 0.3,$$
$$q = 30, \quad \text{and } l = 30.$$

The total energy to be minimised in this scenario is

$$\begin{aligned}\mathcal{J}_\epsilon(u, \mathbf{Q}) = \int_\Omega \Bigg( & \frac{a_1}{2}(u)^2 + \frac{a_2}{3}(u)^3 + \frac{a_3}{4}(u)^4 \\ & + B\left|\mathcal{D}^2 u + q^2\left(\mathbf{Q} + \frac{\mathbf{I}_2}{2}\right)u\right|^2 \\ & + \frac{K}{2}|\nabla \mathbf{Q}|^2 - l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right) + l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right)^2\Bigg) \\ & + \sum_{e \in \mathcal{E}_I} \int_e \frac{1}{2h_e^3}\left(\llbracket \nabla u \rrbracket\right)^2. \end{aligned} \tag{9.2.0.2}$$

We present the bifurcation diagram in Figure 9.1 for this scenario and quantitatively determine which of these solutions is the ground state as a function of $\theta_0$. To give more details on those solution branches with the lowest energy in the bifurcation diagram, we show some computed stationary states in Figure 9.2 as a function of $\theta_0$ by minimising (9.2.0.2). For each state, we display the value of the energy functional

$$\begin{aligned}\mathcal{J}(u, \mathbf{Q}) = \int_\Omega \Bigg( & \frac{a_1}{2}(u)^2 + \frac{a_2}{3}(u)^3 + \frac{a_3}{4}(u)^4 \\ & + B\left|\mathcal{D}^2 u + q^2\left(\mathbf{Q} + \frac{\mathbf{I}_2}{2}\right)u\right|^2 \\ & + \frac{K}{2}|\nabla \mathbf{Q}|^2 - l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right) + l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right)^2\Bigg), \end{aligned}$$

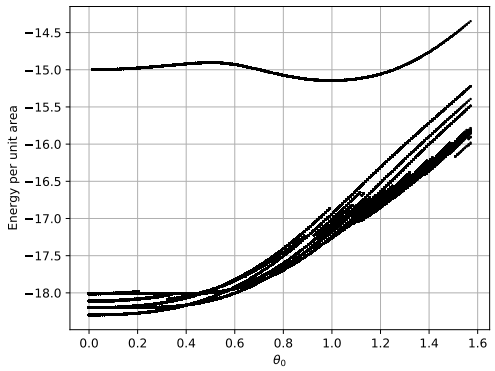



**Figure 9.1:** The bifurcation diagram of the defect-free scenario.

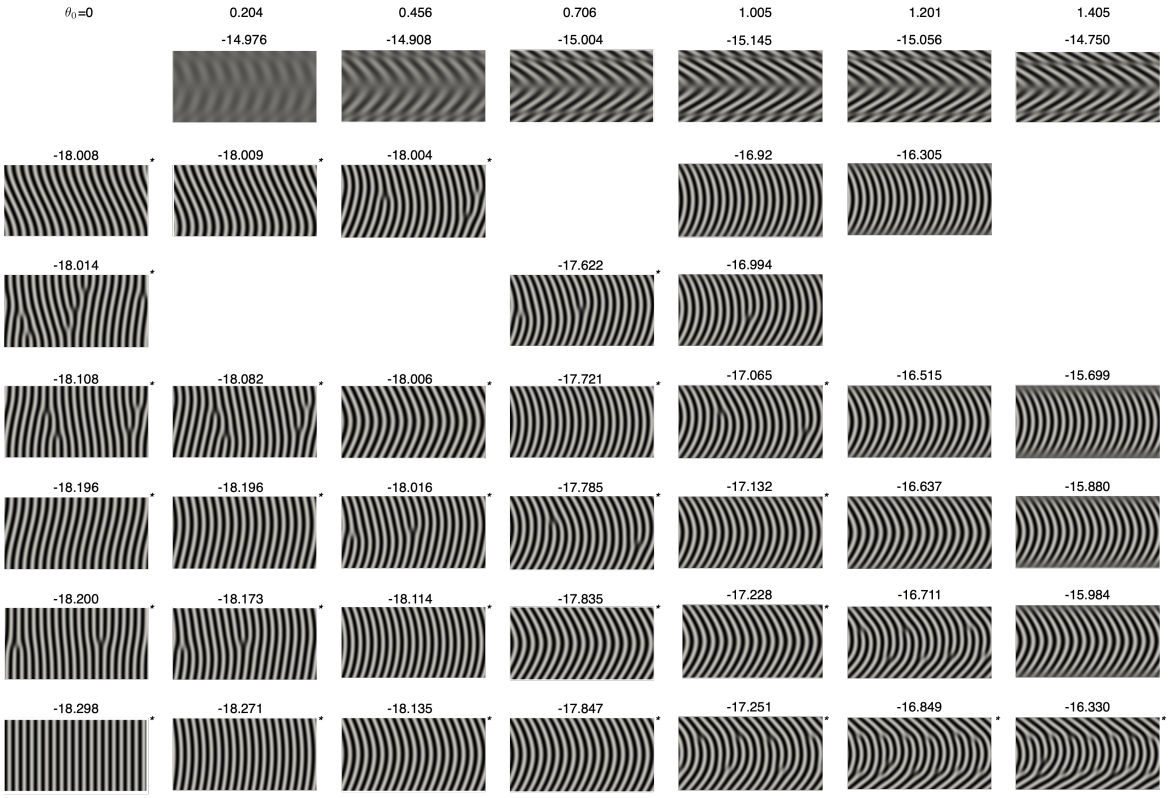



**Figure 9.2:** Stationary states obtained at different values of $\theta_0$ in the defect-free scenario. The visualisation displays the density perturbation $u$. For each solution, the value of the energy functional per unit area is displayed above it and we specify the stable profiles with asterisks. The bottom row depicts the lowest energy solution found for each value of $\theta_0$.

per unit area. For each column (i.e., fixed value of $\theta_0$), we organise the stationary states in an energy-decreasing order and identify stable profiles with asterisks. The bottom row depicts the lowest-energy minimisers found, all of which are stable.

We can observe from Figure 9.2 an energetic competition between the cost of bending and the cost of introducing disclinations from those equilibrium structure as a function of $\theta_0$. More specifically, when $\theta_0 = 0$ (thus the boundary conditions enforce that the director $\mathbf{n}_e$ is horizontally aligned), the resulting configuration is with the layers extending vertically between the substrates in the "bookshelf" geometry. As $\theta_0$ is increased from zero, the boundary conditions impose a bend deformation on the smectic. This can be accommodated in several ways: by distributing the deformation over the vertical direction (see the second picture in the bottom row of Figure 9.2); by localising the bend to a region in the center with the layers flat and tilted in opposite directions in the top and bottom of the domain (see the third and fourth pictures in the bottom row of Figure 9.2); or by introducing edge disclinations to relieve the cost of elastic deformation (see the last three pictures in the bottom row of Figure 9.2).

We also include one video *scenario-i-lowest-energy-in-theta-zero.mp4* in [Xia21a] to illustrate the stationary configurations of lowest energy found as we vary the applied bend deformation $\theta_0 \in [0, \pi/2]$. The profiles shown in this video are all stable.

## 9.3 Scenario II: focal conic domains

Among all defect structures in smectic liquid crystals, the most common one is focal conic domains (FCDs, as illustrated in Figure 1.3): the smectic layers are kept equidistant and parallel, with common normals and same center of curvature along the same normal. Such smectic layers are examples of *Dupin cyclides* which present two types of disclinations: ellipses and hyperbolas (also known as the *fonal conics*). When the ellipse degenerates to a circle and the hyperbola to a straight line, these smectic layers are called toroidal focal conic domains (TFCDs). In this section, we simulate FCDs and TFCDs using our proposed model (7.3.1.2).

We discretise the cuboid $\Omega = [-1.5, 1.5] \times [-1.5, 1.5] \times [0, 2]$ into $6 \times 6 \times 5$ uniform hexahedra, to avoid a directional bias observed in numerical solutions with tetrahedra. To simulate TFCDs or FCDs, we must impose boundary conditions (weakly or strongly) that respect their physical properties. To this end, we label the six boundary faces of $\Omega$ as

$$\begin{aligned} \Gamma_{left} &= \{(x,y,z) : x = -1.5\}, & \Gamma_{right} &= \{(x,y,z) : x = 1.5\}, \\ \Gamma_{back} &= \{(x,y,z) : y = -1.5\}, & \Gamma_{front} &= \{(x,y,z) : y = 1.5\}, \\ \Gamma_{bottom} &= \{(x,y,z) : z = 0\}, & \Gamma_{top} &= \{(x,y,z) : z = 2\}, \end{aligned}$$

and consider the following surface energy

$$F_{surface}(\mathbf{Q}) = \int_{\Gamma_{bottom}} \frac{w}{2} \left|\mathbf{Q} - \mathbf{Q}_{radial}\right|^2 + \int_{\Gamma_{top}} \frac{w}{2} \left|\mathbf{Q} - \mathbf{Q}_{vertical}\right|^2, \tag{9.3.0.1}$$

where $w$ denotes the weak anchoring weight,

$$\mathbf{Q}_{radial} = \begin{bmatrix} \frac{x^2}{x^2+y^2} - \frac{1}{3} & \frac{xy}{x^2+y^2} & 0 \\ \frac{xy}{x^2+y^2} & \frac{y^2}{x^2+y^2} - \frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{bmatrix}$$

represents an in-plane ($x$-$y$ plane) radial configuration of the director, and

$$\mathbf{Q}_{vertical} = \begin{bmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & 0 \\ 0 & 0 & \frac{2}{3} \end{bmatrix}$$

gives a vertical (i.e., along the $z$-axis) alignment configuration of the director. Therefore, the final form of the functional to be minimised in the TFCD scenario is

$$\begin{aligned} \mathcal{J}_\epsilon(u, \mathbf{Q}) = \int_\Omega \Bigg( & \frac{a}{2}(u)^2 + \frac{b}{3}(u)^3 + \frac{c}{4}(u)^4 \\ & + B \left| \mathcal{D}^2 u + q^2 \left( \mathbf{Q} + \frac{\mathbf{I}_3}{3} \right) u \right|^2 + \frac{K}{2} |\nabla \mathbf{Q}|^2 \\ & - \frac{l}{2} \left( \mathrm{tr}(\mathbf{Q}^2) \right) - \frac{l}{3} \left( \mathrm{tr}(\mathbf{Q}^3) \right) + \frac{l}{2} \left( \mathrm{tr}(\mathbf{Q}^2) \right)^2 \Bigg) \\ & + \int_{\Gamma_{bottom}} \frac{w}{2} \left|\mathbf{Q} - \mathbf{Q}_{radial}\right|^2 + \int_{\Gamma_{top}} \frac{w}{2} \left|\mathbf{Q} - \mathbf{Q}_{vertical}\right|^2 \\ & + \sum_{e \in \mathcal{E}_I} \int_e \frac{1}{2h_e^3} \left( [\![ \nabla u ]\!] \right)^2. \end{aligned} \tag{9.3.0.2}$$

For the FCD scenario, we only change the top boundary condition to perturb the preferred tilted director configuration. We perturb the angle $\theta_c$ between the director and the $z$-axis on the top surface $\Gamma_{top}$, thus adopting

$$\mathbf{Q}_c = \begin{bmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & (\sin(\theta_c))^2 - \frac{1}{3} & \sin(\theta_c)\cos(\theta_c) \\ 0 & \sin(\theta_c)\cos(\theta_c) & (\cos(\theta_c))^2 - \frac{1}{3} \end{bmatrix}$$

instead of $\mathbf{Q}_{vertical}$ in (9.3.0.2). Note that when taking $\theta_c = 0$, we return to the TFCD case.

