SIAM J. NUMER. ANAL.
Vol. 53, No. 5, pp. 2226–2254

AN ENERGY-MINIMIZATION FINITE-ELEMENT APPROACH FOR THE FRANK–OSEEN MODEL OF NEMATIC LIQUID CRYSTALS*

J. H. ADLER†, T. J. ATHERTON‡, D. B. EMERSON†, AND S. P. MACLACHLAN§

Abstract. This paper outlines an energy-minimization finite-element approach to the computational modeling of equilibrium configurations for nematic liquid crystals under free elastic effects. The method targets minimization of the system free energy based on the Frank–Oseen free-energy model. Solutions to the intermediate discretized free elastic linearizations are shown to exist generally and are unique under certain assumptions. This requires proving continuity, coercivity, and weak coercivity for the accompanying appropriate bilinear forms within a mixed finite-element framework. Error analysis demonstrates that the method constitutes a convergent scheme. Numerical experiments are performed for problems with a range of physical parameters as well as simple and patterned boundary conditions. The resulting algorithm accurately handles heterogeneous constant coefficients and effectively resolves configurations resulting from complicated boundary conditions relevant in ongoing research.

Key words. nematic liquid crystals, mixed finite elements, saddle-point problem, Newton linearization, energy optimization

AMS subject classifications. 76A15, 65H10, 65N30, 49M15, 65N22

DOI. 10.1137/140956567

1. Introduction.

Liquid crystals, whose discovery is attributed to Reinitzer in 1888 [44], are substances that possess mesophases with properties intermediate between liquids and crystals. The mesophases exist at different temperatures or solvent concentrations. In recent years, research on the novel properties of liquid crystals has rapidly expanded. Modern applications include nanoparticle organization, liquid crystal-functionalized polymer fibers [32], and liquid crystal elastomers designed to produce effective actuator devices such as light-driven motors [49] and artificial muscles [47].

The focus of this paper is on nematic liquid crystal phases, which are formed by rod-like molecules that self-assemble into an ordered structure, such that the molecules tend to align along a preferred orientation. The preferred average direction at any point in a domain, Ω, is known as the director, denoted $\mathbf{n}(x, y, z) = (n_1, n_2, n_3)^T$. The director is taken to be of unit length at every point and headless, that is, $\mathbf{n}$ and $-\mathbf{n}$ are indistinguishable, reflecting the observed experimental symmetry of the phase.

In addition to their self-structuring properties, the orientation of a nematic liquid crystal may be affected by applied electric fields. Moreover, since these materials are birefringent, that is, these materials' refractive indices depend on the polarization of light, they can be used to control the propagation of light through a nematic structure. These traits have led, and continue to lead, to important discoveries in

*Received by the editors February 11, 2014; accepted for publication (in revised form) July 21, 2015; published electronically October 1, 2015.
http://www.siam.org/journals/sinum/53-5/95656.html
†Department of Mathematics, Tufts University, Medford, MA 02155 (James.Adler@tufts.edu, David.Emerson@tufts.edu).
‡Department of Physics, Tufts University, Medford, MA 02155 (Timothy.Atherton@tufts.edu).
§Department of Mathematics and Statistics, Memorial University of Newfoundland and Labrador, St. John's, Newfoundland and Labrador A1C 5S7, Canada (smaclachlan@mun.ca).

display technologies and beyond [32]. Thorough overviews of liquid crystal physics and properties are found in [14, 19, 46].

Many mathematical and computational models of liquid crystal continuum theory lead to complicated systems involving unit-length constrained vector fields. Currently, the complexity of such systems has restricted the existence of known analytical solutions to simplified geometries in one or two dimensions, often under strong simplifying assumptions. When coupled with electric fields and other effects, far fewer analytical solutions exist, even in one dimension [46]. In addition, associated systems of partial differential equations, such as the equilibrium equations [22,46], suffer from nonunique solutions, which must be distinguished via energy arguments. Due to such difficulties, efficient, theoretically supported, numerical approaches to the modeling of nematic liquid crystals under free elastic and augmented electric effects are of great importance. This paper discusses the modeling of free elastic effects. The addition of electric field effects will be the subject of future work. A number of computational techniques for liquid crystal equilibrium [16, 41, 46] and dynamics problems [34, 35, 36, 48, 50] exist, including least-squares finite-element methods [4] and discrete Lagrange multiplier approaches [31, 43].

In this paper, we propose a method that directly targets energy minimization in the continuum, via Lagrange multiplier theory on Banach spaces. The approach is derived absent the often used one-constant approximations [16,31,34,35,36,43,46,48]; that is, the method described here and the accompanying theory are applied for a wide range of physical parameters. This allows for significantly improved modeling of physical phenomena not captured in many models. Furthermore, most models and analytical approaches rely on assumptions to reduce the dimensionality of the problem. Here, the method and theory are suitable for use on two-dimensional (2-D) and 3-D domains and are easily combined with additional energy effects.

After defining the energy functional to be minimized, first-order optimality conditions are computed. These first-order conditions contain highly nonlinear terms and are, therefore, linearized with a generalized Newton's method. The resulting Newton linearization resembles a typical mixed finite-element method formulation [10, 12, 13]. However, these forms present unique difficulties not found, for instance, in the Stokes' problem. In particular, the forms related to the nonlinear unit-length constraint for $\mathbf{n}$ require novel treatment. Additionally, the proofs of continuity and coercivity differ significantly from many standard approaches due to the inherent complexity of the bilinear forms.

In the continuum, it is possible to demonstrate coercivity for the relevant bilinear form with moderate simplifying assumptions. With auxiliary regularity assumptions, continuity of the involved bilinear forms is also established. These results are stated herein without proof [2]. On the other hand, for a pair of discrete spaces, continuity, coercivity, and weak coercivity for the relevant bilinear forms are proved. The main result of this paper proves the existence and uniqueness of solutions to each discrete Newton iteration. Error analysis is also performed to elaborate the convergence order of the update approximations with grid size. The method is implemented and run for a number of configurations, including those relevant to ongoing research. We use nested iteration and damping to handle relatively inaccurate initial guesses and rely on the convergence of Newton's method [20, 45] for efficient iterative convergence. While more rigorous outer iterations, as in [24], could guarantee convergence to global minima in all settings, our numerical results show insensitivity in finding the energy-minimizing solution when the initial guess is varied on the coarsest grid.

This paper is organized as follows. We first introduce the liquid crystal model under consideration, derive the method, and discuss Dirichlet boundary condition simplifications in section 2. In section 3, well-posedness of the Newton iterations for a pair of discrete spaces is proved and an error analysis is performed. The numerical methodology and numerical experiments are detailed in section 4. Finally, section 5 gives some concluding remarks and future work is discussed.

2. Energy model. At equilibrium, absent any external forces, fields, or boundary conditions, the free elastic energy present in a liquid crystal sample is given by an integral functional, $\mathcal{F}$, which depends on the state variables of the system. A liquid crystal sample tends to the state of lowest free energy. While a number of free-energy models exist (cf. [18]), this paper considers the Frank–Oseen free elastic model [26, 46]. The Frank–Oseen equations represent the free elastic energy density, w_F, in a sample as

$$w_F = \frac{1}{2}K_1(\nabla\cdot\mathbf{n})^2+\frac{1}{2}K_2(\mathbf{n}\cdot\nabla\times\mathbf{n})^2+\frac{1}{2}K_3|\mathbf{n}\times\nabla\times\mathbf{n}|^2+\frac{1}{2}(K_2+K_4)\nabla\cdot[(\mathbf{n}\cdot\nabla)\mathbf{n}-(\nabla\cdot\mathbf{n})\mathbf{n}].$$

Throughout this paper, the standard Euclidean inner product and norm are denoted $(\cdot,\cdot)$ and $|\cdot|$, respectively. The K_i, $i = 1, 2, 3, 4$, are known as the Frank elastic constants [26], which vary depending on temperature and liquid crystal type. By Ericksen's inequalities [23], $K_j \geq 0$ for $j = 1, 2, 3$. Each term represents an energy penalty for the presence of splay, twist, bend, and saddle-splay, respectively.

It can be shown that

$$\nabla\cdot[(\mathbf{n}\cdot\nabla)\mathbf{n}-(\nabla\cdot\mathbf{n})\mathbf{n}] = \nabla n_1\cdot\frac{\partial\mathbf{n}}{\partial x}+\nabla n_2\cdot\frac{\partial\mathbf{n}}{\partial y}+\nabla n_3\cdot\frac{\partial\mathbf{n}}{\partial z}-(\nabla\cdot\mathbf{n})^2. \tag{2.1}$$

Additionally, let

$$\mathbf{Z} = \kappa\mathbf{n}\otimes\mathbf{n}+(\mathbf{I}-\mathbf{n}\otimes\mathbf{n}) = \mathbf{I}-(1-\kappa)\mathbf{n}\otimes\mathbf{n}, \tag{2.2}$$

where $\kappa = K_2/K_3$; in general, we consider the case that $K_2, K_3 > 0$. Denote the classical $L^2(\Omega)$ inner product and norm as $\langle\cdot,\cdot\rangle_0$ and $\|\cdot\|_0$, respectively. Employing (2.1), (2.2), and the fact that $\mathbf{n}$ has unit length, the total free energy for a domain, Ω, is

$$\begin{aligned}\int_\Omega w_F\,dV = {}& \frac{1}{2}(K_1-K_2-K_4)\|\nabla\cdot\mathbf{n}\|_0^2+\frac{1}{2}K_3\left\langle\mathbf{Z}\nabla\times\mathbf{n},\nabla\times\mathbf{n}\right\rangle_0\\ &+\frac{1}{2}(K_2+K_4)\left(\left\langle\nabla n_1,\frac{\partial\mathbf{n}}{\partial x}\right\rangle_0+\left\langle\nabla n_2,\frac{\partial\mathbf{n}}{\partial y}\right\rangle_0+\left\langle\nabla n_3,\frac{\partial\mathbf{n}}{\partial z}\right\rangle_0\right).\end{aligned} \tag{2.3}$$

For the special case of full Dirichlet boundary conditions, we consider a fixed director $\mathbf{n}$ at each point on the boundary of Ω. Considering the integration carried out on the terms in (2.1),

$$\begin{aligned}&\frac{1}{2}(K_2+K_4)\int_\Omega\nabla\cdot[(\mathbf{n}\cdot\nabla)\mathbf{n}-(\nabla\cdot\mathbf{n})\mathbf{n}]\,dV\\ &\qquad=\frac{1}{2}(K_2+K_4)\int_{\partial\Omega}[(\mathbf{n}\cdot\nabla)\mathbf{n}-(\nabla\cdot\mathbf{n})\mathbf{n}]\cdot\nu\,dS,\end{aligned} \tag{2.4}$$

by the divergence theorem. Further, since $\mathbf{n}$ is fixed along $\partial\Omega$, the energy contributed by $\mathbf{n}$ on the boundary is constant regardless of the configuration of $\mathbf{n}$ on the interior

of Ω. Thus, in the minimization to follow, the energy contribution from this term is ignored. For this reason, (2.4) is often referred to as a null Lagrangian [46].

A number of methods involving computation of liquid crystal equilibria or dynamics utilize the so-called one-constant approximation that $K_1 = K_2 = K_3$ and $K_4 = 0$ [16, 43, 46, 48], in order to significantly simplify the free elastic energy density to

$$\hat{w}_F = \frac{1}{2} K_1 |\nabla \mathbf{n}|^2, \qquad \text{where } |\nabla \mathbf{n}|^2 = \sum_{i,j=1}^{3} \left(\frac{\partial n_i}{\partial x_j} \right)^2.$$

This expression for the free-energy density is more amenable to theoretical development but ignores significant physical characteristics of the nematic [5, 33]. The following method is derived without such an assumption.

2.1. Free elastic energy minimization. In this section, a general approach to computing the free elastic equilibrium state for $\mathbf{n}$ is derived. This equilibrium state corresponds to the configuration which minimizes the system free energy subject to the local constraint that $\mathbf{n}$ is of unit length throughout the sample volume, Ω. That is, the minimizer must satisfy $\mathbf{n} \cdot \mathbf{n} = 1$ pointwise throughout the volume. To compute this state, define the functional, equivalent to (2.3),

$$\begin{aligned} \mathcal{F}_1(\mathbf{n}) &= (K_1 - K_2 - K_4) \|\nabla \cdot \mathbf{n}\|_0^2 + K_3 \langle \mathbf{Z} \nabla \times \mathbf{n}, \nabla \times \mathbf{n} \rangle_0 \\ &\quad + (K_2 + K_4) \left(\left\langle \nabla n_1, \frac{\partial \mathbf{n}}{\partial x} \right\rangle_0 + \left\langle \nabla n_2, \frac{\partial \mathbf{n}}{\partial y} \right\rangle_0 + \left\langle \nabla n_3, \frac{\partial \mathbf{n}}{\partial z} \right\rangle_0 \right). \end{aligned} \tag{2.5}$$

Define

$$\begin{aligned} H(\mathrm{div}, \Omega) &= \{ \mathbf{v} \in L^2(\Omega)^3 : \nabla \cdot \mathbf{v} \in L^2(\Omega) \}, \\ H(\mathrm{curl}, \Omega) &= \{ \mathbf{v} \in L^2(\Omega)^3 : \nabla \times \mathbf{v} \in L^2(\Omega)^3 \}. \end{aligned}$$

Further, let

$$\begin{aligned} H_0(\mathrm{div}, \Omega) &= \{ \mathbf{v} \in H(\mathrm{div}, \Omega) : \nu \cdot \mathbf{v} = 0 \text{ on } \partial\Omega \}, \\ H_0(\mathrm{curl}, \Omega) &= \{ \mathbf{v} \in H(\mathrm{curl}, \Omega) : \nu \times \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega \}, \end{aligned}$$

where ν is the outward unit normal for $\partial\Omega$. Define

$$\mathcal{H}^{DC}(\Omega) = \{ \mathbf{v} \in H(\mathrm{div}, \Omega) \cap H(\mathrm{curl}, \Omega) : B(\mathbf{v}) = \mathbf{g} \},$$

with norm $\|\mathbf{v}\|_{DC}^2 = \|\mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2$ and appropriate boundary conditions $B(\mathbf{v}) = \mathbf{g}$. Here, we assume that $\mathbf{g}$ satisfies appropriate compatibility conditions for operator B. For example, if B represents Dirichlet boundary conditions and Ω has a Lipschitz continuous boundary, it is assumed that $\mathbf{g} \in H^{\frac{1}{2}}(\partial\Omega)^3$ [28]. Further, let $\mathcal{H}_0^{DC}(\Omega) = \{ \mathbf{v} \in H(\mathrm{div}, \Omega) \cap H(\mathrm{curl}, \Omega) : B(\mathbf{v}) = \mathbf{0} \}$. Note that if Ω is a Lipshitz domain and B imposes Dirichlet boundary conditions, then $\mathcal{H}_0^{DC}(\Omega) = H_0^1(\Omega)^3$ [28, Lemma 2.5]. Finally, denote the unit sphere as $\mathcal{S}^2$. The desired minimization becomes

$$\mathbf{n}_* = \underset{\mathbf{n} \in \mathcal{S}^2 \cap \mathcal{H}^{DC}(\Omega)}{\mathrm{argmin}} \mathcal{F}_1(\mathbf{n}).$$

Foundational theoretical work establishing the existence of minimizing director fields is considered in [29]. However, as noted in [3, 16], the unit-length constraint leads

to nonconvexity in the minimization problem. While this increases the difficulty of finding global minimizers, nested iteration and damped Newton stepping markedly improve convergence in the numerical results presented below. Moreover, in future work, inclusion of the deflation techniques in [24] could efficiently resolve all solutions, including the global minimizer.

