Numerische Mathematik
https://doi.org/10.1007/s00211-020-01133-z

Numerische Mathematik

A structure-preserving FEM for the uniaxially constrained Q-tensor model of nematic liquid crystals

Juan Pablo Borthagaray[1,2] · Ricardo H. Nochetto[3] · Shawn W. Walker[4]

Received: 11 November 2019 / Revised: 11 May 2020

Abstract
We consider the one-constant Landau-de Gennes model for nematic liquid crystals. The order parameter is a traceless tensor field $\mathbf{Q}$, which is constrained to be uniaxial: $\mathbf{Q} = s(\mathbf{n} \otimes \mathbf{n} - d^{-1}\mathbf{I})$ where $\mathbf{n}$ is a director field, $s \in \mathbb{R}$ is the degree of orientation, and $d \geq 2$ is the dimension. Building on similarities with the one-constant Ericksen energy, we propose a structure-preserving finite element method for the computation of equilibrium configurations. We prove stability and consistency of the method without regularization, and Γ-convergence of the discrete energies towards the continuous one as the mesh size goes to zero. We design an alternating direction gradient flow algorithm for the solution of the discrete problems, and we show that such a scheme decreases the energy monotonically. Finally, we illustrate the method's capabilities by presenting some numerical simulations in two and three dimensions including non-orientable line fields.

Keywords Liquid crystals · Finite Element Method · Gamma-convergence · Landau-de Gennes · Defects

Mathematics Subject Classification 65N30 · 65K10 · 35J70

1 Introduction

The liquid crystal state of matter is observed in certain materials as a mesophase between the crystalline and the isotropic liquid phases. Such a state may be obtained as a function of temperature between the two latter phases; in this case, these are

JPB has been supported in part by NSF grant DMS-1411808 and an AMS-Simons Travel Grant. RHN has been supported in part by NSF Grants DMS-1411808 and DMS-1908267. SWW has been supported in part by NSF Grant DMS-1555222 (CAREER).

✉ Juan Pablo Borthagaray
jpborthagaray@unorte.edu.uy

Extended author information available on the last page of the article

Published online: 14 July 2020

called thermotropic liquid crystals. Other classes include lyotropic and metallotropic liquid crystals, in which concentration of the liquid-crystal molecules in a solvent or the ratio between organic and inorganic molecules determine the phase transitions, respectively. In this paper, we consider thermotropic liquid crystals [23].

The physical state of a material can be described in terms of the translational and rotational motion of its constituent molecules. In a crystalline solid, molecules exhibit both long-range ordering of the positions of the centers and orientation of the molecules. As the substance is heated, the molecules gain kinetic energy and large molecular vibrations usually make these two ordering types disappear at the same temperature. This results in a fluid phase. However, in some materials, that typically consist of either rod-like or disc-like molecules, the long-range orientational ordering survives until a higher temperature than the long-range positional ordering. Such a state of matter is called *liquid crystalline*. Moreover, when long-range positional ordering is completely absent, the liquid crystal is regarded as *nematic*.

On average, nematic liquid crystal molecules are aligned with their long axes parallel to each other. At the macroscopic level, this means that there is a preferred direction; often, such a direction is a rotational symmetry axis. In such a case, the nematic liquid crystal phase is *uniaxial*. If, in contrast, there is no such rotational symmetry, then the material is in a *biaxial* state.

Depending on the choice of *order parameter* (cf. Sect. 2.1), several models for nematic liquid crystals have been proposed. Because the vast majority of thermotropic liquid crystals exhibit uniaxial behavior, this is often built into the modeling. If one takes as order parameter the orientation of the molecules $\mathbf{n}(x) \in \mathbb{S}^{d-1}$, for $x \in \Omega \subset \mathbb{R}^d$, then $\mathbf{n}$ is a harmonic mapping in the domain Ω; numerical methods for this model have been proposed, for example, in [1,3,11,19,30,41]. We refer also to [20,33,42,63] for discretizations of liquid crystal flows. It is often the case that liquid crystal configurations display *defects*, that is, that the molecular orientation is not continuous in some regions of the material. Harmonic map models do not allow for point defects if $d = 2$ or line defects if $d = 3$, because the energy is singular.

However, if besides the liquid crystal molecule orientation $\mathbf{n}$ one considers a scalar variable $s(x)$ that represents the degree of alignment that molecules have with respect to $\mathbf{n}$, then the equilibrium configuration minimizes the Ericksen energy [23,24,61]. Minimizers of such an energy can exhibit nontrivial defects, as the parameter s can relax a large contribution from $|\nabla \mathbf{n}|$, and wherever the degree of alignment s vanishes, the resulting Euler-Lagrange equation for $\mathbf{n}$ is degenerate. Finite element methods for the Ericksen model have been used to approximate both equilibrium configurations [46–48] and dynamics [10] of the molecular orientation.

If one considers the probability distribution of the liquid crystal molecules orientation and chooses to use its second moments to define an order parameter, then this leads to the Landau-de Gennes model. In such a model, the order parameter is a tensor field $\mathbf{Q}(x)$ that measures the discrepancy between the probability distribution at $x \in \Omega$ and a uniform distribution on $\mathbb{S}^{d-1}$. Numerical methods for the Landau-de Gennes energy are considered in [6,12,22,29,35,53].

In this work, we shall be concerned with uniaxial nematic liquid crystals in $\mathbb{R}^d$ for $d \geq 2$; we present numerical experiments for $d = 2, 3$. Our goal is to design a finite element method for a uniaxially-constrained $\mathbf{Q}$-tensor model, and to prove stability and

convergence properties. More precisely, we prove that if the corresponding meshes are weakly acute, then our discrete energy Γ-converges to the continuous one as the mesh size goes to zero. Our method can handle the degeneracy introduced by a vanishing degree of orientation without any regularization. Moreover, because the $\mathbf{Q}$-tensor approach incorporates a head-to-tail symmetry into the modeling, our approach is able to capture *non-orientable* equilibrium configurations.

The paper is organized as follows. In Sect. 2, we discuss modeling of the equilibrium states of liquid crystals. We examine the Landau-de Gennes and Ericksen energies, and discuss the capabilities of these models to capture defects. Section 3 is devoted to the formulation of the problem we study in this paper. Such a section includes a discussion on previous work for the Ericksen model [47], which is instrumental for our numerical method. We introduce the discrete setting for the uniaxially-constrained Landau-de Gennes energy and prove key energy inequalities in Sect. 4. Afterwards, in Sect. 5, we prove the Γ-convergence of the discrete energies. For the computation of discrete minimizers, in Sect. 6 we propose a gradient flow and prove a strictly monotone energy decreasing property. Finally, Sect. 7 presents numerical experiments for $d = 2, 3$ illustrating the capabilities of our method.

2 Modeling of nematic liquid crystals

We discuss some elementary properties of the so-called $\mathbf{Q}$-tensors and review three models for the equilibrium states of nematic liquid crystals, which derive from minimizing an energy (see [23,45,61] for more details on the modeling of liquid crystals).

2.1 Order parameters

For a particular material, the transition between phases of different symmetry can be described in terms of an order parameter. Such a parameter represents the extent to which the configuration of the more symmetric phase differs from that of the less symmetric phase.

For the sake of clarity, we fix the dimension to be $d = 3$ in the following discussion. To avoid modeling individual liquid crystal molecules, that is very expensive computationally, we pursue a macroscopic description of liquid crystals. Namely, let us describe the orientation of the nematic molecules by a probability distribution in the unit sphere; this gives raise to a tensor field $\mathbf{Q} : \Omega \to \mathbb{R}^{3\times 3}$, which is required to be symmetric and traceless a.e. [23,45,61].

We can further characterize $\mathbf{Q}$ by its eigenframe and is often written in the form:

$$\mathbf{Q} = s_1(\mathbf{n}_1 \otimes \mathbf{n}_1) + s_2(\mathbf{n}_2 \otimes \mathbf{n}_2) - \frac{1}{3}(s_1 + s_2)\mathbf{I},$$

where $\mathbf{n}_1, \mathbf{n}_2$ are orthonormal eigenvectors of $\mathbf{Q}$, with eigenvalues given by

$$\lambda_1 = \frac{2s_1 - s_2}{3}, \quad \lambda_2 = \frac{2s_2 - s_1}{3}, \quad \lambda_3 = -\frac{s_1 + s_2}{3}, \tag{1}$$

where λ_3 corresponds to the eigenvector $\mathbf{n}_3 \perp \mathbf{n}_1, \mathbf{n}_2$. The eigenvalues of $\mathbf{Q}$ are constrained by

$$-\frac{1}{3} \leq \lambda_i \leq \frac{2}{3}, \quad i = 1, 2, 3. \tag{2}$$

When all eigenvalues are equal, since $\mathbf{Q}$ is traceless, we must have $\lambda_1 = \lambda_2 = \lambda_3 = 0$ and $s_1 = s_2 = 0$, i.e. the distribution of liquid crystal molecules is isotropic. If two eigenvalues are equal, i.e.

$$\begin{aligned} \lambda_1 = \lambda_2 &\quad \Leftrightarrow \quad s_1 = s_2, \\ \lambda_1 = \lambda_3 &\quad \Leftrightarrow \quad s_1 = 0, \\ \lambda_2 = \lambda_3 &\quad \Leftrightarrow \quad s_2 = 0, \end{aligned}$$

then we encounter a *uniaxial* state, in which either molecules prefer to orient in alignment with the simple eigenspace (in case it corresponds to a positive eigenvalue) or perpendicular to it (in case it corresponds to a negative eigenvalue). If all three eigenvalues are distinct, then the state is called *biaxial*.

Remark 1 (biaxial nematics) In this work, we regard liquid crystal molecules as elongated rods. Naturally, most liquid crystal molecules do not possess such an axial symmetry. If the molecules resemble a lath more than a rod, it is expected that the energy interaction can be minimized if the molecules are fully aligned; this necessarily involves a certain degree of biaxiality. Roughly, this was the rationale behind the prediction of the biaxial nematic phase by Freiser [27].

Since that seminal work, empirical evidence of biaxial states in certain lyotropic liquid crystals has been well documented (see [65], for example). Nevertheless, for thermotropic liquid crystals the nematic biaxial phase remained elusive for a long period, and was first reported long after Freiser's original prediction [43,51]. As pointed out by Sonnet and Virga [56, Section 4.1],

> The vast majority of nematic liquid crystals do not, at least in homogeneous equilibrium states, show any sign of biaxiality.

We refer to [14] for further quantitative discussion via computations. In light of Remark 1, in Sect. 3 we shall consider a uniaxially-constrained model. More precisely, we assume that $\mathbf{Q}$ takes the uniaxial state

$$\mathbf{Q} = s\left(\mathbf{n} \otimes \mathbf{n} - \frac{1}{3}\mathbf{I}\right), \tag{3}$$

where $\mathbf{n}$ is the main eigenvector with eigenvalue $\lambda = 2s/3$; the other two eigenvalues equal $-s/3$. The scalar field s is called the *degree of orientation* of the liquid crystal molecules. Taking into account identities (1) and the restrictions (2), it follows that the physically meaningful range is $s \in (-1/2, 1)$. In case $s = 1$, the molecular long axes are in perfect alignment with the direction of $\mathbf{n}$, whereas $s = -1/2$ represents the state in which all molecules are perpendicular to $\mathbf{n}$.

An advantage of the uniaxially constrained model we consider in this work over the standard one-constant Landau-de Gennes model is that the bulk potential acts directly on the parameter s instead of over the $\mathbf{Q}$-tensor. Because s can be directly related to the eigenvalues of $\mathbf{Q}$, this allows one guarantee that the eigenvalues of $\mathbf{Q}$ lie in a physically meaningful range. In contrast, in the low-temperature regime, the Landau-de Gennes model can lead to $\mathbf{Q}$ having physically unrealistic eigenvalues [44].

Remark 2 (problems in $2d$) The discussion above simplifies considerably when $d = 2$. Indeed, since $\mathbf{Q}$ is symmetric and traceless, it must be uniaxial, and writing it as $\mathbf{Q} = s\left(\mathbf{n} \otimes \mathbf{n} - \frac{1}{2}\mathbf{I}\right)$, we deduce that its eigenvalues are $\lambda_1 = s/2$, with eigenvector $\mathbf{n}$, and $\lambda_2 = -\lambda_1$, with eigenvector $\mathbf{n}^{\perp}$. Because eigenvalues are constrained to satisfy $\lambda_i \in (-1/2, 1/2)$, we deduce that the physically meaningful range is $s \in (-1, 1)$. Actually, one can further simplify to $s \in [0, 1)$ by noting that a state with director $\mathbf{n}$ and degree of orientation $s < 0$ is equivalent to a state with director $\mathbf{m} \perp \mathbf{n}$ and degree of orientation $-s$.

Remark 3 (thin films) For simplicity, in this work we consider $\mathbf{Q}$ to be a square tensor with the same dimension as the spatial domain. With minor modifications, our approach carries to the case where these dimensions are different, such as three dimensional tensors on thin films.

2.2 Continuum mechanics

Given the order parameter $\mathbf{Q}$, we still need a model to determine its state as a function of space. For modeling equilibrium states, this amounts to finding minimizers of an energy functional. A common approach from continuum mechanics [34,58,60] is to construct the "simplest" functional possible that is quadratic in gradients of the order parameter while obeying standard laws of physics, such as frame indifference and material symmetries. We assume all equations have been non-dimensionalized; see [28] for the case of the Landau-de Gennes model.

2.2.1 Landau-de Gennes model

Using $\mathbf{Q}$ as the order parameter, we obtain the Landau-de Gennes model, in which the energy is given by [23,56]:

$$\begin{aligned} E_{\mathrm{LdG}}[\mathbf{Q}] &:= \int_{\Omega} \mathcal{W}_{\mathrm{LdG}}(\mathbf{Q}, \nabla\mathbf{Q})\, dx + \frac{1}{\eta_{\mathrm{B}}} \int_{\Omega} \phi_{\mathrm{LdG}}(\mathbf{Q})\, dx, \\ \mathcal{W}_{\mathrm{LdG}}(\mathbf{Q}, \nabla\mathbf{Q}) &:= \frac{1}{2}\left(L_1 |\nabla\mathbf{Q}|^2 + L_2 |\nabla \cdot \mathbf{Q}|^2 + L_3 (\nabla\mathbf{Q})^T : \nabla\mathbf{Q}\right). \end{aligned} \tag{4}$$

Above, $\{L_i\}_{i=1}^3$, η_{B} are material parameters, ϕ_{LdG} is a bulk (thermotropic) potential and

$$|\nabla\mathbf{Q}|^2 := (\partial_k Q_{ij})(\partial_k Q_{ij}), \quad |\nabla \cdot \mathbf{Q}|^2 := (\partial_j Q_{ij})^2, \quad (\nabla\mathbf{Q})^T : \nabla\mathbf{Q} := (\partial_j Q_{ik})(\partial_k Q_{ij}),$$

and we use the convention of summation over repeated indices. This is a relatively simple form for $\mathcal{W}_{\mathrm{LdG}}$; more complicated models can also be considered [23,45,56].

The bulk potential ϕ_{LdG} is a double-well type of function that controls the eigenvalues of $\mathbf{Q}$. The simplest form is given by

$$\phi_{\mathrm{LdG}}(\mathbf{Q}) = K + \frac{A}{2}\mathrm{tr}(\mathbf{Q}^2) - \frac{B}{3}\mathrm{tr}(\mathbf{Q}^3) + \frac{C}{4}\left(\mathrm{tr}(\mathbf{Q}^2)\right)^2, \tag{5}$$

where A, B, C are material parameters such that A has no sign, and B, C are positive; K is a convenient constant. It is typical to let $A \leq 0$ since we are interested in uniaxial states, so throughout this paper we assume that

$$A \leq 0, \quad B, C > 0,$$

which implies that $\phi_{\mathrm{LdG}}(\mathbf{Q}) \geq 0$ assuming K is suitably chosen.

In two dimensions, $\mathrm{tr}(\mathbf{Q}^3) = 0$, because $\mathbf{Q}^2 = \frac{s^2}{4}\mathbf{I}$. Hence, B is irrelevant in $2d$, and it is necessary that A be strictly negative in order to have a stable nematic phase. This also implies that ϕ_{LdG} is an *even* function of s if $\mathbf{Q}$ is uniaxial (see Remark 2).

As a simplification, one can take $L_1 = 1$, $L_2 = L_3 = 0$ in (4) to obtain a *one-constant approximation*

$$E_{\mathrm{LdG,one}}[\mathbf{Q}] := \frac{1}{2}\int_\Omega |\nabla \mathbf{Q}|^2\, dx + \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \phi_{\mathrm{LdG}}(\mathbf{Q})\, dx. \tag{6}$$

2.2.2 Ericksen model

Though the Landau-de Gennes model is quite general, it can be fairly expensive when $d = 3$. In such a case, since $\mathbf{Q} \in \mathbb{R}^{3\times 3}$ and symmetric, it has five independent variables. Moreover, the bulk potential ϕ_{LdG} is a non-linear function of $\mathbf{Q}$, which couples all five variables together when seeking a minimizer of E_{LdG}.

Assuming that $\mathbf{Q}$ is uniaxial (3), we can take s *and* $\mathbf{n}$ as order parameters. In the same way as (6), we have a *one-constant* Ericksen model:

$$E_{\mathrm{erk}}[s, \mathbf{n}] := \frac{\kappa}{2}\int_\Omega |\nabla s|^2\, dx + \frac{1}{2}\int_\Omega s^2 |\nabla \mathbf{n}|^2\, dx + \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \phi_{\mathrm{erk}}(s)\, dx, \tag{7}$$

where $\kappa > 0$ is a single material parameter, and ϕ_{erk} is a double-well potential acting on s, which is taken from the Landau-de Gennes case: $\phi_{\mathrm{erk}}(s) = \phi_{\mathrm{LdG}}(\mathbf{Q}(s))$, where $\mathbf{Q}$ is any matrix having the form (3).

Remark 4 (Oseen–Frank model) In case the degree of orientation is a non-zero constant field, the energy E_{erk} effectively reduces to the Oseen–Frank energy [61]: $E_{\mathrm{OF}}[\mathbf{n}] := \int_\Omega |\nabla \mathbf{n}|^2$. The Oseen–Frank model has been used extensively in the modeling of liquid crystal-based flat panel displays. Minimizers of the one-constant energy in such a model are director fields $\mathbf{n}\colon \Omega \to \mathbb{S}^{d-1}$ satisfying $\Delta \mathbf{n} - \lambda \mathbf{n} = 0$, where λ is a Lagrange multiplier that enforces the unit length constraint.

In the Oseen–Frank model, point defects in three dimensional domains have finite energy. However, this model is incapable of capturing higher-dimensional defects, that is, defects supported either on lines or planes. Since these naturally occur in many liquid crystal systems, this is a major inherent limitation of the Oseen–Frank model.

We point out that (7) is *degenerate,* in the sense that s may vanish; this allows for $\mathbf{n}$ to have discontinuities (i.e. defects) with finite energy. Indeed, the hallmark of this model is to regularize defects using s, but still retain part of the Oseen–Frank model. Discontinuities in $\mathbf{n}$ may still occur in the singular set

$$\mathbb{S} := \{x \in \Omega : \ s(x) = 0\}. \tag{8}$$

For problems in $\mathbb{R}^3$, because $\mathbf{n} \in \mathbb{S}^2$, it is uniquely defined by two parameters. Thus, in such a case the Ericksen model only has three scalar order parameters, as opposed to five in the Landau-de Gennes model. Another advantage of the Ericksen model is that s and $\mathbf{n}$ provide a natural way to split the system which is convenient for numerical purposes. Additionally, the parameter κ in (7) plays a major role in the occurrence of defects. Assuming that s equals a sufficiently large positive constant on $\partial\Omega$, if κ is large, then $\int_\Omega \kappa|\nabla s|^2 dx$ dominates the energy and s stays close to such a positive constant within the domain Ω. Thus, defects are less likely to occur. If κ is small (say $\kappa < 1$), then $\int_\Omega s^2|\nabla \mathbf{n}|^2 dx$ dominates the energy, and s may vanish in regions of Ω and induce a defect. This is confirmed by the numerical experiments in [46,47].

Remark 5 (orientability) Director field models—either Oseen–Frank or Ericksen—are more than adequate in some situations, although in general they introduce a *nonphysical orientational bias* into the problem. Even though liquid crystal molecules may be polar, in nematics one always finds that the states with $\mathbf{n}$ and $-\mathbf{n}$ are equivalent [31]. At the molecular level, this means that the same number of molecules point "up" and "down." Therefore, *line-fields* are more appropriate for modeling nematic liquid crystals.

Another issue with the use of the vector field $\mathbf{n}$ as an order parameter instead of the matrix $\mathbf{Q}$ is that the only allowable defects in such a case are *integer order* defects. On the other hand $\mathbf{Q}$, specifically $\mathbf{n} \otimes \mathbf{n}$ in (3), is able to represent *line fields* having half-integer defects. These have been largely observed and documented in experiments; see for example [17,49] and references therein. We point out that, if a line field is *orientable*, then a vector field representation is essentially equivalent [8,9].

3 Mathematical formulation

In this work, we will be concerned with the one-constant energy for $\mathbf{Q}$, given by (6). Enforcing $\mathbf{Q}$ to be symmetric and traceless, one can, in principle, directly minimize such an energy. For three-dimensional problems, a standard approach to finding minimizers [5,36,56,59] is to express $\mathbf{Q}(x)$ as

$$\mathbf{Q}(x) = \begin{bmatrix} q_1 & q_3 & q_4 \\ q_3 & q_2 & q_5 \\ q_4 & q_5 & -(q_1+q_2) \end{bmatrix}, \tag{9}$$

i.e. minimize (6) with respect to the order parameters $\{q_i(x)\}_{i=1}^5$. This approach has two drawbacks.

First, a basic argument shows that minimizers of $\int_\Omega \phi_{\mathrm{LdG}}$ have the form of a uniaxial nematic (3) [56]. This is *false* for $E_{\mathrm{LdG,one}}$ in (6) with general boundary conditions. Thus, minimizers of the form (9) *violate* the algebraic form of (3) and exhibit a *biaxial escape* [38,50,55]. This is analogous to the escape to the 3rd dimension in liquid crystal director models [61]. This is not desirable if the underlying nematic liquid crystal is guaranteed to be uniaxial (recall Remark 1). Secondly, minimizing (6) with $\mathbf{Q}$ of the form (9) leads to a non-linear system with five coupled variables in 3d, so it is expensive to solve and possibly not robust [39,52,66,67].

These drawbacks motivate us to enforce the uniaxiallity constraint (3) directly in the Landau-de Gennes one-constant energy (6). The ensuing model has similarities with the Ericksen model (7), although it has the advantage of allowing minimizers to exhibit half-integer order defects. Our approach hinges on previous work on the Ericksen model [46–48], which exploits a hidden structure of (7). We next reveal such structure for the Landau-de Gennes model with uniaxial constraint and point out the corresponding counterpart for the Ericksen model when appropriate. Compared to directly minimizing (6) using (9), our algorithm finds a minimizer by *solving a sequence of linear systems of smaller dimension*. However, our approach is equivalent to directly minimizing the energy (6) for two-dimensional problems (see Remark 2).

Reference [14] offers a detailed numerical study of the effect of imposing the uniaxial constraint in the Landau-de Gennes model. There it is shown that in the presence of defects, the pointwise discrepancy between the minimizer from the standard model and the minimizer with uniaxial constraint becomes significant near such regions. Moreover, the uniaxial minimizers may present a significantly higher energy than minimizers of the standard Landau-de Gennes model under the same domain configurations.