Furthermore, we take the initial guesses:

$$u = \cos(6\pi z), \quad \mathbf{Q} = \mathbf{Q}_{ic},$$

where $\mathbf{Q}_{ic} = \left(\mathbf{n}_{ic} \otimes \mathbf{n}_{ic} - \frac{\mathbf{I}_3}{3}\right)$ with

$$\mathbf{n}_{ic} = \frac{1}{m_{ic}} \begin{bmatrix} x\left(\sqrt{x^2+y^2} - R\right) \\ y\left(\sqrt{x^2+y^2} - R\right) \\ z\left(\sqrt{x^2+y^2}\right) \end{bmatrix},$$

and

$$m_{ic} = \sqrt{x^2+y^2}\sqrt{\left(R - \sqrt{x^2+y^2}\right)^2 + z^2}.$$

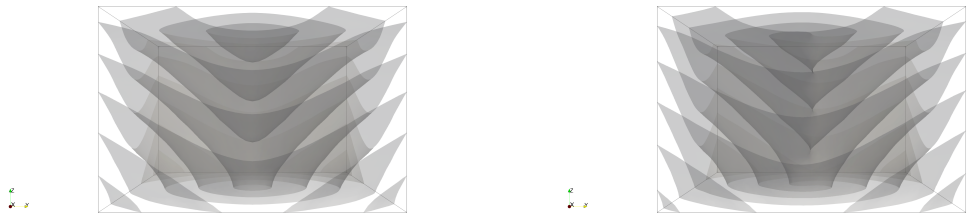
Here, the initial guess for the $\mathbf{Q}$-tensor is computed from the mathematical representation for a family of tori, and we have taken a major radius $R = 1.5$ in our implementation.

We specify the values of parameters used in the (T)FCD experiments:

$$a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-3},\ K = 0.03,$$

$$q = 10,\ l = 30 \text{ and } w = 10.$$

Two numerical solutions of simulating TFCDs are given in Figure 9.3. One can see that these zero isosurfaces of density indeed present a physically reasonable TFCD with two parts of singularities: circles at the bottom and the central line along the cusps. Notice that we are not imposing any periodic conditions of the density $u$ but only weakly enforcing boundary conditions as in (9.3.0.1) on the tensor field $\mathbf{Q}$. It turns out in Figure 9.3 that the smectic layers align themselves to the director field arising from $\mathbf{Q}$ and thus the periodicity on the lateral faces can be observed.



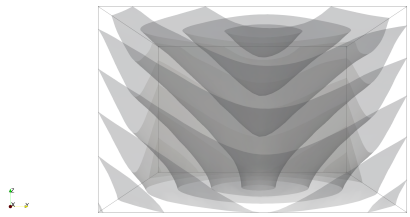
**(a)** TFCD **(b)** Single screw dislocation

**Figure 9.3:** Left: the first converged solution using Newton's method on a mesh of $6 \times 6 \times 5$ hexahedra using the TFCD settings; right: another solution profile with single screw dislocation around the central axis of the cuboid. The solution with screw dislocation has higher energy and both are stable. The gray layers are zero iso-surfaces of the density variation $u$.

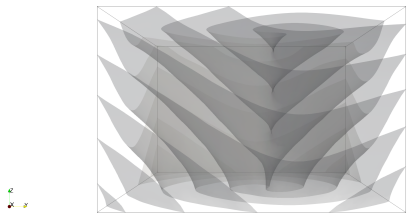
This is due to the coupling term in the model. Other than the TFCD solution as illustrated in Figure 9.3, it also shows another possibility of equilibrium solution with single screw dislocation at the central line, though a theoretical investigation of such interesting structure remains an open problem. We further comment that the single screw dislocation possesses higher energy value than that of the TFCD solution. At this point we are not sure if such a dislocation is physically realistic, but it presents an interesting pattern of defects in this numerical experiment.

In addition, we noticed from some preliminary experiments under the TFCD problem settings that a special case, i.e., the radial configuration of director molecules, of the planar anchoring condition is more likely to give a successful presentation of TFCDs. This may be helpful for a better and more accurate understanding of realistic boundary conditions to be enforced for the appearance of TFCDs.

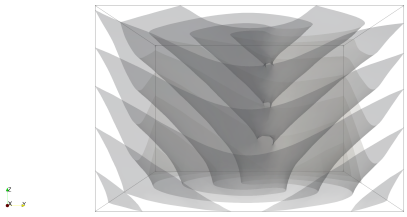
The TFCD profile shown in Figure 9.3 can be generalised into an asymmetric version, thus presenting the Dupin cyclides. We take $\theta_c = \frac{\pi}{12}$ and run the experiment with the other parameters chosen as in the TFCD settings. Three solution examples are shown in Figure 9.4, which includes an FCD solution Figure 9.4a, a single screw dislocation Figure 9.4b and a double screw dislocation structure Figure 9.4c. They are all stable solutions. It can be observed in the FCD solution profile that the smectic layers have deformed asymmetrically when responding to the tilting of the director on the top face. Note here that the FCD solution has the lowest energy due



**(a)** FCD

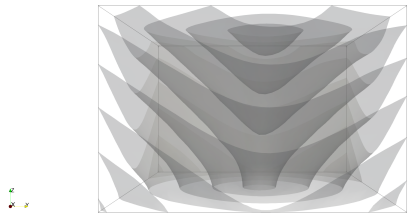


**(b)** Single screw dislocation

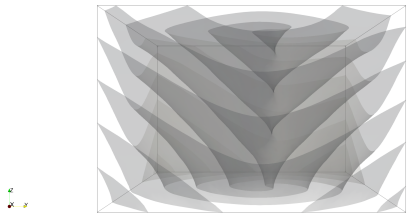


**(c)** Double screw dislocations

**Figure 9.4:** Three numerical solutions for $\theta_c = \frac{\pi}{12}$ on a mesh of $6 \times 6 \times 5$ hexahedra. The solution with double screw dislocations has highest energy while the FCD solution possesses lowest energy. All profiles are stable.



**(a)** FCD

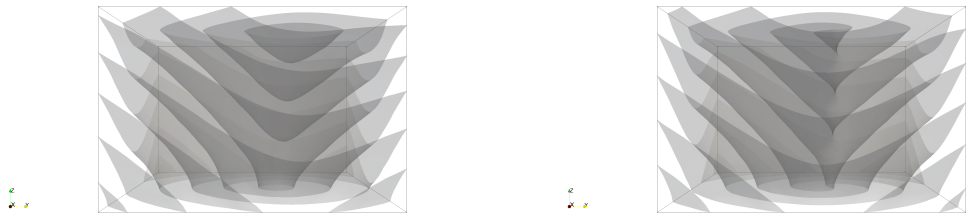


**(b)** Single screw dislocation

**Figure 9.5:** Two solution profiles by taking $\theta_c = \frac{\pi}{10}$ on a mesh of $6 \times 6 \times 5$ hexahedra. The solution with screw dislocation has higher energy. Both profiles are stable.

to the energy cost of the dislocation defects. To depict these three solution structures more closely, we further present an additional video *scenario-ii-pi12.mp4* in [Xia21a], describing the zero-isosurfaces of the smectic density variation field $u$ and colouring the isosurfaces by height (the $z$-coordinate) to assist with depth perception. The time axis of the video is used to illustrate the internal structure of the layers.

If we take $\theta_c = \frac{\pi}{10}$, the first converged solution shows a FCD structure as presented in Figure 9.5a. Another example is also given in Figure 9.5b which yields a single screw dislocation profile possessing higher energy. Again, both profiles are stable equilibrium points of the energy (9.3.0.2).



**(a)** FCD **(b)** Single screw dislocation

**Figure 9.6:** Two numerical solutions for $\theta_c = \frac{\pi}{8}$ on a mesh of $6 \times 6 \times 5$ hexahedra. The solution with screw dislocation has higher energy. Both profiles are stable.

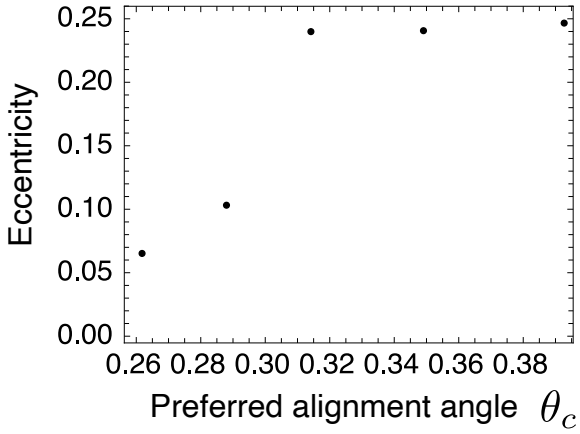



**Figure 9.7:** Eccentricity of FCD solutions as a function of preferred surface alignment angle.

Moreover, as we increase the value of $\theta_c$ to be $\frac{\pi}{8}$, two examples of stable numerical solutions are shown in Figure 9.6, where the focal conic curve in the FCD solution tilts more when compared with that in Figure 9.4a. We also see the screw dislocation structure possessing higher energy than that of the FCD solution in this experiment.

