ABSTRACT

Title of Dissertation: NUMERICAL ANALYSIS AND COMPUTATION OF NONLINEAR VARIATIONAL PROBLEMS IN MATERIALS SCIENCE

Shuo Yang
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Ricardo H. Nochetto
AMSC Program, Department of Mathematics

This dissertation focuses on the numerical analysis and scientific computation of two classes of nonlinear variational problems that originate from materials science: the large deformation of plates with metric constraint and constrained energy minimizations for nematic liquid crystals (LCs).

For the former, we design a local discontinuous Galerkin method (LDG) finite element approach for *prestrained* and *bilayer* plates, and the LDG hinges on the notion of reconstructed Hessian. We consider both Dirichlet and free boundary conditions, the former imposed on part of the boundary. In order to solve the ensuing discrete minimization problems subject to nonconvex metric constraints, we propose discrete gradient flow schemes. We prove Γ-convergence of the discrete energy to the continuous energy for each problem. Then we prove that the discrete gradient flow decreases the energy at each step and computes discrete minimizers with controllable discrete metric constraint violation. We present several insight-

ful numerical experiments for each problem, some of practical interest, and assess various computational aspects of the approximation process.

For LCs we focus on the one-constant Ericksen model that couples a director field with a scalar degree of orientation variable, and allows the formation of various defects with finite energy. We propose a simple but novel finite element approximation of the problem that can be implemented easily within standard finite element packages. Our scheme is projection-free and thus circumvents the use of weakly acute meshes, which are quite restrictive in 3d but are required by recent algorithms for convergence. We prove stability and Γ-convergence properties of the new FEM in the presence of defects. We also design an effective nested gradient flow algorithm for computing minimizers that in turn controls the violation of the unit-length constraint of the director. We present several simulations in 2d and 3d that document the performance of the proposed scheme and its ability to capture quite intriguing defects.

NUMERICAL ANALYSIS AND COMPUTATION OF NONLINEAR VARIATIONAL PROBLEMS IN MATERIALS SCIENCE

by

Shuo Yang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Ricardo H. Nochetto, Chair/Advisor
Professor P. S. Krishnaprasad, Dean's Representative
Professor Howard Elman
Professor Antonie Mellet
Associate Professor Maria Cameron
Associate Professor Tobias von Petersdorff

Table of Contents

List of Tables

List of Figures

List of Abbreviations

$\mathcal{N}_h$	Nodes of triangulation
$\mathbb{P}_k$	Space of polynomial functions of degree at most k
$\mathbb{Q}_k$	Space of polynomial functions of degree at most K in each variable
$\mathcal{T}_h$	Mesh with mesh-size h
Ω	Domain of the problem
DG	Discontinuous Galerkin
DGIP	Discontinuous Galerkin method with interior penalty
dofs	Degree of freedoms
FEM	Finite element method
LDG	Local discontinuous Galerkin
LC	Liquid crystals
PDE	Partial differential equation
SPD	Symmetric positive definite

Chapter 1: Introduction

The study of materials science has evolved to an interdisciplinary area, in which physicists, engineers and applied mathematicians share common interests. In the past several decades, in order to describe the physical essence of material behavior, there have been significant efforts in developing new mathematical models and analysis of materials. Meanwhile, the design of suitable numerical methods for these mathematical models becomes more challenging because they involve such diverse areas as the calculus of variations, convex analysis, optimization, numerical analysis of PDEs, and scientific computation. For example, variational problems with non- linear, non-convex constraints frequently appear in this area, and the nonlinearity usually makes the numerical analysis quite demanding.

This dissertation focuses on two classes of such problems: the large deformation of plates with metric constraints and the constrained energy minimization for nematic liquid crystals. The first part involves the study of *prestrained plates* and *bilayer plates*, which have rich applications in natural and manufactured phenomena such as nematic glasses [56, 57], natural growth of soft tissues [42, 80], manufactured polymer gels [51, 52, 79] and biomedical devices [45, 46, 66, 75]. The second part focuses on a new finite element method designed for the *Ericksen model of nematic*

LCs. It well-known that nematic LCs, a mesophase in between crystalline solid and isotropic liquid, represent a host of numerous potential applications in material science; we refer to [1, 5, 18].

1.1 A general framework

For all the problems considered in this dissertation, mathematical models are constrained minimization problems. In other words, we need to minimize an energy functional $E[\mathbf{y}]$ ($E[s, \boldsymbol{n}]$ respectively in chapter 4) within an admissible set $\mathbf{y} \in \mathbb{A}$ ($(s, \boldsymbol{n}) \in \mathbb{A}$ respectively). The admissible sets $\mathbb{A}$ contain nonconvex nonlinear constraints. Our goal is to design proper discrete energies $E_h[\mathbf{y}_h]$ ($E_h[s_h, \boldsymbol{n}_h]$ respectively) and discrete admissible sets $\mathbf{y}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}$ ($(s_h, \boldsymbol{n}_h) \in \mathbb{A}_{h,\boldsymbol{\epsilon}}$ respectively) that approximate the continuous energies E and admissible sets $\mathbb{A}$.

We prove the Γ-convergence of discrete energies E_h for each problem. In particular, we say $E_h[\mathbf{y}_h]$ Γ-converges to $E[\mathbf{y}]$ if the following conditions are valid:

- *Lim-inf condition*: Let $\{\mathbf{y}_h\} \subset \mathcal{A}_{h,\boldsymbol{\epsilon}}$ be a sequence such that $\mathbf{y}_h \to \mathbf{y}$ in L^2 up to a subsequence as $h, \boldsymbol{\epsilon} \to 0$, then

$$E[\mathbf{y}] \leq \liminf_{h \to 0} E_h[\mathbf{y}_h];$$

- *Lim-sup condition*: For every $\mathbf{y} \in \mathcal{A}$, there exists a *recovery sequence* $\{\mathbf{y}_h\} \subset$

$\mathcal{A}_{h,\boldsymbol{\epsilon}}$ converging to $\mathbf{y}$ in L^2 as $h, \boldsymbol{\epsilon} \to 0$ such that

$$E[\mathbf{y}] \geq \limsup_{h \to 0} E_h[\mathbf{y}_h].$$

Furthermore, in each chapter and in order to compute solutions to the discrete minimization problems, we design iterative schemes. In fact, we relax and linearize the nonconvex constraints at each step. For these schemes, we show the energy stability of discrete energies and prove a control of constraints violations.

1.2 Large deformation of plates with metric constraint

1.2.1 Prestrained plates

Prestrained materials can develop internal stresses at rest, deform out of plane without an external force and exhibit nontrivial 3d shapes. The derivation of dimensionally reduced models for plates is essential in elasticity, and one would express the equilibrium states of plates by 2d deformations of the mid-plane. Starting from the Saint Venant energy in 3d hyperelasticity, a geometrically nonlinear, dimensionally reduced energy for isotropic prestrained plates was derived rigorously via Γ-convergence in [17]. If $\mathbf{y} : \Omega \to \mathbb{R}^3$ is a 2d deformation of the plate Ω, $\mathrm{I}[\mathbf{y}]$ and $\mathrm{II}[\mathbf{y}]$ are the first and second fundamental forms of the deformed plate $\mathbf{y}(\Omega)$, the elastic bending energy reads

$$E[\mathbf{y}] = \frac{\mu}{12} \int_\Omega \left| g^{-\frac{1}{2}} \mathrm{II}[\mathbf{y}] g^{-\frac{1}{2}} \right|^2 + \frac{\lambda}{2\mu + \lambda} \mathrm{tr}\left(g^{-\frac{1}{2}} \mathrm{II}[\mathbf{y}] g^{-\frac{1}{2}} \right)^2, \tag{1.2.1}$$

and is subject to the metric constraint

$$\mathrm{I}[\mathbf{y}](\mathbf{x}) = g(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Omega, \tag{1.2.2}$$

where g is a given 2×2 symmetric positive definite matrix, λ and μ are Lamé parameters of the material. In the special case of $g = I_2$ (i.e. $\mathbf{y}$ is an isometry), a formal derivation of (1.2.1) can be traced back to Kirchhoff in 1850, and an ansatz-free rigorous derivation was carried out in the seminal work [39].

Our goal is to find $\mathbf{y} \in [H^2(\Omega)]^3$ that minimizes the bending energy (1.2.1) subject to the nonlinear, non-convex metric constraint (1.2.2). We conduct a formal asymptotic analysis under a modified Kirchhoff-Love assumption to reproduce (1.2.1) and (1.2.2). We next show an equivalent formulation that basically replaces $\mathrm{II}[\mathbf{y}]$ by the Hessian $D^2\mathbf{y}$ in (1.2.1), which makes the constrained minimization problem amenable to computation.

The case $g = I_2$ has already been discretized with Kirchhoff element [10] and discontinuous Galerkin methods with interior penalty (DGIP) [22] were used to discretize this problem. For prestrained plates, for which g is in general different from I_2, the second term of (1.2.1) cannot be absorbed into the first one, which makes the possible formulation of DGIP too complicated to use.

We introduce a local discontinuous Galerkin (LDG) approach for the discretization of the reduced energy (1.2.1), and to the best of our knowledge, our effort is the first FEM accompanying numerical analysis proposed on the prestrained plates. We summarize the significant ingredients and advantages of our discretiza-

tion as follows.

- *Discrete Hessian.* The fundamental idea of LDG is to replace the Hessian in (1.2.1) by a reconstructed discrete Hessian. Such discrete Hessian $H_h[\mathbf{y}_h]$ consists of three distinct parts: the broken Hessian $D_h^2\mathbf{y}_h$, the lifting $R_h([\nabla_h\mathbf{y}_h])$ of the jump of the broken gradient $\nabla_h\mathbf{y}_h$ of $\mathbf{y}_h$, and the lifting $B_h([\mathbf{y}_h])$ of the jumps of $\mathbf{y}_h$ itself. LDG method was originally introduced in [32]. Lifting operators were introduced in [15] and analyzed in [29, 30]. The definition of R_h and B_h is motivated by the liftings of [62, 63] leading to discrete gradient operators. Discrete Hessians were instrumental to study convergence of DG for the bi-Laplacian in [65] and plates with isometry constraint in [22]. In the present contribution, $H_h[\mathbf{y}_h]$ makes its debut as a chief constituent of the numerical method.

- *Linear solver.* We relax the pointwise metric constraint by using its integral on elements to make it computable. In order to compute minimizers, we propose a discrete H^2-gradient flow with linearized constraint to decrease the discrete energy while keeping the metric constraint defect under control.

- Γ-*convergence.* We prove the Γ-convergence of the discrete energy for both Dirichlet and free boundary conditions. We emphasize that the latter was never discussed in previous numerical analysis works for this type of problem, for instance [13, 22].

- *Initialization.* If an explicit expression of $\mathbf{y}$ that satisfies the metric constraint

(1.2.2) is known, then it is natural to take an interpolation of such a $\mathbf{y}$ as the initialization of the gradient flow. However, this is not readily accessible for a general g. Therefore, we propose a reasonable construction of an initialization that has a small violation of the metric constraint and a bounded discrete energy, which are provably important to control the constraint violation at the end of gradient flow.

- *Numerical experiments.* We present interesting numerical experiments to investigate the performance of the proposed method and the model capabilities for the cases with and without boundary conditions, and some of practical interest. Our simulations are done with the finite element library deal.ii [7].

- *Advantages of LDG.* Since Kirchhoff element and DGIP have been studied for this type of problem, we compare LDG with them. First, a DG method is more standard to implement and requires fewer polynomial degrees than Kirchhoff element. Second, in contrast to DGIP, the stabilization parameters for LDG must be positive for stability but not necessarily large. Last but not least, the formulation of LDG is conceptually simpler and its CPU time is less than DGIP.

The discussion in chapter 2 corresponds to the works presented in [19, 20].

1.2.2 Bilayer plates

Bilayer plates are made of two films with different material properties attached together. These layers react differently to non-mechanical stimuli, such as thermal,

electrical, chemical actuation [49, 50, 73]. Bilayer plates can develop large bending deformations without external force.

Bilayer plates can be modeled as thin 3d elastic bodies as in Fig. 1.1. For bilayer plates, a 2d model for the bending behavior of them has been rigorously derived and analyzed from 3d hyperelasticity in [67, 68]. A formal dimension reduction model allowing for various effects is presented in [13]. The 2d model as thickness $s \to 0$ consists of a nonlinear minimization problem with a nonconvex metric constraint. The resulting bending energy functional involves second order derivatives of deformation $\mathbf{y}$, while the constraint enforces deformations to be isometries, i.e. the first fundamental form of the deformed mid-surface equals to identity.

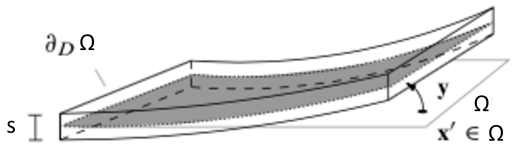

Figure 1.1: Bilayer plates: $\Omega \times (-s/2, s/2)$. $\Omega \subset \mathbb{R}^2$ is the mid-plane (bounded Lipschitz domain) and s is the thickness parameter. $\Omega \times (-s/2, 0)$ and $\Omega \times (0, s/2)$ represent the two layers of different materials.

Mathematically, the bilayer plates develop an intrinsic spontaneous curvature tensor Z and the deformation $\mathbf{y} : \Omega \to \mathbb{R}^3$ of the midplane $\Omega \subset \mathbb{R}^2$ minimizes the elastic energy

$$E[\mathbf{y}] = \frac{1}{2} \int_\Omega \left| \mathrm{I\!I}[\mathbf{y}] - Z \right|^2, \tag{1.2.3}$$

such that $\mathbf{y}$ is an isometry, i.e, $\mathbf{y}$ satisfies (1.2.2) with $g = I_2$. Expanding (1.2.3) and observing that $|\mathrm{I\!I}[\mathbf{y}]|^2 = |D^2\mathbf{y}|^2$ for isometries, we can expect that the nonlinear

term arising from the cross-product

$$\int_\Omega \mathrm{II}[\mathbf{y}] : Z = \sum_{ij=1}^{2} \int_\Omega \partial_{ij}\mathbf{y} \cdot (\partial_1\mathbf{y} \times \partial_2\mathbf{y}) Z_{ij} \tag{1.2.4}$$

is the most demanding part of the problem. This brings additional nonlinearities.

In [12, 13], a discretization based on Kirchhoff element is developed for the bilayer plates and its Γ-convergence is proved in [13]. The analysis of [13] requires a special definition of unit normal to the discrete plate, namely

$$\frac{\partial_1\mathbf{y}_h}{|\partial_1\mathbf{y}_h|} \times \frac{\partial_2\mathbf{y}_h}{|\partial_2\mathbf{y}_h|} \tag{1.2.5}$$

which turns out to complicate the numerical scheme and makes it highly nonlinear. An iterative scheme that decreases energy is proposed in [13], and in each step a fixed point sub-iteration is conducted to solve the discrete nonlinear equation. In contrary, a recent work [14] also considers the Kirchhoff element, but avoids the use of (1.2.5). It then has a fully practical gradient flow scheme that requires only solving linear systems in each step. Energy decreasing property of the new scheme and the Γ-convergence of the new formulation are proved, and the key new ingredient of analysis is an a priori $L^\infty(\Omega)$-bound for the first derivatives $\partial_i\mathbf{y}_h$. Moreover, a recent computational work [23] presents a DGIP approximation of the problem (1.2.3), and also proposes a fully practical scheme as in [14], but without supporting theory.

Motivated by [14, 23] and advantages of LDG indicated in Section 1.2.1, we

design a LDG approach for the problem (1.2.3). We summarize the originalities and strengths of our method as follows.

- *Reduced discrete Hessian.* We recall that the LDG method hinges on the reconstructed Hessian, but for the bilayer problem we further need a reduced discrete Hessian to guarantee the Γ-convergence theory. In fact, we construct the reduced discrete Hessian $\widetilde{H_h}[\mathbf{y}_h]$ by local L^2-projection of $H_h[\mathbf{y}_h]$ and replace (1.2.4) by the discrete counterpart

$$\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\big[(\widetilde{H_h}[\mathbf{y}_h]_{ij}\cdot(\partial_1\mathbf{y}_h\times\partial_2\mathbf{y}_h)Z_{ij}\big](x_T), \tag{1.2.6}$$

  where x_T is the barycenter of each element T.

- *Discrete isometry constraint.* We impose the discrete isometry constraint as

$$\big|\big[\nabla\mathbf{y}_h^T\nabla\mathbf{y}_h - I_2\big](x_T)\big| \le \delta(h) \tag{1.2.7}$$

  at barycenters for any element T with parameter $\delta(h)\to 0$ as $h\to 0$. Compared with [23], this is a novel way of imposing the metric constraint discretely in the DG context, and helps with the Γ-convergence theory.

- Γ-*convergence.* Other than the term (1.2.6) that is cubic in $\mathbf{y}_h$, all remaining terms in the discrete energy can be treated as in Section 1.2.1 using LDG. With this new discretization, we prove the Γ-convergence for the first time in the setting of DG for bilayer plates.

- *Linear solver.* In contrast to [12, 13], we design a fully linear and practical gradient flow scheme. Although the discrete constraint (1.2.7) is nonlinear and nonconvex, and the variational derivative of (1.2.6) in the discrete energy is nonlinear in $\mathbf{y}_h$, we linearize (1.2.7) and treat the variation of (1.2.6) explicitly at each step of iterations. We prove the energy stability with a condition on time-step and the control of constraint violation.

The discussion in chapter 3 corresponds to the work presented in [24].

1.3 Constrained energy minimization for nematic liquid crystals

We consider the one-constant Ericksen model for nematic LCs with variable degree of orientation [33, 37], which lies between the Oseen-Frank director model and the Landau - de Gennes Q-tensor model. The state of the LC is described by a director field $\boldsymbol{n}$ and a scalar function s, which satisfy the constraints $|\boldsymbol{n}| = 1$ and $-1/(d-1) < s < 1$ for the space dimension $d = 2, 3$. The director $\boldsymbol{n}$ indicates the preferred orientations of LC molecules, while s represents the degree of alignment that molecules have with respect to $\boldsymbol{n}$, both in the sense of local probabilistic average. The equilibrium state is given by an admissible pair $(s, \boldsymbol{n})$ that minimizes the Ericksen's energy

$$E[s, \boldsymbol{n}] = \frac{1}{2}\int_\Omega \left(\kappa|\nabla s|^2 + s^2|\boldsymbol{\nabla n}|^2\right) + \int_\Omega \psi(s). \tag{1.3.1}$$

under a unit length constraint $|\boldsymbol{n}| = 1$ and $\kappa > 0$ constant; the constraint on s is enforced by the double well potential ψ. We refer to [4, 54] for early analysis of the Ericksen model.

If s can be approximated by a non-vanishing constant, then the energy (1.3.1) reduces to the Oseen-Frank energy $E[\boldsymbol{n}] = \frac{\kappa}{2}\int_\Omega |\boldsymbol{\nabla n}|^2$, whose minimizers are harmonic maps and have been extensively studied, e.g., in [28, 70]. However, the simpler Oseen-Frank model has severe limitations in capturing defects: it only admits point defects with finite energy for $d = 3$. In contrast, the Ericksen model (1.3.1) allows for $\boldsymbol{n} \notin [H^1(\Omega)]^d$ and compensates blow up of $\nabla \boldsymbol{n}$ by letting s to vanish, which is the mechanism for the formation of a variety of line and surface defects for $d = 2, 3$. This physical process leads to a *degenerate* Euler-Lagrange equation for $\boldsymbol{n}$ that poses serious difficulties to formulate mathematically sound algorithms to approximate (1.3.1) and study their convergence.

Several numerical methods for the Oseen-Frank model have been proposed [3, 9, 55]. Finite element methods (FEMs) for the Ericksen model are designed in [8, 31, 59, 60, 77]; see also the recent review [25]. In contrast to [8], a fundamental structure of (1.3.1) is exploited in [59, 60] to design and analyze FEMs that handle the inherent degeneracy of (1.3.1) without regularization and enforce the constraint $|\boldsymbol{n}| = 1$ robustly. Stability and convergence properties via Γ-convergence are proved in [59, 60], pioneering results in this setting. They hinge on a clever discrete energy that mimics the structure of (1.3.1) discretely but, unfortunately, is cumbersome to implement in standard software packages and requires weakly acute meshes. The latter ensures that the projection of discrete director fields onto the unit sphere is

energy decreasing, and thus compatible with the quasi-gradient flow, but is quite restrictive and difficult to implement for $d = 3$ and domains with non-trivial topology.

We propose a projection-free FEM scheme that avoids dealing with weakly acute meshes. Without the projection step, the unit length constraint $|\boldsymbol{n}| = 1$ is no longer satisfied exactly but instead is relaxed at each step of our iterative solver, a nested gradient flow. The latter guarantees control of the violation of $|\boldsymbol{n}| = 1$ and asymptotic enforcement of it. We summarize the chief novelties and advantages of our approach as follows.

- *Shape regular meshes.* Partitions of Ω are assumed to be only shape regular, which allows for the use of software with general mesh generators such as Netgen [69]. Avoiding weakly acute meshes is important in 3d to deal with interesting but non-trivial geometries. An earlier work achieving this goal is [77], which presents a mass-lumped FEM with a consistent stabilization term involving $s^2 \boldsymbol{\nabla} \boldsymbol{n}^T \boldsymbol{n}$ for the generalized Ericksen energy.

- *Standard algorithm.* Our novel discretization of (1.3.1) is straightforward, requires no stabilization, and is easy to implement in standard software packages such as NGSolve [69]. In contrast to [59, 60], our FEM does no longer exploit the structure of (1.3.1) but its analysis does.

- *Linear solver.* We propose a nested gradient flow that, despite the nonlinear nature of the problem, is *fully linear* to compute minimizers. The inner loop to advance the director field $\boldsymbol{n}$ for fixed degree of orientation s is allowed

to subiterate. This turns out to induce an acceleration mechanism for the computation and motion of defects. For a recent acceleration techniques based on a domain decomposition approach, we refer to [31].

- Γ-*convergence.* The analysis of our FEM hinges heavily on the underlying structure of (1.3.1) and relies on the notion of L^2-gradient on $\boldsymbol{n}$ [38, Theorem 6.2]; see Proposition 4.1.1. Such a notion was already used in [26] in the context of the uniaxial Q-tensor LC model. We prove stability and Γ-convergence. Our results are similar to those in [59, 60, 77] but the way to the discrete structure is new.

- *Numerical experiments.* We present several simulations. Some are meant to compare the new algorithm with the existing literature in terms of performance and ability to capture defects. Other experiments explore 3d intriguing configurations such as the propeller defect and a configuration more challenging than the Saturn ring.

The discussion in chapter 4 corresponds to the work presented in [58].

1.4 Outline

Chapter 2 is concerned with the numerical treatment of the large deformation of prestrained plates. We start with a justification of a 3d elastic energy $E[\mathbf{u}]$ for prestrained plates followed by a formal derivation of (1.2.1) and (1.2.2) as the asymptotic limit of $s^{-3}E[\mathbf{u}]$ as $s \to 0$. Moreover, we show an equivalent formula-

tion that basically replaces the second fundamental form $\mathrm{II}[\mathbf{y}]$ by the Hessian $D^2\mathbf{y}$, which makes the constrained minimization problem amenable to computation.Then, we introduce the LDG type discretization, give the definition of the discrete Hessian $H_h[\mathbf{y}_h]$ and the discrete energy E_h, as well as preliminary key properties of the discrete functions, such as discrete Poincaré-Friedrich type inequalities. Later, we prove weak and strong convergence properties of $H_h[\mathbf{y}_h]$, and apply them to prove the Γ-convergence of E_h for both Dirichlet boundary condition case and *free boundary case*. Next, we introduce the gradient flow scheme used to solve the discrete problem, and prove the unconditional stability and control of violation of the metric constraint for it. We also discuss the effect of flow metric on the discrete solution in the *free boundary case*. Subsequently, we discuss the scheme that is designed for the initialization and illustrate its effectiveness. Finally, we present numerical simulations for prestrained plates and show performance of our algorithms.

We present the new numerical method for large deformation of bilayer plates in Chapter 3, following the LDG type discretization considered in Chapter 2, and especially the use of discrete Hessian $H_h[\mathbf{y}_h]$. We start by introducing the bilayer plates model and its simplification. Then, we present the corresponding discrete energy, and emphasize the novel reduced discrete Hessian $\widetilde{H}_h[\mathbf{y}_h]$ and the new way of imposition of the discrete admissible set. Afterwards, we prove the Γ-convergence of the discrete energy. Subsequently, we design an explicitly linearized, practical discrete gradient flow scheme, and prove its *conditional* energy stability and the control of the constraint. Eventually, we illustrate our method by showing several numerical examples.

Chapter 4 is devoted to the new FEM method for the Ericksen's model of the nematics liquid crystals. First, we describe the Ericksen model for LCs with variable degree of orientation and discuss its key structure. Second, we introduce our discretization of the model and discuss our Γ-convergence result. Then we present our iterative scheme for the computation of discrete local minimizers. Last but not least, we show numerical experiments illustrating effectiveness and efficiency of our method, as well as its flexibility to deal with complex defects in 3d.

Chapter 2: LDG Method of Large Deformations of Prestrained Plates

In this chapter we introduce the model of prestrained plates and the LDG discretization of it. We prove the Γ-convergence of the discrete energy, as well as the energy stability and control of constraint violation property of a discrete gradient flow. We explore the performance of the numerical scheme computationally with several insightful simulations.

2.1 Problem statement

We start by rederiving the 2d elastic energy (1.2.1) from 3d hyperelasticity.

Prestrained plates develop internal stresses, deform out of plane and exhibit nontrivial 3d shapes. A model postulates that these plates may reduce internal stresses by undergoing large out of plane deformations $\mathbf{u}$ as a means to minimize an elastic energy $E[\mathbf{u}]$ that measures the discrepancy between a reference (or target) metric G and the orientation preserving realization $\mathbf{u}$ of it.

Let $\Omega_s := \Omega \times (-s/2, s/2) \subset \mathbb{R}^3$ be a three-dimensional plate at rest, where $s > 0$ denotes the thickness and $\Omega \subset \mathbb{R}^2$ is the (flat) midplane. Given a Riemannian metric $G : \Omega_s \to \mathbb{R}^{3\times 3}$ (symmetric uniformly positive definite matrix), we consider

3d deformations $\mathbf{u} : \Omega_s \to \mathbb{R}^3$ driven by the strain tensor $\boldsymbol{\epsilon}_G(\nabla \mathbf{u})$ given by

$$\boldsymbol{\epsilon}_G(\nabla \mathbf{u}) := \frac{1}{2}\big(\nabla \mathbf{u}^T \nabla \mathbf{u} - G\big), \tag{2.1.1}$$

that measures the discrepancy between $\nabla \mathbf{u}^T \nabla \mathbf{u}$ and G; hence, the 3d elastic energy $E[\mathbf{u}] = 0$ whenever $\boldsymbol{\epsilon}_G(\nabla \mathbf{u}) = \mathbf{0}$. We say that G is the *reference (prestrained or target) metric.* An orientable deformation $\mathbf{u} : \Omega_s \to \mathbb{R}^3$ of class $H^2(\Omega_s)$ satisfying $\boldsymbol{\epsilon}_G(\nabla \mathbf{u}) = \mathbf{0}$ is called an *isometric immersion.* We assume that G does not depend on s and is uniform throughout the thickness, written as follows

$$G(\mathbf{x}', x_3) = G(\mathbf{x}') = \begin{bmatrix} g(\mathbf{x}') & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \quad \forall\, \mathbf{x}' \in \Omega,\ x_3 \in (-s/2, s/2), \tag{2.1.2}$$

with $g : \Omega \to \mathbb{R}^{2\times 2}$ symmetric uniformly positive definite [35, 53]. If $g^{1/2}$ denotes the square root of g, we have

$$G^{\frac{1}{2}} = \begin{bmatrix} g^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}, \quad G^{-\frac{1}{2}} = \begin{bmatrix} g^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{2.1.3}$$

We will use the following notation below. The i^{th} component of a vector $\mathbf{v} \in \mathbb{R}^n$ is denoted v_i while for a matrix $A \in \mathbb{R}^{n\times m}$, we write A_{ij} the coefficient of the i^{th} row and j^{th} column. The gradient of a scalar function is a column vector and for $\mathbf{v} : \mathbb{R}^m \to \mathbb{R}^n$, we set $(\nabla \mathbf{v})_{ij} := \partial_j v_i,\ i = 1, .., n,\ j = 1, ..., m.$ The Euclidean norm of a vector is denoted $|\cdot|$. For matrices $A, B \in \mathbb{R}^{n\times m}$, we write $A : B := \mathrm{tr}(B^T A) = \sum_{i=1}^n \sum_{j=1}^m A_{ij} B_{ij}$ and $|A| := \sqrt{A : A}$ the Frobenius norm of

A. To have a compact notation later, for higher-order tensors we set

$$\mathbf{A} = (A_k)_{k=1}^n \in \mathbb{R}^{n\times m\times m} \;\Rightarrow\; \operatorname{tr}(\mathbf{A}) = \big(\operatorname{tr}(A_k)\big)_{k=1}^n, \quad |\mathbf{A}| = \left(\sum_{k=1}^n |A_k|^2\right)^{\frac{1}{2}}. \tag{2.1.4}$$

Furthermore, we will frequently use the convention

$$B\mathbf{A}B := (BA_kB)_{k=1}^3 \in \mathbb{R}^{3\times 2\times 2}, \tag{2.1.5}$$

for $\mathbf{A} \in \mathbb{R}^{3\times 2\times 2}$ and $B \in \mathbb{R}^{2\times 2}$. In particular, for $\mathbf{y} : \mathbb{R}^2 \to \mathbb{R}^3$, we will often write

$$g^{-1/2}\, D^2\mathbf{y}\, g^{-1/2} = \big(g^{-1/2}\, D^2 y_k\, g^{-1/2}\big)_{k=1}^3, \tag{2.1.6}$$

which, combined with (2.1.4), yields

$$\begin{aligned} \big|g^{-1/2}\, D^2\mathbf{y}\, g^{-1/2}\big| &= \left(\sum_{k=1}^3 \big|g^{-1/2}\, D^2 y_k\, g^{-1/2}\big|^2\right)^{1/2}, \\ \operatorname{tr}\big(g^{-1/2}\, D^2\mathbf{y}\, g^{-1/2}\big) &= \big(\operatorname{tr}\big(g^{-1/2}\, D^2 y_k\, g^{-1/2}\big)\big)_{k=1}^3. \end{aligned} \tag{2.1.7}$$

Finally, I_n will denote the identity matrix in $\mathbb{R}^{n\times n}$.

2.1.1 Elastic energy for prestrained plates

We present, following [35], a simple derivation of the energy density $W(\nabla\mathbf{u}\, G^{-1})$ for prestrained materials. This hinges on the well-established theory of hyperelasticity, and reduces to the classical St. Venant-Kirchhoff model provided $G = I_3$.

Such model for isotropic materials reads

$$W(F) := \mu|\boldsymbol{\epsilon}_I|^2 + \frac{\lambda}{2}\mathrm{tr}(\boldsymbol{\epsilon}_I)^2, \quad \boldsymbol{\epsilon}_I(F) := \frac{1}{2}\left(F^TF - I_3\right). \tag{2.1.8}$$

Here, F is the deformation gradient, $\boldsymbol{\epsilon}_I$ is the Green-Lagrange strain tensor and λ and μ are the (first and second) Lamé constants. This implies

$$D^2W(I_3)(F,F) = 2\mu|e|^2 + \lambda\mathrm{tr}(e)^2, \quad e := \frac{F+F^T}{2}. \tag{2.1.9}$$

We point out that in [39], the strain tensor $\boldsymbol{\epsilon}_I = \boldsymbol{\epsilon}_I(F)$ of (2.1.8) is set to be $\boldsymbol{\epsilon}_I(F) = \sqrt{F^TF} - I_3$, which yields the same relation (2.1.9), and thus the same Γ-limit discussed below.

Given an arbitrary point $\mathbf{x}_0 \in \Omega_s$, we consider the linear transformation $\mathbf{r}_0(\mathbf{x}) := G^{1/2}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)$; hence $\nabla\mathbf{r}_0(\mathbf{x}) = G^{1/2}(\mathbf{x}_0)$. The map $\mathbf{r}_0$ can be viewed as a local re-parametrization of the deformed 3d elastic body, and $\mathbf{z} = \mathbf{r}_0(\mathbf{x})$ is a new local coordinate system. This induces the deformation $\mathbf{U}(\mathbf{z}) := \mathbf{u}(\mathbf{x})$ and

$$\mathbf{u} = \mathbf{U}\circ\mathbf{r}_0 \quad \Rightarrow \quad \nabla\mathbf{u}(\mathbf{x}) = \nabla_{\mathbf{z}}\mathbf{U}(\mathbf{z})\, G^{\frac{1}{2}}(\mathbf{x}_0),$$

where $\nabla_{\mathbf{z}}$ denotes the gradient with respect to the variable $\mathbf{z}$. The deviation of $\nabla\mathbf{u}^T\nabla\mathbf{u}$ from the reference metric G at $\mathbf{x} = \mathbf{x}_0$ is thus given by (2.1.1)

$$\boldsymbol{\epsilon}_G(\nabla\mathbf{u}) = \frac{1}{2}\big(\nabla\mathbf{u}^T\nabla\mathbf{u} - G\big) = \frac{1}{2}G^{\frac{1}{2}}\big(\nabla_{\mathbf{z}}\mathbf{U}^T\nabla_{\mathbf{z}}\mathbf{U} - I_3\big)G^{\frac{1}{2}} = G^{\frac{1}{2}}\boldsymbol{\epsilon}_I(\nabla_{\mathbf{z}}\mathbf{U})G^{\frac{1}{2}}.$$

The energy density $W(\nabla_{\mathbf{z}}\mathbf{U})$ at $\mathbf{z} = \mathbf{r}_0(\mathbf{x})$ with $\mathbf{x} = \mathbf{x}_0$ associated with $\boldsymbol{\epsilon}_I(\nabla_{\mathbf{z}}\mathbf{U})$, which minimizes when $\boldsymbol{\epsilon}_I(\nabla_{\mathbf{z}}\mathbf{U})$ vanishes, is governed by (2.1.8) for isotropic materials according to the theory of hyperelasticity. What we need to do now is to rewrite this energy density in terms of $\nabla\mathbf{u}$ at $\mathbf{x} = \mathbf{x}_0$, namely $W(\nabla_{\mathbf{z}}\mathbf{U}) = W(\nabla\mathbf{u}\, G^{-1/2})$, whence

$$W(\nabla\mathbf{u}\, G^{-1/2}) = \mu\Big|G^{-1/2}\,\boldsymbol{\epsilon}_G(\nabla\mathbf{u})\, G^{-1/2}\Big|^2 + \frac{\lambda}{2}\mathrm{tr}\Big(G^{-1/2}\,\boldsymbol{\epsilon}_G(\nabla\mathbf{u})\, G^{-1/2}\Big)^2. \quad (2.1.10)$$

This motivates the definition of hyperelastic energy for prestrained materials

$$E[\mathbf{u}] := \int_{\Omega_s} W\big(\nabla\mathbf{u}(\mathbf{x})G(\mathbf{x})^{-\frac{1}{2}}\big)d\mathbf{x} - \int_{\Omega_s}\mathbf{f}_s(\mathbf{x})\cdot\mathbf{u}(\mathbf{x})d\mathbf{x}, \quad (2.1.11)$$

where $\mathbf{f}_s : \Omega_s \to \mathbb{R}^3$ is a prescribed forcing term and W is given by (2.1.10).

Note that the pointwise decomposition $G(\mathbf{x}_0) = \nabla\mathbf{r}_0(\mathbf{x}_0)^T\nabla\mathbf{r}_0(\mathbf{x}_0)$ is always possible because $G(\mathbf{x}_0)$ is symmetric positive definite. However, a global transformation $\mathbf{r}$ such that $\nabla\mathbf{r}^T\nabla\mathbf{r} = G$ everywhere need not exist in general because G is not required to be immersible in $\mathbb{R}^3$. This is referred to as *incompatible elasticity* in [35]. Moreover, the infimum of $E[\mathbf{u}]$ in (2.1.11) should be strictly positive if the Riemann curvature tensor associated with G does not vanish identically [53].

2.1.2 Reduced model

It is well-known that the case $E[\mathbf{u}] \sim s$ corresponds to a stretching of the midplane Ω (membrane theory) while pure bending occurs when $E[\mathbf{u}] \sim s^3$ (bending

theory); see [40]. We examine now the formal asymptotic behavior of $s^{-3}E[\mathbf{u}]$ as $s \to 0$; see also [35].

We start with the assumption [17, 41, 53]

$$\mathbf{u}(\mathbf{x}) = \mathbf{y}(\mathbf{x}') + x_3\alpha(\mathbf{x}')\boldsymbol{\nu}(\mathbf{x}') + \frac{1}{2}x_3^2\beta(\mathbf{x}')\boldsymbol{\nu}(\mathbf{x}') \qquad \forall\, \mathbf{x}' \in \Omega,\ x_3 \in (-s/2, s/2), \tag{2.1.12}$$

where $\mathbf{y} : \Omega \to \mathbb{R}^3$ describes the deformation of the mid-surface of the plate, $\boldsymbol{\nu}(\mathbf{x}') := \frac{\partial_1\mathbf{y}(\mathbf{x}')\times\partial_2\mathbf{y}(\mathbf{x}')}{|\partial_1\mathbf{y}(\mathbf{x}')\times\partial_2\mathbf{y}(\mathbf{x}')|}$ is the unit normal vector to the surface $\mathbf{y}(\Omega)$ at the point $\mathbf{y}(\mathbf{x}')$, and $\alpha, \beta : \Omega \to \mathbb{R}$ are functions to be determined. Compared to the usual Kirchhoff-Love assumption

$$\mathbf{u}(\mathbf{x}', x_3) = \mathbf{y}(\mathbf{x}') + x_3\,\boldsymbol{\nu}(\mathbf{x}') \qquad \forall\, \mathbf{x}' \in \Omega,\ x_3 \in (-s/2, s/2), \tag{2.1.13}$$

(2.1.12) not only restricts fibers orthogonal to Ω to remain perpendicular to the surface $\mathbf{y}(\Omega)$ but also allows such fibers to be inhomogeneously stretched. We rescale the forcing term in (2.1.11) as follows

$$\mathbf{f}(\mathbf{x}') := \lim_{s\to 0^+} s^{-3}\int_{-s/2}^{s/2} \mathbf{f}_s(\mathbf{x}', x_3)\,dx_3 \qquad \forall\, \mathbf{x}' \in \Omega, \tag{2.1.14}$$

and assume the limit to be finite. However, for the asymptotics below we omit this term for simplicity from the derivation and focus on the energy density W in (2.1.11).

Denoting by ∇' the gradient with respect to $\mathbf{x}'$ and writing $\mathbf{b}(\mathbf{x}') := \alpha(\mathbf{x}')\boldsymbol{\nu}(\mathbf{x}')$

and $\mathbf{d}(\mathbf{x}') := \beta(\mathbf{x}')\boldsymbol{\nu}(\mathbf{x}')$, we have for all $\mathbf{x} = (\mathbf{x}', x_3) \in \Omega_s$

$$\nabla \mathbf{u}(\mathbf{x}) = \left[\nabla' \mathbf{y}(\mathbf{x}') + x_3 \nabla' \mathbf{b}(\mathbf{x}') + \frac{1}{2} x_3^2 \nabla' \mathbf{d}(\mathbf{x}'), \mathbf{b}(\mathbf{x}') + x_3 \mathbf{d}(\mathbf{x}') \right] \in \mathbb{R}^{3\times 3}.$$

Using the relations

$$\boldsymbol{\nu}^T \boldsymbol{\nu} = 1 \quad \text{and} \quad \boldsymbol{\nu}^T \nabla' \mathbf{y} = \boldsymbol{\nu}^T \nabla' \boldsymbol{\nu} = \mathbf{d}^T \nabla' \boldsymbol{\nu} = \mathbf{d}^T \nabla' \mathbf{y} = \mathbf{b}^T \nabla' \boldsymbol{\nu} = \mathbf{b}^T \nabla' \mathbf{y} = \mathbf{0},$$

we easily get

$$\begin{aligned} \nabla \mathbf{u}^T \nabla \mathbf{u} = & \begin{bmatrix} \nabla' \mathbf{y}^T \nabla' \mathbf{y} & \mathbf{0} \\ \mathbf{0} & \alpha^2 \end{bmatrix} + x_3 \begin{bmatrix} \nabla' \mathbf{y}^T \nabla' \mathbf{b} + \nabla' \mathbf{b}^T \nabla' \mathbf{y} & \nabla' \mathbf{b}^T \mathbf{b} \\ \mathbf{b}^T \nabla' \mathbf{b} & 2\alpha\beta \end{bmatrix} \\ & + x_3^2 \begin{bmatrix} \frac{1}{2}(\nabla' \mathbf{y}^T \nabla' \mathbf{d} + \nabla' \mathbf{d}^T \nabla' \mathbf{y}) + \nabla' \mathbf{b}^T \nabla' \mathbf{b} & \frac{1}{2} \nabla' \mathbf{d}^T \mathbf{b} + \nabla' \mathbf{b}^T \mathbf{d} \\ \frac{1}{2} \mathbf{b}^T \nabla' \mathbf{d} + \mathbf{d}^T \nabla' \mathbf{b} & \beta^2 \end{bmatrix} \\ & + h.o.t. \end{aligned}$$

Moreover, since

$$|\boldsymbol{\nu}|^2 = 1, \quad \partial_j \mathbf{b} = (\partial_j \alpha)\boldsymbol{\nu} + \alpha \partial_j \boldsymbol{\nu} \quad \text{and} \quad \boldsymbol{\nu} \cdot \partial_j \mathbf{y} = 0 \quad \text{for } j = 1, 2,$$

we have

$$\nabla' \mathbf{b}^T \nabla' \mathbf{y} = \alpha \nabla' \boldsymbol{\nu}^T \nabla' \mathbf{y} \quad \text{and} \quad \nabla' \mathbf{b}^T \mathbf{b} = \alpha \nabla' \alpha.$$

Therefore, the expression $2G^{-1/2}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-1/2}$ becomes

$$G^{-\frac{1}{2}}\nabla\mathbf{u}^T\nabla\mathbf{u}G^{-\frac{1}{2}} - I_3 = A_1 + 2x_3A_2 + x_3^2A_3 + \mathcal{O}(x_3^3),$$

where

$$A_1 := \begin{bmatrix} g^{-\frac{1}{2}}\,\mathrm{I}[\mathbf{y}]\,g^{-\frac{1}{2}} - I_2 & \mathbf{0} \\ \mathbf{0} & \alpha^2 - 1 \end{bmatrix},$$

$$A_2 := \begin{bmatrix} -\alpha g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}} & \frac{1}{2}\alpha g^{-\frac{1}{2}}\nabla'\alpha \\ \frac{1}{2}\alpha\nabla'\alpha^T g^{-\frac{1}{2}} & \alpha\beta \end{bmatrix},$$

$$A_3 := \begin{bmatrix} g^{-\frac{1}{2}}(\nabla'\mathbf{b}^T\nabla'\mathbf{b} + \frac{1}{2}(\nabla'\mathbf{y}^T\nabla'\mathbf{d} + \nabla'\mathbf{d}^T\nabla'\mathbf{y}))g^{-\frac{1}{2}} & \frac{1}{2}g^{-\frac{1}{2}}(\nabla'\mathbf{d}^T\mathbf{b} + 2\nabla'\mathbf{b}^T\mathbf{d}) \\ \frac{1}{2}(\nabla'\mathbf{d}^T\mathbf{b} + 2\nabla'\mathbf{b}^T\mathbf{d})^T g^{-\frac{1}{2}} & \beta^2 \end{bmatrix}$$

are independent of x_3 and

$$\mathrm{I}[\mathbf{y}] = \nabla'\mathbf{y}^T\nabla'\mathbf{y} \quad \text{and} \quad \mathrm{II}[\mathbf{y}] = -\nabla'\boldsymbol{\nu}^T\nabla'\mathbf{y}$$

are the first and second fundamental forms of $\mathbf{y}(\Omega)$, respectively. To evaluate the two terms on the right-hand side of (2.1.10), we split them into powers of x_3. We first deal with the pre-asymptotic regime, in which $s > 0$ is small, and next we consider the asymptotic regime $s \to 0$.

Pre-asymptotics. To compute $s^{-3}\int_{\Omega_s}\left|G^{-\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-\frac{1}{2}}\right|^2$, we first note that

$$\left|G^{-\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-\frac{1}{2}}\right|^2 = \frac{1}{4}|A_1|^2 + x_3A_1 : A_2 + \frac{x_3^2}{2}A_1 : A_3 + x_3^2|A_2|^2 + \mathcal{O}(x_3^3),$$

all the terms with odd powers of x_3 integrate to zero on $[-s/2, s/2]$, and those terms hidden in $\mathcal{O}(x_3^3)$ integrate to an $\mathcal{O}(s)$ contribution after rescaling by s^{-3}. We next realize that

$$\begin{aligned} s^{-3}\int_{-s/2}^{s/2} dx_3 \int_\Omega |A_1|^2 d\mathbf{x}' &= s^{-2}\int_\Omega \big|A_1\big|^2 d\mathbf{x}' \\ s^{-3}\int_{-s/2}^{s/2} x_3^2\, dx_3 \int_\Omega A_1 : A_3 d\mathbf{x}' &= \frac{1}{12}\int_\Omega A_1 : A_3 d\mathbf{x}' \\ s^{-3}\int_{-s/2}^{s/2} x_3^2\, dx_3 \int_\Omega |A_2|^2 d\mathbf{x}' &= \frac{1}{12}\int_\Omega \big|A_2\big|^2 d\mathbf{x}', \end{aligned}$$

and exploit that $s^{-3}\int_{\Omega_s} \big|G^{-\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-\frac{1}{2}}|^2 \le \Lambda$ independent of s to find that

$$\Big|\int_\Omega A_1 : A_3 d\mathbf{x}'\Big| \le s\left(s^{-2}\int_\Omega |A_1|^2 d\mathbf{x}'\right)^{\frac{1}{2}}\left(\int_\Omega |A_3|^2 d\mathbf{x}'\right)^{\frac{1}{2}} \le C\Lambda^{\frac{1}{2}} s$$

is a higher order term because $\int_\Omega |A_3|^2 d\mathbf{x}' \le C^2$. We thus obtain the expression

$$s^{-3}\int_{\Omega_s} \big|G^{\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{\frac{1}{2}}\big|^2 = \frac{1}{4s^2}\int_\Omega \big|A_1\big|^2 d\mathbf{x}' + \frac{1}{12}\int_\Omega \big|A_2\big|^2 d\mathbf{x}' + \mathcal{O}(s).$$

We proceed similarly with the second term in (2.1.10) to arrive at

$$\begin{aligned} \operatorname{tr}\big(G^{-\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-\frac{1}{2}}\big)^2 =& \frac{1}{4}\operatorname{tr}(A_1)^2 + x_3\operatorname{tr}(A_1)\operatorname{tr}(A_2) + \frac{1}{2}x_3^2\operatorname{tr}(A_1)\operatorname{tr}(A_3) \\ &+ x_3^2\operatorname{tr}(A_2)^2 + \mathcal{O}(x_3^3), \end{aligned}$$

and

$$s^{-3}\int_{\Omega_s}\operatorname{tr}\big(G^{-\frac{1}{2}}\boldsymbol{\epsilon}_G(\nabla\mathbf{u})G^{-\frac{1}{2}}\big)^2=\frac{1}{4s^2}\int_{\Omega}\operatorname{tr}\big(A_1\big)^2d\mathbf{x}'+\frac{1}{12}\int_{\Omega}\operatorname{tr}\big(A_2\big)^2d\mathbf{x}'+\mathcal{O}(s).$$

In view of (2.1.10) and (2.1.11), we deduce that the rescaled elastic energy $s^{-3}E[\mathbf{u}]\approx E_s[\mathbf{y}]+E_b[\mathbf{y}]$ for s small, where the two leading terms are the *stretching energy*

$$E_s[\mathbf{y}]=\frac{1}{8s^2}\int_{\Omega}\Big(2\mu\big|A_1\big|^2+\lambda\operatorname{tr}\big(A_1\big)^2\Big)d\mathbf{x}' \tag{2.1.15}$$

and the *bending energy*

$$E_b[\mathbf{y}]=\frac{1}{24}\int_{\Omega}\Big(2\mu\big|A_2\big|^2+\lambda\operatorname{tr}\big(A_2\big)^2\Big)d\mathbf{x}' \tag{2.1.16}$$

with A_1 and A_2 depending on $\mathrm{I}[\mathbf{y}]$ and $\mathrm{II}[\mathbf{y}]$, respectively.

Asymptotics. We now let the thickness $s\to 0$ and observe that for the scaled energy to remain uniformly bounded, the integrant of the stretching energy must vanish with a rate at least s^2. By definition of A_1, this implies that the parametrization $\mathbf{y}$ must satisfy the metric constraint $g^{-\frac{1}{2}}\,\mathrm{I}[\mathbf{y}]\,g^{-\frac{1}{2}}=I_2$, or equivalently $\mathbf{y}$ is an *isometric immersion* of g

$$\nabla'\mathbf{y}^T\nabla'\mathbf{y}=g\quad\text{a.e. in }\Omega, \tag{2.1.17}$$

and $\alpha^2\equiv 1$. Since $E_s[\mathbf{y}]=0$, we can take the limit for $s\to 0$ and neglect the higher

order terms to obtain the following expression for the reduced elastic energy

$$\lim_{s\to 0}\frac{1}{s^3}\int_{\Omega_s} W(\nabla \mathbf{u} G^{-\frac{1}{2}})d\mathbf{x} = \frac{1}{24}\int_{\Omega}\Big(\underbrace{2\mu|A_2|^2+\lambda\mathrm{tr}(A_2)^2}_{=:w(\beta)}\Big)d\mathbf{x}', \tag{2.1.18}$$

where, using the definition of A_2, $w(\beta)$ is given by

$$w(\beta) = 2\mu|g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}}|^2 + 2\mu\beta^2 + \lambda(-\mathrm{tr}(g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}})+\beta)^2$$

because $\alpha^2 \equiv 1$. In order to obtain deformations with minimal energies, we now choose $\beta = \beta(\mathbf{x}')$ such that $w(\beta)$ is minimized. Since

$$\frac{dw}{d\beta} = 4\mu\beta + 2\lambda\Big(-\mathrm{tr}\big(g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}}\big)+\beta\Big) = 0 \quad \text{and} \quad \frac{d^2w}{d\beta^2} = 4\mu+2\lambda > 0,$$

we get

$$\beta = \frac{\lambda}{2\mu+\lambda}\mathrm{tr}\big(g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}}\big),$$

which gives

$$w(\beta) = 2\mu\big|g^{-\frac{1}{2}}\,\mathrm{II}[\mathbf{y}]\,g^{-\frac{1}{2}}\big|^2 + \frac{2\mu\lambda}{\lambda+2\mu}\mathrm{tr}\big(g^{-\frac{1}{2}}\mathrm{II}[\mathbf{y}]g^{-\frac{1}{2}}\big)^2.$$

Finally, the right-hand side of (2.1.18) has to be supplemented with the forcing term that we have ignored in this derivation but scales correctly owing to definition (2.1.14). In the sequel, we relabel the bending energy $E_b[\mathbf{y}]$ as $E[\mathbf{y}]$, add the forcing

and replace $\mathbf{x}'$ by $\mathbf{x}$ (and drop the notation $'$ on differential operators)

$$E[\mathbf{y}] = \frac{\mu}{12}\int_\Omega \left(\left|g^{-\frac{1}{2}}\mathrm{II}[\mathbf{y}]g^{-\frac{1}{2}}\right|^2 + \frac{\lambda}{2\mu+\lambda}\mathrm{tr}\left(g^{-\frac{1}{2}}\mathrm{II}[\mathbf{y}]g^{-\frac{1}{2}}\right)^2\right)d\mathbf{x} - \int_\Omega \mathbf{f}\cdot\mathbf{y}d\mathbf{x}. \quad (2.1.19)$$

This formal procedure has been justified via Γ-convergence in [39, 41] for isometries $\mathrm{I}[\mathbf{y}] = I_2$ and in [53, Corollary 2.7], [17, Theorem 2.1] for isometric immersions $\mathrm{I}[\mathbf{y}] = g$. Moreover, as already observed in [39], we mention that using the Kirchhoff-Love assumption (2.1.13) instead (2.1.12) yields a similar bending energy, namely we obtain (2.1.19) but with λ instead of $\frac{\mu\lambda}{2\mu+\lambda}$.

2.1.3 Admissibility

We need to supplement (2.1.19) with suitable boundary conditions for $\mathbf{y}$ for the minimization problem to be well-posed. For simplicity, we consider Dirichlet and free boundary conditions in this chapter, but other types of boundary conditions are possible. Let $\Gamma_D \subset \partial\Omega$ be a (possibly empty) open set on which the following Dirichlet boundary conditions are imposed:

$$\mathbf{y} = \boldsymbol{\varphi} \quad \text{and} \quad \nabla\mathbf{y} = \Phi \quad \text{on } \Gamma_D, \quad (2.1.20)$$

where $\boldsymbol{\varphi} : \Omega \to \mathbb{R}^3$ and $\Phi : \Omega \to \mathbb{R}^{3\times 2}$ are sufficiently smooth and Φ satisfies the compatibility condition $\Phi^T\Phi = g$ a.e. in Ω. The set of *admissible* functions is

$$\mathbb{A}(\boldsymbol{\varphi}, \Phi) := \left\{\mathbf{y} \in \mathbb{V}(\boldsymbol{\varphi}, \Phi) : \nabla\mathbf{y}^T\nabla\mathbf{y} = g \text{ a.e. in } \Omega\right\}, \quad (2.1.21)$$

where the affine manifold $\mathbb{V}(\boldsymbol{\varphi}, \Phi)$ of $H^2(\Omega)$ is defined by

$$\mathbb{V}(\boldsymbol{\varphi}, \Phi) := \left\{ \mathbf{y} \in [H^2(\Omega)]^3 : \mathbf{y}_{|\Gamma_D} = \boldsymbol{\varphi},\ \nabla \mathbf{y}_{|\Gamma_D} = \Phi \right\}. \tag{2.1.22}$$

Our goal is to obtain

$$\mathbf{y}^* := \operatorname{argmin}_{\mathbf{y} \in \mathbb{A}(\boldsymbol{\varphi}, \Phi)} E(\mathbf{y}), \tag{2.1.23}$$

but this minimization problem is highly nonlinear and seems to be out of reach both analytically and geometrically. In fact, whether or not there exists a smooth *global* deformation $\mathbf{y}$ from $\Omega \subset \mathbb{R}^n$ into $\mathbb{R}^N$ satisfying the metric constraint (2.1.17), a so-called *isometric immersion*, is a long standing problem in differential geometry [47]. Note that $\nabla \mathbf{y}$ is full rank if $\mathbf{y}$ is an isometric immersion; if in addition $\mathbf{y}$ is injective, then we say that $\mathbf{y}$ is an *isometric embedding.* For $n = 2$, Nash's theorem guarantees that an isometric embedding exists for $N = 10$ (Nash proved it for $N = 17$, while it was further improved to $N = 10$ by Gromov [43]). When $N = 3$, as in our context, a given metric g may or may not admit an isometric immersion. Some elliptic and hyperbolic metrics with special assumptions have isometric immersions in $\mathbb{R}^3$ [47]. We assume implicitly below that $\mathbb{A}(\boldsymbol{\varphi}, \Phi)$ is non-empty, thus there exists an isometric immersion that satisfies boundary conditions, but now we discuss an illuminating example in polar coordinates [36, 64].

Change of variables and polar coordinates. If $\zeta = (\zeta_1, \zeta_2) : \widetilde{\Omega} \to \Omega$ is a change of variables $\xi \mapsto \mathbf{x}$ into Cartesian coordinates $\mathbf{x} = (x_1, x_2) \in \Omega$ and $\mathbf{J}(\xi)$ is the

Jacobian matrix, then the target metrics $\widetilde{g}(\xi)$ and $g(\mathbf{x}) = g(\zeta(\xi))$ satisfy

$$\widetilde{g}(\xi) = \mathbf{J}(\xi)^T g(\zeta(\xi))\mathbf{J}(\xi), \quad \mathbf{J}(\xi) = \begin{bmatrix} \partial_{\xi_1}\zeta_1(\xi) & \partial_{\xi_2}\zeta_1(\xi) \\ \partial_{\xi_1}\zeta_2(\xi) & \partial_{\xi_2}\zeta_2(\xi) \end{bmatrix}. \tag{2.1.24}$$

Let $\xi = (r, \theta)$ indicate polar coordinates with $r \in I = [0, R]$ and $\theta \in [0, 2\pi)$. If $g = I_2$ is the identity matrix (i.e., $\mathrm{I}[\mathbf{y}] = I_2$) and $\eta(r) = r$, then $\widetilde{g}(\xi)$ reads

$$\widetilde{g}(r, \theta) = \begin{bmatrix} 1 & 0 \\ 0 & \eta(r)^2 \end{bmatrix}. \tag{2.1.25}$$

We now show that some metrics of the form of (2.1.25) with $\eta(r) \neq r$ are still isometric immersible provided η is sufficiently smooth. Consider the case $|\eta'(r)| \leq 1$ along with the parametrization

$$\widetilde{\mathbf{y}}(r, \theta) = (\eta(r)\cos\theta, \eta(r)\sin\theta, \psi(r))^T. \tag{2.1.26}$$

Since $\partial_r\widetilde{\mathbf{y}} \cdot \partial_\theta\widetilde{\mathbf{y}} = 0$ and $|\partial_\theta\widetilde{\mathbf{y}}|^2 = \eta(r)^2$, if ψ satisfies $|\partial_r\widetilde{\mathbf{y}}|^2 = \eta'(r)^2 + \psi'(r)^2 = 1$, we realize that $\widetilde{\mathbf{y}}$ is an isometric embedding compatible with (2.1.25). On the other hand, if $|\eta'(r)| \geq 1$ and $a \geq \max_{r\in I} |\eta'(r)|$ is an integer, then the parametrization

$$\widetilde{\mathbf{y}}(r, \theta) = \Big(\frac{\eta(r)}{a}\cos(a\theta), \frac{\eta(r)}{a}\sin(a\theta), \int_0^r \sqrt{1 - \frac{\eta'(t)^2}{a^2}}dt\Big)^T \tag{2.1.27}$$

is an isometric immersion compatible with (2.1.25) but not an isometric embedding. We will construct later a couple of isometric embeddings computationally.