3.1 The basic structure

We start with the main part (elastic energy) of the one-constant Ericksen model in (7), namely

$$E_{\mathrm{erk-m}}[s,\mathbf{n}] := \frac{1}{2}\int_\Omega \left(\kappa|\nabla s|^2\,dx + s^2|\nabla \mathbf{n}|^2\right)dx. \tag{10}$$

It is clear that a configuration $(s,\mathbf{n})$ with finite elastic energy implies $s \in H^1(\Omega)$ and that the weight s vanishing within the singular set $\mathbb{S}$ of (8) allows for director fields $\mathbf{n}$ with infinite Dirichlet energy and thus for the presence of defects. The hidden structure in (10) becomes apparent upon introducing the auxiliary variable $\mathbf{u} = s\mathbf{n}$ as

proposed first in [4,40]: since $|\mathbf{n}| = 1$ we get $\nabla\mathbf{n}\,\mathbf{n} = \mathbf{0}$ and the pointwise orthogonal decomposition $\nabla\mathbf{u} = \mathbf{n}\otimes\nabla s + s\nabla\mathbf{n}$. Consequently, (10) can be equivalently written

$$E_{\mathrm{erk-m}}[s,\mathbf{n}] = \widetilde{E}_{\mathrm{erk-m}}[s,\mathbf{u}] := \frac{1}{2}\int_\Omega \Big((\kappa-1)|\nabla s|^2 + |\nabla\mathbf{u}|^2\Big)dx, \tag{11}$$

to discover that $\mathbf{u} \in [H^1(\Omega)]^d$. Moreover, it is apparent from (11) that if $\kappa > 1$ the Ericksen energy $\widetilde{E}_{\mathrm{erk-m}}[s,\mathbf{u}]$ is convex with respect to $(s,\mathbf{u})$. The physically relevant case $0 < \kappa < 1$ in terms of defects is more difficult with regard to proving Γ-convergence, because convexity of $\widetilde{E}_{\mathrm{erk-m}}[s,\mathbf{u}]$ is no longer obvious unless we exploit the relation $|s| = |\mathbf{u}|$. This relation can only be enforced at the nodes of a finite element approximation of $(s,\mathbf{u})$, whence convexity as well as weak lower semi-continuity of $\widetilde{E}_{\mathrm{erk-m}}[s,\mathbf{u}]$ become problematic [46–48]; we will refer to this issue later in Lemma 6.

We now turn to the Landau-de Gennes model with uniaxial constraint (3). To this end, we introduce the line field $\boldsymbol{\Theta} = \mathbf{n}\otimes\mathbf{n}$, which will be treated as a control variable in minimizing (6) subject to (3). Since $\nabla\mathbf{Q}$ is a 3-tensor of the form $\nabla\mathbf{Q} = \nabla s\otimes\big(\boldsymbol{\Theta} - \frac{1}{d}\mathbf{I}\big) + s\nabla\boldsymbol{\Theta}$, we have

$$|\nabla\mathbf{Q}|^2 = |\nabla s|^2\left|\boldsymbol{\Theta} - \frac{1}{d}\mathbf{I}\right|^2 + s^2|\nabla\boldsymbol{\Theta}|^2 + 2s\left[\nabla s\otimes\left(\boldsymbol{\Theta} - \frac{1}{d}\mathbf{I}\right)\right]:\nabla\boldsymbol{\Theta}.$$

A direct calculation gives $\left|\boldsymbol{\Theta} - \frac{1}{d}\mathbf{I}\right|^2 = \frac{d-1}{d}$ and $\left[\nabla s\otimes\left(\boldsymbol{\Theta} - \frac{1}{d}\mathbf{I}\right)\right]:\nabla\boldsymbol{\Theta} = 0$ because $\nabla\boldsymbol{\Theta} : \boldsymbol{\Theta} = \nabla\boldsymbol{\Theta} : \mathbf{I} = \mathbf{0}$. Therefore, we obtain the first relation with the Ericksen model

$$|\nabla\mathbf{Q}|^2 = \frac{d-1}{d}|\nabla s|^2 + s^2|\nabla\boldsymbol{\Theta}|^2.$$

The second one comes from the equalities

$$s^2 = C_2\,\mathrm{tr}(\mathbf{Q}^2), \quad s^3 = C_3\,\mathrm{tr}(\mathbf{Q}^3), \quad s^4 = C_4\,(\mathrm{tr}(\mathbf{Q}^2))^2,$$

which are valid for suitable constants $C_2, C_3, C_4 > 0$. Consequently, the double-well potential $\phi_{\mathrm{LdG}}(\mathbf{Q})$ in (5) becomes a quartic function $\psi_{\mathrm{LdG}}(s) = \phi_{\mathrm{LdG}}(\mathbf{Q})$ of s that blows-up at the end points of the interval $[-\frac{1}{d-1}, 1]$ and forces s to remain within this physical range. If we let the main energy be

$$E_{\mathrm{uni-m}}[s,\boldsymbol{\Theta}] := E_{\mathrm{uni-s}}[s] + E_{\mathrm{uni-i}}[s,\boldsymbol{\Theta}],$$

where the orientation, interaction and bulk energies are given by

$$E_{\mathrm{uni-s}}[s] := \frac{d-1}{2d}\int_\Omega |\nabla s|^2, \quad E_{\mathrm{uni-i}}[s,\boldsymbol{\Theta}] := \frac{1}{2}\int_\Omega s^2|\nabla\boldsymbol{\Theta}|^2\,dx,$$
$$E_{\mathrm{LdG,bulk}}[s] := \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \psi_{\mathrm{LdG}}(s)\,dx,$$

then the Landau-de Gennes total energy $E_{\text{uni}-\text{t}}[s, \boldsymbol{\Theta}] = E_{\text{LdG,one}}[\mathbf{Q}]$ in (6) reads

$$E_{\text{uni}-\text{t}}[s, \boldsymbol{\Theta}] = E_{\text{uni}-\text{m}}[s, \boldsymbol{\Theta}] + E_{\text{LdG,bulk}}[s]. \tag{12}$$

We see that this energy has the *same form* as the Ericksen energy (7), except that $\boldsymbol{\Theta}$ replaces $\mathbf{n}$ and $\kappa = (d-1)/d < 1$. This motivates a change of variable analogous to the one in the Ericksen model: we set $\mathbf{U} := s\boldsymbol{\Theta}$ and note that $\nabla \mathbf{U} = \nabla s \otimes \boldsymbol{\Theta} + s\nabla\boldsymbol{\Theta}$ is a d-tensor with orthogonal components, whence

$$|\nabla \mathbf{U}|^2 = |\nabla s|^2 + s^2 |\nabla \boldsymbol{\Theta}|^2$$

and the main and total energies in terms of $(s, \mathbf{U})$ read

$$E_{\text{uni}-\text{m}}[s, \boldsymbol{\Theta}] = \widetilde{E}_{\text{uni}-\text{m}}[s, \mathbf{U}] := -\tfrac{1}{2d} \int_\Omega |\nabla s|^2 \, dx + \tfrac{1}{2} \int_\Omega |\nabla \mathbf{U}|^2 \, dx, \tag{13}$$

$$\widetilde{E}_{\text{uni}-\text{t}}[s, \mathbf{U}] := \widetilde{E}_{\text{uni}-\text{m}}[s, \mathbf{U}] + E_{\text{LdG,bulk}}[s]. \tag{14}$$

Similarly, we could set $\widetilde{s} := |s|$ and $\widetilde{\mathbf{U}} := \widetilde{s}\boldsymbol{\Theta}$ to arrive at $\widetilde{E}_{\text{uni}-\text{m}}[\widetilde{s}, \widetilde{\mathbf{U}}] = \widetilde{E}_{\text{uni}-\text{m}}[s, \mathbf{U}]$ because $|\nabla \widetilde{s}| = |\nabla s|$ a.e. in Ω. We are now able to reach similar conclusions as for the Ericksen model. If $E_{\text{LdG,one}}[\mathbf{Q}] < \infty$, then $(s, \mathbf{U}) \in H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$ but in general $\boldsymbol{\Theta} \notin [H^1(\Omega)]^{d\times d}$ because the presence of the weight s^2 in $E_{\text{uni}-\text{i}}[s, \boldsymbol{\Theta}]$ allows for blow-up of $\nabla\boldsymbol{\Theta}$ in the singular set $\mathbb{S}$ of (8). We intend to preserve this basic structure discretely. In fact, this will be crucial later in Sect. 5 to interpret $\nabla\boldsymbol{\Theta}$ in the Lebesgue L^2 sense and recover the orthogonality relation $|\nabla \mathbf{U}|^2 = |\nabla s|^2 + s^2|\nabla\boldsymbol{\Theta}|^2$ a.e. in $\Omega \setminus \mathbb{S}$, as well as to derive Γ convergence.

In order to define the *admissible class* of functions, we begin with the set of *line fields*

$$\mathbb{L}^{d-1} := \{\mathbf{A} \in \mathbb{R}^{d\times d} : \text{ there exists } \mathbf{n} \in \mathbb{S}^{d-1}, \ \mathbf{A} = \mathbf{n} \otimes \mathbf{n}\}. \tag{15}$$

We say that a triple $(s, \boldsymbol{\Theta}, \mathbf{U})$ satisfies the *structural condition* provided

$$-\frac{1}{d-1} \le s \le 1, \qquad \mathbf{U} = s\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta} \in \mathbb{L}^{d-1} \qquad \text{a.e. } \Omega. \tag{16}$$

We next define the admissible class of functions to be

$$\mathcal{A}_{\text{uni}} := \big\{(s, \boldsymbol{\Theta}, \mathbf{U}) \in H^1(\Omega) \times [L^\infty(\Omega)]^{d\times d} \times [H^1(\Omega)]^{d\times d} : (s, \boldsymbol{\Theta}, \mathbf{U}) \text{ satisfies (16)}\big\}.$$

To enforce boundary conditions, let $(\Gamma_s, \Gamma_{\boldsymbol{\Theta}}, \Gamma_{\mathbf{U}})$ with $\Gamma_{\boldsymbol{\Theta}} = \Gamma_{\mathbf{U}}$ be open subsets of $\partial\Omega$ where we impose Dirichlet conditions. Given functions $(g, \mathbf{M}, \mathbf{R}) \in W^1_\infty(\mathbb{R}^d) \times [L^\infty(\mathbb{R}^d)]^{d\times d} \times [W^1_\infty(\mathbb{R}^d)]^{d\times d}$ that satisfy the structural condition (16) in a neighborhood of $\partial\Omega$, we define the restricted admissible class

$$\mathcal{A}_{\text{uni}}(g, \mathbf{R}) := \big\{(s, \boldsymbol{\Theta}, \mathbf{U}) \in \mathcal{A}_{\text{uni}} : \ s|_{\Gamma_s} = g, \quad \mathbf{U}|_{\Gamma_{\mathbf{U}}} = \mathbf{R}\big\}.$$

Moreover, we assume that for some $\delta_0 > 0$

$$-\frac{1}{d-1} + \delta_0 \le g \le 1 - \delta_0 \quad \text{in } \Omega, \tag{17}$$

and

$$g \ge \delta_0 \ \text{ on } \partial\Omega, \tag{18}$$

so that the function $\mathbf{M}$ is of class W^1_∞ in a neighborhood of $\Gamma_{\boldsymbol{\Theta}}$ and satisfies $\mathbf{M} = g^{-1}\mathbf{R} \in \mathbb{L}^{d-1}$ on $\Gamma_{\boldsymbol{\Theta}}$.

Finally, we assume that the coefficients A, B, C in (5) are such that

$$\begin{aligned} &\psi_{\mathrm{LdG}}(s) \ge \psi_{\mathrm{LdG}}(1-\delta_0) \ \text{ for } s \ge 1-\delta_0, \\ &\psi_{\mathrm{LdG}}(s) \ge \psi_{\mathrm{LdG}}\left(-\frac{1}{d-1} + \delta_0\right) \ \text{ for } s \le -\frac{1}{d-1} + \delta_0. \end{aligned} \tag{19}$$

This will lead to confinement of s with the interval $[-\frac{1}{d-1} + \delta_0, 1 - \delta_0]$.

4 Discretization

Let $\mathcal{T}_h = \{T\}$ be a conforming shape-regular and quasi-uniform triangulation of Ω made of simplices. Let $\mathcal{N}_h = \{x_i\}_{i=1}^N$ be the set of nodes (vertices) x_i of $\mathcal{T}_h$ and N be its cardinality. Let ϕ_i be the standard "hat" basis function associated with the node $x_i \in \mathcal{N}_h$. We indicate with $\omega_i = \text{supp } \phi_i$ the patch of a node x_i (i.e. the "star" of elements in $\mathcal{T}_h$ that contain the vertex x_i). For simplicity we assume that $\Omega = \Omega_h$, so that there is no geometric error caused by domain approximation. We further assume that $\mathcal{T}_h$ is *weakly acute*, namely

$$k_{ij} := -\int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, dx \ge 0 \quad \text{for all } i \ne j. \tag{20}$$

Condition (20) ensures the validity of the discrete maximum principle. However, (20) imposes a severe geometric restriction on $\mathcal{T}_h$ [18,57], especially in three dimensions. Circumventing (20) is an open problem.

We consider three continuous piecewise linear Lagrange finite element spaces on Ω:

$$\begin{aligned} \mathbb{S}_h &:= \{s_h \in H^1(\Omega) : s_h|_T \text{ is affine for all } T \in \mathcal{T}_h\}, \\ \mathbb{U}_h &:= \{\mathbf{U}_h \in [H^1(\Omega)]^{d\times d} : \text{ each entry of } \mathbf{U}_h|_T \text{ is afffine for all } T \in \mathcal{T}_h\}, \\ \mathbb{T}_h &:= \{\boldsymbol{\Theta}_h \in \mathbb{U}_h : \boldsymbol{\Theta}_h(x_i) \in \mathbb{L}^{d-1}, \text{ for all } x_i \in \mathcal{N}_h\}, \end{aligned}$$

where $\mathbb{T}_h$ imposes both the rank-one and unit-norm constraints only at the vertices of the mesh $\mathcal{T}_h$. We say that the discrete triple $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathbb{S}_h \times \mathbb{T}_h \times \mathbb{U}_h$ satisfies the *discrete structural condition* if

$$\mathbf{U}_h = I_h(s_h\boldsymbol{\Theta}_h), \qquad -\frac{1}{d-1} \leq s_h \leq 1, \tag{21}$$

where I_h stands for the Lagrange interpolation operator. All such triples make the *discrete admissible set* $\mathcal{A}^h_{\text{uni}}$. We let $g_h := I_h g$, $\mathbf{R}_h := I_h\mathbf{R}$, and $\mathbf{M}_h := I_h\mathbf{M}$ be the discrete Dirichlet data, and incorporate Dirichlet boundary conditions within the discrete spaces:

$$\begin{aligned}
\mathbb{S}_h(g_h) &:= \{s_h \in \mathbb{S}_h : s_h|_{\Gamma_s} = g_h\},\\
\mathbb{U}_h(\mathbf{R}_h) &:= \{\mathbf{U}_h \in \mathbb{U}_h : \mathbf{U}_h|_{\Gamma_{\mathbf{U}}} = \mathbf{R}_h\},\\
\mathbb{T}_h(\mathbf{M}_h) &:= \{\boldsymbol{\Theta}_h \in \mathbb{T}_h : \boldsymbol{\Theta}_h|_{\Gamma_{\boldsymbol{\Theta}}} = \mathbf{M}_h\}.
\end{aligned}$$

In view of (18), the following compatibility condition must hold: $\mathbf{M}_h = I_h[g_h^{-1}\mathbf{R}_h]$ on $\Gamma_{\boldsymbol{\Theta}}$. This leads to the following discrete admissible class with boundary conditions:

$$\mathcal{A}^h_{\text{uni}}(g_h, \mathbf{R}_h) := \big\{(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathbb{S}_h(g_h) \times \mathbb{T}_h(\mathbf{M}_h) \times \mathbb{U}_h(\mathbf{R}_h)) : \quad (s_h, \mathbf{U}_h, \boldsymbol{\Theta}_n) \text{ satisfies (21)}\big\}.$$

We are now ready to introduce the discrete version of $E_{\text{uni}-\text{m}}[s, \boldsymbol{\Theta}]$ which mimics that of the Ericksen model [46–48]. First note that $\sum_{j=1}^N k_{ij} = 0$ for all $x_i \in \mathcal{N}_h$, and for $s_h = \sum_{i=1}^N s_h(x_i)\phi_i \in \mathbb{S}_h$ we have

$$\int_\Omega |\nabla s_h|^2 dx = -\sum_{i=1}^N k_{ii} s_h(x_i)^2 - \sum_{i,j=1, i\neq j}^N k_{ij} s_h(x_i) s_h(x_j).$$

Using $k_{ii} = -\sum_{j\neq i} k_{ij}$ and the symmetry $k_{ij} = k_{ji}$, we thus obtain

$$\int_\Omega |\nabla s_h|^2 dx = \frac{1}{2}\sum_{i,j=1}^N k_{ij}\big(\delta_{ij} s_h\big)^2, \tag{22}$$

where we have introduced the notation

$$\delta_{ij} s_h := s_h(x_i) - s_h(x_j), \quad \delta_{ij}\boldsymbol{\Theta}_h := \boldsymbol{\Theta}_h(x_i) - \boldsymbol{\Theta}_h(x_j).$$

We next define the main part of the discrete Landau-de Gennes energy to be

$$\begin{aligned}
E^h_{\text{uni}-\text{m}}[s_h, \boldsymbol{\Theta}_h] := \frac{d-1}{4d}&\sum_{i,j=1}^N k_{ij}\left(\delta_{ij} s_h\right)^2\\
&+ \frac{1}{4}\sum_{i,j=1}^N k_{ij}\left(\frac{s_h(x_i)^2 + s_h(x_j)^2}{2}\right)|\delta_{ij}\boldsymbol{\Theta}_h|^2.
\end{aligned} \tag{23}$$

We point out that the first term corresponds to

$$E^h_{\mathrm{uni-s}}[s_h] = \frac{d-1}{2d}\int_\Omega |\nabla s_h|^2 dx = \frac{d-1}{4d}\sum_{i,j=1}^{N} k_{ij}\left(\delta_{ij}s_h\right)^2,$$

while the second term is a first order nonstandard approximation of $E_{\mathrm{uni-i}}[s,\boldsymbol{\Theta}] = \frac{1}{2}\int_\Omega s^2|\nabla\boldsymbol{\Theta}|^2 dx$,

$$E^h_{\mathrm{uni-i}}[s_h,\boldsymbol{\Theta}_h] := \frac{1}{4}\sum_{i,j=1}^{N} k_{ij}\left(\frac{s_h(x_i)^2+s_h(x_j)^2}{2}\right)|\delta_{ij}\boldsymbol{\Theta}_h|^2 \tag{24}$$

introduced in [47]. As we will see below, a key feature of this discretization is that it makes it possible to handle degenerate parameters s_h without regularization. This is due to Lemma 1, which deals with discrete versions of $\widetilde{E}_{\mathrm{uni-m}}[s,\mathbf{U}]$ defined in (13) involving the auxiliary variable $\mathbf{U}_h$:

$$\widetilde{E}^h_{\mathrm{uni-m}}[s_h,\mathbf{U}_h] := -\frac{1}{2d}\int_\Omega |\nabla s_h|^2 dx + \frac{1}{2}\int_\Omega |\nabla\mathbf{U}_h|^2 dx. \tag{25}$$

We finally discretize the nonlinear bulk energy in the usual manner

$$E^h_{\mathrm{LdG,bulk}}[s_h] := \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \psi_{\mathrm{LdG}}(s_h)dx.$$

With the notation introduced above, the formulation of the discrete problem is as follows: find $(s_h,\boldsymbol{\Theta}_h) \in \mathbb{S}_h(g_h)\times\mathbb{T}_h(\mathbf{M}_h)$ such that the following discrete total energy is minimized:

$$E^h_{\mathrm{uni-t}}[s_h,\boldsymbol{\Theta}_h] := E^h_{\mathrm{uni-m}}[s_h,\boldsymbol{\Theta}_h] + E^h_{\mathrm{LdG,bulk}}[s_h]. \tag{26}$$

Because the discrete spaces consist of piecewise linear functions, the structural condition $\mathbf{U}_h = s_h\boldsymbol{\Theta}_h$ is only satisfied at the mesh nodes (cf. (21)). Therefore, there is a variational crime that we need to account for. To this end, we now derive energy inequalities similar to [47, Lemma 2.2]. Although the arguments are the same, we present the proof for completeness. For our analysis, we introduce the functions

$$\widetilde{s}_h = I_h(|s_h|), \qquad \widetilde{\mathbf{U}}_h = I_h(|s_h|\boldsymbol{\Theta}_h), \tag{27}$$

and remark that $(\widetilde{s}_h,\boldsymbol{\Theta}_h,\widetilde{\mathbf{U}}_h)$ satisfies (21).

Lemma 1 (energy inequality) *Let the mesh $\mathcal{T}_h$ satisfy (20). Then, for all $(s_h,\boldsymbol{\Theta}_h,\mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h,\mathbf{R}_h)$, the main part of the discrete Landau-de Gennes energy satisfies*

$$E^h_{\mathrm{uni-m}}[s_h,\boldsymbol{\Theta}_h] - \widetilde{E}^h_{\mathrm{uni-m}}[s_h,\mathbf{U}_h] = \mathcal{E}_h, \tag{28}$$

as well as

$$E^h_{\text{uni}-\text{m}}[s_h, \boldsymbol{\Theta}_h] - \widetilde{E}^h_{\text{uni}-\text{m}}[\widetilde{s}_h, \widetilde{\mathbf{U}}_h] \geq \widetilde{\mathcal{E}}_h, \tag{29}$$

where $\widetilde{E}^h_{\text{uni}-\text{m}}[s_h, \mathbf{U}_h]$ *is defined in* (25) *and*

$$\begin{aligned} \mathcal{E}_h &:= \frac{1}{8} \sum_{i,j=1}^{N} k_{ij} \big(\delta_{ij} s_h\big)^2 \big|\delta_{ij} \boldsymbol{\Theta}_h\big|^2 \geq 0, \\ \widetilde{\mathcal{E}}_h &:= \frac{1}{8} \sum_{i,j=1}^{N} k_{ij} \big(\delta_{ij} \widetilde{s}_h\big)^2 \big|\delta_{ij} \boldsymbol{\Theta}_h\big|^2 \geq 0. \end{aligned} \tag{30}$$

Proof Expanding

$$s_h(x_i)\boldsymbol{\Theta}_h(x_i) - s_h(x_j)\boldsymbol{\Theta}_h(x_j) = \frac{s_h(x_i) + s_h(x_j)}{2} \delta_{ij} \boldsymbol{\Theta}_h + \frac{\boldsymbol{\Theta}_h(x_i) + \boldsymbol{\Theta}_h(x_j)}{2} \delta_{ij} s_h$$

and using the orthogonality relation $(\delta_{ij} \boldsymbol{\Theta}_h) : \big(\boldsymbol{\Theta}_h(x_i) + \boldsymbol{\Theta}_h(x_j)\big) = 0$, we can write

$$\frac{1}{2} \int_\Omega |\nabla \mathbf{U}_h|^2 = \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} \left(\frac{s_h(x_i) + s_h(x_j)}{2} \right)^2 |\delta_{ij} \boldsymbol{\Theta}_h|^2 + \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} (\delta_{ij} s_h)^2 \left| \frac{\boldsymbol{\Theta}_h(x_i) + \boldsymbol{\Theta}_h(x_j)}{2} \right|^2.$$

Next, we utilize the identities $\big(s_h(x_i) + s_h(x_j)\big)^2 = 2\big(s_h(x_i)^2 + s_h(x_j)^2\big) - \big(s_h(x_i) - s_h(x_j)\big)^2$ and $\big|\boldsymbol{\Theta}_h(x_i) + \boldsymbol{\Theta}_h(x_j)\big|^2 = 4 - |\delta_{ij} \boldsymbol{\Theta}_h|^2$, to obtain

$$\frac{1}{2} \int_\Omega |\nabla \mathbf{U}_h|^2 dx = \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} \left(\frac{s_h(x_i)^2 + s_h(x_j)^2}{2} \right) |\delta_{ij} \boldsymbol{\Theta}_h|^2 + \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} (\delta_{ij} s_h)^2 - \mathcal{E}_h.$$

Identity (28) follows immediately.