As the Dupin cyclide has a confocal pair of a hyperbola and an ellipse, we fit a hyperbola to each solution with least squares (data points extracted via ParaView [AGL05]) and calculate its eccentricity (e.g., for a hyperbola expressed as $\frac{y^2}{a_{fit}^2} - \frac{z^2}{b_{fit}^2} = 1$, its eccentricity is defined as $\frac{\sqrt{a_{fit}^2+b_{fit}^2}}{a_{fit}}$). Then the eccentricity of the ellipse is the inverse of that of the confocal hyperbola. Values of eccentricity fitted from the solution set are shown as a function of the preferred surface alignment angle $\theta_c$ in Figure 9.7.

## 9.4 Scenario III: oily streaks

Besides the (T)FCD defects illustrated in the previous section, there is another type of defects that are experimentally observable in films of 8CB deposited in air on crystalline surfaces of molybdenite ($MoS_2$) [Mic+04]: the so-called *oily streaks* (OS). When thin smectic liquid crystal films are subject to competing boundary conditions, they can form interesting patterns. In particular, *planar degenerate anchoring* (i.e., the molecules on the surface are in the plane of the surface) and *homeotropic anchoring* (i.e., the molecules prefer to be perpendicular to the surface) imposed on two opposing surfaces can form a periodic stacking of flattened hemi-cylinders, as shown in Figure 1.2. We simulate this typical defect in this section using our proposed model (7.3.1.2).

Let $r$ denote the aspect ratio of a rectangle $\Omega = [-r, r] \times [0, 2]$ with the boundaries labels

$$\Gamma_l = \{(x, y) : x = -r\}, \qquad \Gamma_r = \{(x, y) : x = r\},$$
$$\Gamma_b = \{(x, y) : y = 0\}, \qquad \Gamma_t = \{(x, y) : y = 2\}.$$

We impose the following surface energy

$$F_{surface}(\mathbf{Q}) = \int_{\Gamma_b} \frac{w}{2} |\mathbf{Q} - \mathbf{Q}_{bottom}|^2 + \int_{\Gamma_t \cup \Gamma_l \cup \Gamma_r} \frac{w}{2} |\mathbf{Q} - \mathbf{Q}_{top}|^2,$$

where $w$ is the weak anchoring weight and two weakly prescribed configurations $\mathbf{Q}_{bottom}$ and $\mathbf{Q}_{top}$ are given by

$$\mathbf{Q}_{bottom} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix},$$

yielding horizontally aligned directors, and

$$\mathbf{Q}_{top} = \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

yielding vertically aligned directors.

In this experiment, we always discretise the domain $\Omega$ into $90 \times 30$ quadrilateral elements, even as we change the domain size by varying the aspect ratio $r$. The final form of the functional to be minimised in this scenario is

$$\begin{aligned}\mathcal{J}_\epsilon(u, \mathbf{Q}) = \int_\Omega \Bigg( & \frac{a_1}{2}(u)^2 + \frac{a_2}{3}(u)^3 + \frac{a_3}{4}(u)^4 \\ & + B\left|\mathcal{D}^2 u + q^2\left(\mathbf{Q} + \frac{\mathbf{I}_2}{2}\right)u\right|^2 \\ & + \frac{K}{2}|\nabla \mathbf{Q}|^2 - l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right) + l\left(\mathrm{tr}\left(\mathbf{Q}^2\right)\right)^2\Bigg) \\ & + \int_{\Gamma_b} \frac{w}{2}|\mathbf{Q} - \mathbf{Q}_{bottom}|^2 + \int_{\Gamma_t \cup \Gamma_l \cup \Gamma_r} \frac{w}{2}|\mathbf{Q} - \mathbf{Q}_{top}|^2 \\ & + \sum_{e \in \mathcal{E}_I} \int_e \frac{1}{2h_e^3}\left(\llbracket \nabla u \rrbracket\right)^2 .\end{aligned} \tag{9.4.0.1}$$

We take the same form of the initial guesses for $u$ and $\mathbf{Q}$ as in (9.2.0.1) but with a larger major radius $R = 1$ in this scenario.

Finally, we specify the values of parameters in this experiment:

$$a_1 = -10,\ a_2 = 0,\ a_3 = 10,\ B = 10^{-5},\ K = 0.3,$$
$$q = 30,\ l = 1 \text{ and } w = 10.$$

Based on X-ray diffraction experiments of thin smectic films, Michel et al. [Lac+07] proposed some approximate structures of oily streaks as illustrated in Figure 9.8A-C. Since some experiments reveal that the smectic layer normals are continuously oriented for smectic layers that are parallel to the plane of substrate for thin films, the authors gave a possible structure in Figure 9.8A depicting periodic units incorporating sections of cylinders joined to planes oriented parallel to the substrate. However, this structure implies significant deformations of the free interface with singular points between units. To avoid so, they proposed a more complex structure as illustrated in Figure 9.8B incorporating curvature walls between units. Moreover, it is observed in the X-ray diffraction of even thinner films that an apparent excess of the planar region is shown, which cannot be explained by either structure discussed so far [MLG06]. Therefore, Figure 9.8C provides a possible structure consistent with the experimental data envisioned in [MLG06], though it is energetically very costly.

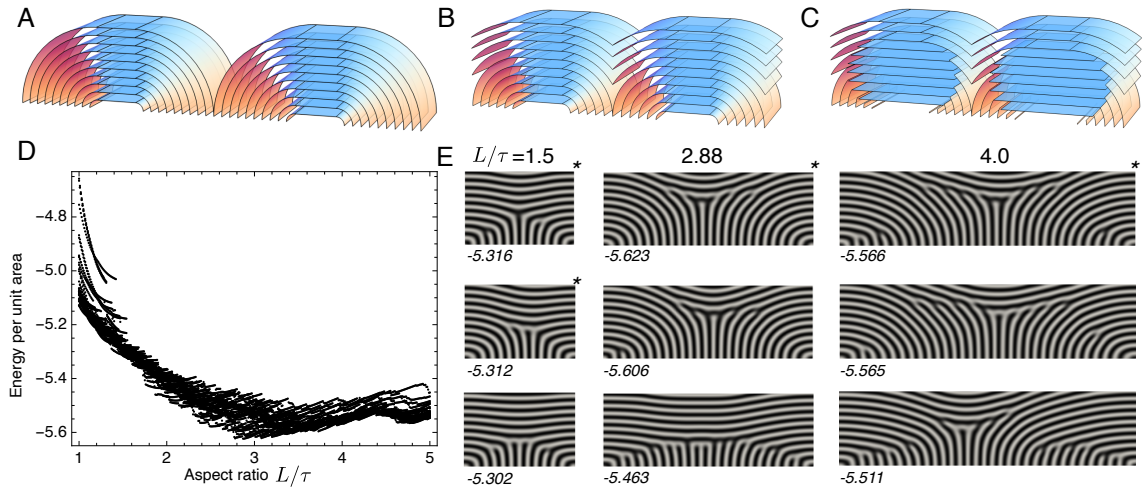



**Figure 9.8:** Oily streaks. **A-C** Candidate structures proposed in Michel et al. [Lac+07] consistent with X-ray diffraction. **D** Bifurcation diagram of structures as a function of aspect ratio. **E** Selected stationary states obtained at different aspect ratio $r$. The top row represents the lowest energy solution found. For each solution, the value of the energy functional per unit area is displayed below it with asterisks indicating stable profiles.

By implementing the proposed mathematical model, we display the partially enumerated energy landscape in Figure 9.8D, showing an extremely dense thicket of solutions. This qualitatively supports earlier work in that an overall minimiser occurs at an aspect ratio of around 3, which is similar to experimental values even with no parameter tuning performed here. Close examination of the energy landscape, together with the corresponding solution set, shows many small discontinuous jumps that result from delicate commensurability effects, whereby certain sizes of domain are compatible with a given periodicity of the layers as well as from variations in the number of defects and their detailed placement. Similar effects have been observed when other periodic liquid crystals such as cholesterics are confined in domains that promote geometric frustration [Eme+18].

The solution set obtained contains examples reminiscent of previously proposed structures (Figure 9.8E). The minimum energy states found at different aspect ratio contain cylindrical sections mediated by a defect-filled region reminiscent of the mesoscopic rotating grain boundaries. Other solutions displayed in the lowest row of Figure 9.8E are quite different from those heretofore proposed, where regions of relatively vertically oriented layers sit atop cylindrical regions interspersed

with defects. Each of these incorporates a greater proportion of vertical layers relative to the hemicylindrical-planar *ansatz* of Figure 9.8A,B and may provide alternative structures for oily streaks in ultrathin films. In future work, the boundary conditions at the top interface should be carefully reconsidered, including the incorporation of a free interface.

We refer readers to the video *scenario-iii-lowest-energy-in-r.mp4* [Xia21a] depicting the lowest-energy configurations discovered as we vary the aspect ratio $r \in [1, 5]$. All presented profiles in this video are stable.

## 9.5 Summary

In this chapter, we simulated three smectic scenarios involving boundary conditions that are incompatible with uniform smectic order to investigate the effectiveness of our proposed mathematical model (7.3.1.2) in characterising the defect structures, e.g., (toroidal) focal conic domains and oily streaks in smectics. Our new model successfully reproduced, even without careful tuning of parameters, a number of experimentally observed and theoretically expected phenomena, as well as producing new candidate structures for thin smectic films that are explicitly stationary states of an energy functional. We believe this success can lead to many other smectic applications in future.

# 10 Conclusions and future work

## Contents



## 10.1 Conclusions

This thesis tackles and implements several energy minimisation problems arising from modelling cholesteric liquid crystals, ferronematics and smectic-A liquid crystals.

In Chapter 2–Chapter 4, we consider the Oseen–Frank model of cholesteric liquid crystals that employs a vector-valued director field as state variable, subject to a unit-length constraint. We apply augmented Lagrangian methods to transform the constrained minimisation problem into an unconstrained one of saddle point type. The benefits of the AL method are twofold: it helps control the Schur complement, enabling fast solvers; and it improves the discrete constraint as we increase the value of the penalty parameter in the implementation. The details of the relevant discussions are illustrated in Chapter 2. The tradeoff is that it complicates the solution of the top-left director block, as it adds a semi-definite

term with a large coefficient arising from the AL formulation. To resolve this issue, our core contribution in Chapter 3 is to develop a robust and efficient multigrid solver. A parameter-robust relaxation method is achieved by developing a space decomposition that stably captures the kernel of the semi-definite terms. Chapter 4 demonstrates the validity of our derived parameter- and mesh-independent solver through several numerical experiments.