We also point out that (2.1.25) accounts for *shrinking* if $0 \leq \eta(r) < r$ and *stretching* if $\eta(r) > r$. To see this, let $\gamma_r(\theta) = (r, \theta)^T$, $\theta \in [0, 2\pi)$, be the parametrization of a circle in Ω centered at the origin and of radius r, and let $\Gamma_r(\theta) = \widetilde{\mathbf{y}}(\gamma_r(\theta))$ be its image on $\widetilde{\mathbf{y}}(\widetilde{\Omega}) = \mathbf{y}(\Omega)$. The length $\ell(\Gamma_r)$ satisfies

$$\ell(\Gamma_r) = \int_0^{2\pi} \left| \frac{d}{d\theta} \Gamma_r(\theta) \right| d\theta = \int_0^{2\pi} \sqrt{\gamma_r'(\theta)^T \widetilde{g}(r,\theta) \gamma_r'(\theta)} d\theta = \int_0^{2\pi} \eta(r) d\theta = \ell(\gamma_r) \frac{\eta(r)}{r},$$

and the ratio $\eta(r)/r$ acts as a shrinking/stretching parameter.

Gaussian curvature. Since $E[\mathbf{y}] > 0$ provided that the Gaussian curvature $\kappa = \det(\mathrm{II}[\mathbf{y}]) \det(\mathrm{I}[\mathbf{y}])^{-1}$ of the surface $\mathbf{y}(\Omega)$ does not vanish identically [17, 53], it is instructive to find κ for a deformation $\widetilde{\mathbf{y}}$ so that $\mathrm{I}[\widetilde{\mathbf{y}}] = \widetilde{g}$ is given by (2.1.25). Since the formula for change of variables for $\mathrm{II}[\widetilde{\mathbf{y}}]$ is the same as that in (2.1.24) for $\widetilde{g} = \mathrm{I}[\widetilde{\mathbf{y}}]$, we realize that κ is independent of the parametrization of the surface. According to Gauss's Theorema Egregium, $\kappa = \det(\mathrm{II}[\widetilde{\mathbf{y}}]) \det(\mathrm{I}[\widetilde{\mathbf{y}}])^{-1}$ can be rewritten as an expression solely depending on $\mathrm{I}[\widetilde{\mathbf{y}}]$. Do Carmo gives an explicit formula for κ in case $\widetilde{g} = \mathrm{I}[\widetilde{\mathbf{y}}]$ is diagonal [34, Exercise 1, p.237], which reduces to

$$\kappa = -\frac{\eta''(r)}{\eta(r)} \tag{2.1.28}$$

for $\widetilde{g}$ of the form (2.1.25). Alternatively, we may express $\mathrm{II}[\widetilde{\mathbf{y}}]_{ij} = \partial_{ij} \widetilde{\mathbf{y}} \cdot \widetilde{\boldsymbol{\nu}}$, where $\widetilde{\boldsymbol{\nu}}(r, \theta)$ is the unit normal vector to the surface $\widetilde{\mathbf{y}}(\widetilde{\Omega})$ at the point $\widetilde{\mathbf{y}}(r, \theta)$, in terms

of the orthonormal basis $\{\widetilde{\boldsymbol{\nu}}, \partial_r\widetilde{\mathbf{y}}, \eta(r)^{-1}\partial_\theta\widetilde{\mathbf{y}}\}$ as follows. First observe that

$$\begin{aligned}
|\partial_r\widetilde{\mathbf{y}}|^2 = 1 \quad &\Rightarrow \quad \partial_{rr}\widetilde{\mathbf{y}}\cdot\partial_r\widetilde{\mathbf{y}} = 0, \quad \partial_{\theta r}\widetilde{\mathbf{y}}\cdot\partial_r\widetilde{\mathbf{y}} = 0,\\
|\partial_\theta\widetilde{\mathbf{y}}|^2 = \eta^2(r) \quad &\Rightarrow \quad \partial_{r\theta}\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = \eta(r)\eta'(r), \quad \partial_{\theta\theta}\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = 0,\\
\partial_r\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = 0 \quad &\Rightarrow \quad \partial_{rr}\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = 0.
\end{aligned}$$

This yields

$$\partial_{rr}\widetilde{\mathbf{y}} = (\partial_{rr}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}})\widetilde{\boldsymbol{\nu}}, \quad \partial_{\theta\theta}\widetilde{\mathbf{y}} = (\partial_{\theta\theta}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}})\widetilde{\boldsymbol{\nu}} + (\partial_{\theta\theta}\widetilde{\mathbf{y}}\cdot\partial_r\widetilde{\mathbf{y}})\partial_r\widetilde{\mathbf{y}},$$

whence

$$\mathrm{II}[\widetilde{\mathbf{y}}]_{rr}\mathrm{II}[\widetilde{\mathbf{y}}]_{\theta\theta} = (\partial_{rr}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}})(\partial_{\theta\theta}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}}) = \partial_{rr}\widetilde{\mathbf{y}}\cdot\partial_{\theta\theta}\widetilde{\mathbf{y}}.$$

We next differentiate $\partial_{rr}\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = 0$ and $\partial_{r\theta}\widetilde{\mathbf{y}}\cdot\partial_\theta\widetilde{\mathbf{y}} = \eta(r)\eta'(r)$ with respect to θ and r, respectively, to obtain

$$\partial_{rr}\widetilde{\mathbf{y}}\cdot\partial_{\theta\theta}\widetilde{\mathbf{y}} = \partial_{r\theta}\widetilde{\mathbf{y}}\cdot\partial_{r\theta}\widetilde{\mathbf{y}} - \eta'(r)^2 - \eta(r)\eta''(r).$$

We finally notice that $\partial_{r\theta}\widetilde{\mathbf{y}} = (\partial_{r\theta}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}})\widetilde{\boldsymbol{\nu}} + \frac{\eta'(r)}{\eta(r)}\partial_\theta\widetilde{\mathbf{y}}$, whence

$$\big(\mathrm{II}[\widetilde{\mathbf{y}}]_{r\theta}\big)^2 = (\partial_{r\theta}\widetilde{\mathbf{y}}\cdot\widetilde{\boldsymbol{\nu}})^2 = \partial_{r\theta}\widetilde{\mathbf{y}}\cdot\partial_{r\theta}\widetilde{\mathbf{y}} - \eta'(r)^2.$$

Therefore, we have derived $\det\mathrm{II}[\widetilde{\mathbf{y}}] = \mathrm{II}[\widetilde{\mathbf{y}}]_{rr}\mathrm{II}[\widetilde{\mathbf{y}}]_{\theta\theta} - \big(\mathrm{II}[\widetilde{\mathbf{y}}]_{r\theta}\big)^2 = -\eta(r)\eta''(r)$ and as $\det\mathrm{I}[\widetilde{\mathbf{y}}] = \eta(r)^2$, we obtain (2.1.28). This expression will be essential in a computa-

tional example that is presented later.

2.1.4 Alternative energy

The expression (2.1.19) involves the second fundamental form $\mathrm{II}[\mathbf{y}] = -\nabla\boldsymbol{\nu}^T\nabla\mathbf{y}$ and is too nonlinear to be practically useful. To render (2.1.23) amenable to computation, we show now that $\mathrm{II}[\mathbf{y}]$ can be replaced by the Hessian $D^2\mathbf{y}$ without affecting the minimizers. This is the subject of next proposition, which uses the notation (2.1.7) for $g^{-1/2}D^2\mathbf{y}g^{-1/2}$.

Proposition 2.1.1 (alternative energy)**.** *Let $\mathbf{y} = (y_k)_{k=1}^3 : \Omega \to \mathbb{R}^3$ be a sufficiently smooth orientable deformation and let $g = \mathrm{I}[\mathbf{y}]$ and $\mathrm{II}[\mathbf{y}]$ be the first and second fundamental forms of $\mathbf{y}(\Omega)$. Then, there exist functions $f_1, f_2 : \Omega \to \mathbb{R}_{\geq 0}$ depending only on g and its derivatives, with precise definitions given in the proof, such that*

$$\left|g^{-\frac{1}{2}}\, D^2\mathbf{y}\, g^{-\frac{1}{2}}\right|^2 = \left|g^{-\frac{1}{2}}\, \mathrm{II}[\mathbf{y}]\, g^{-\frac{1}{2}}\right|^2 + f_1, \tag{2.1.29}$$

and

$$\left|\mathrm{tr}\big(g^{-\frac{1}{2}}\, D^2\mathbf{y}\, g^{-\frac{1}{2}}\big)\right|^2 = \mathrm{tr}\big(g^{-\frac{1}{2}}\, \mathrm{II}[\mathbf{y}]\, g^{-\frac{1}{2}}\big)^2 + f_2. \tag{2.1.30}$$

Proof. First of all, because $\mathbf{y}$ is smooth and orientable, the second derivatives $\partial_{ij}\mathbf{y}$ of the deformation $\mathbf{y}$ can be (uniquely) expressed in the basis $\{\partial_1\mathbf{y}, \partial_2\mathbf{y}, \boldsymbol{\nu}\}$ as

$$\partial_{ij}\mathbf{y} = \sum_{l=1}^{2} \Gamma_{ij}^l\, \partial_l\mathbf{y} + \mathrm{II}_{ij}[\mathbf{y}]\, \boldsymbol{\nu}, \tag{2.1.31}$$

where $\frac{\partial_1\mathbf{y}\times\partial_2\mathbf{y}}{|\partial_1\mathbf{y}\times\partial_2\mathbf{y}|}$ is the unit normal and Γ_{ij}^l are the so-called Christoffel symbols of

$\mathbf{y}(\Omega)$. Since Γ^l_{ij} are intrinsic quantities, they can be computed in terms of the coefficients g_{ij} of g and their derivatives [34]; they do not depend explicitly on $\mathbf{y}$.

We start with the proof of relation (2.1.29). To simplify the notation, let us write $a = g^{-\frac{1}{2}}$. Using (2.1.31) we get

$$\begin{aligned}(a\,\mathrm{II}[\mathbf{y}]\,a)_{ij}\,\boldsymbol{\nu} &= \sum_{m,n=1}^{2} a_{im}\big(\mathrm{II}_{mn}[\mathbf{y}]\,\boldsymbol{\nu}\big)\,a_{nj} \\ &= \sum_{m,n=1}^{2} a_{im}(\partial_{mn}\mathbf{y})a_{nj} - \sum_{m,n=1}^{2} a_{im}\left(\sum_{l=1}^{2}\Gamma^l_{mn}\partial_l\mathbf{y}\right)a_{nj},\end{aligned}$$

or equivalently, rearranging the above expression,

$$(a\,D^2\mathbf{y}\,a)_{ij} = (a\,\mathrm{II}[\mathbf{y}]\,a)_{ij}\boldsymbol{\nu} + \sum_{m,n=1}^{2} a_{im}\left(\sum_{l=1}^{2}\Gamma^l_{mn}\partial_l\mathbf{y}\right)a_{nj}.$$

Since the unit vector $\boldsymbol{\nu}$ is orthogonal to both $\partial_1\mathbf{y}$ and $\partial_2\mathbf{y}$, the right-hand side is an l_2-orthogonal decomposition. Computing the square of the l_2-norms yields

$$\sum_{k=1}^{3}(a\,D^2 y_k\,a)^2_{ij} = (a\,\mathrm{II}[\mathbf{y}]\,a)^2_{ij} + f_{ij} \tag{2.1.32}$$

with

$$f_{ij} := \sum_{l_1,l_2=1}^{2} g_{l_1 l_2} \sum_{m_1,m_2,n_1,n_2=1}^{2} a_{im_1}a_{im_2}\Gamma^{l_1}_{m_1 n_1}\Gamma^{l_2}_{m_2 n_2}a_{n_1 j}a_{n_2 j}.$$

Functions f_{ij} do not depend explicitly on $\mathbf{y}$ but on g and first derivatives of g. Therefore, summing (2.1.32) over i, j from 1 to 2 gives (2.1.29) with $f_1 := \sum_{i,j=1}^{2} f_{ij}$.

The proof of (2.1.30) is similar. Since $\mathrm{tr}(a\,\mathrm{II}[\mathbf{y}]\,a)\,\boldsymbol{\nu} = \sum_{i=1}^{2}(a\,\mathrm{II}[\mathbf{y}]\,a)_{ii}\,\boldsymbol{\nu}$ it suffices to take $i = j$ and sum over i in the previous derivation to arrive at (2.1.30)

with

$$f_2 := \sum_{l_1,l_2=1}^{2} g_{l_1 l_2} \sum_{i_1,i_2,m_1,m_2,n_1,n_2=1}^{2} a_{i_1 m_1} a_{i_2 m_2} \Gamma^{l_1}_{m_1 n_1} \Gamma^{l_2}_{m_2 n_2} a_{n_1 i_1} a_{n_2 i_2}.$$

This completes the proof because f_2 does not dependent explicitly on $\mathbf{y}$. □

Remark 2.1.1 (alternative energy)**.** *As stated, Proposition 2.1.1 is valid for smooth deformations* $\mathbf{y}$ *and metric* g*. It turns out that for* $\mathbf{y} \in [H^2(\Omega)]^3$ *and* $g \in [H^1(\Omega) \cap L^\infty(\Omega)]^{2\times 2}$*, the key relation* (2.1.31) *holds a.e. in* Ω *and so does the conclusion of Proposition 2.1.1.*

Proposition 2.1.1 (alternative energy) shows that the solutions of (2.1.23) with the energy $E[\mathbf{y}]$ given by (2.1.19) are the same as those given by the energy

$$E(\mathbf{y}) := \frac{\mu}{12} \int_\Omega \left(\left| g^{-\frac{1}{2}} D^2\mathbf{y}\, g^{-\frac{1}{2}} \right|^2 + \frac{\lambda}{2\mu+\lambda} \left| \mathrm{tr}\big(g^{-\frac{1}{2}} D^2\mathbf{y}\, g^{-\frac{1}{2}}\big) \right|^2 \right) - \int_\Omega \mathbf{f}\cdot\mathbf{y}. \quad (2.1.33)$$

The Euler-Lagrange equations characterizing local extrema $\mathbf{y} \in [H^2(\Omega)]^3$ of (2.1.33)

$$\delta E[\mathbf{y};\mathbf{v}] = 0 \quad \forall \mathbf{v} \in [H^2(\Omega)]^3, \quad (2.1.34)$$

can be written in terms of the first variation of $E[\mathbf{y}]$ in the direction $\mathbf{v}$ given by

$$\begin{aligned} \delta E[\mathbf{y};\mathbf{v}] :=& \frac{\mu}{6} \int_\Omega \big(g^{-\frac{1}{2}} D^2\mathbf{y}\, g^{-\frac{1}{2}}\big) : \big(g^{-\frac{1}{2}} D^2\mathbf{v}\, g^{-\frac{1}{2}}\big) \\ &+ \frac{\mu\lambda}{6(2\mu+\lambda)} \int_\Omega \mathrm{tr}\big(g^{-\frac{1}{2}} D^2\mathbf{y}\, g^{-\frac{1}{2}}\big) \cdot \mathrm{tr}\big(g^{-\frac{1}{2}} D^2\mathbf{v}\, g^{-\frac{1}{2}}\big) - \int_\Omega \mathbf{f}\cdot\mathbf{v}. \end{aligned} \quad (2.1.35)$$

The presence of the trace term in (2.1.35) makes it problematic to find the governing partial differential equation hidden in (2.1.34) (strong form). However, when $\lambda = 0$,

integration by parts shows that $P_k := g^{-1}\, D^2 y_k\, g^{-1} \in \mathbb{R}^{2\times 2}$ for $k = 1, 2, 3$ satisfies

$$\delta E[\mathbf{y}; \mathbf{v}] = \frac{\mu}{6} \sum_{k=1}^{3} \left(\int_\Omega \operatorname{div}\operatorname{div} P_k\, v_k - \int_{\partial\Omega} \operatorname{div} P_k \cdot \mathbf{n} v_k + \int_{\partial\Omega} P_k \mathbf{n} \cdot \nabla v_k \right) - \int_\Omega \mathbf{f} \cdot \mathbf{v},$$

where $\mathbf{n}$ is the outwards unit normal vector to $\partial\Omega$. On the other hand, if $g = I_2$ in which case $\mathbf{y}$ is an *isometry*, then $E[\mathbf{y}]$ in (2.1.19) and (2.1.33) are equal and reduce to

$$E[\mathbf{y}] = \frac{\alpha}{2} \int_\Omega |D^2 \mathbf{y}|^2 - \int_\Omega \mathbf{f} \cdot \mathbf{y}, \quad \alpha := \frac{\mu(\mu + \lambda)}{3(2\mu + \lambda)} \tag{2.1.36}$$

thanks to the relations for isometries [10, 13, 22]

$$|\mathrm{II}[\mathbf{y}]| = |D^2 \mathbf{y}| = |\Delta \mathbf{y}| = \operatorname{tr}(\mathrm{II}[\mathbf{y}]). \tag{2.1.37}$$

The strong form of the Euler-Lagrange equation for a minimizer of (2.1.36) reads $\alpha \operatorname{div}\operatorname{div} D^2 \mathbf{y} = \alpha \Delta^2 \mathbf{y} = \mathbf{f}$. This problem has been studied numerically in [10, 22].

2.2 Discretization

We propose here a *local discontinuous Galerkin* (LDG) method to approximate the solution of the problem (2.1.23). LDG is inspired by, and in fact improves upon, the previous dG methods [22, 23] but they are conceptually different. LDG hinges on the explicit computation of a discrete Hessian $H_h[\mathbf{y}_h]$ for the discontinuous piecewise polynomial approximation $\mathbf{y}_h$ of $\mathbf{y}$, which allows for a direct discretization of $E_h[\mathbf{y}_h]$ in (2.1.33), including the trace term. A salient feature is that the stability of the LDG method is retained even when the penalty parameters are arbitrarily small.

2.2.1 LDG-type discretization

From now on, we assume that $\Omega \subset \mathbb{R}^2$ is a polygonal domain. Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular but possibly graded elements T, either triangles or quadrilaterals, of diameter $h_T := \text{diam}(T) \leq h$. In order to handle hanging nodes (necessary for graded meshes based on quadrilaterals), we assume that all the elements within each domain of influence have comparable diameters. We refer to Sections 2.2.4 and 6 of Bonito-Nochetto [21] for precise definitions and properties. At this stage, we only point out that sequences of subdivisions made of quadrilaterals with at most one hanging node per side satisfy this assumption.

Let $\mathcal{E}_h = \mathcal{E}_h^0 \cup \mathcal{E}_h^b$ denote the set of edges, where $\mathcal{E}_h^0$ stands for the set of interior edges and $\mathcal{E}_h^b$ for the set of boundary edges. We assume a compatible representation of the Dirichlet boundary Γ_D, i.e., if $\Gamma_D \neq \emptyset$ then Γ_D is the union of (some) edges in $\mathcal{E}_h^b$ for every $h > 0$, which we indicate with $\mathcal{E}_h^D$; note that Γ_D and $\mathcal{E}_h^D$ are empty sets when dealing with a problem with free boundary conditions. Let $\mathcal{E}_h^a := \mathcal{E}_h^0 \cup \mathcal{E}_h^D$ the set of *active edges* on which jumps and averages will be computed. The union of these edges give rise to the corresponding skeletons of $\mathcal{T}_h$

$$\Gamma_h^0 := \cup\{e : e \in \mathcal{E}_h^0\}, \quad \Gamma_h^D := \cup\{e : e \in \mathcal{E}_h^D\}, \quad \Gamma_h^a := \Gamma_h^0 \cup \Gamma_h^D. \tag{2.2.1}$$

If h_e is the diameter of $e \in \mathcal{E}_h$, then we introduce the piecewise constant mesh density function h defined to be equivalent locally to the size h_T of T and h_e of an edge e. From now on, we use the notation $(\cdot,\cdot)_{L^2(\Omega)}$ and $(\cdot,\cdot)_{L^2(\Gamma_h^a)}$ to denote the L^2

inner products over Ω and Γ_h^a, and a similar notation for subsets of Ω and Γ_h^a.

Broken spaces. For an integer $k \geq 0$, we let $\mathbb{P}_k$ (resp. $\mathbb{Q}_k$) be the space of polynomials of total degree at most k (resp. of degree at most k in the each variable). The reference unit triangle (resp. square) is denoted by $\widehat{T}$ and for $T \in \mathcal{T}_h$, we let $F_T : \widehat{T} \to T$ be the generic map from the reference element to the physical element. When $\mathcal{T}_h$ is made of triangles the map is affine, i.e., $F_T \in [\mathbb{P}_1]^2$, while $F_T \in [\mathbb{Q}_1]^2$ when quadrilaterals are used.

If $k \geq 2$, the (*broken*) finite element space $\mathbb{V}_h^k$ to approximate each component of the deformation $\mathbf{y}$ (modulo boundary conditions) reads

$$\mathbb{V}_h^k := \left\{ v_h \in L^2(\Omega) : \ v_{h\,|T} \circ F_T \in \mathbb{P}_k \quad (\text{resp. } \mathbb{Q}_k) \quad \forall T \in \mathcal{T}_h \right\} \tag{2.2.2}$$

if $\mathcal{T}_h$ is made of triangles (resp. quadrilaterals). We define the broken gradient $\nabla_h v_h$ of $v_h \in \mathbb{V}_h^k$ to be the gradient computed elementwise, and use similar notation for other piecewise differential operators such as the broken Hessian $D_h^2 v_h = \nabla_h \nabla_h v_h$.

We now introduce the jump and average operators. To this end, let $\mathbf{n}_e$ be a unit normal to $e \in \mathcal{E}_h^0$ (the orientation is chosen arbitrarily but is fixed once for all), while for a boundary edge $e \in \mathcal{E}_h^b$, $\mathbf{n}_e$ is the outward unit normal vector to $\partial\Omega$. For $v_h \in \mathbb{V}_h^k$ and $e \in \mathcal{E}_h^0$, we set

$$[v_h]|_e := v_h^- - v_h^+, \quad [\nabla_h v_h]|_e := \nabla_h v_h^- - \nabla_h v_h^+, \tag{2.2.3}$$

where $v_h^\pm(\mathbf{x}) := \lim_{s \to 0^+} v_h(\mathbf{x} \pm s\mathbf{n}_e)$ for $\mathbf{x} \in e$. We compute the jumps compo-

nentwise provided the function v_h is vector or matrix-valued. In what follows, the subindex e is omitted when it is clear from the context.

In order to deal with Dirichlet boundary data $(\boldsymbol{\varphi}, \Phi)$ we resort to a Nitsche approach; hence we do not impose essential restrictions on the discrete space $[\mathbb{V}_h^k]^3$. However, to simplify the notation later, it turns out to be convenient to introduce the discrete sets $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ and $\mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ which mimic the continuous counterparts $\mathbb{V}(\boldsymbol{\varphi}, \Phi)$ and $\mathbb{V}(\mathbf{0}, \mathbf{0})$ but coincide with $[\mathbb{V}_h^k]^3$. In fact, we say that $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$ belongs to $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ provided the boundary jumps of $\mathbf{v}_h$ are defined to be

$$[\mathbf{v}_h]_e := \mathbf{v}_h - \boldsymbol{\varphi}, \quad [\nabla_h \mathbf{v}_h]_e := \nabla_h \mathbf{v}_h - \Phi, \quad \forall\, e \in \mathcal{E}_h^D. \tag{2.2.4}$$

We stress that $\|[\mathbf{v}_h]\|_{L^2(\Gamma_h^D)} \to 0$ and $\|[\nabla_h \mathbf{v}_h]\|_{L^2(\Gamma_h^D)} \to 0$ imply $\mathbf{v}_h \to \boldsymbol{\varphi}$ and $\nabla_h \mathbf{v}_h \to \Phi$ in $L^2(\Gamma_D)$ as $h \to 0$; hence the connection between $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ and $\mathbb{V}(\boldsymbol{\varphi}, \Phi)$. Therefore, we emphasize again that the sets $[\mathbb{V}_h^k]^3$ and $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ coincide but the latter carries the notion of boundary jump, namely

$$\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) := \Big\{ \mathbf{v}_h \in [\mathbb{V}_h^k]^3 : \ [\mathbf{v}_h]_e, [\nabla_h \mathbf{v}_h]_e \text{ given by (2.2.4) for all } e \in \mathcal{E}_h^D \Big\}. \tag{2.2.5}$$

When free boundary conditions are imposed, i.e., $\Gamma_D = \emptyset$, then we do not need to distinguish between $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ and $[\mathbb{V}_h^k]^3$. However, we keep the notation $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ in all cases thereby allowing for a uniform presentation.

We define the *average* of $v_h \in \mathbb{V}_h^k$ across an edge $e \in \mathcal{E}_h$ to be

$$\{v_h\}_e := \begin{cases} \frac{1}{2}(v_h^+ + v_h^-) & e \in \mathcal{E}_h^0 \\ v_h^- & e \in \mathcal{E}_h^b, \end{cases} \tag{2.2.6}$$

and apply this definition componentwise to vector and matrix-valued functions. As for the jump notation, the subindex e is drop when it is clear from the context.

Discrete Hessian. To approximate the elastic energy (2.1.33), we propose an LDG approach. Inspired by [22, 65], the idea is to replace the Hessian $D^2\mathbf{y}$ by a discrete Hessian $H_h[\mathbf{y}_h] \in [L^2(\Omega)]^{3\times 2\times 2}$ to be defined now. To this end, let l_1, l_2 be non-negative integers and consider two *local lifting operators* $r_e : [L^2(e)]^2 \to [\mathbb{V}_h^{l_1}]^{2\times 2}$ and $b_e : L^2(e) \to [\mathbb{V}_h^{l_2}]^{2\times 2}$ defined for $e \in \mathcal{E}_h^a$ by

$$r_e(\boldsymbol{\phi}) \in [\mathbb{V}_h^{l_1}]^{2\times 2} : \int_{\omega_e} r_e(\boldsymbol{\phi}) : \tau_h = \int_e \{\tau_h\}\, \mathbf{n}_e \cdot \boldsymbol{\phi} \quad \forall \tau_h \in [\mathbb{V}_h^{l_1}]^{2\times 2}, \tag{2.2.7}$$

$$b_e(\phi) \in [\mathbb{V}_h^{l_2}]^{2\times 2} : \int_{\omega_e} b_e(\phi) : \tau_h = \int_e \{\operatorname{div} \tau_h\} \cdot \mathbf{n}_e \phi \quad \forall \tau_h \in [\mathbb{V}_h^{l_2}]^{2\times 2}. \tag{2.2.8}$$

It is clear that $\operatorname{supp}(r_e(\boldsymbol{\phi})) = \operatorname{supp}(b_e(\phi)) = \omega_e$, where ω_e is the patch associated with e (i.e., the union of two elements sharing e for interior edges $e \in \mathcal{E}_h^0$ or just one single element for boundary edges $e \in \mathcal{E}_h^b$). We extend r_e and b_e to $[L^2(e)]^{3\times 2}$ and $[L^2(e)]^3$, respectively, by component-wise applications.

The corresponding *global lifting operators* are then given by

$$R_h := \sum_{e\in\mathcal{E}_h^a} r_e : [L^2(\Gamma_h^a)]^2 \to [\mathbb{V}_h^{l_1}]^{2\times 2}, \quad B_h := \sum_{e\in\mathcal{E}_h^a} b_e : L^2(\Gamma_h^a) \to [\mathbb{V}_h^{l_2}]^{2\times 2}. \tag{2.2.9}$$

This construction is simpler than that in [22] for quadrilaterals. We now define the *discrete Hessian operator* $H_h : \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) \to [L^2(\Omega)]^{3\times 2\times 2}$ to be

$$H_h[\mathbf{v}_h] := D_h^2 \mathbf{v}_h - R_h([\nabla_h \mathbf{v}_h]) + B_h([\mathbf{v}_h]). \tag{2.2.10}$$

We can prove in a standard way (using trace and inverse inequalities), see for instance [22, 30], the following a priori upper bounds for the $L^2(\Omega)$ norm of the lifting operators R_h and B_h:

Lemma 2.2.1 (L^2-bound of lifting operators)**.** *For any $\mathbf{v}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$, for any l_1, l_2 non-negative we have*

$$\|R_h([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)} \lesssim \|\mathrm{h}^{-1/2}[\nabla_h \mathbf{v}_h]\|_{L^2(\Gamma_h^a)},$$

and

$$\|B_h([\mathbf{v}_h])\|_{L^2(\Omega)} \lesssim \|\mathrm{h}^{-3/2}[\mathbf{v}_h]\|_{L^2(\Gamma_h^a)}.$$

Discrete energies. We are now ready to introduce the discrete energy on $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$

$$\begin{aligned} E_h[\mathbf{y}_h] :=& \frac{\mu}{12} \int_\Omega \left| g^{-\frac{1}{2}} H_h[\mathbf{y}_h] \, g^{-\frac{1}{2}} \right|^2 \\ &+ \frac{\mu\lambda}{12(2\mu+\lambda)} \int_\Omega \left| \mathrm{tr}\big(g^{-\frac{1}{2}} H_h[\mathbf{y}_h] \, g^{-\frac{1}{2}}\big) \right|^2 \\ &+ \frac{\gamma_1}{2} \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h \mathbf{y}_h]\|_{L^2(\Gamma_h^a)}^2 + \frac{\gamma_0}{2} \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{y}_h]\|_{L^2(\Gamma_h^a)}^2 - \int_\Omega \mathbf{f} \cdot \mathbf{y}_h, \end{aligned} \tag{2.2.11}$$

where $\gamma_0, \gamma_1 > 0$ are stabilization parameters; recall the notation (2.1.4) and (2.1.5). One of the most attractive feature of the LDG method is that γ_0, γ_1 are not required

to be sufficiently large as is typical for interior penalty methods [22].

Note that the Euler-Lagrange equation $\delta E_h[\mathbf{y}_h; \mathbf{v}_h] = 0$ in the direction $\mathbf{v}_h$ reads

$$a_h(\mathbf{y}_h, \mathbf{v}_h) = F(\mathbf{v}_h) \quad \forall\, \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}), \tag{2.2.12}$$

where

$$\begin{aligned} a_h(\mathbf{y}_h, \mathbf{v}_h) :=& \frac{\mu}{6} \int_\Omega \left(g^{-\frac{1}{2}} H_h[\mathbf{y}_h] g^{-\frac{1}{2}}\right) : \left(g^{-\frac{1}{2}} H_h[\mathbf{v}_h] g^{-\frac{1}{2}}\right) \\ &+ \frac{\mu\lambda}{6(2\mu+\lambda)} \int_\Omega \operatorname{tr}\left(g^{-\frac{1}{2}} H_h[\mathbf{y}_h] g^{-\frac{1}{2}}\right) \cdot \operatorname{tr}\left(g^{-\frac{1}{2}} H_h[\mathbf{v}_h] g^{-\frac{1}{2}}\right) \\ &+ \gamma_1 \left(\mathrm{h}^{-1}[\nabla_h \mathbf{y}_h], [\nabla_h \mathbf{v}_h]\right)_{L^2(\Gamma_h^a)} + \gamma_0 \left(\mathrm{h}^{-3}[\mathbf{y}_h], [\mathbf{v}_h]\right)_{L^2(\Gamma_h^a)}, \end{aligned} \tag{2.2.13}$$

and

$$F(\mathbf{v}_h) := \int_\Omega \mathbf{f} \cdot \mathbf{v}_h; \tag{2.2.14}$$

compare with (2.1.34) and (2.1.35).

We reiterate that finding the strong form of (2.1.35) is problematic because of the presence of the trace term. Yet, it is a key ingredient in the design of discontinuous Galerkin methods such as the interior penalty method and raises the question how to construct such methods for (2.1.35). The use of reconstructed Hessian in (2.2.13) leads to a numerical scheme without resorting to the strong form of the equation.

Constraints. We now discuss how to impose the Dirichlet boundary conditions (2.1.20) and the metric constraint (2.1.17) discretely. The former is enforced via the Nitsche approach and thus is not included as a constraint in the discrete admissible

set as in (2.1.21); this turns out to be advantageous for the analysis of the method [22]. The latter is too strong to be imposed on a polynomial space. Inspired by [22], we define the *metric defect* as

$$D_h[\mathbf{y}_h] := \sum_{T\in\mathcal{T}_h} \left| \int_T \left(\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g \right) \right| \tag{2.2.15}$$

and, for a positive number ε, we define the *discrete admissible set* to be

$$\mathbb{A}_{h,\varepsilon}^k := \left\{ \mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) : \quad D_h[\mathbf{y}_h] \leq \varepsilon \right\}.$$

Therefore, the discrete minimization problem, discrete counterpart of (2.1.23), reads

$$\min_{\mathbf{y}_h \in \mathbb{A}_{h,\varepsilon}^k} E_h[\mathbf{y}_h]. \tag{2.2.16}$$

Problem (2.2.16) is nonconvex due to the structure of $\mathbb{A}_{h,\varepsilon}^k$.

2.2.2 Discrete inequalities

In this subsection we collect and prove some key definitions, inequalities and properties for discrete functions $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$, which will be useful in the analysis later in the dissertation.

We first introduce the mesh-dependent quantity $\langle \cdot, \cdot \rangle_{H_h^2(\Omega)}$ defined for any

$\mathbf{v}_h, \mathbf{w}_h \in [\mathbb{V}_h^k]^3$ by

$$\langle \mathbf{v}_h, \mathbf{w}_h \rangle_{H_h^2(\Omega)} := (D_h^2 \mathbf{v}_h, D_h^2 \mathbf{w}_h)_{L^2(\Omega)} + (\mathrm{h}^{-1}[\nabla_h \mathbf{v}_h], [\nabla_h \mathbf{w}_h])_{L^2(\Gamma_h^a)} + (\mathrm{h}^{-3}[\mathbf{v}_h], [\mathbf{w}_h])_{L^2(\Gamma_h^a)} \tag{2.2.17}$$

which is a discrete H^2 scalar product on $\mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ when $\Gamma_D \neq \emptyset$. Moreover, we define

$$|||\mathbf{v}_h|||_{H_h^2(\Omega)}^2 := \langle \mathbf{v}_h, \mathbf{v}_h \rangle_{H_h^2(\Omega)}.$$

In the Dirichlet boundary case $\Gamma_D \neq \emptyset$, the latter is a norm on $\mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ but is not a norm on $\mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ due to the prescribed boundary data. Moreover, $|||\cdot|||_{H_h^2(\Omega)}$ is only a seminorm for the *free boundary case* ($\Gamma_D = \emptyset$).

Then, following what is done in [21, 22], when $k \geq 2$ we define a smoothing interpolation operator $\Pi_h : \mathbb{E}(\mathcal{T}_h) := \Pi_{T \in \mathcal{T}_h} H^1(T) \to \mathbb{V}_h^k \cap H^1(\Omega)$ as follows.

Definition 2.2.1 (Smoothing interpolation)**.** *Given the canonical basis functions $\{\phi_i\}_{i=1}^N$ of $\mathbb{V}_h^k \cap H^1(\Omega)$ with supports $\{\omega_i\}_{i=1}^N$ associated with nodes $\{x_i\}_{i=1}^N$, we compute the local L^2 projection $v_{h,i} \in \mathbb{V}_h^k \cap H^1(\Omega)$ of $v_h \in \mathbb{V}_h^k$ on ω_i for each $i = 1, \ldots, N$*

$$\int_{\omega_i} (v_h - v_{h,i}) w_h = 0 \quad \forall w_h \in \mathbb{V}_h^k \cap H^1(\Omega), \tag{2.2.18}$$

and we define $\Pi_h v_h = \sum_{i=1}^N v_{h,i}(x_i)\phi_i$. For functions $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$, Π_h is applied component-wise. For the general case $\mathbf{v}_h \in \mathbb{E}(\mathcal{T}_h)$, we can modify the definition of Π_h by composition with a local L_2-projection from $\mathbb{E}(\mathcal{T}_h)$ to $\mathbb{V}_h^k$, as in [22].

By (2.2.18), it is clear that $\mathbb{V}_h^k \cap H^1(\Omega)$ is invariant under Π_h. We have the following estimates for Π_h. First, as proven in [22] for any $v_h \in \mathbb{E}(\mathcal{T}_h)$ there holds

$$\|\nabla \Pi_h v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-1}(v_h - \Pi_h v_h)\|_{L^2(\Omega)} \lesssim \|\nabla_h v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|_{L^2(\Gamma_h^0)}. \quad (2.2.19)$$

Then, we further have

Lemma 2.2.2 (Estimates of the smoothing interpolant Π_h)**.** *For any $v_h \in \mathbb{V}_h^k$ we have*

$$\|\mathrm{h}^{-1}(\nabla_h v_h - \nabla \Pi_h v_h)\|_{L^2(\Omega)} \lesssim \|D_h^2 v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h v_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[v_h]\|_{L^2(\Gamma_h^0)}, \quad (2.2.20)$$

$$\|\Pi_h v_h\|_{L^2(\Omega)} \lesssim \|v_h\|_{L^2(\Omega)}, \quad (2.2.21)$$

and

$$\|\nabla \Pi_h v_h - \frac{1}{|\Omega|}\int_\Omega \nabla \Pi_h v_h\|_{L^2(\Omega)} \lesssim \|D_h^2 v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h v_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[v_h]\|_{L^2(\Gamma_h^0)} \quad (2.2.22)$$

Proof. We prove the three inequalities following the ideas used in the proof of Lemma 6.6 in [21] and of Lemma 2.1 in [22]. Definition (2.2.18) implies that

$$\|v_{h,i}\|_{L^2(\omega_i)} \leq \|v_h\|_{L^2(\omega_i)}.$$

Then for any $T \in \mathcal{T}_h$, denoting by ω_T the union of all the ω_i that intersect T, we

have

$$\|\Pi_h v_h\|_{L^2(T)} \lesssim \sum_{x_i \in T} \|v_{h,i}(x_i)\|_{L^2(T)} \leq \sum_{x_i \in T} \|v_{h,i}\|_{L^\infty(T)} |T|^{\frac{1}{2}}$$
$$\lesssim \sum_{x_i \in T} \|v_{h,i}\|_{L^2(T)} \lesssim \|v_h\|_{L^2(\omega_T)},$$

where we used that $v_{h,i}$ belongs to a finite dimensional space (and thus $\|v_{h,i} \circ F_T\|_{L^\infty(\widehat{T})}$ and $\|v_{h,i} \circ F_T\|_{L^2(\widehat{T})}$ are equivalent with constant independent of the mesh size) and the shape regularity of the mesh. Then we obtain (2.2.21) by summing last inequality over the elements $T \in \mathcal{T}_h$. To prove (2.2.20), it suffices to prove that

$$\|\mathrm{h}^{-1}(\nabla_h v_h - \nabla v_{h,i})\|_{L^2(\omega_i)} \lesssim \|D_h^2 v_h\|_{L^2(\omega_i)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h v_h]\|_{L^2(\Gamma^0_{h,i})} + \|\mathrm{h}^{-\frac{3}{2}}[v_h]\|_{L^2(\Gamma^0_{h,i})}, \tag{2.2.23}$$

where $\Gamma^0_{h,i}$ denotes the union of the edges $e \in \Gamma^0_h$ that belong to the interior to w_i. Since v_h belongs to a finite dimensional space, then according to standard norm equivalence and scaling arguments, it suffices to show that if the right-hand side of (2.2.23) vanishes then the left-hand side also vanishes. If the right-hand side equals to zero then $v_h \in \mathbb{P}_1(\omega_i)$ which implies that $v_{h,i} = v_h$ in ω_i. Hence, the left-hand side is zero. We emphasize again that the powers of the meshsize result from the scaling argument.

To prove (2.2.22), we first note that for any $w_h \in \mathbb{E}(\mathcal{T}_h)$ there holds

$$\begin{aligned}\|w_h - \frac{1}{|\Omega|}\int_\Omega w_h\|_{L^2(\Omega)} &\le \|w_h - \Pi_h w_h\|_{L^2(\Omega)} + \left\|\frac{1}{|\Omega|}\int_\Omega w_h - \frac{1}{|\Omega|}\int_\Omega \Pi_h w_h\right\|_{L^2(\Omega)} \\ &\quad + \left\|\Pi_h w_h - \frac{1}{|\Omega|}\int_\Omega \Pi_h w_h\right\|_{L^2(\Omega)} \\ &\lesssim \|w_h - \Pi_h w_h\|_{L^2(\Omega)} + \|\nabla \Pi_h w_h\|_{L^2(\Omega)} \\ &\lesssim \|\nabla_h w_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[w_h]\|_{L^2(\Gamma_h^0)}, \end{aligned} \tag{2.2.24}$$

where we use the triangle inequality for the first inequality, the Cauchy-Schwarz and the standard Poincaré-Friedrichs inequalities for the second one, and the estimate (2.2.19) for the third one. The hidden constant depends on Ω and is independent of h.

Then we apply (2.2.24) to $w_h = \nabla \Pi_h v_h \in \mathbb{V}_h^k$ and get

$$\|\nabla \Pi_h v_h - \frac{1}{|\Omega|}\int_\Omega \nabla \Pi_h v_h\|_{L^2(\Omega)} \lesssim \|D_h^2 \Pi_h v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla \Pi_h v_h]\|_{L^2(\Gamma_h^0)}.$$

Then we have

$$\begin{aligned}\|D_h^2 v_{h,i}\|_{L^2(\omega_i)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla v_{h,i}]\|_{L^2(\Gamma_{h,i}^0)} &\lesssim \|D_h^2 v_h\|_{L^2(\omega_i)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h v_h]\|_{L^2(\Gamma_{h,i}^0)} \\ &\quad + \|\mathrm{h}^{-\frac{3}{2}}[v_h]\|_{L^2(\Gamma_{h,i}^0)},\end{aligned}$$

since $v_h \in \mathbb{P}_1(\omega_i)$ implies that $v_{h,i} = v_h$ in ω_i and then implies the left-hand side is 0. Altogether we prove (2.2.22) as a consequence. □

We have the following inequalities related to $|||\cdot|||_{H_h^2(\Omega)}$.

Lemma 2.2.3 (discrete Poincaré-Friedrichs)**.** (i) Case $\Gamma_D \neq \emptyset$. For any discrete function $\mathbf{v}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ we have

$$\|\mathbf{v}_h\|_{L^2(\Omega)} + \|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \lesssim |||\mathbf{v}_h|||_{H_h^2(\Omega)} + \|\boldsymbol{\varphi}\|_{H^1(\Omega)} + \|\Phi\|_{H^1(\Omega)}. \tag{2.2.25}$$

Moreover, for any $\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ we have

$$\|\mathbf{v}_h\|_{L^2(\Omega)} + \|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \lesssim |||\mathbf{v}_h|||_{H_h^2(\Omega)}. \tag{2.2.26}$$

In both (2.2.25) and (2.2.26), the hidden constant depends on Ω and Γ_D.

(ii) Case $\Gamma_D = \emptyset$. For any $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$ we have

$$\|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \lesssim |||\mathbf{v}_h|||_{H_h^2(\Omega)} + \|\mathbf{v}_h\|_{L^2(\Omega)}, \tag{2.2.27}$$

where the hidden constant depends on Ω.

Proof. The proof of (i) is given in [22]. To prove (ii), applying (2.2.20) componentwisely and assuming $h \leq 1$ we have

$$\|\nabla_h \mathbf{v}_h - \nabla \Pi_h \mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h \mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{v}_h]\|_{L^2(\Gamma_h^0)}.$$

By (2.2.22) componentwisely and the triangle inequality we have

$$\|\nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|D_h^2\mathbf{v}_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} \qquad (2.2.28)$$
$$+ \|\frac{1}{|\Omega|}\int_\Omega \nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)}.$$

Note that for $\Pi_h\mathbf{v}_h \in [H^1(\Omega)]^3$, by integration by parts for each component there holds

$$\int_\Omega \nabla\Pi_h\mathbf{v}_h = \int_{\partial\Omega}(\Pi_h\mathbf{v}_h)\otimes\boldsymbol{\nu}_{\partial\Omega},$$

and thus using the Cauchy-Schwarz inequality, the trace inequality and Young's inequality we have

$$\begin{aligned}\|\frac{1}{|\Omega|}\int_\Omega \nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)} &= \frac{1}{|\Omega|^{\frac{1}{2}}}\Big|\int_{\partial\Omega}(\Pi_h\mathbf{v}_h)\otimes\boldsymbol{\nu}_{\partial\Omega}\Big| \lesssim \frac{|\partial\Omega|^{\frac{1}{2}}}{|\Omega|^{\frac{1}{2}}}\|\Pi_h\mathbf{v}_h\|_{L^2(\partial\Omega)}\\ &\lesssim \epsilon\|\nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)} + \frac{C}{\epsilon}\|\Pi_h\mathbf{v}_h\|_{L^2(\Omega)},\end{aligned}$$

with ϵ chosen small enough (depends on Ω but is independent of h) such that $\epsilon\|\nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)}$ can be absorbed into the left-hand side of (2.2.28). As a result,

$$\|\nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|D_h^2\mathbf{v}_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\Pi_h\mathbf{v}_h\|_{L^2(\Omega)},$$

where the hidden constant depends on Ω.

Using (2.2.21), we eventually obtain

$$\|\nabla\Pi_h\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|D_h^2\mathbf{v}_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{v}_h]\|_{L^2(\Gamma_h^0)} + \|\mathbf{v}_h\|_{L^2(\Omega)},$$

where the hidden constant depends on Ω. Combining the above relations we get (2.2.27) with a hidden constant depending on Ω. $\square$

2.3 Γ-convergence

We first introduce some properties of the discrete Hessian $H_h[\mathbf{v}_h]$ that will be used to prove the Γ-convergence of the discrete energy E_h to the continuous one E.

Lemma 2.3.1. *For any $\gamma_1, \gamma_0 > 0$ there exists a constant $C(\gamma_0, \gamma_1) > 0$ such that for any $\mathbf{v}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ and any $l_1, l_2 \geq 0$ we have*

$$C(\gamma_0, \gamma_1)|||\mathbf{v}_h|||^2_{H_h^2(\Omega)} \leq \int_\Omega |H_h[\mathbf{v}_h]|^2 + \gamma_1 \sum_{e \in \mathcal{E}_h^a} \int_e \mathrm{h}^{-1} |[\nabla_h \mathbf{v}_h]|^2 + \gamma_0 \sum_{e \in \mathcal{E}_h^a} \int_e \mathrm{h}^{-3} |[\mathbf{v}_h]|^2. \tag{2.3.1}$$

Moreover, the constant $C(\gamma_0, \gamma_1)$ tends to 0 when γ_0 or γ_1 tends to 0.

Proof. Let us write

$$\int_\Omega |H_h[\mathbf{v}_h]|^2 + \gamma_1 \sum_{e \in \mathcal{E}_h^a} \int_e \mathrm{h}^{-1} |[\nabla_h \mathbf{v}_h]|^2 + \gamma_0 \sum_{e \in \mathcal{E}_h^a} \int_e \mathrm{h}^{-3} |[\mathbf{v}_h]|^2 =: I_1 + I_2 + I_3. \tag{2.3.2}$$

We give a lower bound for the term I_1. We have

$$\begin{aligned} I_1 = \|D_h^2 \mathbf{v}_h\|^2_{L^2(\Omega)} + \| - R_h^{l_1}([\nabla_h \mathbf{v}_h]) + B_h^{l_2}([\mathbf{v}_h])\|^2_{L^2(\Omega)} \\ + 2 \int_\Omega D_h^2 \mathbf{v}_h : (-R_h^{l_1}([\nabla_h \mathbf{v}_h]) + B_h^{l_2}([\mathbf{v}_h])) \end{aligned}$$

Then

$$
\begin{aligned}
I_1 &\geq \|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)}^2 + \| - R_h^{l_1}([\nabla_h \mathbf{v}_h]) + B_h^{l_2}([\mathbf{v}_h])\|_{L^2(\Omega)}^2 \\
&\quad - 2\|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)} \|B_h^{l_2}([\mathbf{v}_h]) - R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)} \\
&\geq (1-\alpha^{-1})\|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)}^2 + (1-\alpha)\|B_h^{l_2}([\mathbf{v}_h]) - R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)}^2,
\end{aligned}
$$

where we used the Cauchy-Schwarz inequality for the first inequality and Young's inequality with $\alpha > 1$ for the second one. Similarly, using triangle and Young's inequalities we get

$$
\|B_h^{l_2}([\mathbf{v}_h]) - R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)}^2 \leq 2\|B_h^{l_2}([\mathbf{v}_h])\|_{L^2(\Omega)}^2 + 2\|R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)}^2.
$$

Since $\alpha > 1$, we have $1 - \alpha < 0$, and thus

$$
\begin{aligned}
I_1 &\geq (1-\alpha^{-1})\|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)}^2 + 2(1-\alpha)\|B_h^{l_2}([\mathbf{v}_h])\|_{L^2(\Omega)}^2 + 2(1-\alpha)\|R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)}^2 \\
&\geq (1-\alpha^{-1})\|D_h^2 \mathbf{v}_h\|_{L^2(\Omega)}^2 + 2(1-\alpha)C_0\|\mathrm{h}^{-3}[\mathbf{v}_h]\|_{L^2(\Gamma_h^a)}^2 \\
&\quad + 2(1-\alpha)C_1\|\mathrm{h}^{-1}[\nabla_h \mathbf{v}_h]\|_{L^2(\Gamma_h^a)}^2,
\end{aligned}
$$

where we used Lemma 2.2.1 for the second inequality with C_0 and C_1 some positive constants. Using the last two terms in (2.3.2) we obtain

$$
I_1 + I_2 + I_3 \geq \min\left\{1-\alpha^{-1}, 2(1-\alpha)C_0 + \gamma_0, 2(1-\alpha)C_1 + \gamma_1\right\} \|\mathbf{v}_h\|_{H_h^2(\Omega)}^2.
$$

Therefore, the result holds with

$$C(\gamma_0, \gamma_1) := \min\left\{1 - \alpha^{-1}, 2(1-\alpha)C_0 + \gamma_0, 2(1-\alpha)C_1 + \gamma_1\right\} \tag{2.3.3}$$

provided that for any $\gamma_0, \gamma_1 > 0$, we can choose $\alpha > 1$ such that

$$2(1-\alpha)C_0 + \gamma_0 > 0 \quad \text{and} \quad 2(1-\alpha)C_1 + \gamma_1 > 0.$$

We easily see that we can take any

$$\alpha \in \left(1, \min\left\{1 + \frac{\gamma_0}{2C_0}, 1 + \frac{\gamma_1}{2C_1}\right\}\right),$$

the interval being non-empty for any $\gamma_0, \gamma_1 > 0$. Finally, the coercivity constant in (2.3.3) tends to 0 since α tends to 1 as γ_0 or γ_1 tends to 0. □

Lemma 2.3.2 (Weak convergence of H_h)**.** *Let $\mathbf{v}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$. If $|||\mathbf{v}_h|||_{H_h^2(\Omega)} \leq C$ for all h and $\mathbf{v}_h \to \mathbf{v} \in [H^2(\Omega)]^3$ in $[L^2(\Omega)]^3$ as $h \to 0$, then for any $l_1, l_2 \geq 0$ we have*

$$H_h[\mathbf{v}_h] \rightharpoonup D^2\mathbf{v} \quad \text{in } \left[L^2(\Omega)\right]^{3\times 2\times 2} \quad \text{as } h \to 0. \tag{2.3.4}$$

Proof. For any $\phi \in [C_0^\infty]^{3\times 2\times 2}$ we have

$$\begin{aligned}
\int_\Omega H_h[\mathbf{v}_h] : \phi &= \int_\Omega D_h^2 \mathbf{v}_h : \phi - R_h^{l_1}([\nabla_h \mathbf{v}_h]) : \phi + B_h^{l_2}([\mathbf{v}_h]) : \phi \\
&= -\int_\Omega \nabla_h \mathbf{v}_h : \operatorname{div} \phi + \sum_{e\in\mathcal{E}_h^0} \int_e [\nabla_h \mathbf{v}_h] : \left\{\phi - \mathcal{I}_h^l \phi\right\} \mathbf{n}_e \\
&\quad - \int_\Omega R_h^{l_1}([\nabla_h \mathbf{v}_h])(\phi - \mathcal{I}_h^l \phi) + \int_\Omega B_h^{l_2}([\mathbf{v}_h]) : \phi \\
&= \int_\Omega \mathbf{v}_h \cdot \operatorname{div}(\operatorname{div} \phi) + \sum_{e\in\mathcal{E}_h^0} \int_e [\nabla_h \mathbf{v}_h] : \left\{\phi - \mathcal{I}_h^l \phi\right\} \mathbf{n}_e \\
&\quad - \int_\Omega R_h^{l_1}([\nabla_h \mathbf{v}_h])(\phi - \mathcal{I}_h^l \phi) + \int_\Omega B_h^{l_2}([\mathbf{v}_h]) : (\phi - \mathcal{I}_h^l \phi) \\
&\quad - \sum_{e\in\mathcal{E}_h^0} \int_e [\mathbf{v}_h] \cdot \left\{\operatorname{div}(\phi - \mathcal{I}_h^l \phi)\right\} \mathbf{n}_e + \sum_{e\in\mathcal{E}_h^D} \int_e (\mathbf{v}_h - \boldsymbol{\varphi}) \cdot \left\{\operatorname{div} \mathcal{I}_h^l \phi\right\} \mathbf{n}_e \\
&=: T_1 + T_2 + T_3 + T_4 + T_5 + T_6.
\end{aligned}$$

Here, $\mathcal{I}_h^l \phi \in [\mathbb{V}_h^l \cap H_0^1(\Omega)]^{3\times 2\times 2}$ denotes the Lagrange interpolant of ϕ, and $l := \min\{l_1, l_2\}$. Note that $T_6 = 0$ when $\Gamma_D = \emptyset$. We treat each term separately. Since $\mathbf{v}_h \to \mathbf{v} \in [H^2(\Omega)]^3$ in $[L^2(\Omega)]^3$ as $h \to 0$, we have

$$T_1 \to \int_\Omega \mathbf{v} \cdot \operatorname{div}(\operatorname{div} \phi) = -\int_\Omega \nabla \mathbf{v} : \operatorname{div} \phi = \int_\Omega D^2 \mathbf{v} : \phi \quad \text{as } h \to 0.$$

For T_3 we have that

$$\begin{aligned}
|T_3| &\le \|R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)} \|\phi - \mathcal{I}_h^l \phi\|_{L^2(\Omega)} \\
&\lesssim \|\mathrm{h}^{-\frac{1}{2}} [\nabla_h \mathbf{v}_h]\|_{L^2(\Gamma_h^a)} \|\phi - \mathcal{I}_h^l \phi\|_{L^2(\Omega)} \\
&\le C \|\mathcal{I}_h^l \phi - \phi\|_{L^2(\Omega)} \to 0
\end{aligned}$$

as $h \to 0$, where we used the uniform boundedness of $|||u_h|||_{H_h^2(\Omega)}$ and Lemma 2.2.1. The proof that T_4 converges to 0 as $h \to 0$ can be done in a similar way.

To bound the term T_2, we use the scaled trace inequality

$$\|\mathcal{I}_h^l\phi - \phi\|_{L^2(e)} \lesssim h_e^{-\frac{1}{2}}\|(\mathcal{I}_h^l\phi - \phi)\|_{L^2(\omega(e))} + h_e^{\frac{1}{2}}\|\nabla(\mathcal{I}_h^l\phi - \phi)\|_{L^2(\omega(e))}, \tag{2.3.5}$$

where $\omega(e)$ denotes the union of the two elements adjacent to $e \in \mathcal{E}_h^0$. We have

$$\begin{aligned}
|T_2| &= \left|\sum_{e\in\mathcal{E}_h^0}\int_e [\nabla_h \mathbf{v}_h] : \left\{\phi - \mathcal{I}_h^l\phi\right\}\mathbf{n}_e\right| \leq \sum_{e\in\mathcal{E}_h^0}\|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|_{L^2(e)}\|\mathrm{h}^{\frac{1}{2}}(\phi - \mathcal{I}_h^l\phi)\|_{L^2(e)} \\
&\lesssim \left(\sum_{e\in\mathcal{E}_h^0}\|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|_{L^2(e)}^2\right)^{\frac{1}{2}}\left(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h^l\phi - \phi\|_{L^2(T)}^2 + h^2\|\nabla(\mathcal{I}_h^l\phi - \phi)\|_{L^2(T)}^2\right)^{\frac{1}{2}} \\
&\to 0
\end{aligned}$$

as $h \to 0$, using again that $|||\mathbf{v}_h|||_{H_h^2(\Omega)}$ is uniformly bounded. We can proceed similarly to show that T_5 tends to 0 as $h \to 0$. Finally, when $\Gamma_D \neq \emptyset$, for T_6 we get

$$\begin{aligned}
|T_6| &\lesssim \sum_{e\in\mathcal{E}_h^D}\|\mathrm{h}^{-\frac{3}{2}}(\mathbf{v}_h - \boldsymbol{\varphi})\|_{L^2(e)}\|\mathrm{h}^{\frac{3}{2}}\operatorname{div}\mathcal{I}_h^l\phi\|_{L^2(e)} \\
&\lesssim \left(\sum_{e\in\mathcal{E}_h^D}\|\mathrm{h}^{-\frac{3}{2}}(\mathbf{v}_h - \boldsymbol{\varphi})\|_{L^2(e)}^2\right)^{\frac{1}{2}}\left(\sum_{T\in\mathcal{T}_h} h^2|\mathcal{I}_h^l\phi|_{H^1(T)}^2 + h^4|\mathcal{I}_h^l\phi|_{H^2(T)}^2\right)^{\frac{1}{2}} \to 0,
\end{aligned}$$

as $h \to 0$ since

$$\sum_{e\in\mathcal{E}_h^D} \|\mathrm{h}^{-\frac{3}{2}}(\mathbf{v}_h - \boldsymbol{\varphi})\|^2_{L^2(e)} = \sum_{e\in\mathcal{E}_h^D} \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{v}_h]\|^2_{L^2(e)} \leq |||\mathbf{v}_h|||^2_{H_h^2(\Omega)} \leq C^2.$$

To sum up, we have $\int_\Omega H_h[\mathbf{v}_h] : \phi \to \int_\Omega D^2\mathbf{v} : \phi$, and thus we have the weak convergence result. □

Lemma 2.3.3. *Let $\mathbf{v} \in [H^2(\Omega)]^3$ be any function such that, when $\Gamma_D \neq \emptyset$, $\mathbf{v} = \boldsymbol{\varphi}$ and $\nabla\mathbf{v} = \Phi$ on Γ_D. Moreover, let $\mathbf{v}_h := \mathcal{I}_h^k\mathbf{v} \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) \cap [H^1(\Omega)]^3$ be the Lagrange interpolant of $\mathbf{v}$. Then for any $l_1, l_2 \geq 0$ we have as $h \to 0$*

$$H_h[\mathbf{v}_h] \to D^2\mathbf{v} \quad \textit{strongly in} \quad [L^2(\Omega)]^{3\times2\times2}. \tag{2.3.6}$$

Proof. To prove (2.3.6), we will show that

$$D_h^2\mathbf{v}_h \to D^2\mathbf{v} \text{ in } [L^2(\Omega)]^{3\times2\times2} \tag{2.3.7}$$

and

$$\|R_h^{l_1}([\nabla_h\mathbf{v}_h])\|^2_{L^2(\Omega)} \to 0 \quad \text{and} \quad \|B_h^{l_2}([\mathbf{v}_h])\|^2_{L^2(\Omega)} \to 0 \tag{2.3.8}$$

as $h \to 0$. For the proof of (2.3.7), we use the H^2-stability

$$\|D^2\mathbf{v}_h\|_{L^2(T)} \lesssim |\mathbf{v}|_{H^2(T)}, \quad \forall\, T \in \mathcal{T}_h, \tag{2.3.9}$$

of the Lagrange interpolant. The inequality (2.3.9) can be shown using the Bramble-

Hilbert lemma and inverse inequalities (as shown in Section 9 of [22]). Moreover, from [22], we have

$$h_T^{-2}\|\mathbf{v}-\mathbf{v}_h\|_{L^2(T)} + h_T^{-1}\|\nabla(\mathbf{v}-\mathbf{v}_h)\|_{L^2(T)} \lesssim |\mathbf{v}|_{H^2(T)}. \tag{2.3.10}$$

Next, we consider $\mathbf{v}^\epsilon \in [C^\infty(\Omega)]^3$ a smooth mollifier of $\mathbf{v}$ such that $\mathbf{v}^\epsilon \to \mathbf{v}$ in $[H^2(\Omega)]^3$ as $\epsilon \to 0$. Then, thanks to (2.3.9) and (2.3.10) there exists a constant $C_1 > 0$ such that

$$\begin{aligned}\|D^2(\mathbf{v}_h-\mathbf{v})\|^2_{L^2(T)} &\lesssim \|D^2(\mathbf{v}_h-\mathbf{v}_h^\epsilon)\|^2_{L^2(T)} + \|D^2(\mathbf{v}_h^\epsilon-\mathbf{v}^\epsilon)\|^2_{L^2(T)} + \|D^2(\mathbf{v}^\epsilon-\mathbf{v})\|^2_{L^2(T)}\\ &\le C_1|\mathbf{v}-\mathbf{v}^\epsilon|^2_{H^2(T)} + C_2h_T^2|\mathbf{v}^\epsilon|^2_{H^3(T)} + |\mathbf{v}^\epsilon-\mathbf{v}|^2_{H^2(T)},\end{aligned}$$

where $\mathbf{v}_h^\epsilon := \mathcal{I}_h^k\mathbf{v}^\epsilon$. Hence, summing over $T \in \mathcal{T}_h$ and using $h_T \le h$ we get

$$\|D_h^2(\mathbf{v}_h-\mathbf{v})\|^2_{L^2(\Omega)} \le (1+C_1)|\mathbf{v}-\mathbf{v}^\epsilon|^2_{H^2(\Omega)} + C_2h^2|\mathbf{v}^\epsilon|^2_{H^3(\Omega)}.$$

Then, since $\|\mathbf{v}-\mathbf{v}^\epsilon\|_{H^2(\Omega)} \to 0$ as $\epsilon \to 0$, given any $\eta > 0$ we can pick ϵ small enough such that $\|\mathbf{v}-\mathbf{v}^\epsilon\|^2_{H^2(\Omega)} \le \eta/2$. Then we pick h small enough such that $C_2h^2|\mathbf{v}^\epsilon|^2_{H^3(\Omega)} \le \eta/2$, and consequently we conclude $\|D_h^2\mathbf{v}_h - D^2\mathbf{v}\|^2_{L^2(\Omega)} \to 0$ as $h \to 0$, which shows (2.3.7).