On the other hand, repeating the argument above with $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h)$ instead of $(s_h, \mathbf{U}_h)$ gives

$$\frac{1}{2} \int_\Omega |\nabla \widetilde{\mathbf{U}}_h|^2 dx = \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} \left(\frac{\widetilde{s}_h(x_i)^2 + \widetilde{s}_h(x_j)^2}{2} \right) |\delta_{ij} \boldsymbol{\Theta}_h|^2 + \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} (\delta_{ij} \widetilde{s}_h)^2 - \widetilde{\mathcal{E}}_h.$$

This yields

$$\widetilde{E}^h_{\text{uni}-\text{m}}[\widetilde{s}_h, \widetilde{\mathbf{U}}_h] = E^h_{\text{uni}-\text{m}}[\widetilde{s}_h, \boldsymbol{\Theta}_h] - \widetilde{\mathcal{E}}_h.$$

The properties $\mathcal{E}_h \geq 0$ and $\widetilde{\mathcal{E}}_h \geq 0$ are a consequence of the mesh acuteness assumption (20). Moreover, since $|\delta_{ij} \widetilde{s}_h| \leq |\delta_{ij} s_h|$ and $\widetilde{s}_h(x_i)^2 = s_h(x_i)^2$ for all $i, j = 1, \ldots, N$, we have

$$\|\nabla \widetilde{s}_h\|_{L^2(\Omega)} = \frac{1}{2}\sum_{i,j=1}^{N} k_{ij}(\delta_{ij}\widetilde{s}_h)^2 \le \frac{1}{2}\sum_{i,j=1}^{N} k_{ij}(\delta_{ij}s_h)^2 = \|\nabla s_h\|_{L^2(\Omega)}.$$

Therefore, $E^h_{\mathrm{uni-m}}[\widetilde{s}_h, \boldsymbol{\Theta}_h] \le E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h]$ and inequality (29) follows. □

5 Γ-convergence of the discrete energies

This section shows that the discrete problems (26) Γ-converge to the continuous problem (12). We set the continuous and discrete spaces

$$\mathbb{X} := L^2(\Omega) \times [L^2(\Omega)]^{d\times d} \times [L^2(\Omega)]^{d\times d}, \qquad \mathbb{X}_h := \mathbb{S}_h \times \mathbb{T}_h \times \mathbb{U}_h,$$

and define $E_{\mathrm{uni-t}}[s, \boldsymbol{\Theta}]$ as in (12) if $(s, \boldsymbol{\Theta}) \in \mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$ and $E_{\mathrm{uni-t}}[s, \boldsymbol{\Theta}] = \infty$ if $(s, \boldsymbol{\Theta}) \in \mathbb{X} \setminus \mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$. In a similar fashion, we define $E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h]$ as in (26) if $(s_h, \boldsymbol{\Theta}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$ and $E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h] = \infty$ if $(s_h, \boldsymbol{\Theta}_h) \in \mathbb{X}_h \setminus \mathcal{A}_{\mathrm{uni}}(g_h, \mathbf{R}_h)$.

5.1 Lim-sup property: existence of a recovery sequence

Our goal is to show the following property: given $(s, \boldsymbol{\Theta}, \mathbf{U}) \in \mathbb{X}$, there exists a sequence $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{M}_h)$ such that

$$\|(s, \mathbf{U}) - (s_h, \mathbf{U}_h)\|_{H^1(\Omega)} \to 0, \quad \|\boldsymbol{\Theta} - \boldsymbol{\Theta}_h\|_{L^2(\Omega\setminus\mathbb{S})} \to 0, \tag{31}$$

as $h \to 0$ and

$$\limsup_{h\to 0} E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h] \le E_{\mathrm{uni-t}}[s, \boldsymbol{\Theta}], \tag{32}$$

where $E_{\mathrm{uni-t}}[s, \boldsymbol{\Theta}]$ is defined in (12).

Truncation Naturally, the interesting case to consider is when $(s, \boldsymbol{\Theta}) \in \mathcal{A}_{\mathrm{uni}}(g, \mathbf{M})$; otherwise the property above is trivially true. As shown in [47, Lemma 3.1], hypotheses (17) and (19) make it possible to assume that the degree of orientation s is sufficiently far from the boundary of the physically meaningful range $[-\frac{1}{d-1}, 1]$. We state this precisely next.

Lemma 2 (truncation) *Given* $(s, \boldsymbol{\Theta}, \mathbf{U}) \in \mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$, *let* $(\hat{s}, \widehat{\mathbf{U}})$ *be the truncations*

$$\hat{s}(x) := \min\left\{1 - \delta_0, \max\left\{-\frac{1}{d-1} + \delta_0,\ s(x)\right\}\right\}, \quad \widehat{\mathbf{U}} := \hat{s}\boldsymbol{\Theta}.$$

Then, $(\hat{s}, \boldsymbol{\Theta}, \widehat{\mathbf{U}}) \in \mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$ *and*

$$E_{\mathrm{uni-m}}[\hat{s}, \boldsymbol{\Theta}] \le E_{\mathrm{uni-m}}[s, \boldsymbol{\Theta}], \qquad E_{\mathrm{LdG,bulk}}[\hat{s}, \boldsymbol{\Theta}] \le E_{\mathrm{LdG,bulk}}[s, \boldsymbol{\Theta}].$$

Moreover, given $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\text{uni}}(g_h, \mathbf{R}_h)$ *and the truncations* $(I_h \hat{s}_h, I_h \widehat{\mathbf{U}}_h)$*, then the same assertion holds for the discrete energies.*

Proof We first observe that (17) implies $(\hat{s}, \boldsymbol{\Theta}, \widehat{\mathbf{U}}) \in \mathcal{A}_{\text{uni}}(g, \mathbf{R})$ whereas (19) yields $\psi_{\text{LdG}}(\hat{s}, \boldsymbol{\Theta}) \leq \psi_{\text{LdG}}(\hat{s}, \boldsymbol{\Theta})$. Moreover, making use of $|\hat{s}| \leq |s|$ and $|\nabla \hat{s}| \leq |\nabla s|$ a.e. in Ω, the inequality $E_{\text{uni}-\text{m}}[\hat{s}, \boldsymbol{\Theta}] \leq E_{\text{uni}-\text{m}}[s, \boldsymbol{\Theta}]$ follows immediately. □

Remark 6 (range of s) For problems in 3d, the admissibility condition $s \in [-1/2, 1]$ is asymmetric with respect to the origin. Since part of our argument below is based on regularizing $|s|$ and afterwards recovering its sign, we need to account for such an asymmetry. A simple way to do so is to consider

$$\check{s} = s_+ - 2s_-. \tag{33}$$

Clearly, the first condition in (16) is equivalent to

$$-1 \leq \check{s} \leq 1.$$

In the next result, we consider the regularization using this modified degree of orientation; for simplicity of notation, we drop the "check" in s.

Rank-one constraint Our regularization method entails smoothing by convolution. This breaks the uniaxial constraint (3), that needs to be rebuilt into the smoothed tensor field; hence, we extract the leading eigenspace. We thus need to account for the dependence of eigenvalues with respect to matrix perturbations. Let $\text{Sym}(d)$ denote the set of symmetric $d \times d$ matrices. Given $\mathbf{A} \in \text{Sym}(d)$, let $\lambda_1 \geq \ldots \geq \lambda_d$ be the eigenvalues of $\mathbf{A}$ including multiplicities and $\lambda_{m(1)} > \cdots > \lambda_{m(n)}$ be the $1 \leq n \leq d$ distinct eigenvalues. Let $\{\mathbf{P}_k\}_{k=1}^n$ be the orthogonal projections onto the eigenspaces associated with $\{\lambda_{m(k)}\}_{k=1}^n$ and let $r(k) \geq 1$ be the rank of $\mathbf{P}_k$; hence $r(k)$ is the multiplicity of $\lambda_{m(k)}$ for $1 \leq k \leq n$. The spectral decomposition of $\mathbf{A}$ reads $\mathbf{A} = \sum_{k=1}^n \lambda_{m(k)} \mathbf{P}_k$ We now consider the set $S^{1,0}(d)$ of non-negative symmetric tensors of rank at most one,

$$S^{1,0}(d) = \left\{ \mathbf{A} \in \text{Sym}(d) \colon \mathbf{A} = \mathbf{u} \otimes \mathbf{u} \text{ for some } \mathbf{u} \in \mathbb{R}^d \right\},$$

and follow [7] to construct the projection operator $\boldsymbol{\Pi} \colon \text{Sym}(d) \to S^{1,0}(d)$ defined by

$$\boldsymbol{\Pi}(\mathbf{A}) = (\lambda_1 - \lambda_2)\mathbf{P}_1. \tag{34}$$

The map $\boldsymbol{\Pi}$ is Lipschitz continuous. This is proven in [7, Lemma 3.4] with an explicit Lipschitz constant $3 + 2^{1+\frac{1}{p}}$ (in the ℓ_p-norm). We give an elementary proof below which relies on the following basic result.

Lemma 3 (C^1 property of $\boldsymbol{\Pi}$) *The map* $Sym(d) \to \mathbb{R}^d$*, given by* $\mathbf{A} \mapsto (\lambda_1(\mathbf{A}), \ldots, \lambda_d(\mathbf{A}))$*, is continuous. Moreover, in the set of symmetric matrices whose first eigenspace has dimension* 1

$$Sym^1(d) := \{\mathbf{A} \in Sym(d) \colon \lambda_1(\mathbf{A}) > \lambda_2(\mathbf{A})\},$$

or equivalently the rank of $\mathbf{P}_1$ *is* 1*, the map* $\mathbf{\Pi}$ *is of class* C^1.

Proof The eigenvalues $\{\lambda_i(\mathbf{A})\}_{i=1}^d$ are the roots of the characteristic polynomial of $\mathbf{A}$ and depend continuously on the coefficients and so on the entries of $\mathbf{A}$. To show the C^1 property around $\mathbf{A}_0 \in \mathrm{Sym}^1(d)$, let $\mathbf{A} \in \mathrm{Sym}(d)$ and $\mathbf{x}_1 = \mathbf{x}_1(\mathbf{A})$ be a normalized eigenvector corresponding to the first eigenvalue $\lambda_1 = \lambda_1(\mathbf{A})$. The equation that defines $(\mathbf{x}_1, \lambda_1)$ and its derivative with respect to $(\mathbf{x}_1, \lambda_1)$ read

$$\mathbf{F}(\mathbf{x}_1, \lambda_1, \mathbf{A}) = \begin{bmatrix} \mathbf{A}\mathbf{x}_1 - \lambda_1 \mathbf{x}_1 \\ \|\mathbf{x}_1\|_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad D_{\mathbf{x}_1,\lambda_1}\mathbf{F}(\mathbf{x}_1, \lambda_1, \mathbf{A}) = \begin{bmatrix} \mathbf{A} - \lambda_1 \mathbf{I} & -\mathbf{x}_1 \\ 2\mathbf{x}_1^T & 0 \end{bmatrix}.$$

Since $\lambda_1(\mathbf{A}_0)$ is single, the matrix $D_{\mathbf{x}_1,\lambda_1}\mathbf{F}(\mathbf{x}_1(\mathbf{A}_0), \lambda_1(\mathbf{A}_0), \mathbf{A}_0)$ is invertible for otherwise if $(\mathbf{y}, \alpha)^T \in \mathbb{R}^{d+1}$ is in the kernel it must necessarily vanish. Therefore, the Implicit Function Theorem (IFT) applies thereby giving the existence of $(\mathbf{x}_1(\mathbf{A}), \lambda_1(\mathbf{A}))$ and its C^1 dependence on $\mathbf{A}$; we refer to [25, Chapter 11.1, Theorem 2] for a different argument. To prove that $\lambda_2(\mathbf{A})$ is also C^1 we proceed similarly but note that this eigenvalue might have multiplicity $r(2) > 1$. We thus form the equation for $\mathbf{P}_2 = \mathbf{P}_2(\mathbf{A}) \in \mathbb{R}^{d\times d}$ being a matrix with rank $r(2)$ and $\lambda_2 = \lambda_2(\mathbf{A})$

$$\mathbf{F}(\mathbf{P}_2, \lambda_2, \mathbf{A}) = \begin{bmatrix} \mathbf{A}\mathbf{P}_2 - \lambda_2 \mathbf{P}_2 \\ \|\mathbf{P}_2\|_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad D_{\mathbf{P}_2,\lambda_2}\mathbf{F}(\mathbf{P}_2, \lambda_2, \mathbf{A}) = \begin{bmatrix} \mathbf{A} - \lambda_2 \mathbf{I} & -\mathbf{P}_2 \\ 2\mathbf{P}_2^T & 0 \end{bmatrix}.$$

and show that the kernel of this matrix is trivial. The IFT gives the asserted C^1 continuity of $\lambda_2(\mathbf{A})$. □

Lemma 4 (Lipschitz property of $\mathbf{\Pi}$) *The map* $\mathbf{\Pi}\colon Sym(d) \to S^{1,0}(d)$ *is uniformly Lipschitz continuous and is invariant on* $S^{1,0}(d)$*, i.e.* $\mathbf{\Pi}(\mathbf{A}) = \mathbf{A}$ *for all* $\mathbf{A} \in S^{1,0}(d)$.

Proof The invariance of $\mathbf{\Pi}$ over $S^{1,0}(d)$ is clear from its definition. Given $\mathbf{A}, \mathbf{B} = \mathbf{A} + \delta\mathbf{A} \in \mathrm{Sym}(d)$, write

$$\begin{aligned}\mathbf{\Pi}(\mathbf{B}) - \mathbf{\Pi}(\mathbf{A}) = &\big[\big(\lambda_1(\mathbf{B}) - \lambda_1(\mathbf{A})\big) - \big(\lambda_2(\mathbf{B}) - \lambda_2(\mathbf{A})\big)\big]\mathbf{P}_1(\mathbf{B}) \\ &+ \big(\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})\big)\big(\mathbf{P}_1(\mathbf{B}) - \mathbf{P}_1(\mathbf{A})\big).\end{aligned}$$

We examine the two terms on the right hand side separately. We split the proof into three steps.

Step 1: Lipschitz property of the first term. We resort to Weyl's inequality for eigenvalues of symmetric matrices [13, Section III.2]

$$\big|\lambda_k(\mathbf{B}) - \lambda_k(\mathbf{A})\big| \le \|\mathbf{B} - \mathbf{A}\|_2 \quad \forall 1 \le k \le d.$$

Since $\|\mathbf{P}_1(\mathbf{B})\|_2 = 1$ because $\mathbf{P}_1(\mathbf{B})$ is an orthogonal projection, this proves the Lipschitz property for the first term with constant 2. If $\lambda_1(\mathbf{A})$ is a multiple eigenvalue, then $\lambda_1(\mathbf{A}) = \lambda_2(\mathbf{A})$, the second term vanishes, and the proof is over. We thus assume that $\lambda_1(\mathbf{A})$ is simple from now on.

Step 2: Bound on $\|D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A}; \delta\mathbf{A})\|_2$. In view of Lemma 3 ($C^1$ property of $\mathbf{\Pi}$), we differentiate the equation $\mathbf{F}(\mathbf{x}_1(\mathbf{A}), \lambda_1(\mathbf{A}), \mathbf{A}) = [0, 1]^T$ with respect to $\mathbf{A}$ in the

direction $\delta\mathbf{A}$ to obtain $D_{\mathbf{x}_1}\mathbf{F}\, D_{\mathbf{A}}\mathbf{x}_1 + D_{\lambda_1}\mathbf{F}\, D_{\mathbf{A}}\lambda_1 + D_{\mathbf{A}}\mathbf{F} : \delta\mathbf{A} = 0$ where $D_{\mathbf{A}}\mathbf{x}_1 = D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A})$ and $D_{\mathbf{A}}\lambda_1 = D_{\mathbf{A}}\lambda_1(\mathbf{A};\delta\mathbf{A})$. Making use of Lemma 3 again, we thus deduce the equation in $\mathbb{R}^{d+1}$

$$\begin{bmatrix} \mathbf{A} - \lambda_1\mathbf{I} \\ 2\mathbf{x}_1^T \end{bmatrix} D_{\mathbf{A}}\mathbf{x}_1 + \begin{bmatrix} -\mathbf{x}_1 \\ 0 \end{bmatrix} D_{\mathbf{A}}\lambda_1 = -\begin{bmatrix} \delta\mathbf{A}\,\mathbf{x}_1 \\ 0 \end{bmatrix}.$$

The last row yields $\mathbf{x}_1^T\, D_{\mathbf{A}}\mathbf{x}_1 = 0$, whence $D_{\mathbf{A}}\mathbf{x}_1$ is perpendicular to $\mathbf{x}_1$ and $D_{\mathbf{A}}\mathbf{x}_1 = \sum_{k=2}^{d} \alpha_k \mathbf{x}_k$ can be expressed in terms of the orthonormal eigenvectors $\{\mathbf{x}_k\}_{k=1}^d$ of $\mathbf{A}$ without component along $\mathbf{x}_1$. Moreover, if $\delta\mathbf{A}\,\mathbf{x}_1 = \sum_{k=1}^{d} \beta_k \mathbf{x}_k$, then the first d rows of the preceding equation give

$$\sum_{k=2}^{d} \big[\alpha_k(\lambda_k - \lambda_1) + \beta_k\big]\mathbf{x}_k = \big(D_{\mathbf{A}}\lambda_1 - \beta_1\big)\mathbf{x}_1.$$

This obviously implies $D_{\mathbf{A}}\lambda_1 = \beta_1$ and

$$\alpha_k = \frac{\beta_k}{\lambda_1 - \lambda_k} \quad \forall\, 2 \le k \le d.$$

Let $\boldsymbol{\alpha} = (\alpha_k)_{k=1}^d$ with $\alpha_1 = 0$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^d$. Since $\|\boldsymbol{\beta}\|_2 \le \|\delta\mathbf{A}\|_2$, we see that

$$\|D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A})\|_2 = \|\boldsymbol{\alpha}\|_2 \le \frac{\|\delta\mathbf{A}\|_2}{\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})},$$

because $0 < \lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A}) \le \lambda_1(\mathbf{A}) - \lambda_k(\mathbf{A})$ for all $2 \le k \le d$.

Step 3: Lipschitz property of the second term. Exploiting that $\mathbf{P}_1(\mathbf{A}) = \mathbf{x}_1(\mathbf{A}) \otimes \mathbf{x}_1(\mathbf{A})$, we readily get

$$D_{\mathbf{A}}\mathbf{P}_1(\mathbf{A};\delta\mathbf{A}) = D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A}) \otimes \mathbf{x}_1(\mathbf{A}) + \mathbf{x}_1(\mathbf{A}) \otimes D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A}).$$

Since $\mathbf{x}_1(\mathbf{A})$ and $D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A})$ are perpendicular, we infer that

$$\|D_{\mathbf{A}}\mathbf{P}_1(\mathbf{A};\delta\mathbf{A})\|_2 \le \|D_{\mathbf{A}}\mathbf{x}_1(\mathbf{A};\delta\mathbf{A})\|_2. \tag{35}$$

Indeed, if $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ are orthonormal and $\mathbf{w} \in \mathbb{R}^d$, then $(\mathbf{u}\otimes\mathbf{v} + \mathbf{v}\otimes\mathbf{u})\mathbf{w} = (\mathbf{v}\cdot\mathbf{w})\mathbf{u} + (\mathbf{u}\cdot\mathbf{w})\mathbf{v}$, and thus, by Bessel's inequality,

$$\|(\mathbf{u}\otimes\mathbf{v} + \mathbf{v}\otimes\mathbf{u})\mathbf{w}\|_2 \le \|\mathbf{w}\|_2;$$

estimate (35) then follows by scaling. Combining this with Step 2 gives

$$\begin{aligned} &\big|\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})\big|\, \|\mathbf{P}_1(\mathbf{A}+\delta\mathbf{A}) - \mathbf{P}_1(\mathbf{A})\|_2 \\ &\qquad \le \|\delta\mathbf{A}\|_2 + \big|\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})\big|\, o(\|\delta\mathbf{A}\|_2) = \big(1 + o(1)\big)\|\delta\mathbf{A}\|_2, \end{aligned}$$

which shows that the desired Lipschitz constant is 1. Altogether the uniform Lipschitz constant of $\mathbf{\Pi}$ (with respect to the ℓ_2-norm) is 3. This concludes the proof. □

Regularization We now have all the tools we need to prove that Lipschitz continuous functions are dense in the Landau - de Gennes restricted admissible class $\mathcal{A}_{\text{uni}}(g, \mathbf{R})$.

Proposition 7 (regularization) *Let* (18), (19) *and* (33) *hold. Given* $\varepsilon > 0$ *and* $(s, \mathbf{\Theta}, \mathbf{U}) \in \mathcal{A}_{\text{uni}}(g, \mathbf{R})$ *with*

$$-1 + \delta_0 \leq s \leq 1 - \delta_0 \quad \text{a.e. } \Omega \tag{36}$$

there exists a sequence $(s_\varepsilon, \mathbf{\Theta}_\varepsilon, \mathbf{U}_\varepsilon) \in \mathcal{A}_{\text{uni}}(g, \mathbf{R})$ *such that* $(s_\varepsilon, \mathbf{U}_\varepsilon) \in W^{1,\infty}(\Omega) \times [W^{1,\infty}(\Omega)]^{d\times d}$, *and*

$$\begin{gathered}\|(s, \mathbf{U}) - (s_\varepsilon, \mathbf{U}_\varepsilon)\|_{H^1(\Omega)} < \varepsilon, \qquad \|\mathbf{\Theta} - \mathbf{\Theta}_\varepsilon\|_{L^2(\Omega\setminus\mathbb{S})} < \varepsilon, \\ -1 + \delta_0 \leq s_\varepsilon \leq 1 - \delta_0.\end{gathered} \tag{37}$$

Proof We proceed in several steps.