Due to the difficulties of (i) solving a constrained minimisation problem and (ii) representing certain defect structures (e.g., half charge defects), we turn from the Oseen–Frank theory to the Landau–de Gennes modelling theory that uses a tensor-valued state variable. We consider a one-dimensional model of ferronematics in Chapters 5 and 6 to study order reconstruction solutions, bifurcations, and multistability. We construct a novel numerical bifurcation analysis in Chapter 6 of theoretical results analysed in [Dal+21] and perform an asymptotic analysis (see Section 6.4) for certain model parameters. We pay special attention to defect structures (domain walls in ferronematics) in our investigation. These numerical studies form a solid basis for validating analytical results and demonstrate the promising potential of capturing defects using the **Q**-tensor theory.

In the last part of this thesis (Chapters 7 to 9), we devote ourselves to proposing a new continuum mathematical model for smectic-A liquid crystals, and developing a convergent finite element discretisation thereof. To represent half charge defects that are likely to happen in smectics, the model is characterised by a tensor-valued nematic order parameter and a real-valued smectic order parameter. We prove an existence result in Chapter 7 for the proposed minimisation problem. Chapter 8 investigates an appropriate finite element formulation for solving the optimality conditions, which are essentially a coupled system involving a fourth order PDE and a second order PDE. For the fourth order problem, we take the common Lagrange elements with an interior penalisation term to avoid the use of more complicated $H^2$-conforming elements. The second order PDE, which comes from the classical Landau–de Gennes model for nematic phases, is simpler and is discretised with standard Lagrange elements. This chapter derives some a priori error estimates

for both variables in the decoupled case, accompanied by numerical verifications of convergence rates in the coupled case. Some interesting applications of the new model are presented in Chapter 9, where some typical defect structures are numerically captured for the first time. This shows promise for further related work in smectic liquid crystals.

## 10.2 Future work I

Regarding the Oseen–Frank model, we have developed in Chapter 3 the theory for the construction of a robust multigrid algorithm for the equal-constant nematic LC. Extensions to the multi-constant case give rise to some additional difficulties, especially in the characterisation of the kernels of the $\nabla\cdot$ and $\nabla\times$ operators in the Frank energy density (2.1.0.3). A potential resolution for this difficulty is to use the de Rham complexes [AFW00]. The smooth de Rham complex in two dimensions is given by

$$\mathbb{R} \xrightarrow{\text{id}} C^\infty(\Omega) \xrightarrow{\nabla\times} [C^\infty(\Omega)]^2 \xrightarrow{\nabla\cdot} C^\infty(\Omega) \xrightarrow{\text{null}} 0,$$

where the kernel $\text{Ker}(\cdot)$ of an operator is the range $\text{Range}(\cdot)$ of the preceding operator on a simply connected domain. For instance, $\text{Range}(\nabla\times) = \text{Ker}(\nabla\cdot)$. This allows us to characterise the divergence-free vector fields as the curls of potentials. However, the above de Rham complex is rather restrictive in implementation as it requires smooth spaces. For our interests in LC problems with directors having $H^1$-regularity, we should instead utilise complexes involving Sobolev spaces, e.g., the so-called Stokes complex in two dimensions:

$$\mathbb{R} \xrightarrow{\text{id}} H^2(\Omega) \xrightarrow{\nabla\times} [H^1(\Omega)]^2 \xrightarrow{\nabla\cdot} L^2(\Omega) \xrightarrow{\text{null}} 0.$$

Discrete versions of these complexes are much harder to construct and often result in high order polynomials due to the high regularity requirements, such as the $H^2$-regularity. The study of an appropriate de Rham complex will help characterise the kernel of $\nabla\cdot$ and $\nabla\times$ operators in the finite element spaces. This will allow for the preconditioner developed in this thesis to be analysed for the multi-constant case.

## 10.3 Future work II

With the success in predicting typical defects in smectic-A liquid crystals, we can extend our result to encompass the smectic-C phase, and thus give a unified model for liquid crystals including isotropic, nematic, smectic-A and C phase transitions.

The idea can be built on the work of Biscari, Calderer and Terentjev [BCT07], who present a de Gennes variational theory based on a complex-valued smectic order parameter $\psi$ and a tensor-valued nematic order parameter $\mathbf{Q}$ to simultaneously describe those transitions. More specifically, the difference between smectic-A and C phases is characterised by a new interaction term

$$\boldsymbol{\chi} \coloneqq \mathbf{Q}\nabla\psi \times \nabla\psi. \tag{10.3.0.1}$$

If the nematic director is aligned to the smectic layer normals as in the smectic-A phase, then $\boldsymbol{\chi} = 0$, otherwise a nonzero $\boldsymbol{\chi}$ represents a smectic-C phase. The following energy from the interaction term characterising smectic-C phases is added to the free energy:

$$\int_\Omega e_{AC}\boldsymbol{\chi}\cdot\boldsymbol{\chi} = \int_\Omega e_{AC}\left|\mathbf{Q}\nabla\psi \times \nabla\psi\right|^2, \tag{10.3.0.2}$$

where $e_{AC}$ is a constant. Note that a negative value of $e_{AC}$ will enforce smectic-C phases in the model and a positive value results in smectic-A phases.

Considering our proposed model of smectic-A LC in Chapter 7, which is based on a real-valued smectic density $u$ and a tensor-valued nematic order parameter $\mathbf{Q}$, we intend to introduce the following interaction term similar to (10.3.0.1) to distinguish the smectic-A and C phases:

$$\boldsymbol{\chi} = \mathbf{Q}\nabla u \times \nabla u,$$

and add

$$\int_\Omega \frac{e_{AC}}{2}\left|\mathbf{Q}\nabla u \times \nabla u\right|^2$$

to our proposed free energy (7.3.1.2).

One important potential application could be simulating smectic-C LC in a wedge, as illustrated in [CSL91, Section 3], where smectic layers are expected to form concentric cylinders with the common axis coinciding with the center of the wedge. This simulation is used there to examine different distortion effects existed in smectic-C LC. Another avenue to pursue is to investigate the chevron structure (see [BCT07, Section IV]), one of the most interesting defects existing in the smectic-C phase.

## 10.4 Future work III

Concerning the smectic-A phase, there are several topics that can be pursued further using our proposed smectic model (7.3.1.2).

The computational time required to solve three-dimensional problems is noticeable longer than for two-dimensional problems. This motivates the use of a faster algorithm to improve computational efficiency. Some choices can be taken, e.g., designing a preconditioner for the model (7.3.1.2) or using the static condensation technique [Guy65; Iro65] to reduce the size of the stiffness matrix. Moreover, due to the similarities of our adopted $\mathcal{C}^0$-IP methods and the weakly over-penalised symmetric interior penalty method illustrated in [BGS10] for biharmonic problems, we may build on [BGS10] for the construction and analysis of efficient solvers for the smectic-smectic block of the matrix.

Since our proposed model characterises both nematic and smectic-A phases, it may be used to investigate the nematic-smectic transition by varying the temperature-dependent parameter $a_1$. Zappone et al. recently confirmed the existence of intermediate LC state analogous to superconductors [Zap+20] for thin smectic films of different thicknesses. In particular, they find the so-called P-texture (see Figure 10.1) only observed when cooling a thin smectic film. It can be seen from this schematic description that the $-\frac{1}{2}$ defects possess similar structures of defect walls as in the oily streaks problem explored in Section 9.4. This motivates us to apply our new model to study the nematic-smectic transition.

From the numerical perspective and inspired by the progress of using our proposed smectic-A model (7.3.1.2) to capture typical defects in smectics, we believe

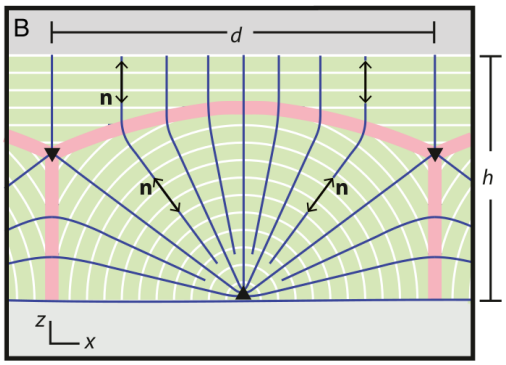



**Figure 10.1:** The P-texture profile taken from [Zap+20] where blue and white lines indicate the director and smectic layers, respectively. Smectic layers penetrate at the pink-shaded region while the upward- and downward-pointing triangles represent defects with $+\frac{1}{2}$ and $-\frac{1}{2}$ charge, respectively.

it can be further applied to more laboratory experiments to help in investigating internal defect structures. For instance, one could use our smectic model to characterise and analyse edge and screw dislocations in a wedge similarly to [LBK06]. We give a preliminary result (see Figure 10.3) related to this wedge problem that is schematically described in Figure 10.2.

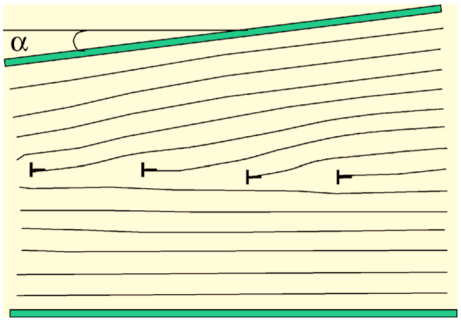
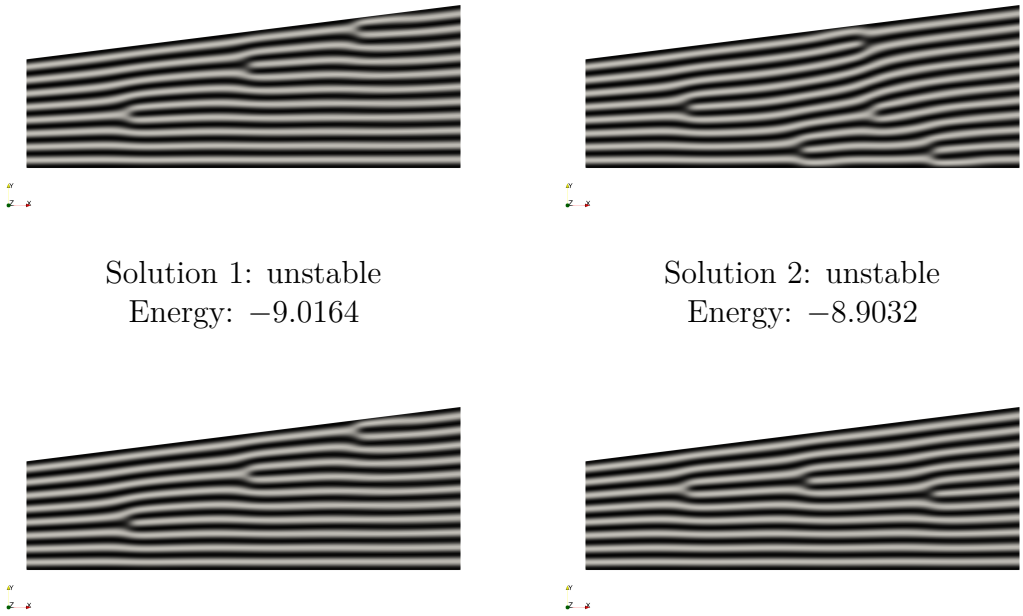



