Report Number 11/30

Numerical Study of Liquid Crystal Elastomers by a Mixed Finite Element Method

by

Chong Luo and Maria-Carme Calderer

Oxford Centre for Collaborative Applied Mathematics
Mathematical Institute
24 - 29 St Giles'
Oxford
OX1 3LB
England

Numerical Study of Liquid Crystal Elastomers by a Mixed Finite Element Method

Chong Luo *
Oxford Center for Collaborative Applied Mathematics
Mathematical Institute
University of Oxford
Oxford, OX1 3LB, United Kingdom,
and

Maria-Carme Calderer †
School of Mathematics
University of Minnesota
Minneapolis, MN 55455, USA.

July 6, 2011

Abstract

Liquid crystal elastomers (LCE) present features not found in ordinary elastic materials, such as semi-soft elasticity and the related stripe domain phenomenon. In this paper, the two-dimensional (2D) Bladon-Terentjev-Warner (BTW) model and the one-constant Oseen-Frank energy expression are combined to study the liquid crystal elastomer. We also impose two material constraints, the incompressibility of the elastomer and the unit director norm of the liquid crystal. We prove existence of minimizer of the energy for the proposed model. Next, we formulate the discrete model, and also prove that it possesses a minimizer of the energy. The inf-sup values of the discrete linearized system are then related to the smallest singular values of certain matrices. Next, the existence and uniqueness of the Lagrange multipliers associated with the two material constraints are proved under the assumption that the inf-sup conditions hold. Finally, numerical simulations of the clamped-pulling experiment are presented for elastomer samples with aspect ratio 1 or 3. The semi-soft elasticity is successfully recovered in both cases. The stripe domain phenomenon, however, is not observed, which might be due to the relative coarse mesh employed in the numerical experiment. Possible improvements are discussed which might lead to the recovery of the stripe domain phenomenon.

1 Introduction

A liquid crystal elastomer is an elastic material containing nematic liquid crystal molecules. Rotation of these liquid crystal molecules may lead to unique mechanical, optical and electrical properties, which are not observed in ordinary elastic materials. By properly exploiting these special properties, one might be able to manufacture new devices such as artificial muscles [23].

The stripe domain formation and the semi-soft elasticity property are two special phenomena exhibited by liquid crystal elastomers [34]. They have been observed in the *clamped-pulling* experiment, in which a piece of rectangular LCE is clamped and pulled in the direction perpendicular to the initial uniform

*This publication was based on work supported in part by Award No KUK-C1-013-04, made by King Abdullah University of Science and Technology(KAUST). This publication was partially supported by the National Science Foundation, grant number: DMS-FRG-0456232 and DMS 1009181.

†This publication was partially supported by the National Science Foundation, grant number: DMS-FRG-0456232 and DMS 1009181.

orientation of the liquid crystal directors. The stripe domain phenomenon refers to the formation of stripes with alternating director angles during the pulling process. In each stripe, the directors align along the same direction, while the directors in adjacent stripes align symmetrically about their middle line. For square shaped polysiloxane LCE, it is observed [25, 35] that during the pulling process, the stripe domain first occurred in the center of the domain, and then broke into two which then migrated towards the two clamped ends. The semi-soft elasticity refers to the unusual stress-strain relationship during the pulling process. The LCE is first *hard*, in a regime such that the stress grows almost linearly with strain; then it reaches the *soft* regime, in which the stress remains almost constant while the strain increases; then the LCE becomes hard again upon further increase of the strain [12, 26].

Several models [3, 33, 7, 18, 6] have been proposed to explain the special behaviors of LCE. A very successful one is that proposed by Bladon, Terentjev and Warner (BTW)[3] which predicts the stripe domain phenomenon and the soft elastic response of the elastomer. However, the stress-strain relationship computed with the BTW model is *ideally soft* [34]. That is, it lacks the initial hard regime typically observed in experiments. Several approaches have been proposed to modify the BTW and fully capture the experimental results. The Verwey-Warner-Terentjev (VWT) model[33] extends the BTW model by adding a term related to the cross-linking state, and successfully recovers the semi-soft phenomenon. Other models[7, 18, 6] extend the BTW model by adding liquid crystal elastic energy terms, such as Oseen-Frank[21], Ericksen[19] and Landau-de Gennes energy terms [15]. Unlike the VWT model, the latter involve first derivatives of the director field.

There are, however, relatively few works in the literature about the numerical simulation of liquid crystal elastomers. This may be due to the complexity caused by the coupling between the displacement of the bulk and the orientation of the liquid crystal directors. An important work on numerical simulation of LCE is that of Conti, DeSimone and Dolzmann [14], who did 3D finite element simulation based on the BTW model. They eliminated the orientation field $\boldsymbol{n}$ from the BTW model by taking it to be the minimizer of the energy for fixed displacement field. The resulting energy as a functional of displacement field was non-convex, and so, they took its polyconvex envelope as the energy functional to analyze. Their simulation successfully recovered the stripe domain phenomenon. However, their model inherited the features of the BTW model, and only recovered ideally soft elasticity. In later work, the same authors applied a similar approach to the VWT model, and did successfully recover the semi-soft elasticity property [13].

In this paper, we extend the BTW energy by adding the Oseen-Frank energy, and do finite element simulation for the full model on a 2D rectangular domain. We use the clamped-pulling as a benchmark problem to check whether our numerical method can recover the special behaviors of LCE such as stripe domain phenomenon and semi-soft elasticity.

The paper is organized as follows. In section 2, we list the notations that we employ. In section 3, we investigate the continuous problem, and proceed to study the discrete one in section 4. In section 5, we present the numerical results, and in section 6, we give the conclusions of the work. Finally, in section 7, we discuss possible improvements of current work to fully capture the stripe domain phenomenon.

2 Notations

In this paper, in addition to the standard notations of Sobolev spaces [1], such as $W^{m,p}$, L^2, H^1, H^1_0 and H^{-1}, we let

$$H^1_{0|\Gamma}(\Omega) = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma \subset \partial\Omega\}. \tag{1}$$

We use $H^1_{g|\Gamma}(\Omega)$ to denote $g + H^1_{0|\Gamma}(\Omega)$, and $\mathbf{H}^1_{\boldsymbol{g}|\Gamma}(\Omega)$ to represent its vector version. We use $H^{-1}_{\Gamma}(\Omega)$ to denote the dual space of $H^1_{0|\Gamma}(\Omega)$. We let $\langle\cdot,\cdot\rangle$ represent the dot product of two vectors, the inner product in Hilbert spaces, or the action of a linear functional on a function. Its actual meaning is made clear by the context.

We let $\mathbb{M}^{m\times n}$ denote the space of real $m \times n$ matrices. For any matrix F, we denote its transpose by F^T. For any square matrix A, $\det(A)$ denotes the determinant, $\mathrm{tr}(A)$ the trace, and $\mathrm{cof}(A)$ represents the cofactor matrix, whose (i, j) entry is equal to $(-1)^{i+j}$ times the determinant of the submatrix obtained by eliminating row i and column j of the matrix A. For any matrix F, we use $\frac{\partial \det}{\partial F}$ to denote $\frac{\partial \det(F)}{\partial F}$, which can be shown to be exactly $\mathrm{cof}(F)$. We let $A : B$ represent the inner product of the two matrices, that is,

$$A : B = \mathrm{tr}(A^T B) = \sum_{i,j} A_{ij} B_{ij}.$$

For any matrix F, we let $|F|$ denote its Frobenius norm, that is

$$|F| = (F : F)^{1/2}.$$

3 Existence and well-posedness of the continuous problem

In this section, we present our analytical results on the continuous problem. We first introduce the energy functional and prove existence of minimizer. Then we derive the Euler-Lagrange equations for the energy, and obtain the corresponding linearized system. Finally, we reduce the linearized system to a standard saddle point framework and discuss its well-posedness.

3.1 The energy functional and the minimization problem

In this subsection, we introduce the energy density as a combination of 2D BTW energy and the Oseen-Frank energy, and prove existence of minimizer.

Throughout this paper, the analysis and computation are carried out in two-dimensional domains. This is justified as follows. In the experiment by Finkelmann et al. [25, 35], the elastomer has a very thin, rectangular shape, and, consequently, it can be assumed that the director vectors lie in the same plane. If the elastomer were compressed, it might buckle, with the directors tilting out of the plane. However, in the pulling experiment, the elastomer sample remains planar. If in addition, we look for configurations with director field confined in that plane, then a 2D model and analysis may be appropriate.

We use $\boldsymbol{X} = (X_1, X_2)^T$ to denote a point in the reference configuration, and $\boldsymbol{x}(\boldsymbol{X}) = (x_1, x_2)^T$ the corresponding point in the deformed configuration. We define the displacement field as $\boldsymbol{u}(\boldsymbol{X}) = \boldsymbol{x} - \boldsymbol{X}$. The deformation gradient tensor is $F = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{X}}$, and satisfies $F = I + \nabla \boldsymbol{u}$. We assume the elastomer is incompressible, thus F satisfies $\det(F) = 1$. We denote the director field as $\boldsymbol{n}(\boldsymbol{X}) = (n_1, n_2)^T$, which is a unit vector representing the average orientation of the relevant liquid crystal molecular units at each point.

The BTW stored energy for LCE, in the form derived by DeSimone et al. [16, 17, 18], can be written as

$$W_{BTW} = \mu \left(|F|^2 - (1-a)|F^T \boldsymbol{n}|^2 - 2a^{1/2} \right), \tag{2}$$

where μ is an elasticity constant. The dimensionless constant a satisfies $0 < a < 1$, and is a measurement of interaction between the bulk displacement $\boldsymbol{u}$ and the director orientation $\boldsymbol{n}$. In the limit $a \to 1$, the BTW model degenerates to the neo-Hookean model, and there is no interaction between $\boldsymbol{u}$ and $\boldsymbol{n}$. On the other hand, in the limit $a \to 0$, there is maximum interaction between $\boldsymbol{u}$ and $\boldsymbol{n}$. Note that the reference configuration (the one with $\boldsymbol{u} = 0$) for (2) is not the stress-free state. If we take the stress-free state as the reference state, the BTW energy will be in a slightly more complicated form. We will elaborate on this issue in the coming sections.

The Oseen-Frank stored energy [21], in its simplest form, can be written as

$$W_{OF} = b|\nabla \boldsymbol{n}|^2. \tag{3}$$

The energy (3) penalizes change in the director field, and the prescribed constant $b > 0$ measures the strength of the penalization.

The non-dimensionalized energy functional is the following,

$$\begin{aligned} \Pi(\boldsymbol{u}, \boldsymbol{n}) = & \int_\Omega \left(|F|^2 - (1-a)|F^T \boldsymbol{n}|^2 \right) + b|\nabla \boldsymbol{n}|^2 \\ & - \int_\Omega \boldsymbol{f} \cdot \boldsymbol{u} - \int_\Gamma \boldsymbol{g} \cdot \boldsymbol{u} da, \end{aligned} \tag{4}$$

where $\boldsymbol{f}$ is a prescribed body force, and $\boldsymbol{g}$ an applied boundary traction on $\Gamma \subset \partial\Omega$. The admissible set for the displacement $\boldsymbol{u}$ is

$$\mathcal{K} = \{\boldsymbol{u} \in H^1(\Omega, \mathbb{R}^2) : \det(I + \nabla \boldsymbol{u}) = 1 \text{ a.e. in } \Omega,\ \boldsymbol{u} = \boldsymbol{u}_0 \text{ on } \Gamma_u \subset \partial\Omega\}, \tag{5}$$

and, the admissible set for the director $\boldsymbol{n}$ is

$$\mathcal{N} = \{\boldsymbol{n} \in H^1(\Omega, \mathbb{R}^2) : |\boldsymbol{n}| = 1 \text{ a.e. in } \Omega,\ \boldsymbol{n} = \boldsymbol{n}_0 \text{ on } \Gamma_n \subset \partial\Omega\}. \tag{6}$$

We let $\mathcal{A} = \mathcal{K} \times \mathcal{N}$. The admissible set $\mathcal{A}$ is non-empty as long as $\boldsymbol{u}_0$ and $\boldsymbol{n}_0$ are both Lipschitz continuous functions [22].

The problem of energy minimization is formulated as follows:

$$\text{Find } (\boldsymbol{u}, \boldsymbol{n}) \in \mathcal{A} \text{ minimizing } \Pi(\boldsymbol{u}, \boldsymbol{n}) \text{ in } \mathcal{A}. \tag{7}$$

Next, we prove existence of minimizer for (7).

Prior to the proof of existence, we summarize several lemmas, some of them well-known in the literature. We include them here for the purpose of self-consistency.

Lemma 1. *Assume* $0 < a < 1$, $|\boldsymbol{n}| = 1$ *and* $\det(F) = 1$. *Then the BTW energy (2) is always non-negative. It is zero if and only if* $\mathrm{eig}(FF^T) = \{a^{1/2}, a^{-1/2}\}$ *and* $\boldsymbol{n}$ *is an eigenvector corresponding to the eigenvalue* $a^{-1/2}$.

Proof. Let the eigenvalues of F^TF be λ_1^2 and λ_2^2 and satisfy $0 \le \lambda_1^2 \le \lambda_2^2$, and let $\boldsymbol{v}_1, \boldsymbol{v}_2$ denote the corresponding (unit) eigenvectors. Also assume that

$$\boldsymbol{n} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2.$$

Since $|\boldsymbol{n}| = 1$, and $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthonormal, we have

$$\alpha_1^2 + \alpha_2^2 = 1.$$

Thus, we can rewrite the BTW energy as

$$\begin{aligned} W_{BTW} &= \mu\left(\mathrm{tr}(FF^T) - (1-a)\boldsymbol{n}^T FF^T \boldsymbol{n} - 2a^{1/2}\right) \\ &= \mu\left(\lambda_1^2 + \lambda_2^2 - (1-a)(\alpha_1^2\lambda_1^2 + \alpha_2^2\lambda_2^2) - 2a^{1/2}\right). \end{aligned}$$

Since $0 \le \lambda_1^2 \le \lambda_2^2$, the values $\alpha_2^2 = 1$, and $\alpha_1^2 = 0$ minimize W_{BTW}. So, $\boldsymbol{n}$ is parallel to $\boldsymbol{v}_2$. Consequently,

$$\begin{aligned} W_{BTW} &\ge \mu\left(\lambda_1^2 + \lambda_2^2 - (1-a)\lambda_2^2 - 2a^{1/2}\right) \\ &= \mu\left(\lambda_1^2 + a\lambda_2^2 - 2a^{1/2}\right). \end{aligned}$$

Since $\det(FF^T) = 1$, we have that

$$\lambda_1^2 \lambda_2^2 = 1. \tag{8}$$

Therefore,

$$\begin{aligned} W_{BTW} &\ge \mu\left(2\sqrt{\lambda_1^2 \cdot a\lambda_2^2} - 2a^{1/2}\right) \\ &= \mu\left(2a^{1/2} - 2a^{1/2}\right) \\ &= 0, \end{aligned}$$

hold. The equality is satisfied if and only if $\lambda_1^2 = a\lambda_2^2$. Combining it with equation (8) yields

$$\mathrm{eig}(FF^T) = \{a^{1/2}, a^{-1/2}\}. \tag{9}$$

□

Lemma 2. *Assume* $|\boldsymbol{n}| = 1$ *and* $0 < a < 1$. *Then*

$$|F|^2 - (1-a)|F^T\boldsymbol{n}|^2 \ge a|F|^2. \tag{10}$$

Proof. Let λ_1^2 and λ_2^2 be as in Lemma 1. Since $|\boldsymbol{n}| = 1$, by the proof of Lemma 1, we have

$$\begin{aligned} |F|^2 - (1-a)|F\boldsymbol{n}|^2 &\ge \lambda_1^2 + a\lambda_2^2 \\ &\ge a(\lambda_1^2 + \lambda_2^2) \\ &= a\mathrm{tr}(FF^T) \\ &= a|F|^2. \end{aligned}$$

□

Lemma 3. *Assume* $|\boldsymbol{n}| = 1$, *and* $0 < a < 1$. *The function*

$$L(F) = |F|^2 - (1-a)|F^T\boldsymbol{n}|^2 \tag{11}$$

is convex with respect to F.