Step 1: Regularization with boundary condition. Consider a zero-extension of $s - g \in H_0^1(\Omega)$ over $\mathbb{R}^d \setminus \Omega$. Given $\delta > 0$, we set

$$\omega_\delta := \{x \in \Omega \colon d(x, \partial\Omega) \leq \delta\},$$

and define $d_\delta(x) = \chi_\Omega(x) \min\left\{\frac{1}{\delta} d(x, \partial\Omega), 1\right\}$, which is a Lipschitz continuous function, with $\operatorname{supp}(\nabla d_\delta) \subset \omega_\delta$ and $|\nabla d_\delta| = \delta^{-1}\chi_{\omega_\delta}$. Let η_δ be a smooth, nonnegative mollifier supported in $B_\delta(0)$, and define

$$\begin{aligned} s_\delta &:= d_\delta(s * \eta_\delta) + (1 - d_\delta)g, \\ \widetilde{\mathbf{U}}_\delta &:= d_\delta\left(\widetilde{\mathbf{U}} * \eta_\delta\right) + (1 - d_\delta)\mathbf{R}, \end{aligned}$$

where $\widetilde{\mathbf{U}} := \operatorname{sgn}(s)\mathbf{U} = |s|\mathbf{\Theta} \in [H^1(\Omega)]^{d\times d}$ coincides with $\mathbf{R}$ on $\partial\Omega$ (because of (18)). We thus have $(s_\delta, \widetilde{\mathbf{U}}_\delta) = (s, \mathbf{R})$ on $\partial\Omega$ and arguing as in [47, Proposition 3.2, Step 1] it follows that

$$s_\delta \to s, \quad \widetilde{\mathbf{U}}_\delta \to \widetilde{\mathbf{U}} \quad \text{a.e. and in } H^1(\Omega).$$

The choice to regularize the field $\widetilde{\mathbf{U}}$ instead of $\mathbf{U}$ is motivated by the next step. Since convolution breaks the uniaxial structure of tensor fields, we cannot preserve the trace condition $s = \operatorname{tr}[\mathbf{U}]$. However, convolution does preserve positive-semidefiniteness, which is a property that $\widetilde{\mathbf{U}}$ satisfies. Additionally, we shall recover the rank-one constraint by means of the map $\mathbf{\Pi}$ defined in (34). Because $\widetilde{\mathbf{U}} \in S^{1,0}(d)$, we have $\mathbf{\Pi}(\widetilde{\mathbf{U}}) = \widetilde{\mathbf{U}}$; in contrast, if $s < 0$, we have $\mathbf{\Pi}(\mathbf{U}) = \mathbf{0}$ when $d > 2$ and $\mathbf{\Pi}(\mathbf{U}) = -s\mathbf{\Theta}^\perp$ when $d = 2$, where $\mathbf{\Theta}^\perp$ is the line field orthogonal to $\mathbf{\Theta}$ a.e. in Ω.

Step 2: Preserve structural conditions. We now rebuild these conditions into the regularized pair $(s_\delta, \widetilde{\mathbf{U}}_\delta)$ by introducing a coarser scale. Our assumption (36) implies

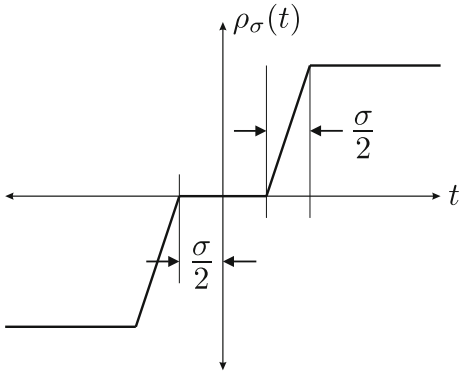

Fig. 1 Regularized sign function

that the extension of s satisfies the same bound on $\mathbb{R}^d \setminus \Omega$. Therefore, we also have $-1 + \delta_0 \leq s_\delta(x) \leq 1 - \delta_0$ on $\mathbb{R}^d$. Moreover, we have $\lambda_1(\widetilde{\mathbf{U}}_\delta) \leq 1 - \delta_0$ since, given any vector $\mathbf{v} \in \mathbb{R}^d$, with $|\mathbf{v}| = 1$, there holds for δ sufficiently small that

$$|\widetilde{\mathbf{U}}_\delta \mathbf{v} \cdot \mathbf{v}| \leq d_\delta |\mathrm{sgn}(s)\mathbf{U}\mathbf{v} \cdot \mathbf{v} * \eta_\delta| + (1 - d_\delta)|\mathbf{R}\mathbf{v} \cdot \mathbf{v}| \leq 1 - \delta_0,$$

because $|\lambda_1(\mathbf{U})| \leq 1 - \delta_0$ a.e. in Ω and $|\lambda_1(\mathbf{R})| \leq 1 - \delta_0$ in a neighborhood of $\partial\Omega$.

We introduce a parameter $\sigma > \delta$ and the following regularization of the sign function (see Fig. 1):

$$\rho_\sigma(t) = \begin{cases} \mathrm{sgn}(t) & \text{if } \sigma < |t|, \\ \frac{2\,\mathrm{sgn}(t)}{\sigma}(|t| - \sigma/2) & \text{if } \sigma/2 < |t| \leq \sigma, \\ 0 & \text{if } |t| \leq \sigma/2. \end{cases}$$

An elementary verification gives

$$\rho_\sigma(s_\delta) \to \rho_\sigma(s) \text{ as } \delta \to 0, \qquad \text{a.e. and in } H^1(\Omega).$$

Next, we use the operator $\mathbf{\Pi}$ given by (34) to define

$$\begin{aligned} s_{\sigma,\delta} &:= \rho_\sigma(s_\delta)\,\mathrm{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)] = \rho_\sigma(s_\delta)|\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)|, \\ \mathbf{U}_{\sigma,\delta} &:= \rho_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta). \end{aligned}$$

Since $\mathrm{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)] = \lambda_1(\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)) \in [0, 1 - \delta_0]$ and $-1 \leq \rho_\sigma \leq 1$, we deduce that $-1 + \delta_0 \leq s_{\sigma,\delta} \leq 1 - \delta_0$; thus, we have $\mathbf{U}_{\sigma,\delta} = s_{\sigma,\delta}\mathbf{\Theta}_{\sigma,\delta}$ for some $\mathbf{\Theta}_{\sigma,\delta} \in \mathbb{L}^{d-1}$ and $(s_{\sigma,\delta}, \mathbf{\Theta}_{\sigma,\delta}, \widetilde{\mathbf{U}}_{\sigma,\delta})$ satisfies the structural condition (16).

Under assumption (18), it follows that if $\sigma < \delta_0$ then $s_\delta = g > \sigma$ on $\partial\Omega$, so that $\rho_\sigma(s_\delta) = 1$ on $\partial\Omega$. Thus,

$$\begin{aligned} s_{\sigma,\delta} &= \mathrm{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)] = \mathrm{tr}(\mathbf{R}) = g \quad \text{on } \partial\Omega, \\ \mathbf{U}_{\sigma,\delta} &= \mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) = \mathbf{R} \quad \text{on } \partial\Omega. \end{aligned}$$

Therefore, $(s_{\sigma,\delta}, \mathbf{\Theta}_{\sigma,\delta}, \mathbf{U}_{\sigma,\delta}) \in \mathcal{A}_{\text{uni}}(g, \mathbf{R}, \mathbf{M})$. We still need to choose σ and δ such that $(s_{\sigma,\delta}, \mathbf{U}_{\sigma,\delta})$ is sufficiently close to $(s, \mathbf{U})$ in $[H^1(\Omega)]^{1+d\times d}$.

Step 3: Convergence as $\delta \to 0$. Since $\mathbf{\Pi}$ is Lipschitz in view of Lemma 4, it is immediate to see that

$$\begin{cases} s_{\sigma,\delta} \to s_\sigma := \rho_\sigma(s)\text{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}})] = \rho_\sigma(s)\text{tr}[\widetilde{\mathbf{U}}] \\ \mathbf{U}_{\sigma,\delta} \to \mathbf{U}_\sigma := \rho_\sigma(s)\mathbf{\Pi}(\widetilde{\mathbf{U}}) = \rho_\sigma(s)\widetilde{\mathbf{U}} \end{cases} \quad \text{a.e. and in } L^2(\Omega),$$

as $\delta \to 0$. Consider now the set $\Lambda_\sigma := \{|s| > \frac{\sigma}{2}\}$ to deal with $\mathbf{\Theta}_{\sigma,\delta}$. The fact that $s_\delta \to s, \widetilde{\mathbf{U}}_\delta \to \widetilde{\mathbf{U}}$ a.e. yields $\rho_\sigma(s_\delta(x)) \neq 0$, $\text{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta(x))] \neq 0$ for a.e. $x \in \Lambda_\sigma$ provided δ is sufficiently small depending on x. Hence

$$\mathbf{\Theta}_{\sigma,\delta} = \frac{\mathbf{U}_{\sigma,\delta}}{s_{\sigma,\delta}} = \frac{\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)}{\text{tr}[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)]} \to \frac{\widetilde{\mathbf{U}}}{\text{tr}[\widetilde{\mathbf{U}}]} = \frac{\widetilde{\mathbf{U}}}{|s|} = \mathbf{\Theta} \quad \text{a.e. in } \Lambda_\sigma \text{ and in } L^2(\Lambda_\sigma), \text{ as } \delta \to 0.$$

We next prove convergence in $H^1(\Omega)$. For $i, j = 1, \dots, d$, we have

$$\begin{cases} \nabla[(\mathbf{U}_{\sigma,\delta})_{ij}] = \rho'_\sigma(s_\delta)\nabla s_\delta \mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} + \rho_\sigma(s_\delta)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij}], \\ \nabla[(\mathbf{U}_\sigma)_{ij}] = \rho'_\sigma(s)\nabla s \mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij} + \rho_\sigma(s)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij}]. \end{cases} \tag{38}$$

It suffices to check convergence term by term in the right hand sides in (38). For the first one, we write

$$\begin{aligned} \rho'_\sigma(s_\delta)\nabla s_\delta \mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} - \rho'_\sigma(s)\nabla s \mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij} &= \nabla(s_\delta - s)\rho'_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} \\ &\quad + \nabla s\left[\rho'_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} - \rho'_\sigma(s)\mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij}\right]. \end{aligned}$$

Since $\nabla(s_\delta - s) \to 0$ in $L^2(\Omega)$ and $|\rho'_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij}|$ is bounded, we deduce that

$$\int_\Omega |\nabla(s_\delta - s)|^2 \left|\rho'_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij}\right|^2 dx \to 0.$$

As for the remaining term, we write

$$\rho'_\sigma(s_\delta)\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} - \rho'_\sigma(s)\mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij} = [\rho'_\sigma(s_\delta) - \rho'_\sigma(s)]\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} + \rho'_\sigma(s)[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} - \mathbf{P}(\widetilde{\mathbf{U}})_{ij}]$$

and notice that

$$\begin{aligned} &\rho'_\sigma(s_\delta) - \rho'_\sigma(s) \to 0 \quad \text{in } L^2(\Omega), \\ &\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} \quad \text{remains bounded}, \\ &|\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)_{ij} - \mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij}| \leq |\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) - \mathbf{\Pi}(\widetilde{\mathbf{U}})| \leq C|\widetilde{\mathbf{U}}_\delta - \widetilde{\mathbf{U}}| \to 0 \quad \text{in } L^2(\Omega), \end{aligned}$$

according to Lemma 4. This shows convergence of the first terms in the right hand sides in (38):

$$\rho_\sigma'(s_\delta)\nabla s_\delta \mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) \to \rho_\sigma'(s)\nabla s \mathbf{\Pi}(\widetilde{\mathbf{U}}) \quad \text{in } L^2(\Omega).$$

To prove that $\rho_\sigma(s_\delta)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)] \to \rho_\sigma(s)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}})]$ in $L^2(\Omega)$, we write

$$\begin{aligned}\rho_\sigma(s_\delta)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)] - \rho_\sigma(s)\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}})] &= (\rho_\sigma(s_\delta) - \rho_\sigma(s))\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}})] \\ &+ \rho_\sigma(s_\delta)D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)\nabla(\widetilde{\mathbf{U}}_\delta - \widetilde{\mathbf{U}}) + \rho_\sigma(s_\delta)(D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) - D\mathbf{\Pi}(\widetilde{\mathbf{U}}))\nabla\widetilde{\mathbf{U}}.\end{aligned} \tag{39}$$

The first term in the right hand side above converges to 0 in $L^2(\Omega)$ because $\nabla[\mathbf{\Pi}(\widetilde{\mathbf{U}})_{ij}] \in L^2(\Omega)$ and $|\rho_\sigma(s_\delta) - \rho_\sigma(s)|$ is bounded and converges to 0 a.e. in Ω. As for the second term in (39), we use Lemma 4 (Lipschitz property of $\mathbf{\Pi}$) and the boundedness of ρ_σ to obtain

$$\int_\Omega \rho_\sigma^2(s_\delta)|D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta)|^2|\nabla(\widetilde{\mathbf{U}}_\delta - \widetilde{\mathbf{U}})|^2 \le \|D\mathbf{\Pi}\|_\infty^2 \int_\Omega |\nabla(\widetilde{\mathbf{U}}_\delta - \widetilde{\mathbf{U}})|^2 \to 0,$$

because $\widetilde{\mathbf{U}}_\delta \to \widetilde{\mathbf{U}}$ in $H^1(\Omega)$.

Finally, to prove that the last term in (39) converges to 0 in $L^2(\Omega)$, we consider Λ_σ as above, namely

$$\Lambda_\sigma = \{|s| > \sigma/2\}, \quad \Omega \setminus \Lambda_\sigma = \{|s| \le \sigma/2\}.$$

In the region $\Omega \setminus \Lambda_\sigma$, we have $\rho_\sigma(s_\delta) \to \rho_\sigma(s) = 0$ a.e.. Using this together with the boundedness of $|\rho_\sigma(s_\delta)|$ and $|D\mathbf{\Pi}|$, and the fact that $\nabla\widetilde{\mathbf{U}} \in L^2(\Omega)$, we obtain

$$\int_{\Omega\setminus\Lambda_\sigma} |\rho_\sigma(s_\delta)|^2|D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) - D\mathbf{\Pi}(\widetilde{\mathbf{U}})|^2|\nabla\widetilde{\mathbf{U}}|^2 \to 0.$$

On the other hand, we have that for a.e. $x \in \Lambda_\sigma$, $\widetilde{\mathbf{U}}(x) = |s(x)|\mathbf{\Theta}(x) \in \mathrm{Sym}^1(d)$. Also, since $\widetilde{\mathbf{U}}_\delta \to \widetilde{\mathbf{U}}$ and $\lambda_1(\widetilde{\mathbf{U}}(x)) = |s(x)| \ge \sigma/2$ a.e. $x \in \Lambda_\sigma$, there exists a δ' (depending on x) such that $\widetilde{\mathbf{U}}_\delta(x) \in \mathrm{Sym}^1(d)$ for all $\delta \le \delta'$. Using that $\mathbf{\Pi}$ is of class C^1 in $\mathrm{Sym}^1(d)$, according to Lemma 3, we deduce that

$$D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) \to D\mathbf{\Pi}(\widetilde{\mathbf{U}}) \quad \text{a.e. in } \Lambda_\sigma.$$

Therefore, applying again the Dominated Convergence Theorem yields

$$\int_{\Lambda_\sigma} |\rho_\sigma(s_\delta)|^2|D\mathbf{\Pi}(\widetilde{\mathbf{U}}_\delta) - D\mathbf{\Pi}(\widetilde{\mathbf{U}})|^2|\nabla\widetilde{\mathbf{U}}|^2 \to 0.$$

We have thus proved that

$$\begin{cases} s_{\sigma,\delta} \to s_\sigma := \rho_\sigma(s)\mathrm{tr}(\widetilde{\mathbf{U}}) \\ \mathbf{U}_{\sigma,\delta} \to \mathbf{U}_\sigma := \rho_\sigma(s)\widetilde{\mathbf{U}} \end{cases} \quad \text{in } H^1(\Omega), \text{ as } \delta \to 0.$$

Step 4: Convergence as $\sigma \to 0$. Because $\widetilde{\mathbf{U}} = |s|\boldsymbol{\Theta}$, a straightforward calculation gives

$$\begin{cases} s_\sigma = \rho_\sigma(s)\mathrm{tr}(\widetilde{\mathbf{U}}) \to s \\ \mathbf{U}_\sigma = \rho_\sigma(s)\widetilde{\mathbf{U}} \to \mathbf{U} \end{cases} \quad \text{a.e. and in } L^2(\Omega), \text{ as } \sigma \to 0.$$

To prove convergence in $H^1(\Omega)$ we observe that $\mathbf{U}_\sigma = \rho_\sigma(s)\widetilde{\mathbf{U}} = \rho_\sigma(s)\,\mathrm{sgn}(s)\,\mathbf{U} = |\rho_\sigma(s)|\,\mathbf{U}$, whence

$$\nabla(\mathbf{U}_\sigma - \mathbf{U}) = \nabla\big[\big(|\rho_\sigma(s)| - 1\big)\mathbf{U}\big] = \nabla|\rho_\sigma(s)|\,\mathbf{U} + \big(|\rho_\sigma(s)| - 1\big)\nabla\mathbf{U}.$$

We show that these two terms tend to zero separately in $L^2(\Omega)$. First note that

$$\big|\nabla|\rho_\sigma(s)|\big| = \rho_\sigma'(s)\big|\nabla s\big| = \frac{2}{\sigma}\chi_{\{\frac{\sigma}{2}<|s|<\sigma\}}\big|\nabla s\big|,$$

whereas $|\mathbf{U}| = |s| < \sigma$ in the set $\{\frac{\sigma}{2} < |s| < \sigma\}$. Since $\chi_{\{\frac{\sigma}{2}<|s|<\sigma\}} \to 0$ a.e. in Ω as $\sigma \to 0$, and $\big|\nabla s\big| \in L^2(\Omega)$, we infer from the Dominated Convergence Theorem that

$$\int_\Omega \big|\nabla|\rho_\sigma(s)|\,\mathbf{U}\big|^2 \to 0 \quad \text{as } \sigma \to 0.$$

On the other hand, in view of the definition of $\rho_\sigma(s)$, we have

$$\int_\Omega \big|\big(|\rho_\sigma(s)| - 1\big)\nabla\mathbf{U}\big|^2 \le \int_\Omega \chi_{\{|s|\le\sigma\}}\big|\nabla\mathbf{U}\big|^2 = \int_\Omega \chi_{\{|\mathbf{U}|\le\sigma\}}\big|\nabla\mathbf{U}\big|^2 \to \int_\Omega \chi_{\{|\mathbf{U}|=0\}}\big|\nabla\mathbf{U}\big|^2 = 0$$

because $\nabla v = 0$ a.e. in $\{v = 0\}$ for any $v \in H^1(\Omega)$ [25, Ch. 5, Exercise 17]. We have thus proved that $\nabla(\mathbf{U}_\sigma - \mathbf{U}) \to 0$ in $L^2(\Omega)$ as $\sigma \to 0$.

It remains to deal with $s_\sigma - s$. We write $s_\sigma = \rho_\sigma(s)\,\mathrm{tr}(\mathrm{sgn}(s)\,\mathbf{U}) = |\rho_\sigma(s)|\,\mathrm{tr}(\mathbf{U})$ to realize that

$$\nabla(s_\sigma - s) = \nabla\big[\big(|\rho_\sigma(s)| - 1\big)\,\mathrm{tr}(\mathbf{U})\big] = \nabla|\rho_\sigma(s)|\,\mathrm{tr}(\mathbf{U}) + \big(|\rho_\sigma(s)| - 1\big)|\nabla\mathrm{tr}(\mathbf{U})|.$$

This expression has the same structure as $\nabla(\mathbf{U}_\sigma - \mathbf{U})$ except that $\mathbf{U}$ is now replaced by $\mathrm{tr}(\mathbf{U})$. Therefore, the same argument as before yields as $\sigma \to 0$

$$\nabla(s_\sigma - s) \to 0 \quad \text{in } L^2(\Omega).$$

Step 5: Choice of σ *and* δ. Given $\varepsilon > 0$, we first choose $\sigma > 0$ such that

$$\|\mathbf{U}_\sigma - \mathbf{U}\|_{H^1(\Omega)} \le \varepsilon/2, \quad \|s_\sigma - s\|_{H^1(\Omega)} \le \varepsilon/2, \quad \|\boldsymbol{\Theta} - \boldsymbol{\Theta}\chi_{\{|s|>\frac{\sigma}{2}\}}\|_{L^2(\Omega\setminus\mathbb{S})} \le \varepsilon/2,$$

because $\chi_{\{|s|>\frac{\sigma}{2}\}} \to \chi_{\{|s|>0\}}$ a.e. as $\sigma \to 0$ and $\Omega\setminus\mathbb{S} = \{|s| > 0\}$. Since $\chi_{\{0<|s|\le\frac{\sigma}{2}\}} \to 0$ a.e. and $|\boldsymbol{\Theta}_{\sigma,\delta}| = 1$, we can further reduce σ so that

$$\|\boldsymbol{\Theta}_{\sigma,\delta}\|_{L^2(\{0<|s|\le\frac{\sigma}{2}\})} \le \varepsilon/4.$$

Finally, take $\delta \leq \sigma$ such that

$$\|\mathbf{U}_{\sigma,\delta} - \mathbf{U}_\sigma\|_{H^1(\Omega)} \leq \varepsilon/2, \quad \|s_{\sigma,\delta} - s_\sigma\|_{H^1(\Omega)} \leq \varepsilon/2, \quad \|\mathbf{\Theta}_{\sigma,\delta} - \mathbf{\Theta}\|_{L^2(\{|s|>\frac{\sigma}{2}\})} \leq \varepsilon/4.$$

The proof concludes upon defining $(s_\varepsilon, \mathbf{\Theta}_\varepsilon, \mathbf{U}_\varepsilon) := (s_{\sigma,\delta}, \mathbf{\Theta}_{\sigma,\delta}, \mathbf{U}_{\sigma,\delta})$. □

With this regularization result at hand, we now address the construction of a recovery sequence. Given $\varepsilon > 0$, let $(s_{\varepsilon,h}, \mathbf{U}_{\varepsilon,h}) := \big(I_h(s_{\varepsilon,h}), I_h(\mathbf{U}_{\varepsilon,h})\big)$ be the Lagrange interpolants of the regularized pair $(s_\varepsilon, \mathbf{U}_\varepsilon)$ constructed in Proposition 7, that are well-defined because $(s_\varepsilon, \mathbf{U}_\varepsilon) \in W^{1,\infty}(\Omega) \times [W^{1,\infty}(\Omega)]^{d\times d}$. We define the line field $\mathbf{\Theta}_{\varepsilon,h} \in \mathbb{T}_h$ so that, at the node $x_i \in \mathcal{N}_h$ it satisfies

$$\mathbf{\Theta}_{\varepsilon,h}(x_i) = \begin{cases} \mathbf{U}_\varepsilon(x_i)/s_\varepsilon(x_i) & \text{if } s_\varepsilon(x_i) \neq 0, \\ \text{any tensor in } \mathbb{L}^{d-1} & \text{if } s_\varepsilon(x_i) = 0. \end{cases}$$

This definition guarantees that $\mathbf{U}_{\varepsilon,h} = I_h(s_{\varepsilon,h}\mathbf{\Theta}_{\varepsilon,h})$, whence the structural condition (21) is satisfied and thus $(s_{\varepsilon,h}, \mathbf{\Theta}_{\varepsilon,h}, \mathbf{U}_{\varepsilon,h}) \in \mathcal{A}^h_{\text{uni}}(g_h, \mathbf{R}_h)$. Because $(s_{\varepsilon,h}, \mathbf{U}_{\varepsilon,h}) \to (s_\varepsilon, \mathbf{U}_\varepsilon)$ in $H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$ as $h \to 0$, we readily deduce that (31) is satisfied. Proving (32) is equivalent to showing that $\mathcal{E}^h \to 0$, the consistency term in (30), and can be done using the same arguments as in [47, Lemma 3.3]. We omit the proof.