In the presence of full Dirichlet boundary conditions, the functional to be minimized is significantly simplified as

$$\mathcal{F}_2(\mathbf{n}) = K_1\|\nabla\cdot\mathbf{n}\|_0^2 + K_3\left\langle \mathbf{Z}\nabla\times\mathbf{n}, \nabla\times\mathbf{n}\right\rangle_0 \tag{2.6}$$

by the application of (2.4). However, the functional still contains nonlinear terms introduced by the presence of $\mathbf{Z} = \mathbf{Z}(\mathbf{n})$. Note that this simplification is also applicable to a rectangular domain with mixed Dirichlet and periodic boundary conditions. Such a domain is considered in the numerical experiments presented here.

We proceed with the functional in (2.5) in building a framework for minimization under general boundary conditions and, therefore, utilize the $\mathcal{H}^{DC}(\Omega)$ notation throughout this paper. However, in the treatment of existence and uniqueness theory, we assume the application of full Dirichlet or mixed Dirichlet and periodic boundary conditions and, thus, utilize the simplified form in (2.6).

2.2. First-order optimality conditions and Newton linearization. Since $\mathbf{n}$ must be of unit length, it is natural to employ a Lagrange multiplier approach. This length requirement represents a pointwise equality constraint, such that $(\mathbf{n},\mathbf{n})-1=0$. Thus, following general constrained optimization theory [37], define the Lagrangian

$$\mathcal{L}(\mathbf{n},\lambda) = \mathcal{F}_1(\mathbf{n}) + \int_\Omega \lambda(\mathbf{x})((\mathbf{n},\mathbf{n})-1)\,dV,$$

where $\lambda\in L^2(\Omega)$. In order to minimize (2.5), we compute the Gâteaux derivatives of $\mathcal{L}$ with respect to $\mathbf{n}$ and λ in the directions $\mathbf{v}\in\mathcal{H}_0^{DC}(\Omega)$ and $\gamma\in L^2(\Omega)$, respectively. Hence, the necessary continuum first-order optimality conditions are

$$\mathcal{L}_{\mathbf{n}}[\mathbf{v}] = \frac{\partial}{\partial\mathbf{n}}\mathcal{L}(\mathbf{n},\lambda)[\mathbf{v}] = 0 \qquad \forall\mathbf{v}\in\mathcal{H}_0^{DC}(\Omega), \tag{2.7}$$

$$\mathcal{L}_{\lambda}[\gamma] = \frac{\partial}{\partial\lambda}\mathcal{L}(\mathbf{n},\lambda)[\gamma] = 0 \qquad \forall\gamma\in L^2(\Omega). \tag{2.8}$$

Computing these derivatives yields

$$\begin{aligned}\mathcal{L}_{\mathbf{n}}[\mathbf{v}] = {}& 2(K_1-K_2-K_4)\langle\nabla\cdot\mathbf{n},\nabla\cdot\mathbf{v}\rangle_0 + 2K_3\left\langle\mathbf{Z}(\mathbf{n})\nabla\times\mathbf{n},\nabla\times\mathbf{v}\right\rangle_0\\ &+ 2(K_2-K_3)\langle\mathbf{n}\cdot\nabla\times\mathbf{n},\mathbf{v}\cdot\nabla\times\mathbf{n}\rangle_0 + 2(K_2+K_4)\left(\left\langle\nabla n_1,\frac{\partial\mathbf{v}}{\partial x}\right\rangle_0\right.\\ &\left.+\left\langle\nabla n_2,\frac{\partial\mathbf{v}}{\partial y}\right\rangle_0+\left\langle\nabla n_3,\frac{\partial\mathbf{v}}{\partial z}\right\rangle_0\right)+2\int_\Omega\lambda(\mathbf{n},\mathbf{v})\,dV\end{aligned}$$

and

$$\mathcal{L}_{\lambda}[\gamma] = \int_\Omega\gamma((\mathbf{n},\mathbf{n})-1)\,dV.$$

The variational system contains nonlinearities in both (2.7) and (2.8). Therefore, Newton iterations are employed by computing a generalized first-order Taylor series expansion, requiring computation of the Hessian [9, 40].

Let $\mathbf{n}_k$ and λ_k be the current approximations for $\mathbf{n}$ and λ, respectively. Additionally, let $\delta\mathbf{n} = \mathbf{n}_{k+1} - \mathbf{n}_k$ and $\delta\lambda = \lambda_{k+1} - \lambda_k$ be updates to these approximations. Then, the Newton iterations are denoted

$$\begin{bmatrix} \mathcal{L}_{\mathbf{nn}} & \mathcal{L}_{\mathbf{n}\lambda} \\ \mathcal{L}_{\lambda\mathbf{n}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \delta\mathbf{n} \\ \delta\lambda \end{bmatrix} = - \begin{bmatrix} \mathcal{L}_{\mathbf{n}} \\ \mathcal{L}_{\lambda} \end{bmatrix}, \tag{2.9}$$

where each of the system components is evaluated at $\mathbf{n}_k$ and λ_k. The matrix-vector multiplication indicates the direction that the derivatives in the Hessian are taken. That is,

$$\mathcal{L}_{\mathbf{nn}}[\mathbf{v}] \cdot \delta\mathbf{n} = \frac{\partial}{\partial \mathbf{n}} \left(\mathcal{L}_{\mathbf{n}}(\mathbf{n}_k, \lambda_k)[\mathbf{v}] \right) [\delta\mathbf{n}], \quad \mathcal{L}_{\mathbf{n}\lambda}[\mathbf{v}] \cdot \delta\lambda = \frac{\partial}{\partial \lambda} \left(\mathcal{L}_{\mathbf{n}}(\mathbf{n}_k, \lambda_k)[\mathbf{v}] \right) [\delta\lambda],$$
$$\mathcal{L}_{\lambda\mathbf{n}}[\gamma] \cdot \delta\mathbf{n} = \frac{\partial}{\partial \mathbf{n}} \left(\mathcal{L}_{\lambda}(\mathbf{n}_k, \lambda_k)[\gamma] \right) [\delta\mathbf{n}],$$

where the partials denote Gâteaux derivatives in the respective variables.

The discrete form of this Hessian leads to a saddle-point matrix, which poses unique difficulties for the efficient computation of the solution to the resulting linear system. Such structures commonly appear in constrained optimization and other settings; for a comprehensive overview of discrete saddle-point problems see [8]. Here, we focus only on the linearization step rather than the underlying linear solvers, which will be investigated in future work. Computing the Gâteaux derivatives yields

$$\mathcal{L}_{\mathbf{n}\lambda}[\mathbf{v}] \cdot \delta\lambda = 2 \int_\Omega \delta\lambda(\mathbf{n}_k, \mathbf{v}) \, dV, \tag{2.10}$$

$$\mathcal{L}_{\lambda\mathbf{n}}[\gamma] \cdot \delta\mathbf{n} = 2 \int_\Omega \gamma(\mathbf{n}_k, \delta\mathbf{n}) \, dV, \tag{2.11}$$

and

$$\begin{aligned} \mathcal{L}_{\mathbf{nn}}[\mathbf{v}] \cdot \delta\mathbf{n} = {} & 2(K_1 - K_2 - K_4) \langle \nabla \cdot \delta\mathbf{n}, \nabla \cdot \mathbf{v} \rangle_0 + 2K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \delta\mathbf{n}, \nabla \times \mathbf{v} \rangle_0 \\ & + 2(K_2 - K_3) \Big(\langle \delta\mathbf{n} \cdot \nabla \times \mathbf{v}, \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k \rangle_0 + \langle \mathbf{n}_k \cdot \nabla \times \mathbf{v}, \delta\mathbf{n} \cdot \nabla \times \mathbf{n}_k \rangle_0 \\ & + \langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \delta\mathbf{n} \rangle_0 + \langle \mathbf{n}_k \cdot \nabla \times \delta\mathbf{n}, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \\ & + \langle \delta\mathbf{n} \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \Big) + 2(K_2 + K_4) \Big(\left\langle \nabla \delta n_1, \frac{\partial \mathbf{v}}{\partial x} \right\rangle_0 \\ & + \left\langle \nabla \delta n_2, \frac{\partial \mathbf{v}}{\partial y} \right\rangle_0 + \left\langle \nabla \delta n_3, \frac{\partial \mathbf{v}}{\partial z} \right\rangle_0 \Big) + 2 \int_\Omega \lambda_k(\delta\mathbf{n}, \mathbf{v}) \, dV. \end{aligned} \tag{2.12}$$

Constructing (2.9) using (2.10)–(2.12) yields a linearized variational system. For these iterations, we compute $\delta\mathbf{n}$ and $\delta\lambda$ satisfying this system for all $\mathbf{v} \in \mathcal{H}_0^{DC}(\Omega)$ and $\gamma \in L^2(\Omega)$ with the current approximations $\mathbf{n}_k$ and λ_k. The current approximations are then corrected with the solutions $\delta\mathbf{n}$ and $\delta\lambda$ to yield $\mathbf{n}_{k+1}$ and λ_{k+1}. While they typically improve robustness and efficiency, we do not consider the use of line searches or trust regions [40] in the work presented here, leaving this for future work.

If we are considering a system with Dirichlet or mixed periodic and Dirichlet boundary conditions, as described above, we eliminate the $(K_2 + K_4)$ terms from (2.9), simplifying the linearization.

2.3. Uniform symmetric positive definiteness of Z. In subsequent sections, theory establishing the existence and uniqueness of solutions to the Newton linearizations is developed. A key property exploited in these proofs is that $\mathbf{Z}$ is uniformly symmetric positive definite (USPD) under reasonable assumptions.

It is clear from the definition that $\mathbf{Z}$ is symmetric. Ericksen's inequalities [23] guarantee that $K_2, K_3 \geq 0$. Throughout this paper, we consider the case where the inequality is strict; thus, $\kappa > 0$. We also assume that, in the Newton iterations, control has been maintained over the director length such that

$$\alpha \leq n_1^2 + n_2^2 + n_3^2 \leq \beta \qquad \forall \mathbf{x} \in \Omega \tag{2.13}$$

with constants $0 < \alpha \leq 1 \leq \beta$.

LEMMA 2.1. *Assume that* $\alpha \leq (\mathbf{n}, \mathbf{n}) \leq \beta$ *for all* $\mathbf{x} \in \Omega$. *If* $\kappa \geq 1$, *then* $\mathbf{Z}$ *is USPD on* Ω. *For* $0 < \kappa < 1$, *if* $\beta < \frac{1}{1-\kappa}$, *then* $\mathbf{Z}$ *is USPD on* Ω.

Proof. Rewrite $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{I} - \frac{\mathbf{n} \otimes \mathbf{n}}{\mathbf{n} \cdot \mathbf{n}} + (1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n}))\frac{\mathbf{n} \otimes \mathbf{n}}{\mathbf{n} \cdot \mathbf{n}}.$$

For any $\mathbf{x} \in \Omega$, consider $\xi \in \mathbb{R}^3$. Decompose ξ as $\xi = a_1 \mathbf{v} + a_2 \mathbf{n}$, where $\mathbf{v} \cdot \mathbf{n} = 0$. Then,

$$\frac{\xi^T \mathbf{Z}(\mathbf{x}) \xi}{\xi^T \xi} = \frac{a_1^2 \mathbf{v} \cdot \mathbf{v} + (1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n}))a_2^2 (\mathbf{n} \cdot \mathbf{n})}{a_1^2 \mathbf{v} \cdot \mathbf{v} + a_2^2 (\mathbf{n} \cdot \mathbf{n})}.$$

Thus,

$$\min(1, 1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n})) \leq \frac{\xi^T \mathbf{Z}(\mathbf{x}) \xi}{\xi^T \xi} \leq \max(1, 1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n})).$$

Case 1. $\kappa \geq 1$. Note that

$$1 \leq 1 + (\kappa - 1)\alpha \leq (1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n})) \leq 1 + (\kappa - 1)\beta \qquad \forall \mathbf{x} \in \Omega.$$

Hence,

$$1 \leq \frac{\xi^T \mathbf{Z}(\mathbf{x}) \xi}{\xi^T \xi} \leq 1 + (\kappa - 1)\beta \qquad \forall \mathbf{x} \in \Omega, \xi \in \mathbb{R}^3.$$

Case 2. $0 < \kappa < 1$. For this case,

$$1 + (\kappa - 1)\beta \leq (1 + (\kappa - 1)(\mathbf{n} \cdot \mathbf{n})) \leq 1 + (\kappa - 1)\alpha \leq 1 \qquad \forall \mathbf{x} \in \Omega.$$

Along with the assumption that $\beta < \frac{1}{1-\kappa}$, this implies that

$$0 < 1 + (\kappa - 1)\beta \leq \frac{\xi^T \mathbf{Z}(\mathbf{x}) \xi}{\xi^T \xi} \leq 1 \qquad \forall \mathbf{x} \in \Omega, \xi \in \mathbb{R}^3. \qquad \square$$

Thus, $\mathbf{Z}$ is USPD for any $\kappa > 0$, as long as sufficient control is maintained on the length of $\mathbf{n}$. The USPD property of $\mathbf{Z}$ plays an important role in the proofs of existence and uniqueness of solutions to the linearization undertaken in the next section.