Proof. Let

$$A(\boldsymbol{n}) = I - (1-a)\boldsymbol{n}\boldsymbol{n}^T. \tag{12}$$

So,

$$L = \mathrm{tr}(FAF^T). \tag{13}$$

For any matrices $F_1, F_2 \in \mathbb{M}^{2\times 2}$, and $0 \le \alpha \le 1$, we have

$$\begin{aligned}
&[\alpha L(F_1) + (1-\alpha)L(F_2)] - L(\alpha F_1 + (1-\alpha)F_2) \\
&= \alpha \mathrm{tr}(F_1 A F_1^T) + (1-\alpha)\mathrm{tr}(F_2 A F_2^T) \\
&\quad - \mathrm{tr}[(\alpha F_1 + (1-\alpha)F_2)A(\alpha F_1 + (1-\alpha)F_2)^T] \\
&= \alpha(1-\alpha)\mathrm{tr}[(F_1 - F_2)A(F_1 - F_2)] \\
&= \alpha(1-\alpha)\left(|F_1 - F_2|^2 - (1-a)|(F_1 - F_2)^T\boldsymbol{n}|^2\right) \\
&\ge \alpha(1-\alpha)a|F_1 - F_2|^2 \\
&\ge 0,
\end{aligned}$$

where we have used Lemma 2. Hence, the result follows. □

Next we quote the following theorem on nonlinear elasticity.

Theorem 4 (Ball, [2])**.** *Let Ω be a nonempty, bounded, open subset of $\mathbb{R}^d$.*

- *If $d = 2$, suppose we have $\boldsymbol{u}_k \rightharpoonup \boldsymbol{u}$ in $W^{1,s}$ with $s > \frac{4}{3}$, then we have $\det(I + \nabla \boldsymbol{u}_k) \to \det(I + \nabla \boldsymbol{u})$ in $\mathcal{D}'(\Omega)$;*
- *If $d = 3$,*
  - *suppose we have $\boldsymbol{u}_k \rightharpoonup \boldsymbol{u}$ in $W^{1,s}$ with $s > \frac{3}{2}$, then we have $\mathrm{adj}(I + \nabla \boldsymbol{u}_k)_{ij} \to \mathrm{adj}(I + \nabla \boldsymbol{u})_{ij}$ in $\mathcal{D}'(\Omega)$;*
  - *suppose we have $\boldsymbol{u}_k \rightharpoonup \boldsymbol{u}$ in $W^{1,s}$, and $\mathrm{adj}(I+\nabla \boldsymbol{u}_k) \rightharpoonup \mathrm{adj}(I+\nabla \boldsymbol{u})$ in $L^q(\Omega; \mathbb{M}^3)$ with $s > 1, q > 1$ and $\frac{1}{s} + \frac{1}{q} < \frac{4}{3}$, then we have $\det(I + \nabla \boldsymbol{u}_k) \to \det(I + \nabla \boldsymbol{u})$ in $\mathcal{D}'(\Omega)$.*

Now we are ready to prove the main theorem of this section.

Theorem 5. *There exists solution to the problem (7).*

Proof. Let m be the infimum of Π in $\mathcal{A}$, and let $(\boldsymbol{u}_k, \boldsymbol{n}_k) \in \mathcal{A}$ be a minimizing sequence of Π. Note that $m < +\infty$. Thus $\Pi(\boldsymbol{u}_k, \boldsymbol{n}_k)$ is bounded above by some constant C. By Lemma 2,

$$\begin{aligned}
C \ge \Pi(\boldsymbol{u}_k, \boldsymbol{n}_k) \ge & \int_\Omega a|(I + \nabla \boldsymbol{u}_k)|^2 + b|\nabla \boldsymbol{n}_k|^2 dx \\
& - \|\boldsymbol{f}\|_{L^2(\Omega)}\|\boldsymbol{u}_k\|_{L^2(\Omega)} - \|\boldsymbol{g}\|_{L^2(\Gamma)}\|\boldsymbol{u}_k\|_{L^2(\Gamma)} \\
\ge & \int_\Omega a|(I + \nabla \boldsymbol{u}_k)|^2 + b|\nabla \boldsymbol{n}_k|^2 dx \\
& - \left(\frac{1}{\varepsilon}\|\boldsymbol{f}\|^2_{L^2(\Omega)} + \varepsilon\|\boldsymbol{u}_k\|^2_{L^2(\Omega)}\right) - \left(\frac{1}{\varepsilon}\|\boldsymbol{g}\|^2_{L^2(\Gamma)} + \varepsilon\|\boldsymbol{u}_k\|^2_{L^2(\Gamma)}\right) \\
\ge & \int_\Omega C_1|(I + \nabla \boldsymbol{u}_k)|^2 + C_2|\nabla \boldsymbol{n}_k|^2 dx - C_3,
\end{aligned} \tag{14}$$

where $\varepsilon > 0$ is small and $C_i > 0, i = 1, 2, 3$ are constants. In the last step, we have applied the generalized Poincaré inequality ([9], p281) and the Trace Theorem ([20], p258). By (14), $F_k = I + \nabla \boldsymbol{u}_k$ and $\nabla \boldsymbol{n}_k$ are bounded in L^2. Since $\nabla(\boldsymbol{u}_k - \boldsymbol{u}_0)$ is bounded in L^2, by the Poincaré inequality, $\boldsymbol{u}_k$ is bounded in H^1. On the other hand, since $\nabla \boldsymbol{n}_k$ is bounded in L^2 and $|\boldsymbol{n}_k| = 1$ a.e. in Ω, $\boldsymbol{n}_k$ is bounded in H^1. Now since H^1 is a reflexive Banach space and $\boldsymbol{u}_k$ and $\boldsymbol{n}_k$ are bounded in H^1, we can find a subsequence of $\boldsymbol{u}_k$ and a subsequence of $\boldsymbol{n}_k$ such that they are weakly convergent in H^1. We still denote them as $(\boldsymbol{u}_k, \boldsymbol{n}_k)$, and assume $\boldsymbol{u}_k \rightharpoonup \boldsymbol{u}$, $\boldsymbol{n}_k \rightharpoonup \boldsymbol{n}$.

Since $\boldsymbol{u}_k \rightharpoonup \boldsymbol{u}$ in H^1, by Theorem 4, $\det(I + \nabla \boldsymbol{u}_k) \to \det(I + \nabla \boldsymbol{u})$ in $\mathcal{D}'(\Omega)$. Moreover, since $\det(I + \nabla \boldsymbol{u}_k) = 1$ a.e., it follows that $\det(I + \nabla \boldsymbol{u}) = 1$ a.e. in Ω as well. [1]. On the other hand, weak convergence

[1] This is because, by definition,

$$\langle \det(I + \nabla \boldsymbol{u}_k), \phi\rangle \to \langle \det(I + \nabla \boldsymbol{u}), \phi\rangle$$

in $\mathbb{R}$, for any $\phi \in \mathcal{D}(\Omega)$. Since $\det(I + \nabla \boldsymbol{u}_k) = 1$ a.e. in Ω for any k,

$$\langle \det(I + \nabla \boldsymbol{u}) - 1, \phi\rangle = 0, \quad \forall \phi \in \mathcal{D}(\Omega)$$

holds. Thus $\det(I + \nabla \boldsymbol{u}) = 1$ a.e. in Ω.

in H^1 implies strong convergence in L^2 [2], thus we can find a subsequence of $\boldsymbol{n}_k$ that converges point-wise almost everywhere. Therefore we have $|\boldsymbol{n}| = 1$ a.e. in Ω. Finally since $\boldsymbol{u}_k - \boldsymbol{u}_0 \in H^1_{0|\Gamma_u}$, which is a closed linear subspace of H^1, by the Mazur's Theorem, it is weakly closed. Therefore $\boldsymbol{u} - \boldsymbol{u}_0$ is also in $H^1_{0|\Gamma_u}$, implying $\boldsymbol{u} = \boldsymbol{u}_0$ on Γ_u. Similarly, $\boldsymbol{n} = \boldsymbol{n}_0$ on Γ_n holds. Therefore, $(\boldsymbol{u}, \boldsymbol{n}) \in \mathcal{A}$.

By Lemma 3, the following function

$$L(F, \boldsymbol{n}, P) = \left(|F|^2 - (1-a)|F^T\boldsymbol{n}|^2\right) + b|P|^2$$

is a convex function of F and P. Therefore by Theorem 1 of section 8.2 of [20], Π is weakly lower semi-continuous. Thus,

$$\begin{aligned}\Pi(\boldsymbol{u}, \boldsymbol{n}) &\leq \liminf_{k\to\infty} \Pi(\boldsymbol{u}_k, \boldsymbol{n}_k) \\ &= m.\end{aligned}$$

Since m is the infimum of Π on $\mathcal{A}$, we conclude that

$$\Pi(\boldsymbol{u}, \boldsymbol{n}) = m. \tag{15}$$

That is, $(\boldsymbol{u}, \boldsymbol{n})$ is the minimizer of Π on $\mathcal{A}$. □

3.2 Equilibrium equation and stress-free state

In this section, we derive the weak form of the equilibrium equation satisfied by the energy minimizer. After discretization, this equilibrium equation can be regarded as a nonlinear equation satisfied by the degrees of freedom (DOF) of the energy minimizer, which can then be solved by a nonlinear solver such as Newton's method. We also discuss, in this subsection, the solution corresponding to the stress-free state.

A common way to convert a constrained minimization problem to an unconstrained one, is by the method of the Lagrange multipliers. In this way, the constraints are moved from the admissible set to the objective function. We gain the flexibility at the cost of solving a larger system. After introducing the Lagrange multipliers, the objective energy functional becomes

$$\begin{aligned}\mathcal{E}(\boldsymbol{u}, \boldsymbol{n}, p, \lambda) = &\int_\Omega \left(|F|^2 - (1-a)|F^T\boldsymbol{n}|^2\right) + b|\nabla \boldsymbol{n}|^2 \\ &- p(\det F - 1) + \lambda(|\boldsymbol{n}|^2 - 1) - \int_\Omega \boldsymbol{f}\cdot\boldsymbol{u} - \int_\Gamma \boldsymbol{g}\cdot\boldsymbol{u},\end{aligned} \tag{16}$$

where $p \in L^2(\Omega)$ is the Lagrange multiplier for the incompressibility constraint $\det(I + \nabla \boldsymbol{u}) = 1$, that can be interpreted as *pressure*, while $\lambda \in L^2(\Omega)$ is the Lagrange multiplier for the unity constraint $|\boldsymbol{n}| = 1$. The admissible set for $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda)$ is

$$\mathcal{S} = \mathbf{H}^1_{\boldsymbol{u}_0|\Gamma_u}(\Omega) \times \mathbf{H}^1_{\boldsymbol{n}_0|\Gamma_n}(\Omega) \times L^2(\Omega) \times L^2(\Omega). \tag{17}$$

Next we use variational principle to derive the weak form of the equilibrium equation, also known as the Euler-Lagrange equation. Assume $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda) \in \mathcal{S}$ minimizes the energy (16). Then for any test function $\boldsymbol{v} \in \mathbf{H}^1_{0|\Gamma_u}(\Omega)$, the 1D function $\mathcal{E}(\varepsilon) = \mathcal{E}(\boldsymbol{u} + \varepsilon\boldsymbol{v}, \boldsymbol{n}, p, \lambda)$ has a minimum at $\varepsilon = 0$. Thus, it follows that

$$0 = \left.\frac{d\mathcal{E}}{d\varepsilon}\right|_{\varepsilon=0},$$

which simplifies to the following equation

$$\begin{aligned}0 = &\int_\Omega 2\left(F : \nabla \boldsymbol{v} - (1-a)\langle F^T\boldsymbol{n}, \nabla \boldsymbol{v}^T\boldsymbol{n}\rangle\right) - p\frac{\partial \det}{\partial F} : \nabla \boldsymbol{v} \\ &- \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v} - \int_\Gamma \boldsymbol{g}\cdot\boldsymbol{v}da.\end{aligned}$$

[2] This is because the embedding $I : W^{1,p} \to L^p$ is compact for $1 \leq p \leq \infty$ ([20], p274), while for any compact operator $A : V \to W$ with V and W Banach spaces, $u_k \rightharpoonup u$ in V implies $Au_k \to Au$ in W ([9], Theorem 7.1-5 on p348).

Similarly by taking the variations $\boldsymbol{n} \to \boldsymbol{n} + \varepsilon \boldsymbol{m}$, $p \to p + \varepsilon q$, or $\lambda \to \lambda + \varepsilon \mu$, we obtain the following Euler-Lagrange equations

$$\begin{aligned} 0 = & \int_\Omega 2\left(F : \nabla \boldsymbol{v} - (1-a)\langle F^T \boldsymbol{n}, \nabla \boldsymbol{v}^T \boldsymbol{n}\rangle\right) - p\frac{\partial \det}{\partial F} : \nabla \boldsymbol{v} \\ & - \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} - \int_\Gamma \boldsymbol{g} \cdot \boldsymbol{v} da, \end{aligned} \tag{18}$$

$$0 = \int_\Omega -2(1-a)\langle F^T \boldsymbol{n}, F^T \boldsymbol{m}\rangle + 2b\, \nabla \boldsymbol{m} : \nabla \boldsymbol{n} + 2\lambda \langle \boldsymbol{n}, \boldsymbol{m}\rangle, \tag{19}$$

$$0 = \int_\Omega -q(\det F - 1), \tag{20}$$

$$0 = \int_\Omega \mu(\langle \boldsymbol{n}, \boldsymbol{n}\rangle - 1), \tag{21}$$

where the solution $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda)$ belongs to $\mathcal{S}$, while the test function $(\boldsymbol{v}, \boldsymbol{m}, q, \mu)$ is in the space $\mathbf{H}^1_{0|\Gamma_u} \times \mathbf{H}^1_{0|\Gamma_n} \times L^2(\Omega) \times L^2(\Omega)$.