Lemma 5 (lim-sup inequality) *Let $(s_\varepsilon, \mathbf{\Theta}_\varepsilon, \mathbf{U}_\varepsilon) \in \mathcal{A}_{\text{uni}}(g, \mathbf{R})$ be the functions constructed in Proposition 7 and $(s_{\varepsilon,h}, \mathbf{\Theta}_{\varepsilon,h}, \mathbf{U}_{\varepsilon,h}) \in \mathcal{A}^h_{\text{uni}}(g_h, \mathbf{R}_h)$ be the discrete functions defined above. Then,*

$$E_{\text{uni}-\text{m}}[s_\varepsilon, \mathbf{\Theta}_\varepsilon] = \lim_{h\to 0} E^h_{\text{uni}-\text{m}}[s_{\varepsilon,h}\mathbf{\Theta}_{\varepsilon,h}] = \lim_{h\to 0} \widetilde{E}^h_{\text{uni}-\text{m}}[s_{\varepsilon,h}\mathbf{U}_{\varepsilon,h}] = \widetilde{E}_{\text{uni}-\text{m}}[s_\varepsilon, \mathbf{U}_\varepsilon].$$

5.2 Lim-inf property: weak lower semicontinuity

This property hinges on convexity of the underlying functional. However, this is not apparent for the main energy in (13)

$$\widetilde{E}_{\text{uni}-\text{m}}[\widetilde{s}, \widetilde{\mathbf{U}}] = -\frac{1}{2d}\int_\Omega |\nabla \widetilde{s}|^2\, dx + \frac{1}{2}\int_\Omega |\nabla \widetilde{\mathbf{U}}|^2\, dx,$$

because of the negative sign. What restores convexity is the structural property (16), which reads $\widetilde{\mathbf{U}} = \widetilde{s}\mathbf{\Theta}$ in terms of the triple $(\widetilde{s}, \mathbf{\Theta}, \widetilde{\mathbf{U}})$, along with $|\widetilde{\mathbf{U}}| = |\widetilde{s}|$ and equalities

$$\big|\nabla \widetilde{s}\big| = \big|\nabla |\widetilde{s}|\big| = \big|\nabla |\widetilde{\mathbf{U}}|\big| = \big|\nabla \widetilde{\mathbf{U}}\big| \qquad \text{a.e. } \Omega.$$

This reveals the fundamental convexity property of $\widetilde{E}_{\text{uni}-\text{m}}[\widetilde{s}, \widetilde{\mathbf{U}}]$, namely

$$\widetilde{E}_{\text{uni}-\text{m}}[\widetilde{s}, \widetilde{\mathbf{U}}] = \frac{d-1}{2d}\int_\Omega \big|\nabla |\widetilde{\mathbf{U}}|\big|^2 dx.$$

The discretization poses a severe challenge to convexity because the discrete variables $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h)$ defined in (27) satisfy $|\widetilde{s}_h| = |\widetilde{\mathbf{U}}_h|$ only at the mesh nodes and $\nabla\widetilde{s}_h \neq \nabla|s_h|$. However, upon flattening the matrix $\mathbf{U}_h$ into a vector and exploiting that the Euclidean norm of the gradient of the flattened matrix coincides with the Fröbenius norm $|\nabla\mathbf{U}_h|$, we resort to [47, Lemma 3.4] to establish the following result.

Lemma 6 (weak lower semi-continuity) *If* $\mathbf{W}_h \in \mathbb{U}_h$ *converges weakly in* $[H^1(\Omega)]^{d\times d}$ *to* $\mathbf{W}$*, then*

$$\liminf_{h\to 0}\left(-\frac{1}{d}\int_\Omega |\nabla I_h|\mathrm{tr}(\mathbf{W}_h)|^2 + \int_\Omega |\nabla\mathbf{W}_h|^2\right) \geq -\frac{1}{d}\int_\Omega |\nabla|\mathrm{tr}(\mathbf{W})|^2 + \int_\Omega |\nabla\mathbf{W}|^2.$$

5.3 Equicoercivity and compactness

The last ingredient to prove the convergence of minimum problems is some form of compactness. This follows by deriving uniform bounds in H^1 for the discrete minimizers $(s_h, \mathbf{U}_h)$ and $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h) = (I_h|s_h|, I_h(|s_h|\boldsymbol{\Theta}_h))$.

Lemma 7 (coercivity) *Given* $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$*, we have*

$$E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h] \geq \frac{d-1}{2d}\max\left\{\|\nabla\mathbf{U}_h\|^2_{L^2(\Omega)}, \|\nabla s_h\|^2_{L^2(\Omega)}\right\},$$

and

$$E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h] \geq \frac{d-1}{2d}\max\left\{\|\nabla\widetilde{\mathbf{U}}_h\|^2_{L^2(\Omega)}, \|\nabla\widetilde{s}_h\|^2_{L^2(\Omega)}\right\}.$$

Proof First of all, definition (23) of $E^h_{\mathrm{uni-m}}$ in conjunction with (22) and (20) readily yields

$$E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h] \geq \frac{d-1}{4d}\sum_{i,j=1}^n k_{ij}(\delta_{ij}s_h)^2 = \frac{d-1}{2d}\|\nabla s_h\|^2_{L^2(\Omega)}.$$

Moreover, because $|\delta_{ij}\widetilde{s}_h| \leq |\delta_{ij}s_h|$ for all $i, j = 1, \ldots, n$, we also have $\|\nabla\widetilde{s}_h\|_{L^2(\Omega)} \leq \|\nabla s_h\|_{L^2(\Omega)}$.

Secondly, combining (25) and (28) with $\mathcal{E}_h \geq 0$, we obtain

$$\frac{1}{2}\|\nabla\mathbf{U}_h\|^2_{L^2(\Omega)} = E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h] + \frac{1}{2d}\|\nabla s_h\|^2_{L^2(\Omega)} - \mathcal{E}_h \leq \frac{d}{d-1}E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h].$$

Estimate $\frac{d-1}{2d}\|\nabla\widetilde{\mathbf{U}}_h\|^2_{L^2(\Omega)} \leq E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h]$ follows similarly from (29). □

Our next goal is to show that, from sequences of discrete functions $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h)$ and $(\widetilde{s}_h, \boldsymbol{\Theta}_h, \widetilde{\mathbf{U}}_h)$ with uniformly bounded energies, it is possible to extract subsequences that converge to admissible functions. For that purpose, we need an elementary auxiliary result.

Lemma 8 (admissible tensors) *Let* $\mathbf{M} \in Sym(d)$ *be such that* $\mathrm{tr}(\mathbf{M}^k) = [\mathrm{tr}(\mathbf{M})]^k$ *for all* $k = 1, \dots, d$. *Then, at least* $d - 1$ *eigenvalues of* $\mathbf{M}$ *are equal to zero, i.e.,* $\mathbf{M}$ *has rank less than or equal to 1.*

We are now ready to pursue our goal. The key point in the next result is to verify that the candidate tensor fields satisfy the rank-one constraint.

Lemma 9 (characterization of limits) *Let a sequence* $(s_h, \mathbf{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$ *satisfy*

$$E^h_{\mathrm{uni-m}}[s_h, \mathbf{\Theta}_h] \leq \Lambda \quad \forall h > 0,$$

for some constant Λ *independent of* h, *and let* $\widetilde{s}_h = I_h(|s_h|)$, $\widetilde{\mathbf{U}}_h = I_h(|s_h|\mathbf{\Theta}_h)$ *as in (27). Then, there exist subsequences (not relabeled)* $(s_h, \mathbf{U}_h) \in \mathbb{X}_h$ *and* $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h) \in \mathbb{X}_h$, *and functions* $(s, \mathbf{U})$, $(\widetilde{s}, \widetilde{\mathbf{U}}) \in H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$ *and* $\mathbf{\Theta} \in L^\infty(\Omega; \mathbb{L}^{d-1})$ *such that:*

- $(s_h, \mathbf{U}_h) \to (s, \mathbf{U})$ *in* $L^2(\Omega) \times [L^2(\Omega)]^{d\times d}$, *a.e in* Ω, $(s_h, \mathbf{U}_h) \rightharpoonup (s, \mathbf{U})$ *in* $H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$;
- $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h) \to (\widetilde{s}, \widetilde{\mathbf{U}})$ *in* $L^2(\Omega) \times [L^2(\Omega)]^{d\times d}$, *a.e in* Ω, $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h) \rightharpoonup (\widetilde{s}, \widetilde{\mathbf{U}})$ *in* $H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$;
- *the limits satisfy* $\widetilde{s} = |s| = \mathrm{tr}[\widetilde{\mathbf{U}}]$, $s = \mathrm{tr}[\mathbf{U}]$, *a.e. in* Ω;
- $\mathbf{\Theta}_h \to \mathbf{\Theta}$ *a.e. in* $\Omega \setminus \mathbb{S}$, *and in* $L^2(\Omega \setminus \mathbb{S})$, *and* $\mathbf{U} = s\mathbf{\Theta}$, $\widetilde{\mathbf{U}} = \widetilde{s}\mathbf{\Theta}$ *a.e. in* Ω;
- $\mathbf{\Theta}$ *admits Lebesgue gradient* $\nabla\mathbf{\Theta}$ *a.e. in* $\Omega \setminus \mathbb{S}$ *and* $|\nabla\widetilde{\mathbf{U}}|^2 = |\nabla\widetilde{s}|^2 + \widetilde{s}^2|\nabla\mathbf{\Theta}|^2$ *is valid a.e. in* $\Omega \setminus \mathbb{S}$;

where $\mathbb{L}^{d-1}$ *is defined in (15) and* $\mathbb{S}$ *in (8).*

Proof Because the discrete energy $E^h_{\mathrm{uni-m}}[s_h, \mathbf{\Theta}_h]$ is uniformly bounded, Lemma 7 guarantees that the sequences $(s_h, \mathbf{U}_h)$ and $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h)$ are bounded in $H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$. Thus, we can extract subsequences (not relabeled) such that

$$(s_h, \mathbf{U}_h) \to (s, \mathbf{U}) \quad \text{and} \quad (\widetilde{s}_h, \widetilde{\mathbf{U}}_h) \to (\widetilde{s}, \widetilde{\mathbf{U}}),$$

strongly in $L^2(\Omega) \times [L^2(\Omega)]^{d\times d}$, a.e. in Ω, and weakly in $H^1(\Omega) \times [H^1(\Omega)]^{d\times d}$. The rest of the proof is about characterizing these limits. We proceed in three steps.

Step 1: Trace constraint. To show that $\widetilde{s} = |s|$, we use a standard approximation estimate for the Lagrange interpolant and the fact that $|\nabla|s_h|| = |\nabla s_h|$ a.e.:

$$\|\widetilde{s}_h - |s_h|\|_{L^2(\Omega)} = \|I_h|s_h| - |s_h|\|_{L^2(\Omega)} \leq Ch\|\nabla|s_h|\|_{L^2(\Omega)} \leq C\Lambda h.$$

This, together with the triangle inequality and the fact that $s_h \to s$, $\widetilde{s}_h \to \widetilde{s}$ in $L^2(\Omega)$, give

$$\big|\widetilde{s} - |s|\big| \leq \big|\widetilde{s} - \widetilde{s}_h v| + \big|\widetilde{s}_h - |s_h|\big| + \big||s_h| - |s|\big| \to 0 \quad \text{as } h \to 0.$$

Using a similar argument, we can show that $s = \mathrm{tr}[\mathbf{U}]$ and $\widetilde{s} = \mathrm{tr}[\widetilde{\mathbf{U}}]$. Indeed, since $s_h = I_h(\mathrm{tr}[\mathbf{U}_h])$, we have

$$\|\mathrm{tr}[\mathbf{U}_h] - s_h\|_{L^2(\Omega)} \le Ch\|\nabla(\mathrm{tr}[\mathbf{U}_h])\|_{L^2(\Omega)} \le C\Lambda h,$$

and thus

$$\big|\mathrm{tr}[\mathbf{U}] - s\big| \le \big|\mathrm{tr}[\mathbf{U}] - \mathrm{tr}[\mathbf{U}_h]\big| + \big|\mathrm{tr}[\mathbf{U}_h] - s_h\big| + \big|s_h - s\big| \to 0 \quad \text{as } h \to 0.$$

Step 2: Rank-one constraint. We now show that both $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ have rank at most 1; this is a new issue relative to [47]. In order to apply Lemma 8, it suffices to check that

$$s^k = \mathrm{tr}[\mathbf{U}^k], \quad \widetilde{s}^k = \mathrm{tr}[\widetilde{\mathbf{U}}^k] \quad \forall k = 2, \dots, d.$$

Since the two identities above follow from the same argument, we just prove the first one. Let $2 \le k \le d$. The discrete admissibility condition (21) implies that $s_h^k(x_i) = \mathrm{tr}[\mathbf{U}_h(x_i)^k]$ for all $x_i \in \mathcal{N}_h$, whence $I_h(s_h^k) = I_h(\mathrm{tr}[\mathbf{U}_h^k])$. In a similar fashion as before, we use the triangle inequality to write

$$\big|s^k - \mathrm{tr}[\mathbf{U}^k]\big| \le \big|s^k - s_h^k\big| + \big|s_h^k - I_h(s_h^k)\big| + \big|I_h(\mathrm{tr}[\mathbf{U}_h^k]) - \mathrm{tr}[\mathbf{U}_h^k]\big| + \big|\mathrm{tr}[\mathbf{U}_h^k] - \mathrm{tr}[\mathbf{U}^k]\big|.$$

The first and last terms in the right hand side tend to 0 a.e., because $s_h \to s$ and $\mathbf{U}_h \to \mathbf{U}$. Next, we note that $|\nabla s_h^k| = k|s_h|^{k-1}|\nabla s_h| \le d|\nabla s_h|$, because $|s_h| \le 1$, whence

$$\|s_h^k - I_h(s_h^k)\|_{L^2(\Omega)} \le C\Lambda h \to 0, \quad \text{as } h \to 0.$$

The estimate

$$\|I_h(\mathrm{tr}[\mathbf{U}_h^k]) - \mathrm{tr}[\mathbf{U}_h^k]\|_{L^2(\Omega)} \le C\Lambda h \to 0, \quad \text{as } h \to 0,$$

follows in a similar fashion. This proves that $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ have rank ≤ 1 a.e.

Step 3: Line field $\boldsymbol{\Theta}$. Because $s = \mathrm{tr}[\mathbf{U}]$, it follows that $\mathrm{rank}(\mathbf{U}) = 1$ if and only if $s \ne 0$. Therefore, we can define a line field $\boldsymbol{\Theta} : \Omega \setminus \mathbb{S} \to \mathbb{L}^{d-1}$ by $\boldsymbol{\Theta} = s^{-1}\mathbf{U}$, and extend $\boldsymbol{\Theta}$ to $\mathbb{S}$ by any arbitrary tensor in $\mathbb{L}^{d-1}$.

We next show that $\boldsymbol{\Theta}_h \to \boldsymbol{\Theta}$ a.e. in $\Omega \setminus \mathbb{S}$ and in $L^2(\Omega \setminus \mathbb{S})$. We note that at every element $T \in \mathcal{T}_h$, the second derivatives of s_h and $\boldsymbol{\Theta}_h$ vanish, because these functions are piecewise linear. Thus, $\|s_h\boldsymbol{\Theta}_h - I_h(s_h\boldsymbol{\Theta}_h)\|_{L^1(T)} \le Ch^2\|\nabla s_h \otimes \nabla\boldsymbol{\Theta}_h\|_{L^1(T)}$, and summing over all elements $T \in \mathcal{T}_h$, we obtain

$$\|s_h\boldsymbol{\Theta}_h - I_h(s_h\boldsymbol{\Theta}_h)\|_{L^1(\Omega)} \le Ch^2\|\nabla s_h \otimes \nabla\boldsymbol{\Theta}_h\|_{L^1(\Omega)} \le Ch^2\|\nabla s_h\|_{L^2(\Omega)}\|\nabla\boldsymbol{\Theta}_h\|_{L^2(\Omega)}.$$

Since $|\boldsymbol{\Theta}_h| \le 1$, an inverse inequality yields $\|\nabla\boldsymbol{\Theta}_h\|_{L^2(\Omega)} \le Ch^{-1}$ and therefore

$$\|s_h\boldsymbol{\Theta}_h - I_h(s_h\boldsymbol{\Theta}_h)\|_{L^1(\Omega)} \le C\Lambda h \to 0 \quad \text{as } h \to 0. \tag{40}$$

Noticing that $I_h(s_h\boldsymbol{\Theta}_h) = \mathbf{U}_h \to \mathbf{U}$, we deduce that $s_h\boldsymbol{\Theta}_h \to \mathbf{U}$ a.e. in Ω as $h \to 0$. Since $s_h \to s$ a.e., for almost every $x \in \Omega \setminus \mathbb{S}$ it holds that $s_h(x) \neq 0$ if h is sufficiently small, and we deduce

$$\boldsymbol{\Theta}_h(x) = \frac{s_h(x)\boldsymbol{\Theta}_h(x)}{s_h(x)} \to \frac{\mathbf{U}(x)}{s(x)} = \boldsymbol{\Theta}(x) \quad \text{as } h \to 0.$$

Convergence $\boldsymbol{\Theta}_h \to \boldsymbol{\Theta}$ in $L^2(\Omega \setminus \mathbb{S})$ now follows by the Dominated Convergence Theorem, as $|\boldsymbol{\Theta}_h| \leq 1$. Finally, to prove that $\widetilde{\mathbf{U}} = \widetilde{s}\boldsymbol{\Theta}$ a.e. in Ω, in the same fashion as (40) we can show that $\|\widetilde{s}_h\boldsymbol{\Theta}_h - I_h(\widetilde{s}_h\boldsymbol{\Theta}_h)\|_{L^1(\Omega)} \to 0$ as $h \to 0$ which, recalling that $\widetilde{\mathbf{U}}_h = I_h(\widetilde{s}_h\boldsymbol{\Theta}_h) \to \widetilde{\mathbf{U}}$, gives $\widetilde{s}_h\boldsymbol{\Theta}_h \to \widetilde{\mathbf{U}}$. Because $\widetilde{s}_h \to \widetilde{s}$ and $\boldsymbol{\Theta}_h \to \boldsymbol{\Theta}$ a.e. in $\Omega \setminus \mathbb{S}$, it follows that $\widetilde{\mathbf{U}} = \widetilde{s}\boldsymbol{\Theta}$ a.e. in Ω.

Step 4: Lebesgue gradient and orthogonality. At the Lebesgue points of $(\widetilde{s}, \widetilde{\mathbf{U}})$ and their weak gradients $(\nabla\widetilde{s}, \nabla\widetilde{\mathbf{U}})$, the first order Taylor expansions exist and define superlinear approximations of $(\widetilde{s}, \widetilde{\mathbf{U}})$ in the L^2 sense [26, Chapter 6.1.2]. This defines L^2-gradients for $(\widetilde{s}.\widetilde{\mathbf{U}})$ which coincide with the weak gradients. At each Lebesgue point $x \in \Omega \setminus \mathbb{S}$ of $(\widetilde{s}, \boldsymbol{\Theta}, \widetilde{\mathbf{U}}, \nabla\widetilde{s}, \nabla\widetilde{\mathbf{U}})$ we define the quantity $\nabla\boldsymbol{\Theta}(x)$ to be

$$\nabla\boldsymbol{\Theta}(x) := \frac{\nabla\widetilde{\mathbf{U}}(x) - \nabla\widetilde{s}(x) \otimes \boldsymbol{\Theta}(x)}{\widetilde{s}(x)}.$$

To verify that $\nabla\boldsymbol{\Theta}(x)$ is the L^2-gradient of $\boldsymbol{\Theta}$ at x, we have to show that the first order Taylor expansion around $y = x$ gives a superlinear approximation of $\boldsymbol{\Theta}(y)$ in the L^2 sense. Therefore, we let $B_\varepsilon(x)$ denote the ball centered at x of radius ε and observe that

$$\begin{aligned}
&\fint_{B_\varepsilon(x)} \left|\boldsymbol{\Theta}(y) - \boldsymbol{\Theta}(x) - \nabla\boldsymbol{\Theta}(x)(y-x)\right|^2 dy \\
&\qquad \lesssim \frac{1}{\widetilde{s}(x)^2} \fint_{B_\varepsilon(x)} \left|\widetilde{\mathbf{U}}(y) - \widetilde{\mathbf{U}}(x) - \nabla\widetilde{\mathbf{U}}(x)(y-x)\right|^2 dy \\
&\qquad + \frac{1}{\widetilde{s}(x)^2} \fint_{B_\varepsilon(x)} \left|\widetilde{s}(y) - \widetilde{s}(x) - \nabla\widetilde{s}(x)(y-x)\right|^2 \left|\boldsymbol{\Theta}(y)\right|^2 dy \\
&\qquad + \frac{\left|\nabla\widetilde{s}(x)\right|^2}{\widetilde{s}(x)^2} \fint_{B_\varepsilon(x)} \left|\boldsymbol{\Theta}(y) - \boldsymbol{\Theta}(x)\right|^2 |y-x|^2 dy = o(\varepsilon^2)
\end{aligned}$$

as $\varepsilon \to 0$ because the first order Taylor expansions of $(\widetilde{s}, \widetilde{\mathbf{U}})$ converge superlinearly at x, which is a Lebesgue point of $\boldsymbol{\Theta}$ that belongs to $L^\infty(\Omega)$, and $\widetilde{s}(x) > 0$ and $\nabla\widetilde{s}(x)$ are fixed.

We next claim that $\nabla\boldsymbol{\Theta} : \nabla\widetilde{s} \otimes \boldsymbol{\Theta} = 0$ and note that this is true if and only if $\nabla\widetilde{\mathbf{U}} : \nabla\widetilde{s} \otimes \boldsymbol{\Theta} = |\nabla\widetilde{s} \otimes \boldsymbol{\Theta}|^2$ at any Lebesgue point $x \in \Omega \setminus \mathbb{S}$ as above. To see this, we compute at x

$$|\nabla\widetilde{s} \otimes \boldsymbol{\Theta}|^2 = \sum_{i,j,k=1}^{d} (\partial_i\widetilde{s})^2 (\boldsymbol{\Theta}_{j,k})^2 \sum_{i=1}^{d} (\partial_i\widetilde{s})^2 = |\nabla\widetilde{s}|^2,$$

and

$$\begin{aligned}\nabla\widetilde{\mathbf{U}} : \nabla\widetilde{s}\otimes\boldsymbol{\Theta} &= \sum_{i,j,k=1}^{d} \partial_i \widetilde{\mathbf{U}}_{j,k}\, \partial_i\widetilde{s}\, \boldsymbol{\Theta}_{j,k} = \frac{1}{\widetilde{s}}\sum_{i=1}^{d}\partial_i\widetilde{s}\sum_{j,k=1}^{d}\partial_i\widetilde{\mathbf{U}}_{j,k}\,\widetilde{s}\boldsymbol{\Theta}_{j,k}\\ &= \frac{1}{\widetilde{s}}\sum_{i=1}^{d}\partial_i\widetilde{s}\sum_{j,k=1}^{d}\partial_i\widetilde{\mathbf{U}}_{j,k}\,\widetilde{\mathbf{U}}_{j,k} = \frac{1}{2\widetilde{s}}\sum_{i=1}^{d}\partial_i\widetilde{s}\,\partial_i|\widetilde{\mathbf{U}}|^2\\ &= \frac{1}{2\widetilde{s}}\sum_{i=1}^{d}\partial_i\widetilde{s}\,\partial_i\widetilde{s}^2 = \sum_{i=1}^{d}(\partial_i\widetilde{s})^2 = |\nabla\widetilde{s}|^2.\end{aligned}$$

This shows the orthogonality relation $|\nabla\widetilde{\mathbf{U}}|^2 = |\nabla\widetilde{s}|^2 + \widetilde{s}^2|\nabla\boldsymbol{\Theta}|^2$ at every Lebesgue point $x \in \Omega\setminus\mathbb{S}$ of $(\widetilde{s}, \boldsymbol{\Theta}, \widetilde{\mathbf{U}}, \nabla\widetilde{s}, \nabla\widetilde{\mathbf{U}})$, and concludes the proof. □

5.4 Γ-convergence

We have collected all the elements needed to prove the main theoretical result of this work. Using a standard argument [15,16,21], we can prove the convergence of discrete global minimizers.