**Figure 10.2:** Figure 2 of [LBK06]. Original caption: *Schematic cross section of a wedge-shaped homeotropic smectic-A sample, containing a tilt subboundary of edge dislocations. $\alpha$ is the wedge angle formed by the glass plates.*

Another avenue of investigation is to compare results from our model with actual experiments and with simulations conducted using other methods (particularly Monte Carlo and Density Functional Theory). This would yield a better understanding of the strengths and weaknesses of the different available smectic modelling theories. We have begun to collaborate with the authors of [Wit+21] to investigate the smectic structures that are predicted by different modelling frameworks in



Solution 1: unstable
Energy: −9.0164

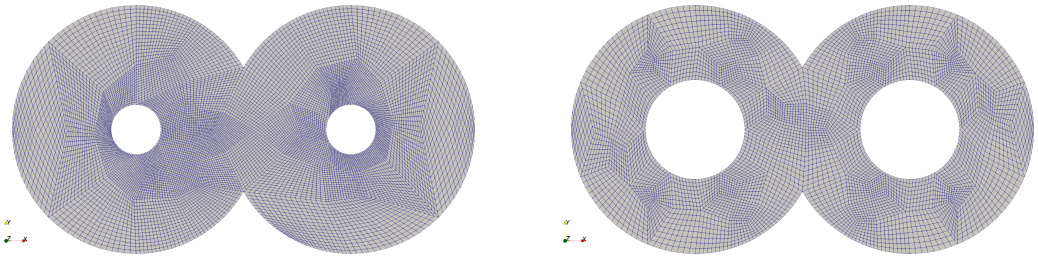
Solution 2: unstable
Energy: −8.9032

Solution 3: stable
Energy: −9.0182

Solution 4: stable
Energy: −9.0286

**Figure 10.3:** Four solution profiles and their stabilities with strongly-enforced Dirichlet data on $\delta\rho = 1$ and strongly-enforced homeotropic boundary conditions of $\mathbf{Q}$ on top and bottom surfaces of a wedge. Solution 4 with three edge dislocations has the lowest energy.

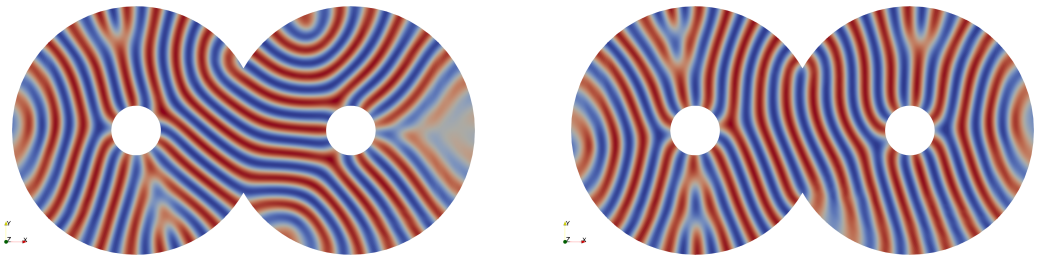
confined geometries with holes. A simple example of a geometry to be considered in this work is two overlapped annuli, as illustrated in Figure 10.4. We present some preliminary results (see Figures 10.5 and 10.6) of obtained profiles when tangential boundary conditions are imposed along both external and inner circles of the annuli. As of writing, laboratory experiments in these geometries are underway, led by Prof. Dirk Aarts of the Oxford Colloid Group.



Ratio = 0.2

Ratio = 0.4

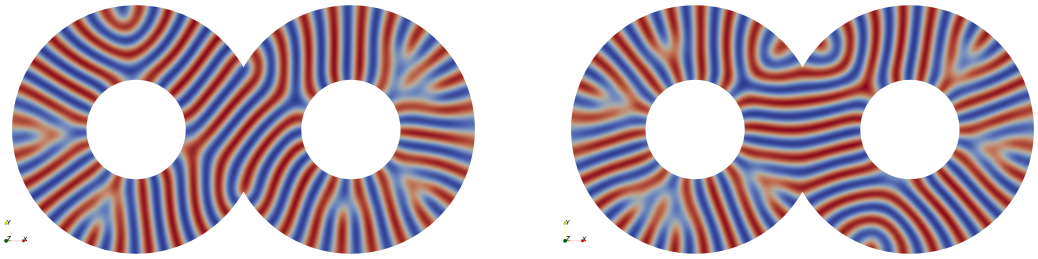
**Figure 10.4:** Meshes of two fused annuli. The domains differ in the sizes of the inclusions.



Stable;
Energy: −16.519

Stable;
Energy: −16.855

**Figure 10.5:** Two solution profiles of the geometry with inclusion ratio 0.2.



Stable;
Energy: −14.696

Unstable;
Energy: −14.449

**Figure 10.6:** Two solution profiles for the geometry with inclusion ratio 0.4.

# Appendices

# A

# Equilibrium equations in two dimensions

To construct the manufactured solution for numerical verification of the theoretical convergence order (see Section 8.2), we need to derive the strong form of the equilibrium equations of the minimisation problem. In two dimensions, the free energy functional to be minimised is

$$\begin{aligned}\mathcal{J}(u,\mathbf{Q}) = \int_\Omega \Bigg( & \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 \\ & + B\left|\mathcal{D}^2 u + q^2\left(\mathbf{Q} + \frac{\mathbf{I}_2}{2}\right)u\right|^2 \\ & + \frac{K}{2}|\nabla\mathbf{Q}|^2 - l\left(\mathrm{tr}(\mathbf{Q}^2)\right) + l\left(\mathrm{tr}(\mathbf{Q}^2)\right)^2\Bigg),\end{aligned}$$

with real parameters $a_1, a_2, a_3, B, q, K, l$. Note that $\mathbf{Q}$ is a symmetric and traceless $2\times 2$ matrix and thus can be represented by two degrees of freedom $(Q_{11}, Q_{12})$ as given by (7.3.1.1). Then, we rewrite the above free energy in terms of variables $(Q_{11}, Q_{12}, u)$ as follows,

$$\begin{aligned}\mathcal{J}(Q_{11},Q_{12},u) = \int_\Omega \Bigg( & \frac{a_1}{2}u^2 + \frac{a_2}{3}u^3 + \frac{a_3}{4}u^4 \\ & + B|\mathcal{D}^2 u|^2 + Bq^4u^2\left(2\left(Q_{11}^2+Q_{12}^2\right)+\frac{1}{2}\right) \\ & + 2Bq^2u\left(\left(Q_{11}+\frac{1}{2}\right)\partial_x^2 u + \left(-Q_{11}+\frac{1}{2}\right)\partial_y^2 u + 2Q_{12}\partial_x\partial_y u\right) \\ & + K\left|\nabla Q_{11}\right|^2 + K\left|\nabla Q_{12}\right|^2 - 2l\left(Q_{11}^2+Q_{12}^2\right) + 4l\left(Q_{11}^2+Q_{12}^2\right)^2\Bigg).\end{aligned} \tag{A.0.0.1}$$

The admissible set for $(Q_{11}, Q_{12}, u)$ based on (7.3.2.1) is denoted as

$$\tilde{\mathcal{A}}^s = \{u \in H^2(\Omega, \mathbb{R}), (Q_{11}, Q_{12}) \in H^1(\Omega, \mathbb{R}^2) \colon (Q_{11}, Q_{12}) = \mathbf{q}_b \text{ on } \partial\Omega\},$$

where $\mathbf{q}_b = (q_{b,1}, q_{b,2})^T$ is the prescribed Dirichlet boundary data arising from $\mathbf{Q}_b$.

**Remark A.1.** *Note that the uniaxiality condition is not included in the admissible set here. This condition is beneficial for the variational analysis in Section 7.3.2, but enforcing the uniaxiality constraint strongly is not a trivial task [BNW20]. Instead, we weakly impose this constraint through the additional nematic bulk density $f_n^b(Q)$ in (7.3.1.4) which possesses a uniaxial minimiser by [MZ10, Proposition 15].*

**Remark A.2.** *Other choices of boundary data can be taken for $(Q_{11}, Q_{12})$; we choose Dirichlet boundary conditions for simplicity.*