The corresponding equations in the strong form are derived using integration by parts, as long as $\boldsymbol{u}$ and $\boldsymbol{n}$ are smooth enough. The resulting system of partial differential equations is

$$\mathrm{div}\sigma + \boldsymbol{f} = 0 \qquad \text{in } \Omega \tag{22}$$

$$b\mathrm{div}(\nabla \boldsymbol{n}) + (1-a)\boldsymbol{n}^T F F^T - \lambda \boldsymbol{n}^T = 0 \qquad \text{in } \Omega \tag{23}$$

$$\det F - 1 = 0 \qquad \text{in } \Omega \tag{24}$$

$$|\boldsymbol{n}|^2 - 1 = 0 \qquad \text{in } \Omega \tag{25}$$

$$\sigma \boldsymbol{\nu} = \boldsymbol{g} \qquad \text{on } \partial\Omega \backslash \Gamma_u \tag{26}$$

$$\frac{\partial \boldsymbol{n}}{\partial \boldsymbol{\nu}} = 0 \qquad \text{on } \partial\Omega \backslash \Gamma_n, \tag{27}$$

where $\boldsymbol{\nu}$ denotes the unit normal vector on the boundary, and

$$\sigma = 2\left(I - (1-a)\boldsymbol{n}\boldsymbol{n}^T\right) F - p\frac{\partial \det}{\partial F} \tag{28}$$

is the Piola-Kirchhoff stress tensor.

Note that when $\boldsymbol{f} = 0$ and $\boldsymbol{g} = 0$, the reference configuration $\boldsymbol{u} \equiv 0$ is not a stress-free state. The reason is as follows. At the reference configuration, $F = I$, so

$$\sigma = (2-p)I - 2(1-a)\boldsymbol{n}\boldsymbol{n}^T. \tag{29}$$

The matrix $(2-p)I$ has rank 0 or 2 according to whether $p = 2$ or $p \neq 2$, while the matrix $2(1-a)\boldsymbol{n}\boldsymbol{n}^T$ has rank 1 for $0 < a < 1$. Thus, the Cauchy stress σ cannot be zero [3].

However, the stress-free state can be achieved by a uniform stretch of the reference state. It is easy to check that the system has zero stress in the case that

$$F = \begin{pmatrix} a^{1/4} & 0 \\ 0 & a^{-1/4} \end{pmatrix}, \tag{30}$$

$\boldsymbol{n} \equiv (0,1)^T$, $p = 2\sqrt{a}$, and $\lambda = (1-a)/\sqrt{a}$.

3.3 Linearized system and well-posedness

In this section, we derive the linearized system of the equilibrium equations, and discuss its well-posedness. This system is closely related to the matrix derivative in Newton's method. Also, the well-posedness of the linearized system is closely related to the stability of the numerical scheme.

To linearize the original system, we fix a solution $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda)$, and letting $(\boldsymbol{w}, \boldsymbol{l}, o, \gamma)$ be a small perturbation, we carry out the corresponding Taylor expansions in equations (18)-(21) about the given solution,

[3] If $a = 1$, and $p = 2$, we indeed get zero stress. In this case, $p = 2$ corresponds to the hydrostatic pressure of a neo-Hookean material.

retaining the linear terms. The resulting linearized system is

$$a_1(\boldsymbol{w},\boldsymbol{v}) + a_2(\boldsymbol{l},\boldsymbol{v}) + b_1(o,\boldsymbol{v}) = L_1(\boldsymbol{v}), \tag{31}$$
$$a_2(\boldsymbol{m},\boldsymbol{w}) + a_3(\boldsymbol{l},\boldsymbol{m}) + b_2(\gamma,\boldsymbol{m}) = L_2(\boldsymbol{m}), \tag{32}$$
$$b_1(q,\boldsymbol{w}) = L_3(q), \tag{33}$$
$$b_2(\mu,\boldsymbol{l}) = L_4(\mu), \tag{34}$$

where a_1, a_2, a_3, b_1 and b_2 denote bilinear forms depending on the solution $(\boldsymbol{u},\boldsymbol{n},p,\lambda)$, and L_1, L_2, L_3 and L_4 are linear functionals of the test functions. The perturbation $(\boldsymbol{w},\boldsymbol{l},o,\gamma)$ is supposed to satisfy the linearized system (31)-(34) for any test function $(\boldsymbol{v},\boldsymbol{m},q,\mu)$. Both the perturbation and the test function belong to the space $\mathbf{H}^1_{0|\Gamma_u}\times\mathbf{H}^1_{0|\Gamma_n}\times L^2(\Omega)\times H^{-1}_{\Gamma_n}$. The bilinear forms are defined by the following equations:

$$\begin{aligned} a_1(\boldsymbol{w},\boldsymbol{v}) = \int_\Omega & 2\,\nabla\boldsymbol{w} : \nabla\boldsymbol{v} - 2(1-a)\langle \nabla\boldsymbol{w}^T\boldsymbol{n}, \nabla\boldsymbol{v}^T\boldsymbol{n}\rangle \\ & - p(\frac{\partial^2 \det}{\partial F^2}\nabla\boldsymbol{w}) : \nabla\boldsymbol{v}, \end{aligned} \tag{35}$$

$$a_2(\boldsymbol{m},\boldsymbol{v}) = \int_\Omega -2(1-a)\langle F^T\boldsymbol{m}, \nabla\boldsymbol{v}^T\boldsymbol{n}\rangle - 2(1-a)\langle F^T\boldsymbol{n}, \nabla\boldsymbol{v}^T\boldsymbol{m}\rangle, \tag{36}$$

$$a_3(\boldsymbol{l},\boldsymbol{m}) = \int_\Omega -2(1-a)\langle F^T\boldsymbol{l}, F^T\boldsymbol{m}\rangle + 2b\,\nabla\boldsymbol{m} : \nabla\boldsymbol{l} + 2\lambda\langle \boldsymbol{l},\boldsymbol{m}\rangle, \tag{37}$$

$$b_1(q,\boldsymbol{w}) = \int_\Omega -q\frac{\partial \det}{\partial F} : \nabla\boldsymbol{w}, \tag{38}$$

$$b_2(\mu,\boldsymbol{l}) = \int_\Omega 2\mu\langle \boldsymbol{l},\boldsymbol{n}\rangle. \tag{39}$$

The linearized system (31)-(34) can now be reduced to a standard saddle point system. In fact, adding (31)-(32) together, and (33)-(34) together as well, yields

$$a(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{v}}) + b(\tilde{o},\tilde{\boldsymbol{v}}) = \tilde{L}_1(\tilde{\boldsymbol{v}}), \tag{40}$$
$$b(\tilde{q},\tilde{\boldsymbol{w}}) = \tilde{L}_2(\tilde{q}), \tag{41}$$

where $\tilde{\boldsymbol{w}} = (\boldsymbol{w},\boldsymbol{l})$, $\tilde{\boldsymbol{v}} = (\boldsymbol{v},\boldsymbol{m})$, $\tilde{o} = (o,\gamma)$, and $\tilde{q} = (q,\mu)$. Moreover

$$a(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{v}}) = a_1(\boldsymbol{w},\boldsymbol{v}) + a_2(\boldsymbol{l},\boldsymbol{v}) + a_2(\boldsymbol{m},\boldsymbol{v}) + a_3(\boldsymbol{l},\boldsymbol{m}), \tag{42}$$
$$b(\tilde{q},\tilde{\boldsymbol{v}}) = b_1(q,\boldsymbol{v}) + b_2(\mu,\boldsymbol{m}), \tag{43}$$
$$\tilde{L}_1(\tilde{\boldsymbol{v}}) = L_1(\boldsymbol{v}) + L_2(\boldsymbol{m}), \tag{44}$$
$$\tilde{L}_2(\tilde{q}) = L_3(q) + L_4(\mu). \tag{45}$$

The system (40)-(41) exhibits a saddle point structure. Its well-posedness is shown in the theorem by Ladyzenskaya-Babuska-Brezzi (LBB) that we describe next as presented in [32].

Theorem 6 (Ladyzenskaya-Babuska-Brezzi)**.** *Consider the following saddle point problem*

$$a(u,v) + b(p,v) = L_V(v) \qquad \forall v\in\mathbb{V}, u\in\mathbb{V} \tag{46}$$
$$b(q,u) = L_P(q) \qquad \forall q\in\mathbb{P}, p\in\mathbb{P} \tag{47}$$

with $\mathbb{V}$ and $\mathbb{P}$ given Hilbert spaces, L_V and L_P belonging to $\mathbb{V}'$ and $\mathbb{P}'$ respectively. Moreover, a and b are continuous bilinear forms defined on $\mathbb{V}\times\mathbb{V}$ and $\mathbb{P}\times\mathbb{V}$, respectively. Define the operators

$$\begin{aligned} &\mathscr{B} : \mathbb{V}\to\mathbb{P}' \\ &\quad v\mapsto \mathscr{B}v \text{ such that } \langle \mathscr{B}v, q\rangle = b(q,v) \qquad \forall q\in\mathbb{P} \end{aligned}$$

$$\begin{aligned} &\mathscr{A} : \mathrm{Ker}\mathscr{B}\to(\mathrm{Ker}\mathscr{B})' \\ &\quad w\mapsto \mathscr{A}w \text{ such that } \langle \mathscr{A}w, v\rangle = a(w,v) \qquad \forall v\in\mathrm{Ker}\mathscr{B} \end{aligned}$$

Then the operator $\mathscr{B}$ is onto if and only if the spaces $\mathbb{V}$ and $\mathbb{P}$ satisfy the following inf-sup condition:

$$\inf_{q\in\mathbb{P},\|q\|=1}\sup_{v\in\mathbb{V},\|v\|=1} b(q,v) \geq \beta > 0 \tag{48}$$

Moreover, the mixed problem is well-posed if and only if $\mathscr{B}$ is onto and $\mathscr{A}$ is invertible.

According to the LBB theorem, the well-posedness of the saddle point system requires both $\mathscr{B}$ being onto and $\mathscr{A}$ being invertible. Moreover $\mathscr{B}$ being onto is equivalent to the inf-sup condition (48) being satisfied. What is the corresponding equivalent condition that guarantees the invertibility of $\mathscr{A}$? It turns out that $\mathscr{A}$ being invertible is also equivalent to an inf-sup condition, namely, the inf-sup condition of $a(\cdot,\cdot)$ on the space $\mathrm{Ker}\mathscr{B}$,

$$\inf_{w\in\mathrm{Ker}\mathscr{B},\|w\|=1}\sup_{v\in\mathrm{Ker}\mathscr{B},\|v\|=1} a(w,v)\geq\alpha>0. \tag{49}$$

This can be easily proved using the LBB theorem, and the fact that a linear operator A on a Hilbert space H is invertible if and only if A is onto and $\mathrm{Ker}(A)=0$ ([31] p104). Therefore, the well-posedness of a standard saddle point system amounts to the verification of the two inf-sup conditions (48) and (49). In practice, the inf-sup condition (49) is often replaced by the following stronger yet easier to verify *ellipticity* condition:

$$\inf_{v\in\mathrm{Ker}\mathscr{B},\|v\|=1} a(v,v)\geq\alpha>0. \tag{50}$$

However, in most situations the spaces $\mathbb{P}$ and $\mathbb{V}$ are different, and so, the inf-sup condition (48) cannot be replaced with a stronger ellipticity condition, and it may be very difficult to verify analytically.

For the LCE problem that we study, the bilinear form $b(\tilde{q},\tilde{\boldsymbol{v}})=b_1(q,\boldsymbol{v})+b_2(\mu,\boldsymbol{m})$ is the sum of two decoupled bilinear forms. We prove that the inf-sup condition for $b(\tilde{q},\tilde{\boldsymbol{v}})$ is actually equivalent to the inf-sup conditions for both b_1 and b_2.

Theorem 7. *The inf-sup condition for $b(\tilde{q},\tilde{\boldsymbol{v}})=b_1(q,\boldsymbol{v})+b_2(\mu,\boldsymbol{m})$ is satisfied if and only if the corresponding inf-sup conditions for $b_1(q,\boldsymbol{v})$ and $b_2(\mu,\boldsymbol{m})$ hold.*

Proof. Assume the bilinear form $b_1(q,\boldsymbol{v})$ is defined on $\mathbb{P}\times\mathbb{V}$ and $b_2(\mu,\boldsymbol{m})$ is defined on $\Lambda\times\mathbb{M}$, where $\mathbb{P},\mathbb{V},\Lambda,\mathbb{M}$ are Hilbert spaces.

First assume the inf-sup condition for $b(\tilde{q},\tilde{\boldsymbol{v}})$ is satisfied. Then it follows from Theorem 6 that the operator

$$\begin{aligned}&\mathscr{B}:\mathbb{V}\times\mathbb{M}\to\mathbb{P}'\times\Lambda'\\ &\quad(\boldsymbol{v},\boldsymbol{m})\mapsto\mathscr{B}(\boldsymbol{v},\boldsymbol{m})\text{ such that }\langle\mathscr{B}(\boldsymbol{v},\boldsymbol{m}),(q,\mu)\rangle=b_1(q,\boldsymbol{v})+b_2(\mu,\boldsymbol{m})\qquad\forall(q,\mu)\in\mathbb{P}\times\Lambda\end{aligned}$$

is onto. Therefore, the operators

$$\begin{aligned}&\mathscr{B}_1:\mathbb{V}\to\mathbb{P}'\\ &\quad\boldsymbol{v}\mapsto\mathscr{B}_1\boldsymbol{v}\text{ such that }\langle\mathscr{B}_1\boldsymbol{v},q\rangle=b_1(q,\boldsymbol{v})\qquad\forall q\in\mathbb{P}\end{aligned}$$

and

$$\begin{aligned}&\mathscr{B}_2:\mathbb{M}\to\Lambda'\\ &\quad\boldsymbol{m}\mapsto\mathscr{B}_2\boldsymbol{m}\text{ such that }\langle\mathscr{B}_2\boldsymbol{m},\mu\rangle=b_2(\mu,\boldsymbol{m})\qquad\forall\mu\in\Lambda\end{aligned}$$

are both onto. Hence, if follows from Theorem 6 that the inf-sup conditions for $b_1(q,\boldsymbol{v})$ and $b_2(\mu,\boldsymbol{m})$ are satisfied.

Conversely, let us assume the inf-sup conditions $b_1(q,\boldsymbol{v})$ and $b_2(\mu,\boldsymbol{m})$ are both satisfied. Then it follows that $\mathscr{B}_1$ and $\mathscr{B}_2$ are both onto, and so is the operator $\mathscr{B}$. Therefore by Theorem 6, $b(\tilde{q},\tilde{\boldsymbol{v}})=b_1(q,\boldsymbol{v})+b_2(\mu,\boldsymbol{m})$ satisfies the inf-sup condition. □

Consequently, to verify the inf-sup condition for $b(\tilde{q},\tilde{\boldsymbol{v}})=b_1(q,\boldsymbol{v})+b_2(\mu,\boldsymbol{m})$, it is sufficient to verify the inf-sup conditions for $b_1(q,\boldsymbol{v})$ and $b_2(\mu,\boldsymbol{m})$ individually. We point out that the bilinear form $b_1(q,\boldsymbol{v})$ in (38) corresponds to that of the incompressible elasticity problem [32], while $b_2(\mu,\boldsymbol{m})$ in (39) corresponds to that of the harmonic map problem [24].

We observe that the inf-sup condition for $b_1(q,\boldsymbol{v})$ is at least satisfied at the strain-free and the stress-free states. In fact, at the strain-free state, $F=I$, it reduces to that of the Stokes problem (for the proof, see for example, [29]):

$$\inf_{q\in L^2(\Omega)}\sup_{\boldsymbol{v}\in\mathbf{H}^1_{0|\Gamma_u}(\Omega)}\frac{\langle q,\mathrm{div}(\boldsymbol{v})\rangle}{\|q\|_0\|\boldsymbol{v}\|_1}\geq\beta_1>0. \tag{51}$$

On the other hand, since the stress-free state has constant F matrix, the inf-sup condition for $b_1(q,\boldsymbol{v})$ can be verified by change of variables [29]. In the general case that $\boldsymbol{u}\neq 0$ and F is not a constant, analytical verification of such a condition can be very challenging.