Theorem 1 (convergence of discrete global minimizers) *Let* $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$ *be a sequence of global minimizers of the discrete total energy* $E^h_{\mathrm{uni-t}}$ *defined in (26). Then, every cluster point* $(s, \boldsymbol{\Theta}, \mathbf{U})$ *belongs to* $\mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$ *and* $(s, \mathbf{U})$ *is a global minimizer of the continuous total energy* $\widetilde{E}_{\mathrm{uni-t}}$ *given in (14). Moreover,* $\boldsymbol{\Theta}$ *admits a Lebesgue gradient a.e. in the set* $\Omega\setminus\mathbb{S}$ *so that the continuous main energy*

$$E_{\mathrm{uni-m}}[s, \boldsymbol{\Theta}] := \frac{d-1}{d}\int_{\Omega\setminus\mathbb{S}}|\nabla s|^2 + \frac{1}{2}\int_{\Omega\setminus\mathbb{S}} s^2|\nabla\boldsymbol{\Theta}|^2$$

is well defined and satisfies $E_{\mathrm{uni-m}}[s, \boldsymbol{\Theta}] = \widetilde{E}_{\mathrm{uni-m}}[s, \mathbf{U}]$.

Proof If $\lim_{h\to 0} E^h_{\mathrm{uni}}[s_h, \boldsymbol{\Theta}_h] = \infty$, then $\mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$ is empty because otherwise Lemma 5 (lim-sup inequality) would imply the existence of a triple $(s_h, \boldsymbol{\Theta}_h, \mathbf{U}_h) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$ with uniformly bounded discrete total energy $E^h_{\mathrm{uni}}[s_h, \boldsymbol{\Theta}_h]$. In this case there is nothing to prove. We thus assume there is some $\Lambda > 0$ such that

$$\limsup_{h\to 0} E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h] \le \Lambda.$$

Applying Lemma 7 (coercivity) and Lemma 9 (characterization of limits), we can extract subsequences $(s_h, \mathbf{U}_h) \to (s, \mathbf{U})$, $(\widetilde{s}_h, \widetilde{\mathbf{U}}_h) \to (\widetilde{s}, \widetilde{\mathbf{U}})$, converging a.e. in Ω, strongly in $L^2(\Omega)\times[L^2(\Omega)]^{d\times d}$ and weakly in $H^1(\Omega)\times[H^1(\Omega)]^{d\times d}$, and such that the limits satisfy the structural condition (16). By Lemma 6 (weak lower semi-continuity) and the energy inequality (29), we have

$$\widetilde{E}_{\mathrm{uni-m}}[\widetilde{s},\widetilde{\mathbf{U}}] = -\frac{1}{2d}\int_\Omega \left|\nabla|\mathrm{tr}[\widetilde{\mathbf{U}}]|\right|^2 dx + \frac{1}{2}\int_\Omega |\nabla\widetilde{\mathbf{U}}|^2 dx$$
$$\leq \liminf_{h\to 0}\left(-\frac{1}{2d}\int_\Omega |\nabla \mathrm{tr}[\widetilde{\mathbf{U}}_h]|^2 dx + \frac{1}{2}\int_\Omega |\nabla\widetilde{\mathbf{U}}_h|^2 dx\right) \leq \liminf_{h\to 0} E^h_{\mathrm{uni-m}}[s_h,\boldsymbol{\Theta}_h].$$

Moreover, $\psi_{\mathrm{LdG}}(s_h) \to \psi_{\mathrm{LdG}}(s)$ a.e. in Ω because $s_h \to s$ a.e., whence applying Fatou's Lemma yields

$$E_{\mathrm{LdG,bulk}}[s] = \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \psi_{\mathrm{LdG}}(s)dx \leq \liminf_{h\to 0}\int_\Omega \frac{1}{\eta_{\mathrm{B}}}\psi_{\mathrm{LdG}}(s_h)dx = \liminf_{h\to 0} E^h_{\mathrm{LdG,bulk}}[s_h].$$

We have thus shown that

$$\begin{aligned}\widetilde{E}_{\mathrm{uni-m}}[\widetilde{s},\widetilde{\mathbf{U}}] + E_{\mathrm{LdG,bulk}}[s] &\leq \liminf_{h\to 0}\left(E^h_{\mathrm{uni-m}}[s_h,\boldsymbol{\Theta}_h] + E^h_{\mathrm{LdG,bulk}}[s_h]\right)\\ &= \liminf_{h\to 0} E^h_{\mathrm{uni-t}}[s_h,\boldsymbol{\Theta}_h].\end{aligned} \tag{41}$$

Next, we prove that $\widetilde{E}_{\mathrm{uni-m}}[\widetilde{s},\widetilde{\mathbf{U}}] = E_{\mathrm{uni-m}}[s,\boldsymbol{\Theta}]$. This follows from the orthogonality relation $|\nabla\widetilde{\mathbf{U}}|^2 = |\nabla\widetilde{s}|^2 + \widetilde{s}^2|\nabla\boldsymbol{\Theta}|^2$ of Lemma 9 (characterization of limits), valid a.e. in $\Omega\setminus\mathbb{S}$, as well as $|\nabla\widetilde{\mathbf{U}}| = |\nabla\widetilde{s}| = 0$ a.e. in $\mathbb{S}$ [25, Ch. 5, Exercise 17]. Therefore, making use of properties $\widetilde{s} = |s|$ (from Lemma 9) and $|\nabla\widetilde{s}| = |\nabla s|$, we infer that

$$\begin{aligned}\widetilde{E}_{\mathrm{uni-m}}[\widetilde{s},\widetilde{\mathbf{U}}] &= -\frac{1}{2d}\int_{\Omega\setminus\mathbb{S}}|\nabla\widetilde{s}|^2 + \frac{1}{2}\int_{\Omega\setminus\mathbb{S}}|\nabla\widetilde{\mathbf{U}}|^2\\ &= \frac{d-1}{2d}\int_{\Omega\setminus\mathbb{S}}|\nabla\widetilde{s}|^2 + \frac{1}{2}\int_{\Omega\setminus\mathbb{S}}\widetilde{s}^2|\nabla\boldsymbol{\Theta}|^2 = E_{\mathrm{uni-m}}[s,\boldsymbol{\Theta}].\end{aligned}$$

This, together with (41), shows that the total energy satisfies

$$E_{\mathrm{uni-t}}[s,\boldsymbol{\Theta}] \leq \liminf_{h\to 0} E^h_{\mathrm{uni-t}}[s_h,\boldsymbol{\Theta}_h]. \tag{42}$$

Next, given $\varepsilon > 0$, we consider $(t,\mathbf{N},\mathbf{V}) \in \mathcal{A}_{\mathrm{uni}}(g,\mathbf{R})$ such that

$$E_{\mathrm{uni-t}}[t,\mathbf{N}] \leq \inf_{(\mathbf{t}',\mathbf{N}'),\in\mathcal{A}_{\mathrm{uni}}(g,\mathbf{R})} E_{\mathrm{uni-t}}[t',\mathbf{N}'] + \varepsilon/2$$

and, in view of Proposition 7, we can take $(t_\varepsilon,\mathbf{N}_\varepsilon,\mathbf{V}_\varepsilon) \in \mathcal{A}_{\mathrm{uni}}(g,\mathbf{R})$ with $(t_\varepsilon,\mathbf{V}_\varepsilon) \in W^{1,\infty}(\Omega)\times[W^{1,\infty}(\Omega)]^{d\times d}$ such that

$$E_{\mathrm{uni-m}}[t_\varepsilon,\mathbf{N}_\varepsilon] = \widetilde{E}_{\mathrm{uni-m}}[t_\varepsilon,\mathbf{V}_\varepsilon] \leq \widetilde{E}_{\mathrm{uni-m}}[t,\mathbf{V}] + \varepsilon/4 = E_{\mathrm{uni-m}}[t,\mathbf{N}] + \varepsilon/4.$$

Moreover, because $t_\varepsilon \to t$ a.e. in Ω so does $\psi_{\mathrm{LdG}}(t_\varepsilon) \to \psi_{\mathrm{LdG}}(t)$. Since (19) and (37) imply that $|\psi_{\mathrm{LdG}}(t_\varepsilon)|$ is uniformly bounded, we can apply the Dominated Convergence Theorem to deduce that

$$E_{\mathrm{LdG,bulk}}[t] = \frac{1}{\eta_{\mathrm{B}}} \int_\Omega \lim_{\varepsilon\to 0} \psi_{\mathrm{LdG}}(t_\varepsilon)\, dx = \lim_{\varepsilon\to 0} \frac{1}{\eta_{\mathrm{B}}} \int_\Omega \psi_{\mathrm{LdG}}(t)\, dx = \lim_{\varepsilon\to 0} E_{\mathrm{LdG,bulk}}[t_\varepsilon].$$

Therefore, we can find $(t_\varepsilon, \mathbf{N}_\varepsilon, \mathbf{V}_\varepsilon) \in \mathcal{A}_{\mathrm{uni}}(g, \mathbf{R})$ such that

$$E_{\mathrm{uni-t}}[t_\varepsilon, \mathbf{N}_\varepsilon] \leq \inf_{(\mathbf{t}', \mathbf{N}'), \in \mathcal{A}_{\mathrm{uni}}(g,\mathbf{R})} E_{\mathrm{uni-t}}[t', \mathbf{N}'] + \varepsilon. \tag{43}$$

We next consider the Lagrange interpolants $t_{\varepsilon,h} = I_h(t_\varepsilon)$, $\mathbf{V}_{\varepsilon,h} = I_h(\mathbf{V}_\varepsilon)$, and set $\mathbf{N}_{\varepsilon,h}(x_i) = \mathbf{V}_\varepsilon(x_i)/t_\varepsilon(x_i)$ if $t_\varepsilon(x_i) \neq 0$ and $\mathbf{N}_{\varepsilon,h}(x_i)$ equal to any tensor in $\mathbb{L}^{d-1}$ otherwise. By the same arguments as before, it follows that

$$E_{\mathrm{LdG,bulk}}[t_\varepsilon] = \frac{1}{\eta_{\mathrm{B}}} \int_\Omega \lim_{h\to 0} \psi_{\mathrm{LdG}}(t_{\varepsilon,h}) dx = \lim_{h\to 0} \frac{1}{\eta_{\mathrm{B}}} \int_\Omega \psi_{\mathrm{LdG}}(t_{\varepsilon,h}) dx = \lim_{h\to 0} E^h_{\mathrm{LdG,bulk}}[t_{\varepsilon,h}].$$

Using Lemma 5 (lim-sup inequality) in conjunction with this estimate, we arrive at

$$E_{\mathrm{uni-t}}[t_\varepsilon, \mathbf{N}_\varepsilon] = \lim_{h\to 0} E^h_{\mathrm{uni-t}}[t_{\varepsilon,h}, \mathbf{N}_{\varepsilon,h}],$$

and therefore, by (42) and (43), the total energies verify

$$E_{\mathrm{uni-t}}[s, \boldsymbol{\Theta}] \leq \liminf_{h\to 0} E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h] \leq \lim_{h\to 0} E^h_{\mathrm{uni-t}}[t_{\varepsilon,h}, \mathbf{N}_{\varepsilon,h}] \leq \inf_{(\mathbf{t}', \mathbf{N}'), \in \mathcal{A}_{\mathrm{uni}}(g,\mathbf{R})} E_{\mathrm{uni-t}}[t', \mathbf{N}'] + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this proves that $(s, \boldsymbol{\Theta})$ is a global minimizer of $E_{\mathrm{uni-t}}$. □

In case there is a unique global minimizer of the continuous total energy $E_{\mathrm{uni-t}}$, Theorem 1 implies that the entire sequence of discrete global energy minimizers converges to it strongly in L^2 and weakly in H^1. We also point out that a well-known result in Γ-convergence theory [37] guarantees that, for every isolated local minimizer of $E_{\mathrm{uni-t}}$ there is a sequence of local minimizers of $E^h_{\mathrm{uni-t}}$ that converges to it in the same sense. However, in either case, because of the lack of continuous dependence on data as well as regularity theory, we cannot derive convergence rates.

6 Computation of discrete minimizers

We next discuss a gradient flow algorithm for the computation of discrete minimizers. Recall that, according to (26), we write the discrete total energy as

$$E^h_{\mathrm{uni-t}}[s_h, \boldsymbol{\Theta}_h] = E^h_{\mathrm{uni-m}}[s_h, \boldsymbol{\Theta}_h] + E^h_{\mathrm{LdG,bulk}}[s_h],$$

with main and bulk energies

$$E^h_{\mathrm{uni-m}}[s_h, \mathbf{\Theta}_h] = \frac{d-1}{4d} \sum_{i,j=1}^{N} k_{ij} \left(\delta_{ij} s_h\right)^2 + E^h_{\mathrm{uni-i}}[s_h, \mathbf{\Theta}_h],$$
$$E^h_{\mathrm{LdG,bulk}}[s_h] = \frac{1}{\eta_{\mathrm{B}}} \int_\Omega \psi_{\mathrm{LdG}}(s_h) dx,$$

where $E^h_{\mathrm{uni-i}}[s_h, \mathbf{\Theta}_h]$ is the interaction energy

$$E^h_{\mathrm{uni-i}}[s_h, \mathbf{\Theta}_h] = \frac{1}{4} \sum_{i,j=1}^{N} k_{ij} \left(\frac{s_h(x_i)^2 + s_h(x_j)^2}{2} \right) |\delta_{ij} \mathbf{\Theta}_h|^2.$$

Tangential variations The algorithm we discuss here is an alternating direction method that, at each step $k \geq 0$, first performs a tangential variation on the current line field $\mathbf{\Theta}_h = \mathbf{n}_h^k \otimes \mathbf{n}_h^k$, then normalizes the update, and finally performs a gradient flow step on the current degree of orientation s_h. The director field $\mathbf{n}_h^k$ belongs to

$$\mathbb{N}_h = \{\mathbf{v}_h \in [\mathbb{S}_h]^d : \mathbf{v}_h(x_i) \in \mathbb{S}^{d-1} \ \forall x_i \in \mathcal{N}_h\},$$

whereas a tangential variation $\mathbf{t}_h^k$ belongs to the space

$$\mathbb{N}_h^\perp(\mathbf{n}_h^k) = \{\mathbf{v}_h \in [\mathbb{S}_h]^d : \mathbf{v}_h(x_i) \cdot \mathbf{n}_h^k(x_i) = 0 \ \forall x_i \in \mathcal{N}_h\}.$$

It is easy to see that tangential variations $\mathbf{T}_h^k$ of $\mathbf{\Theta}_h^k$ are of the form

$$\mathbf{T}_h^k = \mathbf{n}_h^k \otimes \mathbf{t}_h^k + \mathbf{t}_h^k \otimes \mathbf{n}_h^k$$

with $\mathbf{t}_h^k \in \mathbb{N}_h^\perp(\mathbf{n}_h^k)$. However, in our algorithm we shall update the line field $\widehat{\mathbf{\Theta}}_h^{k+1}$ by

$$\widehat{\mathbf{\Theta}}_h^{k+1} = \left(\mathbf{n}_h^k + \mathbf{t}_h^k\right) \otimes \left(\mathbf{n}_h^k + \mathbf{t}_h^k\right) = \mathbf{\Theta}_h^k + \mathbf{T}_h^k + \mathbf{t}_h^k \otimes \mathbf{t}_h^k.$$

The extra quadratic term can be handled if we have control of $\mathbf{t}_h^k$ in an $H^1(\Omega)$-type space. This dictates the metric of the gradient flow. Bartels and Raisch first proposed the metric $H^1(\Omega)$ provided $s_h^k > 0$ is constant [12]. In our case, s_h^k may vary across the domain and may even vanish to allow for the formation of defects. Near the singular set, where s_h^k is small, it is critical to allow for relatively large variations $\mathbf{t}_h^k$ in order to accelerate the algorithm. We achieve this via the weight $\omega = (s_h^k)^2$ and corresponding weighted H^1-norm

$$\|\mathbf{v}\|_{H^1_\omega(\Omega)} := \left(\int_\Omega |\mathbf{v}(x)|^2 \, dx + \int_\Omega |\nabla \mathbf{v}(x)|^2 \, \omega(x) \, dx \right)^{1/2}. \tag{44}$$

Moreover, $\mathbf{t}_h^k$ must vanish on the Dirichlet part $\Gamma_{\boldsymbol{\Theta}} = \Gamma_{\mathbf{U}}$ of the boundary so that $\widehat{\boldsymbol{\Theta}}_h^{k+1} = \mathbf{M}$ on $\Gamma_{\boldsymbol{\Theta}}$. We thus introduce the subspace $H^1_{\Gamma_{\boldsymbol{\Theta}}}(\Omega)$ of $H^1(\Omega)$ of functions with vanishing trace on $\Gamma_{\boldsymbol{\Theta}}$.

Discrete gradient flow The algorithm reads as follows. Given $(s_h^0, \boldsymbol{\Theta}_h^0, \mathbf{U}_h^0) \in \mathcal{A}^h_{\mathrm{uni}}(g_h, \mathbf{R}_h)$, with $\boldsymbol{\Theta}_h^0 = \mathbf{n}_h^0 \otimes \mathbf{n}_h^0$, and a time step $\tau > 0$, iterate Steps 1–3 for $k \geq 0$:

(1) *Weighted tangent flow step for* $\boldsymbol{\Theta}_h$: find $\mathbf{t}_h^k \in \mathbb{N}_h^\perp(\mathbf{n}_h^k) \cap [H^1_{\Gamma_{\boldsymbol{\Theta}}}(\Omega)]^d$ and $\mathbf{T}_h^k = \mathbf{n}_h^k \otimes \mathbf{t}_h^k + \mathbf{t}_h^k \otimes \mathbf{n}_h^k$ such that

$$\frac{1}{\tau}\int_\Omega \left(\mathbf{t}_h^k \cdot \mathbf{v}_h + \nabla \mathbf{t}_h^k \colon \nabla \mathbf{v}_h |s_h^k|^2\right) + \delta_{\boldsymbol{\Theta}_h} E^h_{\mathrm{uni-i}}[s_h^k, \boldsymbol{\Theta}_h^k + \mathbf{T}_h^k; \mathbf{V}_h] = 0 \quad (45)$$

for all $\mathbf{V}_h = \mathbf{n}_h^k \otimes \mathbf{v}_h + \mathbf{v}_h \otimes \mathbf{n}_h^k$, $\mathbf{v}_h \in \mathbb{N}_h^\perp(\mathbf{n}_h^k) \cap [H^1_{\Gamma_{\boldsymbol{\Theta}}}(\Omega)]^d$.

(2) *Projection:* update $\boldsymbol{\Theta}_h^{k+1} \in \mathbb{T}_h$ by

$$\boldsymbol{\Theta}_h^{k+1}(x_i) := \frac{\mathbf{n}_h^k(x_i) + \mathbf{t}_h^k(x_i)}{|\mathbf{n}_h^k(x_i) + \mathbf{t}_h^k(x_i)|} \otimes \frac{\mathbf{n}_h^k(x_i) + \mathbf{t}_h^k(x_i)}{|\mathbf{n}_h^k(x_i) + \mathbf{t}_h^k(x_i)|} \quad \forall x_i \in \mathcal{N}_h. \quad (46)$$

(3) *Gradient flow step for* s_h: find $s_h^{k+1} \in \mathbb{S}_h(g_h)$ such that

$$\frac{1}{\tau}\int_\Omega (s_h^{k+1} - s_h^k)\, z_h + \delta_{s_h} E^h_{\mathrm{uni-t}}[s_h^{k+1}, \boldsymbol{\Theta}_h^{k+1}; z_h] = 0 \quad \forall z_h \in \mathbb{S}_h(0).$$

The symbols $\delta_{\boldsymbol{\Theta}_h} E^h_{\mathrm{uni-m}}$ and $\delta_{s_h} E^h_{\mathrm{uni-m}}$ stand for the standard first variations of these functionals, whereas $\delta_{s_h} E^h_{\mathrm{LdG,bulk}}$ uses the following convex splitting method [54,64] to obtain an unconditionally stable scheme. Let ψ_c, ψ_e be convex functions so that the double-well potential splits as $\psi_{\mathrm{LdG}}(s) = \psi_c(s) - \psi_e(s)$ and take

$$\delta_{s_h} E^h_{\mathrm{LdG,bulk}}[s_h^{k+1}; z_h] := \frac{1}{\eta_{\mathrm{B}}}\int_\Omega \left(\psi_c'(s_h^{k+1}) - \psi_e'(s_h^k)\right) z_h \, dx \quad \forall\, z_h \in \mathbb{S}_h(0). \quad (47)$$

Energy decrease property Note that the discrete interaction energy (24) can be written equivalently as

$$E^h_{\mathrm{uni-i}}[s_h^k, \boldsymbol{\Theta}_h^k] = \frac{1}{8}\sum_{i,j} k_{ij}\left(s_h(x_1)^2 + s_h(x_j)^2\right)\left(1 - \boldsymbol{\Theta}_h^k(x_i) \colon \boldsymbol{\Theta}_h^k(x_j)\right).$$

To show that Step 2 decreases this energy, namely

$$E^h_{\mathrm{uni-i}}[s_h^k, \boldsymbol{\Theta}_h^{k+1}] \leq E^h_{\mathrm{uni-i}}[s_h^k, \widehat{\boldsymbol{\Theta}}_h^{k+1}], \quad (48)$$

we recall that $k_{ij} \geq 0$ if $i \neq j$ and invoke the following result from [12, Lemmas 3 and 4], but omit its proof.

Lemma 10 (monotonicity) *Let the mesh $\mathcal{T}_h$ be weakly acute (cf. (20)) and let $\mathbf{v}_h \in \mathbb{U}_h$ be such that $|\mathbf{v}_h(x_i)| \geq 1$ for all $x_i \in \mathcal{N}_h$. The discrete tensor fields $\mathbf{V}_h, \widetilde{\mathbf{V}}_h \in \mathbb{U}_h$,*

$$\mathbf{V}_h(x_i) = \mathbf{v}_h(x_i) \otimes \mathbf{v}_h(x_i), \quad \widetilde{\mathbf{V}}_h(x_i) = \frac{\mathbf{v}_h(x_i)}{|\mathbf{v}_h(x_i)|} \otimes \frac{\mathbf{v}_h(x_i)}{|\mathbf{v}_h(x_i)|}.$$

satisfy the inequality

$$1 - \mathbf{V}_h(x_i) : \mathbf{V}_h(x_j) \leq \frac{1}{2}|\delta_{ij}\widetilde{\mathbf{V}}_h|^2.$$

We also need the following key property of (47) (cf. [47, Lemma 4.1], for example).