We now prove (2.3.8). By Lemma 2.2.1 we have

$$\|R_h^{l_1}([\nabla_h\mathbf{v}_h])\|^2_{L^2(\Omega)} \lesssim \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h\mathbf{v}_h]\|^2_{L^2(\Gamma_h^a)} = \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h(\mathbf{v}_h-\mathbf{v})]\|^2_{L^2(\Gamma_h^a)},$$

as $[\nabla \mathbf{v}] = 0$ on $e \in \mathcal{E}_h^0$ and $\nabla \mathbf{v} = \Phi$ on $e \in \mathcal{E}_h^D$ (when $\Gamma_D \neq \emptyset$). Moreover, by the scaled trace inequality (2.3.5) and (2.3.10) we obtain for any $e \in \mathcal{E}_h^a$

$$\begin{aligned}\|\mathrm{h}^{-\frac{1}{2}}[\nabla_h(\mathbf{v}_h - \mathbf{v})]\|_{L^2(e)}^2 &\lesssim h_e^{-2}\|\nabla_h(\mathbf{v}_h - \mathbf{v})\|_{L^2(\omega(e))}^2 + \|D_h^2(\mathbf{v}_h - \mathbf{v})\|_{L^2(\omega(e))}^2 \\ &= h_e^{-2}\|\nabla_h(\mathbf{v}_h - \mathbf{v} - \mathcal{I}_h^k(\mathbf{v}_h - \mathbf{v}))\|_{L^2(\omega(e))}^2 \\ &+ \|D_h^2(\mathbf{v}_h - \mathbf{v})\|_{L^2(\omega(e))}^2 \\ &\lesssim \sum_{T\in\omega(e)} \Big(\frac{h_T^2}{h_e^2} + 1\Big)|\mathbf{v}_h - \mathbf{v}|_{H^2(T)}^2,\end{aligned}$$

where $\omega(e)$ reduces to one element if $e \in \mathcal{E}_h^D$ (when $\Gamma_D \neq \emptyset$), and thus further by shape-regularity

$$\|R_h^{l_1}([\nabla_h \mathbf{v}_h])\|_{L^2(\Omega)}^2 \lesssim \sum_{T\in\omega(e)} |\mathbf{v}_h - \mathbf{v}|_{H^2(T)}^2 \to 0 \quad \text{as} \quad h \to 0.$$

Proceeding similarly, we can show that $\|B_h^{l_2}([\mathbf{v}_h])\|_{L^2(\Omega)}^2 \to 0$.

The strong convergence (2.3.6) of the reconstructed Hessians follows from the definition of the reconstructed Hessians and the strong convergence properties of $D_h^2\mathbf{v}_h$, $R_h^{l_1}([\nabla_h \mathbf{v}_h])$ and $B_h^{l_2}([\mathbf{v}_h])$ established previously.

□

The above properties of the discrete Hessian are now used to prove the coercivity and Γ-convergence of discrete energy $E_h[\mathbf{y}_h]$.

Theorem 2.3.1 (Coercivity)**.** *Let* $\mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ *and let* $\gamma_0, \gamma_1 > 0$. *When* $\Gamma_D = \emptyset$,

$$|||\mathbf{y}_h|||_{H_h^2(\Omega)}^2 \lesssim E_h[\mathbf{y}_h]. \tag{2.3.11}$$

When $\Gamma_D \neq \emptyset$,

$$|||\mathbf{y}_h|||^2_{H^2_h(\Omega)} \lesssim E_h[\mathbf{y}_h] + \|\boldsymbol{\varphi}\|^2_{H^1(\Omega)} + \|\Phi\|^2_{H^1(\Omega)} + \|\mathbf{f}\|^2_{L^2(\Omega)}. \tag{2.3.12}$$

The hidden constants of (2.3.11) *and* (2.3.12) *depend only on* μ, g *and the constant* $C(\gamma_0, \gamma_1)$ *that appears in* (2.3.1).

Proof. Let $\mathbb{V}$ denote the range space of $H_h^{l_1,l_2} : \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) \to [L^2(\Omega)]^{3\times 2\times 2}$. We first note that $(\int_\Omega |g^{-\frac{1}{2}} \cdot g^{-\frac{1}{2}}|^2)^{\frac{1}{2}} : \mathbb{V} \to \mathbb{R}$ is a norm as g is a symmetric positive definite matrix. Since $\mathbb{V}$ is a finite dimensional space, and all the norms are equivalent on a finite dimensional space, there exist two positive constants c_g and C_g depending only on g such that

$$C_g \|H_h^{l_1,l_2}(y_{h,k})\|^2_{L^2(\Omega)} \leq \int_\Omega |g^{-\frac{1}{2}} H_h^{l_1,l_2}(y_{h,k}) g^{-\frac{1}{2}}|^2 \leq c_g \|H_h^{l_1,l_2}(y_{h,k})\|^2_{L^2(\Omega)}.$$

Then by Lemma 2.3.1 and the fact that the trace term in (2.2.11) is positive, in the case $\Gamma_D = \emptyset$, for any $\gamma_0, \gamma_1 > 0$ we have

$$\begin{aligned} E_h[\mathbf{y}_h] \;&\geq\; C_g \frac{\mu}{12} \|H_h^{l_1,l_2}(\mathbf{y}_h)\|^2_{L^2(\Omega)} + \frac{\gamma_1}{2} \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h \mathbf{y}_h]\|^2_{L^2(\Gamma_h^a)} \\ &+\; \frac{\gamma_0}{2} \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{y}_h]\|^2_{L^2(\Gamma_h^a)} \geq \frac{1}{2} \min\{C_g \frac{\mu}{6}, 1\} C(\gamma_0, \gamma_1) |||\mathbf{y}_h|||^2_{H^2_h(\Omega)}, \end{aligned} \tag{2.3.13}$$

which proves (2.3.11). Recall that we assume $\mathbf{f} = 0$ for the *free boundary case*.

When $\Gamma_D \neq \emptyset$, the left-hand side of (2.3.13) becomes $E_h[\mathbf{y}_h] + \int_\Omega \mathbf{f} \cdot \mathbf{y}_h$. To estimate the forcing term, using Cauchy-Schwarz and Young's inequalities, as well

as Lemma 2.2.3 (i), we have for any $\tilde{\alpha} > 0$

$$\int_\Omega \mathbf{f} \cdot \mathbf{y}_h \leq \tilde{\alpha} \left(|||\mathbf{y}_h|||^2_{H^2_h(\Omega)} + \|\boldsymbol{\varphi}\|^2_{H^1(\Omega)} + \|\Phi\|^2_{H^1(\Omega)} \right) + \frac{3}{4\tilde{\alpha}} \|\mathbf{f}\|^2_{L^2(\Omega)}.$$

For any $\gamma_0, \gamma_1 > 0$ we can choose $\tilde{\alpha} > 0$ small enough such that

$$C_3 := \frac{1}{2} \min\{C_g \frac{\mu}{6}, 1\} C(\gamma_0, \gamma_1) - \tilde{\alpha} > 0.$$

Therefore, we get

$$C_3 |||\mathbf{y}_h|||^2_{H^2_h(\Omega)} \leq E_h[\mathbf{y}_h] + \tilde{\alpha} \|\boldsymbol{\varphi}\|^2_{H^1(\Omega)} + \tilde{\alpha} \|\Phi\|^2_{H^1(\Omega)} + \frac{3}{4\tilde{\alpha}} \|\mathbf{f}\|^2_{L^2(\Omega)}. \tag{2.3.14}$$

which concludes the proof of (2.3.12). □

Remark 2.3.1. *Note that the coercivity of E_h holds for any positive penalty parameter γ_0 and γ_1, and they do not necessarily have to be sufficiently large as in [22].*

Now, we turn to the Γ-convergence of $E_h[\mathbf{y}_h]$. In the remaining part of this section, we mainly focus on the *free boundary case*. The case with Dirichlet boundary conditions then naturally follows.

In the case $\Gamma_D = \emptyset$, the key point is a compactness result. Indeed, the coercivity of E_h only provides the uniform boundedness of the seminorm $\|\mathbf{y}_h\|_{H^2_h(\Omega)}$, which clearly cannot ensure compactness in $[L^2(\Omega)]^3$. In fact, the compactness can only hold for the sequence $\mathbf{y}_h$ after a proper rescaling for each h. Before proving this, see

Theorem 2.3.2 below, we first make the following observation for functions in the discrete admissible set $\mathbb{A}^k_{h,\boldsymbol{\epsilon}}$.

Lemma 2.3.4. *If* $\mathbf{y}_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$, *then* $\|\nabla_h \mathbf{y}_h\|^2_{L^2(\Omega)} \lesssim \boldsymbol{\epsilon} + \|g\|_{L^1(\Omega)}$.

Proof. Note that by the triangle inequality

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}\left|\int_T \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h\right| &\le \sum_{T\in\mathcal{T}_h}\left|\int_T \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g\right| + \sum_{T\in\mathcal{T}_h}\left|\int_T g\right| \\ &\le \boldsymbol{\epsilon} + \sum_{T\in\mathcal{T}_h}\int_T |g| \\ &\le \boldsymbol{\epsilon} + \|g\|_{L^1(\Omega)}.\end{aligned}$$

Then, we compute

$$\begin{aligned}\left|\int_T \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h\right|^2 = \sum_{i,j=1}^2 \left(\int_T \partial_i \mathbf{y}_h \cdot \partial_j \mathbf{y}_h\right)^2 &\ge \sum_{i=1}^2 \left(\int_T |\partial_i \mathbf{y}_h|^2\right)^2 \\ \ge \frac{1}{2}\left(\sum_{i=1}^2 \int_T |\partial_i \mathbf{y}_h|^2\right)^2 &= \frac{1}{2}\left(\int_T |\nabla \mathbf{y}_h|^2\right)^2.\end{aligned} \tag{2.3.15}$$

Hence,

$$\|\nabla_h \mathbf{y}_h\|^2_{L^2(\Omega)} = \sum_{T\in\mathcal{T}_h}\int_T |\nabla \mathbf{y}_h|^2 \lesssim \sum_{T\in\mathcal{T}_h}\left|\int_T \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h\right| \le \boldsymbol{\epsilon} + \|g\|_{L^1(\Omega)}.$$

This completes the proof. □

Theorem 2.3.2 (Compactness for the *free boundary case*)**.** *Assume that* $\Gamma_D = \emptyset$ *and let* $\{\mathbf{y}_h\} \subset \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$ *be a sequence such that* $E_h[\mathbf{y}_h] \le C$ *uniformly. Then there exists a shifted sequence* $\bar{\mathbf{y}}_h := \mathbf{y}_h - \mathbf{c}_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$ *with constant* $\mathbf{c}_h \in \mathbb{R}^3$ *and* $\mathbf{y} \in [H^2(\Omega)]^3$

such that (up to a subsequence) $\bar{\mathbf{y}}_h \to \mathbf{y}$ *and* $\nabla_h \bar{\mathbf{y}}_h \to \nabla \mathbf{y}$ *in* $[L^2(\Omega)]^3$ *as* $h, \boldsymbol{\epsilon} \to 0$ *provided that* $h \leq 1$.

Proof. Take $\mathbf{c}_h := \frac{1}{|\Omega|}\int_\Omega \mathbf{y}_h$. Then by the Poincaré-Friedrichs inequality (2.2.24) applying to $\mathbf{y}_h \in [\mathbb{V}_h^k]^3$ we have

$$\|\bar{\mathbf{y}}_h\|^2_{L^2(\Omega)} = \|\mathbf{y}_h - \mathbf{c}_h\|^2_{L^2(\Omega)} \lesssim \|\nabla_h \mathbf{y}_h\|^2_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[\mathbf{y}_h]\|^2_{L^2(\Gamma_h^0)}. \tag{2.3.16}$$

By Lemma 2.3.4 and the fact that $\nabla_h \bar{\mathbf{y}}_h = \nabla_h \mathbf{y}_h$, we have that $\|\nabla_h \bar{\mathbf{y}}_h\|_{L^2(\Omega)}$ is uniformly bounded. Since $E_h[\mathbf{y}_h] \leq C$ by assumption and E_h is coercive, we have that $\|\mathbf{y}_h\|_{H_h^2(\Omega)}$ is uniformly bounded and thus $\|D_h^2 \bar{\mathbf{y}}_h\|_{L^2(\Omega)} = \|D_h^2 \mathbf{y}_h\|_{L^2(\Omega)}$ is uniformly bounded. The uniform boundedness of $\|\mathbf{y}_h\|_{H_h^2(\Omega)}$, recall $h \leq 1$ by assumption, also implies that

$$\|\mathrm{h}^{-\frac{1}{2}}[\mathbf{y}_h]\|^2_{L^2(\Gamma_h^0)} \lesssim \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{y}_h]\|^2_{L^2(\Gamma_h^0)} \leq \tilde{C},$$

and thus using (2.3.16) we deduce that $\|\bar{\mathbf{y}}_h\|^2_{L^2(\Omega)}$ is also uniformly bounded. It is clear that $\bar{\mathbf{y}}_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$, $[\mathbf{y}_h] = [\bar{\mathbf{y}}_h]$ on each $e \in \mathcal{E}_h^0$. Moreover, as we assume that $\mathbf{f} = 0$ in the *free boundary case*, we have $E_h[\bar{\mathbf{y}}_h] = E_h[\mathbf{y}_h]$.

Then we can apply the same argument used for the case $\Gamma_D \neq \emptyset$ in [22, Proposition 5.1, step1-step3] to conclude the claimed compactness result. It is therefore only sketched. The uniform bound in L^2 guarantees that $\bar{\mathbf{y}}_h$ converges weakly (up to a subsequence) in $[L^2(\Omega)]^3$ to some $\mathbf{y}$. Setting $\mathbf{z}_h := \Pi_h \mathbf{y}_h - (1/|\Omega|)\int_\Omega \Pi_h \mathbf{y}_h \in [\mathbb{V}_h^k \cap H^1(\Omega)]^3$, we invoke the Poincaré-Friedrichs inequality (2.2.24) coupled with the

$[H^1(\Omega)]^3$ stability (2.2.19) of Π_h to deduce that $\mathbf{z}_h$ is uniformly bounded in $[H^1(\Omega)]^3$. As a consequence, $\mathbf{z}_h$ converges strongly (up to a subsequence) in $[L^2(\Omega)]^3$ to some $\mathbf{z} \in [H^1(\Omega)]^3$. To show that $\mathbf{y} = \mathbf{z}$, we note that the interpolation property (2.2.20), Poincaré-Friedrichs inequality (2.2.24) and the uniform boundedness of $E_h[\mathbf{y}_h]$ yield $\|\bar{\mathbf{y}}_h - \mathbf{z}_h\|_{L^2(\Omega)} \to 0$ as $h \to 0$. Consequently, we also have

$$\|\bar{\mathbf{y}}_h - \mathbf{z}\|_{L^2(\Omega)} \leq \|\bar{\mathbf{y}}_h - \mathbf{z}_h\|_{L^2(\Omega)} + \|\mathbf{z}_h - \mathbf{z}\|_{L^2(\Omega)} \to 0$$

as $h \to 0$. The uniqueness of weak limits guarantees that $\mathbf{y} = \mathbf{z}$ and thus $\bar{\mathbf{y}}_h$ strongly converges (up to a subsequence) in $[L^2(\Omega)]^3$ to $\mathbf{y} \in [H^1(\Omega)]^3$. Repeating this argument for $\nabla_h \mathbf{y}_h = \nabla_h \bar{\mathbf{y}}_h$ yields that $\nabla_h \bar{\mathbf{y}}_h$ strongly converges (up to a subsequence) in $[L^2(\Omega)]^{3\times 2}$ to $\nabla \mathbf{y}$. □

Remark 2.3.2. *In the* free boundary case, *define* $\mathbb{K}_h := \{\mathbf{w}_h \in [\mathbb{V}_h^k]^3 : E_h[\mathbf{y}_h + \mathbf{w}_h] = E_h[\mathbf{y}_h]$ *for all* $\mathbf{y}_h \in [\mathbb{V}_h^k]^3\}$ *and* $\mathbb{D}_h := \{\mathbf{w}_h \in [\mathbb{V}_h^k]^3 : \mathbf{y}_h + \mathbf{w}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ *for all* $\mathbf{y}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k\}$. *Then it is clear that if* $\mathbf{y}_h$ *is a solution to* (2.2.16) *then* $\mathbf{y}_h + \mathbf{w}_h$ *is also a solution for any* $\mathbf{w}_h \in \mathbb{K}_h \cap \mathbb{D}_h$. *Constants belong to* $\mathbb{K}_h \cap \mathbb{D}_h$.

Theorem 2.3.3 (Lim-inf of E_h)**.** *Assume that* $\Gamma_D = \emptyset$. *Let* $l_1, l_2 \geq 0$ *and let the prestrain defect parameter* $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(h) \to 0$ *as* $h \to 0$. *Let* $\{\mathbf{y}_h\} \subset \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ *be a sequence such that* $E_h[\mathbf{y}_h] \leq C$ *uniformly. Then there exists* $\mathbf{y} \in \mathbb{A}$ *and a shifted sequence* $\bar{\mathbf{y}}_h := \mathbf{y}_h - \mathbf{c}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ *with constant* $\mathbf{c}_h \in \mathbb{R}^3$ *such that (up to a subsequence)* $\bar{\mathbf{y}}_h \to \mathbf{y}$ *in* $[L^2(\Omega)]^3$ *as* $h \to 0$, *and* $E[\mathbf{y}] \leq \liminf_{h\to 0} E_h[\bar{\mathbf{y}}_h] = \liminf_{h\to 0} E_h[\mathbf{y}_h]$.

Proof. If $\mathbf{c}_h := \frac{1}{|\Omega|}\int_\Omega \mathbf{y}_h$, by Theorem 2.3.2, we have $\bar{\mathbf{y}}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ and there exists

$\mathbf{y} \in [H^2(\Omega)]^3$ such that $\bar{\mathbf{y}}_h \to \mathbf{y}$ and $\nabla_h \bar{\mathbf{y}}_h \to \nabla \mathbf{y}$ in $[L^2(\Omega)]^3$ as $h \to 0$ up to a subsequence. Moreover, $E_h[\bar{\mathbf{y}}_h] = E_h[\mathbf{y}_h]$, $\|D_h^2 \bar{\mathbf{y}}_h\|_{L^2(\Omega)}$ and $\|\nabla_h \bar{\mathbf{y}}_h\|_{L^2(\Omega)}$ are uniformly bounded.

To prove that $\mathbf{y} \in \mathbb{A}$, we also need to show that $\mathbf{y}$ satisfies the constraint $\nabla \mathbf{y}^T \nabla \mathbf{y} = g$ a.e. in Ω. We proceed in two steps. First, notice that

$$\sum_{T \in \mathcal{T}_h} \left| \int_T (\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g) \right| \leq \boldsymbol{\epsilon}$$

implies that

$$\sum_{T \in \mathcal{T}_h} \|\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g\|_{L^1(T)} \leq ch + \boldsymbol{\epsilon}. \tag{2.3.17}$$

Indeed, we have

$$\begin{aligned}
\sum_{T \in \mathcal{T}_h} \|\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g\|_{L^1(T)} &\leq \sum_{T \in \mathcal{T}_h} \left\| \nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g - \frac{1}{|T|} \int_T (\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g) \right\|_{L^1(T)} \\
&\quad + \sum_{T \in \mathcal{T}_h} \left\| \frac{1}{|T|} \int_T (\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g) \right\|_{L^1(T)} \\
&\lesssim \sum_{T \in \mathcal{T}_h} h_T \|\nabla (\nabla \bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h) - \nabla g\|_{L^1(T)} \\
&\quad + \sum_{T \in \mathcal{T}_h} \left| \int_T \nabla (\bar{\mathbf{y}}_h^T \nabla \bar{\mathbf{y}}_h - g) \right| \\
&\lesssim \sum_{T \in \mathcal{T}_h} h_T \left(\|D^2 \bar{\mathbf{y}}_h\|_{L^2(T)} \|\nabla \bar{\mathbf{y}}_h\|_{L^2(T)} + \|\nabla g\|_{L^1(T)} \right) + D_h[\bar{\mathbf{y}}_h] \\
&\leq h \|D_h^2 \bar{\mathbf{y}}_h\|_{L^2(\Omega)} \|\nabla_h \bar{\mathbf{y}}_h\|_{L^2(\Omega)} + h \|\nabla g\|_{L^1(\Omega)} + D_h[\bar{\mathbf{y}}_h],
\end{aligned}$$

where the second inequality follows from the Poincaré inequality and the third one uses Hölder's inequality and the definition of the prestrain defect given in (2.2.15).

Since $\|D_h^2\bar{\mathbf{y}}_h\|_{L^2(\Omega)}$ and $\|\nabla_h\bar{\mathbf{y}}_h\|_{L^2(\Omega)}$ are uniformly bounded, (2.3.17) follows from the regularity assumption on g and the assumption that $D_h[\bar{\mathbf{y}}_h] \leq \boldsymbol{\epsilon}$.

Then, since

$$\|\nabla\mathbf{y}^T\nabla\mathbf{y} - g\|_{L^1(\Omega)} \leq \sum_{T\in\mathcal{T}_h} \left(\|\nabla\mathbf{y}^T\nabla(\mathbf{y}-\bar{\mathbf{y}}_h)\|_{L^1(T)} + \|\nabla(\mathbf{y}-\bar{\mathbf{y}}_h)^T\nabla\bar{\mathbf{y}}_h\|_{L^1(T)}\right) + \sum_{T\in\mathcal{T}_h} \|\nabla\bar{\mathbf{y}}_h^T\nabla\bar{\mathbf{y}}_h - g\|_{L^1(T)},$$

we have

$$\|\nabla\mathbf{y}^T\nabla\mathbf{y} - g\|_{L^1(\Omega)} \leq \left(\|\nabla\mathbf{y}\|_{L^2(\Omega)} + \|\nabla_h\bar{\mathbf{y}}_h\|_{L^2(\Omega)}\right)\|\nabla_h\bar{\mathbf{y}}_h - \nabla\mathbf{y}\|_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} \|\nabla\bar{\mathbf{y}}_h^T\nabla\bar{\mathbf{y}}_h - g\|_{L^1(T)},$$

which goes to 0 as $h, \boldsymbol{\epsilon} \to 0$. Here, we used the fact that $\nabla_h\bar{\mathbf{y}}_h \to \nabla\mathbf{y}$ in $[L^2(\Omega)]^3$, the uniform boundedness of $\nabla_h\bar{\mathbf{y}}_h$ and $\nabla\mathbf{y}$, and (2.3.17). Thus, $\nabla\mathbf{y}^T\nabla\mathbf{y} = g$ a.e. in Ω, and hence $\mathbf{y} \in \mathbb{A}$.

Now we prove the lim-inf property. Due to Lemma 2.3.2 we have $H_h[\bar{y}_{h,k}] \rightharpoonup D^2 y_k$ as $h \to 0$ for $k = 1, 2, 3$. Therefore, $g^{-\frac{1}{2}}H_h[\bar{y}_{h,k}]g^{-\frac{1}{2}} \rightharpoonup g^{-\frac{1}{2}}D^2y_k g^{-\frac{1}{2}}$ as $h \to 0$. By weakly lower-semicontinuity of the $L^2(\Omega)$ norm, we have that

$$\int_\Omega |g^{-\frac{1}{2}}D^2y_k g^{-\frac{1}{2}}|^2 \leq \liminf_{h\to 0}\int_\Omega |g^{-\frac{1}{2}}H_h[\bar{y}_{h,k}]g^{-\frac{1}{2}}|^2.$$

Now we focus on the trace term of E_h and E. Since we have $g^{-\frac{1}{2}}H_h[\bar{y}_{h,k}]g^{-\frac{1}{2}} \rightharpoonup g^{-\frac{1}{2}}D^2y_k g^{-\frac{1}{2}}$, it suffices to prove the weakly lower semicontinuity of the following

functional:

$$\int_\Omega \mathrm{tr}(F)^2 \leq \liminf_{h\to 0} \int_\Omega \mathrm{tr}(F_h)^2,$$

where $F_h, F \in [L^2(\Omega)]^{2\times 2}$ are such that $F_h \rightharpoonup F$ as $h \to 0$.

Consider the functional $p : [L^2(\Omega)]^{2\times 2} \to \mathbb{R}$ defined by $p(F) := (\int_\Omega \mathrm{tr}(F)^2)^{\frac{1}{2}}$. We can easily see that p satisfies the triangle inequality, is a convex function and is actually a semi-norm. Thus for any $\beta > 0$ the set $S := \{F \in [L^2(\Omega)]^{2\times 2} : p(F) \leq \beta\}$ is convex. Moreover, since the following estimate holds true

$$p(F) = \|\mathrm{tr}(F)\|_{L^2(\Omega)} = \|F : I\|_{L^2(\Omega)} \lesssim \|F\|_{L^2(\Omega)},$$

we infer that S is closed in the topology with respect to the $L^2(\Omega)$ norm. Since the convex set's closure is the closure in the weak topology, S is then also closed in the weak topology. This implies that $p(F) \leq \liminf_{h\to 0} p(F_h)$ whenever $F_h \rightharpoonup F$ as $h \to 0$. Therefore,

$$\int_\Omega \mathrm{tr}(g^{-\frac{1}{2}} D^2 y_k g^{-\frac{1}{2}})^2 \leq \liminf_{h\to 0} \int_\Omega \mathrm{tr}(g^{-\frac{1}{2}} H_h[\bar{y}_{h,k}] g^{-\frac{1}{2}})^2.$$

Finally, since the stabilization terms in E_h are positive, we have $E[\mathbf{y}] \leq \liminf_{h\to 0} E_h[\bar{\mathbf{y}}_h] = \liminf_{h\to 0} E_h[\mathbf{y}_h]$. □

Remark 2.3.3 (Lim-inf for the Dirichlet boundary case)**.** *When $\Gamma_D \neq \emptyset$, thanks to Theorem 2.3.1 (coercivity of E_h) and equation (2.2.25) of Lemma 2.2 (discrete Poincaré-Friedrich inequality), there exists $\mathbf{y} \in [H^2(\Omega)]^3$ satisfying the* given boundary conditions *and such that, up to a subsequence, $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$ and $\nabla_h \mathbf{y}_h \to$*

$\nabla \mathbf{y}$ *in* $[L^2(\Omega)]^{3\times 2}$ *as* $h \to 0$. *The proof of this compactness result for the Dirichlet boundary condition case is the same as Proposition 5.1 in [22]. Then the proof of the lim-inf condition when* $\Gamma_D \neq \emptyset$ *follows naturally as in the proof of Theorem 2.3.3 with* $\bar{\mathbf{y}}_h$ *replaced by* $\mathbf{y}_h$. *Finally, note that the possible presence of a forcing term in this case is not be a problem as* $\int_\Omega \mathbf{f} \cdot \mathbf{y}_h \to \int_\Omega \mathbf{f} \cdot \mathbf{y}$ *when* $\mathbf{y}_h \to \mathbf{y}$ *in* $[L^2(\Omega)]^3$.

Theorem 2.3.4 (Lim-sup of E_h)**.** *Let* $l_1, l_2 \geq 0$. *For any* $\mathbf{y} \in \mathbb{A}$, *there exists* $\{\mathbf{y}_h\} \subset \mathbb{A}^k_{h,\boldsymbol{\epsilon}} \cap [H^1(\Omega)]^3$ *such that*

$$\mathbf{y}_h \to \mathbf{y} \quad \textit{in} \quad [L^2(\Omega)]^3 \qquad \textit{as } h \to 0$$

and

$$E[\mathbf{y}] \geq \limsup_{h\to 0} E_h[\mathbf{y}_h],$$

provided that $\boldsymbol{\epsilon} \geq Ch\|y\|^2_{H^2(\Omega)}$ *for some positive constant* C.

Proof. Assume firstly that $\Gamma_D = \emptyset$. Let $\mathbf{y}_h = \mathcal{I}^k_h \mathbf{y} \in \mathbb{V}^k_h(\boldsymbol{\varphi}, \Phi) \cap [H^1(\Omega)]^3$ be the Lagrange interpolant of $\mathbf{y}$. By Lemma 2.3.3 we have that $H_h[y_{h,k}] \to D^2 y_k$ strongly in $[L^2(\Omega)]^{2\times 2}$ as $h \to 0$, and thus $g^{-\frac{1}{2}} H_h[y_{h,k}] g^{-\frac{1}{2}} \to g^{-\frac{1}{2}} D^2 y_k g^{-\frac{1}{2}}$ in $[L^2(\Omega)]^{2\times 2}$ as $h \to 0$. Therefore,

$$\lim_{h\to 0} \int_\Omega |g^{-\frac{1}{2}} H_h[y_{h,k}] g^{-\frac{1}{2}}|^2 = \int_\Omega |g^{-\frac{1}{2}} D^2 y_k g^{-\frac{1}{2}}|^2.$$

For the trace term, proceeding as in the proof of Theorem 2.3.3, it suffices to show that the condition $F_h \to F$ in $[L^2(\Omega)]^{2\times 2}$ as $h \to 0$ implies $\int_\Omega \mathrm{tr}(F)^2 \geq$

$\limsup_{h\to\infty} \int_\Omega \mathrm{tr}(F_h)^2$. For $p(F) := (\int_\Omega \mathrm{tr}(F)^2)^{\frac{1}{2}}$, we have $p(F_h) \leq p(F) + p(F_h - F)$. If we take lim sup on both sides, we get

$$\begin{aligned}\limsup_{h\to 0} p(F_h) &\leq p(F) + \limsup_{h\to 0} p(F_h - F) \leq p(F) + \limsup_{h\to 0} \|\mathrm{tr}(F_h - F)\|_{L^2} \\ &\leq p(F) + c \limsup_{h\to 0} \|F_h - F\|_{L^2} = p(F),\end{aligned}$$

with $c = \sqrt{2}|\Omega|^{\frac{1}{2}}$, and thus

$$\limsup_{h\to 0} \int_\Omega \mathrm{tr}(g^{-\frac{1}{2}} H_h[y_{h,k}] g^{-\frac{1}{2}})^2 \leq \int_\Omega \mathrm{tr}(g^{-\frac{1}{2}} D^2 y_k g^{-\frac{1}{2}})^2.$$

Finally, as the stabilization terms in E_h tend to 0 for this particular choice of $\mathbf{y}_h$, see Lemma 2.3.3, we have $E[\mathbf{y}] \geq \limsup_{h\to 0} E_h[\mathbf{y}_h]$.

To conclude the proof, we also need to show that $\mathbf{y}_h = \mathcal{I}_h^k \mathbf{y} \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) \cap [H^1(\Omega)]^3$ is in $\mathbb{A}_{h,\boldsymbol{\epsilon}}^k$, namely that $\mathbf{y}_h$ satisfies

$$D_h[\mathbf{y}_h] = \sum_{T\in\mathcal{T}_h} \left| \int_T \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g \right| \leq \boldsymbol{\epsilon}.$$

Since $\nabla \mathbf{y}^T \nabla \mathbf{y} = g$ a.e. in Ω, we have

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} \|\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g\|_{L^1(T)} &\leq \sum_{T\in\mathcal{T}_h} \left(\|\nabla \mathbf{y}\|_{L^2(T)} + \|\nabla \mathbf{y}_h\|_{L^2(T)}\right) \|\nabla(\mathbf{y}-\mathbf{y}_h)\|_{L^2(T)} \\ &\quad + \sum_{T\in\mathcal{T}_h} \|\nabla \mathbf{y}^T \nabla \mathbf{y} - g\|_{L^1(T)}) \\ &\lesssim \sum_{T\in\mathcal{T}_h} h_T \|\nabla \mathbf{y}\|_{L^2(T)} |\mathbf{y}|_{H^2(T)} + \sum_{T\in\mathcal{T}_h} h_T^2 |\mathbf{y}|^2_{H^2(T)} \\ &\leq h\|\nabla \mathbf{y}\|_{L^2(\Omega)} |\mathbf{y}|_{H^2(\Omega)} + h^2 |\mathbf{y}|^2_{H^2(\Omega)} \leq 2h\|y\|^2_{H^2(\Omega)},\end{aligned}$$

where the second inequality follows from the fact that $\|\nabla(\mathbf{y}-\mathbf{y}_h)\|_{L^2(T)} \lesssim h_T |\mathbf{y}|_{H^2(T)}$ and $\|\nabla \mathbf{y}_h\|_{L^2(T)} \lesssim \|\nabla \mathbf{y}\|_{L^2(T)} + h_T |\mathbf{y}|_{H^2(T)}$. Then for $\boldsymbol{\epsilon} \geq Ch\|y\|^2_{H^2(\Omega)}$ we have

$$\sum_T \left| \int_T (\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g) \right| \leq \sum_T \|\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - g\|_{L^1(T)} \leq \boldsymbol{\epsilon},$$

i.e., $\mathbf{y}_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$.

The same procedure can be applied in the case $\Gamma_D \neq \emptyset$, using additionally that $\lim_{h\to 0} \int_\Omega \mathbf{f}_h \cdot \mathbf{y}_h = \int_\Omega \mathbf{f} \cdot \mathbf{y}$. □

Recall that by our assumption of immersibility of g and compatibility of boundary data, we know that $\mathbb{A} \neq \emptyset$. By above argument in the proof of Theorem 2.3.4, we know that the discrete admissible set $\mathbb{A}^k_{h,\boldsymbol{\epsilon}}$ is not empty, provided that $\boldsymbol{\epsilon}$ is large enough. Indeed, for any $\mathbf{y} \in \mathbb{A}$, let $\mathbf{y}_h := \mathcal{I}^k_h \mathbf{y} \in \mathbb{V}^k_h(\boldsymbol{\varphi}, \Phi) \cap [H^1(\Omega)]^3$ be the standard Lagrange interpolant of $\mathbf{y}$. Then there exists a constant $C > 0$ depending only on

the shape regularity of $\mathcal{T}_h$ and Ω such that

$$D_h[\mathbf{y}_h] \leq Ch\|\mathbf{y}\|^2_{H^2(\Omega)}. \tag{2.3.18}$$

In particular, $\mathbb{A}^k_{h,\boldsymbol{\epsilon}} \neq \emptyset$ provided $\boldsymbol{\epsilon} \geq Ch\|\mathbf{y}\|^2_{H^2(\Omega)}$.

This result can then be used to show the existence of a minimizer of the discrete energy E_h within $\mathbb{A}_{h,\boldsymbol{\epsilon}}$, see Proposition 2.3.1 below.

Proposition 2.3.1. *Let $h > 0$ be fixed and let $\boldsymbol{\epsilon}$ be large enough such that $\mathbb{A}^k_{h,\boldsymbol{\epsilon}}$ is not empty. Then there exists at least one solution to Problem (2.2.16).*

Proof. Let us first consider the *free boundary case* $\Gamma_D = \emptyset$. Let $m := \inf_{\mathbf{y}_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}}} E_h[\mathbf{y}_h]$ be finite. Let $\{\mathbf{y}^n_h\}_{n\geq 1} \subset \mathbb{A}^k_{h,\boldsymbol{\epsilon}}$ be a minimizing sequence, i.e., such that

$$\lim_{n\to\infty} E_h[\mathbf{y}^n_h] = m. \tag{2.3.19}$$

Because $E_h[\mathbf{y}^n_h + \mathbf{c}] = E_h[\mathbf{y}^n_h]$ and $D_h[\mathbf{y}^n_h + \mathbf{c}] = D_h[\mathbf{y}^n_h]$ for any constant vector $\mathbf{c} \in \mathbb{R}^3$ (as $\mathbf{f} = 0$ for free boundary case), we can assume without loss of generality that $\int_\Omega \mathbf{y}^n_h = 0$. Consequently, from the Poincaré-Friedrichs estimate (2.2.24), the control of the gradients provided by Lemma 2.3.4 and the coercivity of the energy E_h (Theorem 2.3.1), we deduce that $\|\mathbf{y}^n_h\|_{L^2(\Omega)} \lesssim 1$. Because $[\mathbb{V}^k_h]^3$ is finite dimensional, we have that (up to a subsequence) $\{\mathbf{y}^n_h\}_{n\geq 1}$ converges strongly in $[L^2(\Omega)]^3$ to some $\mathbf{y}^\infty_h \in [\mathbb{V}^k_h]^3$. In turn, the continuity of the quadratic energy E_h and the prestrain

defect D_h guarantee that

$$E_h[\mathbf{y}_h^\infty] = \lim_{n\to\infty} E_h[\mathbf{y}_h^n] = \inf_{\mathbf{y}_h \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k} E_h[\mathbf{y}_h]$$

and $\mathbf{y}_h^\infty \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$. This proves that $\mathbf{y}_h^\infty$ is one solution to the minimization problem (2.2.16).

We can proceed in a similar way for the case $\Gamma_D \neq \emptyset$, except that

$$\sup_{n\geq 1} \|\mathbf{y}_h^n\|_{L^2(\Omega)} < \infty$$

directly follows from Theorem 2.3.1 and the Poincaré-Friedrichs inequality (2.2.25). In particular, the $\mathbf{y}_h^n$ do not need to have mean-value zero. □

To summarize, we can conclude that $E_h[\mathbf{y}_h]$ Γ-converges to $E[\mathbf{y}]$ as $h \to 0$ for both the Dirichlet and free boundary conditions, which implies the convergence (up to a subsequence) of almost global minimizers of the discrete problem (2.2.16) to global minimizer of continuous problem (2.1.23) as in [22].

2.4 Discrete gradient flow

To find a local minimizer $\mathbf{y}_h$ of $E_h[\mathbf{y}_h]$ within $\mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ for a fixed mesh, namely to solve the non-convex constrained minimization problem (2.2.16), we execute a discrete H^2 gradient flow.

The H_h^2 metric is defined by (2.2.17) in the case $\Gamma_D \neq \emptyset$ while in the *free boundary case* $\Gamma_D = \emptyset$, we add the term $\sigma(\cdot,\cdot)_{L^2(\Omega)}$ with $\sigma > 0$. To have a unified

notation, we introduce

$$\begin{aligned}(\mathbf{v}_h, \mathbf{w}_h)_{H^2_h(\Omega)} &:= \sigma(\mathbf{v}_h, \mathbf{w}_h)_{L^2(\Omega)} + (D^2_h \mathbf{v}_h, D^2_h \mathbf{w}_h)_{L^2(\Omega)} \\ &+ (\mathrm{h}^{-1}[\nabla_h \mathbf{v}_h], [\nabla_h \mathbf{w}_h])_{L^2(\Gamma^a_h)} + (\mathrm{h}^{-3}[\mathbf{v}_h], [\mathbf{w}_h])_{L^2(\Gamma^a_h)}, \quad (2.4.1)\end{aligned}$$

where $\sigma = 0$ when $\Gamma_D \neq \emptyset$ and $\sigma > 0$ when $\Gamma_D = \emptyset$. Moreover, we define

$$\|\mathbf{v}_h\|^2_{H^2_h(\Omega)} := (\mathbf{v}_h, \mathbf{v}_h)_{H^2_h(\Omega)}.$$

We emphasize that we have $(\cdot,\cdot)_{H^2_h(\Omega)} = \langle \cdot, \cdot \rangle_{H^2_h(\Omega)}$ and $\|\cdot\|_{H^2_h(\Omega)} = |||\cdot|||_{H^2_h(\Omega)}$ when $\Gamma_D \neq \emptyset$. Note that $\|\cdot\|_{H^2_h(\Omega)}$ defines a norm and $(\cdot,\cdot)_{H^2_h(\Omega)}$ defines a scalar product on $[\mathbb{V}^k_h]^3$ if $\Gamma_D = \emptyset$, and on $\mathbb{V}^k_h(\mathbf{0}, \mathbf{0})$ if $\Gamma_D \neq \emptyset$. They are not norm or scalar product respectively on $V^k_h(\boldsymbol{\varphi}, \Phi)$ if $\Gamma_D \neq \emptyset$ because non-homogenous boundary data are involved in jumps on boundary edges, but this is not problem as we shall see $(\cdot,\cdot)_{H^2_h(\Omega)}$ is only applied to $\mathbb{V}^k_h(\mathbf{0}, \mathbf{0})$ in the gradient flow for Dirichlet boundary case. Moreover, we deduce from the estimate (2.2.27) for the *free boundary case* that

$$\|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{H^2_h(\Omega)}, \quad (2.4.2)$$

where the hidden constant now depends on σ and Ω.

The discrete H^2 gradient reads as follows. Given an initial guess $\mathbf{y}^0_h \in \mathbb{A}^k_{h,\boldsymbol{\epsilon}_0}$ and a pseudo time-step $\tau > 0$, we iteratively compute $\mathbf{y}^{n+1}_h := \mathbf{y}^n_h + \delta \mathbf{y}^{n+1}_h \in \mathbb{V}^k_h(\boldsymbol{\varphi}, \Phi)$ that minimizes

$$\mathbf{y}_h \mapsto \frac{1}{2\tau} \|\mathbf{y}_h - \mathbf{y}^n_h\|^2_{H^2_h(\Omega)} + E_h[\mathbf{y}_h], \quad (2.4.3)$$

under the following *linearized constraint* $\delta \mathbf{y}_h^{n+1} \in \mathcal{F}_h(\mathbf{y}_h^n)$ for the increment, where

$$\mathcal{F}_h(\mathbf{y}_h^n) := \left\{ \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) : L_T[\mathbf{y}_h^n; \mathbf{v}_h] = \int_T \nabla \mathbf{v}_h^T \nabla \mathbf{y}_h^n + (\nabla \mathbf{y}_h^n)^T \nabla \mathbf{v}_h = 0, \forall T \in \mathcal{T}_h \right\}. \tag{2.4.4}$$

This linearized constraint $L_T[\mathbf{y}_h^n; \delta \mathbf{y}_h^{n+1}] = 0$ for all $T \in \mathcal{T}_h$ is enforced using Lagrange multipliers. Here, we consider the space of symmetric piecewise constant matrices defined by

$$\Lambda_h := \left\{ \lambda_h : \Omega \to \mathbb{R}^{2\times 2} : \ \lambda_h^T = \lambda_h, \ \lambda_h \in \left[\mathbb{V}_h^0\right]^{2\times 2} \right\}.$$

To minimize (2.4.3) with the linearized constraint (2.4.4), we thus seek solutions $(\delta \mathbf{y}_h^{n+1}, \lambda_h^{n+1}) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) \times \Lambda_h$ such that

$$\begin{aligned} \tau^{-1}(\delta \mathbf{y}_h^{n+1}, \mathbf{v}_h)_{H_h^2(\Omega)} + \delta E_h[\mathbf{y}_h^n + \delta \mathbf{y}_h^{n+1}](\mathbf{v}_h) + b_h(\mathbf{v}_h, \lambda_h^{n+1}; \mathbf{y}_h^n) &= F(\mathbf{v}_h), \\ b_h(\delta \mathbf{y}_h^{n+1}, \mu_h; \mathbf{y}_h^n) &= 0, \end{aligned} \tag{2.4.5}$$

for any $\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ and $\mu_h \in \Lambda_h$. The bilinear form $b_h(\cdot, \cdot; \mathbf{y}_h^n)$ depends on $\mathbf{y}_h^n$ and is defined for any $(\mathbf{v}_h, \lambda_h) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) \times \Lambda_h$ by

$$b_h(\mathbf{v}_h, \lambda_h^{n+1}; \mathbf{y}_h^n) := \sum_{T \in \mathcal{T}_h} \int_T \lambda_h : (\nabla \mathbf{v}_h^T \nabla \mathbf{y}_h^n + (\nabla \mathbf{y}_h^n)^T \nabla \mathbf{v}_h). \tag{2.4.6}$$

Recall we define the bilinear form a_h as in (2.2.13) and linear form F as in (2.2.14). Also, we assume there is no forcing term in the *free boundary case*, namely

$\mathbf{f} = \mathbf{0}$. Then we have

$$E_h[\mathbf{y}_h] = \frac{1}{2}a_h(\mathbf{y}_h, \mathbf{y}_h) - F_h(\mathbf{y}_h) \quad \text{and} \quad \delta E_h[\mathbf{y}_h](\mathbf{v}_h) = a_h(\mathbf{y}_h, \mathbf{v}_h) - F_h(\mathbf{v}_h), \tag{2.4.7}$$

and (2.4.5) can be rewritten as: find $(\delta\mathbf{y}_h^{n+1}, \lambda_h^{n+1}) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) \times \Lambda_h$ such that

$$\begin{aligned} \tau^{-1}(\delta\mathbf{y}_h^{n+1}, \mathbf{v}_h)_{H_h^2(\Omega)} + a_h(\delta\mathbf{y}_h^{n+1}, \mathbf{v}_h) + b_h(\mathbf{v}_h, \lambda_h^{n+1}; \mathbf{y}_h^n) &= F_h(\mathbf{v}_h) - a_h(\mathbf{y}_h^n, \mathbf{v}_h) \\ b_h(\delta\mathbf{y}_h^{n+1}, \mu_h; \mathbf{y}_h^n) &= 0 \end{aligned} \tag{2.4.8}$$

for all $(\mathbf{v}_h, \mu_h) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) \times \Lambda_h$. Note that $\mathbf{y}_h^n \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$, and when $\Gamma_D \neq \emptyset$, the boundary data are implicitly contained in $a_h(\mathbf{y}_h^n, \mathbf{v}_h)$ through the liftings of the boundary data that appear in $H_h[\mathbf{y}_h]$. Moreover, when $\Gamma_D = \emptyset$, the term $\sigma(\cdot,\cdot)_{L^2(\Omega)}$ in the H_h^2 metric fixes the kernel of the linear problem (2.4.8) and guarantees its solvability.

The proposed strategy is summarized in Algorithm 1.

Algorithm 1: (discrete-H^2 gradient flow) Finding local minima of E_h

Given a target metric defect $\varepsilon > 0$, a pseudo-time step $\tau > 0$ and a target tolerance tol;
Choose initial guess $\mathbf{y}_h^0 \in \mathbb{A}_{h,\varepsilon}^k$;
while $\tau^{-1}|E_h[\mathbf{y}_h^{n+1}] - E_h[\mathbf{y}_h^n]| > tol$ **do**
  Solve (2.4.3)-(2.4.4) for $\delta\mathbf{y}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$;
  Update $\mathbf{y}_h^{n+1} = \mathbf{y}_h^n + \delta\mathbf{y}_h^{n+1}$;
end

We now show that the discrete gradient flow decreases the discrete energy E_h at each step and controls the prestrain defect D_h.

Theorem 2.4.1 (Energy Stability). *Assume that* $\delta\mathbf{y}_h^{n+1}$ *solves* (2.4.8). *If* $\delta\mathbf{y}_h^{n+1}$ *is non-zero, and we iterate* $\mathbf{y}_h^{n+1} = \mathbf{y}_h^n + \delta\mathbf{y}_h^{n+1}$ *for any* $0 \le n \le N-1$. *For any* $N \ge 1$, *we have*

$$E_h[\mathbf{y}_h^N] + \tau^{-1} \sum_{n=0}^{N-1} \|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2(\Omega)} \le E_h[\mathbf{y}_h^0]. \tag{2.4.9}$$

Proof. If we take $\mathbf{v}_h = \delta\mathbf{y}_h^{n+1} \neq \mathbf{0}$ and $\mu_h = \lambda_h^{n+1}$ in (2.4.8), then we obtain when $\Gamma_D \neq \emptyset$

$$\tau^{-1}\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2(\Omega)} + a_h(\mathbf{y}_h^{n+1}, \delta\mathbf{y}_h^{n+1}) = F_h(\delta\mathbf{y}_h^{n+1}).$$

When $\Gamma_D = \emptyset$, the right-hand side is 0 as we assume $\mathbf{f} = 0$ in this case. Since a_h is a symmetric bilinear form, we have

$$\begin{aligned} a_h(\mathbf{y}_h^{n+1}, \delta\mathbf{y}_h^{n+1}) &= \frac{1}{2}a_h(\mathbf{y}_h^{n+1}, \mathbf{y}_h^{n+1} - \mathbf{y}_h^n) + \frac{1}{2}a_h(\mathbf{y}_h^{n+1}, \mathbf{y}_h^{n+1} - \mathbf{y}_h^n) \\ &= \frac{1}{2}a_h(\mathbf{y}_h^{n+1}, \mathbf{y}_h^{n+1}) - \frac{1}{2}a_h(\mathbf{y}_h^{n+1}, \mathbf{y}_h^n) + \frac{1}{2}a_h(\mathbf{y}_h^n + \delta\mathbf{y}_h^{n+1}, \mathbf{y}_h^{n+1} - \mathbf{y}_h^n) \\ &= \frac{1}{2}a_h(\mathbf{y}_h^{n+1}, \mathbf{y}_h^{n+1}) - \frac{1}{2}a_h(\mathbf{y}_h^n, \mathbf{y}_h^n) + \frac{1}{2}a_h(\delta\mathbf{y}_h^{n+1}, \delta\mathbf{y}_h^{n+1}). \end{aligned} \tag{2.4.10}$$

Therefore, using (2.4.7) we get

$$\tau^{-1}\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2(\Omega)} + \frac{1}{2}a_h(\delta\mathbf{y}_h^{n+1}, \delta\mathbf{y}_h^{n+1}) + E_h[\mathbf{y}_h^{n+1}] = E_h[\mathbf{y}_h^n]. \tag{2.4.11}$$

Since $\delta\mathbf{y}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ and $\delta\mathbf{y}_h^{n+1} \neq \mathbf{0}$ we have $a_h(\delta\mathbf{y}_h^{n+1}, \delta\mathbf{y}_h^{n+1}) \ge 0$ and further $\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2(\Omega)} > 0$, and thus $E_h[\mathbf{y}_h^{n+1}] < E_h[\mathbf{y}_h^n]$. Moreover, (2.4.9) follows after we sum over $n = 0, 1, \ldots, N-1$. □

Theorem 2.4.2 (Control of Defect). *Let* $\mathbf{y}_h^0 \in \mathbb{A}_{h,\boldsymbol{\epsilon}_0}^k$. *Then, for any* $N \ge 1$, *the* N^{th}

iterate $\mathbf{y}_h^N$ *of the gradient flow satisfies*

$$D_h[\mathbf{y}_h^n] = \sum_{T\in\mathcal{T}_h} \left| \int_T ((\nabla \mathbf{y}_h^N)^T \nabla \mathbf{y}_h^N - g) \right| \leq \boldsymbol{\epsilon}_0 + c\tau(E_h[\mathbf{y}_h^0] + \tilde{c}) =: \boldsymbol{\epsilon}. \tag{2.4.12}$$

Here $c > 0$ *is the hidden constant of* (2.2.26) *if* $\Gamma_D \neq \emptyset$ *(of* (2.4.2) *if* $\Gamma_D = \emptyset$*), which depends only on* Ω *and* Γ_D *(resp. on* Ω *and* σ*), while* $\tilde{c} \geq 0$ *depends only on* μ*,* g*, the constant* $C(\gamma_0, \gamma_1)$ *that appears in* (2.3.1)*, as well as* $\|\boldsymbol{\varphi}\|_{H^1(\Omega)}$*,* $\|\Phi\|_{H^1(\Omega)}$*, and* $\|\mathbf{f}\|_{L^2(\Omega)}$ *when* $\Gamma_D \neq \emptyset$*.*

Moreover, if $E_h[\mathbf{y}_h^N] \geq 0$ *(which is necessarily the case when* $\mathbf{f} = \mathbf{0}$*, which is assumed when* $\Gamma_D = \emptyset$*), then* (2.4.12) *holds true with* $\tilde{c} = 0$*.*

In particular, if $\mathbf{y}_h := \mathcal{I}_h^k \mathbf{y}^0$ *is the Lagrange interpolant of some* $\mathbf{y}^0 \in \mathbb{A}$*, then*

$$D_h[\mathbf{y}_h^n] \lesssim (h+\tau)(\|\mathbf{y}^0\|_{H^2(\Omega)}^2 + \tilde{c}). \tag{2.4.13}$$

Proof. Note that

$$\begin{aligned} \int_T (\nabla \mathbf{y}_h^{n+1})^T \nabla \mathbf{y}_h^{n+1} &= \int_T (\nabla \mathbf{y}_h^n)^T \nabla \mathbf{y}_h^n + \int_T (\nabla \delta \mathbf{y}_h^{n+1})^T \nabla \mathbf{y}_h^n \\ &\quad + \int_T (\nabla \mathbf{y}_h^n)^T \nabla \delta \mathbf{y}_h^{n+1} + \int_T (\nabla \delta \mathbf{y}_h^{n+1})^T \nabla \delta \mathbf{y}_h^{n+1} \\ &= \int_T (\nabla \mathbf{y}_h^n)^T \nabla \mathbf{y}_h^n + \int_T (\nabla \delta \mathbf{y}_h^n)^T \nabla \delta \mathbf{y}_h^{n+1}, \end{aligned} \tag{2.4.14}$$

where the second equality follows from the linearized constraint (2.4.4). Therefore,

$$\sum_{T\in\mathcal{T}_h}\left|\int_T((\nabla\mathbf{y}_h^N)^T\nabla\mathbf{y}_h^N-g)\right|=\sum_{T\in\mathcal{T}_h}\left|\int_T(\nabla\mathbf{y}_h^0)^T\nabla\mathbf{y}_h^0-g+\sum_{n=0}^{N-1}((\nabla\delta\mathbf{y}_h^{n+1})^T\nabla\delta\mathbf{y}_h^{n+1})\right| \tag{2.4.15}$$

$$\leq\sum_{T\in\mathcal{T}_h}\left|\int_T((\nabla\mathbf{y}_h^0)^T\nabla\mathbf{y}_h^0-g)\right|+\sum_{T\in\mathcal{T}_h}\sum_{n=0}^{N-1}\|\nabla\delta\mathbf{y}_h^{n+1}\|^2_{L^2(T)}$$

$$\leq\boldsymbol{\epsilon}_0+c\sum_{n=0}^{N-1}\|\delta\mathbf{y}_h^{n+1}\|^2_{H^2_{\tilde{h}}(\Omega)},$$

where for the last inequality we used the fact that $\mathbf{y}_h^0\in\mathbb{A}^k_{h,\boldsymbol{\epsilon}_0}$, as well as the discrete Friedrich-Poincaré type inequality (2.2.26) if $\Gamma_D\neq\emptyset$ and the inequality (2.4.2) if $\Gamma_D=\emptyset$. Using (2.4.9) and noting that $E_h[\mathbf{y}_h^N]\geq 0$ (recall that $\mathbf{f}=0$ when $\Gamma_D=\emptyset$, which implies that $E_h[\mathbf{y}_h^N]\geq 0$), then we have

$$\sum_{n=0}^{N-1}\|\delta\mathbf{y}_h^{n+1}\|^2_{H^2_h(\Omega)}\leq\tau E_h[\mathbf{y}_h^0],$$

and (2.4.12) with $\tilde{c}=0$ follows by inserting the last inequality in (2.4.15). For the general case, which can only occur when $\Gamma_D\neq\emptyset$, from (2.3.14) we have

$$C_3|||\mathbf{y}_h^N|||^2_{H^2_h(\Omega)}-\tilde{c}\leq E_h[\mathbf{y}_h^N],$$

where $C_3>0$ and $\tilde{c}\geq 0$ depends only on $\|\mathbf{f}\|_{L^2(\Omega)}$, $\|\boldsymbol{\varphi}\|_{H^1(\Omega)}$, $\|\Phi\|_{H^1(\Omega)}$, μ, g and the constant $C(\gamma_0,\gamma_1)$ that appears in (2.3.1). Inserting the last inequality in (2.4.9),

and using that $|||\mathbf{y}_h^N|||_{H_h^2(\Omega)} \geq 0$, we have

$$\sum_{n=0}^{N-1} \|\delta \mathbf{y}_h^{n+1}\|_{H_h^2(\Omega)}^2 \leq \tau E_h[\mathbf{y}_h^0] + \tau \tilde{c},$$

and (2.4.12) follows by inserting the last inequality in (2.4.15).