The inf-sup condition for $b_2(\mu,\boldsymbol{m})$ holds provided $\boldsymbol{n}$ is sufficiently smooth. This is established in the following theorem, whose proof is a slight modification of that in [24]. The details can be found in [29].

Theorem 8. *Assume* $\boldsymbol{n} \in \mathbf{H}^1_{\boldsymbol{n}_0|\Gamma_n}(\Omega) \bigcap W^{1,\infty}(\Omega)$*, then the inf-sup condition for* $b_2(\mu, \boldsymbol{m})$ *holds. That is*

$$\inf_{\mu \in H^{-1}_{\Gamma_n}(\Omega)} \sup_{\boldsymbol{m} \in \mathbf{H}^1_{\mathbf{0}|\Gamma_n}(\Omega)} \frac{\langle 2\boldsymbol{n} \cdot \boldsymbol{m}, \mu \rangle}{\|\boldsymbol{m}\|_1 \|\mu\|_{-1}} \geq \beta_2 > 0. \tag{52}$$

Finally, to establish the ellipticity condition for the bilinear form $a(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}})$, is, in general, very complicated due to the complication of the expressions of $a_1(\cdot,\cdot)$, $a_2(\cdot,\cdot)$ and $a_3(\cdot,\cdot)$. We found that it actually does *not* hold at the stress-free state ([29]). However, this does not imply that the linearized system (31)-(34) is ill-posed, since as previously mentioned, ellipticity is a sufficient condition instead of a necessary condition.

Although in many situations, the rigorous proof of the inf-sup conditions or ellipticity conditions is not available, the numerical "verification" may be straightforward. We added quotation marks, because numerical verification is not a rigorous argument, and therefore cannot replace the analytical proof. However, it may provide some insights when the analytical proof is not available. We will elaborate on this in later sections.

4 Existence and well-posedness of the discrete problem

In this section, we investigate the existence and well-posedness of the discrete problem. Unlike the usual approach of simply replacing continuous spaces by finite element spaces, following Winther et al. [24], we include an interpolation operator in the discrete formulation. This operator plays an important role in the proof of existence and well-posedness of the discrete problem.

In this section, we first prove existence of minimizer of the discrete problem. We then derive the Euler-Lagrange equations and the corresponding linearized system. We also explain how to numerically compute the constants in the inf-sup and ellipticity conditions as a way of verifying the well-posedness of the linearized system. Next, we prove the existence and uniqueness of the Lagrange multipliers as a consequence of the inf-sup conditions for the discrete problem being satisfied. Finally, we discuss some implementation issues, such as how to deal with the interpolation operator in the software package FEniCS, and how to numerically assemble the H^{-1} norm.

4.1 The discrete problem and existence of minimizer

As indicated in previous sections, the problem of finding $(\boldsymbol{u}, p)$ is very similar to the case of the incompressible elasticity, while finding $(\boldsymbol{n}, \lambda)$ bears an analogy with the harmonic map problem. Thus, we choose the $\mathbf{P}_2 \times P_1$ finite element spaces for $(\boldsymbol{u}, p)$, as in incompressible elasticity, and $\mathbf{P}_1 \times P_1$ for $(\boldsymbol{n}, \lambda)$, following the harmonic map problem.

We let V_h denote the space of continuous piecewise linear functions and, $V_{h,g|\Gamma} = \{v \in V_h \cap H^1 : v = g \text{ on } \Gamma\}$. The symbols $\mathbf{V}_h$ and $\mathbf{V}_{h,\boldsymbol{g}|\Gamma}$ refer to the corresponding vector version. We use π_h as the nodal interpolation operators onto the spaces V_h and $\mathbf{V}_h$. We let W_h denote the space of continuous piecewise quadratic functions and, $W_{h,g|\Gamma} = \{w \in W_h \cap H^1 : w = g \text{ on } \Gamma\}$. The symbols $\mathbf{W}_h$ and $\mathbf{W}_{h,\boldsymbol{g}|\Gamma}$ denote the corresponding vector version as well.

The energy functional is still defined as

$$\begin{aligned}\Pi(\boldsymbol{u}, \boldsymbol{n}) = &\int_\Omega \left(|F|^2 - (1-a)|F^T \boldsymbol{n}|^2 \right) + b|\nabla \boldsymbol{n}|^2 \\ &- \int_\Omega \boldsymbol{f} \cdot \boldsymbol{u} - \int_\Gamma \boldsymbol{g} \cdot \boldsymbol{u} da.\end{aligned} \tag{53}$$

We define the admissible set

$$\mathcal{A}_h = \mathcal{K}_h \times \mathcal{N}_h, \tag{54}$$

where

$$\mathcal{K}_h = \{\boldsymbol{u}_h \in \mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}, \int_\Omega q_h(\det(I + \nabla \boldsymbol{u}_h) - 1)dx = 0, \forall q_h \in V_h\}, \tag{55}$$

and

$$\mathcal{N}_h = \{\boldsymbol{n}_h \in \mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}, \int_\Omega \mu_h \pi_h(|\boldsymbol{n}_h|^2 - 1)dx = 0, \forall \mu_h \in V_{h,0|\Gamma_n}\}. \tag{56}$$

Notice that any piecewise linear function $\boldsymbol{n}_h$ belongs to $\mathcal{N}_h$ if and only if the function $\pi_h(|\boldsymbol{n}_h|^2 - 1) \in V_{h,0|\Gamma_n}$ is identically 0, which means $|\boldsymbol{n}_h| = 1$ at all the mesh nodes.

Our discrete formulation of the minimization problem is

$$\text{Find } (\boldsymbol{u}_h, \boldsymbol{n}_h) \in \mathcal{A}_h, \text{ minimizing } \Pi \text{ in } \mathcal{A}_h. \tag{57}$$

Before proving existence of minimizer, we first establish the following lemma.

Lemma 9. *Assume $\boldsymbol{n} \in N_h$ and $0 < a < 1$, then for any matrix $F \in \mathbb{M}^{2\times 2}$*

$$|F|^2 - (1-a)|F^T\boldsymbol{n}|^2 \geq a|F|^2 \tag{58}$$

holds.

Proof. Take any point $x \in \Omega$, and suppose it is inside the triangle $\triangle P_1P_2P_3$. Since $\boldsymbol{n} \in N_h$, we have

$$\boldsymbol{n}(x) = \lambda_1\boldsymbol{n}(P_1) + \lambda_2\boldsymbol{n}(P_2) + \lambda_3\boldsymbol{n}(P_3)$$

where $\lambda_i, i = 1, 2, 3$ are barycentric coordinates. As indicated above, $\boldsymbol{n} \in N_h$ if and only if $|\boldsymbol{n}| = 1$ at all the mesh nodes. Thus it follows that

$$\begin{aligned}
|\boldsymbol{n}(x)| &= |\lambda_1\boldsymbol{n}(P_1) + \lambda_2\boldsymbol{n}(P_2) + \lambda_3\boldsymbol{n}(P_3)| \\
&\leq \lambda_1|\boldsymbol{n}(P_1)| + \lambda_2|\boldsymbol{n}(P_2)| + \lambda_3|\boldsymbol{n}(P_3)| \\
&= \lambda_1 + \lambda_2 + \lambda_3 \\
&= 1.
\end{aligned}$$

If $|\boldsymbol{n}(x)| = 0$, then the conclusion follows trivially. In the following, we assume that $|\boldsymbol{n}(x)| > 0$.

Let $\hat{\boldsymbol{n}} = \boldsymbol{n}(x)/|\boldsymbol{n}(x)|$, then $|\hat{\boldsymbol{n}}| = 1$. So,

$$\begin{aligned}
&|F|^2 - (1-a)|F^T\boldsymbol{n}|^2 \\
&= |F|^2 - (1-a)|\boldsymbol{n}(x)|^2|F^T\hat{\boldsymbol{n}}|^2 \\
&\geq |F|^2 - (1-a)|F^T\hat{\boldsymbol{n}}|^2 \\
&\geq a|F|^2,
\end{aligned}$$

where we have used Lemma 2 in the last step. □

Now we establish the following existence theorem.

Theorem 10. *There exists a solution to the discrete minimization problem (57).*

Proof. Take any $(\boldsymbol{u}_h, \boldsymbol{n}_h) \in \mathcal{A}_h$. It follows from Lemma 9 that

$$\begin{aligned}
\Pi(\boldsymbol{u}_h, \boldsymbol{n}_h) &\geq \int_\Omega a|(I + \nabla\boldsymbol{u}_h)|^2 + b|\nabla\boldsymbol{n}_h|^2 dx \\
&\quad - \|\boldsymbol{f}\|_{L^2(\Omega)}\|\boldsymbol{u}_h\|_{L^2(\Omega)} - \|\boldsymbol{g}\|_{L^2(\Gamma)}\|\boldsymbol{u}_h\|_{L^2(\Gamma)} \\
&\geq \int_\Omega a|(I + \nabla\boldsymbol{u}_h)|^2 + b|\nabla\boldsymbol{n}_h|^2 dx \\
&\quad - \left(\frac{1}{\varepsilon}\|\boldsymbol{f}\|^2_{L^2(\Omega)} + \varepsilon\|\boldsymbol{u}_h\|^2_{L^2(\Omega)}\right) - \left(\frac{1}{\varepsilon}\|\boldsymbol{g}\|^2_{L^2(\Gamma)} + \varepsilon\|\boldsymbol{u}_h\|^2_{L^2(\Gamma)}\right) \\
&\geq \int_\Omega C_1|(I + \nabla\boldsymbol{u}_h)|^2 + C_2|\nabla\boldsymbol{n}_h|^2 dx - C_3
\end{aligned}$$

where $\varepsilon > 0$ is small, and $C_i > 0, i = 1, 2, 3$ are constants. In the last step, we have applied the generalized Poincaré inequality ([9], p281) and the Trace Theorem ([20], p258). Thus $\Pi(\boldsymbol{u}_h, \boldsymbol{n}_h) \to \infty$ as $\|\boldsymbol{u}_h\|_1$ or $\|\boldsymbol{n}_h\|_1$ goes to ∞. Hence, its minimum must be achieved in a bounded subset of $\mathcal{A}_h$.

On the other hand, the admissible set $\mathcal{A}_h$ is closed. The reason is as follows. Let $\varphi_j, j = 1, \cdots, N$ be a basis of V_h, and $\psi_j, j = 1, \cdots, M$ a basis of $V_{h,0|\Gamma_n}$, and define

$$g_j(\boldsymbol{u}_h, \boldsymbol{n}_h) = \begin{cases} \int_\Omega \varphi_j(\det(I + \nabla\boldsymbol{u}_h) - 1)dx & 1 \leq j \leq N \\ \int_\Omega \psi_{j-N}\pi_h(|\boldsymbol{n}_h|^2 - 1)dx & N+1 \leq j \leq N+M \end{cases} \tag{59}$$

Then g_j is a continuous function on $\left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}\right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}\right)$. Therefore $\mathcal{A}_h$ can be written as the intersection of reciprocal images of 0 by the continuous functions g_j, so it is a closed set.

Since $\Pi(\boldsymbol{u}_h, \boldsymbol{n}_h)$ is a continuous function on a closed, bounded *finite-dimensional* set, the Weierstrass Theorem guarantees the existence of $(\boldsymbol{u}_h, \boldsymbol{n}_h) \in \mathcal{A}_h$ minimizing Π in $\mathcal{A}_h$. □

4.2 Equilibrium equations and linearized system

Similar to the continuous problem, we convert the constrained minimization problem to an unconstrained one by including the Lagrange multipliers. After that, we derive the equilibrium equations and their linearization. These equations are analogous to those of the continuous problem, except for the presence of the interpolation operator π_h.

After including the Lagrange multipliers, the discrete energy functional is given by

$$\begin{aligned}\mathcal{E}(\boldsymbol{u},\boldsymbol{n},p,\lambda) = &\int_\Omega (|F|^2-(1-a)|F^T\boldsymbol{n}|^2)+b|\nabla\boldsymbol{n}|^2\\ &-p(\det(F)-1)+\lambda(\pi_h\langle\boldsymbol{n},\boldsymbol{n}\rangle-1)\\ &-\int_\Omega \boldsymbol{f}\cdot\boldsymbol{u}-\int_\Gamma \boldsymbol{g}\cdot\boldsymbol{u}da,\end{aligned} \tag{60}$$

where $F = I+\nabla\boldsymbol{u}$.

Taking the first variation of the functional (60), we obtain the following equilibrium equations (Euler-Lagrange equations):

$$\begin{aligned}0 = &\int_\Omega 2\left(F:\nabla\boldsymbol{v}-(1-a)\langle F^T\boldsymbol{n},\nabla\boldsymbol{v}^T\boldsymbol{n}\rangle\right)-p\frac{\partial\det}{\partial F}:\nabla\boldsymbol{v}\\ &-\int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}-\int_\Gamma \boldsymbol{g}\cdot\boldsymbol{v}da,\end{aligned} \tag{61}$$

$$0 = \int_\Omega -2(1-a)\langle F^T\boldsymbol{n},F^T\boldsymbol{m}\rangle+2b\,\nabla\boldsymbol{m}:\nabla\boldsymbol{n}+2\lambda\pi_h\langle\boldsymbol{n},\boldsymbol{m}\rangle, \tag{62}$$

$$0 = \int_\Omega -q(\det F-1), \tag{63}$$

$$0 = \int_\Omega \mu\pi_h(\langle\boldsymbol{n},\boldsymbol{n}\rangle-1), \tag{64}$$

where the solution $(\boldsymbol{u},\boldsymbol{n},p,\lambda)\in\mathbf{W}_{h,\boldsymbol{u}_0|\Gamma_u}\times\mathbf{V}_{h,\boldsymbol{n}_0|\Gamma_n}\times V_h\times V_{h,\lambda_0|\Gamma_n}$, and the test function $(\boldsymbol{v},\boldsymbol{m},q,\mu)\in\mathbf{W}_{h,0|\Gamma_u}\times\mathbf{V}_{h,0|\Gamma_n}\times V_h\times V_{h,0|\Gamma_n}$.