By taking the test functions $(p_1, p_2, v) \in H_0^1(\Omega) \times H_0^1(\Omega) \times H^2(\Omega)$ and using integration by parts, we derive the weak form of the Euler–Lagrange equations for the energy functional (A.0.0.1),

$$\begin{aligned}
\mathcal{J}_{Q_{11}}(Q_{11}, Q_{12}, u; p_1) &= \int_\Omega \Big( 4Bq^4u^2Q_{11} + 2Bq^2u\left(\partial_x^2 u - \partial_y^2 u\right) \\
&\quad + 2K\Delta Q_{11} - 4lQ_{11} + 16lQ_{11}\left(Q_{11}^2 + Q_{12}^2\right)\Big) p_1 \\
&= 0 \quad \forall p_1 \in H_0^1(\Omega), \\
\mathcal{J}_{Q_{12}}(Q_{11}, Q_{12}, u; p_2) &= \int_\Omega \Big( 4Bq^4u^2Q_{12} + 4Bq^2u(\partial_x\partial_y u) \\
&\quad + 2K\Delta Q_{12} - 4lQ_{12} + 16lQ_{12}\left(Q_{11}^2 + Q_{12}^2\right)\Big) p_2 \\
&= 0 \quad \forall p_2 \in H_0^1(\Omega), \\
\mathcal{J}_u(Q_{11}, Q_{12}, u; v) &= \int_\Omega \Big( a_1u + a_2u^2 + a_3u^3 + 2B\nabla\cdot(\nabla\cdot(\mathcal{D}^2u)) \\
&\quad + Bq^4\left(4\left(Q_{11}^2 + Q_{12}^2\right) + 1\right)u \\
&\quad + 2Bq^2\left[(Q_{11} + 1/2)\partial_x^2 u + (-Q_{11} + 1/2)\partial_y^2 u + 2Q_{12}(\partial_x\partial_y u)\right] \\
&\quad + 2Bq^2\left[\partial_x^2(u(Q_{11} + 1/2)) + \partial_y^2(u(-Q_{11} + 1/2)) + 2\partial_x\partial_y(uQ_{12})\right]\Big) v \\
&\quad + 2BG_{1,b}(u; v) + 2Bq^2G_{2,b}(Q_{11}, Q_{12}, u; v) \\
&= 0 \quad \forall v \in H^2(\Omega),
\end{aligned}$$

where the boundary integrals $G_{1,b}$ and $G_{2,b}$ are of the form

$$G_{1,b}(u;v) = \int_{\partial\Omega} \nu \cdot \left(\mathcal{D}^2 u \cdot \nabla v\right) - \int_{\partial\Omega} \left(\left(\nabla \cdot (\mathcal{D}^2 u)\right) \cdot \nu\right) v$$

and

$$\begin{aligned} G_{2,b}(u, Q_{11}, Q_{12}; v) = &\int_{\partial\Omega} \left(-v\left(\partial_x(u(Q_{11}+1/2))\nu_x\right) + (\partial_x v)u(Q_{11}+1/2)\nu_x\right) \\ &+ \int_{\partial\Omega} \left(-v\left(\partial_y(u(-Q_{11}+1/2))\nu_y\right) + (\partial_y v)u(-Q_{11}+1/2)\nu_y\right) \\ &+ \int_{\partial\Omega} \left(-v\left(\partial_x(uQ_{12})\nu_y\right) + (\partial_y v)uQ_{12}\nu_x\right) \\ &+ \int_{\partial\Omega} \left(-v\left(\partial_y(uQ_{12})\nu_x\right) + (\partial_x v)uQ_{12}\nu_y\right). \end{aligned}$$

Therefore, the Euler–Lagrange equations for minimising the free energy (A.0.0.1) for $(Q_{11}, Q_{12}, u) \in \tilde{\mathcal{A}}^s$ are

$$\begin{cases} 4Bq^4u^2Q_{11} + 2Bq^2u\left(\partial_x^2 u - \partial_y^2 u\right) - 2K\Delta Q_{11} - 4lQ_{11} + 16lQ_{11}\left(Q_{11}^2 + Q_{12}^2\right) = 0, \\ 4Bq^4u^2Q_{12} + 4Bq^2u\left(\partial_x\partial_y u\right) - 2K\Delta Q_{12} - 4lQ_{12} + 16lQ_{12}\left(Q_{11}^2 + Q_{12}^2\right) = 0, \\ a_1u + a_2u^2 + a_3u^3 + 2B\nabla\cdot\left(\nabla\cdot(\mathcal{D}^2u)\right) + Bq^4\left(4\left(Q_{11}^2+Q_{12}^2\right)+1\right)u + 2Bq^2(t_1+t_2) = 0, \end{cases} \tag{A.0.0.2}$$

subject to the boundary conditions

$$\begin{aligned} (Q_{11}, Q_{12}) &= (q_{b,1}, q_{b,2}) \quad \text{on } \partial\Omega, \\ S^1_{bc}(u, q_{b,1}, q_{b,2}; v) &= 0 \ \forall v \in H^2(\Omega) \quad \text{on } \partial\Omega, \end{aligned}$$

where

$$\begin{aligned} t_1 &\coloneqq (Q_{11}+1/2)\partial_x^2 u + (-Q_{11}+1/2)\partial_y^2 u + 2Q_{12}\partial_x\partial_y u, \\ t_2 &\coloneqq \partial_x^2\left(u\left(Q_{11}+1/2\right)\right) + \partial_y^2(u(-Q_{11}+1/2)) + 2\partial_x\partial_y(uQ_{12}), \\ S^1_{bc}(u, q_{b,1}, q_{b,2}; v) &\coloneqq G_{1,b}(u;v) + q^2 G_{2,b}(u, q_{b,1}, q_{b,2}; v). \end{aligned}$$

These equations (A.0.0.2) are used for the numerical verification of the theoretical convergence rates derived in Chapter 8. Here, we will not derive the equilibrium equations for three dimensional problems due to their complicated form with six coupled degrees of freedom $(Q_{11}, Q_{12}, Q_{13}, Q_{22}, Q_{23}, u)$.

# References


[Adl+15a] J. H. Adler, T. J. Atherton, T. Benson, D. B. Emerson, and S. P. MacLachlan. “Energy minimization for liquid crystal equilibrium with electric and flexoelectric effects”. In: *SIAM J. Sci. Comput.* 37.5 (2015), S157–S176.

[Adl+15b] 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”. In: *SIAM J. Numer. Anal.* 53.5 (2015), pp. 2226–2254.

[Adl+16] J. H. Adler, T. J. Atherton, D. B. Emerson, and S. P. Maclachlan. “Constrained optimization for liquid crystal equilibria”. In: *SIAM J. Sci. Comput.* 38.1 (2016), pp. 50–76.

[AGL05] J. Ahrens, B. Geveci, and C. Law. “ParaView: an end-user tool for large-data visualization”. In: *Visualization Handbook*. Ed. by C. D. Hansen and C. R. Johnson. Burlington: Butterworth-Heinemann, 2005, pp. 717–731.

[ADL00] P. R. Amestoy, I. Duff, and J.-Y. L’Excellent. “Multifrontal parallel distributed symmetric and unsymmetric solvers”. In: *Comput. Methods Appl. Mech. Eng.* 184.2–4 (2000), pp. 501–520.

[AFS68] J. H. Argyris, I. Fried, and D. W. Scharpf. “The TUBA family of plate elements for the matrix displacement method”. In: *Aeronaut. J.* 72 (1968), pp. 701–709.

[AFW00] D. N. Arnold, R. S. Falk, and R. Winther. “Multigrid in H(div) and H(curl)”. In: *Numer. Math.* 85 (2000), pp. 197–217.

[AE11] B. Averill and P. Eldredge. In: *General Chemistry: Principles, Patterns, and Applications*. Minneapolis: Saylor Academy, 2011. Chap. 11.

[Bal+18] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, K. Rupp, B. F. Smith, and H. Zhang. *PETSc users manual*. Tech. rep. ANL-95/11 - Revision 3.9. Argonne National Laboratory, 2018.

[Bal17] J. M. Ball. “Mathematics and liquid crystals”. In: *Mol. Cryst. Liq. Cryst.* 647.1 (2017), pp. 1–27.

[BB15] J. M. Ball and S. J. Bedford. “Discontinuous order parameters in liquid crystal theories”. In: *Mol. Cryst. Liq. Cryst.* 612.1 (2015), pp. 1–23.

[BZ08] J. M. Ball and A. Zarnescu. “Orientable and non-orientable line field models for uniaxial nematic liquid crystals”. In: *Mol. Cryst. Liq. Cryst.* 495 (1 2008), 221/[573]–233/[585].

[Bed14] S. J. Bedford. "Calculus of variations and its application to liquid crystals". Ph.D thesis. University of Oxford, 2014.

[BGL05] M. Benzi, G. H. Golub, and J. Liesen. "Numerical solution of saddle point problems". In: *Acta Numer.* 14 (2005), pp. 1–137.

[BO06] M. Benzi and M. A. Olshanskii. "An augmented Lagrangian-based approach to the Oseen problem". In: *SIAM J. Sci. Comput.* 28.6 (2006), pp. 2095–2113.

[BCT07] P. Biscari, M. C. Calderer, and E. Terentjev. "Landau–de Gennes theory of isotropic-nematic-smectic liquid crystal transitions". In: *Phys. Rev. E* 75 (5 2007), pp. 051707-1–051707-11.

[Bis+19] K. Bisht, V. Banerjee, P. Milewski, and A. Majumdar. "Magnetic nanoparticles in a nematic channel: a one-dimensional study". In: *Phys. Rev. E* 100.012703 (2019), pp. 012703-1–012703-9.

[BB80] H. Blum and R. R. Bonn. "On the boundary value problem of the biharmonic operator on domains with angular corners". In: *Math. Mech. in the Appli. Sci.* 2 (1980), pp. 556–581.

[BNW20] 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". In: *Numer. Math.* 145.4 (2020), pp. 837–881.

[Bra06] A. Braides. "A Handbook of Γ-Convergence". In: *Handbook of Differential Equations: Stationary Partial Differential Equations*. Vol. 3. North-Holland, Amsterdam: Elsevier, 2006, pp. 101–213.

[Bre03] S. C. Brenner. "Poincaré-Friedrichs inequalities for piecewise $H^1$ functions". In: *SIAM J. Numer. Anal.* 41 (2003), pp. 306–324.

[Bre11] S. C. Brenner. "$C^0$ Interior Penalty Methods". In: *Frontiers in Numerical Analysis - Durham 2010. Lecture Notes in Computational Science and Engineering*. Ed. by J. Blowey and M. Jensen. Vol. 85. Berlin, Heidelberg: Springer, 2011.

[BGS10] S. C. Brenner, T. Gudi, and L. Sung. "A weakly over-penalized symmetric interior penalty method for the biharmonic problem". In: *Electon. Trans. Numer. Anal.* 37 (2010), pp. 214–238.

[BS08a] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. 3rd. Vol. 15. Texts in Applied Mathematics. Springer, New York, 2008.

[BS05] S. C. Brenner and L. Sung. "$C^0$ interior penalty methods for fourth order elliptic boundary value problems on polygonal domains". In: *J. Sci. Comput.* 22 (2005), pp. 83–118.

[BS08b] S. C. Brenner and L. Sung. "A weakly over-penalized symmetric interior penalty method". In: *Electon. Trans. Numer. Anal.* 30 (2008), pp. 107–127.

[BG70] F. Brochard and P. G. de Gennes. "Theory of magnetic suspensions in liquid crystals". In: *J. Phys. France* 31.7 (1970), pp. 691–708.

[Bru+15] P. R. Brune, M. G. Knepley, B. F. Smith, and X. Tu. "Composing scalable nonlinear algebraic solvers". In: *SIAM Rev.* 57.4 (2015), pp. 535–565.