3. Existence and uniqueness for the Newton linearizations. Here and in the following subsections, we will routinely make use of the following set of assumptions.

Assumption 3.1. Consider an open, bounded domain, Ω, with Lipschitz-continuous boundary. Further, assume that $\alpha \leq |\mathbf{n}_k|^2 \leq \beta$ such that $\mathbf{Z}(\mathbf{n}_k(\mathbf{x}))$ remains USPD with lower and upper bounds, η and Λ, respectively. Finally, Dirichlet boundary conditions are applied. Therefore, both $\delta\mathbf{n}$ and $\mathbf{v}$ are in $H_0(\mathrm{div}, \Omega) \cap H_0(\mathrm{curl}, \Omega)$.

In the continuum, the above Newton systems are written in a general form as

$$a(\delta\mathbf{n}, \mathbf{v}) + b(\mathbf{v}, \delta\lambda) = F(\mathbf{v}) \qquad \forall \mathbf{v} \in \mathcal{H}_0^{DC}(\Omega), \tag{3.1}$$

$$b(\delta\mathbf{n}, \gamma) = G(\gamma) \qquad \forall \gamma \in L^2(\Omega), \tag{3.2}$$

where $a(\cdot,\cdot)$ is a symmetric bilinear form, $b(\cdot,\cdot)$ is a bilinear form, and F and G are linear functionals. For simplicity, throughout this section, we drop the notation of $\delta\mathbf{n}$, $\delta\lambda$. Thus,

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) =& K_1 \langle \nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v} \rangle_0 + K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{u}, \nabla \times \mathbf{v} \rangle_0 \\ &+ (K_2 - K_3) \Big(\langle \mathbf{u} \cdot \nabla \times \mathbf{v}, \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k \rangle_0 + \langle \mathbf{n}_k \cdot \nabla \times \mathbf{v}, \mathbf{u} \cdot \nabla \times \mathbf{n}_k \rangle_0 \\ &+ \langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{u} \rangle_0 + \langle \mathbf{n}_k \cdot \nabla \times \mathbf{u}, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \\ &+ \langle \mathbf{u} \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \Big) + \int_\Omega \lambda_k(\mathbf{u}, \mathbf{v}) \, dV \end{aligned} \tag{3.3}$$

and

$$b(\mathbf{v}, \gamma) = \int_\Omega \gamma(\mathbf{n}_k, \mathbf{v}) \, dV.$$

Moreover,

$$\begin{aligned} F(\mathbf{v}) = -\Big(& K_1 \langle \nabla \cdot \mathbf{n}_k, \nabla \cdot \mathbf{v} \rangle_0 + K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{n}_k, \nabla \times \mathbf{v} \rangle_0 \\ &+ (K_2 - K_3) \langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 + \int_\Omega \lambda_k(\mathbf{n}_k, \mathbf{v}) \, dV \Big) \end{aligned}$$

and

$$G(\gamma) = -\frac{1}{2} \int_\Omega \gamma((\mathbf{n}_k, \mathbf{n}_k) - 1) \, dV.$$

In this section, we aim to show that the system in (2.9) is well-posed. Therefore, continuity, coercivity, and weak coercivity results are desired for the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. Due to the complexity of the bilinear forms, deriving theoretical results in the continuum is challenging. However, the following lemmas hold, which we state without proof since they are similar to the proofs for the discrete spaces to follow.

LEMMA 3.2. *Under Assumption* 3.1 *and the assumption that* λ_k *is pointwise nonnegative, if* $\kappa = 1$*, there exists an* $\alpha_0 > 0$ *such that* $\alpha_0 \|\mathbf{v}\|_{DC}^2 \leq a(\mathbf{v}, \mathbf{v})$ *for all* $\mathbf{v} \in \mathcal{H}_0^{DC}(\Omega)$.

If additional regularity is asserted, such that $\delta\mathbf{n}$ and $\mathbf{v}$ are elements of $\mathcal{H}_0^{DC^1}(\Omega) = \{\mathbf{w} \in \mathcal{H}_0^{DC}(\Omega) : \nabla \times \mathbf{w} \in H^1(\Omega)^3\}$ with norm $\|\mathbf{w}\|_{DC^1}^2 = \|\mathbf{w}\|_0^2 + \|\nabla \cdot \mathbf{w}\|_0^2 + \|\nabla \times \mathbf{w}\|_1^2$, where $\|\cdot\|_1$ denotes the standard norm on $H^1(\Omega)$, then the next two lemmas hold for arbitrary κ.

LEMMA 3.3. *Under Assumption* 3.1*, F and G are bounded linear functionals on $\mathcal{H}_0^{DC^1}(\Omega)$ and $L^2(\Omega)$, respectively.*

LEMMA 3.4. *Under Assumption* 3.1*, $a(\mathbf{u},\mathbf{v})$ and $b(\mathbf{v},\gamma)$ are continuous for the norms $\|\cdot\|_{DC^1}$ and $\|\cdot\|_0$.*

For proofs of Lemmas (3.2)–(3.4), see [2]. The auxiliary regularity above poses a number of theoretical problems. For the well-posedness of the continuum system, coercivity and weak coercivity must be shown in the more intricate $\mathcal{H}^{DC^1}(\Omega)$ norm. Moreover, conforming finite elements for this space, such as Bogner–Fox–Schmit elements [11], are undesirably cumbersome and present notable difficulties in demonstrating stability for this linearization system. However, in the discrete setting, results guaranteeing the existence and uniqueness of solutions to the discrete Newton systems at each step are attained under less strict assumptions.

3.1. Discrete system preliminaries. Performing the outlined Newton iterations for free elastic effects necessitates solving the above Newton systems for the update functions $\delta\mathbf{n}$ and $\delta\lambda$. Thus, finite elements are used to numerically approximate the updates. Finite dimensional spaces $V_h \subset \mathcal{H}_0^{DC}(\Omega)$ and $\Pi_h \subset L^2(\Omega)$ are considered, yielding the discrete variational problem

$$\mathbf{a}(\delta\mathbf{n}_h, \mathbf{v}_h) + b(\mathbf{v}_h, \delta\lambda_h) = \mathbf{F}(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h, \tag{3.4}$$

$$b(\delta\mathbf{n}_h, \gamma_h) = G(\gamma_h) \qquad \forall \gamma_h \in \Pi_h. \tag{3.5}$$

Throughout the rest of this section, the developed theory applies exclusively to discrete spaces. Therefore, except when necessary for clarity, we drop the subscript h along with the notation $\delta\mathbf{n}$ and $\delta\lambda$. For instance, we write $a(\mathbf{u},\mathbf{v})$ to indicate the bilinear form in (3.3) operating on the discrete space $V_h \times V_h$.

The existence and uniqueness theory in the following subsections is explicitly developed in the presence of full Dirichlet boundary conditions. However, the theory is equally applicable for a rectangular domain with mixed Dirichlet and periodic boundary conditions. Such a domain is considered in the numerical experiments presented herein.

Let $\{\mathcal{T}_h\}$, $0 < h \leq 1$, be a family of quadrilateral subdivisions of Ω, such that

$$\max\{\text{diam } T : T \in \mathcal{T}_h\} \leq h \text{ diam } \Omega. \tag{3.6}$$

Further, assume that $\{\mathcal{T}_h\}$ is quasi-uniform so that there exists a $\rho > 0$, such that

$$\min\{\text{diam } B_T : T \in \mathcal{T}_h\} \geq \rho\ h \text{ diam } \Omega, \tag{3.7}$$

for all $h \in (0,1]$, where B_T is the largest ball contained in T, such that T is star-shaped with respect to B_T [13]. Denote the measure of $T \in \mathcal{T}_h$ as $|T|$. Furthermore, let Q_p denote piecewise C^0 polynomials of degree $p \geq 1$ on $\mathcal{T}_h$ and P_0 denote the space of piecewise constants on $\mathcal{T}_h$. Next, define a bubble space

$$V_h^b = \{\mathbf{v} \in C_c(\Omega)^3 : \mathbf{v}|_T = a_T b_T \mathbf{n}_k|_T \quad \forall T \in \mathcal{T}_h\},$$

where $C_c(\Omega)$ denotes the space of compactly supported continuous functions on Ω, b_T is the bi- or tri-quadratic bubble function [39], depending on dimension, that vanishes on $\partial T \in \mathcal{T}_h$, and a_T is a constant coefficient associated with b_T. The bubble functions are constructed [42] such that

$$\int_T b_T\, dV = 1 \qquad \forall T \in \mathcal{T}_h, \tag{3.8}$$

$$b_T > 0 \qquad \forall \mathbf{x} \in T. \tag{3.9}$$

Then, we consider the pair of spaces

$$\Pi_h = P_0, \tag{3.10}$$

$$V_h = \{\mathbf{v} \in Q_m \times Q_m \times Q_m \oplus V_h^b : \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega\}. \tag{3.11}$$

In the following sections, to demonstrate the existence and uniqueness of solutions to the system given by (3.4) and (3.5), we show that $a(\mathbf{u}, \mathbf{v})$ is a coercive and continuous bilinear form and that $b(\mathbf{v}, \gamma)$ is a continuous and weakly coercive bilinear form [6,10,12,13] for the above spaces, V_h and Π_h. Throughout the remainder of this section, we further assume that $\mathbf{n}_k \in Q_p$, for some $p \geq 1$, so that $V_h \subset Q_l \times Q_l \times Q_l$ for $l = \max(m, p+2)$.

3.2. Discrete continuity. In this section, we show that the right-hand sides of (3.4) and (3.5) are continuous linear functionals and that the bilinear forms $a(\mathbf{u}, \mathbf{v})$ and $b(\mathbf{v}, \gamma)$ are continuous for the assumptions discussed above.

LEMMA 3.5. *Under Assumption* 3.1, *F and G are bounded linear functionals on V_h and Π_h, respectively.*

Proof. A simple application of the Cauchy–Schwarz inequality shows that $G(\gamma)$ is a bounded linear functional.

For $F(\mathbf{v})$, observe that

$$\begin{aligned} |F(\mathbf{v})| \leq{}& K_1 |\langle \nabla \cdot \mathbf{n}_k, \nabla \cdot \mathbf{v} \rangle_0| + K_3 |\langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{n}_k, \nabla \times \mathbf{v} \rangle_0| \\ &+ |K_2 - K_3| |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0| + \left| \int_\Omega \lambda_k(\mathbf{n}_k, \mathbf{v}) \, dV \right| \end{aligned} \tag{3.12}$$

by the triangle inequality. Applying Cauchy–Schwarz inequalities to (3.12), one obtains

$$\begin{aligned} |F(\mathbf{v})| &\leq K_1 \|\nabla \cdot \mathbf{n}_k\|_0 \|\nabla \cdot \mathbf{v}\|_0 + K_3 \|\mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{n}_k\|_0 \|\nabla \times \mathbf{v}\|_0 \\ &\quad + |K_2 - K_3| |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0| + \|\lambda_k \mathbf{n}_k\|_0 \|\mathbf{v}\|_0 \\ &\leq K_1 \|\nabla \cdot \mathbf{n}_k\|_0 \|\mathbf{v}\|_{DC} + K_3 \|\mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{n}_k\|_0 \|\mathbf{v}\|_{DC} \\ &\quad + |K_2 - K_3| |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0| + \|\lambda_k \mathbf{n}_k\|_0 \|\mathbf{v}\|_{DC}. \end{aligned} \tag{3.13}$$

In order to bound $|F(\mathbf{v})|$, consider the final three summands separately. Note that since $|\mathbf{Z}(\mathbf{n}_k)| \leq \Lambda$, where Λ is the relevant upper bound from Lemma 2.1, it is evident that

$$\|\mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{n}_k\|_0 \leq \Lambda \|\nabla \times \mathbf{n}_k\|_0 \tag{3.14}$$

and that

$$\|\lambda_k \mathbf{n}_k\|_0^2 \leq \beta \int_\Omega \lambda_k^2 \, dV = C_1^2, \tag{3.15}$$

where β is the upper bound in (2.13). Finally, consider

$$|\langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0| = |\langle (\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \nabla \times \mathbf{n}_k, \mathbf{v} \rangle_0|.$$

Applying the Cauchy–Schwarz inequality,

$$\begin{aligned} |\langle (\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \nabla \times \mathbf{n}_k, \mathbf{v} \rangle_0| &\leq \|(\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \nabla \times \mathbf{n}_k\|_0 \|\mathbf{v}\|_0 \\ &\leq \|(\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \nabla \times \mathbf{n}_k\|_0 \|\mathbf{v}\|_{DC}. \end{aligned} \tag{3.16}$$

Next, note that

$$\begin{aligned}(\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)\nabla \times \mathbf{n}_k \cdot (\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)\nabla \times \mathbf{n}_k &= (\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)^2(\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \\ &\leq (|\mathbf{n}_k| \cdot |\nabla \times \mathbf{n}_k|)^2(\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k) \\ &\leq \beta \cdot (\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)^2. \qquad (3.17)\end{aligned}$$

Furthermore, $\nabla \times \mathbf{n}_k$ is a vector of piecewise polynomials. Therefore, $\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k \in L^2(\Omega)$. Employing (3.17) and letting $\|\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k\|_0 = C_2$,

$$\begin{aligned}\|(\mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)\nabla \times \mathbf{n}_k\|_0 &\leq \sqrt{\beta}\left(\int_\Omega (\nabla \times \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k)^2\, dV\right)^{1/2} \\ &\leq \sqrt{\beta}C_2. \qquad (3.18)\end{aligned}$$

Therefore, using (3.13)–(3.16), and (3.18),

$$\begin{aligned}|F(\mathbf{v})| \leq & K_1\|\nabla \cdot \mathbf{n}_k\|_0\|\mathbf{v}\|_{DC} + K_3\Lambda\|\nabla \times \mathbf{n}_k\|_0\|\mathbf{v}\|_{DC} \\ & + |K_2 - K_3|\sqrt{\beta}C_2\|\mathbf{v}\|_{DC} + C_1\|\mathbf{v}\|_{DC},\end{aligned}$$

implying $F(\mathbf{v})$ is a bounded linear functional on V_h. ∎

LEMMA 3.6. *Under Assumption* 3.1, $a(\mathbf{u}, \mathbf{v})$ *and* $b(\mathbf{v}, \gamma)$ *are continuous.*

Proof. First consider

$$\begin{aligned}|b(\mathbf{v}, \gamma)| &= \left|\int_\Omega \gamma(\mathbf{v}, \mathbf{n}_k)\, dV\right| \\ &\leq \|\gamma\|_0\|\mathbf{v} \cdot \mathbf{n}_k\|_0 \\ &\leq \|\gamma\|_0\sqrt{\beta}\|\mathbf{v}\|_0\end{aligned}$$

by Hölder's inequality and (2.13). Therefore, $b(\mathbf{v}, \gamma)$ is a continuous bilinear form.