Linearization around a solution $(\boldsymbol{u},\boldsymbol{n},p,\lambda)$ yields the system

$$a_1(\boldsymbol{w},\boldsymbol{v})+a_2(\boldsymbol{l},\boldsymbol{v})+b_1(o,\boldsymbol{v}) = L_1(\boldsymbol{v}), \tag{65}$$

$$a_2(\boldsymbol{m},\boldsymbol{w})+a_3(\boldsymbol{l},\boldsymbol{m})+b_2(\gamma,\boldsymbol{m}) = L_2(\boldsymbol{m}), \tag{66}$$

$$b_1(q,\boldsymbol{w}) = L_3(q), \tag{67}$$

$$b_2(\mu,\boldsymbol{l}) = L_4(\mu), \tag{68}$$

where both the perturbation $(\boldsymbol{w},\boldsymbol{l},o,\gamma)$ and the test function $(\boldsymbol{v},\boldsymbol{m},q,\mu)$ belong to $\mathbf{W}_{h,0|\Gamma_u}\times\mathbf{V}_{h,0|\Gamma_n}\times V_h\times V_{h,0|\Gamma_n}$. Here the bilinear forms a_1, a_2, a_3, b_1 and b_2 depend on the solution $(\boldsymbol{u},\boldsymbol{n},p,\lambda)$. Moreover a_1, a_2 and b_1 are as in the continuous case, while a_3 and b_2 are slightly different, and are given by

$$a_3(\boldsymbol{l},\boldsymbol{m}) = \int_\Omega -2(1-a)\langle F^T\boldsymbol{l},F^T\boldsymbol{m}\rangle+2b\,\nabla\boldsymbol{m}:\nabla\boldsymbol{l}+2\lambda\pi_h\langle\boldsymbol{l},\boldsymbol{m}\rangle, \tag{69}$$

and

$$b_2(\mu,\boldsymbol{m}) = \int_\Omega 2\mu\pi_h\langle\boldsymbol{n},\boldsymbol{m}\rangle. \tag{70}$$

4.3 Well-posedness of the linearized system

As in the continuous case, verifying the well-posedness of the linearized system (65)-(68) can be reduced to verifying the inf-sup conditions for $b_1(q,\boldsymbol{v})$, $b_2(\mu,\boldsymbol{m})$ and $a(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{v}}) = a_1(\boldsymbol{w},\boldsymbol{v})+a_2(\boldsymbol{l},\boldsymbol{v})+a_2(\boldsymbol{m},\boldsymbol{v})+a_3(\boldsymbol{l},\boldsymbol{m})$.

The inf-sup condition for $b_1(q,\boldsymbol{v})$ is also satisfied at least at the strain-free and stress-free state. In fact, at the strain-free state, $F = I$, this condition can be formulated as

$$\inf_{q\in V_h}\sup_{\boldsymbol{v}\in\mathbf{W}_{h,0|\Gamma_u}}\frac{\langle q,\mathrm{div}(\boldsymbol{v})\rangle}{\|q\|_0\|\boldsymbol{v}\|_1}\geq\beta_1>0, \tag{71}$$

which is exactly the inf-sup condition for the Stokes problem with the Taylor-Hood element $\boldsymbol{P}_2\times P_1$ (proof of this inf-sup condition can be found for example in Proposition 6.1 of [5]). The verification of the condition $b_1(q,\boldsymbol{v})$ at the stress-free state follows as in the continuous case.

The proof of the inf-sup condition for $b_2(\mu, \boldsymbol{m})$ is similar to the one in [24]. The only difference is that, in our case, the test functions $\mu, \boldsymbol{m}$ are zero only on part of the boundary. A slight modification of the proof in [24] gives us the following. (The detailed proof can be found in [29].)

Theorem 11. *Assume $\boldsymbol{n} \in \mathbf{H}^1_{\boldsymbol{n}_0|\Gamma_n}(\Omega) \bigcap \boldsymbol{W}^{1,\infty}(\Omega)$, and $\boldsymbol{n}_h \in \mathbf{V}_{h,\boldsymbol{n}_0|\Gamma_n}$ satisfies $|\boldsymbol{n}_h| \geq C > 0$ and $\|\boldsymbol{n}_h - \pi_h \boldsymbol{n}\|_1 \leq \gamma / |\log(h)|^{1/2}$. Then there is a positive constant β_2, independent of h, such that*

$$\inf_{\mu \in V_{h,0|\Gamma_n}} \sup_{\boldsymbol{m} \in \mathbf{V}_{h,0|\Gamma_n}} \frac{\langle \pi_h[\boldsymbol{n}_h \cdot \boldsymbol{m}], \mu \rangle}{\|\mu\|_{-1} \|\boldsymbol{m}\|_1} \geq \beta_2 \tag{72}$$

Theorem 11 states that, if the true solution $\boldsymbol{n}$ is smooth, the approximate solution $\boldsymbol{n}_h$ is close to it, and its norm is bounded below, then the inf-sup condition for $b_2(\mu, \boldsymbol{m})$ is always satisfied.

For the inf-sup condition for $a(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}})$ and the inf-sup conditions for b_1 and b_2 in general cases, analytical verification may turn out to be very difficult. However, in the discrete case, the inf-sup values and the ellipticity constants can be computed numerically. We can compute the inf-sup values for a series of finer and finer meshes. If these inf-sup values are bounded below by a positive constant, we infer some evidence that the inf-sup condition might be satisfied for all meshes. This type of verification known as *inf-sup test* [8] provides a convenient way to get information when analytical results are not available. However, the inf-sup test cannot replace the analytical proof, because we cannot apply the test on infinite number of meshes.

For a general inf-sup condition, the inf-sup value

$$\beta_h = \inf_{q \in \mathbb{P}_h, ||q||=1} \left\{ \sup_{v \in \mathbb{V}_h, ||v||=1} b(q, v) \right\} \tag{73}$$

turns out to be related to the smallest singular value of certain matrix. The following theorem summarizes the results from [5].

Theorem 12. *Let the matrices S, T, B be defined by the following equations*

$$\|q_h\|^2 = \boldsymbol{q}^T S \boldsymbol{q}, \tag{74}$$

$$\|v_h\|^2 = \boldsymbol{v}^T T \boldsymbol{v}, \tag{75}$$

$$b(q_h, v_h) = \boldsymbol{q}^T B \boldsymbol{v}, \tag{76}$$

where $\boldsymbol{q}$, $\boldsymbol{v}$ are the degrees of freedom of q_h and v_h respectively. Then the inf-sup value β_h in (73) is equal to the smallest singular value of the matrix $S^{-\frac{1}{2}} B T^{-\frac{1}{2}}$.

In our case, we also want to compute the inf-sup value or ellipticity constant for the bilinear form $a(\cdot, \cdot)$ on $\mathrm{Ker}(\mathscr{B}_h)$, where $\mathscr{B}_h : \mathbb{V}_h \to \mathbb{P}_h'$ is defined by $b(q, v) = (q, \mathscr{B}_h v)$ for any $q \in \mathbb{P}_h$ and $v \in \mathbb{V}_h$. That is, we want to compute the inf-sup value $\hat{\beta}_h$ in

$$\hat{\beta}_h = \inf_{u \in \mathrm{Ker}(\mathscr{B}_h), \|u\|=1} \sup_{v \in \mathrm{Ker}(\mathscr{B}_h), \|v\|=1} a(u, v) \tag{77}$$

and the ellipticity constant $\hat{\alpha}_h$ in

$$\hat{\alpha}_h = \inf_{v \in \mathrm{Ker}(\mathscr{B}_h), \|v\|=1} a(v, v). \tag{78}$$

We prove the following result, which is similar to Theorem 12.

Theorem 13. *Let n and m be the dimensions of $\mathbb{V}_h$ and $\mathbb{P}_h$, respectively. Let the matrices T, A, B be defined by the following equations*

$$\|v_h\|^2 = \boldsymbol{v}^T T \boldsymbol{v},$$

$$a(u_h, v_h) = \boldsymbol{u}^T A \boldsymbol{v},$$

$$v_h \in \mathrm{Ker}(\mathscr{B}_h) \Leftrightarrow B\boldsymbol{v} = 0,$$

where $\boldsymbol{u}$, $\boldsymbol{v}$ are the degrees of freedom of u_h and v_h respectively. Assume that B is full-rank, and let the matrix Q be defined by the QR decomposition of $(BT^{-1/2})^T$

$$(BT^{-1/2})^T = Q \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

Then the inf-sup value $\hat{\beta}_h$ in (77) and the ellipticity constant $\hat{\alpha}_h$ in (78) are respectively equal to the smallest singular value and the smallest eigenvalue of the matrix A_1, where A_1 is the lower right $(n-m) \times (n-m)$ submatrix of the matrix $Q^T T^{-1/2} A T^{-1/2} Q$.

Proof. First, let $\boldsymbol{x} = T^{\frac{1}{2}}\boldsymbol{u}$, $\boldsymbol{y} = T^{\frac{1}{2}}\boldsymbol{v}$. So

$$\hat{\beta}_h = \inf_{\boldsymbol{x}\in \mathrm{Ker}(\tilde{B})} \sup_{\boldsymbol{y}\in \mathrm{Ker}(\tilde{B})} \frac{\boldsymbol{x}^T \tilde{A} \boldsymbol{y}}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}}\sqrt{\boldsymbol{y}^T\boldsymbol{y}}} \tag{79}$$

where $\tilde{B} = BT^{-1/2}$, and $\tilde{A} = T^{-1/2}AT^{-1/2}$.

Since $\tilde{B}$ is full-rank, the matrix $R \in \mathbb{M}^{m\times m}$ in the QR decomposition

$$\tilde{B}^T = Q \begin{pmatrix} R \\ 0 \end{pmatrix} \tag{80}$$

is non-singular. Let

$$Q^T \boldsymbol{x} = \begin{pmatrix} \boldsymbol{w}_x \\ \boldsymbol{z}_x \end{pmatrix}, \tag{81}$$

where $\boldsymbol{w}_x \in \mathbb{R}^m$ and $\boldsymbol{z}_x \in \mathbb{R}^{n-m}$. Then it is easy to verify that

$$\boldsymbol{x} \in Ker(\tilde{B}) \Leftrightarrow \boldsymbol{w}_x = 0.$$

Thus there is no constraint on $\boldsymbol{z}_x$. Therefore

$$\inf_{\boldsymbol{x}\in \mathrm{Ker}(\tilde{B})} \sup_{\boldsymbol{x}\in \mathrm{Ker}(\tilde{B})} \frac{\boldsymbol{x}^T \tilde{A} \boldsymbol{y}}{\sqrt{\boldsymbol{x}^T\boldsymbol{x}}\sqrt{\boldsymbol{y}^T\boldsymbol{y}}} = \inf_{\boldsymbol{z}_x} \sup_{\boldsymbol{z}_y} \frac{\boldsymbol{z}_x^T A_1 \boldsymbol{z}_y}{\sqrt{\boldsymbol{z}_x^T\boldsymbol{z}_x}\sqrt{\boldsymbol{z}_y^T\boldsymbol{z}_y}} \tag{82}$$

where A_1 is the lower right $(n-m)\times(n-m)$ corner of the matrix $Q^T\tilde{A}Q$. Thus by Theorem 12, $\hat{\beta}_h$ is equal to the smallest singular value of the matrix A_1.

Similarly, we can show that $\hat{\alpha}_h$ is equal to the smallest eigenvalue of the matrix A_1. □

Remark The matrix B is full-rank if and only if the operator $\mathscr{B}_h$ is onto, which is true if and only if the following inf-sup condition holds

$$\inf_{q\in\mathbb{P}_h, ||q||=1} \left\{ \sup_{v\in\mathbb{V}_h, ||v||=1} b(q,v) \right\} \geq \beta_h > 0. \tag{83}$$

4.4 Existence and uniqueness of the Lagrange multipliers for the discrete system

For the incompressible elasticity problem, Le Tallec [27] proved existence and uniqueness of the Lagrange multiplier p given that the inf-sup condition for $b_1(q, \boldsymbol{v})$ is satisfied. In this subsection, we use similar arguments to prove existence and uniqueness of p and λ given that the inf-sup conditions for $b_1(q, \boldsymbol{v})$ and $b_2(\mu, \boldsymbol{m})$ are both satisfied. The proof uses the following result of Clarke [10, 11].

Theorem 14. *Let J denote a finite set of integers. We suppose that the following are given: E a Banach space, g_0, $g_j (j \in J)$ locally Lipschitz functions from E to $\mathbb{R}$, and C a closed subset of E. We consider the following problem:*

$$\begin{aligned} &\textit{Minimize } g_0(x) \\ &\textit{subject to } x \in C, \quad g_j(x) = 0, \quad \forall j \in J. \end{aligned} \tag{84}$$

If $\bar{x}$ is a local solution of (84), then there exist real numbers r_0, s_j not all zero, and a point ξ in the dual space E' of E such that:

$$\xi \in r_0 \partial g_0(\bar{x}) + \sum_j s_j \partial g_j(\bar{x}), \qquad -\xi \in N_c(\bar{x}), \tag{85}$$

where $N_c(\bar{x})$ is the normal cone at C in $\bar{x}$, and ∂g_j is the generalized gradient of $g_j(x)$.

Next we use Theorem 14 to prove the existence and uniqueness of p and λ.

Theorem 15. *Suppose $(\boldsymbol{u}_h, \boldsymbol{n}_h) \in \mathcal{K}_h \times \mathcal{N}_h$, and at $(\boldsymbol{u}_h, \boldsymbol{n}_h)$, the inf-sup conditions for b_1 and b_2 are both satisfied. Then there exist a unique $p_h \in V_h$ and a unique $\lambda_h \in V_{h,\lambda_0|\Gamma_n}$ such that $(\boldsymbol{u}_h, \boldsymbol{n}_h, p_h, \lambda_h)$ is a solution of the discrete equilibrium equations (61)-(64).*

Proof. Let us denote

$$E = C = \left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}\right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}\right) \tag{86}$$

$$g_0(x) = \Pi(\boldsymbol{v}_h, \boldsymbol{m}_h), \qquad g_j(x) = g_j(\boldsymbol{v}_h, \boldsymbol{m}_h), \tag{87}$$

where the functions g_j were defined in (59). It is easy to see that

$$N_C(\bar{x}) = N_E(\bar{x}) = (0,0). \tag{88}$$

Notice that

$$\partial\Pi(\boldsymbol{u}_h, \boldsymbol{n}_h) \subset \left\{Dg_0^1 + Dg_0^2\right\}, \tag{89}$$

where Dg_0^1 and Dg_0^2 are in $\left[\left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}\right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}\right)\right]^*$. We have

$$Dg_0^1(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h) = \begin{pmatrix} f_1(\boldsymbol{v}_h) \\ f_2(\boldsymbol{m}_h) \end{pmatrix},$$

where

$$f_1(\boldsymbol{v}) = \int_\Omega 2(F_h : \nabla \boldsymbol{v} - (1-a)\langle F_h^T \boldsymbol{n}_h, \nabla \boldsymbol{v}^T \boldsymbol{n}_h\rangle),$$

and

$$f_2(\boldsymbol{m}) = \int_\Omega -2(1-a)\langle F_h^T \boldsymbol{n}_h, F_h^T \boldsymbol{m}\rangle + 2b\, \nabla \boldsymbol{m} : \nabla \boldsymbol{n}_h.$$

Also,

$$Dg_0^2(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h) = \begin{pmatrix} -\int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}_h - \int_\Gamma \boldsymbol{g} \cdot \boldsymbol{v}_h da \\ 0 \end{pmatrix}.$$

We also observe that $g_j(\boldsymbol{v}_h, \boldsymbol{m}_h)$ is continuously differentiable in $\left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}\right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}\right)$, and that

$$\partial g_j = \{Dg_j\}, \tag{90}$$

$$Dg_j(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h) = \begin{pmatrix} \int_\Omega -\varphi_j \frac{\partial \det}{\partial F}(I + \nabla \boldsymbol{u}_h) : \nabla \boldsymbol{v}_h \\ 0 \end{pmatrix}$$
$$\text{for } 1 \le j \le N, \tag{91}$$

$$Dg_j(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h) = \begin{pmatrix} 0 \\ \int_\Omega \psi_{j-N}\, \pi_h(2\boldsymbol{n}_h \cdot \boldsymbol{m}_h) dx \end{pmatrix}$$
$$\text{for } N+1 \le j \le N+M. \tag{92}$$