In turn, (2.4.13) follows from (2.3.18) and upon noting that the definition of the discrete Hessians (2.2.10), Lemma 2.2.1 (L^2 bound of lifting operators), a trace inequality, and the local stability of the Lagrange interpolant imply

$$E_h[\mathcal{I}_h^k \mathbf{y}^0] \lesssim \|\mathcal{I}_h^k \mathbf{y}^0\|_{H_h^2(\Omega)}^2 \lesssim \|\mathbf{y}^0\|_{H^2(\Omega)}^2. \tag{2.4.16}$$

This concludes the proof. □

Note that for a general target metric g, it may be difficult (if not impossible) to construct an explicit $\mathbf{y}^0 \in \mathbb{A}$. In this situation, a suitable initial guess $\mathbf{y}_h^0 \in \mathbb{A}_{h,\boldsymbol{\epsilon}_0}^k$ with $\boldsymbol{\epsilon}_0$ small can be generated via a *preprocessing* procedure, which is presented in Section 2.5.1.

We emphasize again that in the case $\Gamma_D = \emptyset$, we modify the definition of $(\cdot,\cdot)_{H_h^2(\Omega)}$ by adding an L^2 term to fix the non-trivial kernel of the linear problem (2.4.8). Indeed, the bilinear form $a_h(.,.)$ has a nontrivial kernel in $\mathcal{F}_h(\mathbf{y}_h^n)$ containing the constant vectors. This is reflected in Theorem 2.3.2 where the sequence $\bar{\mathbf{y}}_h := \mathbf{y}_h - (1/|\Omega|)\int_\Omega \mathbf{y}_h$, and not $\mathbf{y}_h$, is precompact. For this but also to characterize more precisely the limit deformation $\mathbf{y}_h^\infty$ obtained by the gradient flow (see Proposition 2.4.1), we note that our scheme actually controls the evolution of the

deformation averages throughout the gradient flow for the *free boundary case*. This is the object of the next result.

Theorem 2.4.3. *Assume that $\Gamma_D = \emptyset$ (recall that $\mathbf{f} = \mathbf{0}$ in this case). Given any initial deformation $\mathbf{y}_h^0$, the n^{th} iterate $\mathbf{y}_h^n$ of the gradient flow satisfies $\int_\Omega \mathbf{y}_h^n = \int_\Omega \mathbf{y}_h^0$.*

Proof. Let us take $\mathbf{v}_h = \mathbf{C}$ in the first equation of (2.4.8), where $\mathbf{C}$ is a constant vector. Note that $D^2\mathbf{v}_h = 0$, $\nabla \mathbf{v}_h = 0$ and $[\mathbf{v}_h] = 0$ along all interior edges $e \in \mathcal{E}_h^i$. Therefore we have $a_h(\mathbf{y}_h^n, \mathbf{v}_h) = a_h(\delta\mathbf{y}_h^{n+1}, \mathbf{v}_h) = b_h(\lambda_h^{n+1}, \mathbf{v}_h; \mathbf{y}_h^n) = 0$. Moreover, we have

$$(\delta\mathbf{y}_h^{n+1}, \mathbf{C})_{H_h^2(\Omega)} = \sigma(\delta\mathbf{y}_h^{n+1}, \mathbf{C})_{L^2(\Omega)} = \sum_{i=1}^{3} C_i \sigma \int_\Omega (\delta\mathbf{y}_h^{n+1})_i,$$

where the subscript i indicates the extraction of the i^{th} component of a vector.

Then the first equation of (2.4.8) reduces to

$$\sum_{i=1}^{3} C_i \sigma \int_\Omega (\delta\mathbf{y}_h^{n+1})_i = 0,$$

which implies that $\int_\Omega (\delta\mathbf{y}_h^{n+1})_i = 0$ for any n and each component. Hence,

$$\int_\Omega (\mathbf{y}_h^n)_i = \int_\Omega (\mathbf{y}_h^0)_i + \sum_{j=1}^{n} (\delta\mathbf{y}_h^j)_i = \int_\Omega (\mathbf{y}_h^0)_i.$$

This finishes the proof. □

Remark 2.4.1. *If there is a non-zero external force $\mathbf{f}$, then Theorem 2.4.3 does not*

hold. However, at each step of the gradient flow, we have

$$\int_\Omega \delta \mathbf{y}_h^{n+1} = \frac{\tau}{\sigma} \int_\Omega \mathbf{f}.$$

This can be shown easily by using the same calculation as in the proof of Theorem 2.4.3 but adding a term $\int_\Omega \mathbf{f} \cdot \mathbf{v}_h$ *on the right-hand side of the first equation of* (2.4.8). *This means that in the* free boundary case, *if* $\int_\Omega \mathbf{f} \neq 0$ *then it is not possible for the gradient flow to converge. This is physically meaningful: the plate will move endlessly since it is not constrained on the boundary. In the free boundary case, without losing generality, we have assumed the stronger assumption* $\mathbf{f} = \mathbf{0}$.

Remark 2.4.2. *In particular, Theorem 2.4.3 implies that* $\int_\Omega \mathbf{y}_h^n = \mathbf{0}$ *for all the iterates if* $\int_\Omega \mathbf{y}_h^0 = \mathbf{0}$. *The later can easily be achieved by subtracting* $(1/|\Omega|) \int_\Omega \mathbf{y}_h^0$ *to any initial guess* $\mathbf{y}_h^0$ *without affecting* $E_h[\mathbf{y}_h^0]$ *or* $D_h[\mathbf{y}_h^0]$. *The sequence* $\{\mathbf{y}_h^N\}_{h>0}$ *of outputs of the gradient flow is then precompact and satisfies the assumption in Theorem 2.3.3 without further shifting.*

In the *free boundary case*, if $\mathbf{y}_h$ solves (2.2.16), then $R\mathbf{y}_h + \mathbf{c}$ is also a solution for any 3×3 *rotational* matrix R and constant *translation* $\mathbf{c}$. How the *translation* $\mathbf{c}$ is fixed in the gradient flow is already shown in Theorem 2.4.3, and it is related to the average of the initialization. We now show how the rotation R is fixed once an initialization $\mathbf{y}_h^0$ has been chosen.

Theorem 2.4.4. *Assume that* $\Gamma_D = \emptyset$ *(recall that* $\mathbf{f} = \mathbf{0}$ *in this case). Given an initialization* $\mathbf{y}_h^0$ *for the gradient flow, we denote by* $\mathbf{y}_h^n$ *the corresponding* n^{th} *iterate.*

Then, if we take $R\mathbf{y}_h^0$ as a new initialization, the corresponding n^{th} iterate is $R\mathbf{y}_h^n$.

Proof. Let $(\delta\mathbf{y}_h^1, \lambda_h^1)$ be the solution of (2.4.8) for $n = 0$. Then for any test function $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$ we observe that $R^T\mathbf{v}_h \in [\mathbb{V}_h^k]^3$ is also an admissible test function and we thus obtain

$$(\delta\mathbf{y}_h^1, R^T\mathbf{v}_h)_{H_{\tilde{h}}^2(\Omega)} + \tau a_h(\delta\mathbf{y}_h^1, R^T\mathbf{v}_h) + \tau b_h(\lambda_h^1, R^T\mathbf{v}_h; \mathbf{y}_h^0) = -\tau a_h(\mathbf{y}_h^0, R^T\mathbf{v}_h)$$

$$b_h(\mu_h, \delta\mathbf{y}_h^1; \mathbf{y}_h^0) = 0.$$

Now, consider again (2.4.8) for $n = 0$ but with $\mathbf{y}_h^0$ replaced by $\hat{\mathbf{y}}_h^0 := R\mathbf{y}_h^0$ in the right-hand side of the first equation of (2.4.8). Then $(\delta\hat{\mathbf{y}}_h^1, \hat{\lambda}_h^1)$ is such that

$$(\delta\hat{\mathbf{y}}_h^1, \mathbf{v}_h)_{H_{\tilde{h}}^2(\Omega)} + \tau a_h(\delta\hat{\mathbf{y}}_h^1, \mathbf{v}_h) + \tau b_h(\hat{\lambda}_h^1, \mathbf{v}_h; R\mathbf{y}_h^0) = -\tau a_h(R\mathbf{y}_h^0, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$$

$$b_h(\mu_h, \delta\hat{\mathbf{y}}_h^1; R\mathbf{y}_h^0) = 0 \quad \forall \mu_h \in \Lambda_h.$$

Since $(\delta\mathbf{y}_h^1, R^T\mathbf{v}_h)_{H_{\tilde{h}}^2(\Omega)} = (R\delta\mathbf{y}_h^1, \mathbf{v}_h)_{H_{\tilde{h}}^2(\Omega)}$,

$$a_h(\mathbf{y}_h^0, R^T\mathbf{v}_h) = a_h(R\mathbf{y}_h^0, \mathbf{v}_h), \quad a_h(\delta\mathbf{y}_h^1, R^T\mathbf{v}_h) = a_h(R\delta\mathbf{y}_h^1, \mathbf{v}_h),$$

and

$$b_h(\lambda_h^1, R^T\mathbf{v}_h; \mathbf{y}_h^0) = b_h(\lambda_h^1, \mathbf{v}_h; R\mathbf{y}_h^0), \quad b_h(\mu_h, \delta\hat{\mathbf{y}}_h^1; R\mathbf{y}_h^0) = b_h(\mu_h, R^T\delta\hat{\mathbf{y}}_h^1; \mathbf{y}_h^0),$$

we have $\delta\hat{\mathbf{y}}_h^1 = R\delta\mathbf{y}_h^1$ and $\hat{\lambda}_h^1 = \lambda_h^1$. This implies that $\hat{\mathbf{y}}_h^1 := \hat{\mathbf{y}}_h^0 + \delta\hat{\mathbf{y}}_h^1 = R(\mathbf{y}_h^0 + \delta\mathbf{y}_h^1) = R\mathbf{y}_h^1$.

Inductively, we can conclude that $\hat{\mathbf{y}}_h^n = R\mathbf{y}_h^n$ for any n if $\hat{\mathbf{y}}_h^n$ and $\mathbf{y}_h^n$ correspond to n^{th} iterate with initialization $\hat{\mathbf{y}}_h^0 = R\mathbf{y}_h^0$ and $\mathbf{y}_h^0$ respectively. □

Energy decreasing gradient flow algorithms are generally not guaranteed to converge to absolute minimizers. We address this aspect in the next proposition and show that the gradient flow reaches a deformation $\mathbf{y}_h^\infty$, which is a local minimum for E_h in the direction $\mathcal{F}_h(\mathbf{y}_h^\infty)$.

Proposition 2.4.1 (Limit of gradient flow)**.** *For a fixed h, let $\mathbf{y}_h^0 \in \mathbb{A}_{h,\varepsilon_0}^k$ be such that $E_h[\mathbf{y}_h^0] < \infty$ and let $\{\mathbf{y}_h^n\}_{n\geq 1} \subset \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ be the sequence produced by the discrete gradient flow. Suppose that for all $n \geq 0$, there exists a constant $\beta_h > 0$ independent of n such that*

$$\inf_{\mu_h \in \Lambda_h} \sup_{\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})} \frac{b_h(\mathbf{v}_h, \mu_h; \mathbf{y}_h^n)}{\|\mathbf{v}_h\|_{H_h^2(\Omega)} \|\mu_h\|_{L^2(\Omega)}} \geq \beta_h. \tag{2.4.17}$$

Then there exists a subsequence (not relabeled) and $\mathbf{y}_h^\infty \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ such that $\mathbf{y}_h^n \to \mathbf{y}_h^\infty$ as $n \to \infty$ and $\mathbf{y}_h^\infty$ is a local minimum for E_h in the direction $\mathcal{F}_h(\mathbf{y}_h^\infty)$, namely

$$E_h[\mathbf{y}_h^\infty] \leq E_h[\mathbf{y}_h^\infty + \mathbf{v}_h] \quad \forall \mathbf{v}_h \in \mathcal{F}_h(\mathbf{y}_h^\infty). \tag{2.4.18}$$

Proof. Thanks to (2.4.9) we have that $\sup_{n\geq 1} E_h[\mathbf{y}_h^n] < \infty$. Arguing as in the proof of Proposition 2.3.1, we can deduce that a subsequence (not relabeled) converges to some $\mathbf{y}_h^\infty \in \mathbb{A}_{h,\boldsymbol{\epsilon}}^k$ in any norm defined on $[\mathbb{V}_h^k]^3$ (in the *free boundary case*, we can replace $\mathbf{y}_h^0$ by $\mathbf{y}_h^0 - \frac{1}{|\Omega|}\int_\Omega \mathbf{y}_h^0$, which does not affect $E_h[\mathbf{y}_h^0]$ and $D_h[\mathbf{y}_h^0]$, and use Proposition 2.4.3 to deduce that all the iterates have mean-value zero). This proves the first part of the Proposition. We now show (2.4.18). Since E_h is quadratic and

convex, for any $\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})$ we have

$$E_h[\mathbf{y}_h^\infty + \mathbf{v}_h] = E_h[\mathbf{y}_h^\infty] + \delta E_h[\mathbf{y}_h^\infty](\mathbf{v}_h) + \delta^2 E_h[\mathbf{y}_h^\infty](\mathbf{v}_h,\mathbf{v}_h) \geq E_h[\mathbf{y}_h^\infty] + \delta E_h[\mathbf{y}_h^\infty](\mathbf{v}_h)$$

and thus it is enough to show that

$$\delta E_h[\mathbf{y}_h^\infty](\mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathcal{F}_h(\mathbf{y}_h^\infty). \tag{2.4.19}$$

We start by showing that (up to a subsequence) $\{\lambda_h^{n+1}\}_{n\geq 0}$ converges. Recall that $\|\cdot\|_{H_h^2(\Omega)}$ is a norm on $\mathbb{V}_h^k(\mathbf{0},\mathbf{0})$ for both cases $\Gamma_D \neq \emptyset$ and $\Gamma_D = \emptyset$. Therefore, from (2.4.9) we also have that $\lim_{n\to\infty} \|\delta\mathbf{y}_h^{n+1}\|_{H_h^2(\Omega)} = 0$ which in turn implies that $\lim_{n\to\infty} \delta\mathbf{y}_h^{n+1} = \mathbf{0}$. Taking the limit $n \to \infty$ in the first equation of (2.4.8), we get

$$\lim_{n\to\infty} b_h(\mathbf{v}_h, \lambda_h^{n+1}; \mathbf{y}_h^n) = F_h(\mathbf{v}_h) - a_h(\mathbf{y}_h^\infty, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0}). \tag{2.4.20}$$

Now using the inf-sup condition (2.4.17), we have for any $n \geq 0$

$$\|\lambda_h^{n+1}\|_{L^2(\Omega)} \leq \frac{1}{\beta_h} \sup_{\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})} \frac{b_h(\mathbf{v}_h, \lambda_h^{n+1}; \mathbf{y}_h^n)}{\|\mathbf{v}_h\|_{H_h^2(\Omega)}}.$$

We deduce that $\sup_{n\geq 0} \|\lambda_h^{n+1}\|_{L^2(\Omega)} < \infty$ since the upper bound in the above inequality converges thanks to (2.4.20), and thus a subsequence (not relabeled) converges to some $\lambda_h^\infty \in \Lambda_h$ in any norm defined on the finite dimensional space Λ_h. Recalling

(2.4.6) and (2.4.7), from (2.4.20) we thus have

$$\sum_{T\in\mathcal{T}_h}\int_T \lambda_h^\infty : (\nabla \mathbf{v}_h^T \nabla \mathbf{y}_h^\infty + (\nabla \mathbf{y}_h^\infty)^T \nabla \mathbf{v}_h) = -\delta E_h[\mathbf{y}_h^\infty](\mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0}).$$

Since λ_h^∞ is piecewise constant, by (2.4.4) we have that the left-hand side of the last relation vanishes for any $\mathbf{v}_h \in \mathcal{F}_h(\mathbf{y}_h^\infty)$ and thus (2.4.19) is proved. □

A schematic illustration of the limit of gradient flow is given in Fig. 2.1.

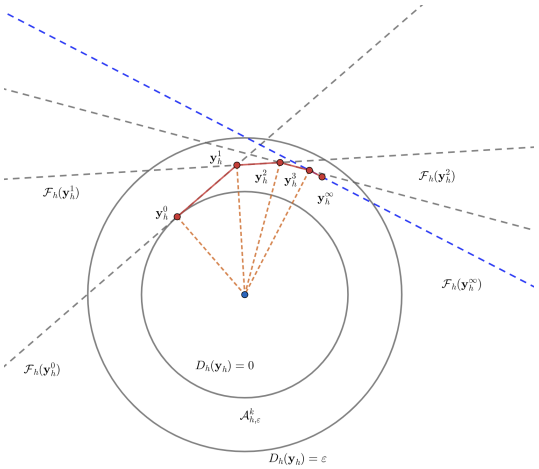

Figure 2.1: A schematic illustration of the gradient flow. For illustration purpose, without losing generality, we let $\mathbf{y}_h^0$ satisfy the metric constraint. Here, the two circles denote $D_h = 0$ and $D_h = \boldsymbol{\epsilon}$ respectively. The annulus between the two circles represent the discrete admissible set $\mathbb{A}_{h,\boldsymbol{\epsilon}}^k$. In each step of the gradient flow, the increment $\delta\mathbf{y}_h^{n+1}$ (red line segement) is searched along the tangent plane $\mathcal{F}_h(\mathbf{y}_h^n)$ (grey dotted line in the figure) such that $E_h[\mathbf{y}_h^{n+1}] < E_h[\mathbf{y}_h^n]$. In the limit as $n \to \infty$, $\delta\mathbf{y}_h^{n+1} \to \mathbf{0}$, and consequently $\mathcal{F}_h(\mathbf{y}_h^n)$ also changes little and moves asymptotically to $\mathcal{F}_h(\mathbf{y}_h^\infty)$. Thus the gradient flow searches new iterates on tangential directions to intermediate circles to decrease the energy asymptotically.

Remark 2.4.3. *Although proving the inf-sup condition (2.4.17) for the proposed method is open, we observe from numerical experiments that it should be satisfied with at least a constant β_h depending on h.*

We will prove the inf-sup condition with β_h depending on h later in chapter 3, which deals with the LDG method for the bilayer problem. The key difference in chapter 3, which plays a significant role, is that we impose the discrete metric constraint pointwise at barycenters of elements instead of the ℓ^1 notion in (2.2.15).

2.5 Initialization

The gradient flow (2.4.8) starts with an initial deformation $\mathbf{y}_h^0$ in $\mathbb{A}^k_{h,\boldsymbol{\epsilon}_0}$ with $\boldsymbol{\epsilon}_0 = D_h[\mathbf{y}_h^0]$ and $E_h[\mathbf{y}_h^0]$ affecting the prestrain defect of the successive iterates controlled by $\boldsymbol{\epsilon}_0 + C\tau(E_h[\mathbf{y}_h^0] + \tilde{c})$ in view of (2.4.12). We also point out that the monotone decay property (2.4.9) could require many iterations of the gradient flow which are considerably reduced when starting with a small energy $E_h[\mathbf{y}_h^0]$. Therefore, the role of the preprocessing algorithm is to construct an initial deformation with $\boldsymbol{\epsilon}_0$ relatively small and $E_h[\mathbf{y}_h^0]$ uniformly bounded. The proposed strategy produces a deformation with small prestrain defect satisfying approximate boundary conditions (if any).

The numerical strategy is divided in two steps: a *boundary conditions preprocessing* step followed by a *metric preprocessing* step.

2.5.1 Preprocessing: scheme

Boundary conditions preprocessing. When $\Gamma_D \neq \emptyset$, we consider the bi-Laplacian problem

$$\begin{cases} \Delta^2 \widehat{\mathbf{y}} &= \widehat{\mathbf{f}} \quad \text{in } \Omega \\ \nabla \widehat{\mathbf{y}} &= \Phi \quad \text{on } \Gamma_D \\ \widehat{\mathbf{y}} &= \boldsymbol{\varphi} \quad \text{on } \Gamma_D. \end{cases} \tag{2.5.1}$$

where typically $\widehat{\mathbf{f}} = \mathbf{0}$. This vector-valued problem is well-posed and gives, in general, a non-flat surface $\widehat{\mathbf{y}}(\Omega)$. We use the LDG method with boundary conditions imposed *à la Nitsche* to approximate the solution $\widehat{\mathbf{y}} \in \mathbb{V}(\boldsymbol{\varphi}, \Phi)$ of (2.5.1):

$$\widehat{\mathbf{y}}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi): \quad c_h(\widehat{\mathbf{y}}_h, \mathbf{v}_h) = (\widehat{\mathbf{f}}, \mathbf{v}_h)_{L^2(\Omega)} \quad \forall\, \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}). \tag{2.5.2}$$

Here, $c_h(\widehat{\mathbf{y}}_h, \mathbf{v}_h)$ is defined similarly to (2.2.13) using the discrete Hessian (2.2.10), i.e.,

$$\begin{aligned} c_h(\mathbf{w}_h, \mathbf{v}_h) := \int_\Omega H_h[\mathbf{w}_h] : H_h[\mathbf{v}_h] \\ + \widehat{\gamma}_1(\mathrm{h}^{-1}[\nabla_h \mathbf{w}_h], [\nabla_h \mathbf{v}_h])_{L^2(\Gamma_h^a)} + \widehat{\gamma}_0(\mathrm{h}^{-3}[\mathbf{w}_h], [\mathbf{v}_h])_{L^2(\Gamma_h^a)}, \end{aligned} \tag{2.5.3}$$

where $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are positive penalty parameters that may not necessarily be the same as their counterparts γ_0 and γ_1 used in the definition of E_h. Then $\widehat{\mathbf{y}}_h$ satisfies (approximately) the given boundary conditions on Γ_D and $\widehat{\mathbf{y}}_h(\Omega)$ is, in general, non-flat.

In the case $\Gamma_D = \emptyset$ (free boundary conditions), then an obvious choice is

$\widehat{\mathbf{y}} = (id, 0)^T$, where $id(x) = x$ for $x \in \Omega$, which gives a flat surface $\widehat{\mathbf{y}}(\Omega) = \Omega \times 0$. However, the *metric preprocessing* algorithm described below will not generate a deformation out of plane if the initial configuration is flat, see Corollary 2.5.1. As a consequence, only metrics g admitting flat immersion could be achieved. To get a surface out of plane, we consider a somewhat ad-hoc procedure: we solve (2.5.1) with a fictitious forcing $\widehat{\mathbf{f}} \neq \mathbf{0}$ supplemented with the Dirichlet boundary condition $\boldsymbol{\varphi}(x) = (x, 0)^T$ for $x \in \partial\Omega$ but obviating Φ and jumps of $\nabla_h \widehat{\mathbf{y}}_h$ on Γ_h^b in (2.5.3). This corresponds to enforcing discretely a variational (Neumann) boundary condition $\Delta\widehat{\mathbf{y}} = 0$ on $\partial\Omega$.

Notice that when a deformation $\widehat{\mathbf{y}}$ satisfying $\nabla\widehat{\mathbf{y}}^T\nabla\widehat{\mathbf{y}} = g$ is known, one can simply use the interpolation of $\widehat{\mathbf{y}}$ into $[\mathbb{V}_h^k]^3$ for $\widehat{\mathbf{y}}_h$. However, this situation is not likely to occur in general.

Metric preprocessing. In this step, an H^2 discrete gradient flow is designed to minimize the energy

$$\tilde{E}_h[\tilde{\mathbf{y}}_h] := E_h^s[\tilde{\mathbf{y}}_h] + \boldsymbol{\epsilon}_b E_h^b[\tilde{\mathbf{y}}_h], \tag{2.5.4}$$

where

$$E_h^s[\tilde{\mathbf{y}}_h] := \frac{1}{2}\int_\Omega |\nabla_h \tilde{\mathbf{y}}_h^T \nabla_h \tilde{\mathbf{y}}_h - g|^2, \tag{2.5.5}$$

$$E_h^b[\tilde{\mathbf{y}}_h] := \frac{1}{2}\left(\int_\Omega |g^{-\frac{1}{2}} H_h[\tilde{\mathbf{y}}_h] g^{-\frac{1}{2}}|^2 + \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h \tilde{\mathbf{y}}_h]\|^2_{L^2(\Gamma_h^0)} + \|\mathrm{h}^{-\frac{3}{2}}[\tilde{\mathbf{y}}_h]\|^2_{L^2(\Gamma_h^0)}\right), \tag{2.5.6}$$

and $0 < \boldsymbol{\epsilon}_b \ll 1$ is a small parameter. At first glance, to produce a deformation with a small prestrain defect D_h, we only need the term E_h^s. Indeed, for any $\tilde{\mathbf{y}}_h \in [\mathbb{V}_h^k]^3$

we have

$$D_h[\tilde{\mathbf{y}}_h] \leq \|(\nabla_h \tilde{\mathbf{y}}_h)^T \nabla_h \tilde{\mathbf{y}}_h - g\|_{L^1(\Omega)} \lesssim \|(\nabla_h \tilde{\mathbf{y}}_h)^T \nabla_h \tilde{\mathbf{y}}_h - g\|_{L^2(\Omega)} \approx E_h^s[\tilde{\mathbf{y}}_h]^{\frac{1}{2}}, \quad (2.5.7)$$

and a small prestrain defect D_h can thus be ensured by an energy decaying gradient flow for E_h^s. However, in order to guarantee the uniform boundedness of E_h, see Remark 2.5.1 below, we need the second term E_h^b that involves the discrete Hessian as a regularization.

Additionally, as discussed in Section 2.1.2, the elastic energy rescaled with s^{-1} (s is the thickness parameter in the 3d model of plates) can be expressed as stretching energy plus bending energy multiplied by s^2. We may view the metric preprocessing energy (2.5.4) as a discrete analogue of this pre-asymptotic decomposition of elastic energy for a small numerical thickness parameter $\sqrt{\boldsymbol{\epsilon}_b}$. We note that the first term of (2.5.4) can be considered as a discrete stretching energy, and the second accounts for the bending. We emphasize that E_h^b is different from the discrete bending energy E_h we minimize in the main gradient flow, but they are equal if Lamé coefficients are $\lambda = 0$ and $\mu = 6$, the forcing term $\mathbf{f} = \mathbf{0}$, and stabilization parameters are $\gamma_0 = \gamma_1 = 1$, and they are equivalent in any event up to a multiplicative constants depending on $\lambda, \mu, \gamma_0, \gamma_1$.

Since the E_h^s is quartic in $\tilde{\mathbf{y}}_h$, we need an explicit treatment on its variational derivative in each step of the gradient flow; the gradient direction is linearized at the previous iterate. The procedure is similar to the main gradient flow of Section 2.4. Recursively, given $\tilde{\mathbf{y}}_h^n$ we compute $\tilde{\mathbf{y}}_h^{n+1} := \tilde{\mathbf{y}}_h^n + \delta\tilde{\mathbf{y}}_h^{n+1}$ by seeking the increment

$\delta\tilde{\mathbf{y}}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})$ satisfying

$$\tilde{\tau}^{-1}(\delta\tilde{\mathbf{y}}_h^{n+1},\mathbf{v}_h)_{H_h^2(\Omega)} + a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1},\mathbf{v}_h;\tilde{\mathbf{y}}_h^n) + \boldsymbol{\epsilon}_b a_h^b(\delta\tilde{\mathbf{y}}_h^{n+1},\mathbf{v}_h) = -a_h^s(\tilde{\mathbf{y}}_h^n,\mathbf{v}_h;\tilde{\mathbf{y}}_h^n) - \boldsymbol{\epsilon}_b a_h^b(\tilde{\mathbf{y}}_h^n,\mathbf{v}_h) \quad (2.5.8)$$

for all $\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})$, where $\tilde{\tau}$ is a pseudo time-step parameter that is not necessarily the same as τ in the main gradient flow and

$$a_h^s(\tilde{\mathbf{y}}_h,\mathbf{v}_h;\tilde{\mathbf{y}}_h^n) := \int_\Omega (\nabla_h \mathbf{v}_h^T \nabla_h \tilde{\mathbf{y}}_h + \nabla_h \tilde{\mathbf{y}}_h^T \nabla_h \mathbf{v}_h) : ((\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \tilde{\mathbf{y}}_h^n - g) \quad (2.5.9)$$

$$a_h^b(\tilde{\mathbf{y}}_h,\mathbf{v}_h) := \int_\Omega (g^{-\frac{1}{2}} H_h[\tilde{\mathbf{y}}_h] g^{-\frac{1}{2}}) : (g^{-\frac{1}{2}} H_h[\mathbf{v}_h] g^{-\frac{1}{2}}) + (\mathrm{h}^{-\frac{1}{2}}[\nabla_h \tilde{\mathbf{y}}_h],[\nabla_h \mathbf{v}_h])_{L^2(\Gamma_h^0)} + (\mathrm{h}^{-\frac{3}{2}}[\tilde{\mathbf{y}}_h],[\mathbf{v}_h])_{L^2(\Gamma_h^0)}. \quad (2.5.10)$$

For the stopping criteria, the flow is ended when either of the following two conditions is satisfied:

1. the prestrain defect D_h reaches a prescribed value $\tilde{\boldsymbol{\epsilon}}_0$, i.e, $D_h[\tilde{\mathbf{y}}_h^{n+1}] \leq \tilde{\boldsymbol{\epsilon}}_0$;

2. the energy $\tilde{E}_h$ becomes stationary, i.e., $\tilde{\tau}^{-1}|\tilde{E}_h[\tilde{\mathbf{y}}^{n+1}] - \tilde{E}_h[\tilde{\mathbf{y}}^n]| \leq \tilde{tol}$.

Summary. We summarize the previous discussion of preprocessing in Algorithm 2, which consists of two separate steps: the *boundary conditions* and *metric* preprocessing steps.

It is conceivable that more efficient or physically motivated algorithms could

Algorithm 2: Initialization step for Algorithm 1.

Given $\widetilde{tol}$ and $\widetilde{\varepsilon}_0$;
if $\Gamma_D \neq \emptyset$ *(Dirichlet boundary condition)* **then**
  Solve (2.5.2) for $\widehat{\mathbf{y}}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ with $\widehat{\mathbf{f}} = \mathbf{0}$;
else
  Solve (2.5.2) for $\widehat{\mathbf{y}}_h$ with $\widehat{\mathbf{f}} \neq \mathbf{0}$, $\boldsymbol{\varphi} = (id, 0)$ and without Φ;
end
Set $\widetilde{\mathbf{y}}_h^0 = \widehat{\mathbf{y}}_h$;
while $\widetilde{\tau}^{-1}\big|\tilde{E}_h[\widetilde{\mathbf{y}}_h^{n+1}] - \tilde{E}_h[\widetilde{\mathbf{y}}_h^{n}]\big| > \widetilde{tol}$ and $D_h[\widetilde{\mathbf{y}}_h^{n+1}] > \widetilde{\varepsilon}_0$ **do**
  Solve (2.5.8) for $\delta\widetilde{\mathbf{y}}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$;
  Update $\widetilde{\mathbf{y}}_h^{n+1} = \widetilde{\mathbf{y}}_h^{n} + \delta\widetilde{\mathbf{y}}_h^{n+1}$;
end
Set $\mathbf{y}_h^0 = \widetilde{\mathbf{y}}_h^{n+1}$.

be designed to construct initial guesses. We leave these considerations for future research. As we shall see later, different initial deformations can lead to different equilibrium configurations corresponding to distinct local minima of the energy E_h in (2.2.11). These minima are generally physically meaningful.

2.5.2 Preprocessing: analysis

Now, we prove that the gradient flow (2.5.8) is conditionally energy stable, namely the $\tilde{E}_h$ decays at each step provided $\tilde{\tau}$ is small enough and the increment is nonzero. Note that if we have an implicit scheme in each step of gradient flow, the unconditional energy decreasing is naturally guaranteed as in Theorem 2.4.1. However, this is not practical due to the nonlinearity brought by E_h^s. Since we treat the nonlinearity explicitly, we break the structure of gradient flow that is presented in Theorem 2.4.1, and we would rather anticipate an energy decay property with an additional restriction on time-step for this preprocessing scheme.

Before proving it, we first introduce three lemmas: Lemma 2.5.1 contains a

discrete Sobolev inequality, which has been proved in [62], and we reproduce it here in a simpler way using the smoothing interpolation; in Lemma 2.5.2 we show the continuity of a_h^s and "coercivity" of the left-hand side of (2.5.8) at each step for $\tilde{\tau}$ sufficiently small, which guarantees the solvability of the system (2.5.8) at each step; finally in Lemma 2.5.3 we prove that the $L^2(\Omega)$ norm of the broken gradient is controlled by the energy $\tilde{E}_h$ and the $L^1(\Omega)$ norm of the prestrain metric g.

Lemma 2.5.1 (Discrete Sobolev inequality)**.** *For any $v_h \in \mathbb{E}(\mathcal{T}_h)$ there holds*

$$\|v_h\|_{L^4(\Omega)}^2 \lesssim \|\nabla_h v_h\|_{L^2(\Omega)}^2 + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|_{L^2(\Gamma_h^0)}^2 + \|v_h\|_{L^2(\Omega)}^2 \tag{2.5.11}$$

Proof. Let $\Pi_h v_h \in H^1(\Omega) \cap \mathbb{V}_h^k$ be the smoothing interpolant introduced in Definition 2.2.1. Thanks to (2.2.19) the and the triangle inequality, we have

$$\begin{aligned}
\|v_h\|_{L^4(\Omega)}^2 &\lesssim \|v_h - \Pi_h v_h\|_{L^4(\Omega)}^2 + \|\Pi_h v_h\|_{L^4(\Omega)}^2 \\
&\lesssim \|\mathrm{h}^{-1}(v_h - \Pi_h v_h)\|_{L^2(\Omega)}^2 + \|\Pi_h v_h\|_{H^1(\Omega)}^2 \\
&\lesssim \|\nabla_h v_h\|_{L^2(\Omega)}^2 + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|_{L^2(\Gamma_h^0)}^2 + \|\Pi_h v_h\|_{L^2(\Omega)}^2 \\
&\lesssim \|\nabla_h v_h\|_{L^2(\Omega)}^2 + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|_{L^2(\Gamma_h^0)}^2 + \|v_h\|_{L^2(\Omega)}^2,
\end{aligned}$$

where we used the inverse inequality (recall that $\Omega \subset \mathbb{R}^2$)

$$\|w_h\|_{L^4(T)} \lesssim h_T^{-\frac{1}{2}} \|w_h\|_{L^2(T)} \quad \forall T \in \mathcal{T}_h \tag{2.5.12}$$

for the discrete function $w_h = v_h - \Pi_h v_h$ and the standard Sobolev inequality for

$\Pi_h v_h \in H^1(\Omega)$ for the second inequality, the estimate (2.2.19) for the third inequality, and (2.2.21) for the last one (note that (2.2.21) can be naturally extended to the case when $\mathbf{v} \in \mathbb{E}(\mathcal{T}_h)$, if we recall the definition of Π_h when $\mathbf{v} \in \mathbb{E}(\mathcal{T}_h)$ is modified by applying a further L^2-projection from $\mathbb{E}(\mathcal{T}_h)$ to $\mathbb{V}_h^k$). □

The next proposition concerns the form a_h^s and is key to guarantee that the hypothesis of the Lax-Milgram theorem are satisfied.

Lemma 2.5.2. *Let $\tilde{\mathbf{y}}_h^n \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$. We have*

$$|a_h^s(\mathbf{v}_h, \mathbf{w}_h; \tilde{\mathbf{y}}_h^n)| \leq C_1 E_h^s[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} \|\mathbf{v}_h\|_{H_h^2(\Omega)} \|\mathbf{w}_h\|_{H_h^2(\Omega)} \quad \forall \mathbf{v}_h, \mathbf{w}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}), \tag{2.5.13}$$

where C_1 is a positive constant independent of n and h.

Moreover, when $\tilde{\tau} \leq (1 + C_1 \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}})^{-1}$ with C_1 independent of n and h we have

$$\|\mathbf{v}_h\|_{H_h^2(\Omega)}^2 \leq \frac{1}{\tilde{\tau}} (\mathbf{v}_h, \mathbf{v}_h)_{H_h^2(\Omega)} + a_h^s(\mathbf{v}_h, \mathbf{v}_h; \tilde{\mathbf{y}}_h^n) \quad \forall \mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}). \tag{2.5.14}$$

As a consequence, there exists unique solution to the variational problem (2.5.8).

Proof. Let $\mathbf{v}_h, \mathbf{w}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$. It is clear that

$$|a_h^s(\mathbf{v}_h, \mathbf{w}_h; \tilde{\mathbf{y}}_h^n)| \leq 2 \|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{w}_h\|_{L^2(\Omega)} \|(\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \tilde{\mathbf{y}}_h^n - g\|_{L^2(\Omega)}.$$

Thanks to the Cauchy-Schwarz inequality we have

$$\|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{w}_h\|_{L^2(\Omega)} \leq \|\nabla_h \mathbf{v}_h\|_{L^4(\Omega)} \|\nabla_h \mathbf{w}_h\|_{L^4(\Omega)}.$$

Moreover, applying Lemma 2.5.1 to $\nabla_h \mathbf{z}_h \in \mathbb{E}(\mathcal{T}_h)$, we infer that

$$\|\nabla_h \mathbf{z}_h\|_{L^4(\Omega)} \lesssim \|\nabla_h \mathbf{z}_h\|_{L^2(\Omega)} + \|D_h^2 \mathbf{z}_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}} [\nabla_h \mathbf{z}_h]\|_{L^2(\Gamma_h^0)}, \quad \text{for } \mathbf{z}_h = \mathbf{v}_h, \mathbf{w}_h.$$

Also, by definition and by Lemma 2.2.3 (relation (2.4.2) if $\Gamma_D = \emptyset$), we have the inequalities

$$\|D_h^2 \mathbf{z}_h\|_{L^2(\Omega)} \leq \|\mathbf{z}_h\|_{H_h^2(\Omega)} \quad \text{and} \quad \|\nabla_h \mathbf{z}_h\|_{L^2(\Omega)} \lesssim \|\mathbf{z}_h\|_{H_h^2(\Omega)}, \quad \text{for } \mathbf{z}_h = \mathbf{v}_h, \mathbf{w}_h,$$

and thus

$$\|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{w}_h\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{H_h^2(\Omega)} \|\mathbf{w}_h\|_{H_h^2(\Omega)}.$$

Therefore, by the definition (2.5.5), we get

$$|a_h^s(\mathbf{v}_h, \mathbf{w}_h; \tilde{\mathbf{y}}_h^n)| \quad \leq \quad C_1 E_h^s[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} \|\mathbf{v}_h\|_{H_h^2(\Omega)} \|\mathbf{w}_h\|_{H_h^2(\Omega)}$$

for some positive constant C_1 independent of n and h. This shows (2.5.13). Now, taking $\mathbf{w}_h = \mathbf{v}_h$ in the last inequality we easily get

$$\tilde{\tau}^{-1} \|\mathbf{v}_h\|_{H_h^2(\Omega)}^2 + a_h^s(\mathbf{v}_h, \mathbf{v}_h; \tilde{\mathbf{y}}_h^n) \geq \tilde{\tau}^{-1} \|\mathbf{v}_h\|_{H_h^2(\Omega)}^2 - C_1 E_h^s[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} \|\mathbf{v}_h\|_{H_h^2(\Omega)}^2.$$

Therefore, since $E_h^s[\tilde{\mathbf{y}}_h^n] < \tilde{E}_h[\tilde{\mathbf{y}}_h^n]$, if $\tilde{\tau} \leq (1+C_1\tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}})^{-1}$ we have the claimed result (2.5.14).

Thanks to the two estimates (2.5.13), (2.5.14) and the coercivity of the discrete Hessian (Lemma 2.3.1), the Lax-Milgram theory applies to guarantee the existence and uniqueness of a solution to (2.5.8). □

Note that if we further have $\tilde{E}_h[\tilde{\mathbf{y}}_h^n] \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^0]$ for any $n \geq 0$, then we could have a uniform upper bound for $\tilde{\tau}$ in Lemma 2.5.2.

Lemma 2.5.3. *For any $\mathbf{v}_h \in [\mathbb{V}_h^k]^3$ we have*

$$\|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)}^2 \lesssim E_h^s[\mathbf{v}_h]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)}.$$

Proof. By (2.3.15), we get

$$\begin{aligned}
\|\nabla_h \mathbf{v}_h\|_{L^2(\Omega)}^2 \leq & \sqrt{2}\|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{v}_h\|_{L^1(\Omega)} \leq \sqrt{2}\left(\|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{v}_h - g\|_{L^1(\Omega)} + \|g\|_{L^1(\Omega)}\right) \\
\leq & \sqrt{2}\left[|\Omega|^{\frac{1}{2}}\|(\nabla_h \mathbf{v}_h)^T \nabla_h \mathbf{v}_h - g\|_{L^2(\Omega)} + \|g\|_{L^1(\Omega)}\right] \\
= & \sqrt{2}\left[\sqrt{2}|\Omega|^{\frac{1}{2}} E_h^s[\mathbf{v}_h]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)}\right]
\end{aligned}$$

which concludes the proof. □

Theorem 2.5.1 (Energy stability for prestrain preprocessing)**.** *Let $\tilde{\mathbf{y}}_h^0 \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ with $\tilde{E}_h[\tilde{\mathbf{y}}_h^0] \leq C$ for some constant C, where $\tilde{E}_h$ is defined as (2.5.4). There exists a constant c_0 depending on $h_{\min} := \min_{T\in\mathcal{T}_h} h_T$, $\tilde{E}_h[\tilde{\mathbf{y}}_h^0]$, g, and Ω, but independent of N, such that if $\tilde{\tau} < c_0/2$ then (2.5.8) has a unique solution $\delta\tilde{\mathbf{y}}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$, and*

if we further let $\tilde{\mathbf{y}}_h^{n+1} := \tilde{\mathbf{y}}_h^n + \delta\tilde{\mathbf{y}}_h^{n+1}$ *for any* $n \geq 0$*, then we have the energy stability estimate for any* $N \geq 0$ *as follows:*

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{N+1}] + \frac{1}{2\tilde{\tau}} \sum_{n=0}^{N} \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^0]. \tag{2.5.15}$$

Proof. For any $n \geq 0$, if $\tilde{\tau} \leq (1 + C_1\tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}})^{-1}$, there exists an unique solution $\delta\tilde{\mathbf{y}}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ to (2.5.8) by Lemma 2.5.2. Under this assumption, we take $\mathbf{v}_h = \delta\tilde{\mathbf{y}}_h^{n+1}$ in (2.5.8) to obtain

$$\tilde{\tau}^{-1}\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 + a_h^s(\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) + \boldsymbol{\epsilon}_b a_h^b(\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}) = 0 \tag{2.5.16}$$

and proceed in several steps to rewrite this expression in terms of energies and prove (2.5.15). The main difficulty is that a_h^s is quadratic in its third argument. We will also show that the assumption on $\tilde{\tau}$ can be replaced by a uniform condition.

Step (i): energy relation. Because $a_h^s(\cdot, \cdot; \tilde{\mathbf{y}}_h^n)$ is bilinear and symmetric, using the identity $(a-b)b = \frac{1}{2}a^2 - \frac{1}{2}b^2 - \frac{1}{2}(a-b)^2$, we have

$$a_h^s(\tilde{\mathbf{y}}_h^n, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) = \frac{1}{2}a_h^s(\tilde{\mathbf{y}}_h^{n+1}, \tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) - \frac{1}{2}a_h^s(\tilde{\mathbf{y}}_h^n, \tilde{\mathbf{y}}_h^n; \tilde{\mathbf{y}}_h^n) - \frac{1}{2}a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n). \tag{2.5.17}$$

Furthermore, using the same identity, we have

$$\begin{aligned}\frac{1}{2}a_h^s(\tilde{\mathbf{y}}_h^{n+1}, \tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) - \frac{1}{2}a_h^s(\tilde{\mathbf{y}}_h^n, \tilde{\mathbf{y}}_h^n; \tilde{\mathbf{y}}_h^n) &= \int_\Omega W_h^n : ((\nabla_h\tilde{\mathbf{y}}_h^n)^T\nabla_h\tilde{\mathbf{y}}_h^n - g) \\ &= E_h^s[\tilde{\mathbf{y}}_h^{n+1}] - E_h^s[\tilde{\mathbf{y}}_h^n] - \frac{1}{2}\|W_h^n\|_{L^2(\Omega)}^2,\end{aligned}$$

where

$$W_h^n := (\nabla_h \tilde{\mathbf{y}}_h^{n+1})^T \nabla_h \tilde{\mathbf{y}}_h^{n+1} - (\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \tilde{\mathbf{y}}_h^n. \tag{2.5.18}$$

Therefore, we are able to express $a_h^s(\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n)$ in terms of energies

$$\begin{aligned} a_h^s(\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) &= a_h^s(\tilde{\mathbf{y}}_h^n, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) + a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) \\ &= E_h^s[\tilde{\mathbf{y}}_h^{n+1}] - E_h^s[\tilde{\mathbf{y}}_h^n] + \frac{1}{2} a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) - \frac{1}{2}\|W_h^n\|_{L^2(\Omega)}^2. \end{aligned}$$

Similarly for $a_h^b(\cdot,\cdot)$ and noting that $E_h^b[\tilde{\mathbf{y}}_h^n] = \frac{1}{2} a_h^b(\tilde{\mathbf{y}}_h^n, \tilde{\mathbf{y}}_h^n)$, we obtain

$$a_h^b(\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}) = E_h^b[\tilde{\mathbf{y}}_h^{n+1}] - E_h^b[\tilde{\mathbf{y}}_h^n] + \frac{1}{2} a_h^b(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}) \geq E_h^b[\tilde{\mathbf{y}}_h^{n+1}] - E_h^b[\tilde{\mathbf{y}}_h^n].$$

Using these two relations in (2.5.16), we arrive at

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^n] - \tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] \geq \tilde{\tau}^{-1}\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 - R_h^n, \tag{2.5.19}$$

where

$$R_h^n := \frac{1}{2}\|W_h^n\|_{L^2(\Omega)}^2 - \frac{1}{2} a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n). \tag{2.5.20}$$

This concludes the step.

Step (ii): estimate of R_h^n. Next, our goal is to estimate R_h^n in terms of $\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2$. On the one hand, since $\delta\tilde{\mathbf{y}}_h^{n+1} \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ and $\tilde{\mathbf{y}}_h^n \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$, the

continuity property (2.5.13) of a_h^s guarantees that

$$|a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n)| \leq C_1 E_h^s[\tilde{\mathbf{y}}_h^n]^{1/2} \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 < C_1 \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2} \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2. \tag{2.5.21}$$

On the other hand, we note that

$$W_h^n = (\nabla_h \delta\tilde{\mathbf{y}}_h^{n+1})^T \nabla_h \tilde{\mathbf{y}}_h^n + (\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \delta\tilde{\mathbf{y}}_h^{n+1} + (\nabla_h \delta\tilde{\mathbf{y}}_h^{n+1})^T \nabla_h \delta\tilde{\mathbf{y}}_h^{n+1},$$

and so, taking advantage of the discrete Sobolev inequality (2.5.11), we obtain

$$\|W_h^n\|_{L^2(\Omega)}^2 \lesssim \|\nabla_h \tilde{\mathbf{y}}_h^n\|_{L^4(\Omega)}^2 \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 + \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^4. \tag{2.5.22}$$

However, we need to further estimate $\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^4$. Note that by Lemma 2.5.2 when $\tilde{\tau} \leq (1 + C_1 \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2})^{-1}$, the fact that $a_h^b(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}) \geq 0$ as well as (2.5.8) we have

$$\begin{aligned}\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 &\leq \tilde{\tau}^{-1} \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 + a_h^s(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) + \boldsymbol{\epsilon}_b a_h^b(\delta\tilde{\mathbf{y}}_h^{n+1}, \delta\tilde{\mathbf{y}}_h^{n+1}) \\ &= -a_h^s(\tilde{\mathbf{y}}_h^n, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n) - \boldsymbol{\epsilon}_b a_h^b(\tilde{\mathbf{y}}_h^n, \delta\tilde{\mathbf{y}}_h^{n+1}).\end{aligned}$$

Proceeding as in the proof of Lemma 2.5.2 we get

$$\begin{aligned}|a_h^s(\tilde{\mathbf{y}}_h^n, \delta\tilde{\mathbf{y}}_h^{n+1}; \tilde{\mathbf{y}}_h^n)| &\leq 2\|(\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \delta\tilde{\mathbf{y}}_h^{n+1}\|_{L^2(\Omega)} \|(\nabla_h \tilde{\mathbf{y}}_h^n)^T \nabla_h \tilde{\mathbf{y}}_h^n - g\|_{L^2(\Omega)} \\ &\lesssim E_h^s[\tilde{\mathbf{y}}_h^n]^{1/2} \|\nabla_h \tilde{\mathbf{y}}_h^n\|_{L^4(\Omega)} \|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)},\end{aligned}$$

and thus by continuity of a_h^b

$$\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \lesssim E_h^s[\tilde{\mathbf{y}}_h^n]^{1/2}\|\nabla_h\tilde{\mathbf{y}}_h^n\|_{L^4(\Omega)}\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)} + \boldsymbol{\epsilon}_b|||\tilde{\mathbf{y}}_h^n|||_{H_h^2(\Omega)}|||\delta\tilde{\mathbf{y}}_h^{n+1}|||_{H_h^2(\Omega)}.$$

Since $E_h^s[\tilde{\mathbf{y}}_h^n]^{1/2} < \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2}$, $\boldsymbol{\epsilon}_b|||\tilde{\mathbf{y}}_h^n|||_{H_h^2(\Omega)} \lesssim \boldsymbol{\epsilon}_b E_h^b[\tilde{\mathbf{y}}_h^n]^{1/2} < \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2}$, and the fact that $|||\cdot|||_{H_h^2(\Omega)} \leq \|\cdot\|_{H_h^2(\Omega)}$, there holds

$$\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \lesssim \tilde{E}_h[\tilde{\mathbf{y}}_h^n](\|\nabla_h\tilde{\mathbf{y}}_h^n\|_{L^4(\Omega)}^2 + 1).$$

Inserting this last relation in (2.5.22), using the inverse estimate (2.5.12) for $\nabla_h\tilde{\mathbf{y}}_h^n$ and Lemma 2.5.3 we get

$$\begin{aligned}
\|W_h^n\|_{L^2(\Omega)}^2 &\lesssim \left(1 + \tilde{E}_h[\tilde{\mathbf{y}}_h^n]\right)\left(\|\nabla_h\tilde{\mathbf{y}}_h^n\|_{L^4(\Omega)}^2 + 1\right)\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \\
&\lesssim \left(1 + \tilde{E}_h[\tilde{\mathbf{y}}_h^n]\right)(h_{\min}^{-1}\|\nabla_h\tilde{\mathbf{y}}_h^n\|_{L^2(\Omega)}^2 + 1)\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \\
&\lesssim h_{\min}^{-1}\left(1 + \tilde{E}_h[\tilde{\mathbf{y}}_h^n]\right)\left(E_h^s[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)} + 1\right)\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2
\end{aligned}$$

from which we deduce that

$$\|W_h^n\|_{L^2(\Omega)}^2 \leq C_2 h_{\min}^{-1}\left(\tilde{E}_h[\tilde{\mathbf{y}}_h^n] + 1\right)\left(\tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)} + 1\right)\|\delta\tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \quad (2.5.23)$$

for some constant C_2 independent of n and h. In summary, as in Lemma 2.5.2, as

long as $\tilde{\tau} \leq (1 + C_1 \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2})^{-1}$ we conclude that

$$|R_h^n| \leq \left(\frac{C_1}{2} \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} + \frac{C_2}{2} h_{\min}^{-1} \big(\tilde{E}_h[\tilde{\mathbf{y}}_h^n] + 1 \big) \big(\tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)} + 1 \big) \right) \|\delta \tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2. \tag{2.5.24}$$

Step(iii): conditional energy dissipation for one step. We now derive the energy dissipation at an arbitrary step, if $\tilde{\tau}$ is sufficiently small. We first define

$$c_n := \min \left\{ \left(1 + C_1 \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2} \right)^{-1}, d_n \right\}, \tag{2.5.25}$$

where

$$d_n := \left(\frac{C_1}{2} \tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{1/2} + \frac{C_2}{2} h_{\min}^{-1} \big(\tilde{E}_h[\tilde{\mathbf{y}}_h^n] + 1 \big) \big(\tilde{E}_h[\tilde{\mathbf{y}}_h^n]^{\frac{1}{2}} + \|g\|_{L^1(\Omega)} + 1 \big) \right)^{-1}. \tag{2.5.26}$$

Then if $\tilde{\tau} < c_n$, inserting (2.5.24) into (2.5.19), we conclude that

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] + (\tilde{\tau}^{-1} - c_n^{-1}) \|\delta \tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] + (\tilde{\tau}^{-1} - d_n^{-1}) \|\delta \tilde{\mathbf{y}}_h^{n+1}\|_{H_h^2(\Omega)}^2 \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^n], \tag{2.5.27}$$

for any $n \geq 0$, which further implies that $\tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] < \tilde{E}_h[\tilde{\mathbf{y}}_h^n]$ if $\delta \tilde{\mathbf{y}}_h^{n+1} \neq \mathbf{0}$.

Step(iv): uniform condition on $\tilde{\tau}$. We proceed by induction to prove that for every $n \in \mathbb{N}$, $\delta \tilde{\mathbf{y}}_h^{n+1}$ is well defined, and $c_{n+1} > c_n$, if $\tilde{\tau} < c_0$. We start with $n = 0$. In that case, by assumption $\tilde{\tau} < c_0$, Lemma 2.5.2 guarantees that $\delta \tilde{\mathbf{y}}_h^1$ is well defined and step (i)-(iii) guarantees that

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^1] < \tilde{E}_h[\tilde{\mathbf{y}}_h^0].$$

From the expression (2.5.25) of c_n, we also deduce that $c_1 > c_0$. For the induction step, we assume that $\{\delta\tilde{\mathbf{y}}_h^j\}_{j=1}^n$ is well defined, $c_j > c_{j-1}$, $j = 1, .., n$. From the latter, we see that $\tilde{\tau} < c_0 < c_n$. Therefore, using Lemma 2.5.2 and step (i)-(iii) again guarantees that $\delta\tilde{\mathbf{y}}_h^{n+1}$ is well defined and

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] < \tilde{E}_h[\tilde{\mathbf{y}}_h^n] \qquad \Longrightarrow \qquad c_{n+1} > c_n.$$

This is the desired property at step $n+1$. This concludes the induction argument.

Furthermore, from (2.5.27) we deduce that

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] + (\tilde{\tau}^{-1} - c_0^{-1})\|\delta\tilde{\mathbf{y}}_h^{n+1}\|^2_{H_h^2(\Omega)} \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^n], \tag{2.5.28}$$

for any $n \geq 0$. Moreover, if we further assume $\tilde{\tau} < c_0/2$, then we can observe that $\tilde{\tau}^{-1} - c_0^{-1} > (2\tilde{\tau})^{-1}$, and consequently

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] + \frac{1}{2\tilde{\tau}}\|\delta\tilde{\mathbf{y}}_h^{n+1}\|^2_{H_h^2(\Omega)} \leq \tilde{E}_h[\tilde{\mathbf{y}}_h^n], \tag{2.5.29}$$

for any $n \geq 0$. Then we sum (2.5.29) over $n = 0, \ldots, N$ to get (2.5.15). This concludes the proof. □

From the condition $\tilde{\tau} < c_0/2$ and definitions (2.5.25) and (2.5.26) with $n = 0$, we see that $\tilde{\tau} \to 0$ as $h_{\min} \to 0$.

Remark 2.5.1 (Choice of $\boldsymbol{\epsilon}_b$)**.** *Under the assumptions of Theorem 2.5.1, the preprocessing gradient flow produces a sequence of deformations* $\{\tilde{\mathbf{y}}_h^n\}_{n\in\mathbb{N}}$ *with decreasing*

energies $\tilde{E}_h[\tilde{\mathbf{y}}_h^n]$. *We now choose* $\boldsymbol{\epsilon}_b \sim h^2$ *and assume that the* N_h*-th iterate of the preprocessing gradient flow, denoted* $\tilde{\mathbf{y}}_h^{N_h}$, *is such that*

$$\tilde{E}_h[\tilde{\mathbf{y}}_h^{N_h}] \lesssim h^2.$$

Then, according to (2.5.7), *the prestrain defect of* $\mathbf{y}_h^{N_h}$ *satisfies*

$$D_h(\tilde{\mathbf{y}}_h^{N_h}) \lesssim E_h^s[\tilde{\mathbf{y}}_h^{N_h}]^{1/2} \lesssim h, \qquad \textit{i.e. } \tilde{\mathbf{y}}_h^{N_h} \in \mathbb{A}_{h,\boldsymbol{\epsilon}_0}^k$$

for $\boldsymbol{\epsilon}_0 \sim h$. *Moreover, we have*

$$E_h^b[\tilde{\mathbf{y}}_h^{N_h}] \lesssim \boldsymbol{\epsilon}_b^{-1}\tilde{E}_h[\tilde{\mathbf{y}}_h^{N_h}] \lesssim 1.$$

This implies that $E_h[\tilde{\mathbf{y}}_h^{N_h}]$ *is also uniformly bounded by continuity of* E_h *and coercivity of* E_h^b. *As a consequence, the main gradient flow (Section 2.4) with initial conditions* $\mathbf{y}_h^0 = \tilde{\mathbf{y}}_h^{N_h}$ *has uniformly bounded initial energy and small initial prestrain defect, and thus controls the final prestrain defect; see Theorem 2.4.1. In addition, recall the main gradient flow (Theorem 2.4.1) is energy decreasing, and hence the final iterate of it is also uniformly bounded. Therefore, if the main gradient flow leads to an almost global minimizer* $\mathbf{y}_h^{n_h}$ *of the energy* E_h, *then the sequence* $\{\mathbf{y}_h^{n_h}\}_{h>0}$ *has uniformly bounded energies.*

Remark 2.5.2. *Note that when* $\Gamma^D \neq \emptyset$, *the boundary conditions are enforced in the following sense. As in Proposition 2.4.1, but now due to* (2.5.15), *we can conclude*

that there exists the limit $\tilde{\mathbf{y}}_h^\infty \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ *such that* $\tilde{\mathbf{y}}_h^n$ *converges to it in any norm and up to a subsequence as* $n \to \infty$. *Due to Remark* (2.5.1), *note that* $E_h^b[\tilde{\mathbf{y}}_h^\infty] \leq C$ *is valid. Consequently, we have* $\|[\tilde{\mathbf{y}}_h^\infty]\|^2_{L^2(\Gamma^D)} \lesssim Ch^3$ *and* $\|[\nabla_h \tilde{\mathbf{y}}_h^\infty]\|^2_{L^2(\Gamma^D)} \lesssim Ch$, *for* $\tilde{\mathbf{y}}_h^\infty \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$.