Lemma 11 (convex-concave splitting) *Given $s_h^k, s_h^{k+1} \in \mathbb{S}_h$, we have*

$$\int_\Omega \psi_{\mathrm{LdG}}(s_h^{k+1})\,dx - \int_\Omega \psi_{\mathrm{LdG}}(s_h^k)\,dx \leq \delta_{s_h} E^h_{\mathrm{LdG,bulk}}[s_h^{k+1}; s_h^{k+1} - s_h^k].$$

Next, we prove that the discrete gradient flow algorithm is energy-decreasing.

Theorem 2 (energy decrease) *If the meshes are weakly acute and $\tau \leq C_0 h^{d/2}$, with C_0 proportional to $E^h_{\mathrm{uni-t}}[s_h^0, \mathbf{\Theta}_h^0]^{-1/2}$, then it holds that*

$$E^h_{\mathrm{uni-t}}[s_h^K, \mathbf{\Theta}_h^K] + \frac{1}{2\tau}\left(\sum_{k=0}^{K-1} \|\mathbf{t}_h^k\|^2_{H^1_\omega(\Omega)} + \|s_h^{k+1} - s_h^k\|^2_{L^2(\Omega)}\right) \leq E^h_{\mathrm{uni-t}}[s_h^0, \mathbf{\Theta}_h^0] \quad \forall K \geq 1,$$

where $H^1_\omega(\Omega)$ is the weighted Sobolev space defined in (44). Therefore, the algorithm stops in a finite number of steps: given a tolerance ε, there exists $K = K_\varepsilon \geq 1$ such that $\frac{1}{\tau}(\|\mathbf{t}_h^K\|^2_{H^1_\omega(\Omega)} + \|s_h^K - s_h^{K-1}\|^2) < \varepsilon$.

Proof We proceed as in [12, Lemma 6] except for the presence of the variable order parameter s_h^k and the weighted $H^1_\omega(\Omega)$ metric. We make the induction assumption that

$$E^h_{\mathrm{uni-t}}[s_h^k, \mathbf{\Theta}_h^k] \leq \Lambda := E^h_{\mathrm{uni-t}}[s_h^0, \mathbf{\Theta}_h^0].$$

for $k \geq 0$ and show the estimate

$$\frac{1}{2\tau}\left(\|\mathbf{t}_h^k\|^2_{H^1_\omega(\Omega)} + \|s_h^{k+1} - s_h^k\|^2_{L^2(\Omega)}\right) + E^h_{\mathrm{uni-t}}[s_h^{k+1}, \mathbf{\Theta}_h^{k+1}] \leq E^h_{\mathrm{uni-t}}[s_h^k, \mathbf{\Theta}_h^k].$$

Upon summation on k this implies the asserted estimate. We split the proof into several steps.

Step 1: Explicit expression for the solution to (45). In order to simplify the notation, we write

$$\begin{aligned} \sigma_{ij} &:= k_{ij}\,\frac{s_h^k(x_i)^2 + s_h^k(x_j)^2}{2} \geq 0, \ \text{if } i \neq j, \\ \widetilde{\mathbf{\Theta}}_h^{k+1} &:= \mathbf{\Theta}_h^k + \mathbf{T}_h^k. \end{aligned} \tag{49}$$

We set $\mathbf{v}_h = \mathbf{t}_h^k$ in (45), and thus $\mathbf{V}_h = \mathbf{T}_h^k$, to obtain

$$\frac{1}{\tau}\|\mathbf{t}_h^k\|^2_{H^1_\omega(\Omega)} + \frac{1}{2}\sum_{i,j}\sigma_{ij}(\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k+1}) : (\delta_{ij}\mathbf{T}_h^k) = 0.$$

The elementary equality $2(\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k+1}) : (\delta_{ij}\mathbf{T}_h^k) = |\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k+1}|^2 - |\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k}|^2 + |\delta_{ij}\mathbf{T}_h^k|^2$ and (24) yield

$$\frac{1}{2}\sum_{i,j}\sigma_{ij}(\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k+1}) : (\delta_{ij}\mathbf{T}_h^k) = E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}] - E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{\Theta}_h^k] + E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{T}_h^k],$$

and therefore we deduce

$$\frac{1}{\tau}\|\mathbf{t}_h^k\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}] + E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{T}_h^k] = E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{\Theta}_h^k]. \tag{50}$$

Step 2: Monotonicity of projection. We define the updated line field to be

$$\widehat{\mathbf{\Theta}}_h^{k+1} = (\mathbf{n}_h^k + \mathbf{t}_h^k) \otimes (\mathbf{n}_h^k + \mathbf{t}_h^k),$$

and recall that $\mathbf{\Theta}^{k+1}$ defined in (46) is its nodewise normalization. From Lemma 10, we have the monotonocity relation (48):

$$E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{\Theta}_h^{k+1}] \le E^h_{\mathrm{uni-i}}[s_h^k, \widehat{\mathbf{\Theta}}_h^{k+1}].$$

Step 3: Bound of the energy $E^h_{\mathrm{uni-i}}[s_h^k, \widehat{\mathbf{\Theta}}_h^{k+1}]$. Expanding the expression for $\widehat{\mathbf{\Theta}}_h^{k+1}$, we have

$$\widehat{\mathbf{\Theta}}_h^{k+1} = \widetilde{\mathbf{\Theta}}_h^{k+1} + \mathbf{t}_h^k \otimes \mathbf{t}_h^k.$$

Therefore, by Cauchy-Schwarz,

$$\begin{aligned} E^h_{\mathrm{uni-i}}[s_h^k, \widehat{\mathbf{\Theta}}_h^{k+1}] &= E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}] + E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{t}_h^k \otimes \mathbf{t}_h^k] + \frac{1}{2}\sum_{i,j}\sigma_{ij}(\delta_{ij}\widetilde{\mathbf{\Theta}}_h^{k+1}) : \delta_{ij}(\mathbf{t}_h^k \otimes \mathbf{t}_h^k) \\ &\le E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}] + E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{t}_h^k \otimes \mathbf{t}_h^k] + 2E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}]^{1/2} E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{t}_h^k \otimes \mathbf{t}_h^k]^{1/2}, \end{aligned}$$

whence

$$\begin{aligned} E^h_{\mathrm{uni-i}}[s_h^k, \widehat{\mathbf{\Theta}}_h^{k+1}] &\le E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}] \\ &\quad + E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{t}_h^k \otimes \mathbf{t}_h^k]^{1/2}\left(E^h_{\mathrm{uni-i}}[s_h^k, \mathbf{t}_h^k \otimes \mathbf{t}_h^k]^{1/2} + 2E^h_{\mathrm{uni-i}}[s_h^k, \widetilde{\mathbf{\Theta}}_h^{k+1}]^{1/2}\right). \end{aligned} \tag{51}$$

Invoking the induction hypothesis, we readily see that $E^h_{\mathrm{uni-i}}[s^k_h, \boldsymbol{\Theta}^k_h] \leq \Lambda$, and using (50) gives

$$\frac{1}{\tau}\|\mathbf{t}^k_h\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-i}}[s^k_h, \widetilde{\boldsymbol{\Theta}}^{k+1}_h] \leq \Lambda. \tag{52}$$

To bound $E^h_{\mathrm{uni-i}}[s^k_h, \mathbf{t}^k_h \otimes \mathbf{t}^k_h]$, we write

$$\delta_{ij}(\mathbf{t}^k_h \otimes \mathbf{t}^k_h) = \delta_{ij}\mathbf{t}^k_h \otimes \mathbf{t}^k_h(x_j) + \mathbf{t}^k_h(x_i) \otimes \delta_{ij}\mathbf{t}^k_h,$$

and thereby obtain

$$\delta_{ij}(\mathbf{t}^k_h \otimes \mathbf{t}^k_h) : \delta_{ij}(\mathbf{t}^k_h \otimes \mathbf{t}^k_h) \leq C|\delta_{ij}\mathbf{t}^k_h|^2 \max\left\{|\mathbf{t}^k_h(x_i)|, |\mathbf{t}^k_h(x_j)|\right\}^2.$$

Using (49), we deduce

$$\begin{aligned} E^h_{\mathrm{uni-i}}[s^k_h, \mathbf{t}^k_h \otimes \mathbf{t}^k_h] &\leq C\sum_{i,j}\sigma_{ij}|\delta_{ij}\mathbf{t}^k_h|^2 \max\left\{|\mathbf{t}^k_h(x_i)|, |\mathbf{t}^k_h(x_j)|\right\}^2 \\ &\leq C\sum_{T\in\mathcal{T}_h}|\mathbf{t}^k_h|^2_{H^1_\omega(T)}\left\|\mathbf{t}^k_h\right\|^2_{L^\infty(T)} \leq C|\mathbf{t}^k_h|^2_{H^1_\omega(\Omega)}\|\mathbf{t}^k_h\|^2_{L^\infty(\Omega)}. \end{aligned}$$

Since the mesh $\mathcal{T}_h$ is shape regular and quasi-uniform, we resort to the inverse inequality $\|\mathbf{t}^k_h\|_{L^\infty(\Omega)} \leq Ch^{-d/2}\|\mathbf{t}^k_h\|_{L^2(\Omega)}$ and rewrite the above expression as follows:

$$E^h_{\mathrm{uni-i}}[s^k_h, \mathbf{t}^k_h \otimes \mathbf{t}^k_h] \leq Ch^{-d}|\mathbf{t}^k_h|^2_{H^1_\omega(\Omega)}\|\mathbf{t}^k_h\|^2_{L^2(\Omega)} \leq Ch^{-d}\|\mathbf{t}^k_h\|^4_{H^1_\omega(\Omega)}.$$

Consequently, (52) yields the bound

$$E^h_{\mathrm{uni-i}}[s^k_h, \mathbf{t}^k_h \otimes \mathbf{t}^k_h]^{1/2} + 2E^h_{\mathrm{uni-i}}[s^k_h, \widetilde{\boldsymbol{\Theta}}^{k+1}_h]^{1/2} \leq Ch^{-d/2}\tau\Lambda + 2\Lambda^{1/2} \leq 4\Lambda^{1/2},$$

provided $\tau \leq C\Lambda^{-1/2}h^{d/2}$. Inserting this expression into (51) results in

$$E^h_{\mathrm{uni-i}}[s^k_h, \widehat{\boldsymbol{\Theta}}^{k+1}_h] \leq E^h_{\mathrm{uni-i}}[s^k_h, \widetilde{\boldsymbol{\Theta}}^{k+1}_h] + Ch^{-d/2}\Lambda^{1/2}\|\mathbf{t}^k_h\|^2_{H^1_\omega(\Omega)}.$$

Step 4: Bound of the energy $E^h_{\mathrm{uni}}[s^k_h, \boldsymbol{\Theta}^{k+1}_h]$. Combining this estimate with (50) and (48), we find

$$\begin{aligned} E^h_{\mathrm{uni-i}}[s^k_h, \boldsymbol{\Theta}^k_h] &\geq \frac{1}{\tau}\|\mathbf{t}^k_h\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-i}}[s^k_h, \widetilde{\boldsymbol{\Theta}}^{k+1}_h] \\ &\geq \frac{1}{\tau}\left(1 - C\Lambda^{1/2}h^{-d/2}\tau\right)\|\mathbf{t}^k_h\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-i}}[s^k_h, \widehat{\boldsymbol{\Theta}}^{k+1}_h] \\ &\geq \frac{1}{2\tau}\|\mathbf{t}^k_h\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-i}}[s^k_h, \boldsymbol{\Theta}^{k+1}_h], \end{aligned}$$

provided $\tau \leq C\Lambda^{-1/2}h^{d/2}$ with a geometric constant C perhaps smaller than before. Since the scalar variable s_h^k remains fixed in the gradient flow for $\boldsymbol{\Theta}_h^{k+1}$, adding $E^h_{\mathrm{uni-s}}[s_h^k]$ to both sides of the above inequality gives

$$\frac{1}{2\tau}\|\mathbf{t}_h^k\|^2_{H^1_\omega(\Omega)} + E^h_{\mathrm{uni-m}}[s_h^k, \boldsymbol{\Theta}_h^{k+1}] \leq E^h_{\mathrm{uni-m}}[s_h^k, \boldsymbol{\Theta}_h^k]. \tag{53}$$

Step 5: Gradient flow for s_h. Taking $z_h = s_h^{k+1} - s_h^k \in \mathbb{S}_h(0)$ in step 3 of the algorithm, and using the elementary identity

$$2s_h^{k+1}\big(s_h^{k+1} - s_h^k\big) = \big|s_h^{k+1}\big|^2 - \big|s_h^k\big|^2 + \big|s_h^{k+1} - s_h^k\big|^2,$$

we readily obtain

$$E^h_{\mathrm{uni-m}}[s_h^{k+1}, \boldsymbol{\Theta}_h^{k+1}] - E^h_{\mathrm{uni-m}}[s_h^k, \boldsymbol{\Theta}_h^{k+1}] \leq \delta_{s_h} E^h_{\mathrm{uni-m}}[s_h^{k+1}, \boldsymbol{\Theta}_h^{k+1}; s_h^{k+1} - s_h^k].$$

In addition, applying Lemma 11 leads to

$$E^h_{\mathrm{LdG,bulk}}[s_h^{k+1}] - E^h_{\mathrm{LdG,bulk}}[s_h^k] \leq \delta_{s_h} E^h_{\mathrm{LdG,bulk}}[s_h^{k+1}; s_h^{k+1} - s_h^k],$$

and together with the previous inequality implies

$$E^h_{\mathrm{uni-t}}[s_h^{k+1}, \boldsymbol{\Theta}_h^{k+1}] - E^h_{\mathrm{uni-t}}[s_h^k, \boldsymbol{\Theta}_h^{k+1}] \leq \delta_s E^h_{\mathrm{uni-t}}[s_h^{k+1}, \boldsymbol{\Theta}_h^{k+1}; s_h^{k+1} - s_h^k] = -\frac{1}{\tau}\|s_h^{k+1} - s_h^k\|^2_{L^2(\Omega)}.$$

Adding this expression to (53) yields the desired estimate and completes the proof. □

Remark 8 (CFL condition) The stability constraint $\tau \leq C E_{\mathrm{uni-t}}[s_h^0, \boldsymbol{\Theta}_h^0]^{-1/2}h^{d/2}$ is due to the weighted $H^1_\omega(\Omega)$ norm and the use of an inverse estimate between $L^\infty(\Omega)$ and $L^2(\Omega)$. If the weight $\omega = (s_h^k)^2$ is bounded away from zero, then the CFL condition is milder, namely $\tau \leq C E_{\mathrm{uni-t}}[s_h^0, \boldsymbol{\Theta}_h^0]^{-1/2}h^{d/2-1}|\log h|$ [12]. The weight ω is critical because it accelerates the algorithm upon allowing large variations of Θ_h^k near defects where it becomes small.

Remark 9 (convergence to stationary points) Even though the discrete gradient flow algorithm described in this section is energy-decreasing, it is not guaranteed to reach a global minimizer of $E^h_{\mathrm{uni-t}}$. The discrete equilibrium configuration may be a local minimizer but is very unlikely to be a saddle point. This is because the latter would not be stable under small perturbations induced by round-off errors.

7 Numerical experiments

To illustrate our method, we present computational experiments carried out with the MATLAB/C++ toolbox FELICITY [62]. We first consider a problem for the Landau - de Gennes energy with orientable Dirichlet boundary conditions. In such a case, the

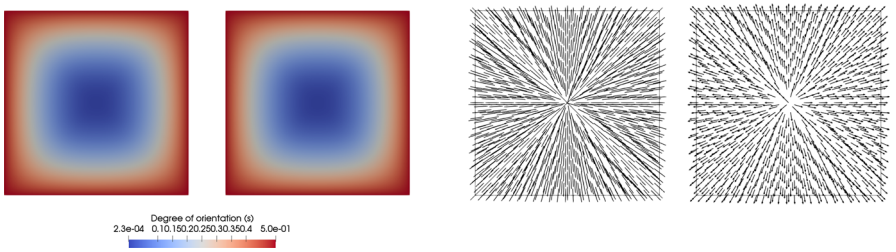

Fig. 2 Minimizing configurations for the Landau-de Gennes and Ericksen energies in 2-D for the setting discussed in Sect. 7.1. Left: degree of orientation for both models (left is uniaxial Landau-de Gennes, right is Ericksen). Right: line field $\boldsymbol{\Theta}$ (left) and director field $\mathbf{n}$ (right) are displayed. In this case, the line field is orientable, so both the Ericksen model and uniaxially constrained model give the same result

resulting line field of degree $+1$ is orientable, and the energy minimization problem is equivalent to the one given by minimizing the Ericksen energy; this allows us to compare with [47]. Afterwards, we illustrate the method's ability to capture non-orientable defects of degree $+1/2$ in two and three dimensional experiments, the latter leading to a non-straight line defect. We conclude with a Saturn-ring defect of degree $-1/2$ around a colloidal spherical inclusion. In all our experiments, meshes were taken to be weakly acute (cf. (20)).

7.1 Ericksen versus Landau de Gennes

It is known that, if the line field is orientable, then a director field representation is equivalent. Thus, we compare the solutions for the Ericksen and the Landau-de Gennes model with orientable boundary conditions. In this first experiment we are *not taking into account the double-well potential.* If $\boldsymbol{\Theta} = \mathbf{m} \otimes \mathbf{m}$ is an orientable line field, then a straightforward calculation gives $|\nabla \boldsymbol{\Theta}|^2 = 2|\nabla \mathbf{m}|^2$, and therefore

$$E_{\text{uni}-\text{m}}[s, \boldsymbol{\Theta}] = \frac{d-1}{2d} \int_\Omega |\nabla s|^2 dx + \int_\Omega s^2 |\nabla \mathbf{m}|^2 dx = 2E_{\text{erk}-\text{m}}[s, \mathbf{m}],$$

where the Ericksen energy corresponds to $\kappa = \frac{d-1}{2d}$.

We consider $\Omega = (0, 1)^2$, and impose the Dirichlet boundary conditions on $\partial\Omega$:

$$s = \frac{1}{2}, \quad \mathbf{n} = \frac{(x, y) - (1/2, 1/2)}{|(x, y) - (1/2, 1/2)|}, \quad \boldsymbol{\Theta} = \mathbf{n} \otimes \mathbf{n},$$

and compare the minimizers of the discrete energies $E^h_{\text{erk}-\text{m}}$ (with $\kappa = \frac{1}{4}$) and $E^h_{\text{uni}-\text{m}}$. We initialize both gradient flows with $s = 1/2$ and a point defect away from the center. Figure 2 shows the equilibrium configurations for both models. For the solutions displayed, we computed $E^h_{\text{uni}-\text{m}}[s_{h,LdG}, \boldsymbol{\Theta}] = E^h_{\text{erk}-\text{m}}[s_{h,Erk}, \mathbf{n}] \approx 1.234$, although $\min(s_{h,LdG}) \approx 2.3 \times 10^{-4}$ while $\min(s_{h,Erk}) \approx 5.8 \times 10^{-5}$.

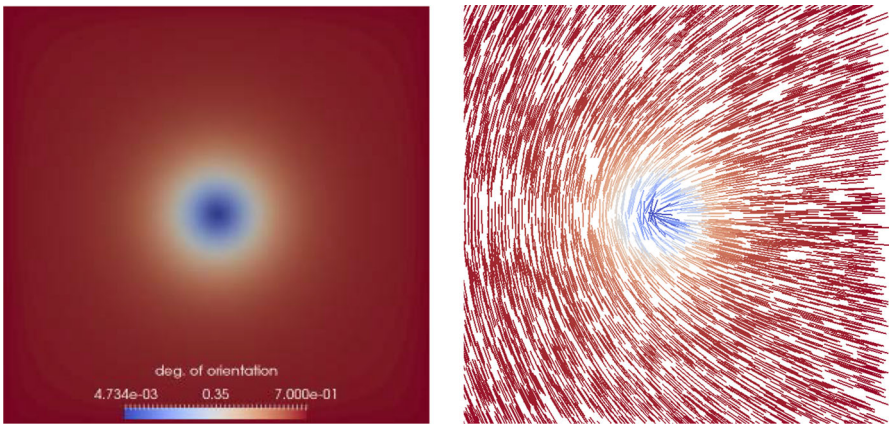

Fig. 3 A $+1/2$ degree point defect in 2-D (Sect. 7.2). Left: the degree-of-orientation s is plotted with the singular region at the center. Right: the line field $\boldsymbol{\Theta}$ is plotted and colored based on s. The time step for the gradient flow was $\tau = 10^{-2}$. This configuration cannot be captured by the Ericksen (director) model

7.2 Non-orientable field in two dimensions

Next, we simulate a non-orientable defect in the unit square $\Omega = (0, 1)^2$. We set the double-well potential with a convex splitting

$$\begin{aligned}\psi_{\mathrm{LdG}}(s) &= \psi_c(s) - \psi_e(s)\\ &:= (26.20577s^2 + 1) - (-4.1649313s^4 + 30.2874s^2),\end{aligned}$$

with $\eta_{\mathrm{B}} = 1/16$, and note that ψ_{LdG} has a local maximum at $s = 0$ and a global minimum at $s = s^* := 0.7$ with $\psi_{\mathrm{LdG}}(s^*) = 0$ (by symmetry in two dimensions, $\psi_{\mathrm{LdG}}(-s^*) = 0$). We impose Dirichlet boundary conditions for both s and $\boldsymbol{\Theta}$ on $\Gamma_s = \Gamma_{\boldsymbol{\Theta}} = \partial\Omega$,

$$\begin{aligned}s = s^*, \quad \mathbf{n}(x, y) = (\cos\theta, \sin\theta), \quad \boldsymbol{\Theta} = \mathbf{n} \otimes \mathbf{n},\\ \theta(x, y) = \frac{1}{2}\mathrm{atan2}\left(\frac{y - 1/2}{x - 1/2}\right),\end{aligned} \tag{54}$$

where atan2 is the four-quadrant inverse tangent function, i.e. the boundary conditions for $\boldsymbol{\Theta}$ correspond to a $+1/2$ degree defect centered at $(0.5, 0.5)$. We initialize the gradient flow with $s = s^*$ and $\boldsymbol{\Theta}$ corresponding to a $+1/2$ degree defect located at $(0.7167, 0.2912)$, which has initial energy $E^h_{\mathrm{uni}}[s_h, \boldsymbol{\Theta}_h] = 18.5468$. We show the final equilibrium configurations of s and the tensor field $\boldsymbol{\Theta}$ in Fig. 3. The method clearly captures the non-orientable defect at the domain center. The final state has $E^h_{\mathrm{uni}}[s_h, \boldsymbol{\Theta}_h] = 2.1192$ and $\min(s_h) \approx 4.734 \times 10^{-3}$.