For the continuity of $a(\mathbf{u}, \mathbf{v})$, observe that

$$\begin{aligned}|a(\mathbf{u}, \mathbf{v})| \leq & K_1|\langle \nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}\rangle_0| + K_3|\,\langle \mathbf{Z}(\mathbf{n}_k)\nabla \times \mathbf{u}, \nabla \times \mathbf{v}\rangle_0\,| \\ & + |K_2 - K_3|\Big(|\langle \mathbf{u} \cdot \nabla \times \mathbf{v}, \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k\rangle_0| + |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{v}, \mathbf{u} \cdot \nabla \times \mathbf{n}_k\rangle_0| \\ & + |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{u}\rangle_0| + |\langle \mathbf{n}_k \cdot \nabla \times \mathbf{u}, \mathbf{v} \cdot \nabla \times \mathbf{n}_k\rangle_0| \\ & + |\langle \mathbf{u} \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k\rangle_0|\Big) + \left|\int_\Omega \lambda_k(\mathbf{u}, \mathbf{v})\, dV\right| \qquad (3.19)\end{aligned}$$

by the triangle inequality. For simplicity, consider the components of the sum above. Note that

$$|\langle \nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v}\rangle_0| \leq \|\nabla \cdot \mathbf{u}\|_0\|\nabla \cdot \mathbf{v}\|_0 \leq \|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \qquad (3.20)$$

Considering $|\,\langle \mathbf{Z}(\mathbf{n}_k)\nabla \times \mathbf{u}, \nabla \times \mathbf{v}\rangle_0\,|$, using (3.14) implies that

$$\begin{aligned}|\,\langle \mathbf{Z}(\mathbf{n}_k)\nabla \times \mathbf{u}, \nabla \times \mathbf{v}\rangle_0\,| &\leq \|\nabla \times \mathbf{v}\|_0\|\mathbf{Z}(\mathbf{n}_k)\nabla \times \mathbf{u}\|_0 \\ &\leq \Lambda\|\mathbf{v}\|_{DC}\|\nabla \times \mathbf{u}\|_0 \\ &\leq \Lambda\|\mathbf{v}\|_{DC}\|\mathbf{u}\|_{DC}. \qquad (3.21)\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} |\langle \mathbf{u}\cdot\nabla\times\mathbf{v}, \mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0| &= |\langle (\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k)\mathbf{u}, \nabla\times\mathbf{v}\rangle_0| \\ &\le \|(\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k)\mathbf{u}\|_0\|\nabla\times\mathbf{v}\|_0. \end{aligned} \tag{3.22}$$

Note that

$$(\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k)^2 \le |\mathbf{n}_k|^2|\nabla\times\mathbf{n}_k|^2 \le \beta|\nabla\times\mathbf{n}_k|^2.$$

Furthermore, since $\nabla\times\mathbf{n}_k$ is a vector of piecewise polynomials, $|\nabla\times\mathbf{n}_k|^2$ is bounded. Letting $C_{\sup} = \sup_{\mathbf{x}\in\Omega}|\nabla\times\mathbf{n}_k|^2$,

$$\begin{aligned} \|(\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k)\mathbf{u}\|_0 &= \left(\int_\Omega (\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k)^2(\mathbf{u}\cdot\mathbf{u})\,dV\right)^{1/2} \\ &\le \sqrt{\beta}\left(\int_\Omega |\nabla\times\mathbf{n}_k|^2(\mathbf{u}\cdot\mathbf{u})\,dV\right)^{1/2} \\ &\le \sqrt{\beta C_{\sup}}\|\mathbf{u}\|_0. \end{aligned}$$

Hence,

$$|\langle \mathbf{u}\cdot\nabla\times\mathbf{v}, \mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0| \le \sqrt{\beta C_{\sup}}\|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \tag{3.23}$$

The next summand from (3.19) is

$$|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{v}, \mathbf{u}\cdot\nabla\times\mathbf{n}_k\rangle_0| \le \|\mathbf{n}_k\cdot\nabla\times\mathbf{v}\|_0\|\mathbf{u}\cdot\nabla\times\mathbf{n}_k\|_0$$

with

$$\|\mathbf{n}_k\cdot\nabla\times\mathbf{v}\|_0 \le \sqrt{\beta}\|\mathbf{v}\|_{DC}.$$

Furthermore,

$$\|\mathbf{u}\cdot\nabla\times\mathbf{n}_k\|_0 \le \sqrt{C_{\sup}}\|\mathbf{u}\|_0.$$

Therefore,

$$|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{v}, \mathbf{u}\cdot\nabla\times\mathbf{n}_k\rangle_0| \le \sqrt{\beta C_{\sup}}\|\mathbf{v}\|_{DC}\|\mathbf{u}\|_{DC}. \tag{3.24}$$

Now consider $|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{n}_k, \mathbf{v}\cdot\nabla\times\mathbf{u}\rangle_0|$ and note that this inner product is the same as that in (3.22) with the roles of $\mathbf{u}$ and $\mathbf{v}$ reversed. Since $\mathbf{u}$ and $\mathbf{v}$ are from the same space, the steps for deriving (3.23) are equally valid. Thus,

$$|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{n}_k, \mathbf{v}\cdot\nabla\times\mathbf{u}\rangle_0| \le \sqrt{\beta C_{\sup}}\|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \tag{3.25}$$

Similarly, the inequality for $|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{u}, \mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0|$ is derived in an analogous manner to that of (3.24). Thus,

$$|\langle \mathbf{n}_k\cdot\nabla\times\mathbf{u}, \mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0| \le \sqrt{\beta C_{\sup}}\|\mathbf{v}\|_{DC}\|\mathbf{u}\|_{DC}. \tag{3.26}$$

Next, examine

$$|\langle \mathbf{u}\cdot\nabla\times\mathbf{n}_k, \mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0| \le \|\mathbf{u}\cdot\nabla\times\mathbf{n}_k\|_0\|\mathbf{v}\cdot\nabla\times\mathbf{n}_k\|_0.$$

Since $\nabla \times \mathbf{n}_k$ is a vector of piecewise polynomials,

$$\|\mathbf{u} \cdot \nabla \times \mathbf{n}_k\|_0 \leq \sqrt{C_{\sup}}\|\mathbf{u}\|_0,$$
$$\|\mathbf{v} \cdot \nabla \times \mathbf{n}_k\|_0 \leq \sqrt{C_{\sup}}\|\mathbf{v}\|_0.$$

Thus,

$$|\langle \mathbf{u} \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0| \leq C_{\sup}\|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \tag{3.27}$$

Finally, since λ_k is piecewise constant, λ_k^2 is bounded. Letting $C_\lambda = \sup_{\mathbf{x}\in\Omega} \lambda_k^2$,

$$\begin{aligned}\left|\int_\Omega \lambda_k(\mathbf{u}, \mathbf{v})\, dV\right| &\leq \|\lambda_k \mathbf{u}\|_0 \|\mathbf{v}\|_0 \\ &\leq \sqrt{C_\lambda}\|\mathbf{u}\|_0\|\mathbf{v}\|_{DC} \\ &\leq \sqrt{C_\lambda}\|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \end{aligned} \tag{3.28}$$

Combining (3.20), (3.21), and (3.23)–(3.28),

$$a(\mathbf{u}, \mathbf{v}) \leq \left(K_1 + K_3\Lambda + |K_2 - K_3|\left(4\sqrt{\beta C_{\sup}} + C_{\sup}\right) + \sqrt{C_\lambda}\right)\|\mathbf{u}\|_{DC}\|\mathbf{v}\|_{DC}. \qquad \square$$

3.3. Discrete coercivity. In this section, two proofs of the coercivity of $a(\mathbf{u}, \mathbf{v})$ are given. The first is for the case when $\kappa = 1$. The second addresses coercivity when κ lies in a neighborhood of unity. For both proofs, we use the additional assumption that the approximation is close enough to the solution such that the Lagrange multiplier, λ_k, is pointwise nonnegative. This assumption is reasonable since at the solution, $\mathbf{n}_*$, λ_* may be chosen arbitrarily.

LEMMA 3.7. *Under Assumption* 3.1 *and the assumption that* λ_k *is pointwise nonnegative, if* $\kappa = 1$*, there exists an* $\alpha_0 > 0$ *such that* $\alpha_0\|\mathbf{v}\|_{DC}^2 \leq a(\mathbf{v}, \mathbf{v})$ *for all* $\mathbf{v} \in V_h$*.*

Proof. Note that since $\kappa = 1$, $(K_2 - K_3) = 0$, and

$$a(\mathbf{v}, \mathbf{v}) = K_1\langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v}\rangle_0 + K_3 \langle \nabla \times \mathbf{v}, \nabla \times \mathbf{v}\rangle_0 + \int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV.$$

Thus, it remains to show that there exists $\alpha_0 > 0$ such that

$$\alpha_0\|\mathbf{v}\|_{DC}^2 \leq K_1\langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v}\rangle_0 + K_3 \langle \nabla \times \mathbf{v}, \nabla \times \mathbf{v}\rangle_0 + \int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV.$$

From Remark 2.7 in [28], there exists $C_3 > 0$ such that

$$\|\nabla \mathbf{v}\|_0^2 \leq C_3^2\left(\|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2\right).$$

Moreover, recall that $\|\mathbf{v}\|_0^2 \leq C_4\|\nabla \mathbf{v}\|_0^2$ by the classical Poincaré–Friedrichs inequality. Hence, for $C = C_4C_3^2 > 0$,

$$\|\mathbf{v}\|_0^2 \leq C\left(\|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2\right). \tag{3.29}$$

Since $\|\mathbf{v}\|_{DC}^2 = \|\mathbf{v}\|_0^2 + \|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2$, then

$$\|\mathbf{v}\|_{DC}^2 \leq (C + 1)\left(\|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2\right).$$

Letting $K = \min(K_1, K_3) > 0$ and $\alpha_0 = K/(C+1)$, it follows that

$$\alpha_0 \|\mathbf{v}\|_{DC}^2 \leq K\big(\|\nabla \cdot \mathbf{v}\|_0^2 + \|\nabla \times \mathbf{v}\|_0^2\big) \leq K_1 \|\nabla \cdot \mathbf{v}\|_0^2 + K_3 \|\nabla \times \mathbf{v}\|_0^2. \tag{3.30}$$

Finally, it was assumed that λ_k is pointwise nonnegative, implying

$$\int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV \geq 0.$$

Therefore, (3.30) implies that

$$\alpha_0 \|\mathbf{v}\|_{DC}^2 \leq K_1 \langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v} \rangle_0 + K_3 \langle \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0 + \int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV. \qquad \square$$

The assumption that $\kappa = 1$ is a common modeling approach. In fact, this supposition represents a weaker constraint than is seen in the many models that utilize the one-constant approximation; cf. [16, 43, 46, 48]. However, it is possible to loosen the restriction that $\kappa = 1$ and still maintain the coercivity of $a(\mathbf{u}, \mathbf{v})$ with a small data type assumption on κ. That is, we assume that κ varies within a certain, possibly small, range of unity. Small data assumptions are common, for instance, in the study of solutions to the Navier–Stokes equations [27, 38], where bounds are imposed on certain norms of the initial data in order to demonstrate existence and uniqueness of solutions.

LEMMA 3.8 (small data). *Under Assumption* 3.1 *and the assumption that λ_k is pointwise nonnegative, there exists $\epsilon_1, \epsilon_2 > 0$, dependent on $\beta = \max |\mathbf{n}_k|^2$, such that if $\kappa \in (1 - \epsilon_2, 1 + \epsilon_1)$, then $a(\mathbf{u}, \mathbf{v})$ is coercive.*

Proof. Since $\mathbf{Z}(\mathbf{n}_k)$ is USPD by assumption,

$$\eta K_3 \langle \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0 \leq K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0,$$

where η is the relevant lower bound from Lemma 2.1. Defining $K' = \min(K_1, \eta K_3) > 0$ and $\alpha_1 = K'/(C+1)$, where $C = C_4 C_3^2$ is the constant defined in (3.29), then

$$\alpha_1 \|\mathbf{v}\|_{DC}^2 \leq K_1 \langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v} \rangle_0 + \eta K_3 \langle \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0.$$

Thus, using the assumption that λ_k is pointwise nonnegative,

$$\alpha_1 \|\mathbf{v}\|_{DC}^2 \leq K_1 \langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v} \rangle_0 + K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0 + \int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV. \tag{3.31}$$

It should be noted that the constant η may depend on κ. Thus, the following three cases are considered.