Another observation is that if we start from a flat initialization (x_3-coordinate of $\tilde{\mathbf{y}}_h^0$ is 0) in the flow (2.5.8), then the resulting stationary solution $\tilde{\mathbf{y}}_h^N$ is flat; we show this next. This justifies taking $\tilde{\mathbf{y}}_h^0$ as a discrete solution of the bi-Laplacian equation (2.5.2) as proposed in Algorithm 2 with a non-zero fictitious force.

Corollary 2.5.1. *If* $\tilde{\mathbf{y}}_h^n$ *is of form* $(f(x_1, x_2), g(x_1, x_2), 0)$ *and* $\delta\tilde{\mathbf{y}}_h^{n+1}$ *is the solution of the gradient flow* (2.4.8), *then* $\tilde{\mathbf{y}}_h^{n+1} := \tilde{\mathbf{y}}_h^n + \delta\tilde{\mathbf{y}}_h^{n+1}$ *is also of form*

$$(\tilde{f}(x_1, x_2), \tilde{g}(x_1, x_2), 0).$$

Proof. Assume $\delta\tilde{\mathbf{y}}_h^{n+1} = (d_1, d_2, d_3) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$, where the d_i are functions of x_1, x_2. Let $\phi \in \mathbb{V}_h^k$ be an arbitrary scalar function such that $\mathbf{v}_h = (0, 0, \phi) \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$, then we have

$$(\nabla \mathbf{v}_h)^T \nabla \tilde{\mathbf{y}}_h^n = \begin{bmatrix} 0 & 0 & \partial_1\phi \\ 0 & 0 & \partial_2\phi \end{bmatrix} \begin{bmatrix} \partial_1 f & \partial_2 f \\ \partial_1 g & \partial_2 g \\ 0 & 0 \end{bmatrix} = \mathbf{0}. \tag{2.5.30}$$

Hence, we deduce $a_h^s(\tilde{\mathbf{y}}_h^n, \mathbf{v}_h; \tilde{\mathbf{y}}_h^n) = 0$. Similarly, invoking (2.5.10) which multiplies Hessians of the arguments componentwise we see that $a_h^b(\tilde{\mathbf{y}}_h^n, \mathbf{v}_h) = 0$. Consequently, the right hand side of (2.5.8) is zero.

Also, we have

$$(\nabla \mathbf{v}_h)^T \nabla \delta \tilde{\mathbf{y}}_h^{n+1} = \begin{bmatrix} 0 & 0 & \partial_1 \phi \\ 0 & 0 & \partial_2 \phi \end{bmatrix} \begin{bmatrix} \partial_1 d_1 & \partial_2 d_1 \\ \partial_1 d_2 & \partial_2 d_2 \\ \partial_1 d_3 & \partial_2 d_3 \end{bmatrix} = \begin{bmatrix} \partial_1 \phi \partial_1 d_3 & \partial_1 \phi \partial_2 d_3 \\ \partial_2 \phi \partial_1 d_3 & \partial_2 \phi \partial_2 d_3 \end{bmatrix}, \quad (2.5.31)$$

as well as,

$$(\delta \tilde{\mathbf{y}}_h^{n+1}, \mathbf{v}_h)_{H_h^2(\Omega)} = (d_3, \phi)_{H_h^2(\Omega)}, \quad (2.5.32)$$

whence $a_h^b(\delta \tilde{\mathbf{y}}_h^{n+1}, \mathbf{v}_h)$ reduces to the bilinear form only on d_3 and ϕ.

As a result, taking $\phi = d_3$, and using Lemma 2.5.2 and non-negativity of the a_h^b term, we have the following when $\tilde{\tau}$ small enough as in Lemma 2.5.2:

$$\|d_3\|_{H_h^2(\Omega)}^2 \lesssim 0. \quad (2.5.33)$$

This means that $d_3 = 0$ as $\|\cdot\|_{H_h^2(\Omega)}$ defines a norm. As this third component of the increment $\delta \tilde{\mathbf{y}}_h^{n+1}$ is 0, we have the claimed result. □

2.6 Numerical experiments

We first make a few comments on the implementation of the gradient flow (2.4.8), built in Algorithm 1, and the resulting linear algebra solver used at each step.

Let $\{\boldsymbol{\varphi}_h^i\}_{i=1}^N$ be a basis for $\mathbb{V}_h^k(\mathbf{0}, \mathbf{0})$ and let $\{\boldsymbol{\psi}_h^i\}_{i=1}^M$ be a basis for Λ_h. The

discrete problem (2.4.8) is a *saddle-point problem* of the form

$$\begin{bmatrix} A & B_n^T \\ B_n & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta Y}_h^{n+1} \\ \boldsymbol{\Lambda}_h^{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_n \\ \mathbf{0} \end{bmatrix}. \tag{2.6.1}$$

Here, $(\boldsymbol{\delta Y}_h^{n+1}, \boldsymbol{\Lambda}_h^{n+1})$ are the nodal values of $(\delta \mathbf{y}_h^{n+1}, \boldsymbol{\lambda}_h^{n+1})$ in these bases, while $A = (A_{ij})_{i,j=1}^N \in \mathbb{R}^{N\times N}$ is the matrix corresponding to the first two terms of (2.4.8)

$$A_{ij} := \tau^{-1}(\boldsymbol{\varphi}_h^j, \boldsymbol{\varphi}_h^i)_{H_h^2(\Omega)} + \widetilde{A}_{ij} \quad \text{with} \quad \widetilde{A}_{ij} := a_h(\boldsymbol{\varphi}_h^j, \boldsymbol{\varphi}_h^i), \quad i,j = 1,\dots,N,$$

while the matrix $B_n \in \mathbb{R}^{M\times N}$ corresponds to the bilinear form $b_h(\cdot,\cdot;\mathbf{y}_h^n)$ and is given by

$$(B_n)_{ij} := b_h(\boldsymbol{\varphi}_h^j, \boldsymbol{\psi}_h^i; \mathbf{y}_h^n) \quad i = 1,\dots,M,\ j = 1,\dots,N.$$

The vector $\mathbf{F}_n \in \mathbb{R}^N$ accounts for the right-hand-side of (2.4.8). It reads $\mathbf{F}_n = \mathbf{F} + \mathbf{L} - \widetilde{A}\mathbf{Y}^n$, where $\mathbf{Y}^n$ contains the nodal values of $\mathbf{y}_h^n$ in the basis $\{\boldsymbol{\varphi}_h^i\}_{i=1}^N$ while $\mathbf{F} = (F_i)_{i=1}^N$ and $\mathbf{L} = (L_i)_{i=1}^N$ are defined by

$$F_i := F_h(\boldsymbol{\varphi}_h^i) \quad \text{and} \quad L_i := -a_h(\bar{\mathbf{0}}, \boldsymbol{\varphi}_h^i), \quad i = 1,\dots,N.$$

Here, $\bar{\mathbf{0}}$ denotes the zero function in the space $\mathbb{V}_h(\boldsymbol{\varphi}, \Phi)$, which is 0 everywhere but equal to $\boldsymbol{\varphi}$ and its gradient equal to Φ on $e \in \mathcal{E}_h^D$; $\mathbf{L}$ contains the liftings of the boundary data. Since B_n and $\mathbf{F}_n$ depend explicitly on the current deformation $\mathbf{y}_h^n$, they have to be re-computed at each iteration of Algorithm 1 (gradient flow). In contrast, the matrices A and $\widetilde{A}$ and the vector $\mathbf{L}$, which are the most costly to

assemble because of the reconstructed Hessians, are independent of the iteration number n and can thus be computed once for all.

More precisely, to compute the element-wise contribution on a cell T, the discrete Hessian (2.2.10) of each basis function associated with T along with those associated with the neighboring cells is computed. Recall that for any interior edge $e \in \mathcal{E}_h^i$, the support of the liftings r_e and b_e in (2.2.7) and (2.2.8) is the union of the two cells sharing e as an edge. We employ direct solvers for these small systems. We proceed similarly for the computation of the liftings of the boundary data $\boldsymbol{\varphi}$ and Φ. Once the discrete Hessians are computed, the rest of the assembly process is standard. Incidentally, we note that the proposed LDG approach couples the degree of freedom (DoFs) of all neighboring cells (not only the cell with its neighbors). As a consequence, the sparsity pattern of LDG is slightly larger than that for a standard symmetric interior penalty dG (SIPG) method. However, the stability properties of LDG are superior to those of SIPG.

System (2.6.1) can be solved using the *Schur complement method*. Denoting $S_n := B_n A^{-1} B_n^T$ the Schur complement matrix, the first step determines $\mathbf{\Lambda}_h^{n+1}$ satisfying

$$S_n \mathbf{\Lambda}_h^{n+1} = B_n A^{-1} \mathbf{F}_n, \tag{2.6.2}$$

followed by the computation of $\delta \mathbf{Y}_h^{n+1}$ solving

$$A \delta \mathbf{Y}_h^{n+1} = \mathbf{F}_n - B_n^T \mathbf{\Lambda}_h^{n+1}. \tag{2.6.3}$$

Because the matrix A is independent of the iterations, we pre-compute its LU decomposition once for all and use it whenever the action of A^{-1} is needed in (2.6.2) and (2.6.3). Furthermore, a conjugate gradient algorithm is utilized to compute $\mathbf{\Lambda}_h^{n+1}$ in (2.6.2) to avoid assembling S_n. The efficiency of the latter depends on the condition number of the matrix S_n, which in turn depends on the inf-sup constant of the saddle-point problem (2.6.1). Leaving aside the preprocessing step, we observe in practice that solving the Schur complement problem (2.6.2) is the most time consuming part of the simulation. Finally, we point out that the stabilization parameters γ_0 and γ_1 influence the number of Schur complement iterations: more iterations of the gradient conjugate algorithm are required for larger stabilization parameter values. We refer to Tables 2.1 and 2.2 below for more details.

Then we present a collection of numerical experiments to illustrate the performance of the proposed methodology. We consider several prestrain tensors g, as well as both $\Gamma_D \neq \emptyset$ (Dirichlet boundary condition) and $\Gamma_D = \emptyset$ (free boundary condition). The Algorithms 1 and 2 are implemented using the deal.ii library [7] and the visualization is performed with paraview [6]. The color code is the following: (multicolor figures) dark blue indicates the lowest value of the deformation's third component while dark red indicate the largest value of the deformation's third component; (unicolor figures) magnitude of the deformation's third component.

For all the simulations, we fix the polynomial degree k of the deformation $\mathbf{y}_h$ and l_1, l_2 for the two liftings of the discrete Hessian $H_h[\mathbf{y}_h]$ to be

$$k = l_1 = l_2 = 2.$$

Moreover, unless otherwise specified, we set the Lamé coefficients to $\lambda = 8$ and $\mu = 6$, and the stabilization parameters for (2.2.13) and (2.5.3) to be

$$\gamma_0 = \gamma_1 = 1, \qquad \widehat{\gamma}_0 = \widehat{\gamma}_1 = 1.$$

In striking contrast to [22, 23], these parameters do not need to be large for stability purposes. When $\Gamma_D = \emptyset$, we set $\sigma = 1$ in (2.4.1). Finally, we choose $tol = 10^{-6}$ for the stopping criteria in Algorithm 1 (gradient flow). In this section, the results are presented for $\boldsymbol{\epsilon}_b = 0$ (regularization parameter introduced in Section 2.5.1 for metric preprocessing), but we tested them for $\boldsymbol{\epsilon}_b = 10^{-4}$ and notice that the difference is negligible computationally.

To record the energy E_h and metric defect D_h after the three key procedures described in Algorithms 1 and 2, we resort to the following notation: *BC PP* (boundary conditions preprocessing); *Metric PP* (metric preprocessing); *Final* (gradient flow).

2.6.1 Vertical load and isometry constraint

This first example has been already investigated in [10, 22]. We consider the square domain $\Omega = (0, 4)^2$, the metric $g = I_2$ (isometry) and a vertical load $\mathbf{f} = (0, 0, 0.025)^T$. Moreover, the plate is clamped on $\Gamma_D = \{0\} \times [0, 4] \cup [0, 4] \times \{0\}$, i.e., we prescribe the Dirichlet boundary condition (2.1.20) with

$$\boldsymbol{\varphi}(x_1, x_2) = (x_1, x_2, 0)^T, \quad \Phi = [I_2, \mathbf{0}]^T \qquad (x_1, x_2) \in \Gamma_D.$$

Finally, we set the Lamé constant $\lambda = 0$ thereby removing the trace term in (2.2.11).

No preprocessing step is required because the flat plate, which corresponds to the identity deformation $\mathbf{y}_h^0(\Omega) = \Omega$, satisfies the metric constraint and the boundary conditions. For the discretization of Ω, we use $\ell = 0, 1, 2, \cdots$ to denote the refinement level and consider uniform partitions $\mathcal{T}_\ell$ consisting of squares T of side-length $4/2^\ell$ and diameters $h_T = h = \sqrt{2}/2^{\ell-2}$. The pseudo-time step used for the discretization of the gradient flow is chosen so that $\tau = h$. The discrete energy $E_h[\mathbf{y}_h]$ and metric defect $D_h[\mathbf{y}_h]$ for $\ell = 3, 4, 5$ are report in Table 2.1 along with the number of gradient flow iterations (GF Iter) required to reach the targeted stationary tolerance and the range of number of iterations (Schur Iter) needed to solve the Schur complement problem (2.6.2). Note that in this case we have $D_h[\mathbf{y}_h^0] = 0$, namely $\mathbf{y}_h^0 \in \mathbb{A}^k_{h,\varepsilon_0}$ with $\varepsilon_0 = 0$.

Nb. cells	DoFs	$\tau = h$	E_h	D_h	GF Iter	Schur Iter
64	1920	$\sqrt{2}/2$	-1.002E-2	1.062E-2	11	[60,65]
256	7680	$\sqrt{2}/4$	-9.709E-3	5.967E-3	17	[85,101]
1024	30720	$\sqrt{2}/8$	-8.762E-3	2.962E-3	28	[118,148]

Table 2.1: Effect of the numerical parameters h and $\tau = h$ on the energy and prestrain defect for the vertical load example using $\gamma_0 = \gamma_1 = 1$. As expected [10, 13, 22], we observe that $D_h[\mathbf{y}_h]$ is $\mathcal{O}(h)$. The number of iterations needed by the gradient flow and for each Schur complement solver increases with the resolution.

We point out that the SIPG method analyzed in [22] requires $\gamma_0 = 5000$ and $\gamma_1 = 1100$ in this example. We report in Table 2.2 the performance of both methods with this choice of stabilization parameters.

Based on Table 2.2, we see that the two methods give similar results. The advantage of the LDG approach is that there is no constraint on the stabilization

	LDG				SIPG			
$\tau = h$	E_h	D_h	GF	Schur	E_h	D_h	GF	Schur
$\sqrt{2}/2$	-8.28E-3	7.71E-3	7	[302,321]	-8.30E-3	7.72E-3	7	[284,307]
$\sqrt{2}/4$	-6.63E-3	3.45E-3	14	[557,605]	-6.64E-3	3.46E-3	13	[556,600]
$\sqrt{2}/8$	-4.88E-3	1.34E-3	37	[788,831]	-4.90E-3	1.34E-3	35	[787,833]

Table 2.2: Comparison of the LDG and SIPG methods using the penalization parameters $\gamma_0 = 5000$, $\gamma_1 = 1100$ required by the SIPG. The results are similar.

parameters γ_0 and γ_1 other than being positive. In contrast, the coercivity of the energy discretized with the SIPG method requires γ_0 and γ_1 to be sufficiently large (depending on the maximum number of edges of the elements in the subdivision $\mathcal{T}$ and the constant in the trace inequality) [22]. For instance, the choice $\gamma_0 = \gamma_1 = 1$ for the SIPG method yields an unstable scheme and the problem (2.4.8) becomes singular after a few iterations of the gradient flow. Moreover, the large values of γ_0, γ_1 are mainly dictated by the penalty of the boundary terms in $E_h[\mathbf{y}_h^0]$ and the need to produce moderate values of $E_h[\mathbf{y}_h^0]$ to prevent very small time steps τ in (2.4.12). Furthermore, within each gradient flow iteration, the solution of the Schur complement problem (2.6.2) using the LDG approach with $\gamma_0 = \gamma_1 = 1$ (reported in Table 2.1) requires less than a fifth of the iterations (Schur Iter) for SIPG with $\gamma_0 = 5000$ and $\gamma_1 = 1100$ (reported in Table 2.2) at the expense of slightly larger number of iterations of the gradient flow (GF Iter); compare Tables 2.1 and 2.2. This documents a superior performance of LDG relative to SIPG.

Note that there is an artificial displacement along the diagonal $x_1 + x_2 = 4$ [10, 22] for this example, which does not correspond to the actual physics of the problem, namely $y = 0$ for $x_1 + x_2 \leq 4$. The artificial displacements obtained by

the two methods for various meshes are compared in Figure 2.2 and Table 2.3.

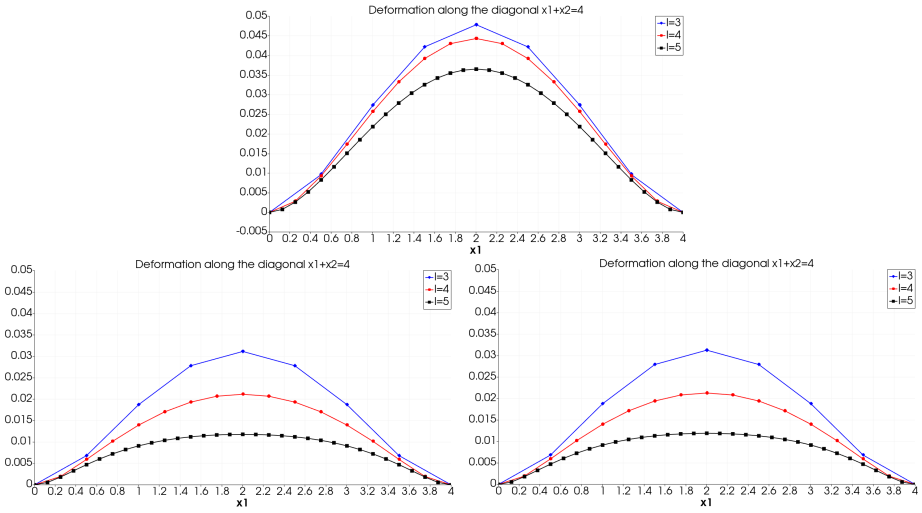

Figure 2.2: Deformation along the diagonal $x_1 + x_2 = 4$. Top: LDG with $\gamma_0 = \gamma_1 = 1$; bottom-left: LDG with $\gamma_0 = 5000$ and $\gamma_1 = 1100$; bottom-right: SIPG with $\gamma_0 = 5000$ and $\gamma_1 = 1100$. The deflection is slightly larger when $\gamma_0 = \gamma_1 = 1$ while both methods yield similar results when $\gamma_0 = 5000$ and $\gamma_1 = 1100$; see Table 2.3.

	LDG		SIPG
♯ ref.	$\gamma_0 = \gamma_1 = 1$	$\gamma_0 = 5000, \gamma_1 = 1100$	$\gamma_0 = 5000, \gamma_1 = 1100$
$l = 3$	0.0478	0.0311	0.0312
$l = 4$	0.0443	0.0211	0.0213
$l = 5$	0.0365	0.0118	0.0119

Table 2.3: Deflection y_3 along the diagonal $x_1 + x_2 = 4$ for both LDG and SIPG

2.6.2 Rectangle with *cylindrical* metric

The domain is the rectangle $\Omega = (-2, 2) \times (-1, 1)$ and the Dirichlet boundary is $\Gamma_D = \{-2\} \times (-1, 1) \cup \{2\} \times (-1, 1)$. The mesh $\mathcal{T}_h$ is uniform and made of 1024 rectangular cells of diameter $h_T = h = \sqrt{5}/4$ (30720 DoFs) and the pseudo time-step is fixed to $\tau = 0.1$.

2.6.2.1 *One mode*

We first consider the immersible metric

$$g(x_1, x_2) = \begin{bmatrix} 1 + \frac{\pi^2}{4} \cos\left(\frac{\pi}{4}(x_1 + 2)\right)^2 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.6.4}$$

for which

$$\mathbf{y}(x_1, x_2) = (x_1, x_2, 2\sin(\frac{\pi}{4}(x_1 + 2)))^T \tag{2.6.5}$$

is a compatible deformation (isometric immersion), i.e., $\mathrm{I}[\mathbf{y}] = g$. We impose the boundary conditions $\boldsymbol{\varphi} = \mathbf{y}|_{\Gamma_D}$ and $\Phi = \nabla\mathbf{y}|_{\Gamma_D}$, so that $\mathbf{y} \in \mathbb{V}(\boldsymbol{\varphi}, \Phi)$ is an admissible deformation and also a global minimizer of the energy.

To challenge our algorithm, we start from a flat initial plate and obtain an admissible initial deformation $\mathbf{y}_h^0$ using the two preprocessing steps (BC PP and Metric PP) in Algorithm 2 with parameters

$$\widetilde{\tau} = 0.05, \quad \widetilde{\varepsilon}_0 = 0.1 \quad \text{and} \quad \widetilde{tol} = 10^{-6}.$$

The deformation obtained after applying Algorithms 2 and 1 are displayed in Figure 2.3. Moreover, the corresponding energy and prestrain defect are reported in Table 2.4. Notice that the target metric defect $\widetilde{\varepsilon}_0$ is reached in 49 iterations while 380 iterations of the gradient flow are needed to reach the stationary deformation.

Interestingly, when no Dirichlet boundary conditions are imposed, i.e., the free

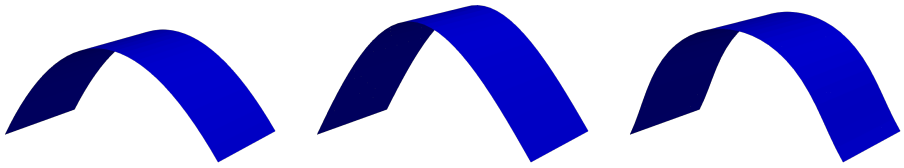

Figure 2.3: Deformed plate for the cylinder metric with one mode. Left: BC PP; middle: Metric PP; right: Final.

	Initial	BC PP	Metric PP	Final
E_h	120.3590	1.1951	2.5464	1.7707
D_h	9.8696	3.2899	9.8609E-2	9.5183E-2

Table 2.4: Energy and prestrain defect for the cylinder metric with one mode. All the algorithms behave as intended: the boundary conditions preprocessing (BC PP) reduces the energy by constructing a deformation with compatible boundary conditions, the metric preprocessing (Metric PP) reduces the metric defect and the gradient flow (Final) reduced the energy to its minimal value while keeping a control on the metric defect.

boundary case, then the flat deformation (pure stretching)

$$\mathbf{y}(x_1, x_2) = \left(\int_{-2}^{x_1} \sqrt{1 + \frac{\pi^2}{4} \cos\left(\frac{\pi}{4}(s+2)\right)^2}\, ds, x_2, 0 \right)^T$$

is also compatible with the metric (2.6.4) and has a smaller energy. We observe that $y_1(2, x_2) - y_1(-2, x_2) \approx 5.85478$ for $x_2 \in (-2, 2)$ corresponds to a stretching ratio of approximately 1.5. The outcome of Metric PP in Algorithm 2 starting from the flat plate produces an initial deformation with $E_h = 0.81755$ and $D_h = 0.09574$ using 37 iterations. The stationary solution of the main gradient flow is reached in 68 iterations and produces a flat plate with energy $E_h = 0.376257$ and metric defect $D_h = 0.0957329$.

2.6.2.2 *Two modes*

This example is similar to that of Section 2.6.2.1 but with one additional *mode* of higher frequency, namely we consider the immersible metric

$$g(x_1, x_2) = \begin{bmatrix} 1 + \left(\frac{\pi}{2}\cos\left(\frac{\pi}{4}(x_1+2)\right) + \frac{5\pi}{8}\cos\left(\frac{5\pi}{4}(x_1+2)\right)\right)^2 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case, the deformation

$$\mathbf{y}(x_1, x_2) = \left(x_1, x_2, 2\sin\left(\frac{\pi}{4}(x_1+2)\right) + \frac{1}{2}\sin\left(\frac{5\pi}{4}(x_1+2)\right)\right)^T$$

is compatible (isometric immersion) with the metric and we impose the corresponding Dirichlet boundary conditions on Γ_D as in Section 2.6.2.1.

Using the same setup as in Section 2.6.2.1, Algorithm 2 produced a suitable initial guess in 1271 iterations, while Algorithm 1 terminated after 1833 steps. The deformations obtained after each of the three main procedures are given in Figure 2.4. The corresponding energy and prestrain defect are reported in Table 2.5. We see that the main gradient flow decreases the energy upon bending the shape but keeping the metric defect roughly constant.

	Initial	BC PP	Metric PP	Final
E_h	413.7400	5.5344	28.9184	13.0706
D_h	25.2909	26.1854	9.9997E-2	1.0178E-1

Table 2.5: Energy and metric defect for the cylinder metric with two modes. Compare with Table 2.4 corresponding to one mode.

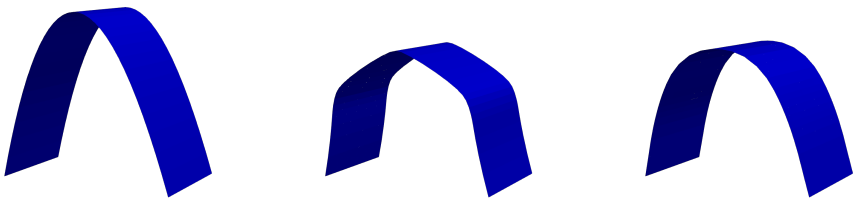

Figure 2.4: Deformed plate for the cylinder metric with two modes. Left: BC PP; middle: Metric PP; right: Final. Compare with Figure 2.3 corresponding to the metric (2.6.4) (one mode).

2.6.3 Rectangle with a *catenoidal-helicoidal* metric

Let Ω be a rectangle to be specified later and let the metric be

$$g(x_1, x_2) = \begin{bmatrix} \cosh(x_2)^2 & 0 \\ 0 & \cosh(x_2)^2 \end{bmatrix}. \tag{2.6.6}$$

Notice that the family of deformations $\mathbf{y}^\alpha : \Omega \to \mathbb{R}^3$, $0 \leq \alpha \leq \frac{\pi}{2}$, defined by

$$\mathbf{y}^\alpha := \cos(\alpha)\bar{\mathbf{y}} + \sin(\alpha)\tilde{\mathbf{y}} \tag{2.6.7}$$

with

$$\bar{\mathbf{y}}(x_1, x_2) = \begin{bmatrix} \sinh(x_2)\sin(x_1) \\ -\sinh(x_2)\cos(x_1) \\ x_1 \end{bmatrix}, \quad \tilde{\mathbf{y}}(x_1, x_2) = \begin{bmatrix} \cosh(x_2)\cos(x_1) \\ \cosh(x_2)\sin(x_1) \\ x_2 \end{bmatrix},$$

are all compatible with the metric (2.6.6). The parameter $\alpha = 0$ corresponds to an *helicoid* while $\alpha = \pi/2$ represents a *catenoid*. Furthermore, the energy $E[\mathbf{y}^\alpha]$ defined

in (2.1.33) (or equivalently $E[\mathbf{y}^\alpha]$ given in (2.1.19)) has the same value for all α. To see this, it suffices to note that the second fundamental form of $\mathbf{y}^\alpha$ is given by

$$\mathrm{I\!I}[\mathbf{y}^\alpha] = \begin{bmatrix} -\cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}, \quad D^2 y_k^\alpha = \cos(\alpha) D^2 \bar{y}_k + \sin(\alpha) D^2 \tilde{y}_k,$$

where $y_k^\alpha = (\mathbf{y}^\alpha)_k$ is the kth component of $\mathbf{y}^\alpha$ for $k = 1, 2, 3$.

In the following sections, we show how the two extreme deformations can be obtained either by imposing the adequate boundary conditions or by starting with an initial configuration sufficiently close to the energy minima.

2.6.3.1 *Catenoid case*

We consider the domain $\Omega = (0, 6.25) \times (-1, 1)$. The mesh $\mathcal{T}_h$ consists of 896 (almost square) rectangular cells of diameter $h_T = h \approx 0.17$ (26880 DoFs). We do not impose any boundary conditions on the deformations, which corresponds to $\Gamma_D = \emptyset$ (free boundary condition). We apply Algorithm 2 (initialization) and start the metric preprocessing with $\widetilde{\mathbf{y}}_h^0 = \widehat{\mathbf{y}}_h$, the solution to the bi-Laplacian problem (2.5.1) with fictitious force $\widehat{\mathbf{f}} = (0, 0, 4)^T$ and boundary condition $\boldsymbol{\varphi}(\mathbf{x}) = (\mathbf{x}, 0)$ on $\partial\Omega$ (but without Φ). Moreover, we use three tolerances $\widetilde{tol} = 0.1,\ 0.025,\ 0.01$ for this preprocessing to investigate the effect on Algorithm 1 (gradient flow). Figure 2.5 depicts final configurations produced by Algorithm 1 with the outputs of Algorithm 2. Corresponding energies and metric defects are given in Table 2.6. We see that the metric defect diminishes, as $\widetilde{tol}$ decreases, and the surface tends to a full (closed)

catenoid as expected from the relation (2.6.7) with $\alpha = \pi/2$.

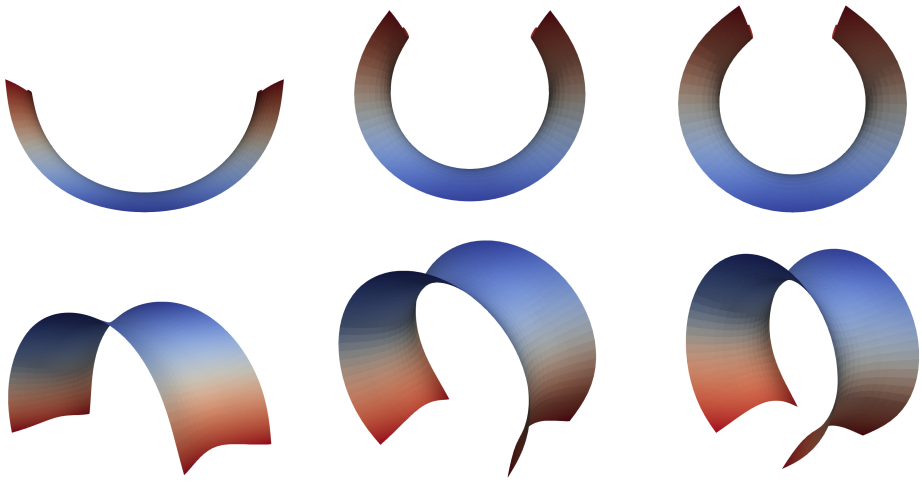

Figure 2.5: Final configurations for the *catenoidal-helicoidal* metric with free boundary condition using tolerances $\widetilde{tol} = 0.1$ (left), 0.025 (middle) and 0.01 (right) for the metric preprocessing of Algorithm 2. The second row offers a different view of the final deformations.

	$\widetilde{tol} = 0.1$		$\widetilde{tol} = 0.025$		$\widetilde{tol} = 0.01$	
	Algo 2	Algo 1	Algo 2	Algo 1	Algo 2	Algo 1
E_h	36.9461	4.01094	103.838	7.42946	146.215	8.78622
D_h	2.62428	3.19839	1.36864	2.69258	0.853431	1.83427

Table 2.6: Energies E_h and metric defects D_h produced by Algorithms 2 and 1 for the *catenoidal-helicoidal* metric with free boundary condition. We see that the tolerance $\widetilde{tol}$ of Algorithms 2 controls D_h and that Algorithm 1 does not increase D_h much but reduces E_h substantially. The smaller $\widetilde{tol}$ is the closer the computed surface gets to the catenoid, which is closed (see Figure 2.5).

2.6.3.2 ***Helicoid shape***

All the deformations $\mathbf{y}^\alpha$ in (2.6.7) are global minima of the energy but the final deformation is not always catenoid-like as in the previous section. In fact, starting with an initial deformation close to $\mathbf{y}^\alpha$ with $\alpha = 0$ leads to an helicoid-like shape.

We postpone such an approach to Section 2.6.4.3. An alternative to achieve an helicoid-like shape is to enforce the appropriate boundary conditions as described now.

We consider the domain $\Omega = (0, 4.5) \times (-1, 1)$ and enforce Dirichlet boundary conditions on $\Gamma_D = \{0\} \times (-1, 1)$ compatible with $\mathbf{y}^\alpha$ given by (2.6.7) with $\alpha = 0$. The mesh $\mathcal{T}_h$ consists of 640 (almost square) rectangular cells of diameter $h_T = h \approx 0.17$ (19200 DoFs) and the pseudo time-step is $\tau = 0.01$.

We apply Algorithm 2 (preprocessing) with $\widetilde{\tau} = 0.01$, $\widetilde{\varepsilon}_0 = 0.1$ and $\widetilde{tol} = 10^{-3}$ to obtain the initial deformation $\mathbf{y}_h^0$. The preprocessing stopped after 2555 iterations, meeting the criteria $\widetilde{\tau}^{-1}|\widetilde{E}_h[\tilde{\mathbf{y}}_h^{n+1}] - \widetilde{E}_h[\tilde{\mathbf{y}}_h^n]| \leq \widetilde{tol}$, while 2989 iterations of Algorithm 1 (gradient flow) were needed to reach the stationary deformation. Figure 2.6 displays the output of the boundary conditions preprocessing and the metric preprocessing, the two stages of Algorithm 2, as well as two views of the output of Algorithm 1. The corresponding energies and metric defects are reported in Table 2.7.

	Initial	BC PP	Metric PP	Final
E_h	138020	0.658342	202.144	7.7461
D_h	5.17664	5.16565	0.248419	1.15764

Table 2.7: Energies and metric defects for the helicoid-like shape with Dirichlet boundary conditions on the bottom side.

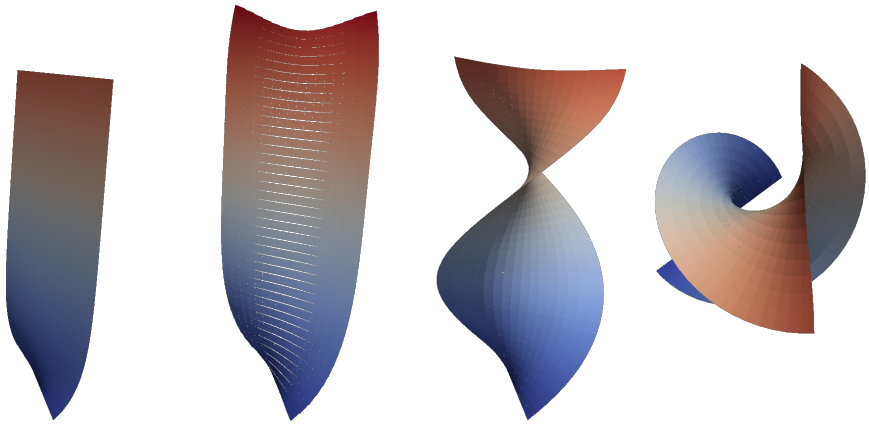

Figure 2.6: Deformed plate for the *catenoidal-helicoidal* with Dirichlet boundary conditions on the bottom side corresponding to $\{0\} \times (-1, 1)$. From left to right: BC PP, Metric PP, and two views (the last from the top) of the output of Algorithm 1.

2.6.4 Disc with positive or negative Gaussian curvature

We now consider a plate consisting of a disc of radius 1

$$\Omega = \Big\{(x_1, x_2) \in \mathbb{R}^2 : \quad x_1^2 + x_2^2 < 1\Big\}.$$

We prescribe several immersible metrics g and impose no boundary conditions.

The mesh $\mathcal{T}_h$ consists of 320 quadrilateral cells of diameter $0.103553 \leq h_T \leq 0.208375$ (9600 DoFs) and the pseudo time-step is $\tau = 0.01$. Moreover, we initialize the metric preprocessing of Algorithm 2 with the identity function $\widetilde{\mathbf{y}}_h^0(\mathbf{x}) = (\mathbf{x}, 0)^T$ for $\mathbf{x} \in \Omega$, and $\widetilde{\tau} = 0.05$, $\widetilde{\varepsilon}_0 = 0.1$, $\widetilde{tol} = 10^{-6}$.

2.6.4.1 *Bubble - positive Gaussian curvature*

To obtain a bubble-like shape, we consider for any $\alpha > 0$ the metric

$$g(x_1, x_2) = \begin{bmatrix} 1 + \alpha\frac{\pi^2}{4}\cos\left(\frac{\pi}{2}(1-r)\right)^2 \frac{x_1^2}{r^2} & \alpha\frac{\pi^2}{4}\cos\left(\frac{\pi}{2}(1-r)\right)^2 \frac{x_1 x_2}{r^2} \\ \alpha\frac{\pi^2}{4}\cos\left(\frac{\pi}{2}(1-r)\right)^2 \frac{x_1 x_2}{r^2} & 1 + \alpha\frac{\pi^2}{4}\cos\left(\frac{\pi}{2}(1-r)\right)^2 \frac{x_2^2}{r^2} \end{bmatrix} \tag{2.6.8}$$

with $r := \sqrt{x_1^2 + x_2^2}$. A compatible deformation is given by

$$\mathbf{y}(x_1, x_2) = \left(x_1, x_2, \sqrt{\alpha}\sin\left(\frac{\pi}{2}(1-r)\right)\right)^T,$$

i.e., $\mathbf{y}$ is an isometric immersion $\mathrm{I}[\mathbf{y}] = g$. In the following, we choose $\alpha = 0.2$.

In the absence of boundary conditions and forcing term, the flat configuration $\tilde{\mathbf{y}}_h^0(\Omega) = \Omega$ has zero energy but has a metric defect of $D_h = 1.0857$. Algorithm 2 (preprocessing) performs 877 iterations to deliver an energy $E_h = 35.3261$ and a metric defect $D_h = 0.0999797$. Algorithm 2 only stretches the plate which remains flat; see Figure 2.7 (left and middle). Algorithm 1 (gradient flow) then deforms the plate out of plane, and reaches a stationary state after 918 iterations with $E_h = 2.08544$, while keeping the metric defect $D_h = 0.087839$; see Figure 2.7-right.

We point out Corollary 2.5.1 also applies to Algorithm 1, i.e., a flat initial configuration ($y_3 = 0$) will theoretically lead to flat deformations throughout the gradient flow. However, in this example and the ones in Section 2.6.5, the initial deformation produced by Algorithm 2 has a non-vanishing third component y_3 (order of machine precision). Furthermore, Algorithm 2 may also produce discontinuous

configurations (as for the initial deformation in Figure 2.7 left and middle) to accommodate for the constraint and will thus have a relatively large energy due to the jump penalty term. These two aspects combined may be responsible for the main gradient flow Algorithm 1 to produce out of plane deformations even when starting with a theoretical flat initial configuration. This is the case when starting with a disc with positive Gaussian curvature metric as in Figure 2.7.

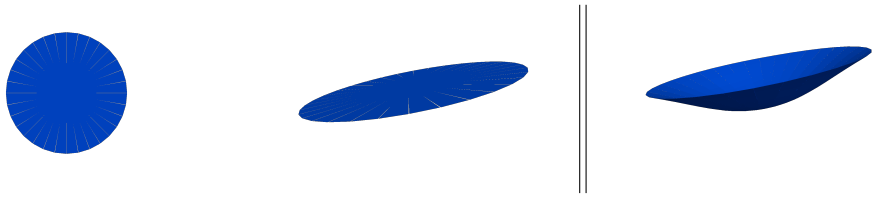

Figure 2.7: Deformed plate for the disc with positive Gaussian curvature metric. Algorithm 2 stretches the plate but keeps it flat (left and middle). Algorithm 1 gives rise to an ellipsoidal shape (right).

2.6.4.2 *Hyperbolic paraboloid - negative Gaussian curvature*

We consider the immersible metric g with negative Gaussian curvature

$$g(x_1, x_2) = \begin{bmatrix} 1 + x_2^2 & x_1 x_2 \\ x_1 x_2 & 1 + x_1^2 \end{bmatrix}. \tag{2.6.9}$$

A compatible deformation is given by $\mathbf{y}(x_1, x_2) = (x_1, x_2, x_1 x_2)^T$, i.e., $\mathrm{I}[\mathbf{y}] = g$.

In this setting, the flat configuration has a prestrain defect of $D_h = 1.56565$ (still vanishing energy). Algorithm 2 (preprocessing) performs 856 iterations to reach the energy $E_h = 50.3934$ and metric defect $D_h = 0.0999757$. Algorithm 1 (gradient flow) executes 1133 iterations to deliver an energy $E_h = 1.83112$ and met-

ric defect $D_h = 0.0980273$. Again, the metric defect remains basically constant throughout the main gradient flow, while the energy is significantly decreased. Figure 2.8 shows the initial (left) and final (middle) deformations of Algorithm 2 and the output of Algorithm 1 (right) which exhibit a saddle point structure.

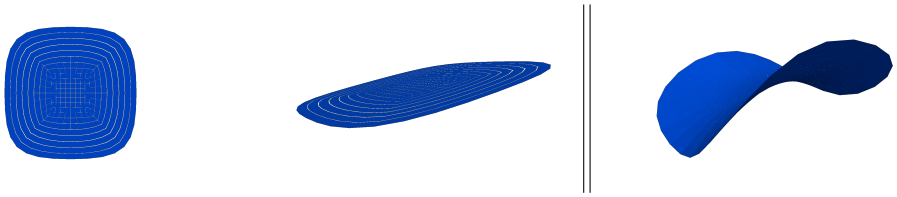

Figure 2.8: Deformed plate for the disc with negative Gaussian curvature. Algorithm 2 stretches the plate but keeps it flat (left and middle). Algorithm 1 gives rise to a saddle shape (right). Compare with Figure 2.7.

We point out that Algorithm 2 gives rise to little gaps between elements of the deformed subdivisions as a consequence of not including jump stabilization terms in the bilinear form (2.5.9). These gaps are reduced by Algorithm 1.

2.6.4.3 *Oscillating boundary*

We construct an immersible metric in polar coordinates (r, θ) with a six-fold oscillation near the boundary of the disc Ω. Let $\widetilde{g}(r,\theta) = \mathrm{I}[\widetilde{\mathbf{y}}(r,\theta)]$ be the first fundamental form of the deformation

$$\widetilde{\mathbf{y}}(r,\theta) = \big(r\cos(\theta), r\sin(\theta), 0.2r^4\sin(6\theta)\big). \tag{2.6.10}$$

The expression of the prestrain metric $g = \mathrm{I}[\mathbf{y}]$ in Cartesian coordinates is then given by (2.1.24) and $\mathbf{y}(x_1, x_2) = \widetilde{\mathbf{y}}(r,\theta)$.

We set the parameters

$$\tau = 0.05, \quad \widetilde{\tau} = 0.05, \quad \widetilde{\varepsilon}_0 = 0.1, \quad \widetilde{tol} = 10^{-4}, \quad tol = 10^{-6},$$

and note that Algorithm 1 (gradient flow) does not necessarily stop at global minima of the energy. Local extrema are frequently achieved and they are, in fact, of particular interest in many applications. To illustrate this property, we consider a couple of initial deformations and run Algorithms 2 and 1.

Case 1: boundary oscillation. We choose $\widetilde{\mathbf{y}}_h^0$ to be the local nodal interpolation of $\mathbf{y} = \widetilde{\mathbf{y}} \circ \boldsymbol{\psi}$ into $[\mathbb{V}_h^k]^3$, with $\widetilde{\mathbf{y}}$ given by (2.6.10). The output deformations of Algorithms 2 and 1 are depicted in Figure 2.9. The former becomes the initial configuration $\mathbf{y}_h^0$ of Algorithm 1 and is almost the same as $\widetilde{\mathbf{y}}_h^0$, which is approximately a disc with six-fold oscillations; see Figure 2.9 (left). This is due to the fact that $\mathrm{I}[\widetilde{\mathbf{y}}_h^0]$ is already close to the target metric g. Algorithm 1 (gradient flow) breaks the symmetry: two peaks are amplified while the other four are reduced. After the preprocessing, the energy is $E_h = 18.0461$ and metric defect is $D_h = 0.00208473$. The final energy is $E_h = 13.6475$ while the final metric defect is $D_h = 0.00528294$.

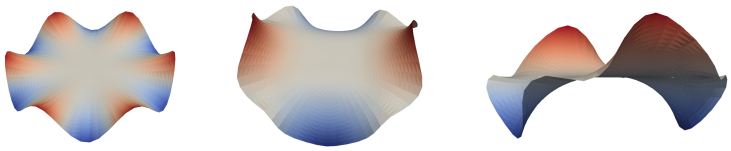

Figure 2.9: Deformed plate for the disc with oscillation boundary using the initial deformation described in **Case 1**. Left: output of Algorithm 2 (preprocessing); Middle: output of Algorithm 1 (gradient flow); Right: another view of output of Algorithm 1.

Case 2: no boundary oscillation. We run Algorithm 2 with the bi-Laplacian problem (2.5.1) with fictitious force $\widehat{\mathbf{f}} = (0, 0, 1)^T$ and boundary condition $\boldsymbol{\varphi}(\mathbf{x}) = (\mathbf{x}, 0)$ on $\partial\Omega$ (but without Φ). The output of Algorithm 2 is an ellipsoid without oscillatory boundary as in Case 1.

This corresponds to an underlying metric rather different from the target g. Algorithm 1 (gradient flow) is unable to improve on the metric defect because it is designed to decrease the bending energy. Therefore, the output of Algorithm 1 is again an ellipsoidal surface totally different from that of Case 1 that is displayed in Figure 2.9. In this case, $D_h = 0.801464$ and $E_h = 0.0377544$ leading to a smaller bending energy but larger metric defect when compared with Case 1.

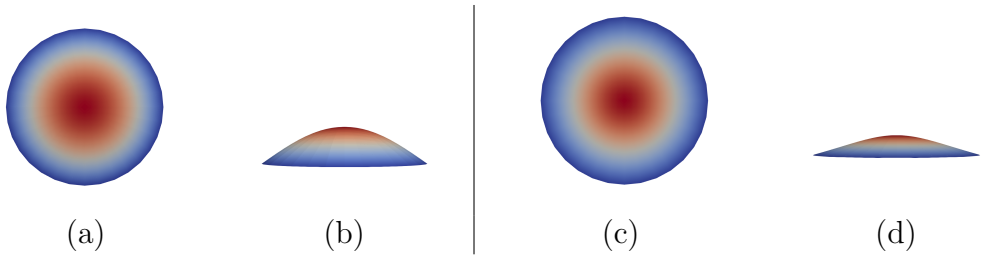

Figure 2.10: Ellipsoidal-like deformation of a disc without boundary oscillation when using the initial deformation described in **Case 2**. (a)-(b): output of Algorithm 2 (preprocessing) with maximal third component y_3 of the deformation about 7.8×10^{-2}; (c)-(d): output of Algorithm 1 (gradient flow) with maximal $y_3 \approx 4.4 \times 10^{-2}$. (a) and (c) are views from the top while (b) and (d) are views from the side where the third component of the deformation is scaled by a factor 10.

2.6.5 Gel discs

Discs made of a NIPA gel with various monomer concentrations can be manufactured in laboratories [52, 71]. NIPA gels undergo a differential shrinking in warm environments depending on the concentration. Monomer concentrations injected at

the center of the disc generate prestrain metrics depending solely on the distance to the center. We thus propose, inspired by [71, Section 4.2], prestrained metrics $\widetilde{g}(r,\theta)$ in polar coordinates of the form (2.1.25) with

$$\eta(r) = \begin{cases} \frac{1}{\sqrt{K}}\sin(\sqrt{K}r) & K > 0, \\ \frac{1}{\sqrt{-K}}\sinh(\sqrt{-K}r) & K < 0. \end{cases} \tag{2.6.11}$$

In view of Section 2.1.3, these metrics are immersible, namely there exist compatible deformations $\mathbf{y}$ such that $\mathrm{I}[\mathbf{y}] = g$ (isometric immersions). We now construct computationally isometric embeddings $\mathbf{y}$ for both $K > 0$ (elliptic) and $K < 0$ (hyperbolic). It turns out that they possess a constant Gaussian curvature $\kappa = K$ according to (2.1.28).

We let the domain Ω be the unit disc centered at the origin, do not enforce any boundary conditions and let $\mathbf{f} = \mathbf{0}$. The partition of Ω is as in Section 2.6.4 and

$$\widetilde{\tau} = 0.05, \quad \widetilde{\varepsilon}_0 = 0.1, \quad \widetilde{tol} = 10^{-4}, \quad tol = 10^{-6}.$$

Case $K = 2$ (elliptic): We use the fictitious force $\widehat{\mathbf{f}} = (0,0,1)^T$ in Algorithm 2 (preprocessing) and the pseudo-time step $\tau = 0.05$ in Algorithm 1 (gradient flow). We obtain a *spherical-like* final deformation; see Figure 2.11 and Table 2.8 for the results.

Case $K = -2$ (hyperbolic): We experiment with two different initial deformations for the metric preprocessing of Algorithm 2: (i) we take the identity map or (ii)

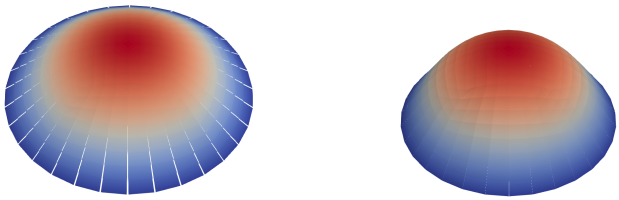

Figure 2.11: Deformed plate for the disc with constant Gaussian curvature $K = 2$ (elliptic). Outputs of Algorithm 2 (left) and Algorithm 1 (right).

	Algorithm 2	Algorithm 1
E_h	156.404	9.35368
D_h	0.0999494	0.188454

Table 2.8: Energy and prestrain defect for disc with constant curvature $K = 2$ (elliptic).

we solve the bi-Laplacian problem (2.5.1) with a fictitious force $\widehat{\mathbf{f}} = (0, 0, 1)^T$ and boundary condition $\boldsymbol{\varphi}(\mathbf{x}) = (\mathbf{x}, 0)$ on $\partial\Omega$ (but without Φ). Algorithm 2 produces *saddle-like* surfaces in both cases but with a different number of waves; see Figure 2.12. Algorithm 1 uses the pseudo-time steps $\tau = 0.00625$ and $\tau = 0.0125$ for (i) and (ii), respectively, while the other parameters remain unchanged. Table 2.9 documents the results.

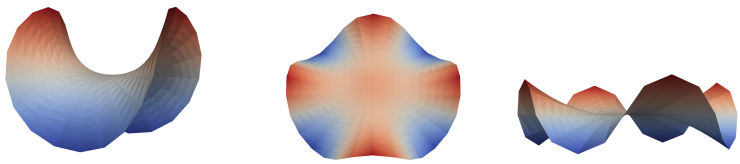

Figure 2.12: Deformed plate for the disc with constant Gaussian curvature $K = -2$ (hyperbolic). Outputs of Algorithm 1 with initialization (i) (left) and initialization (ii) (middle) and another view with initialization (ii) (right).

It is worth mentioning that for the 3d slender model described in [71], it is

	Initialization (i)		Initialization (ii)	
	Algorithm 2	Algorithm 1	Algorithm 2	Algorithm 1
E_h	699.396	6.92318	699.399	12.0978
D_h	0.0998791	0.245552	0.0999183	0.232627

Table 2.9: Energy and metric defect for disc with constant Gaussian curvature $K = -2$ (hyperbolic) for two different initial deformations of Algorithm 2: (i) identity map and (ii) solution to bi-Laplacian with fictitious force.

shown that when $K < 0$, the thickness s of the disc influences the number of waves of the minimizing deformation for $K < 0$. Our reduced model is asymptotic as $s \to 0$ whence it cannot match this feature. However, it reproduces a variety of deformations upon starting Algorithm 2 with suitable initial configurations.

Chapter 3: LDG Method of Large Deformations of Bilayer Plates

In this chapter, we consider a local discontinuous Galerkin (LDG) type numerical method for the approximation of large deformations of bilayer plates. With this new discretization, we prove the Γ-convergence and design a fully practical gradient flow scheme. We also prove the energy stability and the control of constraint defect for this scheme. Moreover, we also illustrate the efficiency and effectiveness of the method by numerical simulations. The key novel ingredients of analysis are an a priori L^∞-bound for the first derivatives of the approximated deformation, a reduced discrete Hessian, and an imposition of linearized discrete isometry constraint at barycenters of elements.

3.1 Problem statement and discretization

3.1.1 Bilayer plates model

Mathematically, to find the 2d equilibrium deformations $\mathbf{y} : \Omega \to \mathbb{R}^3$ of bilayer plates, one needs to solve the constrained minimization problem:

$$\min_{\mathbf{y}\in\mathbb{A}} E[\mathbf{y}] := \frac{1}{2}\int_\Omega \left|\mathrm{I\!I}[\mathbf{y}] - Z\right|^2, \tag{3.1.1}$$

where the admissible set is defined as

$$\mathbb{A} := \left\{ \mathbf{y} \in [H^2(\Omega)]^3 : \mathrm{I}[\mathbf{y}] = I_2 \text{ in } \Omega, \mathbf{y} = \boldsymbol{\varphi} \ \nabla \mathbf{y} = \Phi \text{ on } \Gamma_D \right\}. \tag{3.1.2}$$

Here, $\mathrm{I}[\mathbf{y}] := \nabla \mathbf{y}^T \nabla \mathbf{y}$ is the first fundamental form, and $\mathrm{II}[\mathbf{y}] := -\nabla \boldsymbol{\nu}^T \nabla \mathbf{y} = \boldsymbol{\nu}^T D^2 \mathbf{y}$ is the second fundamental form of the mid-surface $\mathbf{y}(\Omega)$, where $\boldsymbol{\nu}$ is its unit normal vector, and I_2 denotes the 2×2 identity matrix. Dirichlet boundary conditions are imposed on $\Gamma_D \subset \partial\Omega$ with boundary data $\boldsymbol{\varphi}$ and Φ can be extended to $[H^1(\Omega)]^3$ and $[H^1(\Omega)]^{3\times 2}$ respectively. To have boundary conditions compatible with the isometry constraint, Φ is also assumed to satisfy $\Phi^T \Phi = I_2$ a.e. in Ω. Moreover, Γ_D is allowed to be empty in this work and it corresponds to *free boundary* case.

$Z : \Omega \to \mathbb{R}^{2\times 2}$ is *spontaneous curvature* and encodes the material properties of the bilayer plates, i.e, the difference in reactions to the environmental stimuli between two thin layers. It drives plates to undergo large deformation without external forcing term. If the material is homogenous and isotropic, $Z = \alpha I_2$ with a constant α. Otherwise, the material is anisotropic or inhomogeneous. Z is the given data to the problem. In particular, when $Z = \mathbf{0}$ there is no difference between properties of two layers, and thus in this case the model reduces to *single layer plates* [22], which coincides with the classical Kirchhoff plate theory.

Energy functional $E[\mathbf{y}]$ can be further simplified. As $\mathbf{y}$ is an isometry (i.e. $\mathrm{I}[\mathbf{y}] = I_2$), there holds [11]

$$|\mathrm{II}[\mathbf{y}]|^2 = |D^2 \mathbf{y}|^2 = |\Delta \mathbf{y}|^2 = \mathrm{tr}(\mathrm{II}[\mathbf{y}])^2. \tag{3.1.3}$$

As a result, expanding the energy and using (3.1.3),

$$E\left[\mathbf{y}\right] := \frac{1}{2}\int_{\Omega}\left|D^2\mathbf{y}\right|^2 - \int_{\Omega}\mathrm{II}[\mathbf{y}] : Z + \frac{1}{2}\int_{\Omega}\left|Z\right|^2. \tag{3.1.4}$$

The term $\int_{\Omega}\mathrm{II}[\mathbf{y}] : Z$ is the challenging term, as it brings nonlinearity to variational derivative of $E\left[\mathbf{y}\right]$. Indeed, exploiting the definition of $\mathrm{II}[\mathbf{y}]$ and constraint $\mathrm{I}[\mathbf{y}] = I_2$, we can write

$$\boldsymbol{\nu} = \frac{\partial_1\mathbf{y}\times\partial_2\mathbf{y}}{|\partial_1\mathbf{y}\times\partial_2\mathbf{y}|} = \frac{\partial_1\mathbf{y}}{|\partial_1\mathbf{y}|}\times\frac{\partial_2\mathbf{y}}{|\partial_2\mathbf{y}|} = \partial_1\mathbf{y}\times\partial_2\mathbf{y} \tag{3.1.5}$$

and

$$\int_{\Omega}\mathrm{II}[\mathbf{y}] : Z = \sum_{i,j=1}^{2}\int_{\Omega}\partial_{ij}\mathbf{y}\cdot(\partial_1\mathbf{y}\times\partial_2\mathbf{y})Z_{ij}, \tag{3.1.6}$$

and therefore this term is cubic in $\mathbf{y}$. Moreover, since the term $\frac{1}{2}\int_{\Omega}\left|Z\right|^2$ depends only on Z, it is equivalent to minimize the following energy for $\mathbf{y}\in\mathbb{A}$:

$$E\left[\mathbf{y}\right] := \frac{1}{2}\int_{\Omega}\left|D^2\mathbf{y}\right|^2 - \sum_{i,j=1}^{2}\int_{\Omega}\partial_{ij}\mathbf{y}\cdot(\partial_1\mathbf{y}\times\partial_2\mathbf{y})Z_{ij}, \tag{3.1.7}$$

and we keep the same notation for simplicity. We further define

$$\widetilde{E}[\mathbf{y}] := \frac{1}{2}\int_{\Omega}\left|D^2\mathbf{y}\right|^2, \tag{3.1.8}$$

which is the bending energy functional for single layer plates as in [22]. This is consistent with the bending energy for prestrained plates (2.1.33) up to multiplicative constants when the target metric for prestrained plates is $g = I_2$ (no prestrain) and there is no external force ($\mathbf{f} = \mathbf{0}$).

3.1.2 Discretization

In this chapter, we adopt notations of meshes, jumps, derivatives, finite element spaces, norms, lifting operators, and the discrete Hessian from Chapter 2. For a detailed introduction, we refer to Section 2.2.1 and Section 2.4.

We first recall the definition of the discrete Hessian $H_h : \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) \to [L^2(\Omega)]^{3\times 2\times 2}$ as

$$H_h[\mathbf{v}_h] := D_h^2 \mathbf{v}_h - R_h^{l_1}([\nabla_h \mathbf{v}_h]) + B_h^{l_2}([\mathbf{v}_h]). \tag{3.1.9}$$

The weak and strong convergence properties satisfied by H_h are discussed in Chapter 2, while in this chapter we restate the strong convergence property in a more general set-up as follows. Note that in Lemma 2.3.3 we prove the strong convergence of $H_h[\mathbf{v}_h]$ for $\mathbf{v}_h$ as the Lagrange interpolation of $\mathbf{v}$, which satisfies the conditions in Lemma 3.1.1 and thus a special case of this general set-up. Moreover, the proof of Lemma 3.1.1 is verbatim the same as Lemma 2.3.3.