Therefore applying Theorem 14, we have that there exists real numbers r_0, s_j, not all zero, such that

$$0 \in r_0 \partial\Pi(\boldsymbol{u}_h, \boldsymbol{n}_h) + \sum_{j=1}^{N+M} s_j \partial g_j(\boldsymbol{u}_h, \boldsymbol{n}_h). \tag{93}$$

Using (89) and (90), equation (93) becomes

$$r_0\{Dg_0^1(\boldsymbol{u}_h, \boldsymbol{n}_h) + Dg_0^2(\boldsymbol{u}_h, \boldsymbol{n}_h)\} + \sum_{j=1}^{N+M} s_j Dg_j(\boldsymbol{u}_h, \boldsymbol{n}_h) = 0,$$
$$\text{in } \left[\left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h}\right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h}\right)\right]^*. \tag{94}$$

Assume now $r_0 = 0$. By the linearity property, and using equations (91), (92), we rewrite (94) as follows:

$$\int_\Omega \left(\sum_{j=1}^{N} s_j \varphi_j\right) \frac{\partial \det}{\partial F}(I + \nabla \boldsymbol{u}_h) : \nabla \boldsymbol{v}_h dx = 0, \quad \forall \boldsymbol{v}_h \in \mathbf{W}_{h,0|\Gamma_u}, \tag{95}$$

$$\int_\Omega \left(\sum_{j=1}^{M} s_{N+j} \psi_j\right) \pi_h(2\boldsymbol{n}_h \cdot \boldsymbol{m}_h) dx = 0, \quad \forall \boldsymbol{m}_h \in \mathbf{V}_{h,0|\Gamma_n}. \tag{96}$$

Since at least one s_j is nonzero, at least one of the equations (95) or (96) is in contradiction with the inf-sup conditions. Thus r_0 cannot be zero. We can then divide (94) by r_0 to get

$$Dg_0^1(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h) + 1/r_0 \sum_{j=1}^{M} s_j Dg_j(\boldsymbol{u}_h, \boldsymbol{n}_h) = -Dg_0^2(\boldsymbol{u}_h, \boldsymbol{n}_h) \cdot (\boldsymbol{v}_h, \boldsymbol{m}_h),$$
$$\forall (\boldsymbol{v}_h, \boldsymbol{m}_h) \in \left[\left(\mathbf{W}_{h,0|\Gamma_u} + \boldsymbol{u}_{0h} \right) \times \left(\mathbf{V}_{h,0|\Gamma_n} + \boldsymbol{n}_{0h} \right) \right]. \tag{97}$$

That is

$$f_1(\boldsymbol{v}_h) - \int_\Omega p_h \frac{\partial \det}{\partial F}(I + \nabla \boldsymbol{u}_h) : \nabla \boldsymbol{v}_h dx = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}_h + \int_\Gamma \boldsymbol{g} \cdot \boldsymbol{v}_h da, \tag{98}$$

$$f_2(\boldsymbol{m}_h) + \int_\Omega \lambda_h \pi_h (2\boldsymbol{n}_h \cdot \boldsymbol{m}_h) dx = 0, \tag{99}$$

where we have denoted

$$p_h = \left(\sum_{j=1}^{N} s_j \varphi_j \right) / r_0, \tag{100}$$

$$\lambda_h = \left(\sum_{j=1}^{M} s_{N+j} \psi_j \right) / r_0. \tag{101}$$

Equations (98) and (99) are precisely (61)-(62). Since $(\boldsymbol{u}_h, \boldsymbol{n}_h) \in \mathcal{K}_h \times \mathcal{N}_h$, we conclude that $(\boldsymbol{u}_h, \boldsymbol{n}_h, p_h, \lambda_h)$ is a solution of (61)-(64).

Finally, if there were two distinct values p_h, their difference would violate the inf-sup condition for b_1. Likewise, if there existed two distinct values λ_h, their difference would also violate the inf-sup condition for b_2. So we have the uniqueness of both p_h and λ_h. □

4.5 Some implementation issues

In this subsection, we discuss issues related to the implementation of our numerical scheme, such as how to solve the nonlinear problem using the software package FEniCS, how to deal with the interpolation operator π_h in FEniCS, and how to assemble the H^{-1} norm when we assess the rate of convergence.

FEniCS [28] is an open source finite element package. It is very convenient to solve variational problems such as

$$a(u, v) = L(v), \quad \forall v \in \mathbb{V} \tag{102}$$

using FEniCS. To use FEniCS with C++, we just need to specify the finite element space $\mathbb{V}_h$, the expressions for the bilinear form $a(u, v)$ and the linear form $L(v)$ in a form file such as "Poisson.ufl", and compile the form file into a C++ header file "Poisson.h". Then in the C++ source file "main.cpp", we specify the boundary conditions and let FEniCS proceed with the work, which includes assembling the matrix $(a(\phi_i, \phi_j))$ and the right hand side $(L(\phi_j))$, and calling solvers for the linear system[4]. To use FEniCS with Python is even simpler. The form file and the boundary conditions can be specified in the same Python script, and we can call FEniCS interactively in the Python shell.

Our problem (61)-(64) however, is a nonlinear variational problem, and we cannot directly apply the above procedure in FEniCS. We explain here how to solve our nonlinear problem using FEniCS. Let N be the dimension of the space $\mathbf{W}_{h,0|\Gamma_u} \times \mathbf{V}_{h,0|\Gamma_n} \times V_h \times V_{h,0|\Gamma_n}$ for the test function $(\boldsymbol{v}, \boldsymbol{m}, q, \mu)$. The equations (61)-(64) can be regarded as as a system of N nonlinear equations for the degrees of freedom of the solution $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda)$. We can solve it using a nonlinear solver such as Newton's method. It turns out that each iteration of Newton's method is equivalent to solving the following linear variational problem for the increment $(\boldsymbol{w}, \boldsymbol{l}, o, \gamma)$:

$$a((\boldsymbol{w}, \boldsymbol{l}, o, \gamma), (\boldsymbol{v}, \boldsymbol{m}, q, \mu)) = L(\boldsymbol{v}, \boldsymbol{m}, q, \mu), \tag{103}$$

where the bilinear form is

$$\begin{aligned} a((\boldsymbol{w}, \boldsymbol{l}, o, \gamma), (\boldsymbol{v}, \boldsymbol{m}, q, \mu)) =& a_1(\boldsymbol{w}, \boldsymbol{v}) + a_2(\boldsymbol{l}, \boldsymbol{v}) + a_2(\boldsymbol{m}, \boldsymbol{w}) + a_3(\boldsymbol{l}, \boldsymbol{m}) \\ &+ b_1(o, \boldsymbol{v}) + b_1(q, \boldsymbol{w}) + b_2(\gamma, \boldsymbol{m}) + b_2(\mu, \boldsymbol{l}), \end{aligned} \tag{104}$$

and the linear form is

$$L(\boldsymbol{v}, \boldsymbol{m}, q, \mu) = -\left(F_1(\boldsymbol{v}) + F_2(\boldsymbol{m}) + F_3(q) + F_4(\mu) \right). \tag{105}$$

[4]Here we have used ϕ_i to denote basis function of the finite element space $\mathbb{V}_h$.

Here F_1, F_2, F_3 and F_4 are the right hand sides of (61)-(64) respectively. The above observation can be verified by computing the derivative matrix and the right hand side of Newton's method, and comparing them with the matrix and the right hand side of the above linear variational problem.

Another complication is that FEniCS does not support the interpolation operator π_h in their form file, at least not for the version 11.02 that we have used. We overcomed this issue in the following way: we first let FEniCS assemble the matrix and the right hand side without the π_h terms, then we manually assembled those terms and updated the matrix and the right hand side. It turns out that we do not have to do numerical integration ourselves, instead we can compute those π_h terms using the degrees of freedom of $\boldsymbol{n}_h$ and λ_h, and the matrix $S = ((\varphi_i, \phi_j))$, where the φ_i's denote the basis functions for the finite element space V_h of piecewise linear functions. The details can be found in [29].

Finally, to compute the order of convergence, we need to compute the H^{-1} norm for any function in $V_{h,0|\Gamma_n}$. In the rest of this subsection, we explain how to assemble the H^{-1} norm.

We first relate the $H^{-1}_{\Gamma_n}$ norm of a function in $V_{h,0|\Gamma_n}$ to the H^1 norm of some other function in $V_{h,0|\Gamma_n}$. For any function v_h in $V_{h,0|\Gamma_n}$, we can define a linear functional g on $H^1_{0|\Gamma_n}$ by

$$g(w) = \langle v_h, w\rangle_{L^2}, \quad \forall w \in H^1_{0|\Gamma_n}.$$

The $H^{-1}_{\Gamma_n}$ norm of v_h is the same as the norm of the functional g. By the Riesz Representation Theorem, we can find $v \in H^1_{0|\Gamma_n}$, such that

$$g(w) = \langle w, v\rangle_{H^1}, \quad \forall w \in H^1_{0|\Gamma_n}.$$

Thus the norm of g is just the H^1 norm of v. Therefore we get

$$\|v_h\|_{H^{-1}_{\Gamma_n}} = \|v\|_{H^1}. \tag{106}$$

Let $\hat{v}_h$ be the L^2 projection of v into $V_{h,0|\Gamma_n}$, then the H^1 norm of v can be approximated by the H^1 norm of $\hat{v}_h$.

Next, we explain how to calculate the H^1 norm of $\hat{v}_h$, which can be used to approximate the $H^{-1}_{\Gamma_n}$ norm of v_h. Let $\{\varphi_i, i = 1, ..., n\}$ be a basis of $V_{h,0|\Gamma_n}$. We want to assemble the matrix S such that

$$\|v_h\|_{H^{-1}_{\Gamma_n}} \approx \|\hat{v}_h\|_{H^1} = \boldsymbol{v}^T S \boldsymbol{v},$$

where $\boldsymbol{v} \in \mathbb{R}^n$ is the degree of freedom for v_h.

Theorem 16. *Let A and B be the matrices that satisfy*

$$\|v_h\|_{L^2} = \boldsymbol{v}^T A \boldsymbol{v},$$
$$\|v_h\|_{H^1} = \boldsymbol{v}^T B \boldsymbol{v},$$

for any v_h in $V_{h,0|\Gamma_n}$, where $\boldsymbol{v} \in \mathbb{R}^n$ is the degree of freedom for v_h. Then the matrix $S = AB^{-1}A$.

Proof. Let $f : H^{-1}_{\Gamma_n} \to V_{h,0|\Gamma_n}$ be the map taking any $v_h \in V_{h,0|\Gamma_n}$ to $\hat{v}_h \in V_{h,0|\Gamma_n}$, and let $\hat{\varphi}_i = f(\varphi_i)$. It is easy to see that

$$S_{ij} = \langle \hat{\varphi}_i, \hat{\varphi}_j\rangle_{H^1}. \tag{107}$$

By definition of $\hat{\varphi}_i$, we have

$$\int \varphi_i \varphi_j = \int D\hat{\varphi}_i D\varphi_j + \int \hat{\varphi}_i \varphi_j \qquad \forall 1 \le i, j \le n. \tag{108}$$

Since $\hat{\varphi}_i \in V_{h,0|\Gamma_n}$, we can write

$$\hat{\varphi}_i = \sum_k G_{ik} \varphi_k.$$

Substituting it into (108) gives

$$\int \varphi_i \varphi_j = \sum_k G_{ik} \left(\int D\varphi_k \cdot D\varphi_j + \int \varphi_k \varphi_j \right).$$

That is $A = GB$, or $G = AB^{-1}$. Therefore

$$\begin{aligned} S &= (\langle \hat{\varphi}_i, \hat{\varphi}_j \rangle_{H^1}) \\ &= (\langle D\hat{\varphi}_i, D\hat{\varphi}_j \rangle + \langle \hat{\varphi}_i, \hat{\varphi}_j \rangle) \\ &= \left(\sum_{p,q} G_{ip} G_{jq} \left[\langle D\varphi_p, D\varphi_q \rangle + \langle \varphi_p, \varphi_q \rangle \right] \right) \\ &= \left(\sum_{p,q} G_{ip} B_{pq} G^T_{qj} \right) \\ &= GBG^T \\ &= (AB^{-1})B(B^{-1}A) \\ &= AB^{-1}A. \end{aligned}$$

□

Remark For any v_h in $V_{h,0|\Gamma_n}$, although $\boldsymbol{v}^T S \boldsymbol{v}$ gives approximate estimate of its H^{-1} norm, the difference goes to zero when h goes to 0.

5 Numerical results

In this section, we present results of the numerical simulation of the clamped-pulling experiment.

The simulation setup is as follows (Figure 1). The LCE is initially rectangular shaped and the directors align in the vertical direction. It is then clamped on the left and right edges and pulled in the horizontal direction.

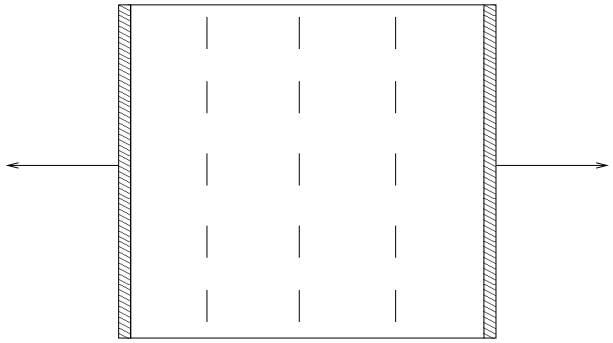

Figure 1: The elastomer is clamped and pulled on both sides.

As previously pointed out, in our model the stress-free state is different from the reference state. In the model and subsequent computation, $\boldsymbol{u}$ represents the displacement relative to the reference domain. However the LCE should be in the stress-free state before it is clamped and pulled. This is not a big issue, because the stress-free state actually has constant deformation gradient matrix F, which means that it can be achieved by a uniform stretch from the reference state. Thus, at both the reference and the stress-free states, the LCE is rectangular, and displacements relative to the reference state or the stress-free state can be easily converted into each other. We take our reference domain to be the rectangle $[0, L] \times [0, 1]$. It can be verified that for the aspect ratio at the stress-free state to be AR, we should take

$$L = \frac{1}{\sqrt{a}} \mathrm{AR}. \tag{109}$$

We give the following starting values for $(\boldsymbol{u}, \boldsymbol{n}, p, \lambda)$, so that they correspond to the stress-free state:

$$u_X = (a^{1/4} - 1)(X - 0.5L), \tag{110}$$

$$u_Y = (a^{-1/4} - 1)(Y - 0.5), \tag{111}$$

$$\boldsymbol{n} \equiv (0, 1)^T, \tag{112}$$

$$p = 2\sqrt{a}, \tag{113}$$

$$\lambda = (1 - a)/\sqrt{a}, \tag{114}$$

where u_X and u_Y are the components of $\boldsymbol{u}$.

The physics of "clamped-pulling" can be modeled by the following boundary conditions: at the two clamped edges, u_Y and $\boldsymbol{n}$ remain at the starting values, while u_X decreases or increases uniformly (that is, independent of Y). Although our model was not formulated as a time-dependent problem, we can still obtain information on the dynamical behavior by solving a series of static problems, each of which only differ slightly from the previous one in the u_X boundary condition.