Case 1. $\kappa = 1 + \epsilon_1$ for $\epsilon_1 > 0$. If this case holds, then $\eta = 1$. Hence, α_1, defined for (3.31), is independent of κ. Since $K_2 - K_3 = K_3(\kappa - 1)$, the discrete bilinear form of (3.3) becomes

$$\begin{aligned} a(\mathbf{v}, \mathbf{v}) =& K_1 \langle \nabla \cdot \mathbf{v}, \nabla \cdot \mathbf{v} \rangle_0 + K_3 \langle \mathbf{Z}(\mathbf{n}_k) \nabla \times \mathbf{v}, \nabla \times \mathbf{v} \rangle_0 \\ &+ \epsilon_1 K_3 \Big(2 \langle \mathbf{v} \cdot \nabla \times \mathbf{v}, \mathbf{n}_k \cdot \nabla \times \mathbf{n}_k \rangle_0 + 2 \langle \mathbf{n}_k \cdot \nabla \times \mathbf{v}, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \\ &+ \langle \mathbf{v} \cdot \nabla \times \mathbf{n}_k, \mathbf{v} \cdot \nabla \times \mathbf{n}_k \rangle_0 \Big) + \int_\Omega \lambda_k(\mathbf{v}, \mathbf{v})\, dV. \end{aligned} \tag{3.32}$$

Observe that from (3.31),

$$\begin{aligned}\alpha_1\|\mathbf{v}\|_{DC}^2 \leq & K_1\langle\nabla\cdot\mathbf{v},\nabla\cdot\mathbf{v}\rangle_0 + K_3\left\langle\mathbf{Z}(\mathbf{n}_k)\nabla\times\mathbf{v},\nabla\times\mathbf{v}\right\rangle_0 + \int_\Omega \lambda_k(\mathbf{v},\mathbf{v})\,dV \\ & + \epsilon_1 K_3\langle\mathbf{v}\cdot\nabla\times\mathbf{n}_k,\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0. \end{aligned} \tag{3.33}$$

Consider the magnitude of the terms in (3.32) not bounded from below in (3.33), denoted as $\mathcal{G}(\mathbf{v},\mathbf{v})$,

$$\begin{aligned}|\mathcal{G}(\mathbf{v},\mathbf{v})| &= |2\epsilon_1 K_3\big(\langle\mathbf{v}\cdot\nabla\times\mathbf{v},\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0 + \langle\mathbf{n}_k\cdot\nabla\times\mathbf{v},\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0\big)| \\ &\leq 2\epsilon_1 K_3\big(|\langle\mathbf{v}\cdot\nabla\times\mathbf{v},\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0| + \|\mathbf{n}_k\cdot\nabla\times\mathbf{v}\|_0\|\mathbf{v}\cdot\nabla\times\mathbf{n}_k\|_0\big).\end{aligned}$$

Using bounds derived in the proof of Lemma 3.6,

$$|\mathcal{G}(\mathbf{v},\mathbf{v})| \leq 4\epsilon_1 K_3\sqrt{\beta C_{\sup}}\|\mathbf{v}\|_{DC}^2.$$

Denoting $\alpha_3 = 4K_3\sqrt{\beta C_{\sup}}$, then

$$|\mathcal{G}(\mathbf{v},\mathbf{v})| \leq \epsilon_1\alpha_3\|\mathbf{v}\|_{DC}^2.$$

Utilizing (3.33),

$$a(\mathbf{v},\mathbf{v}) \geq \alpha_1\|\mathbf{v}\|_{DC}^2 - \epsilon_1\alpha_3\|\mathbf{v}\|_{DC}^2 = (\alpha_1 - \epsilon_1\alpha_3)\|\mathbf{v}\|_{DC}^2.$$

It is, thus, sufficient to have $\epsilon_1 < \alpha_1/\alpha_3$, guaranteeing that $(\alpha_1 - \epsilon_1\alpha_3) > 0$.

Case 2. $\kappa = 1-\epsilon_2 > 0$, for $\epsilon_2 > 0$, and $K_1 < K_3$. Since $\kappa < 1$, $\eta = 1+(\kappa-1)\beta = (1-\epsilon_2\beta)$. For $K_1 < K_3$, there exists an ϵ_2 small enough such that $K_1 < (1-\epsilon_2\beta)K_3$. This implies that, for small enough ϵ_2,

$$\alpha_1 = \frac{\min(K_1,(1-\epsilon_2\beta)K_3)}{(C+1)} = \frac{K_1}{(C+1)}.$$

Therefore, α_1 is again independent of κ. Since $K_2 - K_3 = K_3(\kappa-1)$, the discrete bilinear form of (3.3) becomes

$$\begin{aligned}a(\mathbf{v},\mathbf{v}) =& K_1\langle\nabla\cdot\mathbf{v},\nabla\cdot\mathbf{v}\rangle_0 + K_3\left\langle\mathbf{Z}(\mathbf{n}_k)\nabla\times\mathbf{v},\nabla\times\mathbf{v}\right\rangle_0 \\ & - \epsilon_2 K_3\Big(2\langle\mathbf{v}\cdot\nabla\times\mathbf{v},\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0 + 2\langle\mathbf{n}_k\cdot\nabla\times\mathbf{v},\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0 \\ & + \langle\mathbf{v}\cdot\nabla\times\mathbf{n}_k,\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0\Big) + \int_\Omega\lambda_k(\mathbf{v},\mathbf{v})\,dV.\end{aligned} \tag{3.34}$$

The terms of (3.34), not already bounded from below in (3.31), are bounded as

$$\begin{aligned}|\mathcal{G}(\mathbf{v},\mathbf{v})| &= |\epsilon_2 K_3\big(2\langle\mathbf{v}\cdot\nabla\times\mathbf{v},\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0 \\ &\quad + 2\langle\mathbf{n}_k\cdot\nabla\times\mathbf{v},\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0 + \langle\mathbf{v}\cdot\nabla\times\mathbf{n}_k,\mathbf{v}\cdot\nabla\times\mathbf{n}_k\rangle_0\big)| \\ &\leq \epsilon_2 K_3\big(2|\langle\mathbf{v}\cdot\nabla\times\mathbf{v},\mathbf{n}_k\cdot\nabla\times\mathbf{n}_k\rangle_0| \\ &\quad + 2\|\mathbf{n}_k\cdot\nabla\times\mathbf{v}\|_0\|\mathbf{v}\cdot\nabla\times\mathbf{n}_k\|_0 + \|\mathbf{v}\cdot\nabla\times\mathbf{n}_k\|_0\|\mathbf{v}\cdot\nabla\times\mathbf{n}_k\|_0\big).\end{aligned}$$

Again using the bounds derived in the proof of Lemma 3.6,

$$|\mathcal{G}(\mathbf{v},\mathbf{v})| \leq \epsilon_2 K_3\big(4\sqrt{\beta C_{\sup}} + C_{\sup}\big)\|\mathbf{v}\|_{DC}^2.$$

Denoting $\alpha_4 = K_3\big(4\sqrt{\beta C_{\sup}} + C_{\sup}\big)$, then

$$|\mathcal{G}(\mathbf{v},\mathbf{v})| \le \epsilon_2\alpha_4\|\mathbf{v}\|_{DC}^2.$$

Using (3.31) implies

$$a(\mathbf{v},\mathbf{v}) \ge \alpha_1\|\mathbf{v}\|_{DC}^2 - \epsilon_2\alpha_4\|\mathbf{v}\|_{DC}^2 = (\alpha_1 - \epsilon_2\alpha_4)\|\mathbf{v}\|_{DC}^2.$$

Thus, possibly requiring ϵ_2 to be even smaller, $\epsilon_2 < \alpha_1/\alpha_4$, so that $(\alpha_1 - \epsilon_2\alpha_4) > 0$.

In the case that $\kappa < 1$, the additional restriction that $\beta < \frac{1}{1-\kappa}$ for $\mathbf{Z}$ to be USPD is necessary, which implies that $\epsilon_2\beta < 1$ is required. Therefore, for any fixed choice of β, ϵ_2 must also be taken small enough to satisfy this condition. Hence,

$$\epsilon_2 < \min\left(\frac{\alpha_1}{\alpha_4}, \frac{K_3 - K_1}{\beta K_3}, \frac{1}{\beta}\right).$$

Case 3. $\kappa = 1 - \epsilon_2 > 0$, for $\epsilon_2 > 0$, and $K_3 \le K_1$. Here, again, $\eta = (1 - \epsilon_2\beta)$. For this case, it is clear that $(1 - \epsilon_2\beta)K_3 < K_1$. Thus,

$$\alpha_1 = \frac{(1 - \epsilon_2\beta)K_3}{(C+1)}.$$

Using the same α_4 as in the previous case and similar arguments,

$$a(\mathbf{v},\mathbf{v}) \ge \alpha_1\|\mathbf{v}\|_{DC}^2 - \epsilon_2\alpha_4\|\mathbf{v}\|_{DC}^2 = (\alpha_1 - \epsilon_2\alpha_4)\|\mathbf{v}\|_{DC}^2.$$

Hence, in order for $(\alpha_1 - \epsilon_2\alpha_4) > 0$ to hold, it is necessary that

$$\epsilon_2 < \frac{K_3}{K_3\beta + \alpha_4(C+1)}.$$

Finally, ϵ_2 must still be chosen sufficiently small with respect to β such that $\epsilon_2\beta < 1$, as in Case 2. Therefore,

$$\epsilon_2 < \min\left(\frac{K_3}{K_3\beta + \alpha_4(C+1)}, \frac{1}{\beta}\right).$$

Thus, if ϵ_1, $\epsilon_2 > 0$ satisfy the applicable conditions in the cases above, then at each Newton iteration, $a(\mathbf{u},\mathbf{v})$ is coercive for $\kappa \in (1 - \epsilon_2, 1 + \epsilon_1)$. ∎

Remark 3.9. The size of the interval about $\kappa = 1$ depends, in part, on $C_{\sup}$. At each iteration, $C_{\sup}$ is a well-defined constant for the variational problem given in (3.4)–(3.5). However, no uniformity is guaranteed without stronger assumptions; notably, if the iterates, $\mathbf{n}_k$, approach a function with singular curl, then $C_{\sup}$ may grow either as the iteration proceeds or with grid refinement. To achieve uniformity, an assumption that the true solution, $\mathbf{n}_*$, is suitably smooth is needed. By the Sobolev embedding theorem [1], if Ω is a bounded Lipschitz domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ and $\mathbf{n}_* \in H^3(\Omega)^3$, then $\nabla \times \mathbf{n}_*$ is continuous and bounded. However, further assumptions would be necessary to ensure that the Newton iterations remain in a neighborhood of $\mathbf{n}_*$ where their curls could be uniformly bounded.

3.4. Discrete weak coercivity. For this section, we consider the weak coercivity of $b(\cdot,\cdot)$, under Assumption 3.1, with the restriction that Ω is a polyhedral domain. That is, we show that there exists a $\zeta > 0$ such that

$$\zeta\|\gamma\|_0 \leq \sup_{\mathbf{v}\in V_h} \frac{|b(\mathbf{v},\gamma)|}{\|\mathbf{v}\|_{DC}} \qquad \forall \gamma \in \Pi_h. \tag{3.35}$$

Before proving the weak coercivity result for V_h and Π_h, we prove two critical lemmas. Let $N = 2, 3$ denote the dimension of Ω.

LEMMA 3.10. *For the bubble functions, b_T, satisfying* (3.8) *and* (3.9) *on a rectangle T, $\sup_{\mathbf{x}\in T} b_T = C_d/|T|$, where $C_d = (\frac{3}{2})^N$.*

Proof. For $N = 2$, without loss of generality, assume that T is a rectangle at the origin given by $[0,a]\times[0,b]$. Let $\bar{b}_T = xy(a-x)(b-y)$ on T and zero elsewhere. Note that $\bar{b}_T$ is the bubble function on T that has not been normalized such that (3.8) holds. Integrating over T yields

$$\int_T \bar{b}_T \, dV = \frac{|T|^3}{36}. \tag{3.36}$$

Computing the maximum value of $\bar{b}_T$ shows that $\sup_{\mathbf{x}\in T} \bar{b}_T = \frac{|T|^2}{16}$. Normalizing $\bar{b}_T$, using (3.36), to define b_T implies that

$$\sup_{\mathbf{x}\in T} b_T = \frac{|T|^2/16}{|T|^3/36} = \frac{9}{4|T|}.$$

The case for $N = 3$ is derived analogously for T, the rectangular box $[0,a]\times[0,b]\times[0,c]$, and $\bar{b}_T = xyz(a-x)(b-y)(c-z)$. The corresponding b_T satisfies

$$\sup_{\mathbf{x}\in T} b_T = \frac{|T|^2/64}{|T|^3/216} = \frac{27}{8|T|}. \qquad \square$$

Following the notation in [13], consider two finite elements $(T,\mathcal{P},\mathcal{N})$ and $(\hat{T},\hat{\mathcal{P}},\hat{\mathcal{N}})$, where T and $\hat{T}$ are element domains, $\mathcal{P}$ and $\hat{\mathcal{P}}$ are the respective sets of basis functions, and $\mathcal{N}$ and $\hat{\mathcal{N}}$ are the associated dual bases. We say that $(\hat{T},\hat{\mathcal{P}},\hat{\mathcal{N}})$ is affine equivalent to $(T,\mathcal{P},\mathcal{N})$ if there exists an affine mapping, $G : T \to \hat{T}$, such that for $\mathbf{x} \in T$

$$G\mathbf{x} = \mathbf{x}_0 + M\mathbf{x},$$

with nonsingular matrix M, satisfying

- $G(T) = \hat{T}$,
- $G^*\hat{\mathcal{P}} = \mathcal{P}$, and
- $G_*\mathcal{N} = \hat{\mathcal{N}}$.

Here, the pullback G^* is defined by $G^*(\hat{f}) := \hat{f}\circ G$, and the push-forward G_* is defined by $(G_*N)(\hat{f}) := N(G^*(\hat{f}))$.

LEMMA 3.11. *Consider a rectangular reference element $(T,\mathcal{P},\mathcal{N})$, where $\mathcal{P}$ is the basis of shape functions for T associated with $V_h \times \Pi_h$, defined above. If for all $\hat{T} \in \mathcal{T}_h$, $(\hat{T},\hat{\mathcal{P}},\hat{\mathcal{N}})$ is affine equivalent to $(T,\mathcal{P},\mathcal{N})$, then $\sup_{\hat{\mathbf{x}}\in\hat{T}} b_{\hat{T}} = C_d/|\hat{T}|$, where $b_{\hat{T}}$ is the normalized bubble function satisfying* (3.8) *and* (3.9) *on $\hat{T}$.*

Proof. Note that the nonnormalized bubble function on $\hat{T}$, $\bar{b}_{\hat{T}}$, is given by

$$\bar{b}_{\hat{T}} = b_T \circ G^{-1},$$

where b_T is the normalized bubble function on T. Therefore, the maximum value for $\bar{b}_{\hat{T}}$ corresponds to the maximum value for b_T, which, as shown in Lemma 3.10, is $C_d/|T|$. Observe that

$$\begin{aligned}\int_{\hat{T}} \bar{b}_{\hat{T}}\, dV &= \int_T b_T |\det M|\, dV \\ &= |\det M|,\end{aligned}$$

where $\det M$ denotes the determinant of the matrix M. Thus, $b_{\hat{T}}$ is given by dividing $\bar{b}_{\hat{T}}$ by $|\det M|$. Therefore,

$$\begin{aligned}\sup_{\hat{\mathbf{x}}\in\hat{T}} b_{\hat{T}} &= \frac{1}{|\det M|} \sup_{\mathbf{x}\in T} b_T \\ &= \frac{C_d}{|\det M||T|} \\ &= \frac{C_d}{|\hat{T}|}. \qquad \square\end{aligned}$$

In the following, we will make use of the second set of assumptions below when necessary.