Lemma 3.1.1 (Strong convergence of H_h)**.** *Let $\mathbf{v} \in [H^2(\Omega)]^3$ be any function such that, when $\Gamma_D \neq \emptyset$, $\mathbf{v} = \boldsymbol{\varphi}$ and $\nabla \mathbf{v} = \Phi$ on Γ_D. Moreover, let $\mathbf{v}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$ be such that*

$$\|D^2 \mathbf{v}_h\|_{L^2(T)} \lesssim |\mathbf{v}|_{H^2(T)} \text{ for any } T \in \mathcal{T}_h \text{ and } \sum_{T\in\mathcal{T}_h} |\mathbf{v}_h - \mathbf{v}|^2_{H^2(T)} \to 0 \text{ as } h \to 0. \tag{3.1.10}$$

Then for any $l_1, l_2 \geq 0$ we have as $h \to 0$

$$H_h[\mathbf{v}_h] \to D^2 \mathbf{v} \quad \text{strongly in } [L^2(\Omega)]^{3\times 2\times 2}. \tag{3.1.11}$$

Discrete minimization problem. Consider a discrete energy approximating $E[\mathbf{y}]$ using a LDG-type discretization:

$$E_h[\mathbf{y}_h] = \widetilde{E}_h[\mathbf{y}_h] - \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \big[(\widetilde{H_h}[\mathbf{y}_h])_{ij} \cdot (\partial_1 \mathbf{y}_h \times \partial_2 \mathbf{y}_h) Z_{ij}\big](x_T). \tag{3.1.12}$$

Here, $\widetilde{E}_h[\mathbf{y}_h]$ is defined as

$$\widetilde{E}_h[\mathbf{y}_h] = \frac{1}{2}\int_\Omega \big|H_h[\mathbf{y}_h]\big|^2 + \gamma_1 \|\mathrm{h}^{-\frac{1}{2}}[\nabla_h \mathbf{y}_h]\|^2_{L^2(\Gamma_h^a)} + \gamma_0 \|\mathrm{h}^{-\frac{3}{2}}[\mathbf{y}_h]\|^2_{L^2(\Gamma_h^a)}, \tag{3.1.13}$$

where the last two terms are stabilization terms with penalty parameters $\gamma_0, \gamma_1 > 0$. Moreover, (3.1.13) is the discrete bending energy for the single layer problem with the LDG method. Indeed, it is a natural discretization of (3.1.8) that hinges on the discrete Hessian in the same spirit of Chapter 2. $\widetilde{E}_h[\mathbf{y}_h]$ is the quadratic part (quadratic in $\mathbf{y}_h$) of (3.1.12) approximating (3.1.8), while the second term is the cubic part approximating (3.1.6). When considering the variational derivative of (3.1.12), the quadratic part $\widetilde{E}_h[\mathbf{y}_h]$ will only result in linear terms, while the cubic term brings additional nonlinearity and requires more effort to deal with.

For the second term of (3.1.12), x_T denotes the barycenter of any element $T \in \mathcal{T}_h$, and $\widetilde{H_h}[\mathbf{y}_h]$ denotes a *reduced discrete Hessian*, whose definition and properties are discussed later.

We consider the discrete admissible set $\mathbb{A}_{h,\delta}$ defined as

$$\mathbb{A}_{h,\delta} := \{\mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi) : \quad \Big|\big[\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2\big](x_T)\Big| \le \delta \quad \forall T \in \mathcal{T}_h\}, \tag{3.1.14}$$

and a discrete counterpart of (3.1.1) reads

$$\min_{\mathbf{y}_h \in \mathbb{A}_{h,\delta}} E_h[\mathbf{y}_h]. \tag{3.1.15}$$

For $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$, there holds that $\max_{T\in\mathcal{T}_h} \left|\left[\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2\right](x_T)\right| \leq \delta$, and this implies that a discrete version of L^∞ norm of $\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2$ is controlled by δ; this is a relaxed isometry constraint for the discrete function $\mathbf{y}_h$.

Lemma 3.1.2 (pointwise isometry constraint)**.** *If* $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$*, then*

$$1-\delta \leq |\partial_i \mathbf{y}_h(x_T)|^2 \leq 1+\delta \text{ for } i=1,2, \text{ and } |\partial_1 \mathbf{y}(x_T)\cdot \partial_2 \mathbf{y}(x_T)| \leq \delta, \tag{3.1.16}$$

for all $T \in \mathcal{T}_h$.

Proof. Note that for any $i,j=1,2$

$$\left[\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h\right]_{ij}(x_T) = \partial_i \mathbf{y}(x_T)\cdot \partial_j \mathbf{y}(x_T).$$

By the definition (3.1.14), we deduce that for any $i,j=1,2$

$$\left|\left[\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h\right]_{ij}(x_T) - I_{2,ij}(x_T)\right| \leq \delta.$$

Since $I_{2,ij}=1$ when $i=j$ and $I_{2,ij}=0$ when $i\neq j$, we conclude (3.1.16). □

Reduced discrete Hessian. The *reduced discrete Hessian* $\widetilde{H_h}[\mathbf{y}_h]$ is defined as

follows:

$$\widetilde{H_h}[\mathbf{y}_h] := P_h(H_h[\mathbf{y}_h]), \tag{3.1.17}$$

where P_h is the local L^2 projection onto $[\mathbb{P}_0(T)]^{3\times 2\times 2}$ for each element $T \in \mathcal{T}_h$. Indeed, $P_h(H_h[\mathbf{y}_h])|_T$ is defined as

$$P_h(H_h[\mathbf{y}_h])|_T = \frac{1}{|T|}\int_T H_h[\mathbf{y}_h]. \tag{3.1.18}$$

The second term of (3.1.12) is an approximation (one-point quadrature rule) of the integral $\int_\Omega (\widetilde{H_h}[\mathbf{y}_h])_{ij}\cdot(\partial_1\mathbf{y}_h\times\partial_2\mathbf{y}_h)Z_{ij}$. At first glance, using $\widetilde{H_h}[\mathbf{y}_h] = H_h[\mathbf{y}_h]$ would be the most natural choice for this term in the discrete energy, but it hinders the Γ-convergence proof; the use of $\widetilde{H_h}[\mathbf{y}_h]$ will be justified later in Section 3.2. Note that here $\partial_i\mathbf{y}_h$ $(i = 1, 2)$ denotes columns of the broken gradient $\nabla_h\mathbf{y}_h$.

Next we explore properties of the reduced discrete Hessian $\widetilde{H_h}[\mathbf{y}_h]$.

Lemma 3.1.3 (Upper bound of reduced discrete Hessian)**.** *For any* $\mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$, *there holds*

$$\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(\Omega)} \le c_{stab}\|\mathbf{y}_h\|_{H_h^2}, \tag{3.1.19}$$

where the constant c_{stab} *is independent of* h.

Proof. Due to Lemma 2.2.1, it suffices to prove $\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(\Omega)} \lesssim \|H_h[\mathbf{y}_h]\|_{L^2(\Omega)}$. This is a consequence of definition (3.1.18), because we have the following on each element $T \in \mathcal{T}_h$:

$$\|P_h(H_h[\mathbf{y}_h])\|_{L^2(T)} \le \|H_h[\mathbf{y}_h]\|_{L^2(T)}.$$

□

Similar to Lemma 2.3.2, we have the following result.

Lemma 3.1.4 (Weak convergence for reduced discrete Hessian)**.** *Let* $\mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$. *If* $\|\mathbf{y}_h\|_{H_h^2(\Omega)} \leq C$ *for all* h *and* $\mathbf{y}_h \to \mathbf{y}$ *in* $[L^2(\Omega)]^3$ *for* $\mathbf{y} \in [H^2(\Omega)]^3$, *we have* $\widetilde{H_h}[\mathbf{y}_h]$ *converges weakly to* $D^2\mathbf{y}$ *in* $[L^2(\Omega)]^3$.

Proof. For any $\phi \in [C_c^\infty(\Omega)]^{3\times 2\times 2}$, we have

$$\int_\Omega P_h(H_h[\mathbf{y}_h]) : \phi = \sum_{T\in\mathcal{T}_h} \int_T H_h[\mathbf{y}_h] : P_h\phi = \sum_{T\in\mathcal{T}_h} \int_T H_h[\mathbf{y}_h] : \phi + H_h[\mathbf{y}_h] : (P_h\phi - \phi).$$

For the first term, by Lemma 2.3.2 we have

$$\int_\Omega H_h[\mathbf{y}_h] : \phi \to \int_\Omega D^2\mathbf{y} : \phi.$$

For the second term we can estimate

$$\begin{aligned}
\Big| \sum_{T\in\mathcal{T}_h} H_h[\mathbf{y}_h] : (P_h\phi - \phi) \Big| &\leq \sum_{T\in\mathcal{T}_h} \|H_h[\mathbf{y}_h]\|_{L^2(T)} \|P_h\phi - \phi\|_{L^2(T)} \\
&\leq \sum_{T\in\mathcal{T}_h} \|H_h[\mathbf{y}_h]\|_{L^2(T)} h_T \|\nabla\phi\|_{L^2(T)} \\
&\leq h \|H_h[\mathbf{y}_h]\|_{L^2(\Omega)} \|\nabla\phi\|_{L^2(\Omega)}.
\end{aligned}$$

Due to Lemma 2.2.1 we have $\|H_h[\mathbf{y}_h]\|_{L^2(\Omega)} \lesssim \|\mathbf{y}_h\|_{H_h^2}$ is uniformly bounded, then the second term converges to 0. □

Remark 3.1.1. *If* P_h *is defined to be the local* L^2*-projection onto* $[\mathbb{P}_k(T)]^{3\times 2\times 2}$ *for*

any nonnegative integer k, Lemma 3.1.4 is still correct. Indeed, P_h is defined as for any $T \in \mathcal{T}_h$

$$(P_h(\mathbf{v}), \mathbf{w}_h)_{L^2(T)} = (\mathbf{v}, \mathbf{w}_h)_{L^2(T)}, \ \forall \mathbf{w}_h \in [\mathbb{P}_k(T)]^{3\times 2\times 2}. \tag{3.1.20}$$

Then,

$$\begin{aligned}
\int_\Omega [P_h(H_h[\mathbf{y}_h]) - H_h[\mathbf{y}_h]] : \phi &= \sum_{T\in\mathcal{T}_h} \int_T [P_h(H_h[\mathbf{y}_h]) - H_h[\mathbf{y}_h]] : (\phi - P_h\phi) \\
&+ \sum_{T\in\mathcal{T}_h} \int_T [P_h(H_h[\mathbf{y}_h]) - H_h[\mathbf{y}_h]] : P_h\phi \\
&= \sum_{T\in\mathcal{T}_h} \int_T [P_h(H_h[\mathbf{y}_h]) - H_h[\mathbf{y}_h]] : (\phi - P_h\phi) \\
&\lesssim \sum_{T\in\mathcal{T}_h} \|H_h[\mathbf{y}_h]\|_{L^2(T)} \|\phi - P_h\phi\|_{L^2(T)} \\
&\lesssim \|H_h[\mathbf{y}_h]\|_{L^2(\Omega)} \|\phi - P_h\phi\|_{L^2(\Omega)},
\end{aligned}$$

where the second equality comes from the definition (3.1.20) and the inequality comes from the L^2-stability of the operator P_h. Moreover, since $\|H_h[\mathbf{y}_h]\|_{L^2(\Omega)}$ is uniformly bounded and $\|\phi - P_h\phi\|_{L^2(\Omega)} \to 0$ as $h \to 0$, together with the weak convergence of $H_h[\mathbf{y}_h]$ to $D^2\mathbf{y}$, we can prove Lemma 3.1.4 for P_h defined as (3.1.20).

However, as we shall see in Section 3.2 we need to require $P_h(H_h[\mathbf{y}_h]) \in [\mathbb{P}_1(T)]^{3\times 2\times 2}$ to prove the Γ-convergence, and thus k in (3.1.20) can only be 0 and 1 here. Moreover, since we define the cubic term in (3.1.12) by using a one-point quadrature rule, using $k = 0$ for P_h helps with the efficiency and simplicity, and does not harm the accuracy.

3.2 Γ-convergence

First we recall coercivity, compactness and lim-inf property (Theorem 3.2.1 and Theorem 3.2.2) of the quadratic part $\widetilde{E}_h[\mathbf{y}_h]$. Note that such results with LDG discretization for prestrained plates are proved in Chapter 2 in the case of both Dirichlet boundary and free boundary. Since single layer plates are only special cases of prestrained plates (i.e, $g = I_2$), the proofs in Chapter 2 carry over to $\widetilde{E}_h[\mathbf{y}_h]$. Here, for simplicity of presentation we focus on the case $\Gamma_D \neq \emptyset$, while the case $\Gamma_D = \emptyset$ can be treated as in Chapter 2. Note that $H_h[\mathbf{y}_h]$ in $\widetilde{E}_h[\mathbf{y}_h]$ can have any choice of lifting polynomial degrees $l_1, l_2 \geq 0$.

Theorem 3.2.1 (Coercivity of $\widetilde{E}_h$)**.** *Let $\mathbf{y}_h \in \mathbb{V}_h^k(\boldsymbol{\varphi}, \Phi)$, and $\gamma_0, \gamma_1 > 0$. If $\Gamma_D \neq \emptyset$, then*

$$\|\mathbf{y}_h\|^2_{H^2_h(\Omega)} \leq c_{coer}(\widetilde{E}_h[\mathbf{y}_h] + C^2_{\boldsymbol{\varphi},\Phi}), \tag{3.2.1}$$

where the constant c_{coer} of (3.2.1) depends only on γ_0, γ_1, and the constant $C_{\boldsymbol{\varphi},\Phi}$ is given by $C_{\boldsymbol{\varphi},\Phi} := (\|\boldsymbol{\varphi}\|^2_{H^1(\Omega)} + \|\Phi\|^2_{H^1(\Omega)})^{\frac{1}{2}}$.

Theorem 3.2.2 (Compactness and Lim-inf of $\widetilde{E}_h$)**.** *Assume that $\Gamma_D \neq \emptyset$. Let $l_1, l_2 \geq 0$ and let $\delta = \delta(h) \to 0$ as $h \to 0$. Let $\{\mathbf{y}_h\} \subset \mathbb{A}_{h,\delta}$ be a sequence such that $\widetilde{E}_h[\mathbf{y}_h] \leq C$ uniformly. Then there exists $\mathbf{y} \in \mathbb{A}$ such that (up to a subsequence) $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$ as $h \to 0$, and $\widetilde{E}[\mathbf{y}] \leq \liminf_{h\to 0} \widetilde{E}_h[\mathbf{y}_h]$.*

Moreover, the following Lemma reveals the conditions that a recovery sequence $\{\mathbf{y}_h\}_h$ should satisfy, for the lim-sup condition of $\widetilde{E}_h$. A proof can be carried out as in Theorem 2.3.4.

Lemma 3.2.1 (Conditions of recovery sequence for $\widetilde{E}_h$)**.** *Assume that* $\Gamma_D \neq \emptyset$*. For any* $\mathbf{y} \in \mathbb{A}$*, if there exists* $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ *such that* $H_h[\mathbf{y}_h]$ *converges to* $D^2\mathbf{y}$ *strongly in* $[L^2(\Omega)]^{3\times 2\times 2}$*, then* $\widetilde{E}[\mathbf{y}] = \lim_{h\to 0} \widetilde{E}_h[\mathbf{y}_h]$.

Indeed, to prove lim-sup condition one needs to construct a recovery sequence $\{\mathbf{y}_h\}_h \subset \mathbb{A}_{h,\delta}$ for any $\mathbf{y} \in \mathbb{A}$, which should satisfy the conditions in Lemma 3.1.1 to guarantee the strong convergence of the discrete Hessian. For example, such a specific construction exploiting Lagrange interpolation are used in Theorem 2.3.4, but we emphasize that a novel construction of recovery sequence is considered in this work as in Theorem 3.2.4, corresponding to the brand new discrete constraint (3.1.14).

Then, we turn to properties of the complete discrete energy $E_h[\mathbf{y}_h]$. With the help of Theorem 3.2.1, we can further establish the equicoercivity of energy E_h:

Theorem 3.2.3 (Coercivity of E_h)**.** *Assume that* $\Gamma_D \neq \emptyset$*. If* $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ *and* $E_h[\mathbf{y}_h]$ *is uniformly bounded, then* $\widetilde{E}_h[\mathbf{y}_h]$ *is also uniformly bounded, and thus* $\|\mathbf{y}_h\|_{H_h^2}$ *is uniformly bounded as well.*

Proof. Note that

$$\begin{aligned}
&\sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1\mathbf{y}_h \times \partial_2\mathbf{y}_h) Z_{ij}\big](x_T) \\
&\leq \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \|\widetilde{H_h}[\mathbf{y}_h]_{ij}\|_{L^\infty(T)} |\partial_1\mathbf{y}_h(x_T)| |\partial_2\mathbf{y}_h(x_T)| \|Z_{ij}\|_{L^\infty(T)}.
\end{aligned}$$

Using $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ and Lemma 3.1.2, we get that $|\partial_i\mathbf{y}_h(x_T)| \leq (\delta+1)^{\frac{1}{2}}$ for $i = 1, 2$.

Thus,

$$
\begin{aligned}
&\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\big[\widetilde{H_h}[\mathbf{y}_h]_{ij}\cdot(\partial_1\mathbf{y}_h\times\partial_2\mathbf{y}_h)Z_{ij}\big](x_T)\\
&\leq(\delta+1)\|Z_{ij}\|_{L^\infty(\Omega)}\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\|\widetilde{H_h}[\mathbf{y}_h]_{ij}\|_{L^\infty(T)}\\
&\leq(\delta+1)\|Z_{ij}\|_{L^\infty(\Omega)}c_{inv}\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|^{\frac{1}{2}}\|\widetilde{H_h}[\mathbf{y}_h]_{ij}\|_{L^2(T)}\\
&\leq(\delta+1)\|Z_{ij}\|_{L^\infty(\Omega)}c_{inv}|\Omega|^{\frac{1}{2}}\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(\Omega)}\\
&\leq(\delta+1)\|Z_{ij}\|_{L^\infty(\Omega)}c_{inv}c_{stab}|\Omega|^{\frac{1}{2}}\|\mathbf{y}_h\|_{H_h^2}\\
&\leq(2c_{coer})^{-1}\|\mathbf{y}_h\|_{H_h^2}^2+\frac{1}{2}c_{coer}(\delta+1)^2\|Z_{ij}\|_{L^\infty(\Omega)}^2c_{inv}^2c_{stab}^2|\Omega|\\
&\leq\frac{1}{2}\widetilde{E}_h[\mathbf{y}_h]+\frac{1}{2}C_{\boldsymbol{\varphi},\Phi}^2+\frac{1}{2}c_{coer}(\delta+1)^2\|Z_{ij}\|_{L^\infty(\Omega)}^2c_{inv}^2c_{stab}^2|\Omega|,
\end{aligned}
$$

where we use Young's inequality, inverse inequality, Lemma 3.1.3 and Theorem 3.2.1.

We have proved that

$$
2E_h[\mathbf{y}_h]\geq\widetilde{E}_h[\mathbf{y}_h]-\widetilde{C}_{coer},\tag{3.2.2}
$$

where positive constant $\widetilde{C}_{coer}$ is defined as

$$
\widetilde{C}_{coer}:=\widetilde{C}_{coer}(\delta)=C_{\boldsymbol{\varphi},\Phi}^2+c_{coer}(\delta+1)^2\|Z_{ij}\|_{L^\infty(\Omega)}^2c_{inv}^2c_{stab}^2|\Omega|.\tag{3.2.3}
$$

Moreover, we have

$$
2c_{coer}E_h[\mathbf{y}_h]\geq\|\mathbf{y}_h\|_{H_h^2}^2-c_{coer}C_{\boldsymbol{\varphi},\Phi}^2-c_{coer}\widetilde{C}_{coer}:=\|\mathbf{y}_h\|_{H_h^2}^2-\hat{c}_{coer},\tag{3.2.4}
$$

where the positive constant $\hat{c}_{coer}$ is defined as

$$\hat{c}_{coer} := \hat{c}_{coer}(\delta) = 2c_{coer}C^2_{\boldsymbol{\varphi},\Phi} + c^2_{coer}(\delta+1)^2\|Z_{ij}\|^2_{L^\infty(\Omega)}c^2_{inv}c^2_{stab}|\Omega|. \tag{3.2.5}$$

Consequently, the uniform boundedness of $\|\mathbf{y}_h\|_{H^2_h}$ and $\widetilde{E}_h[\mathbf{y}_h]$ follows from the uniform boundedness of $E_h[\mathbf{y}_h]$. □

Now we prove the Γ-convergence of E_h:

Theorem 3.2.4 (Γ-convergence of E_h)**.** *Assume that $\Gamma_D \neq \emptyset$. Let $l_1, l_2 \geq 0$ and let $\delta = \delta(h) \to 0$ as $h \to 0$.*

(i) Let $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ be a sequence such that $E_h[\mathbf{y}_h]$ is uniformly bounded. Then there exists $\mathbf{y} \in \mathbb{A}$ such that $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$ and $E[\mathbf{y}] \leq \liminf_{h\to 0} E_h[\mathbf{y}_h]$.

(ii) For any $\mathbf{y} \in \mathbb{A}$ there exists $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ such that $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$ and $E[\mathbf{y}] \geq \limsup_{h\to 0} E_h[\mathbf{y}_h]$.

Proof. Note that the same results have been established for $\widetilde{E}_h$. So it suffices to deal with the term $\sum_{i,j=1}^2 \sum_{T\in\mathcal{T}_h} |T|\big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1\mathbf{y}_h \times \partial_2\mathbf{y}_h)Z_{ij}\big](x_T)$ and the novel way of imposing the discrete isometry constraint in $\mathbb{A}_{h,\delta}$.

Step (i): lim-inf property. By Theorem 3.2.3 and Theorem 3.2.2, we have already shown that there exists $\mathbf{y} \in [H^2(\Omega)]^3$ such that boundary conditions are satisfied and $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$ and $\widetilde{E}[\mathbf{y}] \leq \liminf_{h\to 0} \widetilde{E}_h[\mathbf{y}_h]$. Moreover, there further holds the strong convergence of $\nabla_h\mathbf{y}_h$ to $\nabla\mathbf{y}$ in $[L^2(\Omega)]^{3\times 2}$ as in Chapter 2. To check that $\mathbf{y}$ satisfies the isometry constraint as $h \to 0$, we take advantages of the

L^2 convergence of $\nabla_h \mathbf{y}_h$ to $\nabla \mathbf{y}$. Noting $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ we can verify that

$$\begin{aligned}
\sum_{T\in\mathcal{T}_h} \Big| \int_T (\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2) \Big| &\leq \sum_{T\in\mathcal{T}_h} |T| \big| \big[\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I\big](x_T) \big| \\
&\quad + \sum_{T\in\mathcal{T}_h} \int_T \big| \nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2 - [\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2](x_T) \big| \\
&\leq \sum_{T\in\mathcal{T}_h} |T|\delta + \sum_{T\in\mathcal{T}_h} h \int_T |\nabla(\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h)| \\
&\leq |\Omega|\delta + h \|D_h^2 \mathbf{y}_h\|_{L^2(\Omega)} \|\nabla_h \mathbf{y}_h\|_{L^2(\Omega)} \\
&\to 0,
\end{aligned}$$

as $h \to 0$, where we use the uniform boundedness of $\|\mathbf{y}_h\|_{H_h^2}$. Applying Poincaré-Friedrichs inequality we deduce that

$$\|\nabla_h \mathbf{y}_h^T \nabla_h \mathbf{y}_h - I_2\|_{L^1(\Omega)} \lesssim h \|D_h^2 \mathbf{y}_h\|_{L^2(\Omega)} \|\nabla_h \mathbf{y}_h\|_{L^2(\Omega)} + \sum_{T\in\mathcal{T}_h} \Big| \int_T (\nabla \mathbf{y}_h^T \nabla \mathbf{y}_h - I_2) \Big| \to 0,$$

as $h \to 0$. Moreover, considering

$$\big(\nabla_h \mathbf{y}_h^T \nabla_h \mathbf{y}_h - I_2\big) - \big(\nabla \mathbf{y}^T \nabla \mathbf{y} - I_2\big) = \nabla_h(\mathbf{y}_h - \mathbf{y})^T \nabla_h \mathbf{y}_h + \nabla \mathbf{y}^T \nabla_h(\mathbf{y}_h - \mathbf{y}), \quad (3.2.6)$$

we can further conclude that

$$\begin{aligned}
\|\nabla \mathbf{y}^T \nabla \mathbf{y} - I_2\|_{L^1(\Omega)} &\leq \big(\|\nabla \mathbf{y}\|_{L^2(\Omega)} + \|\nabla_h \mathbf{y}_h\|_{L^2(\Omega)}\big) \|\nabla_h \mathbf{y}_h - \nabla \mathbf{y}\|_{L^2(\Omega)} \\
&\quad + \|\nabla_h \mathbf{y}_h^T \nabla_h \mathbf{y}_h - I_2\|_{L^1(\Omega)} \to 0,
\end{aligned}$$

as $h \to 0$. This shows that $\mathbf{y}$ satisfies the isometry constraint, and thus $\mathbf{y} \in \mathbb{A}$.

Now to prove the lim-inf condition for $E_h[\mathbf{y}_h]$, it suffices to check the term tends to 0, where $\mathbf{y} \in \mathbb{A}$ is the limit of $\mathbf{y}_h$ that is just obtained.

$$
\begin{aligned}
&\sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1 \mathbf{y}_h \times \partial_2 \mathbf{y}_h) Z_{ij}\big](x_T) - \sum_{i,j=1}^{2} \int_\Omega \partial_{ij}\mathbf{y} \cdot (\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij} \\
&= \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} \int_T (\widetilde{H_h}[\mathbf{y}_h]_{ij} - \partial_{ij}\mathbf{y}) \cdot (\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij} \\
&+ \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1 \mathbf{y}_h \times \partial_2 \mathbf{y}_h - \partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij}\big](x_T) \\
&+ \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} \Big\{ |T| \big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij}\big](x_T) - \int_T \widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot (\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij} \Big\} \\
&=: R_1 + R_2 + R_3.
\end{aligned}
$$

First, due to the Lemma 3.1.4, we have $R_1 \to 0$ as $h \to 0$.

Then, the following inverse inequality for any function $\mathbf{w}_h$ defined on a finite dimensional space over $T \in \mathcal{T}_h$

$$
|T|^{\frac{1}{2}} |\mathbf{w}_h(x_T)| \le |T|^{\frac{1}{2}} \|\mathbf{w}_h\|_{L^\infty(T)} \le c_{inv} \|\mathbf{w}_h\|_{L^2(T)}, \tag{3.2.7}
$$

yields

$$
\begin{aligned}
|R_2| &= \Big| \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} |T| \big[\widetilde{H_h}[\mathbf{y}_h]_{ij} \cdot ((\partial_1 \mathbf{y}_h - \partial_1 \mathbf{y}) \times \partial_2 \mathbf{y}_h + \partial_1 \mathbf{y} \times (\partial_2 \mathbf{y}_h - \partial_2 \mathbf{y})) Z_{ij}\big](x_T) \Big| \\
&\le \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} \|\widetilde{H_h}[\mathbf{y}_h]_{ij}\|_{L^2(T)} \|Z_{ij}\|_{L^\infty(T)} \Big(\|\partial_1 \mathbf{y}_h - \partial_1 \mathbf{y}\|_{L^2(T)} |\partial_2 \mathbf{y}_h(x_T)| \\
&+ \|\partial_2 \mathbf{y}_h - \partial_2 \mathbf{y}\|_{L^2(T)} \|\partial_1 \mathbf{y}\|_{L^\infty(T)} \Big).
\end{aligned}
$$

Then, by Lemma 3.1.2 and that $\mathbf{y} \in \mathbb{A}$ we know $|\partial_2 \mathbf{y}_h(x_T)|$ and $\|\partial_1 \mathbf{y}\|_{L^\infty(T)}$ are uniformly bounded. Together with Lemma 3.1.3, the fact that $\|\mathbf{y}_h\|_{H_h^2}$ is uniformly bounded and the strong convergence of $\nabla_h \mathbf{y}_h$ to $\nabla \mathbf{y}$ in $[L^2(\Omega)]^{3\times 2}$, we have $R_2 \to 0$ as $h \to 0$.

Since R_3 is a quadrature error, we have

$$
\begin{aligned}
|R_3| &\lesssim \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} h^2 \int_T |D^2[\widetilde{H}_h[\mathbf{y}_h]_{ij} \cdot (\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij}]| \\
&\lesssim \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} h^2 \Big(\|D^2(\widetilde{H}_h[\mathbf{y}_h]_{ij})\|_{L^1(T)} \|(\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij}\|_{L^\infty(T)} \\
&+ \|\nabla(\widetilde{H}_h[\mathbf{y}_h]_{ij})\|_{L^2(T)} \|\nabla((\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij})\|_{L^2(T)} \\
&+ \|\widetilde{H}_h[\mathbf{y}_h]_{ij}\|_{L^\infty(T)} \|D^2((\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij})\|_{L^1(T)} \Big) := I_1 + I_2 + I_3.
\end{aligned}
$$

Note that

$$
\begin{aligned}
I_2 &= \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} h^2 \|\nabla(\widetilde{H}_h[\mathbf{y}_h]_{ij})\|_{L^2(T)} \|\nabla((\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}) Z_{ij})\|_{L^2(T)} \\
&\lesssim \sum_{T\in\mathcal{T}_h} h \|\widetilde{H}_h[\mathbf{y}_h]\|_{L^2(T)} \|D^2 \mathbf{y}\|_{L^2(T)} \|\nabla \mathbf{y}\|_{L^\infty(T)} \|Z\|_{L^\infty(T)} \\
&\lesssim h \|\widetilde{H}_h[\mathbf{y}_h]\|_{L^2(\Omega)} \|D^2 \mathbf{y}\|_{L^2(\Omega)} \lesssim h,
\end{aligned}
$$

as $\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(\Omega)} \lesssim \|\mathbf{y}_h\|_{H_h^2} \leq C$ in view of Lemma 3.1.3. So $I_2 \to 0$ as $h \to 0$. Also,

$$\begin{aligned} I_3 &= \sum_{i,j=1}^{2} \sum_{T\in\mathcal{T}_h} h^2 \|\widetilde{H_h}[\mathbf{y}_h]_{ij}\|_{L^\infty(T)} \|D^2((\partial_1\mathbf{y}\times\partial_2\mathbf{y})Z_{ij})\|_{L^1(T)} \\ &\lesssim C_Z \sum_{T\in\mathcal{T}_h} h\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(T)}(\|D^3\mathbf{y}\|_{L^2(T)}\|\nabla\mathbf{y}\|_{L^2(T)} + \|D^2\mathbf{y}\|_{L^2(T)}^2) \\ &\lesssim Ch\|\widetilde{H_h}[\mathbf{y}_h]\|_{L^2(\Omega)}\|\mathbf{y}\|_{H^3(\Omega)}^2, \end{aligned}$$

where we use $\sum_i a_i b_i c_i \leq (\sum_i a_i^2)^{\frac{1}{2}}(\sum_i b_i^2 c_i^2)^{\frac{1}{2}} \leq (\sum_i a_i^2)^{\frac{1}{2}}(\sum_i b_i^2 \sum_j c_j^2)^{\frac{1}{2}}$. Note that here we use a higher regularity of $\mathbf{y}$, which requires a regularization argument for $\mathbf{y} \in \mathbb{A}$. In fact, isometries $\mathbf{y}$ in $[H^2(\Omega)]^3$ can be approximated with arbitrary precision in the H^2-norm by smooth mollifier $\mathbf{y}^\epsilon$. Hence,

$$\begin{aligned} &\Big|\int_\Omega \partial_{ij}\mathbf{y}\cdot(\partial_1\mathbf{y}\times\partial_2\mathbf{y})Z_{ij} - \int_\Omega \partial_{ij}\mathbf{y}^\epsilon\cdot(\partial_1\mathbf{y}^\epsilon\times\partial_2\mathbf{y}^\epsilon)Z_{ij}\Big| \\ &\lesssim \|\mathbf{y}-\mathbf{y}^\epsilon\|_{H^2(\Omega)}\|\partial_1\mathbf{y}\|_{L^2(\Omega)}\|\partial_2\mathbf{y}\|_{L^\infty(\Omega)} \\ &+ \|\mathbf{y}^\epsilon\|_{H^2(\Omega)}\|\mathbf{y}-\mathbf{y}^\epsilon\|_{H^1(\Omega)}(\|\partial_2\mathbf{y}\|_{L^\infty(\Omega)} + \|\partial_1\mathbf{y}^\epsilon\|_{L^\infty(\Omega)}), \end{aligned}$$

which can be arbitrarily small when $\mathbf{y}^\epsilon$ is close to $\mathbf{y}$ arbitrarily in $[H^2(\Omega)]^3$. Upon choosing first $\mathbf{y}^\epsilon$ and next h, we can make I_3 arbitrarily small.

For I_1, since in each element $T \in \mathcal{T}_h$ we have

$$D^2(\widetilde{H_h}[\mathbf{y}_h]_{ij}) = D^2\big((P_h(H_h[\mathbf{y}_h]))_{ij}\big) = 0,$$

as $P_h(H_h[\mathbf{y}_h]) \in \mathbb{P}_0(T)$. Then we conclude $I_1 = 0$. This justifies the use of reduced

Hessian. Otherwise, if one considers $H_h[\mathbf{y}_h]_{ij}$, $D^2(H_h[\mathbf{y}_h]_{ij})$ is not necessarily 0 and also there is no proper way to control I_1 by h^α with $\alpha > 0$ (using inverse inequalities can only guarantee that I_1 is uniformly bounded).

To sum up, $R_3 \to 0$ as $h \to 0$, and finally we have $E[\mathbf{y}] \leq \liminf_{h\to 0} E_h[\mathbf{y}_h]$.

Step (ii): lim-sup property. We need to apply the regularization argument: by [48], isometries $\mathbf{y}$ in $[H^2(\Omega)]^3$ can be approximated with arbitrary precision in the H^2-norm by smooth isometries $\mathbf{y}^\epsilon$. Given $\mathbf{y} \in \mathbb{A}$ we take $\mathbf{y}^\epsilon$ be such approximation. Then we can define the recovery sequence as

$$\mathbf{y}_h(x) = \mathbf{y}^\epsilon(x_T) + \nabla \mathbf{y}^\epsilon(x_T)(x - x_T) + \frac{1}{2}(x - x_T)^T Q^\epsilon (x - x_T), \tag{3.2.8}$$

for $x \in T$ for any $T \in \mathcal{T}_h$. Here x_T is the barycenter of T, and $Q^\epsilon := \frac{1}{|T|}\int_T D^2\mathbf{y}^\epsilon$. It is clear that $\mathbf{y}_h \in [\mathbb{V}_h^k]^3$ for $k \geq 2$.

Now, we have

$$\|\mathbf{y}_h - \mathbf{y}^\epsilon\|_{L^2(T)} \lesssim h^3 \|D^3\mathbf{y}^\epsilon\|_{L^2(T)}, \tag{3.2.9}$$

by Bramble-Hilbert lemma, and

$$\|D^2\mathbf{y}_h - D^2\mathbf{y}^\epsilon\|_{L^2(T)} = \|Q^\epsilon - D^2\mathbf{y}^\epsilon\|_{L^2(T)} \lesssim h\|D^3\mathbf{y}^\epsilon\|_{L^2(T)} \tag{3.2.10}$$

by Poincaré inequality.

Moreover,

$$\|D^2\mathbf{y}_h\|_{L^2(T)} = \|Q^\epsilon\|_{L^2(T)} \leq |T|^{1/2}|\frac{1}{|T|}\int_T D^2\mathbf{y}^\epsilon| \leq |T|^{-1/2}\|D^2\mathbf{y}^\epsilon\|_{L^1(T)} \lesssim \|D^2\mathbf{y}^\epsilon\|_{L^2(T)}. \tag{3.2.11}$$

Also, it is easy to compute that $\nabla\mathbf{y}_h(x_T) = \nabla\mathbf{y}^\epsilon(x_T)$ for any T. Since $\mathbf{y}^\epsilon$ is an isometry, it holds trivially that $\mathbf{y}_h \in \mathbb{A}_{h,\delta}$ for any $\delta > 0$.

Summing over elements and recalling the arbitrary precision approximation property of $\mathbf{y}^\epsilon$, properties (3.2.9),(3.2.10), (3.2.11) imply that $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$, $D_h^2\mathbf{y}_h \to D^2\mathbf{y}$ in $[L^2(\Omega)]^{3\times2\times2}$ and $\|D_h^2\mathbf{y}_h\|_{L^2(\Omega)} \lesssim \|D^2\mathbf{y}\|_{L^2(\Omega)}$. Consequently, Lemma 3.1.1 and Lemma 3.2.1 guarantees that $\lim_{h\to0}\widetilde{E}_h[\mathbf{y}_h] = \widetilde{E}[\mathbf{y}]$.

Since $\widetilde{E}_h[\mathbf{y}_h]$ is uniformly bounded, then by Theorem 3.2.1, there holds $\|\mathbf{y}_h\|_{H_h^2(\Omega)}$ is uniformly bounded. Therefore, using $\mathbf{y}_h \to \mathbf{y}$ in $[L^2(\Omega)]^3$, we have $\widetilde{H_h}[\mathbf{y}_h]$ converges weakly to $D^2\mathbf{y}$ in $[L^2(\Omega)]^3$ by Theorem 3.1.4. Consequently, convergence of the remaining cubic term can be proved as in Step (i).

□

Remark 3.2.1. *In this remark we discuss the motivation of using one-point quadrature for cubic term and imposing discrete constraint on barycenters.*

1. L^∞***-control****: First, we need to impose the discrete constraint so that a discrete L^∞ norm of $\nabla_h\mathbf{y}_h^T\nabla\mathbf{y}_h - I_2$ is controlled; hence uniform boundedness of $\partial_i\mathbf{y}_h$ in the discrete L^∞ norm can be obtained from admissibility. This is used several times in this section.*

2. ***Compatibility****: We note that the cubic term in E_h should be compatible with*

the discrete constraint. For example, this is necessary to bound the R_2 term using the admissibility in the lim-inf proof. More specifically, as we impose the discrete constraint on barycenters, it is natural to use the one-point quadrature (at barycenter) for the cubic term in E_h. Otherwise, if we consider the exact integral of the cubic term, then the estimate of R_2 is problematic, because it is impossible to get a uniform $L^\infty(T)$ bound for gradients from point values at x_T without exploiting $W^{s,\infty}$ regularity ($s > 1$).

3. ***Options for the discrete constraint**: Unfortunately, using one-point quadrature for the cubic term may hinder the accuracy if one considers high degree of FEM space (the case $k > 2$), although an error analysis for this type of nonlinear nonconvex problem is beyond reach. Therefore, a natural question is: can we use more quadrature points or other alternative treatments for the cubic term, as well as discrete constraint? On one hand, it is not obvious how to construct a recovery sequence such that it satisfies the discrete isometry constraint at several quadrature points. On the other hand, if one considers an alternative way of defining the discrete constraint such as $|\frac{1}{|T|}\int_T(\nabla_h \mathbf{y}_h^T \nabla_h \mathbf{y}_h - I)| \leq \delta$ for each $T \in \mathcal{T}_h$, it is always a challenge to construct a suitable recovery sequence (satisfying (3.2.9),(3.2.10), and (3.2.11)), because $\nabla_h \mathbf{y}_h$ may not be closed to $\nabla \mathbf{y}$ in L^∞. This is a limitation of the current work, and to be enhanced in the future.*

3.3 Discrete gradient flow

To find a solution to (3.1.15), we conduct a discrete H_h^2 gradient flow. Recall from Chapter 2 the definition of H_h^2 metric as follows:

$$\begin{aligned}(\mathbf{v}_h, \mathbf{w}_h)_{H_h^2(\Omega)} &:= \sigma(\mathbf{v}_h, \mathbf{w}_h)_{L^2(\Omega)} + (D_h^2\mathbf{v}_h, D_h^2\mathbf{w}_h)_{L^2(\Omega)} \\ &\quad +(\mathrm{h}^{-1}[\nabla_h\mathbf{v}_h], [\nabla_h\mathbf{w}_h])_{L^2(\Gamma_h^a)} + (\mathrm{h}^{-3}[\mathbf{v}_h], [\mathbf{w}_h])_{L^2(\Gamma_h^a)},\end{aligned}$$

where $\sigma = 0$ when $\Gamma_D \neq \emptyset$ and $\sigma > 0$ when $\Gamma_D = \emptyset$. Moreover, we recall the Discrete Poincaré inequality from Lemma 2.2.3 and (2.4.2), and for simplicity of presentation, we denote the hidden constants in (2.4.2), (2.2.25), and (2.2.26) to be a unified constant c_p, which is independent of h.

The discrete H_h^2 gradient flow reads as follows. Given $\mathbf{y}_h^0 \in \mathbb{A}_{h,0}$ (i.e, $\mathbf{y}_h^0$ satisfies the isometry constraint exactly), in each step knowing $\mathbf{y}_h^n$, we seek $\delta\mathbf{y}_h^{n+1} \in \mathcal{F}_{h,b}(\mathbf{y}_h^n)$, where $\mathcal{F}_{h,b}(\mathbf{y}_h^n)$ is defined as

$$\mathcal{F}_{h,b}(\mathbf{y}_h^n) := \Big\{\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0}, \mathbf{0}) : \big[(\nabla\mathbf{y}_h^n)^T\nabla\mathbf{v}_h + (\nabla\mathbf{v}_h)^T\nabla\mathbf{y}_h^n\big](x_T) = 0 \ \forall\ T \in \mathcal{T}_h\Big\}. \tag{3.3.1}$$

Note that we impose a linearized discrete constraint in practice by requiring $\delta\mathbf{y}_h^{n+1} \in \mathcal{F}_{h,b}(\mathbf{y}_h^n)$ at each step. Indeed, $\mathcal{F}_{h,b}(\mathbf{y}_h^n)$ defines a discrete tangent space of the isometry constraint at $\mathbf{y}_h^n$ for each barycenter x_T.

We compute $\delta \mathbf{y}_h^{n+1} \in \mathcal{F}_{h,b}(\mathbf{y}_h^n)$ such that

$$\frac{1}{\tau}(\delta \mathbf{y}_h^{n+1}, \mathbf{v}_h)_{H_h^2} + a_h(\delta \mathbf{y}_h^{n+1}, \mathbf{v}_h) = -a_h(\mathbf{y}_h^n, \mathbf{v}_h) + \ell[\mathbf{y}_h^n](\mathbf{v}_h), \tag{3.3.2}$$

for any $\mathbf{v}_h \in \mathcal{F}_{h,b}(\mathbf{y}_h^n)$. Here, τ is a pseudo time step, and a_h is the bilinear form corresponding to the variational derivative of $\widetilde{E}_h[\mathbf{y}_h]$. More precisely

$$\begin{aligned} a_h(\mathbf{w}_h, \mathbf{v}_h) = \int_\Omega H_h[\mathbf{w}_h] : H_h[\mathbf{v}_h] &+ \gamma_1(\mathrm{h}^{-1}[\nabla_h \mathbf{w}_h], [\nabla_h \mathbf{v}_h])_{L^2(\Gamma_h^a)} \\ &+ \gamma_0(\mathrm{h}^{-3}[\mathbf{w}_h], [\mathbf{v}_h])_{L^2(\Gamma_h^a)}. \end{aligned}$$

The linear form $\ell[\mathbf{y}_h^n](\mathbf{v}_h)$ on $\mathbf{v}_h$ provides a linearization on the nonlinear term of variational derivative of the $E_h[\mathbf{y}_h]$ (comes from the cubic term), and it is defined as

$$\begin{aligned} \ell[\mathbf{y}_h^n](\mathbf{v}_h) = &\sum_{i,j=1}^{2} \sum_{T \in \mathcal{T}_h} |T| \big[(\widetilde{H_h}[\mathbf{v}_h])_{ij} \cdot (\partial_1 \mathbf{y}_h^n \times \partial_2 \mathbf{y}_h^n) \big](x_T) Z_{ij}(x_T) \\ &+ \sum_{i,j=1}^{2} \sum_{T \in \mathcal{T}_h} |T| \big[(\widetilde{H_h}[\mathbf{y}_h^n])_{ij} \cdot (\partial_1 \mathbf{v}_h \times \partial_2 \mathbf{y}_h^n) \big](x_T) Z_{ij}(x_T) \\ &+ \sum_{i,j=1}^{2} \sum_{T \in \mathcal{T}_h} |T| \big[(\widetilde{H_h}[\mathbf{y}_h^n])_{ij} \cdot (\partial_1 \mathbf{y}_h^n \times \partial_2 \mathbf{v}_h) \big](x_T) Z_{ij}(x_T). \end{aligned}$$

We note that this is an explicit treatment on the nonlinear term in the iterative scheme, as only $\mathbf{y}_h^n$ is involved.

3.3.1 Energy stability and admissibility

Motivated by [14] we have the following energy stability and admissibility property Theorem 3.3.1. Although we relax and linearize the isometry constraint in the iterative scheme, we prove that violation of the constraint is indeed controlled properly if τ is small enough.

First we prove a discrete inverse inequality, which will be useful in the proof of Theorem 3.3.1. Define $\mathbb{E}(\mathcal{T}_h) := \Pi_{T\in\mathcal{T}_h} H^1(T)$, and $h_{\min} := \min_{T\in\mathcal{T}_h} h_T$.

Lemma 3.3.1 (discrete Sobolev inequality)**.** *For any $v_h \in \mathbb{E}(\mathcal{T}_h) \cap W_h$, where W_h is any finite dimensional space, there holds*

$$\|v_h\|^2_{L^\infty(\Omega)} \lesssim (1 + |\log h_{\min}|)(\|\nabla_h v_h\|^2_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|^2_{L^2(\Gamma_h^0)} + \|v_h\|^2_{L^2(\Omega)}) \quad (3.3.3)$$

Proof. We consider a smoothing interpolation operator $\Pi_h : \mathbb{E}(\mathcal{T}_h) \to \mathbb{V}_h^k \cap H^1(\Omega)$ constructed originally in [21, 22] and discussed in Chapter 2, which satisfies (2.2.19) and (2.2.21), i.e,

$$\|\nabla \Pi_h v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-1}(v_h - \Pi_h v_h)\|_{L^2(\Omega)} \lesssim \|\nabla_h v_h\|_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|_{L^2(\Gamma_h^0)},$$

and

$$\|\Pi_h v_h\|_{L^2(\Omega)} \lesssim \|v_h\|_{L^2(\Omega)}.$$

Now, to prove (3.3.3) we consider

$$\begin{aligned}\|v_h\|^2_{L^\infty(\Omega)} &\lesssim \|v_h - \Pi_h v_h\|^2_{L^\infty(\Omega)} + \|\Pi_h v_h\|^2_{L^\infty(\Omega)} \\ &\lesssim \|\mathrm{h}^{-1}(v_h - \Pi_h v_h)\|^2_{L^2(\Omega)} + \|\Pi_h v_h\|^2_{L^\infty(\Omega)},\end{aligned}$$

where the second inequality uses an inverse estimate. Using (2.2.19), (2.2.21) and discrete inverse inequality for $\Pi_h v_h \in H^1(\Omega)$ [27]:

$$\|\Pi_h v_h\|^2_{L^\infty(\Omega)} \lesssim (1 + |\log h_{\min}|)\|\Pi_h v_h\|^2_{H^1(\Omega)}, \tag{3.3.4}$$

we get

$$\begin{aligned}\|v_h\|^2_{L^\infty(\Omega)} &\lesssim \|\nabla_h v_h\|^2_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|^2_{L^2(\Gamma_h^0)} + (1 + |\log h_{\min}|)\|\Pi_h v_h\|^2_{H^1(\Omega)} \\ &\lesssim (1 + |\log h_{\min}|)(\|\nabla_h v_h\|^2_{L^2(\Omega)} + \|\mathrm{h}^{-\frac{1}{2}}[v_h]\|^2_{L^2(\Gamma_h^0)} + \|v_h\|^2_{L^2(\Omega)}).\end{aligned}$$

This concludes the proof. □

Theorem 3.3.1. *Given $\mathbf{y}_h^0 \in \mathbb{A}_{h,0}$, assume $E_h(\mathbf{y}_h^0) \leq c_0$ with c_0 independent of h. There exists a constant α_1 independent of N and h such that if $\tau < (2\alpha_1|\log h_{\min}|)^{-1}$, then for any iterate $\mathbf{y}_h^N$ we have*

$$E_h[\mathbf{y}_h^N] + \frac{1}{2\tau}\sum_{n=0}^{N-1}\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2} \leq E_h[\mathbf{y}_h^0], \tag{3.3.5}$$

and there exists constants α_2 and α_3 independent of N and h such that

$$\left|\left[(\nabla \mathbf{y}_h^N)^T \nabla \mathbf{y}_h^N - I_2\right](x_T)\right| \leq \alpha_3 \tau |\log h_{\min}| \big(E_h[\mathbf{y}_h^0] + \alpha_2\big), \tag{3.3.6}$$

for any $T \in \mathcal{T}_h$. Moreover, α_1 and α_2 depend on the data $\boldsymbol{\varphi}$, Φ and Z, while α_3 is independent of the data.

Proof. We proceed by induction, and first we state the induction hypothesis as follows:

Induction hypothesis 1: assume the estimate (3.3.5) holds for any $N \leq k$ ($k \geq 0$) with the constant α_1;

Induction hypothesis 2: assume the estimate (3.3.6) holds for any $N \leq k$ ($k \geq 0$) with the constant α_2 and α_3.

Now we prove (3.3.5) and (3.3.6) are valid for $N = k+1$ with the same constants α_1, α_2 and α_3 that are discovered in the process, and the constants are independent of k and h. We split the proof into several steps.

Step (i): intermediate estimate on $\|\delta \mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \tau a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1})$ in terms

of $\mathbf{y}_h^k$. Taking the test function $\mathbf{v}_h = \delta\mathbf{y}_h^{k+1}$ in (3.3.2), we have

$$\begin{aligned}
&\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \frac{\tau}{2}a_h(\mathbf{y}_h^{k+1},\mathbf{y}_h^{k+1}) - \frac{\tau}{2}a_h(\mathbf{y}_h^{k},\mathbf{y}_h^{k}) + \frac{\tau}{2}a_h(\delta\mathbf{y}_h^{k+1},\delta\mathbf{y}_h^{k+1}) \\
&= \tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\big[\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\cdot(\partial_1\mathbf{y}_h^k\times\partial_2\mathbf{y}_h^k)\big](x_T)Z_{ij}(x_T) \\
&+ \tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\big[\widetilde{H_h}[\mathbf{y}_h^{k}]_{ij}\cdot(\partial_1\delta\mathbf{y}_h^{k+1}\times\partial_2\mathbf{y}_h^k)\big](x_T)Z_{ij}(x_T) \\
&+ \tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\big[\widetilde{H_h}[\mathbf{y}_h^{k}]_{ij}\cdot(\partial_1\mathbf{y}_h^k\times\partial_2\delta\mathbf{y}_h^{k+1})\big](x_T)Z_{ij}(x_T).
\end{aligned}$$

Then, by Cauchy-Schwarz inequality we have

$$\begin{aligned}
&\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \frac{\tau}{2}a_h(\mathbf{y}_h^{k+1},\mathbf{y}_h^{k+1}) - \frac{\tau}{2}a_h(\mathbf{y}_h^{k},\mathbf{y}_h^{k}) + \frac{\tau}{2}a_h(\delta\mathbf{y}_h^{k+1},\delta\mathbf{y}_h^{k+1}) \qquad (3.3.7)\\
&\le \tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}\Big[|T|^{\frac{1}{2}}|\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}(x_T)||T|^{\frac{1}{2}}|\partial_1\mathbf{y}_h^k(x_T)||\partial_2\mathbf{y}_h^k(x_T)| \\
&+ |T|^{\frac{1}{2}}|\widetilde{H_h}[\mathbf{y}_h^{k}]_{ij}(x_T)||T|^{\frac{1}{2}}|\partial_1\delta\mathbf{y}_h^{k+1}(x_T)||\partial_2\mathbf{y}_h^k(x_T)| \\
&+ |T|^{\frac{1}{2}}|\widetilde{H_h}[\mathbf{y}_h^{k}]_{ij}(x_T)||T|^{\frac{1}{2}}|\partial_2\delta\mathbf{y}_h^{k+1}(x_T)||\partial_1\mathbf{y}_h^k(x_T)|\Big]\|Z_{ij}\|_{L^\infty(T)}.
\end{aligned}$$

By *Induction hypothesis 2* we have

$$\big|\big[(\nabla\mathbf{y}_h^k)^T\nabla\mathbf{y}_h^k - I_2\big](x_T)\big| \le \alpha_3\tau|\log h_{\min}|\big(E_h[\mathbf{y}_h^0] + \alpha_2\big) < C_1, \qquad (3.3.8)$$

where $C_1 = \frac{\alpha_3(c_0+\alpha_2)}{2\alpha_1}$ due to $\tau < (2\alpha_1|\log h_{\min}|)^{-1}$ and $E_h[\mathbf{y}_h^0] < c_0$. By Lemma 3.1.2 with $\delta = C_1$, this implies that

$$|\partial_i\mathbf{y}_h^k(x_T)|^2 \le C_2 := C_1 + 1, \qquad (3.3.9)$$

for $i = 1, 2$. Moreover, recalling (3.2.7) and noting that $a_h(\delta\mathbf{y}_h^{k+1}, \delta\mathbf{y}_h^{k+1}) \geq 0$, together with Lemma 3.1.3 and Lemma 2.2.3, from (3.3.7) we get

$$\begin{aligned}
&\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \frac{\tau}{2}a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1}) - \frac{\tau}{2}a_h(\mathbf{y}_h^k, \mathbf{y}_h^k) \qquad (3.3.10)\\
&\leq c_{inv}^2\|Z\|_{L^\infty(\Omega)}\sqrt{C_2}\tau\Big[\|\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]\|_{L^2(\Omega)}\|\nabla_h\mathbf{y}_h^k\|_{L^2(\Omega)}\\
&+ 2\|\widetilde{H_h}[\mathbf{y}_h^k]\|_{L^2(\Omega)}\|\nabla_h\delta\mathbf{y}_h^{k+1}\|_{L^2(\Omega)}\Big]\\
&\leq c_{inv}^2 c_{stab}\|Z\|_{L^\infty(\Omega)}\sqrt{C_2}c_p\tau\Big[\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}(\|\mathbf{y}_h^k\|_{H_h^2} + C_{\boldsymbol{\varphi},\Phi}) + 2\|\mathbf{y}_h^k\|_{H_h^2}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}\Big].
\end{aligned}$$

Then, denoting

$$C_3 := c_{inv}^2 c_{stab}\|Z\|_{L^\infty(\Omega)}\sqrt{C_2}c_p, \qquad (3.3.11)$$

by Young's inequality the term $\frac{1}{2}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2$ can be absorbed to the left hand side of (3.3.10), and we get

$$\frac{1}{2}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \frac{\tau}{2}a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1}) \leq \frac{\tau}{2}a_h(\mathbf{y}_h^k, \mathbf{y}_h^k) + \frac{1}{2}C_3^2\tau^2(3\|\mathbf{y}_h^k\|_{H_h^2} + C_{\boldsymbol{\varphi},\Phi})^2. \qquad (3.3.12)$$

Step (ii): intermediate estimate on $\tau^{-1}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1})$ in terms of constants. By *Induction hypothesis 1*, it is clear that

$$E_h[\mathbf{y}_h^k] \leq E_h[\mathbf{y}_h^0] < c_0. \qquad (3.3.13)$$

By Theorem 3.2.3 and in particular (3.2.4),

$$\|\mathbf{y}_h^k\|_{H_h^2}^2 \leq 2c_{coer}E_h[\mathbf{y}_h^k] + \hat{c}_{coer}(C_1), \qquad (3.3.14)$$

and we emphasize that $\hat{c}_{coer}(C_1)$ means the δ in (3.2.5) is replaced by C_1, because $\mathbf{y}_h^k \in \mathbb{A}_{h,C_1}$ by (3.3.8). Moreover, the constant $\hat{c}_{coer}(C_1)$ is independent of h and k, but related to data $\boldsymbol{\varphi}$, Φ and Z. Then combining (3.3.13) and (3.3.14), we conclude that

$$\|\mathbf{y}_h^k\|_{H_h^2}^2 \leq 2c_{coer}c_0 + \hat{c}_{coer}(C_1). \tag{3.3.15}$$

By continuity of the bilinear form $a_h(\mathbf{y}_h^k, \mathbf{y}_h^k) \leq c_{cont}\|\mathbf{y}_h^k\|_{H_h^2}^2$, where the constant c_{cont} is independent of h, k, and α_i $(i = 1, 2, 3)$, we have

$$a_h(\mathbf{y}_h^k, \mathbf{y}_h^k) \leq 2c_{cont}c_{coer}c_0 + c_{cont}\hat{c}_{coer}(C_1). \tag{3.3.16}$$

Substituting (3.3.16) and (3.3.15) into (3.3.12), we divide both sides of (3.3.12) by τ. Without losing of generality we further assume that $h_{\min}$ is small enough so that $|\log h_{\min}| > 1$ and $\tau < (2\alpha_1)^{-1}$. We get the intermediate estimate

$$\frac{1}{2\tau}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \frac{1}{2}a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1}) \leq C_4, \tag{3.3.17}$$

where constant C_4 is defined as

$$C_4 := c_{cont}c_{coer}c_0 + \frac{1}{2}c_{cont}\hat{c}_{coer}(C_1) + \frac{C_3^2}{4\alpha_1}(36c_{coer}c_0 + 18\hat{c}_{coer}(C_1) + 2C_{\boldsymbol{\varphi},\Phi}^2), \tag{3.3.18}$$

where only C_3 and $\hat{c}_{coer}(C_1)$ are related to α_i $(i = 1, 2, 3)$ and all the terms are independent of k and h.