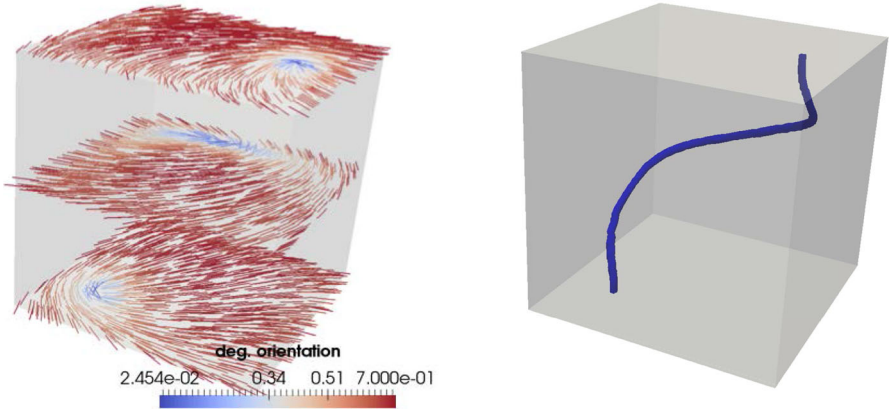

Fig. 4 A $+1/2$ degree line defect in a 3-D cube domain (Sect. 7.3). Left: line field $\boldsymbol{\Theta}$ is shown at levels $z = 0.0, 0.5, 1.0$ (colored by s). Right: The $s = 0.05$ iso-surface is shown that contains the line defect. In each horizontal plane, the line field exhibits a $+1/2$ degree *point* defect in 2-D. The twisting of the line defect is due to the choice of boundary conditions

7.3 Line defect in three dimensions

We simulate a non-orientable line defect in the unit cube $(0, 1)^3$. The double-well potential with a convex splitting is given by

$$\begin{aligned}\psi_{\mathrm{LdG}}(s) &= \psi_c(s) - \psi_e(s)\\ &:= (36.7709s^2 + 1) - (-7.39101s^4 + 4.51673s^3 + 39.27161s^2),\end{aligned}$$

with $\eta_{\mathrm{B}} = 1/16$, and note that ψ_{LdG} has a local maximum at $s = 0$ and a global minimum at $s = s^* := 0.700005531$ with $\psi_{\mathrm{LdG}}(s^*) = 0$.

The boundary conditions for $\boldsymbol{\Theta}$ were constructed in the following way. Let $\theta_0(x, y)$ define a $+1/2$ degree defect in the plane, located at $(0.3, 0.3)$ similar to (54). Likewise, let $\theta_1(x, y)$ define a $+1/2$ degree defect in the plane, located at $(0.7, 0.7)$. Next, define the Dirichlet boundary $\Gamma_s = \Gamma_{\boldsymbol{\Theta}} = \partial\Omega \setminus \Gamma_o$, where $\Gamma_o := \overline{\Omega} \cap (\{z = 0\} \cup \{z = 1\})$. Then, the Dirichlet conditions are

$$\begin{aligned}s = s^*, \quad \mathbf{n}(x, y) = (\cos\theta, \sin\theta, 0), \quad \boldsymbol{\Theta} = \mathbf{n} \otimes \mathbf{n},\\ \theta(x, y, z) = (1 - z)\theta_0(x, y) + z\theta_1(x, y) + \pi z,\end{aligned}$$

with vanishing Neumann condition on Γ_o. Basically, the boundary conditions consist of rotating a planar $+1/2$ degree point defect as a function of z. The solution is computed with the gradient flow approach (45) and time step $\tau = 10^{-3}$, and initialized with

$$s = s^*, \quad \mathbf{n} = (\cos\alpha, \sin\alpha, 0), \quad \boldsymbol{\Theta} = \mathbf{n} \otimes \mathbf{n}, \quad \alpha(x, y, z) = \theta_2(x, y) + \pi z,$$

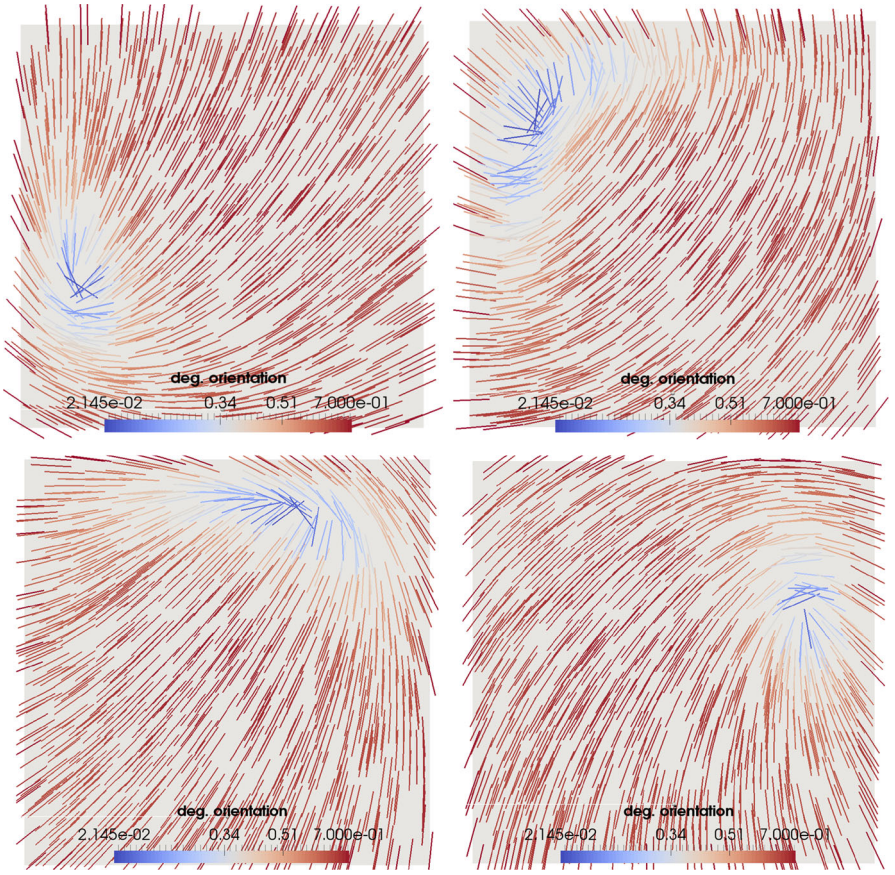

Fig. 5 Horizontal slices of the $+1/2$ degree line defect in a 3-D cube domain shown in Fig. 4 (Sect. 7.3). Top: left is $z = 0.2$, right is $z = 0.4$. Bottom: left is $z = 0.6$, right is $z = 0.8$. The location of the point defect in each plane rotates with the boundary conditions

where $\theta_2(x, y)$ corresponds to a $+1/2$ degree defect centered at $(0.5, 0.5)$; this configuration has an initial energy of $E^h_{\text{uni}}[s_h, \mathbf{\Theta}_h] = 10.013214$.

Figure 4 shows three dimensional views of the minimizing configuration, where as Fig. 5 shows four horizontal slices of the solution. A non-orientable line defect is observed, with final energy $E^h_{\text{uni}}[s_h, \mathbf{\Theta}_h] = 5.2042593769$ and $\min(s_h) \approx 2.145 \times 10^{-2}$.

7.4 Saturn-ring defect

Next, we simulate the Saturn-ring defect [2,32] using the double well potential from Sect. 7.3 with $\eta_{\text{B}} = 0.09$. The domain Ω is a "prism" type of cylindrical domain with square cross-section $[-0.25\sqrt{2}, 0.75\sqrt{2}]^2$, is centered about the $z = 0$ plane, and has

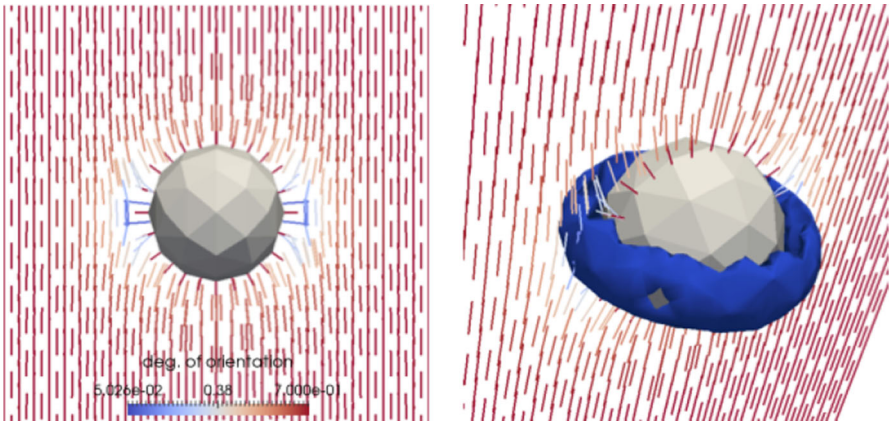

Fig. 6 Saturn-ring defect in 3-D (Sect. 7.4). Left: the line field is plotted with color scale based on the degree-of-orientation s; the $-1/2$ degree defect is visible on the left and right sides of the spherical inclusion (discretized sphere). Right: a view of the $s = 0.25$ isosurface in blue that contains the ring defect. The time step for the gradient flow was $\tau = 10^{-3}$. The configuration is symmetric about the vertical axis. Away from the sphere, the solution is $s = 0.7$ and $\mathbf{\Theta} = (0, 0, 1) \otimes (0, 0, 1)$

height 6. The domain contains a spherical inclusion, with boundary Γ_i, centered at $(\sqrt{2}/4, \sqrt{2}/4, 0)$ with radius $0.283/\sqrt{2}$. See [48, Sec. 5.1.1] for a precise definition.

We use the following Dirichlet boundary conditions on $\Gamma_s = \Gamma_{\mathbf{\Theta}} = \partial\Omega$,

$$\mathbf{n} = \boldsymbol{\nu}, \text{ on } \Gamma_i, \quad \mathbf{n} = (0, 0, 1)^T, \text{ on } \Gamma_o, \quad \mathbf{\Theta} = \mathbf{n} \otimes \mathbf{n}, \text{ on } \partial\Omega, \quad s = s^*, \text{ on } \partial\Omega,$$

where Γ_o is the outer boundary of Ω, $\boldsymbol{\nu}$ is the outer normal vector of the spherical inclusion, and s^* is the global minimum of the double well potential ψ. The initial conditions in Ω for the gradient flow are: $s = s^*$ and $\mathbf{n} = (0, 0, 1)^T$, which have initial energy $E^h_{\text{uni}}[s_h, \mathbf{\Theta}_h] = 7.59906$.

We show the final equilibrium configurations of s and the tensor field $\mathbf{\Theta}$ in Fig. 6. A cross-section of the solution is shown that illustrates the $-1/2$ degree nature of the Saturn-ring defect (note: the defect set of a ring about the equator of the inclusion). The final state has $E^h_{\text{uni}}[s_h, \mathbf{\Theta}_h] = 2.98004$ and $\min(s_h) \approx 5.026 \times 10^{-2}$. In contrast to our previous experiments using the Ericksen model [48], this new simulation is consistent with the physics of liquid crystals [2,32].

8 Conclusions

We introduced a structure-preserving finite element method for a uniaxially-constrained $\mathbf{Q}$-tensor model of nematic liquid crystals. In such a model, the energy is a degenerate functional of a tensor that must satisfy a rank-one constraint a.e. in the physical domain. We proved the Γ-convergence of the discrete energies as the mesh size tends to zero and developed an energy-decreasing gradient flow algorithm for the computation of discrete solutions. The numerical experiments show that this method is capable of capturing high-dimensional and non-orientable defect structures.

Acknowledgements The authors thank Wenbo Li for pointing out reference [7] and suggesting an idea for the proof of Lemma 4.

References

1. Adler, J.H., Atherton, T.J., Emerson, D.B., MacLachlan, S.P.: An energy-minimization finite-element approach for the Frank–Oseen model of nematic liquid crystals. SIAM J. Numer. Anal. **53**(5), 2226–2254 (2015)
2. Alama, S., Bronsard, L., Lamy, X.: Analytical description of the Saturn-ring defect in nematic colloids. Phys. Rev. E **93**, 012705 (2016)
3. Alouges, F.: A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case. SIAM J. Numer. Anal. **34**(5), 1708–1726 (1997)
4. Ambrosio, L.: Existence of minimal energy configurations of nematic liquid crystals with variable degree of orientation. Manuscripta Math. **68**(1), 215–228 (1990)
5. Araki, T., Tanaka, H.: Colloidal aggregation in a nematic liquid crystal: topological arrest of particles by a single-stroke disclination line. Phys. Rev. Lett. **97**, 127801 (2006)
6. Bajc, I., Hecht, F., Žumer, S.: A mesh adaptivity scheme on the Landau-de Gennes functional minimization case in 3d, and its driving efficiency. J. Comput. Phys. **321**, 981–996 (2016)
7. Balan, R., Zou, D.: On Lipschitz analysis and Lipschitz synthesis for the phase retrieval problem. Linear Algebra Appl. **496**, 152–181 (2016)
8. Ball, J.M., Zarnescu, A.: Orientable and non-orientable director fields for liquid crystals. Proc. Appl. Math. Mech. (PAMM) **7**(1), 1050701–1050704 (2007)
9. Ball, J.M., Zarnescu, A.: Orientability and energy minimization in liquid crystal models. Arch. Ration. Mech. Anal. **202**(2), 493–535 (2011)
10. Barrett, J.W., Feng, X., Prohl, A.: Convergence of a fully discrete finite element method for a degenerate parabolic system modelling nematic liquid crystals with variable degree of orientation. M2AN Math. Model. Numer. Anal. **40**, 175–199 (2006)
11. Bartels, S.: Numerical analysis of a finite element scheme for the approximation of harmonic maps into surfaces. Math. Comput. **79**(271), 1263–1301 (2010)
12. Bartels, S., Raisch, A.: Simulation of Q-tensor fields with constant orientational order parameter in the theory of uniaxial nematic liquid crystals. In: Griebel, M. (ed.) Singular Phenomena and Scaling in Mathematical Models, pp. 383–412. Springer, Berlin (2014)
13. Bhatia, R.: Matrix Analysis, Volume 169 of Graduate Texts in Mathematics. Springer, New York (1997)
14. Borthagaray, J.P., Walker, S.W.: The **Q**-tensor Model with Uniaxial Constraint. *ArXiv e-prints* (2020)
15. Braides, A.: Γ-Convergence for Beginners, Volume 22 of Oxford Lecture Series in Mathematics and Its Applications. Oxford Scholarship, Oxford (2002)
16. Braides, A.: Local Minimization, Variational Evolution and Γ-Convergence. Lecture Notes in Mathematics, vol. 2094. Springer, Berlin (2014)
17. Brinkman, W.F., Cladis, P.E.: Defects in liquid crystals. Phys. Today **35**, 48–56 (1982)
18. Ciarlet, P.G., Raviart, P.-A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Eng. **2**(1), 17–31 (1973)
19. Cohen, R., Lin, S.-Y., Luskin, M.: Relaxation and gradient methods for molecular orientation in liquid crystals. Comput. Phys. Commun. **53**(1–3), 455–465 (1989)
20. Cruz, P.A., Tomé, M.F., Stewart, I.W., McKee, S.: Numerical solution of the Ericksen–Leslie dynamic equations for two-dimensional nematic liquid crystal flows. J. Comput. Phys. **247**, 109–136 (2013)
21. Dal Maso, G.: An introduction to Γ-Convergence. Progress in Nonlinear Differential Equations and their Applications, vol. 8. Birkhäuser Boston, Inc., Boston (1993)
22. Davis, T., Gartland, E.C.: Finite element analysis of the Landau-de Gennes minimization problem for liquid crystals. SIAM J. Numer. Anal. **35**(1), 336–362 (1998)
23. de Gennes, P.G., Prost, J.: The Physics of Liquid Crystals: International Series of Monographs on Physics, vol. 83, 2nd edn. Oxford Science Publication, Oxford (1995)
24. Ericksen, J.L.: Liquid crystals with variable degree of orientation. Arch. Ration. Mech. Anal. **113**(2), 97–120 (1991)
25. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (1998)

26. Evans, L.C., Gariepy, R.F.: Measure Theory and Fine Properties of Functions. Textbooks in Mathematics, revised edn. CRC Press, Boca Raton (2015)
27. Freiser, M.J.: Ordered states of a nematic liquid. Phys. Rev. Lett. **24**(19), 1041 (1970)
28. Gartland, E.C.: Scalings and limits of Landau-de Gennes models for liquid crystals: a comment on some recent analytical papers. Math. Model. Anal. **23**(3), 414–432 (2018)
29. Gartland, E.C., Palffy-Muhoray, P., Varga, R.S.: Numerical minimization of the Landau-de Gennes free energy: defects in cylindrical capillaries. Mol. Cryst. Liq. Cryst. **199**(1), 429–452 (1991)
30. Gartland, E.C., Ramage, A.: A renormalized Newton method for liquid crystal director modeling. SIAM J. Numer. Anal. **53**(1), 251–278 (2015)
31. Gramsbergen, E.F., Longa, L., de Jeu, W.H.: Landau theory of the nematic-isotropic phase transition. Phys. Rep. **135**(4), 195–257 (1986)
32. Gu, Y., Abbott, N.L.: Observation of saturn-ring defects around solid microspheres in nematic liquid crystals. Phys. Rev. Lett. **85**, 4719–4722 (2000)
33. Guillén-González, F.M., Gutiérrez-Santacreu, J.V.: A linear mixed finite element scheme for a nematic Ericksen–Leslie liquid crystal model. M2AN Math. Model. Numer. Anal. **47**, 1433–1464 (2013)
34. Holzapfel, G.A.: Nonlinear Solid Mechanics: A Continuum Approach For Engineering. Wiley, Hoboken (2000)
35. James, R., Willman, E., FernandezFernandez, F.A., Day, S.E.: Finite-element modeling of liquid–crystal hydrodynamics with a variable degree of order. IEEE Trans. Electron Devices **53**(7), 1575–1582 (2006)
36. Kim, Y.-K., Shiyanovskii, S.V., Lavrentovich, O.D.: Morphogenesis of defects and tactoids during isotropic-nematic phase transition in self-assembled lyotropic chromonic liquid crystals. J. Phys.: Condens. Matter **25**(40), 404202 (2013)
37. Kohn, R.V., Sternberg, P.: Local minimisers and singular perturbations. Proc. Roy. Soc. Edinb. Sect. A **111**(1–2), 69–84 (1989)
38. Lamy, X.: A New Light on the Breaking of Uniaxial Symmetry in Nematics. arXiv:1307.0295 (2013)
39. Lee, G.-D., Anderson, J., Bos, P.J.: Fast Q-tensor method for modeling liquid crystal director configurations with defects. Appl. Phys. Lett. **81**(21), 3951–3953 (2002)
40. Lin, F.H.: On nematic liquid crystals with variable degree of orientation. Commun. Pure Appl. Math. **44**(4), 453–468 (1991)
41. Lin, S.-Y., Luskin, M.: Relaxation methods for liquid crystal problems. SIAM J. Numer. Anal. **26**(6), 1310–1324 (1989)
42. Liu, C., Walkington, N.: Approximation of liquid crystal flows. SIAM J. Numer. Anal. **37**(3), 725–741 (2000)
43. Madsen, L.A., Dingemans, T.J., Nakata, M., Samulski, E.T.: Thermotropic biaxial nematic liquid crystals. Phys. Rev. Lett. **92**, 145505 (2004)
44. Majumdar, Apala: Equilibrium order parameters of nematic liquid crystals in the landau-de gennes theory. Eur. J. Appl. Math. **21**(2), 181–203 (2010)
45. Mottram, N.J., Newton, C.J.P.: Introduction to Q-Tensor Theory. *ArXiv e-prints* (2014)
46. Nochetto, R.H., Walker, S.W., Zhang, W.: Numerics for liquid crystals with variable degree of orientation. In *Symposium NN - Mathematical and Computational Aspects of Materials Science*, volume 1753 of *MRS Proceedings* (2015)
47. Nochetto, R.H., Walker, S.W., Zhang, W.: A finite element method for nematic liquid crystals with variable degree of orientation. SIAM J. Numer. Anal. **55**(3), 1357–1386 (2017)
48. Nochetto, R.H., Walker, S.W., Zhang, W.: The Ericksen model of liquid crystals with colloidal and electric effects. J. Comput. Phys. **352**, 568–601 (2018)
49. Ohzono, T., Katoh, K., Wang, C., Fukazawa, A., Yamaguchi, S., Fukuda, J.: Uncovering different states of topological defects in schlieren textures of a nematic liquid crystal. Sci. Rep. **7**(1), 16814 (2017)
50. Palffy-Muhoray, P., Gartland, E.C., Kelly, J.R.: A new configurational transition in inhomogeneous nematics. Liq. Cryst. **16**(4), 713–718 (1994)
51. Prasad, V., Kang, S.-W., Suresh, K.A., Joshi, L., Wang, Q., Kumar, S.: Thermotropic uniaxial and biaxial nematic and smectic phases in bent-core mesogens. J. Am. Chem. Soc. **127**(49), 17224–17227 (2005)
52. Ravnik, M., Žumer, S.: Landau-deGennes modelling of nematic liquid crystal colloids. Liquid Cryst. **36**(10–11), 1201–1214 (2009)
53. Schopohl, N., Sluckin, T.J.: Defect core structure in nematic liquid crystals. Phys. Rev. Lett. **59**(22), 2582 (1987)

54. Shen, J., Yang, X.: A phase-field model and its numerical approximation for two-phase incompressible flows with different densities and viscosities. SIAM J. Sci. Comput. **32**(3), 1159–1179 (2010)
55. Sonnet, A., Kilian, A., Hess, S.: Alignment tensor versus director: description of defects in nematic liquid crystals. Phys. Rev. E **52**, 718–722 (1995)
56. Sonnet, A.M., Virga, E.: Dissipative Ordered Fluids: Theories for Liquid Crystals. Springer, Berlin (2012)
57. Strang, G., Fix, G.: An Analysis of the Finite Element Method, 2nd edn. Wellesley-Cambridge, Cambridge (2008)
58. Temam, R.M., Miranville, A.M.: Mathematical Modeling in Continuum Mechanics, 2nd edn. Cambridge University Press, Cambridge (2005)
59. Tojo, K., Furukawa, A., Araki, T., Onuki, A.: Defect structures in nematic liquid crystals around charged particles. Eur. Phys. J. E **30**(1), 55–64 (2009)
60. Truesdell, C.A.: A First Course in Rational Continuum Mechanics: Pure and Applied Mathematics, A Series of Monographs and Textbooks. Academic Press, Cambridge (1976)
61. Virga, E.G.: Variational Theories for Liquid Crystals, vol. 8, 1st edn. Chapman and Hall, London (1994)
62. Walker, S.W.: FELICITY: A Matlab/C++ toolbox for developing finite element methods and simulation modeling. SIAM J. Sci. Comput. **40**(2), C234–C257 (2018)
63. Walkington, N.J.: Numerical approximation of nematic liquid crystal flows governed by the Ericksen–Leslie equations. M2AN Math. Model. Numer. Anal. **45**, 523–540 (2011)
64. Wise, S.M., Wang, C., Lowengrub, J.S.: An energy-stable and convergent finite-difference scheme for the phase field crystal equation. SIAM J. Numer. Anal. **47**(3), 2269–2288 (2009)
65. Yu, L.J., Saupe, A.: Observation of a biaxial nematic phase in potassium laurate-1-decanol-water mixtures. Phys. Rev. Lett. **45**(12), 1000 (1980)
66. Zhao, J., Wang, Q.: Semi-discrete energy-stable schemes for a tensor-based hydrodynamic model of nematic liquid crystal flows. J. Sci. Comput. **68**(3), 1241–1266 (2016)
67. Zhao, J., Yang, X., Shen, J., Wang, Q.: A decoupled energy stable scheme for a hydrodynamic phase-field model of mixtures of nematic liquid crystals and viscous fluids. J. Comput. Phys. **305**, 539–556 (2016)

Affiliations

Juan Pablo Borthagaray[1,2] · Ricardo H. Nochetto[3] · Shawn W. Walker[4]

Ricardo H. Nochetto
rhn@umd.edu

Shawn W. Walker
walker@math.lsu.edu

1 Department of Mathematics, University of Maryland, College Park, MD 20742, USA

2 Departamento de Matemática y Estadística del Litoral, Universidad de la República, Salto, Uruguay

3 Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA

4 Department of Mathematics and Center for Computation and Technology (CCT) Louisiana State University, Baton Rouge, LA 70803, USA