Assumption 3.12. Let $\{\mathcal{T}_h\}$ be a family of quadrilateral subdivisions of a polyhedral domain Ω satisfying (3.6) and (3.7). Moreover, assume that for each $T \in \mathcal{T}_h$, the element $(T, \mathcal{P}_T, \mathcal{N}_T)$ is affine equivalent to a rectangular reference element for all h.

Prior to considering the following lemma, recall that α and β are the bounds on the length of $\mathbf{n}$ in (2.13), ρ is the quasi-uniform mesh parameter defined in (3.7), and C_d is the constant derived in Lemma 3.10 depending on N, the dimension of Ω.

LEMMA 3.13. *Under Assumptions* 3.1 *and* 3.12, V_h *and* Π_h *constitute a pair satisfying* (3.35) *with constant* $\zeta = h[\frac{2\alpha\rho^N}{9C_f C_* \sqrt{\beta C_d}}]$ *for* C_f *and* C_* *defined below.*

Proof. Since $V_h \subset Q_l \times Q_l \times Q_l$, by [13, Theorem 4.5.11] there exists $C_* > 0$ depending only on ρ such that

$$\|\mathbf{v}\|_1 \le C_* h^{-1} \|\mathbf{v}\|_0.$$

Furthermore, using the fact that $\|\mathbf{v}\|_{DC} \le C_f \|\mathbf{v}\|_1$,

$$\sup_{\mathbf{v}\in V_h} \frac{|b(\mathbf{v},\gamma)|}{\|\mathbf{v}\|_{DC}} \ge \sup_{\mathbf{v}\in V_h} \frac{|b(\mathbf{v},\gamma)|}{C_f\|\mathbf{v}\|_1} \ge \sup_{\mathbf{v}\in V_h} \frac{|b(\mathbf{v},\gamma)|}{C_f C_* h^{-1}\|\mathbf{v}\|_0}. \tag{3.37}$$

Therefore, (3.35) is reduced to finding $\zeta > 0$ such that

$$\zeta\|\gamma\|_0 \le \sup_{\mathbf{v}\in V_h} \frac{|b(\mathbf{v},\gamma)|}{C_f C_* h^{-1}\|\mathbf{v}\|_0} \qquad \forall \gamma \in \Pi_h.$$

Now consider constructing $\mathbf{v}_0$ on each $T \in \mathcal{T}_h$ by letting $a_T = \gamma|_T$, where this denotes the restriction of γ to the element T, and defining

$$\mathbf{v}_0|_T = a_T b_T \mathbf{n}_k|_T.$$

Observe that, as defined, $\mathbf{v}_0 \in V_h$. Let $C_m = \max_{T\in\mathcal{T}_h} |T|$. Then,

$$\begin{aligned} b(\mathbf{v}_0,\gamma) = \sum_{T\in\mathcal{T}_h} \int_T \gamma(\mathbf{v}_0,\mathbf{n}_k) &\ge \alpha \sum_{T\in\mathcal{T}_h} \gamma^2 \int_T b_T\, dV \\ &= \alpha \sum_{T\in\mathcal{T}_h} \gamma^2 \ge \frac{\alpha}{C_m}\|\gamma\|_0^2. \end{aligned} \tag{3.38}$$

It is also the case that

$$\|\mathbf{v}_0\|_0^2 = \sum_{T\in\mathcal{T}_h} \int_T a_T^2 b_T^2(\mathbf{n}_k, \mathbf{n}_k)\, dV \le \beta \sum_{T\in\mathcal{T}_h} \gamma^2 \int_T b_T^2\, dV.$$

Since the bubble functions are fixed, let

$$C_b = \max_{T\in\mathcal{T}_h} \int_T b_T^2\, dV, \qquad C_T = \min_{T\in\mathcal{T}_h} |T|.$$

Thus,

$$\|\mathbf{v}_0\|_0^2 \le \beta C_b \sum_{T\in\mathcal{T}_h} \gamma^2 \le \frac{\beta C_b}{C_T}\|\gamma\|_0^2. \tag{3.39}$$

Therefore, combining (3.38) and (3.39),

$$\begin{aligned}\sup_{\mathbf{v}\in V_h} \frac{\int_\Omega \gamma(\mathbf{v}, \mathbf{n}_k)\, dV}{\|\mathbf{v}\|_0} &\ge \frac{\int_\Omega \gamma(\mathbf{v}_0, \mathbf{n}_k)\, dV}{\|\mathbf{v}_0\|_0} \\ &\ge \frac{\frac{\alpha}{C_m}\|\gamma\|_0^2}{\sqrt{\frac{\beta C_b}{C_T}}\|\gamma\|_0} = \frac{\alpha\sqrt{C_T}}{C_m\sqrt{\beta C_b}}\|\gamma\|_0. \end{aligned} \tag{3.40}$$

Note that the final constant in (3.40) is mesh dependent. Let $N = 2, 3$ denote the dimension of Ω. Observe that

$$C_b \le \max_{T\in\mathcal{T}_h} \sup_{\mathbf{x}\in T} b_T \int_T b_T\, dV = \max_{T\in\mathcal{T}_h} \sup_{\mathbf{x}\in T} b_T.$$

From Lemma 3.11, for arbitrary $T \in \mathcal{T}_h$,

$$\sup_{\mathbf{x}\in T} b_T = C_d/|T|,$$

where C_d depends only on the dimension of Ω. Therefore,

$$\max_{T\in\mathcal{T}_h} \sup_{\mathbf{x}\in T} b_T = \frac{C_d}{C_T}.$$

Hence,

$$\frac{\sqrt{C_T}}{C_m\sqrt{C_b}} \ge \frac{C_T}{C_m\sqrt{C_d}}. \tag{3.41}$$

Define the constants

$$\begin{aligned} C_{2,1} &= \frac{\pi}{4}, & C_{2,2} &= \pi & &\text{for } N = 2, \\ C_{3,1} &= \frac{\pi}{6}, & C_{3,2} &= \frac{3\pi}{4} & &\text{for } N = 3. \end{aligned}$$

Using Properties (3.6) and (3.7) with the constants above, it is straightforward to show that

$$\begin{aligned} C_T &\ge C_{N,1}(\min\{\text{diam } B_T : T \in \mathcal{T}_h\})^N \ge C_{N,1}\rho^N(h\text{diam }\Omega)^N, \\ C_m &\le C_{N,2}(\max\{\text{diam } T : T \in \mathcal{T}_h\})^N \le C_{N,2}(h\text{diam }\Omega)^N. \end{aligned}$$

Therefore,

$$\frac{C_T}{C_m} \geq \frac{C_{N,1}\rho^N}{C_{N,2}}. \tag{3.42}$$

Utilizing (3.41) and (3.42)

$$\frac{\alpha\sqrt{C_T}}{C_m\sqrt{\beta C_b}}\|\gamma\|_0 \geq \frac{\alpha C_{N,1}\rho^N}{C_{N,2}\sqrt{\beta C_d}}\|\gamma\|_0 \geq \frac{2\alpha\rho^N}{9\sqrt{\beta C_d}}\|\gamma\|_0,$$

where C_d depends only on the dimension of Ω. Hence, (3.35) is satisfied with constant $\zeta = h[\frac{2\alpha\rho^N}{9C_f C_*\sqrt{\beta C_d}}]$. Thus, V_h and Π_h represent a pair of spaces on which $b(\cdot,\cdot)$ is weakly coercive. ∎

For $\mathbf{n}_k \in Q_p$, with $V_h \subset Q_m \times Q_m \times Q_m \oplus V_h^b$, as in (3.11), and $l = \max(m, p+2)$, the above lemma yields an immediate corollary.

COROLLARY 3.14. *Under Assumptions* 3.1 *and* 3.12, $\mathbf{n}_k \in Q_p$ *implies that* $b(\cdot,\cdot)$ *is weakly coercive for the pair* Q_l–P_0. *The special case that* $\mathbf{n}_k \in P_0$ *implies that* $b(\cdot,\cdot)$ *is weakly coercive on the pair* $Q_{\max(m,2)}$–P_0.

Proof. Note that if $\mathbf{n}_k \in Q_p$, the bubble space defined above satisfies $V_h^b \subset Q_{p+2} \times Q_{p+2} \times Q_{p+2}$, since $b_T \in Q_2$. This implies that $V_h \subset Q_l \times Q_l \times Q_l$. Therefore, since $b(\cdot,\cdot)$ is weakly coercive for the pair V_h–P_0, weak coercivity must also hold for the pair Q_l–P_0. If $\mathbf{n}_k \in P_0$, then $V_h^b \subset Q_2 \times Q_2 \times Q_2$. Hence, $V_h \subset Q_{\max(m,2)} \times Q_{\max(m,2)} \times Q_{\max(m,2)}$. The lemma above is equally valid for $\mathbf{n}_k \in P_0$. Therefore, $b(\cdot,\cdot)$ is weakly coercive on the pair $Q_{\max(m,2)}$–P_0 for the given $\mathbf{n}_k$. ∎

In light of the lemmas discussed above, verification of weak coercivity allows for the formulation and proof of this paper's main theorem.

THEOREM 3.15. *Under Assumptions* 3.1 *and* 3.12, *existence of discrete solutions* $(\delta\mathbf{n}_h, \delta\lambda_h)$ *for each Newton linearization are guaranteed for the pair* V_h–Π_h. *In the case that* $\kappa = 1$ *or that* κ *satisfies the small data conditions of Lemma* 3.8, *such solutions are unique.*

Proof. Following a mixed formulation approach based on [10, 12, 13], Lemmas 3.5 and 3.6 guarantee the existence of a solution to the system given by (3.4) and (3.5). In the event that $\kappa = 1$ or that κ satisfies the small data assumptions, Lemma 3.7 or 3.8 coupled with Lemma 3.13 implies that the solution is also unique. ∎

3.5. Error analysis. In the previous section, the derived weak coercivity constant depends on the mesh parameter h. Therefore, as h approaches zero so too does the weak coercivity constant for the pair V_h and Π_h. However, the convergence of the scheme for the enriched Lagrangian finite-element spaces composing V_h is only slightly compromised. In this section, we derive approximation error bounds for the discrete solution. Throughout this section, it is assumed that Assumptions 3.1 and 3.12 apply. Let $(\mathbf{u}, q)$ represent a solution to the continuum variational system given by (3.1) and (3.2), and let $(\mathbf{u}_h, q_h)$ be the unique solution to the discrete system in (3.4) and (3.5). As above, denote the dimension of Ω by $N = 2, 3$.

LEMMA 3.16. *Let* Π_h *and* V_h *be defined as in* (3.10) *and* (3.11) *with* $m = 2$. *Under Assumptions* 3.1 *and* 3.12, *for* $\mathbf{u} \in H^3(\Omega)^3$ *and* $q \in H^1(\Omega)$ *there exists* $C_a > 0$ *such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_{DC} \leq C_a h\big(\|\mathbf{u}\|_3 + \|q\|_1\big). \tag{3.43}$$

Proof. Let α_0 denote the coercivity constant from either Lemma 3.7 or 3.8. Furthermore, let ζ denote the h-dependent weak coercivity constant derived in Lemma

3.13. By Theorem 5.2.2 in [10],

$$\|\mathbf{u}-\mathbf{u}_h\|_{DC} \le \frac{4C_AC_B}{\alpha_0\zeta}E_u + \frac{C_B}{\alpha_0}E_q, \tag{3.44}$$

where C_A and C_B are the continuity constants associated with $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, respectively, and

$$E_u = \inf_{v_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_{DC}, \qquad E_q = \inf_{\gamma_h\in\Pi_h}\|q-\gamma_h\|_0.$$

Note that

$$\inf_{v_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_{DC} \le C_f \inf_{v_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_1,$$

where C_f is the constant used in (3.37). Let $\mathcal{I}^h f$ denote the global interpolant of f over the appropriate finite-element space. Since $\{\mathcal{T}_h\}$ is quasi-uniform, it is, in particular, nondegenerate. Therefore, applying [13, Theorem 4.4.24] to the discrete space V_h, there exists a $C_5>0$, such that

$$\left(\sum_{T\in\mathcal{T}_h}\|\mathbf{v}-\mathcal{I}^h\mathbf{v}\|^2_{H^1(T)}\right)^{1/2} = \|\mathbf{v}-\mathcal{I}^h\mathbf{v}\|_1 \le C_5h^2\|\mathbf{v}\|_3 \qquad \forall\mathbf{v}\in H^3(\Omega).$$

This implies that if $\mathbf{u}\in H^3(\Omega)^3$, then

$$\inf_{\mathbf{v}_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_{DC} \le C_fC_5h^2\|\mathbf{u}\|_3. \tag{3.45}$$

For Π_h, Theorem 3.1.6 in [15] implies that there exists a $C_6>0$ such that

$$\|\gamma-\mathcal{I}^h\gamma\|_0 \le C_6h\|\gamma\|_1 \qquad \forall\gamma\in H^1(\Omega).$$

Hence, if $q\in H^1(\Omega)$,

$$\inf_{\gamma_h\in\Pi_h}\|q-\gamma_h\|_0 \le C_6h\|q\|_1. \tag{3.46}$$

Combining (3.45) and (3.46) with (3.44) yields the error estimate

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}_h\|_{DC} &\le \frac{4C_AC_B}{\alpha_0\zeta}C_fC_5h^2\|\mathbf{u}\|_3 + \frac{C_B}{\alpha_0}C_6h\|q\|_1\\ &= \frac{18C_AC_BC_f^2C_*\sqrt{\beta C_d}C_5}{\alpha\rho^N\alpha_0}h\|\mathbf{u}\|_3 + \frac{C_BC_6}{\alpha_0}h\|q\|_1.\end{aligned}$$

Taking $C_a=\max(\frac{18C_AC_BC_f^2C_*\sqrt{\beta C_d}C_5}{\alpha\rho^N\alpha_0}, \frac{C_BC_6}{\alpha_0})$, (3.43) is obtained. ∎

Thus, the approximation is convergent for V_h–Π_h but with an order of suboptimality, due to the weak coercivity constant's dependence on the mesh parameter. However, use of a discrete $H^{-1}(\Omega)$ norm for the space Π_h is currently being considered as a means of eliminating this mesh dependence.