Notice that the problem is completely symmetric about the two center lines $X = 0.5L$ and $Y = 0.5$. Therefore we only need to do the computation on the upper-right quarter of the reference domain. The solution on the rest of the domain can be obtained by reflection.

Based on the discussion above, we list here the boundary conditions on the computation domain $[0.5L, L] \times [0.5, 1]$. Fist, to model the clamped-pulling set up, we impose the following Dirichlet boundary conditions at the clamped edge $X = L$:

$$u_X = 0.5L[a^{1/4}(1 + Mt) - 1], \tag{115}$$

$$u_Y = (a^{-1/4} - 1)(Y - 0.5), \tag{116}$$

$$\boldsymbol{n} = (0, 1)^T. \tag{117}$$

That is, both u_Y and $\boldsymbol{n}$ remain at their starting values, while u_X varies with t. Here $t \in [0, 1]$ is the percentage of the loading. When $t = 1$, the LCE reaches its maximum elongation $1 + M$, where *elongation* is defined as the current length divided by the starting length (length at the stress-free state). Note that, by symmetry, the vertical center line remains at $X = 0.5L$, while the horizontal center line stays at $Y = 0.5$. Also by symmetry, the directors at these center lines must be either strictly vertical or strictly horizontal. We assume that the directors change continuously during the pulling process, thus the directors at the two center lines must stay at their starting values. Therefore we impose the following boundary conditions at the two center lines:

$$u_X = 0 \qquad \text{on } X = 0.5L, \tag{118}$$

$$u_Y = 0 \qquad \text{on } Y = 0.5, \tag{119}$$

$$\boldsymbol{n} = (0, 1)^T \qquad \text{on } X = 0.5L \text{ and } Y = 0.5. \tag{120}$$

Finally, to ensure that $|\boldsymbol{n}| = 1$ at all the mesh nodes, we need to impose Dirichlet boundary condition for λ on the same boundary as $\boldsymbol{n}$. Thus the boundary condition for λ is

$$\lambda = (1 - a)/\sqrt{a} \quad \text{ on } X = 0.5L, X = L \text{ and } Y = 0.5. \tag{121}$$

In our computation, we take $a = 0.6$, $b = 0.0015$, and $M = 0.4$. We slowly increase "load" t from 0 to 1 in a step size $\triangle t = 0.01$. We take the initial aspect ratio AR to be either 1 or 3. We use uniform mesh of size $(\mathrm{AR} \cdot N) \times N$, where N is an integer. Each small rectangle of the mesh contains two triangles which are split by the lower-left to upper-right diagonal.

Table 1 lists the numerical errors and orders of convergence. Here e_u, e_n, e_p, and e_λ are the numerical errors for $\boldsymbol{u}$, $\boldsymbol{n}$, p and λ, respectively. And $\|\cdot\|_0$, $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ represent the L^2, H^1 and H^{-1} norms, respectively. The numerical errors are calculated for the solutions of adjacent meshes. For instance, let us consider e_u. We first compute the solution $\boldsymbol{u}^{(N)}$ and $\boldsymbol{u}^{(2N)}$ on the mesh N and $2N$ respectively, next we interpolate the solution $\boldsymbol{u}^{(N)}$ to the mesh $2N$, and finally we compute the difference of that interpolation with the solution $\boldsymbol{u}^{(2N)}$ and obtain e_u. The order of convergence is calculated in the usual sense. Take $\|e_u\|_0$ for example, the order of convergence is calculated by

$$\frac{\log(\|e_u\|_0^{(N/2)} / \|e_u\|_0^{(N)})}{\log(2)}.$$

	AR = 1				AR = 3			
	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 4$	$N = 8$	$N = 16$	$N = 32$
$\|e_u\|_0$	1.91E-03	8.39E-04	2.69E-04	6.99E-05	3.88E-03	1.51E-03	5.18E-04	1.41E-04
order	-	1.18	1.64	1.95	-	1.36	1.55	1.88
$\|e_u\|_1$	3.77E-02	2.02E-02	7.66E-03	3.32E-03	5.06E-02	2.14E-02	8.35E-03	3.57E-03
order	-	0.90	1.40	1.21	-	1.24	1.36	1.23
$\|e_n\|_0$	9.70E-02	3.05E-02	8.25E-03	2.12E-03	1.16E-01	3.79E-02	1.16E-02	3.20E-03
order	-	1.67	1.88	1.96	-	1.61	1.71	1.86
$\|e_n\|_1$	1.91E+00	1.19E+00	6.23E-01	3.14E-01	2.53E+00	1.48E+00	7.64E-01	3.81E-01
order	-	0.69	0.93	0.99	-	0.78	0.95	1.00
$\|e_p\|_0$	7.93E-02	2.38E-02	8.99E-03	3.34E-03	5.60E-02	2.14E-02	8.22E-03	2.79E-03
order	-	1.74	1.41	1.43	-	1.39	1.38	1.56
$\|e_\lambda\|_{-1}$	4.41E-03	1.51E-03	5.22E-04	1.70E-04	4.85E-03	1.95E-03	6.15E-04	1.92E-04
order	-	1.55	1.54	1.62	-	1.32	1.66	1.68

Table 1: The numerical errors and orders of convergence.

	AR = 1				AR = 3			
$t = 0$	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 2$	$N = 4$	$N = 8$	$N = 16$
β_1	0.5836	0.5875	0.5879	0.5880	0.5883	0.5877	0.5879	0.5880
β_2	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
β	3.60E-03	2.70E-04	5.69E-05	1.62E-04	5.78E-04	1.61E-04	1.21E-05	3.60E-05
α	-3.60E-03	-1.27E-02	-1.78E-02	-1.21E-02	-4.96E-02	-7.87E-02	-5.42E-02	-5.32E-02
$t = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 16$	$N = 2$	$N = 4$	$N = 8$	$N = 16$
β_1	0.6549	0.6431	0.6287	0.6163	0.6465	0.6229	0.6125	0.6025
β_2	1.9967	1.9503	1.9065	1.8711	1.9688	1.8737	1.7804	1.7517
β	2.91E-03	1.20E-03	5.82E-04	4.88E-05	2.51E-03	3.48E-04	1.46E-04	2.55E-04
α	-2.91E-03	-2.58E-03	-5.82E-04	-4.88E-05	-1.87E-02	-2.50E-03	-3.09E-03	-3.34E-03

Table 2: The inf-sup values and ellipticity constants.

Table 1 shows that the L^2 errors of $\boldsymbol{u}$ and $\boldsymbol{n}$ converge at rates close to 2, while their H^1 errors converge at rates close to 1. The L^2 error of p and the H^{-1} error of λ converge at rates of at least 1.

Table 2 lists the inf-sup values and the ellipticity constants at both the initial state ($t = 0$) and at the final state ($t = 1$) of the pulling process. Here β_1 and β_2 are the inf-sup values of the bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$, respectively, while β and α are the inf-sup value and ellipticity constant, respectively, of the bilinear form $a(\cdot,\cdot)$ on $\mathrm{Ker}(\mathscr{B})$. The eigenvalue decomposition, singular value decomposition and QR decomposition were done using the open source library ALGLIB 2.6 [4]. Notice that for all cases in Table 2, the α's are negative, while β_1, β_2 and β's are all positive. This means that, in all these cases, although the ellipticity conditions for $a(\cdot,\cdot)$ are not satisfied, the inf-sup conditions for $b_1(\cdot,\cdot)$, $b_2(\cdot,\cdot)$ and $a(\cdot,\cdot)$ are all satisfied, and therefore, the linearized system is well-posed. Furthermore, for both $t = 0$ and $t = 1$, the inf-sup values β_1 and β_2 do not seem to change very much as the mesh refines. This suggests that the inf-sup values for $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ might have a constant positive lower bound during the whole pulling process, for all uniform meshes. On the other hand, this is not the case for the inf-sup value β. There is no obvious constant positive lower bound for β. In the case that AR = 1 and $t = 1$, the β values even seem to go to zero as the mesh keeps on refining.

Next we check the stress-strain curve for semi-soft elasticity. Figure 2 and 3 show the stress-strain curves for AR = 1 and AR = 3. In these figures, the $x-$axis is the strain, which is calculated by Mt, while the $y-$axis is the nominal stress, which is calculated by

$$\int_\Gamma \sigma(t)\boldsymbol{\nu}\cdot\boldsymbol{\nu}da, \tag{122}$$

	AR = 1				AR = 3			
	$N = 4$	$N = 8$	$N = 16$	$N = 32$	$N = 4$	$N = 8$	$N = 16$	$N = 32$
left	0.096	0.076	0.076	0.076	0.048	0.040	0.036	0.036
right	0.288	0.272	0.264	0.264	0.276	0.288	0.292	0.292

Table 3: The endpoints of the soft regime.

where Γ is the clamped edge $X = L$, and $\boldsymbol{\nu}$ is the normal vector on Γ. In both figures, the LCE is first hard, then soft, then hard again. Therefore we have successfully recovered the semi-soft elasticity.

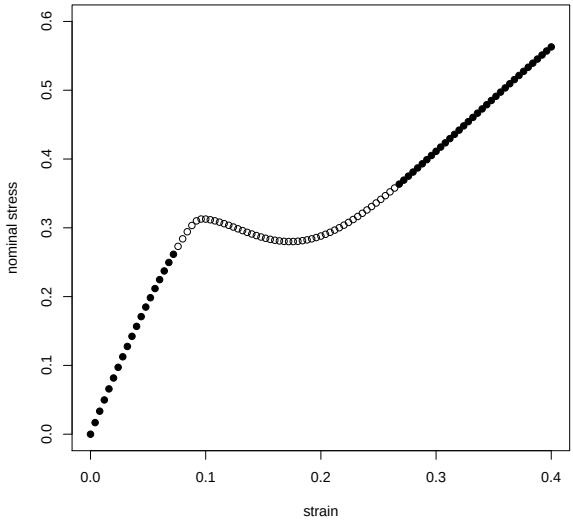

Figure 2: Nominal stress versus strain for AR = 1. Mesh size $N = 32$. The empty circles correspond to the soft regime $[0.076, 0.264]$, which is determined using a curvature criteria.

To check how the soft regime changes with the meshes, we list in Table 3 the endpoints of the soft regime. Since the hard regimes have relatively small curvature, while the soft regimes have relatively large curvature, we choose the endpoints of the soft regime to be those strain values when $|\kappa|$ is first and last bigger than 1, where κ is the curvature of the stress-strain curve. The curvature κ is calculated by

$$\kappa = \frac{f''}{(1 + f'^2)^{3/2}},$$

where f' is the first derivative approximated using forward difference, while f'' is the second derivative approximated using central difference. From Table 3, we can see that as the mesh refines, the soft regime for AR = 1 converges to $[0.076, 0.264]$, while the soft regime for AR = 3 converges to $[0.036, 0.292]$.

To see what the solutions in different regimes of the stress-strain curve look like, we plot some typical solutions in Figure 4 and Figure 5. In both figures, the top-left is a solution in the first hard regime, the top-right is a solution at the start of the soft regime, the bottom-left is a solution at the end of the soft regime, while the bottom-right is a solution in the second hard regime. We can see that the solutions in the first hard regime have most directors vertical, the solutions in the second hard regime have most directors horizontal, while the solutions in the soft regime have directors rotating from vertical to horizontal. This suggests that the soft regime in the stress-strain curve might be related to the rotating of the directors. Also, we can see that the solutions in the soft regime maintain relative low BTW energy, while the solutions in the second hard regime have much higher BTW energy.

Finally we see from Figure 4 and Figure 5 that stripe domain is not observed in these solutions. Instead, the solutions look very smooth. This might be due to the relatively coarse meshes that we have used. Due

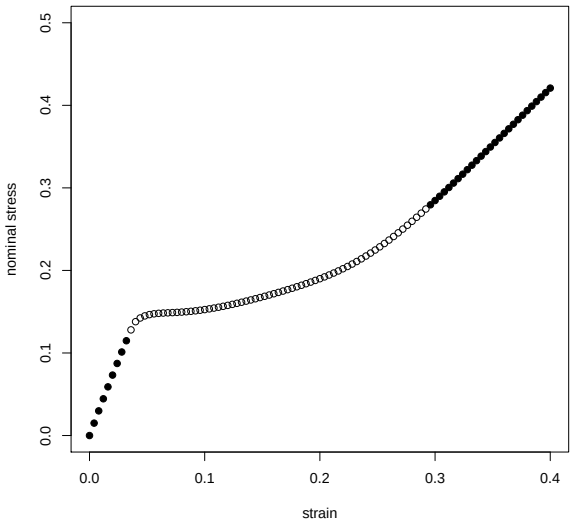

Figure 3: Nominal stress versus strain for AR = 3. Mesh size $N = 32$.The empty circles correspond to the soft regime $[0.036, 0.292]$, which is determined using a curvature criteria.

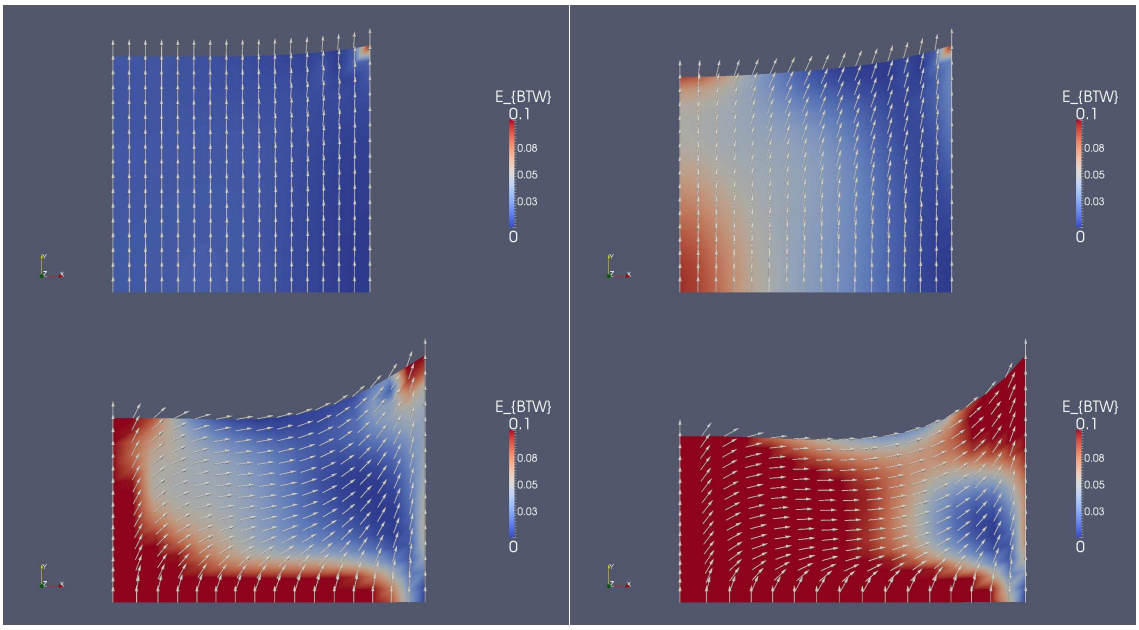

Figure 4: The solutions for AR = 1 with mesh size $N = 16$. From top-left to bottom-right, the strains are 0.040, 0.100, 0.264, 0.400. The domain is colored by the BTW energy with blue corresponds to low BTW energy, while red corresponds to high BTW energy.