[BR95] S. V. Burylov and Y. L. Raikher. "Macroscopic Properties of Ferronematics Caused by Orientational Interactions on the Particle Surfaces. I. Extended Continuum Model". In: *Mol. Cryst. Liq. Cryst. Sci. Technol. Sect. A* 258 (1 1995), pp. 107–122.

[Cal+14] M. C. Calderer, A. DeSimone, D. Golovaty, and A. Panchenko. "An Effective Model for Nematic Liquid Crystal Composites with Ferromagnetic Inclusions". In: *SIAM J. Appl. Math.* 74 (2 2014), pp. 237–262.

[CP00] M. C. Calderer and P. Palffy-Muhoray. "Ericksen's bar and modeling of the smectic A-nematic phase transition". In: *SIAM J. Appl. Math.* 60.3 (2000), pp. 1073–1098.

[CMS19] G. Canevari, A. Majumdar, and A. Spicer. "Order reconstruction for nematics on squares and hexagons: a Landau–de Gennes study". In: *SIAM J. Appl. Math.* 77 (1 2019), pp. 267–293.

[CSL91] T. Carlsson, I. W. Stewart, and F. M. Leslie. "Theoretical studies of smectic C liquid crystals confined in a wedge. Stability considerations and Frederiks transitions". In: *Liq. Cryst.* 9.5 (1991), pp. 661–678.

[Cha92] S. Chandrasekhar. *Liquid Crystals*. 2nd. Cambridge University Press, 1992.

[CHH00] X. Cheng, W. Han, and H. Huang. "Some mixed finite element methods for the biharmonic equation". In: *J. Comp. Appl. Math.* 126 (2000), pp. 91–109.

[Cia78] P. G. Ciarlet. *The Finite Element for Elliptic Problems*. North-Holland, Amsterdam, New York, Oxford, 1978.

[Clé75] P. Clément. "Approximation by finite element functions using local regularization". In: *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge Anal. Numér.* 9.R-2 (1975), pp. 77–84.

[Dal+21] J. Dalby, P. E. Farrell, A. Majumdar, and J. Xia. *One-dimensional ferronematics in a channel: order reconstruction, bifurcations and multistability*. 2021. arXiv: 2102.06347. URL: https://arxiv.org/abs/2102.06347.

[Dav94] T. A. Davis. "Finite element analysis of the Landau–de Gennes minimization problem for liquid crystals in confinement". Ph.D thesis. Kent State University, 1994.

[DG98] T. A. Davis and JR. E. C. Gartland. "Finite element analysis of the Landau-de Gennes minimization problem for liquid crystals". In: *SIAM J. Numer. Anal.* 35 (1 1998), pp. 336–362.

[Den+20] A. Dener, A. Denchfield, H. Suh, T. Munson, J. Sarich, S. Wild, S. Benson, and L. Curfman McInnes. *Toolkit for Advanced Optimization (TAO) Users Manual*. Tech. rep. ANL/MCS-TM-322 - Revision 3.14. Argonne National Laboratory, 2020.

[DS11] D. Dunmur and T. Sluckin. *Soap, Science, and Flat-Screen TVs*. Oxford University Press, 2011.

[E97] W. E. "Nonlinear continuum theory of smectic-A liquid crystals". In: *Arch. Rational Mech. Anal.* 137 (1997), pp. 159–175.

[ESW14] H. C. Elman, D. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*. 2nd. Oxford University Press, Oxford, UK, 2014.

[Eme15] D. B. Emerson. "Advanced discretizations and multigrid methods for liquid crystal configurations". Ph.D thesis. Tufts University, 2015.

[Eme+18] 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". In: *Liq. Cryst.* 45.3 (2018), pp. 341–350.

[Eng+02] 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". In: *Comput. Methods Appl. Mech. Engrg.* 191 (2002), pp. 3669–3750.

[Eva10] L. C. Evans. *Partial Differential Equations*. 2nd. Vol. 19 of Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2010.

[Far17] P. E. Farrell. *Defcon*. https://bitbucket.org/pefarrell/defcon/src/master/. 2017.

[FBF15] P. E. Farrell, Á. Birkisson, and S. W. Funke. "Deflation techniques for finding distinct solutions of nonlinear partial differential equations". In: *SIAM J. Sci. Comput.* 37.4 (2015), A2026–A2045.

[Far+21] P. E. Farrell, M. G. Knepley, F. Wechsung, and L. Mitchell. "PCPATCH: software for the topological construction of multigrid relaxation methods". In: *ACM Trans. Math. Softw.* (2021). In press. arXiv: 1912.08516. URL: https://arXiv.org/abs/1912.08516.

[FMW19] P. E. Farrell, L. Mitchell, and F. Wechsung. "An augmented Lagrangian preconditioner for the 3D stationary incompressible Navier–Stokes equations at high Reynolds number". In: *SIAM J. Sci. Comput.* 41 (5 2019), A3073–A3096.

[Fir20] Firedrake-Zenodo. *Software used in 'Augmented Lagrangian preconditoners for the Oseen–Frank model of nematic and cholesteric liquid crystals'*. 2020. URL: https://doi.org/10.5281/zenodo.4249051.

[Fir21a] Firedrake-Zenodo. *Software used in 'One-dimensional ferronematics in a channel - order reconstruction, bifurcations and multistability'*. 2021. URL: https://doi.org/10.5281/zenodo.4449535.

[Fir21b] Firedrake-Zenodo. *Software used in 'Structural Transitions in Geometrically Frustrated Smectics'*. 2021. URL: https://doi.org/10.5281/zenodo.4441123.

[FG83] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. Vol. 15. Studies in Mathematics and Its Applications. Elsevier Science Ltd, 1983.

[Fra58] F. C. Frank. "Liquid crystals". In: *Faraday Discuss.* 25 (1958), pp. 19–28.

[Fri22] V. Friedel. "Les états mésomorphes de la matiére". In: *Ann. Phys.* 18 (1922), pp. 273–474.

[Gen69] P. G. de Gennes. “Phenomenology of short-range-order effects in the isotropic phase of nematic materials”. In: *Phys. Lett.* 30A.8 (1969), pp. 454–455.

[Gen72] P. G. de Gennes. “An analogy between superconductors and smectic A”. In: *Solid State Commun.* 10 (9 1972), pp. 753–756.

[Gen73] P. G. de Gennes. “Some remarks on the polymorphism of smectics”. In: *Mol. Cryst. Liq. Cryst.* 21 (1973), pp. 49–76.

[Gen74] P. G. de Gennes. *The Physics of Liquid Crystals*. Oxford University Press, Oxford, 1974.

[GR09] C. Geuzaine and J.-F. Remacle. “Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities”. In: *Int. J. Numer. Methods Eng.* 79.11 (2009), pp. 1309–1331.

[Gia83] M. Giaquinta. *Multiple Integrals in the Calculus of Variations and Nonlinear Elliptic Systems*. Princeton University Press, 1983.

[GR11] V. Girault and P. A. Raviart. *Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms*. 1st. Springer, 2011.

[GL89] R. Glowinski and P. Le Tallec. “Augmented Lagrangian Methods for the Solution of Variational Problems”. In: *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Studies in Applied Mathematics. SIAM, 1989. Chap. 3, pp. 45–121.

[GLP03] R. Glowinski, P. Lin, and X. B. Pan. “An operator-splitting method for a liquid crystal model”. In: *Comput. Phys. Commun.* 152.3 (2003), pp. 242–252.

[Gri85] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. 1st. Pitman Advanced Publishing Program, 1985.

[Guy65] R. J. Guyan. “Reduction of stiffness and mass matrices”. In: *AIAA J.* 3.2 (1965), p. 380.

[Hag89] W. W. Hager. “Updating the inverse of a matrix”. In: *SIAM Rev.* 31.2 (1989), pp. 221–239.

[HL74] B. I. Halperin and T. C. Lubensky. “On the analogy between smectic A liquid crystals and superconductors”. In: *Solid State Commun.* 14 (10 1974), pp. 997–1001.

[Han+15] J. Han, Y. Luo, W. Wang, P. Zhang, and Z. Zhang. “From microscopic theory to macroscopic theory: a systematic study on modeling for liquid crystals”. In: *Arch. Rational Mech. Anal.* 215 (2015), pp. 741–809.

[HVK18] X. He, C. Vuik, and C. M. Klaij. “Combining the augmented Lagrangian Preconditioner with the Simple Schur Complement Approximation”. In: *SIAM J. Sci. Comput.* 40.3 (2018), A1362–A1385.

[HR12] T. Heister and G. Rapin. “Efficient augmented Lagrangian-type preconditioning for the Oseen problem using Grad-Div stabilization”. In: *Int. J. Numer. Meth. Fl.* 71.1 (2012), pp. 118–134.

[HRV05] V. Hernandez, J. E. Roman, and V. Vidal. “SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems”. In: *ACM Transactions on Mathematical Software* 31.3 (2005), pp. 351–362.

[HTW09] Q. Hu, X. Tai, and R. Winther. “A saddle point approach to the computation of harmonic maps”. In: *SIAM J. Numer. Anal.* 47.2 (2009), pp. 1500–1523.

[Iro65] B. Irons. “Structural eigenvalue problems: elimination of unwanted variables”. In: *AIAA J.* 3.5 (1965), pp. 961–962.

[Joh+17] V. John, A. Linke, C. Merdon, M. Neilan, and L. Rebholz. “On the divergence constraint in mixed finite element methods for incompressible flows”. In: *SIAM Rev.* 59.3 (2017), pp. 492–544.

[Kes89] S. Kesavan. *Topics in Functional Analysis and Applications*. New York: John Wiley & Sons, 1989.

[KA00] P. Knabner and L. Angermann. *Numerik partieller Differentialgleichungen*. Springer-Verlag: Berlin, Heidelberg, New York, 2000.

[Lac+07] E. Lacaze, J.-P. Michel, M. Alba, and M. Goldmann. “Planar anchoring and surface melting in the smectic-A phase”. In: *Phys. Rev. E* 76.4 (2007), p. 041702.

[LS12] J. P. F. Lagerwall and G. Scalia. “A new era for liquid crystal research: Applications of liquid crystals in soft matter nano-, bio- and microtechnology”. In: *Curr. Appl. Phys* 12 (6 2012), pp. 1387–1412.

[Lam14] X. Lamy. “Bifurcation analysis in a frustrated nematic cell”. In: *J. Nonlinear Sci.* 24 (2014), pp. 1197–1230.

[Lee+07] Y. Lee, J. Wu, J. Xu, and L. Zikatanov. “Robust subspace correction methods for nearly singular systems”. In: *Math. Mod. Meth. Appl. S.* 17.11 (2007), pp. 1937–1963.