Remark 3.17. The constant C_A is dependent on C_{sup} and, therefore, complexities across iterations similar to those discussed in Remark 3.9 may arise. The error analysis presented above deals with a fixed $\mathbf{n}_k$ across grids and a variational problem with true Newton correction $\mathbf{u}$. It demonstrates that, for the fixed variational problem, the solution on successively finer grids converges to $\mathbf{u}$ with order h. Assumptions similar to those in Remark 3.9 would be needed to ensure uniformity across iterations or grid refinements.

3.6. Practical choice of finite elements. The bubble enrichment discussed above is nonstandard in its incorporation of $\mathbf{n}_k$ in the construction of the bubbles. Therefore, during numerical implementation, it was desirable to find an experimentally stable, standard, finite-element pair closely related to the spaces discussed above. It was observed that Q_1–Q_1 finite-element discretizations resulted in singular matrices. This implies that Q_1–Q_1 is not a pair for which $b(\cdot,\cdot)$ is weakly coercive. Such a phenomenon is not unique. For example, instabilities arise for equal order elements in Galerkin approaches to both the Stokes equations [21] and the Navier–Stokes equations [25].

On the other hand, in the numerical experiments to be discussed below, mixed finite-element approaches, such as Q_2–P_0 discretizations, experimentally appear to admit weak coercivity without the need for rising order finite-element implementations or bubble enrichments. Corollary 3.14 implies that for a piecewise constant initial iterate, the update element space Q_2–P_0 implies weak coercivity for the first Newton iteration. With this assurance, coupled with the empirical weak coercivity evidence for Q_2–P_0, we employ Q_2–P_0 spaces for all iterations in the experiments below. In the event that singular matrices occur for the Q_2–P_0 discretization of a particular problem, the bubble enriched finite-element pair V_h–Π_h, defined in (3.10) and (3.11), may be implemented and is particularly attractive because the rising order of the bubble functions, $b_T\mathbf{n}_k|_T$, on each element does not increase the number of unknowns at each Newton iteration.

4. Numerical methodology. The algorithm to perform the minimization discussed in previous sections has three stages; see Algorithm 1. The outermost phase is nested iteration [45], which begins on a specified coarsest grid level. Newton iterations are performed on each grid, updating the current approximation after each step. The stopping criterion for the Newton iterations at each level is based on a specified tolerance for the current approximation's conformance to the first-order optimality conditions in the standard Euclidean l_2 norm. In the numerical experiments to follow, this tolerance was always 10^{-3}. The resulting approximation is then interpolated to a finer grid. The current implementation performs uniform grid refinement after each set of Newton iterations. As noted above, nonconvexity of the minimization problem increases the difficulty of finding global minimizers. As with many Newton-based optimization algorithms, poor initial guesses can lead to only locally optimal solutions or divergence. However, the combination of nested iteration and damped Newton

Algorithm 1. Newton's method minimization algorithm with nested iteration.

0. Initialize $(\mathbf{n}_0, \lambda_0)$ on coarse grid.
while *Refinement limit not reached* **do**
  while *First-order optimality conformance threshold not satisfied* **do**
    1. Set up discrete linear system (2.9) on current grid, H.
    2. Solve for $\delta\mathbf{n}_H$ and $\delta\lambda_H$.
    3. Compute $\mathbf{n}_{k+1}$ and λ_{k+1} as in (4.1).
  end
  4. Uniformly refine the grid.
  5. Interpolate $\mathbf{n}_H \to \mathbf{n}_h$ and $\lambda_H \to \lambda_h$.
end

stepping greatly improves reliability and convergence to global minima in the numerical results to follow.

The Newton iteration systems are constructed by applying finite-element discretizations on each grid. The resulting, relatively sparse, matrix has the anticipated saddle-point block structure

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix}.$$

The matrix is inverted using LU decomposition in order to solve for the discrete updates $\delta\mathbf{n}_h$ and $\delta\lambda_h$. Finally, an incomplete Newton correction is performed. That is, the new iterates are given by

$$\begin{bmatrix} \mathbf{n}_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{n}_k \\ \lambda_k \end{bmatrix} + \omega \begin{bmatrix} \delta\mathbf{n}_h \\ \delta\lambda_h \end{bmatrix}, \tag{4.1}$$

where $\omega \leq 1$. This is to ensure relatively strict adherence to the constraint manifold, which is necessary for the well-posedness discussed above. For this algorithm, ω is chosen to begin at 0.2 on the coarsest grid and increases by 0.2, to a maximum of 1, after each grid refinement, so that as the approximation converges, larger Newton steps are taken. For complicated boundary conditions, such damped Newton steps are important in preventing method divergence. The grid management and discretizations are implemented using the deal.II finite-element library, which is an aggressively optimized and parallelized open-source library widely used in scientific computing [7]. In practice, as discussed above, Q_2–P_0 discretizations were observed to experimentally admit weak coercivity. Therefore, Q_2–P_0 elements were used to approximate $\delta\mathbf{n}_h$ and $\delta\lambda_h$ on each grid for the numerical tests.

4.1. Free elastic numerical results. The general test problem in this section considers a classical domain with two parallel substrates placed at distance $d = 1$ apart. The substrates run parallel to the xz-plane and perpendicular to the y-axis. It is assumed that this domain represents a uniform slab in the xy-plane. That is, $\mathbf{n}$ may have a nonzero z component but $\frac{\partial \mathbf{n}}{\partial z} = \mathbf{0}$. Hence, we consider the 2-D domain $\Omega = \{(x, y) \mid 0 \leq x, y \leq 1\}$. The problem assumes periodic boundary conditions at the edges $x = 0$ and $x = 1$. Dirichlet boundary conditions are enforced on the y-boundaries. As discussed above, the simplification outlined in (2.4) is relevant for this domain and boundary conditions.

The first numerical experiment is run on one of the simplest configurations of this type. Along each of the substrates the liquid crystal rods are uniformly aligned parallel to the x-axis. The relevant Frank constants are $K_1 = K_2 = K_3 = 1$. The problem is solved on a 4×4 coarse grid with five successive uniform refinements resulting in a 128×128 fine grid. The initial guess and computed, converged solution are displayed in Figure 1.

The final minimized functional energy is $\mathcal{F}_1 = 0$, compared to an initial energy of 5.467. In Table 1, the number of Newton iterations per grid is detailed as well as the conformance of the solution to the first-order optimality conditions after the first and final Newton steps, respectively, on each grid. Assuming the presence of solvers that scale linearly with the number of nonzeros in the matrix, the work required in these iterations is roughly 1.34 times that of assembling and solving a single linearization step on the finest grid. In contrast, without nested iteration, the algorithm requires 21 damped Newton steps on the 128×128 finest grid alone, to satisfy the tolerance limit.

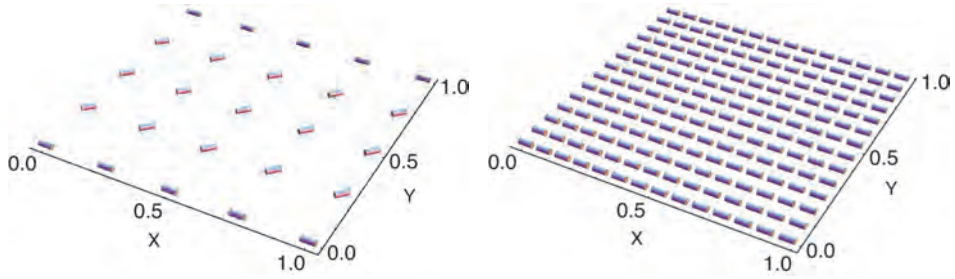

FIG. 1. *Initial guess (left) on* 4×4 *mesh with initial free energy of* 5.467 *and resolved solution (right) on* 128×128 *mesh (restricted for visualization) with final free energy of* 0 *for a uniformly aligned boundary.*

TABLE 1

Grid and solution progression for uniform free elastic boundary conditions with initial and final residuals for the first-order optimality conditions, minimum and maximum director deviations from unit length at the quadrature nodes, and final functional energy on each grid.

Grid dim.	Newton iter.	Init. res.	Final res.	Deviation in $\lvert\mathbf{n}\rvert^2$	Final energy
4×4	18	4.35e-00	4.39e-04	6.17e-06, 5.54e-05	4.941e-08
8×8	1	2.44e-04	9.74e-05	1.25e-06, 2.26e-05	7.905e-09
16×16	1	5.48e-05	1.10e-05	1.26e-07, 4.55e-06	3.162e-10
32×32	1	6.42e-06	1.35e-11	4.20e-14, 4.30e-11	7.932e-21
64×64	1	6.77e-12	6.37e-14	-4.00e-16, 0	0
128×128	1	1.30e-13	1.14e-13	-4.00e-16, 0	0

The application of damped Newton steps becomes even more important when beginning on finer grids with a rough initial guess, as divergence can be more prevalent. Table 1 also reveals the performance of the algorithm with respect to the pointwise constraint, presenting the increasingly tighter minimum and maximum director deviations from unit length at the quadrature nodes. The computed equilibrium solution behaves as expected with the rods uniformly aligning parallel to the x-axis.

The second test, run for the free elastic slab problem, incorporates twist boundary conditions and unequal Frank constants. On the lower slab, along $y = 0$, the nematic rods are aligned parallel to the x-axis. For the upper slab, the rods are uniformly aligned along the z-axis. The relevant constants for this run are $K_1 = 1$, $K_2 = 1.2$, and $K_3 = 1$. This implies that $\kappa = K_2/K_3 > 1$. The solves are again performed on a 4×4 coarse grid, uniformly ascending to a 128×128 fine grid. The expected configuration for such boundary conditions is a twisted equilibrium solution along the y-axis. Indeed, the numerically resolved solution in Figure 2, displayed alongside the initial guess, demonstrates such a twist. The final minimized functional energy is $\mathcal{F}_1 = 1.480$, compared to the initial guess energy of 12.534. Table 2 enumerates the algorithm run attributes.

As in Table 1 above, a sizable majority of the Newton iteration computations are isolated to the coarsest grids, with the finest grids requiring only one Newton iteration to reach the residual tolerance limit. Therefore, most of the computational cost is also isolated to the cheaper coarse grids rather than the finer levels. Here, the total work required is approximately 1.43 times that of assembling and solving a single linearization step on the finest grid. Without nested iteration, 22 damped Newton steps are required on the finest grid to compute the equilibrium solution.

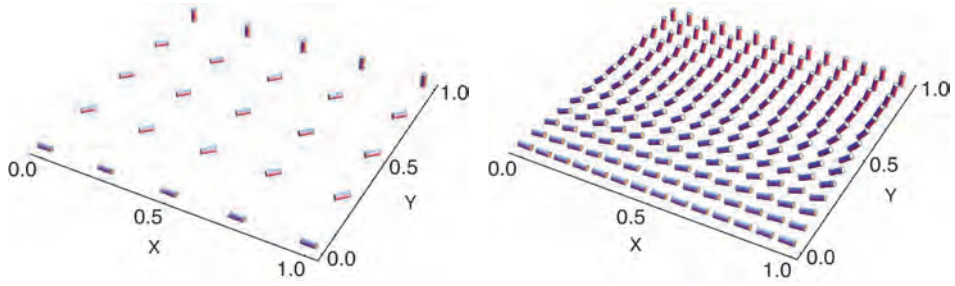

FIG. 2. *Initial guess (left) on* 4×4 *mesh with initial free energy of* 12.534 *and resolved solution (right) on* 128×128 *mesh (restricted for visualization) with final free energy of* 1.480 *for a twist boundary.*

TABLE 2

Grid and solution progression for the free elastic problem with twist boundary conditions with initial and final residuals for the first-order optimality conditions, minimum and maximum director deviations from unit length at the quadrature nodes, and final functional energy on each grid.

Grid dim.	Newton iter.	Init. res.	Final res.	Deviation in $\lvert\mathbf{n}\rvert^2$	Final energy
4×4	19	6.71e-00	3.97e-04	-5.69e-05, 1.50e-04	1.481
8×8	5	1.80e-02	1.84e-04	-4.10e-06, 2.57e-06	1.480
16×16	2	4.51e-03	1.80e-04	-3.27e-07, 1.51e-07	1.480
32×32	2	1.13e-03	2.09e-14	-1.47e-08, 6.88e-09	1.480
64×64	1	2.82e-04	4.31e-11	-9.21e-10, 4.31e-10	1.480
128×128	1	7.05e-05	1.36e-12	-5.75e-11, 2.69e-11	1.480

In the final numerical runs, letting $r = 0.25$ and $s = 0.95$, the boundary conditions considered are

$$\begin{aligned} n_1 &= 0, \\ n_2 &= \cos\left(r(\pi + 2\tan^{-1}(X_m) - 2\tan^{-1}(X_p))\right), \\ n_3 &= \sin\left(r(\pi + 2\tan^{-1}(X_m) - 2\tan^{-1}(X_p))\right), \end{aligned}$$

where $X_m = \frac{-s\sin(2\pi(x+r))}{-s\cos(2\pi(x+r))-1}$ and $X_p = \frac{-s\sin(2\pi(x+r))}{-s\cos(2\pi(x+r))+1}$. Such boundary conditions are meant to simulate nano-patterned surfaces important in current research [4, 5]. Even in the absence of electric fields, such patterned surfaces result in complicated director configurations throughout the interior of Ω.