Step (iii): proof of (3.3.5) for $N = k + 1$. Now, we rewrite the following

term in (3.3.7):

$$
\begin{aligned}
&|T|\Big[\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\cdot(\partial_1\mathbf{y}_h^k\times\partial_2\mathbf{y}_h^k) \qquad (3.3.19)\\
&+\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\cdot(\partial_1\delta\mathbf{y}_h^{k+1}\times\partial_2\mathbf{y}_h^k)+\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\cdot(\partial_1\mathbf{y}_h^k\times\partial_2\delta\mathbf{y}_h^{k+1})\Big](x_T)Z_{ij}(x_T)\\
&=|T|\big[\widetilde{H_h}[\mathbf{y}_h^{k+1}]_{ij}\cdot(\partial_1\mathbf{y}_h^{k+1}\times\partial_2\mathbf{y}_h^{k+1})Z_{ij}\big](x_T)-|T|\big[\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\cdot(\partial_1\mathbf{y}_h^k\times\partial_2\mathbf{y}_h^k)Z_{ij}\big](x_T)\\
&-|T|\big[\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\cdot(\partial_1\mathbf{y}_h^{k+1}\times\partial_2\mathbf{y}_h^{k+1}-\partial_1\mathbf{y}_h^k\times\partial_2\mathbf{y}_h^k)Z_{ij}\big](x_T)\\
&+|T|\big[\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\cdot(\partial_1\delta\mathbf{y}_h^{k+1}\times(\partial_2\mathbf{y}_h^k-\partial_2\mathbf{y}_h^{k+1}))Z_{ij}\big](x_T),
\end{aligned}
$$

where we use the identity

$$
a^{k+1}b^{k+1}c^{k+1}-a^kb^kc^k=(a^{k+1}-a^k)b^{k+1}c^{k+1}+a^k(b^{k+1}-b^k)c^{k+1}+a^kb^k(c^{k+1}-c^k).
$$

Note that the first two terms contribute to the energy $E_h[\mathbf{y}_h^{k+1}]$ while the last two terms are considered as a remainder, and then we can substitute (3.3.19) into (3.3.7) and reach

$$
\begin{aligned}
&\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2+\tau E_h[\mathbf{y}_h^{k+1}]-\tau E_h[\mathbf{y}_h^k]\\
&\le\tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\Big|\big[\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\cdot(\partial_1\mathbf{y}_h^{k+1}\times\partial_2\mathbf{y}_h^{k+1}-\partial_1\mathbf{y}_h^k\times\partial_2\mathbf{y}_h^k)Z_{ij}\big](x_T)\Big|\\
&+\tau\sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}|T|\Big|\big[\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\cdot(\partial_1\delta\mathbf{y}_h^{k+1}\times(\partial_2\mathbf{y}_h^k-\partial_2\mathbf{y}_h^{k+1}))Z_{ij}\big](x_T)\Big|,
\end{aligned}
$$

and then due to (3.2.7) we have

$$
\begin{aligned}
&\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \tau E_h[\mathbf{y}_h^{k+1}] - \tau E_h[\mathbf{y}_h^k] \qquad (3.3.20)\\
&\leq c_{inv}^2\tau \sum_{i,j=1}^{2}\sum_{T\in\mathcal{T}_h}\Big[\|\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\|_{L^2(T)}\|\partial_1\delta\mathbf{y}_h^{k+1}\|_{L^\infty(T)}\|\partial_2\mathbf{y}_h^{k+1}\|_{L^2(T)}\\
&+ \|\widetilde{H_h}[\delta\mathbf{y}_h^{k+1}]_{ij}\|_{L^2(T)}|\partial_1\mathbf{y}_h^k(x_T)|\|\partial_2\delta\mathbf{y}_h^{k+1}\|_{L^2(T)}\\
&+ \|\widetilde{H_h}[\mathbf{y}_h^k]_{ij}\|_{L^2(T)}\|\partial_1\delta\mathbf{y}_h^{k+1}\|_{L^\infty(T)}\|\partial_2\delta\mathbf{y}_h^{k+1}\|_{L^2(T)}\Big]\|Z_{ij}\|_{L^\infty(T)}\\
&\leq c_{inv}^2\tau\|Z\|_{L^\infty(\Omega)}\Big[c_{stab}c_p\sqrt{C_2}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2\\
&+ 4c_{stab}c_pc_{sob}|\log h_{\min}|\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}^2(\|\mathbf{y}_h^{k+1}\|_{H_h^2} + C_{\boldsymbol{\varphi},\Phi} + \|\mathbf{y}_h^k\|_{H_h^2})\Big],
\end{aligned}
$$

where for the last inequality we apply (3.3.3) to $\partial_i\delta\mathbf{y}_h^{k+1}$ and consider further discrete Poincaré inequality (2.2.26) to get that

$$
\|\partial_i\delta\mathbf{y}_h^{k+1}\|_{L^\infty(\Omega)} \leq c_{sob}(1+|\log h_{\min}|)^{\frac{1}{2}}\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2} \leq 2c_{sob}|\log h_{\min}|\|\delta\mathbf{y}_h^{k+1}\|_{H_h^2}, \qquad (3.3.21)
$$

for $i = 1, 2$. Moreover, by (3.2.1), (3.3.15) and (3.3.17) with the fact that $\widetilde{E}_h[\mathbf{y}_h^{k+1}] = \frac{1}{2}a_h(\mathbf{y}_h^{k+1}, \mathbf{y}_h^{k+1})$ we have

$$
\begin{aligned}
\|\mathbf{y}_h^{k+1}\|_{H_h^2} + C_{\boldsymbol{\varphi},\Phi} + \|\mathbf{y}_h^k\|_{H_h^2} &\leq c_{coer}^{\frac{1}{2}}(\widetilde{E}_h[\mathbf{y}_h^{k+1}] + C_{\boldsymbol{\varphi},\Phi}^2)^{\frac{1}{2}} + C_{\boldsymbol{\varphi},\Phi} + \|\mathbf{y}_h^k\|_{H_h^2} \qquad (3.3.22)\\
&\leq c_{coer}^{\frac{1}{2}}(C_4 + C_{\boldsymbol{\varphi},\Phi}^2)^{\frac{1}{2}} + C_{\boldsymbol{\varphi},\Phi} + \big(2c_{coer}c_0 + \hat{c}_{coer}(C_1)\big)^{\frac{1}{2}}
\end{aligned}
$$

Finally, substitute (3.3.22) into (3.3.20) and recall the definition of C_3 (3.3.11) we

have

$$\|\delta \mathbf{y}_h^{k+1}\|_{H_h^2}^2 + \tau E_h[\mathbf{y}_h^{k+1}] - \tau E_h[\mathbf{y}_h^k] \leq C_5 \tau |\log h_{\min}| \|\delta \mathbf{y}_h^{k+1}\|_{H_h^2}^2, \tag{3.3.23}$$

where

$$C_5 = C_3 + 4c_{sob} C_2^{-\frac{1}{2}} C_3 \big[c_{coer}^{\frac{1}{2}} (C_4 + C_{\boldsymbol{\varphi},\Phi}^2)^{\frac{1}{2}} + C_{\boldsymbol{\varphi},\Phi} + \big(2c_{coer} c_0 + \hat{c}_{coer}(C_1)\big)^{\frac{1}{2}} \big], \tag{3.3.24}$$

where only constants C_2, C_3, C_4, and $\hat{c}_{coer}(C_1)$ depend on α_i $(i = 1, 2, 3)$, and all the terms are independent of h, k. Moreover, if α_1, α_2 and α_3 are such that

$$C_5 \leq \alpha_1, \tag{3.3.25}$$

then together with $\tau < (2\alpha_1 |\log h_{\min}|)^{-1}$ we prove (3.3.5) for $N = k + 1$. The validity of (3.3.25) will be justified in the last step.

Step (iv): intermediate estimate of $\big|\big[(\nabla \mathbf{y}_h^{k+1})^T \nabla \mathbf{y}_h^{k+1} - I_2\big](x_T)\big|$. Now for $\delta \mathbf{y}_h^{k+1} \in \mathcal{F}_{h,b}(\mathbf{y}_h^k)$ and any $T \in \mathcal{T}_h$, by definition (3.3.1) there holds

$$\begin{aligned} \big|\big[(\nabla \mathbf{y}_h^{k+1})^T \nabla \mathbf{y}_h^{k+1} - I_2\big](x_T)\big| \leq{}& \big|\big[(\nabla \mathbf{y}_h^{k})^T \nabla \mathbf{y}_h^{k} - I_2\big](x_T)\big| \\ &+ \big|(\nabla \delta \mathbf{y}_h^{k+1}(x_T))^T \nabla \delta \mathbf{y}_h^{k+1}(x_T)\big|. \end{aligned} \tag{3.3.26}$$

By (3.3.3) and discrete Poincaré inequality (2.2.26),

$$|(\nabla\delta\mathbf{y}_h^{k+1}(x_T))^T\nabla\delta\mathbf{y}_h^{k+1}(x_T)| \leq \|\nabla_h\delta\mathbf{y}_h^{k+1}\|^2_{L^\infty(T)} \leq c_{sob}(1+|\log h_{\min}|)\|\delta\mathbf{y}_h^{k+1}\|^2_{H_h^2}. \tag{3.3.27}$$

Using the intermediate estimate (3.3.17) and $\tau < (2\alpha_1|\log h_{\min}|)^{-1}$, we deduce that

$$|(\nabla\delta\mathbf{y}_h^{k+1}(x_T))^T\nabla\delta\mathbf{y}_h^{k+1}(x_T)| \leq c_{sob}(1+|\log h_{\min}|)2\tau C_4 \leq 2C_4c_{sob}\alpha_1^{-1}. \tag{3.3.28}$$

Together with (3.3.8), we conclude that

$$\big|\big[(\nabla\mathbf{y}_h^{k+1})^T\nabla\mathbf{y}_h^{k+1} - I_2\big](x_T)\big| \leq C_1 + 2C_4c_{sob}\alpha_1^{-1} =: C_6. \tag{3.3.29}$$

This implies that $\mathbf{y}_h^{k+1} \in \mathbb{A}_{h,C_6}$.

Step (v): proof of (3.3.6) for $N = k+1$. As $\mathbf{y}_h^0$ satisfies isometry constraint on x_T, recursively using (3.3.26) and (3.3.27) we have

$$|\big[(\nabla\mathbf{y}_h^{k+1})^T\nabla\mathbf{y}_h^{k+1} - I_2\big](x_T)| \leq c_{sob}(1+|\log h_{\min}|)\sum_{n=0}^{k}\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2}, \tag{3.3.30}$$

Since in step (iii) we prove (3.3.5) for $N = k+1$, we conclude that

$$\sum_{n=0}^{k}\|\delta\mathbf{y}_h^{n+1}\|^2_{H_h^2} \leq 2\tau(E_h[\mathbf{y}_h^0] - E_h[\mathbf{y}_h^{k+1}]).$$

Moreover, by (3.2.4) and the fact that $\mathbf{y}_h^{k+1} \in \mathbb{A}_{h,C_6}$, a lower bound

$$E_h[\mathbf{y}_h^{k+1}] \geq -(2c_{coer})^{-1}\hat{c}_{coer}(C_6)$$

holds. Therefore,

$$\sum_{n=0}^{k} \|\delta\mathbf{y}_h^{n+1}\|_{H_h^2}^2 \leq 2\tau(E_h[\mathbf{y}_h^0] + (2c_{coer})^{-1}\hat{c}_{coer}(C_6)).$$

Hence, noting $|\log h_{\min}| > 1$ we have

$$|\big[(\nabla\mathbf{y}_h^{k+1})^T\nabla\mathbf{y}_h^{k+1} - I_2\big](x_T)| \leq 4c_{sob}|\log h_{\min}|\tau E_h[\mathbf{y}_h^0] + 2c_{sob}c_{coer}^{-1}\hat{c}_{coer}(C_6)|\log h_{\min}|\tau, \tag{3.3.31}$$

and thus this leads to (3.3.6) if

$$\alpha_3 \geq 4c_{sob} \text{ and } \alpha_2\alpha_3 \geq 2c_{sob}c_{coer}^{-1}\hat{c}_{coer}(C_6). \tag{3.3.32}$$

Step (vi): choice of the constants α_i. Our goal is to show that there exists at least a set of constants $\{\alpha_1, \alpha_2, \alpha_3\}$ that is independent of h, k and satisfies the system of inequalities (3.3.25) and (3.3.32). We first fix $\alpha_3 = 5c_{sob}$ by the first inequality in (3.3.32), and α_3 is independent of h, k and the data $\boldsymbol{\varphi}$, Φ and Z.

Now it suffice to solve the following system of inequalities for α_1 and α_2:

$$\alpha_2 \geq \frac{2}{5}c_{coer}^{-1}\hat{c}_{coer}(C_6) \text{ and } \alpha_1 \geq C_5. \tag{3.3.33}$$

Recall the constant C_1 is defined as $C_1 = \alpha_3(c_0 + \alpha_2)\alpha_1^{-1}$, and note that $\hat{c}_{coer}(\delta)$ is quadratic in δ as in (3.2.5). Moreover, we observe that C_i $(i = 2, \ldots, 6)$, $\hat{c}_{coer}(C_1)$ and $\hat{c}_{coer}(C_6)$ are expressed in only positive powers of C_1 and constants independent of $h, k, \alpha_1, \alpha_2, \alpha_3$, which further implies that they can be expressed in only negative powers of α_1, positive powers of $r := \alpha_2\alpha_1^{-1}$ and constants unrelated to $h, k, \alpha_1, \alpha_2, \alpha_3$. If we consider a particular case $\alpha_2 = \sqrt{\alpha_1}$ and we let $\alpha_1 \to \infty$, then we first notice that $r \to 0$. Therefore, in the limit of $\alpha_1 \to \infty$, we can observe in the order that $C_1 \to 0$, $C_2 \to 1$, C_3 tends to a constant depending on Z, $\hat{c}_{coer}(C_1)$, C_4 and C_5 converges to a constant depending on data $\boldsymbol{\varphi}$, Φ and Z, $C_6 \to 0$, and $\hat{c}_{coer}(C_6)$ also tends to a constant depending on the data. Consequently, we conclude that when α_1 is sufficiently large and $\alpha_2 = \sqrt{\alpha_1}$ they satisfy the system of inequalities (3.3.33). In particular, they are independent of h, k, but depends on the data $\boldsymbol{\varphi}$, Φ and Z implicitly. In summary, this shows the existence of such a set of constants $\{\alpha_1, \alpha_2, \alpha_3\}$ and concludes the proof. □

3.3.2 Inf-sup stability

To deal with the constraint $\delta\mathbf{y}_h^{n+1} \in \mathcal{F}_{h,b}(\mathbf{y}_h^n)$, we apply the method of Lagrange multipliers. Indeed, we use piecewise constant Lagrange multipliers in the space

$$\Lambda_h := \left\{ \lambda_h : \Omega \to \mathbb{R}^{2\times 2} : \ \lambda_h^T = \lambda_h, \ \lambda_h \in \left[\mathbb{V}_h^0\right]^{2\times 2} \right\}.$$

We define the bilinear form $b_h(\cdot,\cdot;\mathbf{y}_h^n)$ for any $(\mathbf{v}_h,\boldsymbol{\mu}_h)\in\mathbb{V}_h^k(\mathbf{0},\mathbf{0})\times\Lambda_h$ to be

$$b_h(\mathbf{v}_h,\boldsymbol{\mu}_h;\mathbf{y}_h^n):=\sum_{T\in\mathcal{T}_h}|T|(\nabla\mathbf{v}_h^T\nabla\mathbf{y}_h^n+(\nabla\mathbf{y}_h^n)^T\nabla\mathbf{v}_h)(x_T):\boldsymbol{\mu}_h. \tag{3.3.34}$$

We observe that $b_h(\cdot,\cdot;\mathbf{y}_h^n)$ depends on $\mathbf{y}_h^n$ and that $b_h(\delta\mathbf{y}_h^{n+1},\boldsymbol{\mu}_h;\mathbf{y}_h^n)=0$ for all $\boldsymbol{\mu}_h\in\Lambda_h$ implies (3.3.1) for all $T\in\mathcal{T}_h$. Therefore, in each gradient flow step with the linearized metric constraint we seek $(\delta\mathbf{y}_h^{n+1},\boldsymbol{\lambda}_h^{n+1})\in\mathbb{V}_h^k(\mathbf{0},\mathbf{0})\times\Lambda_h$ such that

$$\begin{aligned}\tau^{-1}(\delta\mathbf{y}_h^{n+1},\mathbf{v}_h)_{H_h^2}+a_h(\delta\mathbf{y}_h^{n+1},\mathbf{v}_h)+b_h(\mathbf{v}_h,\boldsymbol{\lambda}_h^{n+1};\mathbf{y}_h^n)&=\ell[\mathbf{y}_h^n](\mathbf{v}_h)-a_h(\mathbf{y}_h^n,\mathbf{v}_h)\\ b_h(\delta\mathbf{y}_h^{n+1},\boldsymbol{\mu}_h;\mathbf{y}_h^n)&=0\end{aligned} \tag{3.3.35}$$

for all $(\mathbf{v}_h,\boldsymbol{\mu}_h)\in\mathbb{V}_h^k(\mathbf{0},\mathbf{0})\times\Lambda_h$. Since $\mathbf{y}_h^n\in\mathbb{V}_h^k(\boldsymbol{\varphi},\Phi)$, whence $\mathbf{y}_h^{n+1}=\mathbf{y}_h^n+\delta\mathbf{y}_h^{n+1}\in\mathbb{V}_h^k(\boldsymbol{\varphi},\Phi)$, the Dirichlet condition of $\mathbf{y}_h^{n+1}$ is inherited from that of $\mathbf{y}_h^n$, which in turn appears on the right-hand side $a_h(\mathbf{y}_h^n,\mathbf{v}_h)$ when Γ_D is not empty.

The proposed strategy is summarized in Algorithm 3.

Algorithm 3: (discrete-H^2 gradient flow) Finding local minima of E_h

Given a pseudo-time step $\tau>0$ and a target tolerance *tol*;
Choose initial guess $\mathbf{y}_h^0\in\mathbb{A}_{h,0}$;
while $\tau^{-1}|E_h[\mathbf{y}_h^{n+1}]-E_h[\mathbf{y}_h^n]|>tol$ **do**
  Solve (3.3.35) for $(\delta\mathbf{y}_h^{n+1},\boldsymbol{\lambda}_h^{n+1})\in\mathbb{V}_h^k(\mathbf{0},\mathbf{0})\times\Lambda_h$;
  Update $\mathbf{y}_h^{n+1}=\mathbf{y}_h^n+\delta\mathbf{y}_h^{n+1}$;
end

We solve (3.3.35) using the *Schur complement approach* as in Chapter 2. As indicated in [61], the solvability of Schur complement matrix relies on the continuity of a_h and inf-sup stability of b_h, and also as in [16] the conditioning of the Schur complement matrix depends on the coercivity of a_h, boundedness of b_h and inf-sup

stability of b_h.

To prove inf-sup stability of b_h, we first show a lemma in the context of linear algebra as follows.

Lemma 3.3.2 (solvability of auxiliary matrix equation)**.** *Given a 2×2 symmetric matrix C, and given a 3×2 matrix B with strictly positive smallest singular value $\sigma_{min}(B) > 0$ such that $BC \neq 0$, we can find a 3×2 matrix A that solves the equation*

$$(A^T B + B^T A) : C = |C|^2, \tag{3.3.36}$$

and there holds $|A| \lesssim |C|$. Here $|\cdot|$ denotes the Frobenius norm of matrices.

Proof. Note that

$$A^T B : C = \mathrm{tr}(B^T A C) = \mathrm{tr}(C A^T B) = \mathrm{tr}(A^T B C) = B^T A : C = A : BC.$$

Thus (3.3.36) is equivalent to

$$A : BC = \frac{1}{2}|C|^2,$$

and so

$$A = \frac{(BC)|C|^2}{2|BC|^2}$$

is clearly a solution to (3.3.36). Also,

$$|A| = \frac{|C|^2}{2|BC|}.$$

Considering the singular value decomposition of B, we have $B = U\Sigma V^T$, where U is 3×3 orthogonal matrix, V is a 2×2 orthogonal matrix, and Σ is a 3×2 matrix of form $[\sigma_1(B), 0; 0, \sigma_2(B); 0, 0]$ with singular values $\sigma_i(B)$ of B. Hence,

$$|BC|^2 = |U\Sigma V^T C|^2 = |\Sigma V^T C|^2 = |\Sigma \widetilde{C}|^2 \geq \sigma_{min}(B)^2 |\widetilde{C}|^2 = \sigma_{min}(B)^2 |C|^2,$$

where $\widetilde{C} = V^T C$ and thus $|\widetilde{C}| = |C|$. Note that $|C|$ cannot be 0, otherwise $C = 0$ and then $BC = 0$ that contradicts the assumption. As a result,

$$|A| \leq \frac{|C|}{2\sigma_{min}(B)}.$$

This finishes the proof. □

Therefore the following inf-sup condition for b_h defined in (3.3.34) holds.

Theorem 3.3.2. *For all $n \geq 0$, there exists a constant $\beta_h = \beta h_{\min} > 0$ independent of n such that*

$$\inf_{\mu_h \in \Lambda_h} \sup_{\mathbf{v}_h \in \mathbb{V}_h^k(\mathbf{0},\mathbf{0})} \frac{b_h(\mathbf{v}_h, \boldsymbol{\mu}_h; \mathbf{y}_h^n)}{\|\mathbf{v}_h\|_{H_h^2(\Omega)} \|\boldsymbol{\mu}_h\|_{L^2(\Omega)}} \geq \beta_h. \tag{3.3.37}$$

Proof. Due to the Lemma 3.3.36, for any element $K \in \mathcal{T}_h$, we conclude that there exists 3×2 matrix A_K such that

$$\boldsymbol{\mu}_{h,K} : (A_K^T \nabla \mathbf{y}_h^n(x_K) + \nabla \mathbf{y}_h^n(x_K)^T A_K) = |\boldsymbol{\mu}_{h,K}|^2,$$

and $|A_K| \lesssim \frac{|\boldsymbol{\mu}_{h,K}|}{\sigma_{\min}(\nabla \mathbf{y}_h^n(x_K))}$, where $\boldsymbol{\mu}_{h,K} = \boldsymbol{\mu}_h|_K$ is a constant symmetric 2×2 matrix.

Note that by definition

$$\sigma_{\min}(\nabla \mathbf{y}_h^n(x_K)) = \big(\lambda_{\min}((\nabla \mathbf{y}_h^n(x_K))^T \nabla \mathbf{y}_h^n(x_K))\big)^{\frac{1}{2}},$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. Note that $\mathbf{y}_h^n \in \mathbb{A}_{h,\delta}$, and thus

$$\big|(\nabla \mathbf{y}_h^n(x_K))^T \nabla \mathbf{y}_h^n(x_K) - I_2\big| \leq \delta.$$

One can show that the function $\lambda(\cdot) : \mathbb{M}^{2\times 2} \to \mathbb{R}$ from the space of symmetric matrices $\mathbb{M}^{2\times 2}$ into $\mathbb{R}$ such that $\lambda(A)$ gives the (real) eigenvalues of A is continuous, say in the Frobenius norm $|\cdot|$, because all norms are equivalent. Hence,

$$|\lambda_{\min}\big((\nabla \mathbf{y}_h^n(x_K))^T \nabla \mathbf{y}_h^n(x_K) - I_2\big)| \lesssim \delta,$$

and therefore

$$\lambda_{\min}((\nabla \mathbf{y}_h^n(x_K))^T \nabla \mathbf{y}_h^n(x_K)) > 1 - c\delta.$$

Consequently, for δ small enough we have

$$\sigma_{\min}(\nabla \mathbf{y}_h^n(x_K)) := \big(\lambda_{\min}((\nabla \mathbf{y}_h^n(x_K))^T \nabla \mathbf{y}_h^n(x_K))\big)^{\frac{1}{2}}$$

is bounded away from 0. Therefore $|A_K| \lesssim \frac{|\mu_{h,K}|}{(1-c\delta)^{\frac{1}{2}}}$.

We define $\mathbf{v}_h$ as $\mathbf{v}_h(\mathbf{x}) := A_K\mathbf{x} - \frac{1}{|K|}\int_K A_K\mathbf{x}$ on each K, and it is clear that

$\mathbf{v}_h \in \mathbb{V}_h^k$ for $k \geq 2$ and $\nabla \mathbf{v}_h = A_K$ for all $K \in \mathcal{T}_h$. Thus, we have

$$\begin{aligned} b_h(\mathbf{v}_h, \boldsymbol{\mu}_h; \mathbf{y}_h^n) &= \sum_{K \in \mathcal{T}_h} |K| \boldsymbol{\mu}_{h,K} : [A_K^T \nabla \mathbf{y}_h^n(x_K) + \nabla \mathbf{y}_h^n(x_K)^T A_K] \\ &= \sum_{K \in \mathcal{T}_h} |\boldsymbol{\mu}_{h,K}|^2 |K| = \|\boldsymbol{\mu}_h\|_{L^2(\Omega)}^2. \end{aligned}$$

Note that with this choice of $\mathbf{v}_h$, $D^2 \mathbf{v}_h = 0$ on each element K. Now we compute:

$$\begin{aligned} \|\mathbf{v}_h\|_{h,2}^2 &= \sum_{e \in \mathcal{E}_h} h^{-3} \|[\mathbf{v}_h]\|_{L^2(e)}^2 + h^{-1} \|[\nabla \mathbf{v}_h]\|_{L^2(e)}^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} h^{-4} \|\mathbf{v}_h\|_{L^2(K)}^2 + h^{-2} \|\nabla \mathbf{v}_h\|_{L^2(K)}^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} h^{-2} \|\nabla \mathbf{v}_h\|_{L^2(K)}^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} h^{-2} \int_K |A_K|^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} h^{-2} \int_K \frac{|\boldsymbol{\mu}_{h,K}|^2}{(1 - c\delta)^2} \\ &\lesssim h_{\min}^{-2} (1 - c\delta)^{-2} \|\boldsymbol{\mu}_h\|_{L^2(\Omega)}^2, \end{aligned}$$

where we use trace inequality and Poincaré inequality ($\mathbf{v}_h$ is zero mean). □

3.4 Numerical experiments

In this section we present several numerical experiments, compared with the work [12, 13, 14, 23], and motivated by experimental work [2, 49, 72]. We conduct

simulations with various spontaneous curvature matrix Z, and we consider both *Dirichlet and free boundary conditions*. Different aspect ratios of rectangular domain are also taken into accounts. Our numerical simulations illustrate the effectiveness and efficiency of our algorithm.

3.4.1 Implementation

We start with a few comments on the implementation of the gradient flow (3.3.35) and Algorithm 3.

First, the discrete problem (3.3.35) is a *saddle-point system*, which can be solved efficiently by *Schur complement method*. For more details of its implementation, please refer to Chapter 2, where there is a similar linear algebra structure to solve the discrete problem in each step. Here, we emphasize that we use a conjugate gradient iterative solver to compute the inverse of the Schur complement matrix. In numerical experiments shown in this section, we observe that the number of iterations needed in the conjugate gradient solver is roughly order h, which is related to the condition number of the Schur complement matrix.

We emphasize that the scalar product $(\cdot,\cdot)_{H_h^2}$ and the bilinear form a_h are assembled once for all before the main loop, while the bilinear form b_h^n and right hand side are assembled at each step of the loop as they depend on the previous iterate $\mathbf{y}_h^n$. We notice that computing the discrete Hessian $H_h[\mathbf{y}_h]$ is the most expensive part in the assemble process, as it requires solving the linear system (2.2.7) and (2.2.8) for lifting operators. However, we manage to compute the discrete Hessian only one

time before the main loop by storing $H_h[\boldsymbol{\varphi}_h^i]$ globally in the code, where $\{\boldsymbol{\varphi}_h^i\}_{i=1}^N$ is a basis for $[\mathbb{V}_h^k]^3$. In this way, although $\ell[\mathbf{y}_h^n](\mathbf{v}_h)$ contains the discrete Hessians $\widetilde{H}_h[\mathbf{v}_h]$ and $\widetilde{H}_h[\mathbf{y}_h^n]$, the cost of assembling the right hand side of (3.3.35) is tiny, since the reduced discrete Hessian $\widetilde{H}_h$ are simply L^2 projection of discrete Hessian H_h, which is only a light computation. Moreover, evaluating $a_h(\mathbf{y}_h^n, \mathbf{v}_h)$ is also only a standard matrix-vector operation since we have assembled a_h and degrees of freedoms of $\mathbf{y}_h^n$ is known.

To summarize, the time-consuming parts of our algorithm mainly consists of two parts: i) computing discrete Hessian; ii) solving the linear system (3.3.35) by Schur complement method. ii) eventually dominates since it needs to be done in each iteration while i) is conducted once for all (empirically, it costs as much as several iterations). By storing $H_h[\boldsymbol{\varphi}_h^i]$ we need to sacrifice certain memory of the machine, but this is worthy as the speed and efficiency of the algorithm are improved a lot.

The implementation is carried out within the software platform deal.ii [7] and the visualization is performed with paraview [6].

For all the simulations, we fix the polynomial degree k of the deformation $\mathbf{y}_h$ and l_1, l_2 for the two liftings of the discrete Hessian $H_h[\mathbf{y}_h]$ to be

$$k = l_1 = l_2 = 2.$$

Moreover the stabilization parameters are taken to be

$$\gamma_0 = \gamma_1 = 1.$$

In contrast to [22, 23], these parameters do not need to be large for stability purposes.

In the following numerical simulations, we consider both free and clamped Dirichlet boundary conditions. In either situation, a natural choice for the initialization is $\mathbf{y}_h^0(x_1, x_2) = (x_1, x_2, 0)$ for $(x_1, x_2) \in \Omega$, which corresponds to a flat plane, and satisfies the isometry constraint everywhere so that $\mathbf{y}_h^0 \in \mathbb{A}_{h,0}$ and the clamped boundary condition. We notice that for a more complicated boundary condition not satisfied by $\mathbf{y}_h^0(x_1, x_2) = (x_1, x_2, 0)$, one can apply the *boundary condition preprocessing* and the *metric preprocessing* as in Chapter 2 to generate a suitable initialization $\mathbf{y}_h^0$.

3.4.2 Clamped plate: $Z = I_2$

We consider a rectangular plate $\Omega = (-5, 5) \times (-2, 2)$, clamped on the side $\{-5\} \times [-2, 2]$, with spontaneous curvature given by $Z = I_2$. The deformation with minimal energy corresponds to a cylinder of radius 1 and energy 20 [67]. We report iterations of the gradient flow in Fig.3.1, with number of elements is 1024 (30720 dofs), $\tau = 5 \times 10^{-3}$ and $tol = 10^{-4}$. A cylindrical equilibrium configuration is reached confirming the results in [67].

In contrast to [23], but similar to [14], we notice that self-intersecting appears here in the evolution. It takes fewer iterations for our algorithm to reach the equi-

librium configuration in this case, with the same or even smaller time step τ, than methods in [13, 14, 23].

Moreover, with fixed $\tau = 5 \times 10^{-3}$, we look at the mesh where there are 256 and 1024 elements respectively, and we observe that $E_h = 16.8627$ and $E_h = 17.8038$ respectively. The error in energy is smaller than the Kirchhoff method described in [13], for which $E_h = 15.961$ and $E_h = 16.544$ with the same mesh-size and time step τ. Also, the error is smaller than the new Kirchhoff method in [14], which computes with the case $Z = 2.5I_2$ and produce a 36% relative error with a smaller mesh (5120 triangular elements).

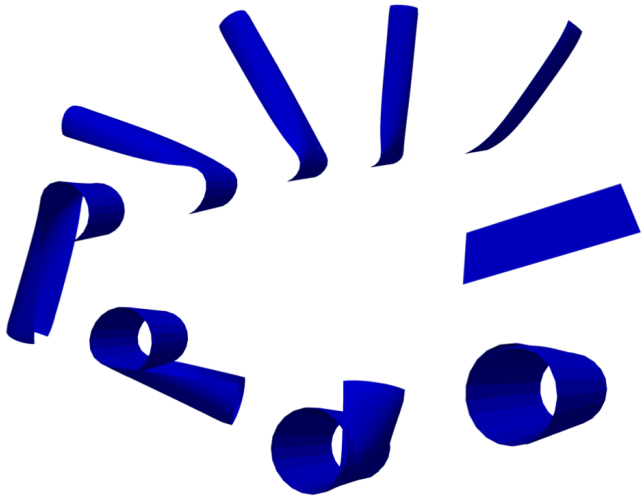

Figure 3.1: Evolution (counter-clockwise) towards the equilibrium of a clamped rectangular plate with spontaneous curvature $Z = I$. The bilayer plate is depicted at times $0, 50, 1000, 9000, 18000, 36050, 48100, 56050, 72100$ of the gradient flow.

3.4.3 Free plate:Anisotropic Curvature

We now explore a cigar-type configuration motivated by [49]. The plate $\Omega = (-5, 5) \times (-2, 2)$, and no boundary condition is imposed, with an anisotropic

spontaneous curvature

$$Z = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}.$$

We expect that the plate deforms at 45 degrees with respect to the cartesian axes in a symmetric way and eventually reaches a cigar-like configuration, as in [23], since two eigenvectors of Z are $[1,1]^T$ and $[1,-1]^T$. We confirm this in Figure 3.2. The computation is conducted with 1024 elements (30720 dofs) and $\tau = 5 \times 10^{-3}$. The final energy is $E_h = 46.3898$. It also takes fewer iterations to reach the equilibrium configuration than [23].

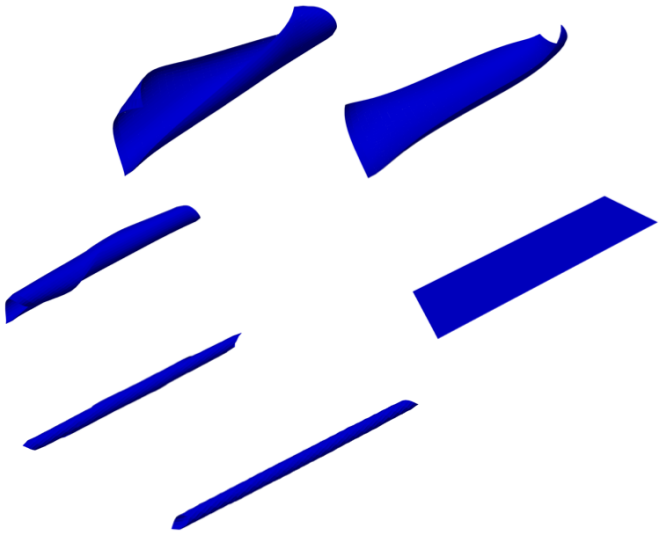

Figure 3.2: Evolution (counter-clockwise) towards the equilibrium of a free rectangular plate. The bilayer plate is depicted at times $0, 50, 200, 1000, 10000, 30000$ of the gradient flow.

3.4.4 Free plate: Helix Shape

We present the example with a helix-type shape motivated by [72], which is a DNA-like configuration. We consider a high aspect ratio plate $\Omega = (-8, 8) \times (-0.5, 0.5)$, and no boundary condition is imposed, with an anisotropic spontaneous curvature

$$Z = \begin{bmatrix} 1 & -3/2 \\ -3/2 & 1 \end{bmatrix}.$$

We expect that this choice of spontaneous curvature corresponds to principal directions that form an angle of 45 degrees with the coordinate axes, and together with the high aspect ratio this leads to a deformation that resembles the twisting of DNA molecules, as in [23]. We confirm this in Figure 3.3. The computation is conducted with 256 elements and $\tau = 10^{-2}$. The final energy is $E_h = 6.75379$. It also takes fewer iterations to reach the equilibrium configuration than [23].

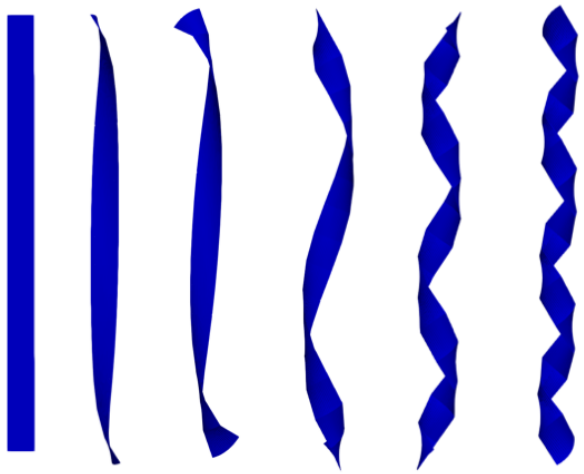

Figure 3.3: Evolution (left to right) towards the equilibrium of a free rectangular strip. The bilayer plate is depicted at times $0, 50, 200, 1000, 4000, 13380$ of the gradient flow.

Chapter 4: Γ-convergent projection-free finite element methods for nematic liquid crystals: The Ericksen model

In this chapter, we design a new FEM to compute the equilibrium state of nematic liquid crystals.

4.1 Problem formulation and discretization

4.1.1 Ericksen model

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded Lipschitz domain. In the Ericksen model (see, e.g., [37] or [76, Section 6.2]), the state of liquid crystals is described in terms of a unit-length vector field $\boldsymbol{n} : \Omega \to \mathbb{S}^{d-1}$ and a scalar function $s : \Omega \to [-1/(d-1), 1]$, usually referred to as *director* and *degree of orientation*, respectively. Equilibrium configurations of the liquid crystals are minimizers of the energy $E[s, \boldsymbol{n}] = E_1[s, \boldsymbol{n}] + E_2[s]$ in (1.3.1), where

$$E_1[s, \boldsymbol{n}] := \frac{1}{2} \int_\Omega \big(\kappa |\nabla s|^2 + s^2 |\boldsymbol{\nabla} \boldsymbol{n}|^2\big), \qquad E_2[s] := \int_\Omega \psi(s). \tag{4.1.1}$$

Here, $\kappa > 0$ is constant, while the double well potential $\psi : (-1/(d-1), 1) \to \mathbb{R}_{\geq 0}$ satisfies the following properties:

- $\psi \in C^2(-1/(d-1), 1)$,
- $\lim_{s\to 1^-} \psi(s) = +\infty = \lim_{s\to -1/(d-1)^+} \psi(s)$,
- $\psi(0) > \psi(s^*) = \min_{s\in(-1/(d-1),1)} \psi(s) = 0$ for some $s^* \in (0, 1)$,
- $\psi'(0) = 0$.

In (4.1.1), $E_1[s, \boldsymbol{n}]$ is the so-called *one-constant* approximation of the elastic energy proposed in [37], while $E_2[s]$ is a potential energy which confines the variable s within the physically admissible interval $(-1/(d-1), 1)$. The presence of the weight s^2 in the second term of $E_1[s, \boldsymbol{n}]$ allows for blow-up of $\boldsymbol{\nabla n}$, namely $\boldsymbol{n} \notin \boldsymbol{H}^1(\Omega)$, in the *singular set* Σ where defects may occur

$$\Sigma := \{x \in \Omega : s(x) = 0\}. \tag{4.1.2}$$

To complete the setting, we define the set of admissible functions where we seek minimizers of (4.1.1). Note that, allowing for a director $\boldsymbol{n} \notin \boldsymbol{H}^1(\Omega)$, one encounters at least two difficulties: On the one hand, it is not clear how to interpret the gradient of $\boldsymbol{n}$ appearing in $E_1[s, \boldsymbol{n}]$. On the other hand, the trace of $\boldsymbol{n}$ on the boundary of Ω is not well-defined, so that one cannot impose Dirichlet conditions on $\boldsymbol{n}$ in the standard way. To cope with these problems, following [4, 54], we introduce the auxiliary variable $\boldsymbol{u} = s\boldsymbol{n}$. Then, the product rule formally yields that

$$\boldsymbol{\nabla u} = \boldsymbol{n} \otimes \nabla s + s\boldsymbol{\nabla n}. \tag{4.1.3}$$

Since $|\boldsymbol{n}| = 1$, the identities $\boldsymbol{\nabla n}^\top \boldsymbol{n} = \boldsymbol{0}$ and $|\boldsymbol{n} \otimes \nabla s| = |\nabla s|$ are valid. It follows that the above decomposition of $\boldsymbol{\nabla u}$ is orthogonal, i.e.,

$$|\boldsymbol{\nabla u}|^2 = |\boldsymbol{n} \otimes \nabla s|^2 + s^2 |\boldsymbol{\nabla n}|^2 = |\nabla s|^2 + s^2 |\boldsymbol{\nabla n}|^2. \tag{4.1.4}$$

In particular, $E_1[s, \boldsymbol{n}]$ can be rewritten in terms of s and $\boldsymbol{u} = s\boldsymbol{n}$ as

$$E_1[s, \boldsymbol{n}] = \widetilde{E}_1[s, \boldsymbol{u}] = \frac{1}{2} \int_\Omega \big((\kappa - 1) |\nabla s|^2 + |\boldsymbol{\nabla u}|^2 \big). \tag{4.1.5}$$

In the latter, the degree of orientation and the auxiliary field are decoupled. In particular, this reveals that, for $(s, \boldsymbol{n})$ such that $E_1[s, \boldsymbol{n}] < \infty$, $\boldsymbol{u} = s\boldsymbol{n} \in \boldsymbol{H}^1(\Omega)$ even though $\boldsymbol{n} \notin \boldsymbol{H}^1(\Omega)$.

We say that a triple $(s, \boldsymbol{n}, \boldsymbol{u})$ satisfies the *structural condition* if

$$-\frac{1}{d-1} < s < 1, \quad |\boldsymbol{n}| = 1, \quad \text{and} \quad \boldsymbol{u} = s\boldsymbol{n} \quad \text{a.e. in } \Omega. \tag{4.1.6}$$

In view of the above discussion, we are therefore led to consider the following admissible class:

$$\mathbb{A} := \big\{ (s, \boldsymbol{n}, \boldsymbol{u}) \in H^1(\Omega) \times \boldsymbol{L}^\infty(\Omega) \times \boldsymbol{H}^1(\Omega) : (s, \boldsymbol{n}, \boldsymbol{u}) \text{ satisfies (4.1.6)} \big\}. \tag{4.1.7}$$

For triples $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}$, it is possible to characterize the gradient of $\boldsymbol{n}$ occurring in $E_1[s, \boldsymbol{n}]$ using a weaker notion of differentiability. To this end, we recall the following definition: We say that $\boldsymbol{n}$ is L^2-differentiable at $x \in \Omega$, and we denote its

L^2-gradient at x by $\boldsymbol{\nabla n}(x)$, if

$$⨍_{B_r(x)} |\boldsymbol{n}(y) - \boldsymbol{n}(x) - \boldsymbol{\nabla n}(x)(y-x)|^2 \,\mathrm{d}y = o(r^2) \quad \text{as } r \to 0.$$

It is well-known that the notion of L^2-differentiability is weaker than the existence of a L^2-integrable weak gradient, in the sense that every H^1-function is L^2-differentiable almost everywhere and its L^2-gradient coincides with the weak gradient; see, e.g., [38, Theorem 6.2].

In the following proposition, we establish that if $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}$, then $\boldsymbol{n}$ is L^2-differentiable and the decomposition (4.1.4) holds almost everywhere outside of the singular set Σ in (4.1.2).

Proposition 4.1.1 (orthogonal decomposition)**.** *Let $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}$. Then, $\boldsymbol{n}$ is L^2-differentiable a.e. in $\Omega \setminus \Sigma$. In particular, its L^2-gradient is given by*

$$\boldsymbol{\nabla n} = s^{-1}(\boldsymbol{\nabla u} - \boldsymbol{n} \otimes \nabla s) \quad \textit{a.e. in } \Omega \setminus \Sigma. \tag{4.1.8}$$

Moreover, the following identity holds

$$|\boldsymbol{\nabla u}|^2 = |\nabla s|^2 + s^2 |\boldsymbol{\nabla n}|^2 \quad \textit{a.e. in } \Omega \setminus \Sigma. \tag{4.1.9}$$

Proof. Since $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}$, we have that $s \in H^1(\Omega)$ and $\boldsymbol{u} = s\boldsymbol{n} \in \boldsymbol{H}^1(\Omega)$. Then, for almost all $x \in \Omega$ (specifically, for all Lebesgue points of $(s, \boldsymbol{u}, \nabla s, \boldsymbol{\nabla u})$), s and $\boldsymbol{u}$ are L^2-differentiable and their L^2-gradients coincide with their respective weak

gradients for a.e. $x \in \Omega$, i.e., as $r \to 0$, it holds that

$$\fint_{B_r(x)} |s(y) - s(x) - \nabla s(x) \cdot (y-x)|^2 \,\mathrm{d}y = o(r^2),$$
$$\fint_{B_r(x)} |\boldsymbol{u}(y) - \boldsymbol{u}(x) - \boldsymbol{\nabla u}(x)(y-x)|^2 \,\mathrm{d}y = o(r^2);$$

see [38, Theorem 6.2]. For almost all $x \in \Omega \setminus \Sigma$ (specifically, for all Lebesgue points of $(s, \boldsymbol{n}, \boldsymbol{u}, \nabla s, \boldsymbol{\nabla u})$ in $x \in \Omega \setminus \Sigma$), in view of the identity (4.1.3), we define the quantity

$$\boldsymbol{\nabla n}(x) := \frac{\boldsymbol{\nabla u}(x) - \boldsymbol{n}(x) \otimes \nabla s(x)}{s(x)}. \tag{4.1.10}$$

Let $r > 0$. It holds that

$$\begin{aligned}
&\fint_{B_r(x)} |\boldsymbol{n}(y) - \boldsymbol{n}(x) - \boldsymbol{\nabla n}(x)(y-x)|^2 \,\mathrm{d}y \\
&\quad \lesssim \frac{1}{s(x)^2} \fint_{B_r(x)} |\boldsymbol{u}(y) - \boldsymbol{u}(x) - \boldsymbol{\nabla u}(x)(y-x)|^2 \mathrm{d}y \\
&\qquad + \frac{1}{s(x)^2} \fint_{B_r(x)} |s(y) - s(x) - \nabla s(x) \cdot (y-x)|^2 |\boldsymbol{n}(y)|^2 \mathrm{d}y \\
&\qquad + \frac{|\nabla s(x)|^2}{s(x)^2} \fint_{B_r(x)} |\boldsymbol{n}(y) - \boldsymbol{n}(x)|^2 |y-x|^2 \mathrm{d}y = o(r^2)
\end{aligned}$$

as $r \to 0$. This shows that $\boldsymbol{\nabla n}(x)$ is the L^2-gradient of $\boldsymbol{n}$ at x. Moreover, (4.1.9) follows from a direct computation. In fact, in view of (4.1.10), there holds that

$$\begin{aligned}
s(x)^2 |\boldsymbol{\nabla n}(x)|^2 &= |\boldsymbol{\nabla u}(x) - \boldsymbol{n}(x) \otimes \nabla s(x)|^2 \\
&= |\boldsymbol{\nabla u}(x)|^2 + |\boldsymbol{n}(x) \otimes \nabla s(x)|^2 - 2\boldsymbol{\nabla u}(x) : [\boldsymbol{n}(x) \otimes \nabla s(x)] \\
&= |\boldsymbol{\nabla u}(x)|^2 - |\nabla s(x)|^2,
\end{aligned}$$

where the last equality follows from the identities

$$|\boldsymbol{n}(x) \otimes \nabla s(x)|^2 = \sum_{i,j=1}^{d} n_i(x)^2 \, \partial_j s(x)^2 = \sum_{j=1}^{d} \partial_j s(x)^2 = |\nabla s(x)|^2$$

and for almost all $x \in \Omega \setminus \Sigma$

$$\begin{aligned}
\boldsymbol{\nabla u}(x) : [\boldsymbol{n}(x) \otimes \nabla s(x)] &= \sum_{i,j=1}^{d} \partial_j u_i(x) \, n_i(x) \, \partial_j s(x) = \frac{1}{s(x)} \sum_{i,j=1}^{d} \partial_j u_i(x) \, u_i(x) \, \partial_j s(x) \\
&= \frac{1}{2s(x)} \sum_{i,j=1}^{d} \partial_j |u_i(x)|^2 \, \partial_j s(x) = \frac{1}{2s(x)} \sum_{j=1}^{d} \partial_j |\boldsymbol{u}(x)|^2 \, \partial_j s(x) \\
&= \frac{1}{2s(x)} \sum_{j=1}^{d} \partial_j \big[s(x)^2 \big] \, \partial_j s(x) = \sum_{j=1}^{d} \partial_j s(x)^2 = |\nabla s(x)|^2.
\end{aligned}$$

This concludes the proof. □

This allows us to give a precise meaning to $E_1[s, \boldsymbol{n}]$ in (4.1.1). Depending on the context, we interpret $\nabla \boldsymbol{n}$ in the sense of L^2-gradient in $\Omega \setminus \Sigma$ and $\int_\Sigma s^2 |\nabla \boldsymbol{n}|^2 = 0$, or we alternatively replace Ω by $\Omega \setminus \Sigma$ as domain of integration or even use the representation $\widetilde{E}_1[s, \boldsymbol{u}]$ of (4.1.5).

Turning to boundary conditions, let $\Gamma_D \subseteq \partial\Omega$ be a relatively open subset of the boundary such that $|\Gamma_D| > 0$, where we aim to impose Dirichlet boundary conditions. These, in the context of LCs, are usually referred to as *strong anchoring* conditions. To this end, given a triple $(g, \boldsymbol{q}, \boldsymbol{r}) \in W^{1,\infty}(\mathbb{R}^3) \times \boldsymbol{L}^\infty(\mathbb{R}^3) \times \boldsymbol{W}^{1,\infty}(\mathbb{R}^3)$ satisfying the structural condition (4.1.6), we consider the following restricted admissible class

that incorporates boundary conditions:

$$\mathbb{A}(g, \boldsymbol{r}) := \big\{(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A} : s|_{\Gamma_D} = g|_{\Gamma_D} \text{ and } \boldsymbol{u}|_{\Gamma_D} = \boldsymbol{r}|_{\Gamma_D}\big\}. \tag{4.1.11}$$

Overall, we are interested in the following constrained minimization problem: Find $(s^*, \boldsymbol{n}^*, \boldsymbol{u}^*) \in \mathbb{A}(g, \boldsymbol{r})$ such that

$$(s^*, \boldsymbol{n}^*, \boldsymbol{u}^*) = \underset{(s,\boldsymbol{n},\boldsymbol{u})\in\mathbb{A}(g,\boldsymbol{r})}{\arg\min} E[s, \boldsymbol{n}]. \tag{4.1.12}$$

To conclude this section, let $\delta_0 > 0$ be sufficiently small. Some of our results below will require the following technical assumptions on the Dirichlet data, namely

$$-\frac{1}{d-1} + \delta_0 \le g(x) \le 1 - \delta_0 \quad \text{for all } x \in \mathbb{R}^d \tag{4.1.13}$$

and

$$g \ge \delta_0 \quad \text{on } \Gamma_D, \tag{4.1.14}$$

and on the double well potential, namely

$$\begin{aligned} &\psi(s) \ge \psi(1-\delta_0) && \text{for all } s \ge 1-\delta_0, \\ &\psi(s) \ge \psi\left(-\frac{1}{d-1} + \delta_0\right) && \text{for all } s \le -\frac{1}{d-1} + \delta_0, \end{aligned} \tag{4.1.15}$$

and ψ in monotone in $(-\frac{1}{d-1}, -\frac{1}{d-1} + \delta_0)$ and $(1-\delta_0, 1)$. Note that (4.1.14) implies that $\boldsymbol{q} = g^{-1}\boldsymbol{r}$ is $\boldsymbol{W}^{1,\infty}$ in a neighborhood of Γ_D and hence $\boldsymbol{n}$ is $\boldsymbol{H}^1$ in a neighborhood of Γ_D, so that in this case one can impose the Dirichlet conditions $\boldsymbol{n}|_{\Gamma_D} = \boldsymbol{q}|_{\Gamma_D}$

directly on $\boldsymbol{n}$. Finally, the property (4.1.15) is consistent with the fact that $\psi(s) \to +\infty$ as $s \to -1/(d-1)$ and $s \to 1$.

4.1.2 Discretization

We assume Ω be a polytopal domain and consider a shape-regular family $\{\mathcal{T}_h\}$ of simplicial meshes of Ω parametrized by the mesh size $h = \max_{K\in\mathcal{T}_h} h_K$, where $h_K = \operatorname{diam}(K)$. We denote by $\mathcal{N}_h$ the set of vertices of $\mathcal{T}_h$. For any $K \in \mathcal{T}_h$, we denote by $\mathbb{P}_1(K)$ the space of first-order polynomials on K and by $\mathcal{N}_h(K)$ the set of vertices of K. We consider the space of $\mathcal{T}_h$-piecewise affine and globally continuous functions

$$V_h := \left\{ v_h \in C^0(\overline{\Omega}) : v_h|_K \in \mathbb{P}_1(K) \text{ for all } K \in \mathcal{T}_h \right\}.$$

Let $\mathbf{V}_h := (V_h)^d$ be the corresponding space of vector-valued polynomials. We denote by I_h both the nodal interpolant $I_h : C^0(\overline{\Omega}) \to V_h$ and its vector-valued counterpart $I_h : \boldsymbol{C}^0(\overline{\Omega}) \to \mathbf{V}_h$.

For $s_h \in V_h$ and $\boldsymbol{n}_h \in \mathbf{V}_h$, let discrete energy be $E^h[s_h, \boldsymbol{n}_h] = E_1^h[s_h, \boldsymbol{n}_h] + E_2^h[s_h]$ with

$$E_1^h[s_h, \boldsymbol{n}_h] := \frac{1}{2}\int_\Omega \left(\kappa|\boldsymbol{n}_h \otimes \nabla s_h|^2 + s_h^2|\boldsymbol{\nabla} \boldsymbol{n}_h|^2\right), \qquad E_2^h[s_h] := \int_\Omega \psi(s_h). \quad (4.1.16)$$

Note that E^h is consistent, in the sense that $E^h[s, \boldsymbol{n}] = E[s, \boldsymbol{n}]$ if $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$.

We say that a triple $(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h) \in V_h \times \mathbf{V}_h \times \mathbf{V}_h$ satisfies the *discrete structural*

condition if

$$-\frac{1}{d-1} < s_h(z) < 1, \quad |\boldsymbol{n}(z)| \geq 1, \quad \text{and} \quad \boldsymbol{u}_h(z) = s_h(z)\boldsymbol{n}_h(z) \quad \text{for all } z \in \mathcal{N}_h. \tag{4.1.17}$$

In (4.1.17), the requirements prescribed by the continuous structural condition (4.1.6) are imposed only at the vertices of the mesh, which is practical. Moreover, the unit-length constraint for the director is relaxed, since $\boldsymbol{n}_h$ may attain also values outside of the unit sphere.

Let $\varepsilon > 0$, $g_h = I_h[g]$, and $\boldsymbol{r}_h = I_h[\boldsymbol{r}]$. We consider the following discrete minimization problem: Find $(s_h^*, \boldsymbol{n}_h^*, \boldsymbol{u}_h^*) \in \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ such that

$$(s_h^*, \boldsymbol{n}_h^*, \boldsymbol{u}_h^*) = \operatorname*{arg\,min}_{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h) \in \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)} E_h[s_h, \boldsymbol{n}_h], \tag{4.1.18}$$

where the discrete restricted admissible class is defined as

$$\begin{aligned}\mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h) := \big\{ &(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h) \in V_h \times \mathbf{V}_h \times \mathbf{V}_h : \\ &(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h) \text{ satisfies (4.1.17)}, \ \|I_h\big[|\boldsymbol{n}_h|^2\big] - 1\|_{L^1(\Omega)} \leq \varepsilon, \\ &s_h(z) = g_h(z), \text{ and } u_h(z) = r_h(z) \text{ for all } z \in \mathcal{N}_h \cap \Gamma_D \big\}. \end{aligned} \tag{4.1.19}$$

4.2 Γ-convergence

In this section, we show that the discrete energy (4.1.16) converges towards the continuous one (4.1.1) in the sense of Γ-convergence.

Theorem 4.2.1 (Γ-convergence)**.** *Suppose that $\varepsilon \to 0$ as $h \to 0$. Then, the following two properties are satisfied:*

(i) *Lim-sup: Let $\Gamma_D = \partial\Omega$. Let the assumptions (4.1.13)–(4.1.15) hold. If $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$, then there exists a sequence $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\} \subset \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ such that $s_h \to s$ in $H^1(\Omega)$, $\boldsymbol{n}_h \to \boldsymbol{n}$ in $\boldsymbol{L}^2(\Omega \setminus \Sigma)$, and $\boldsymbol{u}_h \to \boldsymbol{u}$ in $\boldsymbol{H}^1(\Omega)$, as $h \to 0$, and*

$$E[s, \boldsymbol{n}] \geq \limsup_{h \to 0} E^h[s_h, \boldsymbol{n}_h]. \tag{4.2.1}$$

(ii) *Lim-inf: Let $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\} \subset \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ be a sequence such that $E^h[s_h, \boldsymbol{n}_h] \leq C$ and $\|\boldsymbol{n}_h\|_{\boldsymbol{L}^\infty(\Omega)} \leq C$, where $C \geq 1$ is a constant independent of h. Then, there exist $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$ and a subsequence of $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\}$ (not relabeled) such that $s_h \rightharpoonup s$ in $H^1(\Omega)$, $\boldsymbol{n}_h \to \boldsymbol{n}$ in $\boldsymbol{L}^2(\Omega \setminus \Sigma)$, and $\boldsymbol{u}_h \rightharpoonup \boldsymbol{u}$ in $\boldsymbol{H}^1(\Omega)$ as $h \to 0$, and*

$$E[s, \boldsymbol{n}] \leq \liminf_{h \to 0} E^h[s_h, \boldsymbol{n}_h]. \tag{4.2.2}$$

4.2.1 Lim-sup inequality

We start with two results from [59] that we state without proofs. The first one shows that the degree of orientation s can be truncated near the end points of the domain of definition $(-(d-1)^{-1}, 1)$ of ψ without increasing the energy $E[s, \boldsymbol{n}]$. We refer to [59, Lemma 3.1] for a proof.

Lemma 4.2.1 (truncation of s)**.** *Let the assumptions (4.1.13) and (4.1.15) hold.*

Let $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$. *For all* $0 < \rho \leq \delta_0$, *define for a.e.* $x \in \Omega$

$$s_\rho(x) := \min\left\{1-\rho, \max\left\{-\frac{1}{d-1}+\rho, s(x)\right\}\right\} \quad \textit{and} \quad \boldsymbol{u}_\rho(x) := s_\rho(x)\boldsymbol{n}(x).$$

Then, $(s_\rho, \boldsymbol{n}, \boldsymbol{u}_\rho) \in \mathbb{A}(g, \boldsymbol{r})$ *and* $E_1[s_\rho, \boldsymbol{n}] \leq E_1[s_\rho, \boldsymbol{n}]$, $E_2[s_\rho] \leq E_2[s]$.

A simple consequence of Lemma 4.2.1, based on convergence of the characteristic function $\chi_{\{s_\rho \neq s\}} \to_{\rho\to 0} \chi_\Omega$, is that $\|(s, \boldsymbol{u}) - (s_\rho, \boldsymbol{u}_\rho)\|_{H^1(\Omega)^{1+d}} \to_{\rho\to 0} 0$. The second result is about regularization of admissible functions but preserving the structural condition (4.1.6) and boundary values. This is a rather tricky two-scale process fully discussed in [59, Proposition 3.2].