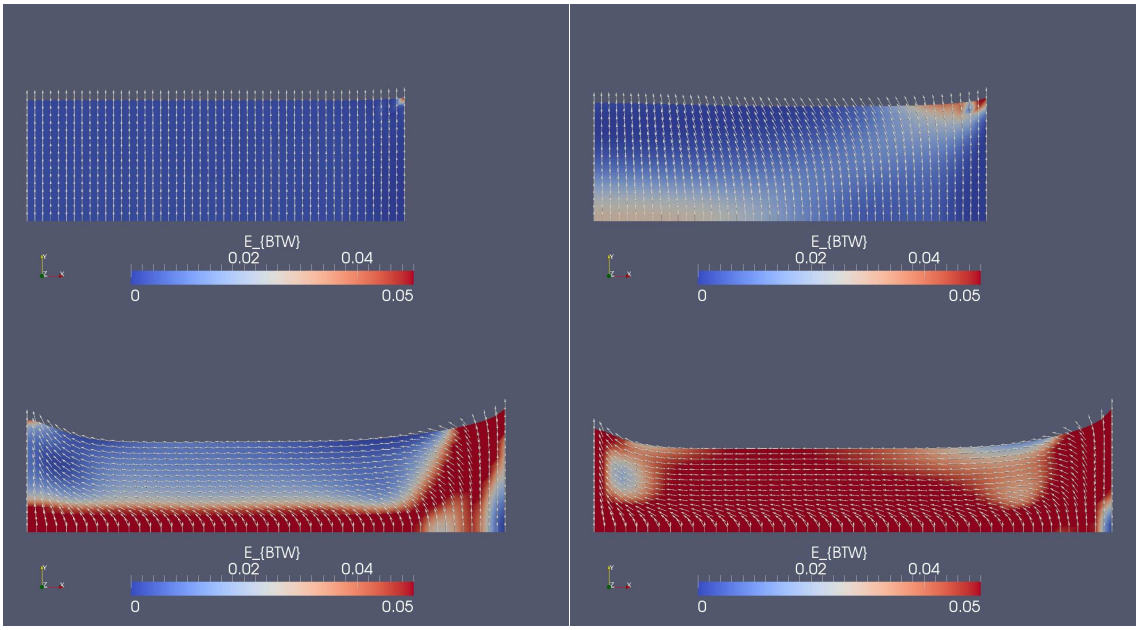

Figure 5: The solutions for AR = 3 with mesh size $N = 16$. From top-left to bottom-right, the strains are 0.020, 0.060, 0.292, 0.400. The domain is colored by the BTW energy with blue corresponds to low BTW energy, while red corresponds to high BTW energy.

to the intrinsic high degress of freedom of our model, the finest mesh we have used has $N = 64$, which might still be too coarse for the development of stripe domains. Another possible reason is that $b = 0.0015$ might be relatively too large, which penalizes the changing in $\boldsymbol{n}$, thus prevents the formation of stripe domains.

6 Conclusion

In this paper, we modeled the liquid crystal elastomer using 2D BTW energy and one-constant Oseen-Frank energy. We imposed the constraint of incompressibility of the bulk, and the unity of the directors to the admissible set of $(\boldsymbol{u}, \boldsymbol{n})$. We proved the existence of minimizer for this energy minimization problem. Then we converted the constrained minimization problem to an unconstrained minimization problem, by introducing Lagrange multipliers p and λ. Next we derived the equilibrium equation and its linearization. We reduced the linearized system to a standard saddle point system, and verified the well-posedness for some simple cases.

Next, we proposed the corresponding discrete problem, which used the $\mathbf{P}_2 \times P_1$ Taylor-Hood element for the $(\boldsymbol{u}, p)$ combination and the $\mathbf{P}_1 \times P_1$ element for the $(\boldsymbol{n}, \lambda)$ combination. We also imposed the constraints that the L^2 projection of $\det(F) - 1$ is zero and that $|\boldsymbol{n}| - 1$ is zero at all the mesh nodes. We proved the existence of minimizer for this discrete energy minimization problem. Similar to the continuous case, we then introduced the Lagrange multipliers p and λ, derived the equilibrium equation and its linearization, and reduced the linearized system to a standard saddle point system. We verified the well-posedness for some simple cases. For the general cases, we explained how to reduce the verification of inf-sup conditions to the computation of smallest singular value of certain matrices. Next, we proved the existence and uniqueness of the Lagrange multipliers p and λ, under the condition that both inf-sup conditions are satisfied.

Finally, we used finite element method on our model to simulate the clamped-pulling experiment, for elastomer samples with aspect ratio AR = 1 or 3. The orders of convergence and inf-sup values were listed. The stress-strain curves were plotted. For both AR = 1 and 3, the semi-soft elasticity was observed. However, the stripe domain phenomenon was not observed, which might due to the relative coarse meshes in the computation and the relative large Oseen-Frank coefficient $b = 0.0015$.

7 Discussion

Although we have successfully recovered the semi-soft elasticity phenomenon, the exclusion of the stripe domain phenomenon is tentative. This is mainly because we only applied computation for meshes with size up to $N = 64$. That is, the ratio of the edge length of the triangular elements to the edge length of the domain is around 10^{-2}. However, in the experiment of Finkelmann et al. [25, 35], the ratio of the width of the stripe domains to the edge length of the domain was around 10^{-3}. Thus our mesh might be too coarse to resolve the stripe domains. We did not use finer mesh because the computational cost was already very high. Even for the mesh $N = 64$, we have 50,182 degrees of freedom to solve in the case $\mathrm{AR} = 1$, and 149,510 degrees of freedom to solve in the case $\mathrm{AR} = 3$. Another possible reason is that the Oseen-Frank coefficient $b = 0.0015$ was too large. The zig-zag pattern of the stripe domain phenomenon naturally has very rapid change of $\boldsymbol{n}$ across the domain. However, a relatively large b value penalizes such rapid change, and suppresses the occurrence of stripe domains. We did not take much smaller b values than 0.0015, because that would require much finer mesh and much smaller $\triangle t$ to stabilize, which was computationally too demanding.

Stripe domain phenomenon might still be observable with our current model, if we try some more sophisticated numerical techniques. As remarked above, the main obstacle might be the computational cost. One way to get around this obstacle is to use adaptive mesh refinement. Since the stripe domains only occur in part of the elastomer domain, while $\boldsymbol{n}$ in the rest of the domain is quite smooth, we can save computational cost by refining the mesh only on part of the domain. Another way to reduce the computaional cost, is to replace $\boldsymbol{n}$ by $(\cos(\theta), \sin(\theta))^T$, where θ is the azimuthal angle of the director. This is perfectly fine because our model is in 2D. In this way, we can reduce a 2D variable to a 1D variable, and also eliminate the need to use the Lagrange multiplier λ.

Another direction is to replace the Oseen-Frank model by more advanced models such as Ericksen model [19] or Landau-de Gennes model [15]. Oseen-Frank energy only allows point defects, while Ericksen and Landau-de Gennes model allow line and surface defects, as well [30]. The stripe domains might have line or surface defects in the transition area between the stripes, thus using Ericksen or Landau-de Gennes model might have a better chance of capturing the stripe domain phenomenon.

References

[1] R. Adams and J. Fournier, *Sobolev spaces*, vol. 65, Academic press New York, 1975.

[2] J. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Archive for rational mechanics and Analysis, 63 (1976), pp. 337–403.

[3] P. Bladon, E. Terentjev, and M. Warner, *Transitions and instabilities in liquid crystal elastomers*, Physical Review E, 47 (1993), pp. 3838–3840.

[4] S. Bochkanov and V. Bystritsky, *ALGLIB*. Available from `http://www.alglib.net/`.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, Berlin, 1991.

[6] M. Calderer, C. Liu, and B. Yan, *A mathematical theory for nematic elastomers with non-uniform prolate spheroids*, Preprint.

[7] P. Cesana and A. DeSimone, *Strain-order coupling in nematic elastomers: equilibrium configurations*, Math. Models Methods Appl. Sci, (2008).

[8] D. Chapelle and K. Bathe, *The inf-sup test*, Computers and Structures, 47 (1993), pp. 537–537.

[9] P. Ciarlet, *Mathematical Elasticity, Vol 1*, North-Holland, 1987.

[10] F. Clarke, *Generalized gradients and applications*, Transactions of the American Mathematical Society, 205 (1975), pp. 247–262.

[11] ——, *A new approach to Lagrange multipliers*, Mathematics of Operations Research, 1 (1976), pp. 165–174.

[12] S. Clarke, A. Hotta, A. Tajbakhsh, and E. Terentjev, *Effect of crosslinker geometry on equilibrium thermal and mechanical properties of nematic elastomers*, Physical Review E, 64 (2001), p. 61702.

[13] S. Conti, A. DeSimone, and G. Dolzmann, *Semisoft elasticity and director reorientation in stretched sheets of nematic elastomers*, Physical Review E, 66 (2002), p. 061710.

[14] ——, *Soft elastic response of stretched sheets of nematic elastomers: a numerical study*, Journal of the Mechanics and Physics of Solids, 50 (2002), pp. 1431–1451.

[15] P. de Gennes and J. Prost, *The physics of liquid crystals*, Oxford University Press, USA, 1995.

[16] A. Desimone, *Energetics of fine domain structures*, Ferroelectrics, 222 (1999), pp. 533–542.

[17] A. DeSimone and G. Dolzmann, *Material instabilities in nematic elastomers*, Physica D: Nonlinear Phenomena, 136 (2000), pp. 175–191.

[18] A. DeSimone and L. Teresi, *Elastic energies for nematic elastomers*, The European Physical Journal E: Soft Matter and Biological Physics, 29 (2009), pp. 191–204.

[19] J. Ericksen, *Liquid crystals with variable degree of orientation*, Archive for Rational Mechanics and Analysis, 113 (1991), pp. 97–120.

[20] L. Evans, *Partial Differential Equations*, American Mathematical Society, 1998.

[21] F. Frank, *I. Liquid crystals. On the theory of liquid crystals*, Discussions of the Faraday Society, 25 (1958), pp. 19–28.

[22] R. Hardt, D. Kinderlehrer, and F. Lin, *Existence and partial regularity of static liquid crystal configurations*, Communications in Mathematical Physics, 105 (1986), pp. 547–570.

[23] M. Hébert, R. Kant, and P. De Gennes, *Dynamics and thermodynamics of artificial muscles based on nematic gels*, Journal de Physique I, 7 (1997), pp. 909–919.

[24] Q. Hu, X. Tai, and R. Winther, *A saddle point approach to the computation of harmonic maps*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 1500–1523.

[25] I. Kundler and H. Finkelmann, *Strain-induced director reorientation in nematic liquid single crystal elastomers*, Macromolecular rapid communications, 16 (1995), pp. 679–686.

[26] J. Küpfer and H. Finkelmann, *Liquid crystal elastomers: influence of the orientational distribution of the crosslinks on the phase behaviour and reorientation processes*, Macromolecular chemistry and physics, 195 (1994), pp. 1353–1367.

[27] P. Le Tallec, *Compatibility condition and existence results in discrete finite incompressible elasticity*, Computer Methods in Applied Mechanics and Engineering, 27 (1981), pp. 239–259.

[28] A. Logg and G. Wells, *DOLFIN: Automated finite element computing*, ACM Transactions on Mathematical Software (TOMS), 37 (2010), pp. 1–28. Available from `https://launchpad.net/fenics`.

[29] C. Luo, *Modeling, analysis and numerical simulation of liquid crystal elastomer*, PhD thesis, University of Minnesota, 2010.

[30] A. Majumdar and A. Zarnescu, *Landau–De Gennes Theory of Nematic Liquid Crystals: the Oseen–Frank Limit and Beyond*, Archive for rational mechanics and analysis, 196 (2010), pp. 227–280.

[31] W. Rudin, *Functional analysis. International Series in Pure and Applied Mathematics*, 1991.

[32] P. L. Tallec, *Numerical methods for nonlinear three-dimensional elasticity*, Handbook of Numerical Analysis, 3 (1994), pp. 465–622.

[33] G. Verwey, M. Warner, and E. Terentjev, *Elastic instability and stripe domains in liquid crystalline elastomers*, J. Phys. II France, 6 (1996), pp. 1273–1290.

[34] M. Warner and E. Terentjev, *Liquid crystal elastomers*, Oxford University Press, USA, 2007.

[35] E. Zubarev, S. Kuptsov, T. Yuranova, R. Talroze, and H. Finkelmann, *Monodomain liquid crystalline networks: reorientation mechanism from uniform to stripe domains*, Liquid crystals, 26 (1999), pp. 1531–1540.

RECENT REPORTS

07/11	Strong stability preserving two-step Runke-Kutta methods	Ketcheson Gotlieb MacDonald
08/11	Hysteresis and Post Walrasian Economics	Cross McNamara Kalachev Pokrovskii
09/11	A locally adaptive time-stepping algorithm for petroleum reservoir simulations	McNamara Bowen Dellar
10/11	On the predictions and limitations of the BeckerDoring model for reaction kinetics in micellar surfactant solutions	Griffiths Bain Breward Colegate Howell Waters
11/11	Dynamics of the Tear Film	Braun
12/11	The infuence of receptor-mediated interactions on reaction-diffusion mechanisms of cellular self-organisation	Klikaa Baker Headon Gaffney
13/11	Quasi-steady state analysis of two-dimensional random intermittent search processes	Bressloff Newby
14/11	A Constrained Approach to Multiscale Stochastic Simulation of Chemically Reacting Systems	Cotter Zygalakis Kevrekidis Erban
15/11	The Two Regime Method for optimizing stochastic reaction-diffusion simulations	Flegg Chapman Erban
16/11	Recombination via tail states in polythiophene:fullerene solar cells	Kirchartz Pieters Kirkpatrick Rau Nelson
17/11	Energy versus electron transfer in organic solar cells: a comparison of the photophysics of two indenofluorene: fullerene blend films	Soon Clarke Zhang Agostinelli Kirkpatrick Dyer-Smith McCulloch Nelson Durrant
18/11	Asymptotic analysis of a pile-up of edge dislocation	Hall
19/11	A perturbation analysis of spontaneous action potential initiation by stochastic ion channels	Keener1 Newby

23/11	Positive or negative Poynting efect? The role of adscititious inequalities in hyperelastic materials	Mihai Goriely McCue McElwain
24/11	On approaches to modelling lattice dislocations	Hall Markenscoff
25/11	Nonlinear waves in heterogeneous elastic rods via homogenization	de Luna Emptage Goriely Bressloff
26/11	Synaptic bistability due to nucleation and evaporation of receptor clusters	Burlakov Duričković Goriely
27/11	Particle trapping and banding in rapid solidification	Elliot Peppin
28/11	Growth of confined cancer spheroids: a combined experimental and mathematical modelling approach	Loessner Flegg Byrne Hall Moroney Clements McElwain Hutmacher
29/11	Floating carpets and the delamination of elastic sheets	Wagner Vella

Copies of these, and any other OCCAM reports can be obtained from:

Oxford Centre for Collaborative Applied Mathematics
Mathematical Institute
24 - 29 St Giles'
Oxford
OX1 3LB
England
www.maths.ox.ac.uk/occam