A similar grid progression to the cases above is applied. The Frank elastic constants for the experiment are $K_1 = 1$, $K_2 = 0.62903$, and $K_3 = 1.32258$. This results in $\kappa < 1$. The final solution, as well as the initial guess, are displayed in Figure 3. Table 3, again, details the relevant output data. The computed configuration demonstrates the expected alignment and symmetries given the patterned surfaces.

The minimized functional energy is $\mathcal{F}_1 = 3.890$, compared to the initial guess energy of 13.242. The work required is approximately 3.06 times that of assembling and solving a single linearization step on the finest grid. On the other hand, without nested iterations, 22 damped Newton steps are required on the finest grid. Therefore, in all cases discussed, nested iteration is successful in significantly reducing the computational work necessary to compute an equilibrium solution.

Finally, to exhibit the robust convergence of the minimization algorithm, the approach is applied with varying initial guesses on the coarsest grid. The initial conditions considered include the tilt configuration in Figure 3, uniform horizontal

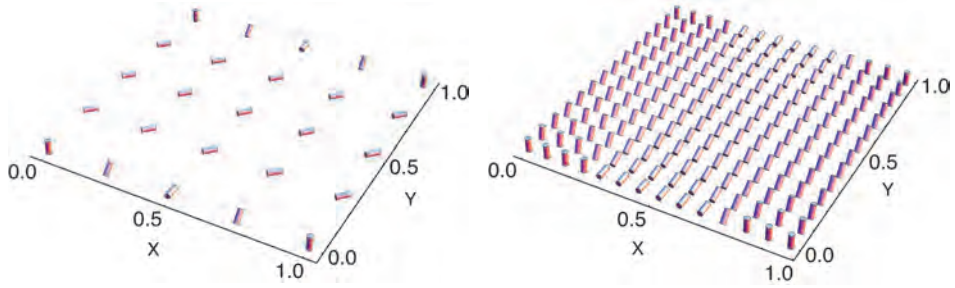

FIG. 3. *Initial guess (left) on* 4×4 *mesh with initial free energy of* 13.242 *and resolved solution (right) on* 128×128 *mesh (restricted for visualization) with final free energy of* 3.890 *for a nano-patterned boundary.*

TABLE 3

Grid and solution progression for patterned boundary conditions with initial and final residuals for the first-order optimality conditions, minimum and maximum director deviations from unit length at the quadrature nodes, and final functional energy on each grid.

Grid dim.	Newton iter.	Init. res.	Final res.	Deviation in $\lvert\mathbf{n}\rvert^2$	Final energy
4×4	19	7.04e-00	4.72e-04	-9.07e-02, 4.67e-02	2.521
8×8	9	1.20e-00	3.14e-04	-8.20e-02, 4.58e-02	3.194
16×16	6	1.06e-00	6.71e-05	-6.69e-02, 3.96e-02	3.674
32×32	3	8.22e-01	3.42e-12	-4.31e-02, 2.78e-02	3.885
64×64	3	5.04e-01	4.75e-14	-1.73e-02, 1.26e-02	3.900
128×128	2	2.24e-01	3.00e-09	-3.51e-03, 2.81e-03	3.890

TABLE 4

Run statistics displaying the initial and final free energy, final residual, and computational work in terms of assembling and solving a single linearization step on the finest grid for various initial conditions in the presence of nano-patterned surfaces.

Init. guess	Tilt	Horizontal	Vertical	Patterned	Rand. ($\omega = 0.1$)
Init. energy	13.24	20.47	10.28	7.18	24.16
Final energy	3.890	3.890	3.890	3.890	3.890
Final res.	3.00e-09	3.00e-09	3.00e-09	3.00e-09	3.00e-09
Work	3.059	3.060	3.059	3.059	3.733

or vertical alignments, a continuation of the patterned surface on the interior, and a random guess with periodic boundaries. The associated statistics are recorded in Table 4. Note that while the free energy across initial guesses differs greatly, the algorithm successfully converges to the global minimum in each case. Moreover, the computational work remains relatively steady for each guess. In all cases, nested iteration and damped Newton stepping are highly effective at improving convergence, especially for the case of randomized initial conditions. With the random initial guess, a reduced damping factor of 0.1 is necessary to promote convergence. Furthermore, such damping only yields convergence in the presence of nested iteration.

5. Summary and future work. We have discussed a constrained minimization approach for liquid crystal equilibrium configurations in the presence of free elastic effects. Due to the nonlinearity of the continuum first-order optimality conditions, Newton linearizations were derived. The resulting discrete systems were analyzed, and it was shown that solutions to the discretized Newton iterations exist. If $\kappa = 1$ or κ satisfies the conditions of the small data assumption in Lemma 3.8 and the assumptions of Lemma 3.13 hold, then unique solutions to the discrete Newton iterations are

guaranteed for the prescribed discrete spaces. Overall method convergence relies on the well-established convergence of Newton's method. Nested iteration and damping were effectively used to deal with inaccurate initial guesses. Finally, error analysis was conducted to demonstrate discrete update convergence with grid size.

Numerical results demonstrate the accuracy and efficiency of the algorithm in resolving some difficult features for free elastic effects. The experiments address problems that include unequal Frank constants and nano-patterned boundary conditions. The experiments also reveal the necessity for a mixed finite-element approach. Such a requirement exposes an interesting parallel to other problems with similar instabilities such as the Stokes and Navier–Stokes equations. The minimization approach overcomes some difficulties inherent to the liquid crystal equilibrium problem, such as the nonlinear unit-length constraint, and effectively deals with heterogeneous Frank constants. The algorithm also productively utilizes nested iteration to reduce computational costs by isolating much of the computational work to the coarsest grids. Such work allocation significantly reduces the effective number of Newton iterations on the finest grid, even for the nano-patterned boundary conditions example.

The above method is currently being extended to include electric and flexoelectric effects in order to more accurately capture physical phenomenon important to many applications, such as the study of bistable devices [17]. The rising complexity involved in these extensions presents interesting challenges, such as the appearance of more complicated saddle-point structures. Development and implementation of specifically tailored solvers for the systems encountered above, as well as those anticipated in future problems, is a priority.

Additionally, following [30], investigation into the use of $H^{-1}(\Omega)$ norms for the Lagrange multiplier to achieve discrete inf-sup stability independent of the mesh parameter, h, is being pursued. Furthermore, analysis of the Newton linearizations for the electric and flexoelectric augmentations will be undertaken. Future work will also include study of effective adaptive refinement and linearization tolerance schemes. Because the energy minimization formulation does not yield an obvious a priori error estimator, new techniques will be explored to flag cells for refinement and determine when grid refinement should occur.

Acknowledgments. The authors would like to thank Professors Thomas Manteuffel, Johnny Guzmán, and Ludmil Zikatanov for their useful contributions and suggestions.

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, 2nd ed., Elsevier, New York, 2003.
[2] J. H. Adler, T. J. Atherton, D. B. Emerson, and S. P. MacLachlan, *An Energy-Minimization Finite-Element Approach for the Frank-Oseen Model of Nematic Liquid Crystals: Continuum and Discrete Analysis*, Technical report, Tufts University, Boston, MA, 2014.
[3] F. Alouges, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.
[4] T. J. Atherton and J. H. Adler, *Competition of elasticity and flexoelectricity for bistable alignment of nematic liquid crystals on patterned surfaces*, Phys. Rev. E, 86 (2012).
[5] T. J. Atherton and J. R. Sambles, *Orientational transition in a nematic liquid crystal at a patterned surface*, Phys. Rev. E, 74 (2006).
[6] I. Babuska, *Error-bounds for finite element methods*, Numer. Math., 16 (1971), pp. 322–333.
[7] W. Bangerth, R. Hartmann, and G. Kanschat, *deal.II—a general purpose object oriented finite element library*, ACM Trans. Math. Softw., 33 (2007), pp. 24/1–24/27.
[8] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer., (2005), pp. 1–137.

[9] M. Benzi, E. Haber, and L. Taralli, *A preconditioning technique for a class of PDE-constrained optimization problems*, Adv. Comput. Math., 35 (2011), pp. 149–173.
[10] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*, Springer, New York, 2013.
[11] F. K. Bogner, R. L. Fox, and L. A. Schmit, *The generation of interelement compatible stiffness and mass matrices by the use of interpolation formulas*, in Proceedings of the Conference on Matrix Methods in Structural Mechanics, Wright Patterson AFB, Dayton, OH, 1965, pp. 397–444.
[12] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 1997.
[13] S. C. Brenner and L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1996.
[14] S. Chandrasekhar, *Liquid Crystals*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
[15] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Vol. 4, North-Holland, Amsterdam, 1978.
[16] R. Cohen, R. Hardt, D. Kinderlehrer, S. Lin, and M. Luskin, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, Springer-Verlag, New York, 1987, pp. 99–121.
[17] A. J. Davidson and N. J. Mottram, *Flexoelectric switching in a bistable nematic device*, Phys. Rev. E, 65 (2002).
[18] T. A. Davis and E. C. Gartland, Jr., *Finite element analysis of the Landau-de Gennes minimization problem for liquid crystals*, SIAM J. Numer. Anal., 35 (1998), pp. 336–362.
[19] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Clarendon Press, Oxford, UK, 1993.
[20] P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer, Berlin, 2004.
[21] C. R. Dohrmann and P. B. Bochev, *A stabilized finite element method for the Stokes problem based on polynomial pressure projections*, Internat. J. Numer. Methods Fluids, (2000).
[22] J. L. Ericksen, *Hydrostatic theory of liquid crystals*, Arch. Ration. Mech. Anal., 9 (1962), pp. 371–378.
[23] J. L. Ericksen, *Inequalities in liquid crystal theory*, Phys. Fluids, 9 (1966), pp. 1205–1207.
[24] P. E. Farrell, A. Birkisson, and S. W. Funke, *Deflation techniques for finding distinct solutions of nonlinear partial differential equations*, SIAM J. Sci. Comput., 37 (2015), pp. A2026–A2045.
[25] L. P. Franca and S. L. Frey, *Stabilized finite element methods:* II. *The incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.
[26] F. C. Frank, *On the theory of liquid crystals*, Discuss. Faraday Soc., 25 (1958), pp. 19–28.
[27] H. Fujita and T. Kato, *On the Navier-Stokes initial values problem* I, Arch. Ration. Mech. Anal., 16 (1964), pp. 269–315.
[28] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[29] R. Hardt, D. Kinderlehrer, and F.-H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Commun. Math. Phys., 105 (1986), pp. 547–570.
[30] Q. Hu, X.-C. Tai, and R. Winther, *A saddle point approach to the computation of harmonic maps*, SIAM J. Numer. Anal., 47 (2009), pp. 1500–1523.
[31] E. C. Gartland Jr. and A. Ramage, *A renormalized Newton method for liquid crystal director models with pointwise unit-vector constraints*, SIAM J. Numer. Anal., 53 (2015), pp. 251–278.
[32] J. P. F. Lagerwall and G. Scalia, *A new era for liquid crystal research: Applications of liquid crystals in soft matter, nano-, bio- and microtechnology*, Current Appl. Phys., 12 (2012), pp. 1387–1412.
[33] B. W. Lee and N. A. Clark, *Alignment of liquid crystals with patterned isotropic surfaces*, Science, 291 (2001), pp. 2576–2580.
[34] P. Lin, C. Liu, and H. Zhang, *An energy law preserving C0 finite element scheme for simulating the kinematic effects in liquid crystal flow dynamics*, J. Comput. Phys., (2007), pp. 1411–1427.
[35] C. Liu and H. Sun, *On energetic variational approaches in modeling the nematic liquid crystal flows*, Discrete Contin. Dyn. Syst., 23 (2009), pp. 455–475.
[36] C. Liu, H. Zhang, and S. Zhang, *Numerical simulations of hydrodynamics of nematic liquid crystals: Effects of kinematic transports*, Commun. Comput. Phys., 9 (2011), pp. 974–993.
[37] D. G. Luenberger, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[38] E. Marusic-Paloka, *Solvability of the Navier-Stokes system with L^2 boundary data*, Appl. Math. Optim., 41 (2000), pp. 365–375.
[39] H. M. Mourad, J. Dolbow, and I. Harari, *A bubble-stabilized finite element method for Dirichlet constraints on embedded interfaces*, Internat. J. Numer. Methods Engng, 69 (2006), pp. 1–21.
[40] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[41] A. Pandolfi and G. Napoli, *A numerical investigation of configurational distortions in nematic liquid crystals*, J. Nonlinear Sci., 21 (2011), pp. 785–809.
[42] R. Pierre, *Simple C^0 approximations for the computation of incompressible flows*, Comput. Methods Appl. Mech. Engrg, 68 (1988), pp. 205–227.
[43] A. Ramage and E. C. Gartland, Jr., *A preconditioned nullspace method for liquid crystal director modeling*, SIAM J. Sci. Comput., 35 (2013), pp. B226–B247.
[44] F. Reinitzer, *Beitrage zur kenntnis des cholesterins*, Monatsh. Chem., 9 (1888), pp. 421–441.
[45] G. Starke, *Gauss-Newton multilevel methods for least-squares finite element computations of variably saturated subsurface flow*, Computing, 64 (2000), pp. 323–338.
[46] I. W. Stewart, *The Static and Dynamic Continuum Theory of Liquid Crystals: A Mathematical Introduction*, Taylor and Francis, London, 2004.
[47] D. Thomsen, P. Keller, J. Naciri, R. Pink, H. Jeon, D. Shenoy, and B. Ratna, *Liquid crystal elastomers with mechanical properties of a muscle*, Macromolecules, 34 (2001), pp. 5868–5875.
[48] H. Wu, X. Xu, and C. Liu, *On the general Ericksen-Leslie system: Parodi's relation, well-posedness and stability*, Arch. Ration. Mech. Anal., (2013), pp. 59–107.
[49] M. Yamada, M. Kondo, J. Mamiya, Y. Yu, M. Kinoshita, C. Barrett, and T. Ikeda, *Photomobile polymer materials: Towards light-driven plastic motors*, Angew. Chem. Int., 47 (2008), pp. 4986–4988.
[50] H. Zhang and Q. Bai, *Numerical investigation of tumbling phenomena based on a macroscopic model for hydrodynamic nematic liquid crystals*, Commun. Comput. Phys., 7 (2010), pp. 317–332.