[LBK06] I. Lelidis, C. Blanc, and M. Klèman. “Optical and confocal microscopy observations of screw dislocations in smectic-A liquid crystals”. In: *Phys. Rev. E* 74.051710 (2006), pp. 1–5.

[Lin89] F. Lin. “Nonlinear theory of defects in nematic liquid crystals; phase transition and flow phenomena”. In: *Commun. Pur. Appl. Math.* 42.6 (1989), pp. 789–814.

[LR07] P. Lin and T. Richter. “An adaptive homotopy multi-grid method for molecule orientations of high dimensional liquid crystals”. In: *J. Comput. Phys.* 225.2 (2007), pp. 2069–2082.

[LT14] P. Lin and X. Tai. “An Augmented Lagrangian Method for the Microstructure of a Liquid Crystal Model”. In: *Modeling, Simulation and Optimization for Science and Technology*. Ed. by W. Fitzgibbon, Y. Kuznetsov, P. Neittaanmäki, and O. Pironneau. Vol. 34. Springer, Dordrecht, 2014. Chap. 7, pp. 123–137.

[LS91] A. Linhananta and D. E. Sullivan. “Phenomenological theory of smectic-A liquid crystals”. In: *Phys. Rev. A* 44.12 (1991), pp. 8189–8197.

[MMN20] R. R. Maity, A. Majumdar, and N. Nataraj. "Discontinuous Galerkin finite element methods for the Landau–de Gennes minimization problem of liquid crystals". In: *IMA J. Numer. Anal.* 00 (2020), pp. 1–34.

[MZ10] A. Majumdar and A. Zarnescu. "Landau-de Gennes theory of nematic liquid crystals: the Oseen–Frank limit and beyond". In: *Arch. Ration. Mech. Anal.* 196 (2010), pp. 227–280.

[MZ15] S. Mei and P. Zhang. "On a molecular based Q-tensor model for liquid crystals with density variations". In: *Multiscale Model. Simul.* 13.3 (2015), pp. 977–1000.

[Mer+13] A. Mertelj, D. Lisjak, M. Drofenik, and M. Čopič. "Ferromagnetism in suspensions of magnetic platelets in liquid crystals". In: *Nature* 504 (2013), pp. 237–241.

[Mic+04] J.-P. Michel, E. Lacaze, M. Alba, M. de Boissieu, M. Gailhanou, and M. Goldmann. "Optical gratings formed in thin smectic films frustrated on a single crystalline substrate". In: *Phys. Rev. E* 70.011709 (2004), pp. 1011709-1–011709-12.

[MLG06] J.-P. Michel, E. Lacaze, and M. Goldmann. "Structure of Smectic Defect Cores: X-Ray Study of 8CB Liquid Crystal Ultrathin Films". In: *Phys. Rev. Lett.* 96.2 (2006), p. 027803.

[MN14] N. J. Mottram and C. J. P. Newton. *Introduction to Q-tensor theory*. 2014. arXiv: 1409.3502. URL: https://arxiv.org/abs/1409.3542.

[NW99] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[OU06] H. Ogawa and N. Uchida. "Numerical simulation of the twist-grain-boundary phase of chiral liquid crystals". In: *Phys. Rev. E* 73 (6 2006), pp. 060701-1–060701-4.

[Ols02] M. A. Olshanskii. "A low order Galerkin finite element method for the Navier–Stokes equations of steady incompressible flow: a stabilization issue and iterative methods". In: *Comput. Method. Appl. M.* 191.47 (2002), pp. 5515–5536.

[Ose33] C. W. Oseen. "The theory of liquid crystals". In: *Trans. Faraday Soc.* 29.140 (1933), pp. 883–899.

[PSS14] M. Y. Pevnyi, J. Selinger, and T. J. Sluckin. "Modeling smectic layers in confined geometries: order parameter and defects". In: *Physics Review E* (2014), pp. 1–8.

[PT74] V. T. Polyak and N. V. Tret'yakov. "The method of penalty estimates for conditional extremum problems". In: *USSR Comput. Math. Math. Phys.* 13.1 (1974), pp. 42–58.

[PS91] A. Poniewierski and T. J. Sluckin. "Phase diagram for a system of hard spherocylinders". In: *Phys. Rev. A* 43.12 (1991), pp. 6837–6842.

[Rat+17] F. Rathgeber, D. A. Ham, L. Mitchell, M. Lange, F. Luporini, A. T. T. McRae, G. T. Bercea, G. R. Markall, and P. H. J. Kelly. "Firedrake: automating the finite element method by composing abstractions". In: *ACM T. Math. Software* 43.24 (2017), pp. 1–27.

[RCB70] J. Rault, P. E. Cladis, and J. P. Burger. "Ferronematics". In: *Phys. Lett. A* 32 (3 1970), pp. 199–200.

[Rei88] F. Reinitzer. "Beiträge zur Kenntnis des Cholesterins". In: *Monatsh. Chem.* 9 (1888), pp. 421–441.

[Saa93] Y. Saad. "A flexible inner-outer preconditioned GMRES algorithm". In: *SIAM J. Sci. Comput.* 14.2 (1993), pp. 461–469.

[SS86] Y. Saad and M. Schultz. "GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems". In: *SIAM J. Sci. Statist. Comput.* 7.3 (1986), pp. 856–869.

[SK07] C. D. Santangelo and R. D. Kamien. "Triply periodic smectic liquid crystals". In: *Phys. Rev. E* 75.011702 (2007), pp. 011702-1–011702-12.

[Sch99a] J. Schöberl. "Multigrid methods for a parameter dependent problem in primal variables". In: *Numer. Math.* 84.1 (1999), pp. 97–119.

[Sch99b] J. Schöberl. "Robust multigrid methods for parameter dependent problems". Ph.D thesis. Johannes Kepler University Linz, 1999.

[Sch78] R. Scholtz. "A mixed method for fourth-order problems using linear finite elements". In: *RAIRO Numer. Anal.* 15 (1978), pp. 85–90.

[Sci18] SciencebyDegrees. *An introduction to liquid crystals*. 2018. URL: https://sciencebydegrees.com/2018/08/10/liquid-crystals/.

[SW94] D. Silvester and A. J. Wathen. "Fast iterative solution of stabilised Stokes systems. Part II: Using general block preconditioners". In: *SIAM J. Numer. Anal.* 31 (1994), pp. 1352–1367.

[Ste02] G. W. Stewart. "A Krylov–Schur algorithm for large eigenproblems". In: *SIAM Journal on Matrix Analysis and Applications* 23.3 (2002), pp. 601–614.

[Ste04] I. W. Stewart. *The Static and Dynamic Continuum Theory of Liquid Crystals: A Mathematical Introduction*. CPC Press, 2004.

[SM07] E. Süli and I. Mozolevski. "*hp*-version interior penalty DGFEMs for the biharmonic equation". In: *Comput. Methods Appl. Mech. Engrg.* 196 (2007), pp. 1851–1863.

[Van86] S. P. Vanka. "Block-implicit multigrid calculation of two-dimensional recirculating flows". In: *Comput. Method. Appl. M.* 59.1 (1986), pp. 29–48.

[Wal20] H. G. Walton. *Electro-optical effects in liquid crystals*. https://www.britannica.com/technology/liquid-crystal-display/Projection-displays. July 2020.

[WCM19] Y. Wang, G. Canevari, and A. Majumdar. "Order reconstruction for nematics on squares with isotropic inclusions: a Landau–de Gennes study". In: *SIAM J. Appl. Math.* 79 (4 2019), pp. 1314–1340.

[WS91] A. J. Wathen and D. Silvester. "Fast iterative solution of stabilised Stokes systems. Part I: Using simple diagonal preconditioners". In: *SIAM J. Numer. Anal.* 30.3 (1991), pp. 630–649.

[WK75] C. Williams and M. Kléman. "Dislocations, grain boundaries and focal conics in smectics A". In: *J. Phys. Colloq.* 36.C1 (1975), pp. C1-315–C1-320.

[Wit+21] R. Wittmann, L. B. G. Cortes, H. Löwen, and D. G. A. L. Aarts. "Particle-resolved topological defects of smectic colloidal liquid crystals in extreme confinement". In: *Nat. Comm.* 12.623 (2021), pp. 1–20.

[Xia21a] J. Xia. *Structural Transitions in Geometrically Frustrated Smectics (Supplementary Materials)*. Jan. 2021. URL: https://www.youtube.com/playlist?list=PLr8tas_dt-wwd81QWCyNZe51L7NckSfwo.

[XFW21] J. Xia, P. E. Farrell, and F. Wechsung. "Augmented Lagrangian preconditioners for Oseen–Frank models in nematic and cholesteric liquid crystals". In: *BIT Numer. Math.* (2021), pp. 1–38. URL: https://doi.org/10.1007/s10543-020-00838-9.

[Xia+21] J. Xia, S. MacLachlan, T. J. Atherton, and P. E. Farrell. "Structural landscapes on geometrically frustrated smectics". In: *Phys. Rev. Lett.* 126 (17 2021), pp. 177801-1–1779801-6.

[Xu92] J. Xu. "Iterative methods by space decomposition and subspace correction". In: *SIAM Rev.* 34.4 (1992), pp. 581–613.

[ZL08] B. Zappone and E. Lacaze. "Surface-frustrated periodic textures of smectic-A liquid crystals on crystalline surfaces". In: *Phys. Rev. E* 061704 (78 2008), pp. 061704-1–061704-9.

[Zap+20] B. Zappone, A. E. Mamuk, I. Gryn, V. Arima, A. Zizzari, R. Bartolino, E. Lacaze, and R. Petschek. "Analogy between periodic patterns in thin smectic liquid crystal films and the intermediate state of superconductors". In: *Proc. Natl. Acad. Sci. U.S.A.* 117.30 (2020), pp. 17643–17649.

[Xia21b] J. Xia. *Ferronematics-numerics*. 2021. URL: https://doi.org/10.5281/zenodo.4616745.

[Xia20] J. Xia. *ALpaper-numerics*. 2020. URL: https://doi.org/10.5281/zenodo.4257094.

[Xia21c] J. Xia. *Smectic-A numerics*. 2021. URL: https://doi.org/10.5281/zenodo.4607849.

[Zha09] S. Zhang. "A family of 3D continuously differentiable finite elements on tetrahedral grids". In: *Appl. Numer. Math.* 59 (1 2009), pp. 219–233.