Lemma 4.2.2 (regularization of functions in $\mathbb{A}(g, \boldsymbol{r})$)**.** *Let the assumptions* (4.1.13) *and* (4.1.14) *hold, and suppose that* $\Gamma_D = \partial\Omega$. *Let* $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$ *and* $\rho \leq \delta_0$ *such that*

$$-\frac{1}{d-1}+\rho \leq s(x) \leq 1-\rho \quad \textit{for a.e. } x \in \Omega.$$

Then, for all $\delta > 0$, *there exists a triple* $(s_\delta, \boldsymbol{n}_\delta, \boldsymbol{u}_\delta) \in \mathbb{A}(g, \boldsymbol{r})$ *such that* $s_\delta \in W^{1,\infty}(\Omega)$ *and* $\boldsymbol{u}_\delta \in \boldsymbol{W}^{1,\infty}(\Omega)$. *Moreover, there holds* $\|(s, \boldsymbol{u}) - (s_\delta, \boldsymbol{u}_\delta)\|_{H^1(\Omega)^{1+d}} \leq \delta$, $\|\boldsymbol{n} - \boldsymbol{n}_\delta\|_{\boldsymbol{L}^2(\Omega\setminus\Sigma)} \leq \delta$, *and*

$$-\frac{1}{d-1}+\rho \leq s_\delta(x) \leq 1-\rho \quad \textit{for all } x \in \Omega.$$

It is well known that the Lagrange interpolation operator $I_h : C(\overline{\Omega}) \to V_h$ is not stable in $H^1(\Omega)$ unless $d = 1$. We exploit stability in $L^\infty(\Omega)$ to derive stability

in $W^{1,p}(\Omega)$ for $p > d$.

Lemma 4.2.3 ($W^{1,p}$-stability of Lagrange interpolant)**.** *Let $v \in W^{1,p}(\Omega)$ for $d < p \leq \infty$. Then*

$$\|\nabla I_h v\|_{L^p(K)} \lesssim \|\nabla v\|_{L^p(K)} \quad \textit{for all } K \in \mathcal{T}_h. \tag{4.2.3}$$

Proof. Let $K \in \mathcal{T}_h$ be an arbitrary element and let $\overline{v}_K = \fint_K v$. An inverse estimate gives

$$\|\nabla I_h v\|_{L^p(K)}^p \leq |K| \, \|\nabla I_h (v - \overline{v}_K)\|_{L^\infty(K)}^p \lesssim h_K^{d-p} \|v - \overline{v}_K\|_{L^\infty(K)}^p.$$

The Bramble-Hilbert estimate yields $\|v - \overline{v}_K\|_{L^\infty(K)} \lesssim h_K^{1-d/p} \|\nabla v\|_{L^p(K)}$ and ends the proof. □

Applying a standard density argument in $W^{1,p}(\Omega)$, for $d < p < \infty$, we deduce

$$\lim_{h \to 0} \|\nabla (v - I_h v)\|_{L^p(\Omega)} = 0 \quad \text{for all } \ v \in W^{1,p}(\Omega). \tag{4.2.4}$$

We have collected all the ingredients to show the existence of a recovery sequence.

Proof of Theorem 4.2.1(i)**.** For the sake of clarity, we decompose the proof into seven steps.

Step 1: Setup. Let $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$. For all $k \in \mathbb{N}$ such that $1/k \leq \delta_0$, let $0 < \delta_k \leq 1/k$ be sufficiently small. Applying successively Lemma 4.2.1 (with $\rho = 1/k$) and Lemma 4.2.2 (with $\delta = \delta_k$), we obtain $(s_k, \boldsymbol{n}_k, \boldsymbol{u}_k) \in \mathbb{A}(g, \boldsymbol{r})$ satisfying

$(s_k, \boldsymbol{u}_k) \in [W^{1,\infty}(\Omega)]^{1+d}$ and $-1/(d-1) + 1/k \leq s_k \leq 1 - 1/k$ in Ω for all k. Moreover, we have that

$$\|(s, \boldsymbol{u}) - (s_k, \boldsymbol{u}_k)\|_{H^1(\Omega)^{1+d}} \to 0, \qquad \|\boldsymbol{n} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Omega\setminus\Sigma)} \to 0.$$

Since $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$, Proposition 4.1.1 guarantees that $\boldsymbol{n}$ is L^2-differentiable a.e. in $\Omega\setminus\Sigma$, with its L^2-gradient given by (4.1.8) and that the identity (4.1.9) holds. The same result is valid for $\boldsymbol{n}_k$ a.e. in $\Omega \setminus \Sigma_k$, where $\Sigma_k := \{x \in \Omega : s_k(x) = 0\}$.

Let $s_{k,h} := I_h[s_k]$ and $\boldsymbol{u}_{k,h} := I_h[\boldsymbol{u}_k]$. Let $\boldsymbol{n}_{k,h} \in \mathbf{V}_h$ be defined, for all $z \in \mathcal{N}_h$, as

$$\boldsymbol{n}_{k,h}(z) := \begin{cases} \boldsymbol{u}_{k,h}(z)/s_{k,h}(z) = \boldsymbol{u}_k(z)/s_k(z) & \text{if } z \in \Omega \setminus \Sigma_k, \\ \text{an arbitrary unit vector} & \text{if } z \in \Sigma_k. \end{cases}$$

Note that, by construction, $(s_{k,h}, \boldsymbol{n}_{k,h}, \boldsymbol{u}_{k,h})$ satisfies the discrete structural condition (4.1.17), and $\|\boldsymbol{n}_{k,h}\|_{L^\infty(\Omega)} \leq C$. Moreover, since $0 = \|I_h\big[|\boldsymbol{n}_{k,h}|^2\big] - 1\|_{L^1(\Omega)} \leq \varepsilon$ as well as $s_{k,h}(z) = g_h(z)$ and $\boldsymbol{u}_{k,h}(z) = \boldsymbol{r}_h(z)$ for all $z \in \mathcal{N}_h \cap \Gamma_D$, we deduce $(s_{k,h}, \boldsymbol{n}_{k,h}, \boldsymbol{u}_{k,h}) \in \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$.

Given $\delta > 0$, we consider the sets

$$\Sigma_{k,\delta} := \{x \in \Omega : \ |s_k(x)| \leq \delta\} \quad \text{and} \quad \Omega^h_{k,\delta} := \bigcup \{K : \ K \in \mathcal{T}_h, K \cap \Sigma_{k,\delta} = \emptyset\}.$$

Note that, by construction, there holds $\Omega^h_{k,\delta} \subset \Omega \setminus \Sigma_{k,\delta}$; see Figure 4.1.

Let $K \in \mathcal{T}_h$ such that $K \cap \Sigma_{k,\delta} \neq \emptyset$. In particular, there exists $x_0 \in K \cap \Sigma_{k,\delta}$.

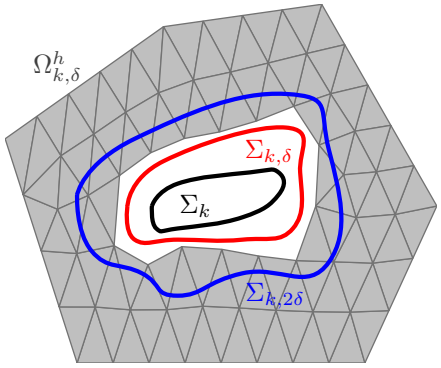

Figure 4.1: A schematic illustration of the mutual relations of the sets defined in Step 1 of the proof of Theorem 4.2.1(i). Note that the set $\Sigma_k \subset \Omega$ is closed, as it is the preimage of a closed set with respect to the continuous function s_k, but it might be more topologically complicated than in the picture.

For $x_1 \in K$ arbitrary, Lipschitz continuity of s_k yields

$$|s_k(x_1)| \leq |s_k(x_0)| + |s_k(x_1) - s_k(x_0)| \leq \delta + C_k h.$$

In particular, $\Omega \setminus \Omega^h_{k,\delta} \subset \Sigma_{k,2\delta}$ provided h is sufficiently small so that $C_k h \leq \delta$. We refer again to Figure 4.1.

Now, for any $x \in \Omega^h_{k,\delta}$, we infer that

$$|s_k(x) - s_{k,h}(x)| \leq \|s_k - s_{k,h}\|_{L^\infty(\Omega^h_{k,\delta})} = \|s_k - I_h[s_k]\|_{L^\infty(\Omega^h_{k,\delta})} \lesssim h\|\nabla s_k\|_{L^\infty(\Omega^h_{k,\delta})},$$

whence

$$|s_{k,h}(x)| \geq |s_k(x)| - |s_k(x) - s_{k,h}(x)| > \delta - Ch\|\nabla s_k\|_{L^\infty(\Omega^h_{k,\delta})} > \delta/2$$

provided the mesh size h is chosen to be sufficiently small. Hence, for those h, we can define $\widetilde{\boldsymbol{n}}_k := \boldsymbol{u}_{k,h}/s_{k,h}$ in $\Omega^h_{k,\delta}$. Note that, by definition, the relation $\boldsymbol{n}_{k,h} = I_h[\widetilde{\boldsymbol{n}}_k]$

in $\Omega^h_{k,\delta}$ holds.

To conclude this step, we observe that the L^2-gradient $\boldsymbol{\nabla} \boldsymbol{n}_k$ of $\boldsymbol{n}_k$ exists a.e. in $\Omega \setminus \Sigma_k$ and

$$\int_{\Omega^h_{k,\delta}} |\boldsymbol{\nabla} \boldsymbol{n}_k - \boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 \lesssim \int_{\Omega^h_{k,\delta}} |\boldsymbol{\nabla} \boldsymbol{n}_k - \boldsymbol{\nabla} \widetilde{\boldsymbol{n}}_k|^2 + \int_{\Omega^h_{k,\delta}} |\boldsymbol{\nabla} \widetilde{\boldsymbol{n}}_k - \boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2, \tag{4.2.5}$$

where $\boldsymbol{\nabla} \widetilde{\boldsymbol{n}}_k$ and $\boldsymbol{\nabla} \boldsymbol{n}_{k,h}$ denote the weak gradients of $\widetilde{\boldsymbol{n}}_k$ and $\boldsymbol{n}_{k,h}$, respectively, which coincide elementwise with their classical gradients in $\Omega^h_{k,\delta}$. In the following steps, we will show that, for fixed $k \in \mathbb{N}$ and $\delta > 0$, both two terms on the right-hand side of (4.2.5) converge to 0 as $h \to 0$.

Step 2: Proof of $\lim_{h\to 0} \int_{\Omega^h_{k,\delta}} |\widetilde{\boldsymbol{n}}_k - \boldsymbol{n}_{k,h}|^2 + |\boldsymbol{\nabla} \widetilde{\boldsymbol{n}}_k - \boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 = 0$. Since $\boldsymbol{n}_{k,h} = I_h[\widetilde{\boldsymbol{n}}_k]$ in $\Omega^h_{k,\delta}$, a classical local interpolation estimate yields that

$$\int_{\Omega^h_{k,\delta}} |\boldsymbol{\nabla} \widetilde{\boldsymbol{n}}_k - \boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 = \sum_{\substack{K \in \mathcal{T}_h \\ K \cap \Sigma_{k,\delta} = \emptyset}} \int_K \big|\boldsymbol{\nabla} (\widetilde{\boldsymbol{n}}_k - I_h[\widetilde{\boldsymbol{n}}_k])\big|^2 \lesssim \sum_{\substack{K \in \mathcal{T}_h \\ K \cap \Sigma_{k,\delta} = \emptyset}} h_K^2 \big\| D^2 \widetilde{\boldsymbol{n}}_k \big\|^2_{\boldsymbol{L}^2(K)}.$$

Moreover, in view of $\widetilde{\boldsymbol{n}}_k = \boldsymbol{u}_{k,h}/s_{k,h}$ in $\Omega^h_{k,\delta}$, explicit computations reveal that

$$\partial_i \widetilde{\boldsymbol{n}}_k = s_{k,h}^{-1}\, \partial_i \boldsymbol{u}_{k,h} - s_{k,h}^{-2}\, \partial_i s_{k,h}\, \boldsymbol{u}_{k,h} = s_{k,h}^{-1} \big(\partial_i \boldsymbol{u}_{k,h} - \partial_i s_{k,h}\, \widetilde{\boldsymbol{n}}_k \big),$$

$$\partial_j \partial_i \widetilde{\boldsymbol{n}}_k = s_{k,h}^{-1} \big(s_{k,h}^{-1}\, \partial_j s_{k,h}\, \partial_i s_{k,h}\, \widetilde{\boldsymbol{n}}_k - \partial_i s_h\, \partial_j \widetilde{\boldsymbol{n}}_k - s_{k,h}^{-1}\, \partial_j s_{k,h}\, \partial_i \boldsymbol{u}_{k,h} \big),$$

for all $1 \leq i, j \leq d$. Several applications of the generalized Hölder inequality, in

conjunction with the lower bound $|s_{k,h}| > \delta/2$ in $\Omega^h_{k,\delta}$, thus yield

$$\begin{aligned}\big\|D^2\widetilde{\boldsymbol{n}}_k\big\|_{\boldsymbol{L}^2(K)} &\lesssim \delta^{-3}\|\nabla s_{k,h}\|^2_{L^8(K)}\|\boldsymbol{u}_{k,h}\|_{L^4(K)}\\ &\quad+\delta^{-1}\|\nabla s_{k,h}\|_{L^4(K)}\big(\delta^{-1}\|\boldsymbol{\nabla}\boldsymbol{u}_{k,h}\|_{L^4(K)}+\delta^{-2}\|\boldsymbol{u}_{k,h}\|_{L^8(K)}\|\nabla s_{k,h}\|_{L^8(K)}\big)\\ &\quad+\delta^{-2}\|\nabla s_{k,h}\|_{L^4(K)}\|\boldsymbol{\nabla}\boldsymbol{u}_{k,h}\|_{L^4(K)}.\end{aligned}$$

In view of (4.2.3), $s_{k,h}$ (resp., $\boldsymbol{u}_{k,h}$) is uniformly bounded in $W^{1,p}(\Omega)$ (resp., $\boldsymbol{W}^{1,p}(\Omega)$) when $d < p \leq \infty$. Altogether, we thus obtain the desired estimate

$$\int_{\Omega^h_{k,\delta}}|\boldsymbol{\nabla}\widetilde{\boldsymbol{n}}_k-\boldsymbol{\nabla}\boldsymbol{n}_{k,h}|^2+\sum_{\substack{K\in\mathcal{T}_h\\ K\cap\Sigma_{k,\delta}=\emptyset}}h_K^{-2}\int_K|\widetilde{\boldsymbol{n}}_k-\boldsymbol{n}_{k,h}|^2\lesssim\sum_{\substack{K\in\mathcal{T}_h\\ K\cap\Sigma_{k,\delta}=\emptyset}}h_K^2\big\|D^2\widetilde{\boldsymbol{n}}_k\big\|^2_{\boldsymbol{L}^2(K)}\lesssim h^2.$$

Step 3: Proof of $\lim_{h\to 0}\int_{\Omega^h_{k,\delta}}|\boldsymbol{n}_k-\widetilde{\boldsymbol{n}}_k|^2+|\boldsymbol{\nabla}\boldsymbol{n}_k-\boldsymbol{\nabla}\widetilde{\boldsymbol{n}}_k|^2=0$. We first observe that

$$\begin{aligned}\|\widetilde{\boldsymbol{n}}_k-\boldsymbol{n}_k\|_{\boldsymbol{L}^q(\Omega^h_{k,\delta})}&=\|s_{k,h}^{-1}\boldsymbol{u}_{k,h}-s_k^{-1}\boldsymbol{u}_k\|_{\boldsymbol{L}^q(\Omega^h_{k,\delta})}\\ &\leq\delta^{-2}\|s_k-s_{k,h}\|_{L^q(\Omega^h_{k,\delta})}\|\boldsymbol{u}_{k,h}\|_{\boldsymbol{L}^\infty(\Omega^h_{k,\delta})}+\delta^{-1}\|\boldsymbol{u}_{k,h}-\boldsymbol{u}_k\|_{\boldsymbol{L}^q(\Omega^h_{k,\delta})},\end{aligned}$$

for all $q \geq 1$. This shows, in view of (4.2.4), that $\|\widetilde{\boldsymbol{n}}_k-\boldsymbol{n}_k\|_{\boldsymbol{L}^q(\Omega^h_{k,\delta})}\to 0$ as $h\to 0$ for $d < q < \infty$. To deal with the gradient part, we resort to available expressions of $\boldsymbol{\nabla}\boldsymbol{n}_k$ and $\boldsymbol{\nabla}\widetilde{\boldsymbol{n}}_k$ to write

$$\begin{aligned}\int_{\Omega^h_{k,\delta}}|\boldsymbol{\nabla}\boldsymbol{n}_k-\boldsymbol{\nabla}\widetilde{\boldsymbol{n}}_k|^2&=\int_{\Omega^h_{k,\delta}}|s_k^{-1}(\boldsymbol{\nabla}\boldsymbol{u}_k-\boldsymbol{n}_k\otimes\nabla s_k)-s_{k,h}^{-1}(\boldsymbol{\nabla}\boldsymbol{u}_{k,h}-\widetilde{\boldsymbol{n}}_k\otimes\nabla s_{k,h})|^2\\ &=T_1+T_2+T_3\end{aligned}$$

where

$$T_1 := \int_{\Omega^h_{k,\delta}} |s_{k,h}^{-1}(\boldsymbol{\nabla} \boldsymbol{u}_k - \boldsymbol{\nabla} \boldsymbol{u}_{k,h})|^2$$

$$T_2 := \int_{\Omega^h_{k,\delta}} |s_{k,h}^{-1}(\widetilde{\boldsymbol{n}}_k \otimes \nabla s_{k,h} - \boldsymbol{n}_k \otimes \nabla s_k)|^2$$

$$T_3 := \int_{\Omega^h_{k,\delta}} |(s_k^{-1} - s_{k,h}^{-1})(\boldsymbol{\nabla} \boldsymbol{u}_k - \boldsymbol{n}_k \otimes \nabla s_k)|^2.$$

Recalling again $|s_k|, |s_{k,h}| > \delta/2$ in $\Omega^h_{k,\delta}$, as well as (4.2.4), the asserted estimate follows from

$$T_1 \lesssim \delta^{-2} \|\boldsymbol{\nabla}(\boldsymbol{u}_k - I_h \boldsymbol{u}_k)\|^2_{\boldsymbol{L}^2(\Omega)},$$

$$T_2 \lesssim \delta^{-2} \|\nabla s_{k,h}\|^2_{\boldsymbol{L}^4(\Omega^h_{k,\delta})} \|\widetilde{\boldsymbol{n}}_k - \boldsymbol{n}_k\|^2_{\boldsymbol{L}^4(\Omega^h_{k,\delta})} + \delta^{-2} \|\boldsymbol{n}_k\|^2_{\boldsymbol{L}^4(\Omega^h_{k,\delta})} \|\nabla(s_k - I_h s_k)\|^2_{\boldsymbol{L}^4(\Omega^h_{k,\delta})},$$

$$T_3 \lesssim \delta^{-4} \|s_k - I_h s_k\|^2_{L^4(\Omega)} \|\boldsymbol{\nabla} \boldsymbol{u}_k - \boldsymbol{n}_k \otimes \nabla s_k\|^2_{\boldsymbol{L}^4(\Omega)}.$$

Step 4: Proof of $\lim_{h\to 0} \int_\Omega s_{k,h}^2 |\boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 = \int_{\Omega \setminus \Sigma_k} s_k^2 |\boldsymbol{\nabla} \boldsymbol{n}_k|^2$. Combining Steps 2 and 3 gives

$$\lim_{h\to 0} \int_{\Omega^h_{k,\delta}} |\boldsymbol{\nabla} \boldsymbol{n}_k - \boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 = 0. \tag{4.2.6}$$

In order to exploit this property, we split the integral under consideration as

$$\int_\Omega s_{k,h}^2 |\boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 = \int_{\Omega^h_{k,\delta}} s_{k,h}^2 |\boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 + \int_{\Omega \setminus \Omega^h_{k,\delta}} s_{k,h}^2 |\boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2. \tag{4.2.7}$$

The fact that $s_{k,h} \to s_k$ strongly in $L^p(\Omega)$ as $h \to 0$ for $d < p < \infty$, according to (4.2.4), together with $s_{k,h} \in L^\infty(\Omega)$ uniformly in h, $\boldsymbol{\nabla} \boldsymbol{n}_k \in L^\infty(\Omega \setminus \Sigma_{k,\delta})$ and (4.2.6),

yields

$$\lim_{h\to 0}\left|\int_{\Omega^h_{k,\delta}} s^2_{k,h}|\boldsymbol{\nabla}\boldsymbol{n}_{k,h}|^2 - \int_{\Omega^h_{k,\delta}} s^2_k|\boldsymbol{\nabla}\boldsymbol{n}_k|^2\right| = 0.$$

Since $\Omega\setminus\Sigma_{k,2\delta}\subset\Omega^h_{k,\delta}\subset\Omega\setminus\Sigma_{k,\delta}$, we deduce

$$\lim_{\delta\to 0}\lim_{h\to 0}\int_{\Omega^h_{k,\delta}} s^2_{k,h}|\boldsymbol{\nabla}\boldsymbol{n}_{k,h}|^2 = \int_{\Omega\setminus\Sigma_k} s^2_k|\boldsymbol{\nabla}\boldsymbol{n}_k|^2.$$

Now, we consider the second term on the right-hand side of (4.2.7). Since $\Omega\setminus\Omega_{k,\delta}\subset\Sigma_{k,2\delta}$ and $s_{k,h}\boldsymbol{\nabla}\boldsymbol{n}_{k,h} = \boldsymbol{\nabla}(s_{k,h}\boldsymbol{n}_{k,h}) - \boldsymbol{n}_{k,h}\otimes\nabla s_{k,h}$, using $\boldsymbol{u}_{k,h} = I_h(s_{k,h}\boldsymbol{n}_{k,h})$, we see that

$$\begin{aligned}\int_{\Omega\setminus\Omega^h_{k,\delta}} s^2_{k,h}|\boldsymbol{\nabla}\boldsymbol{n}_{k,h}|^2 &\lesssim \int_{\Sigma_{k,2\delta}}|\boldsymbol{\nabla}(s_{k,h}\boldsymbol{n}_{k,h})|^2 + \int_{\Sigma_{k,2\delta}}|\boldsymbol{n}_{k,h}\otimes\nabla s_{k,h}|^2\\ &\le \int_{\Sigma_{k,2\delta}}|\boldsymbol{\nabla}(s_{k,h}\boldsymbol{n}_{k,h}) - \boldsymbol{\nabla}I_h(s_{k,h}\boldsymbol{n}_{k,h})|^2 + \int_{\Sigma_{k,2\delta}}|\boldsymbol{\nabla}\boldsymbol{u}_{k,h}|^2\\ &\quad + \int_{\Sigma_{k,2\delta}}|\nabla s_{k,h}|^2.\end{aligned}$$

Combining an interpolation estimate with the fact that $s_{k,h}$ and $\boldsymbol{n}_{k,h}$ are piecewise affine, and exploiting an inverse estimate to bound $\|\boldsymbol{\nabla}\boldsymbol{n}_{k,h}\|_{\boldsymbol{L}^\infty(K)}$ in terms of $\|\boldsymbol{n}_{k,h}\|_{\boldsymbol{L}^\infty(K)}\le C$, yields

$$\begin{aligned}\int_{\Sigma_{k,2\delta}}|\boldsymbol{\nabla}(s_{k,h}\boldsymbol{n}_{k,h}) - \boldsymbol{\nabla}I_h(s_{k,h}\boldsymbol{n}_{k,h})|^2 &\lesssim \sum_{\substack{K\in\mathcal{T}_h\\ K\cap\Sigma_{k,2\delta}\neq\emptyset}} h_K^2\|D^2(s_{k,h}\boldsymbol{n}_{k,h})\|^2_{\boldsymbol{L}^2(K)}\\ \lesssim \sum_{\substack{K\in\mathcal{T}_h\\ K\cap\Sigma_{k,2\delta}\neq\emptyset}} h_K^2\|\nabla s_{k,h}\|^2_{\boldsymbol{L}^2(K)}\|\boldsymbol{\nabla}\boldsymbol{n}_{k,h}\|^2_{\boldsymbol{L}^\infty(K)} &\lesssim \sum_{\substack{K\in\mathcal{T}_h\\ K\cap\Sigma_{k,2\delta}\neq\emptyset}}\|\nabla s_{k,h}\|^2_{\boldsymbol{L}^2(K)}.\end{aligned}$$

Using the $W^{1,p}$-stability (4.2.3) of the nodal interpolant with $p > d$ for elements

$K \cap \Sigma_{k,2\delta} \neq \emptyset$, which thus satisfy $K \subset \Sigma_{k,3\delta}$ when h is sufficiently small by Lipschitz continuity of s_k, we end up with the following as $\delta \to 0$

$$\int_{\Omega \setminus \Omega^h_{k,\delta}} s^2_{k,h} |\boldsymbol{\nabla} \boldsymbol{n}_{k,h}|^2 \lesssim \|\boldsymbol{\nabla} \boldsymbol{u}_k\|^2_{\boldsymbol{L}^p(\Sigma_{k,3\delta})} + \|\nabla s_k\|^2_{\boldsymbol{L}^p(\Sigma_{k,3\delta})}$$

$$\to \|\boldsymbol{\nabla} \boldsymbol{u}_k\|^2_{\boldsymbol{L}^p(\Sigma_k)} + \|\nabla s_k\|^2_{\boldsymbol{L}^p(\Sigma_k)} = 0.$$

Step 5: Proof of $\lim_{h\to 0} \int_\Omega |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2 = \int_\Omega |\nabla s_k|^2$. We split the integral as

$$\int_\Omega |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2 = \int_{\Omega^h_{k,\delta}} |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2 + \int_{\Omega \setminus \Omega^h_{k,\delta}} |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2.$$

Exploiting the identity $\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h} - \boldsymbol{n}_k \otimes \nabla s_k = (\boldsymbol{n}_{k,h} - \boldsymbol{n}_k) \otimes \nabla s_k + \boldsymbol{n}_{k,h} \otimes (\nabla s_{k,h} - \nabla s_k)$, and using the convergence results for $s_{k,h}$ and $\boldsymbol{n}_{k,h}$ in $\Omega^h_{k,\delta}$ from Steps 1-3, we readily see that

$$\lim_{\delta \to 0} \lim_{h \to 0} \int_{\Omega^h_{k,\delta}} |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2 = \int_{\Omega \setminus \Sigma_k} |\boldsymbol{n}_k \otimes \nabla s_k|^2 = \int_\Omega |\nabla s_k|^2.$$

Moreover, employing $\Omega \setminus \Omega_{k,\delta} \subset \Sigma_{k,2\delta}$ together with (4.2.3) implies

$$\int_{\Omega \setminus \Omega^h_{k,\delta}} |\boldsymbol{n}_{k,h} \otimes \nabla s_{k,h}|^2 \lesssim \|\nabla I_h s_k\|^2_{\boldsymbol{L}^2(\Sigma_{k,2\delta})} \lesssim \|\nabla I_h s_k\|^2_{\boldsymbol{L}^p(\Sigma_{k,2\delta})} \lesssim \|\nabla s_k\|^2_{\boldsymbol{L}^p(\Sigma_{k,3\delta})}.$$

Finally, taking $\delta \to 0$ yields the desired limit.

Step 6: Convergence of $\{s_{k,h}\}$, $\{\boldsymbol{n}_{k,h}\}$, and $\{\boldsymbol{u}_{k,h}\}$. The triangle inequality

gives

$$\|s_{k,h} - s\|_{H^1(\Omega)} \leq \|s_{k,h} - s_k\|_{H^1(\Omega)} + \|s_k - s\|_{H^1(\Omega)} \to 0 \quad \text{as } h \to 0 \text{ and } k \to \infty.$$

Likewise, $\boldsymbol{u}_{k,h} \to \boldsymbol{u}$ in $\boldsymbol{H}^1(\Omega)$ as $h \to 0$ and $k \to \infty$. Turning to $\boldsymbol{n}$, we observe that

$$\|\boldsymbol{n}_{k,h} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Omega\setminus\Sigma)} \lesssim \|\boldsymbol{n}_{k,h} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Omega^h_{k,\delta})} + \|\boldsymbol{n}_{k,h} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Sigma_{k,2\delta}\setminus\Sigma)},$$

and $\|\boldsymbol{n}_{k,h} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Omega^h_{k,\delta})} \to 0$ as $h \to 0$ from Steps 2–3. Instead, for the second term we have

$$\|\boldsymbol{n}_{k,h} - \boldsymbol{n}_k\|_{\boldsymbol{L}^2(\Sigma_{k,2\delta}\setminus\Sigma)} \leq 2|\Sigma_{k,2\delta} \setminus \Sigma|^{1/2} \to 0 \quad \text{as } \delta \to 0 \text{ and } k \to \infty.$$

The convergence of $\boldsymbol{n}_{k,h}$ to $\boldsymbol{n}$ in $\boldsymbol{L}^2(\Omega \setminus \Sigma)$ then follows from the triangle inequality.

Step 7: Convergence of energy. The previous steps yield

$$\lim_{h\to 0} E_1^h[s_{k,h}, \boldsymbol{n}_{k,h}] = E_1[s_k, \boldsymbol{n}_k] = \frac{1}{2}\int_{\Omega\setminus\Sigma_k} \kappa|\nabla s_k|^2 + s_k^2|\nabla \boldsymbol{n}_k|^2.$$

To prove that $E_1[s_k, \boldsymbol{n}_k] \to E_1[s, \boldsymbol{n}]$ as $k \to \infty$ we resort to (4.1.5), namely

$$\begin{aligned} E_1[s_k, \boldsymbol{n}_k] = \widetilde{E}_1[s_k, \boldsymbol{u}_k] &= \frac{1}{2}\int_\Omega (\kappa - 1)|\nabla s_k|^2 + |\nabla \boldsymbol{u}_k|^2 \\ &\to \frac{1}{2}\int_\Omega (\kappa - 1)|\nabla s|^2 + |\nabla \boldsymbol{u}|^2 = E_1[s, \boldsymbol{n}]. \end{aligned}$$

We now deal with E_2. Since $-1/(d-1) + 1/k \leq s_k \leq 1 - 1/k$ in Ω, assump-

tion (4.1.15) guarantees that $0 \le \psi(s_{k,h}) \le \max\{\psi(-1/(d-1)+1/k), \psi(1-1/k)\}$. Hence, the dominated convergence theorem implies that

$$\lim_{h\to 0} E_2^h[s_{k,h}] = \lim_{h\to 0} \int_\Omega \psi(I_h s_k) = \int_\Omega \lim_{h\to 0} \psi(I_h s_k) = \int_\Omega \psi(s_k) = E_2[s_k].$$

Moreover, the monotonicity of ψ in $(-\frac{1}{d-1}, -\frac{1}{d-1} + \delta_0)$ and $(1-\delta_0, 1)$ translates into $\psi(s_k) \ge 0$ increasing and converging pointwise to $\psi(s)$, whence the monotone convergence theorem gives

$$E_2[s_k] = \int_\Omega \psi(s_k) \to \int_\Omega \psi(s) = E_2[s].$$

Consequently, the sequence $(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h) := (s_{k,h}, \boldsymbol{n}_{k,h}, \boldsymbol{u}_{k,h}) \in \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ for k sufficiently large depending on h converges to $(s, \boldsymbol{n}, \boldsymbol{u})$ in $H^1(\Omega) \times \boldsymbol{L}^2(\Omega \setminus \Sigma) \times \boldsymbol{H}^1(\Omega)$ as $h \to 0$ and satisfies

$$\lim_{h\to 0} E^h[s_h, \boldsymbol{n}_h] = E[s, \boldsymbol{n}].$$

This implies the lim-sup inequality (4.2.1) and concludes the proof. □

4.2.2 Lim-inf inequality

To show the lim-inf inequality, we first prove that admissible discrete pairs $(s_h, \boldsymbol{n}_h)$ with uniformly bounded energy are uniformly bounded in H^1. In constrast to [59], we do not need to assume that $\mathcal{T}_h$ is weakly acute.

Lemma 4.2.4 (coercivity)**.** *Let* $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\} \subset V_h \times \mathbf{V}_h \times \mathbf{V}_h$ *satisfy* $\boldsymbol{u}_h = I_h[s_h \boldsymbol{n}_h]$ *and* $|\boldsymbol{n}_h(z)| \ge 1$ *for all* $z \in \mathcal{N}_h$*. Then, there exists a constant* $C > 0$ *depending only*

on the shape-regularity of $\{\mathcal{T}_h\}$ *and* κ *such that*

$$C \max\left\{ \|\boldsymbol{\nabla} \boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(\Omega)}, \|\boldsymbol{\nabla}(s_h \boldsymbol{n}_h)\|^2_{\boldsymbol{L}^2(\Omega)}, \|\nabla s_h\|^2_{\boldsymbol{L}^2(\Omega)} \right\} \leq E_1^h[s_h, \boldsymbol{n}_h].$$

Proof. Since $\|\boldsymbol{n}_h\|_{\boldsymbol{L}^\infty(K)} \geq 1$ for all $K \in \mathcal{T}_h$ and ∇s_h is piecewise constant, it holds that

$$\begin{aligned} \|\nabla s_h\|^2_{\boldsymbol{L}^2(\Omega)} &\leq \sum_{K \in \mathcal{T}_h} \|\boldsymbol{n}_h\|^2_{\boldsymbol{L}^\infty(K)} \|\nabla s_h\|^2_{\boldsymbol{L}^2(K)} = \sum_{K \in \mathcal{T}_h} |K| \, \|\boldsymbol{n}_h\|^2_{\boldsymbol{L}^\infty(K)} |\nabla s_h|_K|^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} |\nabla s_h|_K|^2 \|\boldsymbol{n}_h\|^2_{\boldsymbol{L}^2(K)} = \|\boldsymbol{n}_h \otimes \nabla s_h\|^2_{\boldsymbol{L}^2(\Omega)} \leq \frac{2}{\kappa} E_1^h[s_h, \boldsymbol{n}_h], \end{aligned}$$

where the hidden multiplicative constant depends only on the shape-regularity of $\{\mathcal{T}_h\}$. Let $\widetilde{\boldsymbol{u}}_h = s_h \boldsymbol{n}_h$ and use (4.2.3) for $p > d$ in conjunction with an inverse estimate to obtain for all $K \in \mathcal{T}_h$

$$\|\nabla I_h \widetilde{\boldsymbol{u}}_h\|_{\boldsymbol{L}^2(K)} \lesssim |K|^{\frac{p-2}{2p}} \|\nabla I_h \widetilde{\boldsymbol{u}}_h\|_{\boldsymbol{L}^p(K)} \lesssim |K|^{\frac{p-2}{2p}} \|\nabla \widetilde{\boldsymbol{u}}_h\|_{\boldsymbol{L}^p(K)} \lesssim \|\nabla \widetilde{\boldsymbol{u}}_h\|_{\boldsymbol{L}^2(K)}$$

Consequently, for $\boldsymbol{u}_h = I_h \widetilde{\boldsymbol{u}}_h$ we deduce

$$\|\boldsymbol{\nabla} \boldsymbol{u}_h\|^2_{\boldsymbol{L}^2(\Omega)} \lesssim \|\boldsymbol{\nabla} \widetilde{\boldsymbol{u}}_h\|^2_{\boldsymbol{L}^2(\Omega)} \lesssim \|\boldsymbol{n}_h \otimes \nabla s_h\|^2_{\boldsymbol{L}^2(\Omega)} + \|s_h \boldsymbol{\nabla} \boldsymbol{n}_h\|^2_{\boldsymbol{L}^2(\Omega)} \lesssim E_1^h[s_h, \boldsymbol{n}_h].$$

This completes the proof. □

We are now ready to extract convergent subsequences and characterize their limits.

Lemma 4.2.5 (characterization of limits)**.** *Let* $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\} \subset \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ *be a sequence such that* $E_1^h[s_h, \boldsymbol{n}_h] \leq C$ *and* $\|\boldsymbol{n}_h\|_{\boldsymbol{L}^\infty(\Omega)} \leq C$*, where* $C > 0$ *is a constant independent of* h*. Then, there exist a triple* $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$ *and a subsequence (not relabeled) of* $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\}$ *satisfying the following properties:*

- *As* $h \to 0$*,* $(s_h, \boldsymbol{u}_h, s_h\boldsymbol{n}_h)$ *converges towards* $(s, \boldsymbol{u}, \boldsymbol{u})$ *weakly in* $H^1(\Omega) \times \boldsymbol{H}^1(\Omega) \times \boldsymbol{H}^1(\Omega)$*, strongly in* $L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times \boldsymbol{L}^2(\Omega)$*, and pointwise a.e. in* Ω*;*
- $\boldsymbol{n}_h$ *converges towards* $\boldsymbol{n}$ *strongly in* $\boldsymbol{L}^2(\Omega \setminus \Sigma)$ *and pointwise a.e. in* $\Omega \setminus \Sigma$ *as* $h \to 0$ *and* $\varepsilon \to 0$*;*
- $\boldsymbol{n}$ *is* L^2*-differentiable a.e. in* $\Omega \setminus \Sigma$ *and the orthogonal decomposition* $|\boldsymbol{\nabla} \boldsymbol{u}|^2 = |\nabla s|^2 + s^2 |\boldsymbol{\nabla} \boldsymbol{n}|^2$ *is valid a.e. in* $\Omega \setminus \Sigma$*,*

where $\Sigma \subset \Omega$ *is given by* (4.1.2)*.*

Proof. For the sake of clarity, we divide the proof into 3 steps.

Step 1: Convergence of $\{s_h\}$*,* $\{\boldsymbol{u}_h\}$*, and* $\{s_h\boldsymbol{n}_h\}$*.* Since the energy $E_1^h[s_h, \boldsymbol{n}_h]$ is uniformly bounded, Lemma 4.2.4 (coercivity) gives uniform bounds in $H^1(\Omega) \times \boldsymbol{H}^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ for the the sequence $\{(s_h, \boldsymbol{u}_h, s_h\boldsymbol{n}_h)\}$. With successive extractions of subsequences (not relabeled), one can show that there exists a limit $(s, \boldsymbol{u}, \widetilde{\boldsymbol{u}}) \in H^1(\Omega) \times \boldsymbol{H}^1(\Omega) \times \boldsymbol{H}^1(\Omega)$ such that $(s_h, \boldsymbol{u}_h, s_h\boldsymbol{n}_h)$ converges to $(s, \boldsymbol{u}, \widetilde{\boldsymbol{u}})$ weakly in $H^1(\Omega) \times \boldsymbol{H}^1(\Omega) \times \boldsymbol{H}^1(\Omega)$, strongly in $L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times \boldsymbol{L}^2(\Omega)$, and pointwise a.e. in Ω. Moreover, weak H^1-convergence guarantees attainment of traces, namely $s = g$ and $\boldsymbol{u} = \widetilde{\boldsymbol{u}} = \boldsymbol{r}$ on Γ_D. To see this, note that $g_h = I_h g \to g$ in $W^{1,p}(\Omega)$ for $p > d$,

according to (4.2.4), and so in $H^1(\Omega)$. Therefore $s_h - g_h \in H_0^1(\Omega)$ satisfies

$$s_h - g_h \rightharpoonup s - g \in H_0^1(\Omega),$$

because $H_0^1(\Omega)$ is closed under weak convergence. Hence $s = g$ on Γ_D in the sense of traces, as asserted. Dealing with $\boldsymbol{u}_h$ and $\widetilde{\boldsymbol{u}}_h$ is identical. Since $\boldsymbol{u}_h = I_h[s_h \boldsymbol{n}_h]$, interpolation and inverse estimates, yield

$$\begin{aligned}\|\boldsymbol{u}_h - s_h\boldsymbol{n}_h\|_{\boldsymbol{L}^2(\Omega)}^2 &\lesssim \sum_{K\in\mathcal{T}_h} h_K^4 \|D^2(s_h\boldsymbol{n}_h)\|_{\boldsymbol{L}^2(K)}^2 \\ &\lesssim \sum_{K\in\mathcal{T}_h} h_K^2 \|\boldsymbol{\nabla}(s_h\boldsymbol{n}_h)\|_{\boldsymbol{L}^2(K)}^2 \lesssim h^2 E_1^h[s_h, \boldsymbol{n}_h] \le Ch^2.\end{aligned}$$

This shows that $s_h\boldsymbol{n}_h$ and $\boldsymbol{u}_h$ converge strongly in $\boldsymbol{L}^2(\Omega)$ towards the same limit i.e., $\widetilde{\boldsymbol{u}} = \boldsymbol{u}$. Moreover, $s_h\boldsymbol{n}_h$ converges to $\boldsymbol{u}$ weakly in $\boldsymbol{H}^1(\Omega)$ and pointwise a.e. in Ω.

Step 2: $|s| = |\boldsymbol{u}|$ *a.e. in* Ω. The triangle inequality yields

$$\begin{aligned}\big\||\boldsymbol{u}_h|^2 - |s_h|^2\big\|_{L^1(\Omega)} &\le \big\||\boldsymbol{u}_h|^2 - I_h\big[|\boldsymbol{u}_h|^2\big]\big\|_{L^1(\Omega)} \\ &\quad + \big\|I_h\big[|\boldsymbol{u}_h|^2 - |s_h|^2\big]\big\|_{L^1(\Omega)} + \big\||s_h|^2 - I_h\big[|s_h|^2\big]\big\|_{L^1(\Omega)}.\end{aligned}$$

For the first and third terms on the right-hand side, standard interpolation estimates yield

$$\big\||s_h|^2 - I_h\big[|s_h|^2\big]\big\|_{L^1(\Omega)} \lesssim h^2\|\nabla s_h\|_{\boldsymbol{L}^2(\Omega)}^2, \quad \big\||\boldsymbol{u}_h|^2 - I_h\big[|\boldsymbol{u}_h|^2\big]\big\|_{L^1(\Omega)} \lesssim h^2\|\boldsymbol{\nabla}\boldsymbol{u}_h\|_{\boldsymbol{L}^2(\Omega)}^2.$$

On the other hand, since $\{s_h\}$ is uniformly bounded in $L^\infty(\Omega)$, we infer that

$$\begin{aligned}\|I_h\big[|\boldsymbol{u}_h|^2 - |s_h|^2\big]\|_{L^1(\Omega)} &= \|I_h\big[|s_h|^2(|\boldsymbol{n}_h|^2 - 1)\big]\|_{L^1(\Omega)} \\ &\leq \|s_h\|^2_{L^\infty(\Omega)}\|I_h\big[|\boldsymbol{n}_h|^2 - 1\big]\|_{L^1(\Omega)} \leq \varepsilon\|s_h\|^2_{L^\infty(\Omega)} \to 0,\end{aligned}$$

as $\varepsilon \to 0$. As $|s_h| \to |s|$ and $|\boldsymbol{u}_h| \to |\boldsymbol{u}|$ a.e. in Ω, we conclude that $|s| = |\boldsymbol{u}|$ a.e. in Ω.

Step 3: Convergence of $\{\boldsymbol{n}_h\}$. We now define $\boldsymbol{n} : \Omega \to \mathbb{R}^3$ as $\boldsymbol{n} := s^{-1}\boldsymbol{u}$ in $\Omega \setminus \Sigma$ and as an arbitrary unit vector in Σ. Step 2 implies, by construction, that $|\boldsymbol{n}| = 1$ a.e. in Ω. This shows that $(s, \boldsymbol{n}, \boldsymbol{u})$ satisfies the structural condition (4.1.6), i.e., $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}$.

We now observe that $s(x) \neq 0$ for a.e. $x \in \Omega \setminus \Sigma$ by definiton of Σ. Since $s_h(x) \to s(x)$ as $h \to 0$, if h is sufficiently small (depending on x), then $s_h(x) \neq 0$ is valid. Consequently,

$$\boldsymbol{n}_h(x) = \frac{s_h(x)\boldsymbol{n}_h(x)}{s_h(x)} \to \frac{\boldsymbol{u}(x)}{s(x)} = \boldsymbol{n}(x),$$

i.e., $\boldsymbol{n}_h \to \boldsymbol{n}$ pointwise a.e. in $\Omega \setminus \Sigma$. Since $\{\boldsymbol{n}_h\}$ is uniformly bounded in $\boldsymbol{L}^\infty(\Omega)$, the Lebesgue dominated convergence theorem yields $\boldsymbol{n}_h \to \boldsymbol{n}$ strongly in $\boldsymbol{L}^2(\Omega \setminus \Sigma)$.

Finally, the L^2-differentiability of $\boldsymbol{n}$ and the orthogonal decomposition of $\boldsymbol{\nabla u}$, both valid a.e. in $\Omega \setminus \Sigma$, follow from Proposition 4.1.1 (orthogonal decomposition). This concludes the proof. □

We are now in the position to prove the lim-inf inequality.

Proof of Theorem 4.2.1(ii)**.** The sequence $\{(s_h, \boldsymbol{n}_h, \boldsymbol{u}_h)\} \subset \mathbb{A}_{h,\varepsilon}(g_h, \boldsymbol{r}_h)$ satisfies the assumptions of Lemma 4.2.5 (characterization of limits). Hence, we can apply it to obtain subsequences (not relabeled) converging to the respective limits $(s, \boldsymbol{n}, \boldsymbol{u}) \in \mathbb{A}(g, \boldsymbol{r})$. Moreover, since also the sequences $\{\boldsymbol{n}_h \otimes \nabla s_h\}$ and $\{s_h \boldsymbol{\nabla} \boldsymbol{n}_h\}$ are uniformly bounded in $\boldsymbol{L}^2(\Omega)$, there exist subsequences (not relabeled) and functions $\boldsymbol{M}, \boldsymbol{N}$ in $\boldsymbol{L}^2(\Omega)$ such that $\boldsymbol{n}_h \otimes \nabla s_h \rightharpoonup \boldsymbol{M}$ and $s_h \boldsymbol{\nabla} \boldsymbol{n}_h \rightharpoonup \boldsymbol{N}$ weakly in $\boldsymbol{L}^2(\Omega)$. Combining the equality $s_h \boldsymbol{\nabla} \boldsymbol{n}_h = \boldsymbol{\nabla}(s_h \boldsymbol{n}_h) - \boldsymbol{n}_h \otimes \nabla s_h$, which is valid in every element of $\mathcal{T}_h$, with $s_h \boldsymbol{n}_h \rightharpoonup \boldsymbol{u}$ weakly in $\boldsymbol{H}^1(\Omega)$, helps identify the limits $\boldsymbol{N} = \boldsymbol{\nabla} \boldsymbol{u} - \boldsymbol{M}$.

Let $\boldsymbol{\Phi} \in \boldsymbol{C}_c^\infty(\Omega \setminus \Sigma)$ be an arbitrary $d \times d$ tensor field. We can thus write

$$\langle \boldsymbol{n}_h \otimes \nabla s_h - \boldsymbol{n} \otimes \nabla s, \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma} = \langle (\boldsymbol{n}_h - \boldsymbol{n}) \otimes \nabla s_h, \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma} + \langle \boldsymbol{n} \otimes (\nabla s_h - \nabla s), \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma}.$$

We note that $\boldsymbol{n}_h \to \boldsymbol{n}$ strongly in $L^2(\Omega \setminus \Sigma)$ implies

$$\langle (\boldsymbol{n}_h - \boldsymbol{n}) \otimes \nabla s_h, \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma} \leq \|\boldsymbol{n}_h - \boldsymbol{n}\|_{\boldsymbol{L}^2(\Omega \setminus \Sigma)} \|\nabla s_h\|_{\boldsymbol{L}^2(\Omega)} \|\boldsymbol{\Phi}\|_{\boldsymbol{L}^\infty(\Omega \setminus \Sigma)} \to 0$$

whereas $s_h \rightharpoonup s$ weakly in $H^1(\Omega)$ yields

$$\langle \boldsymbol{n} \otimes (\nabla s_h - \nabla s), \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma} \to 0.$$

Hence, we infer that

$$\langle \boldsymbol{n}_h \otimes \nabla s_h - \boldsymbol{n} \otimes \nabla s, \boldsymbol{\Phi} \rangle_{\Omega \setminus \Sigma} \to 0,$$

whence $\boldsymbol{n}_h \otimes \nabla s_h \rightharpoonup \boldsymbol{n} \otimes \nabla s$ weakly in $\boldsymbol{L}^2(\Omega \setminus \Sigma)$. This in turn identifies the limit $\boldsymbol{M} = \boldsymbol{n} \otimes \nabla s$ and gives the equality a.e. in $\Omega \setminus \Sigma$

$$\boldsymbol{N} = \nabla \boldsymbol{u} - \boldsymbol{n} \otimes \nabla s \quad \Rightarrow \quad \nabla \boldsymbol{n} = \frac{\boldsymbol{N}}{s}$$

where $\nabla \boldsymbol{n}$ is understood in the L^2-sense according to Proposition 4.1.1. Exploiting the fact that norms are weakly lower semicontinuous, along with $|\boldsymbol{n} \otimes \nabla s|^2 = |\nabla s|^2$ a.e. in $\Omega \setminus \Sigma$, and $\nabla s = \boldsymbol{0}$ a.e. in Σ, it holds that

$$\begin{aligned}
\liminf_{h \to 0} E_1^h[s_h, \boldsymbol{n}_h] &= \liminf_{h \to 0} \left\{ \frac{\kappa}{2} \|\boldsymbol{n}_h \otimes \nabla s_h\|^2_{\boldsymbol{L}^2(\Omega)} + \frac{1}{2} \|s_h \boldsymbol{\nabla} \boldsymbol{n}_h\|^2_{\boldsymbol{L}^2(\Omega)} \right\} \\
&\geq \liminf_{h \to 0} \left\{ \frac{\kappa}{2} \|\boldsymbol{n}_h \otimes \nabla s_h\|^2_{\boldsymbol{L}^2(\Omega \setminus \Sigma)} + \frac{1}{2} \|s_h \boldsymbol{\nabla} \boldsymbol{n}_h\|^2_{\boldsymbol{L}^2(\Omega \setminus \Sigma)} \right\} \\
&\geq \frac{\kappa}{2} \|\boldsymbol{n} \otimes \nabla s\|^2_{\boldsymbol{L}^2(\Omega \setminus \Sigma)} + \frac{1}{2} \|s \boldsymbol{\nabla} \boldsymbol{n}\|^2_{\boldsymbol{L}^2(\Omega \setminus \Sigma)} = E_1[s, \boldsymbol{n}].
\end{aligned}$$

Since $s_h \to s$ a.e. in Ω and ψ is continuous, $\psi(s_h) \to \psi(s)$ a.e. in Ω. The Fatou lemma yields

$$E_2[s] = \int_\Omega \psi(s) = \int_\Omega \lim_{h \to 0} \psi(s_h) \leq \liminf_{h \to 0} \int_\Omega \psi(s_h) = \liminf_{h \to 0} E_2^h[s]$$

Altogether, we thus obtain the lim-inf inequality (4.2.2). This finishes the proof. □

4.3 Iterative scheme

In this section, we propose an effective algorithm to compute discrete local minimizers of (4.1.16). The method is based on a discretization of the energy-

decreasing dynamics driven by the system of gradient flows

$$\partial_t \boldsymbol{n} + \delta_{\boldsymbol{n}} E^h[s, \boldsymbol{n}] = 0,$$

$$\partial_t s + \delta_s E^h[s, \boldsymbol{n}] = 0,$$

where $\delta_{\boldsymbol{n}} E^h[s, \boldsymbol{n}]$ and $\delta_s E^h[s, \boldsymbol{n}]$ denote the Gâteaux derivatives of the energy with respect to the order parameters, i.e.,

$$\begin{aligned}
\left\langle \delta_{\boldsymbol{n}} E^h[s, \boldsymbol{n}], \boldsymbol{\phi} \right\rangle &= \left\langle \delta_{\boldsymbol{n}} E_1^h[s, \boldsymbol{n}], \boldsymbol{\phi} \right\rangle = \kappa \langle \boldsymbol{n} \otimes \nabla s, \boldsymbol{\phi} \otimes \nabla s \rangle + \langle s \boldsymbol{\nabla} \boldsymbol{n}, s \boldsymbol{\nabla} \boldsymbol{\phi} \rangle, \\
\left\langle \delta_s E^h[s, \boldsymbol{n}], w \right\rangle &= \left\langle \delta_s E_1^h[s, \boldsymbol{n}], w \right\rangle + \left\langle \delta_s E_2^h[s, \boldsymbol{n}], w \right\rangle \\
&= \kappa \langle \boldsymbol{n} \otimes \nabla s, \boldsymbol{n} \otimes \nabla w \rangle + \langle s \boldsymbol{\nabla} \boldsymbol{n}, w \boldsymbol{\nabla} \boldsymbol{n} \rangle + \langle \psi'(s), w \rangle.
\end{aligned}$$

Let us introduce the ingredients of the scheme. First, let

$$V_{h,D} := \{ v_h \in V_h : v_h(z) = 0 \text{ for all } z \in \mathcal{N}_h \cap \Gamma_D \} \quad \text{and} \quad \mathbf{V}_{h,D} := (V_{h,D})^d$$

be the spaces of discrete functions satisfying homogeneous Dirichlet conditions on Γ_D. Given $\boldsymbol{n}_h \in \mathbf{V}_h$, we consider the subspace of $\mathbf{V}_{h,D}$ consisting of all discrete functions with nodal values orthogonal to those of $\boldsymbol{n}_h$ at all vertices:

$$\boldsymbol{\mathcal{K}}_h[\boldsymbol{n}_h] := \{ \boldsymbol{\phi}_h \in \mathbf{V}_{h,D} : \boldsymbol{n}_h(z) \cdot \boldsymbol{\phi}_h(z) = 0 \text{ for all } z \in \mathcal{N}_h \}.$$

For the treatment of the double well potential, we follow a convex splitting approach (see, e.g., [78]): we assume the splitting $\psi = \psi_c - \psi_e$, where ψ_c and ψ_e are both

convex, and ψ_c is quadratic.

The *time* discretization of the gradient flow for the director and the degree of orientation are based on the constant time-step sizes $\tau_{\boldsymbol{n}} > 0$ and $\tau_s > 0$, respectively. Moreover, we consider the difference quotient $d_t s_h^{i+1} := (s_h^{i+1} - s_h^i)/\tau_s$.

In the following algorithm, we state the proposed numerical scheme for the computation of discrete local minimizers of (4.1.16). We assume that assumption (4.1.14) is satisfied so that imposing Dirichlet boundary conditions directly for the director is allowed. Let $\mathsf{tol} > 0$ denote a tolerance.

Algorithm 4.3.1 (alternating direction discrete gradient flow). *Input:* $s_h^0 \in V_h$, $\boldsymbol{n}_h^0 \in \mathbf{V}_h$ *such that* $|\boldsymbol{n}_h^0(z)| = 1$ *for all* $z \in \mathcal{N}_h$, $\boldsymbol{n}_h^0(z) = \boldsymbol{r}_h(z)/g_h(z)$ *and* $s_h^0(z) = g_h(z)$ *for all* $z \in \mathcal{N}_h \cap \Gamma_D$.

Outer loop: For all $i \in \mathbb{N}_0$, *iterate* (i)–(ii)*:*

(i) *Inner loop: Given* $(\boldsymbol{n}_h^i, s_h^i)$, *let* $\boldsymbol{n}_h^{i,0} = \boldsymbol{n}_h^i$. *For all* $\ell \in \mathbb{N}_0$, *iterate* (i-a)–(ii-b)*:*

(i-a) *Compute* $\boldsymbol{t}_h^{i,\ell} \in \boldsymbol{\mathcal{K}}_h\big[\boldsymbol{n}_h^{i,\ell}\big]$ *such that*

$$\begin{aligned} &\langle \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h\rangle_* + \tau_{\boldsymbol{n}}\,\kappa\langle \boldsymbol{t}_h^{i,\ell} \otimes \nabla s_h^i, \boldsymbol{\phi}_h \otimes \nabla s_h^i\rangle + \tau_{\boldsymbol{n}}\langle s_h^i \boldsymbol{\nabla} \boldsymbol{t}_h^{i,\ell}, s_h^i \boldsymbol{\nabla} \boldsymbol{\phi}_h\rangle \\ &\quad = -\kappa\langle \boldsymbol{n}_h^{i,\ell} \otimes \nabla s_h^i, \boldsymbol{\phi}_h \otimes \nabla s_h^i\rangle - \langle s_h^i \boldsymbol{\nabla} \boldsymbol{n}_h^{i,\ell}, s_h^i \boldsymbol{\nabla} \boldsymbol{\phi}_h\rangle \end{aligned} \tag{4.3.1}$$

for all $\boldsymbol{\phi}_h \in \boldsymbol{\mathcal{K}}_h\big[\boldsymbol{n}_h^{i,\ell}\big]$*;*

(i-b) *Update* $\boldsymbol{n}_h^{i,\ell+1} := \boldsymbol{n}_h^{i,\ell} + \tau_{\boldsymbol{n}}\,\boldsymbol{t}_h^{i,\ell}$*;*

until

$$\big|E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell+1}] - E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell}]\big| < \mathsf{tol}. \tag{4.3.2}$$

If $\ell_i \in \mathbb{N}_0$ denotes the smallest integer for which the stopping criterion (4.3.2) is satisfied, define $\boldsymbol{n}_h^{i+1} := \boldsymbol{n}_h^{i,\ell_i+1}$.

(ii) *Compute $s_h^{i+1} \in V_h$ such that $s_h^{i+1}(z) = g_h(z)$ for all $z \in \mathcal{N}_h \cap \Gamma_D$ and*

$$\begin{aligned}&\langle d_t s_h^{i+1}, w_h\rangle + \kappa\langle \boldsymbol{n}_h^{i+1} \otimes \nabla s_h^{i+1}, \boldsymbol{n}_h^{i+1} \otimes \nabla w_h\rangle + \langle s_h^{i+1}\boldsymbol{\nabla}\boldsymbol{n}_h^{i+1}, w_h \boldsymbol{\nabla}\boldsymbol{n}_h^{i+1}\rangle \\ &\quad + \langle \psi_c'(s_h^{i+1}), w_h\rangle = \langle \psi_e'(s_h^i), w_h\rangle\end{aligned} \tag{4.3.3}$$

for all $w_h \in V_{h,D}$.

Output: Sequence of approximations $\{(s_h^i, \boldsymbol{n}_h^i)\}_{i\in\mathbb{N}_0}$.

In Algorithm 4.3.1, $\langle\cdot,\cdot\rangle_*$ denotes the scalar product of the metric used in the discrete gradient flow (4.3.1) for the director. In this work, we consider the following two choices for $\langle\cdot,\cdot\rangle_*$:

$$\langle \boldsymbol{\phi}, \boldsymbol{\psi}\rangle_* = \langle \boldsymbol{\phi}, \boldsymbol{\psi}\rangle \qquad (L^2\text{-metric}), \tag{4.3.4}$$

$$\langle \boldsymbol{\phi}, \boldsymbol{\psi}\rangle_* = \langle h^\alpha \boldsymbol{\nabla}\boldsymbol{\phi}, \boldsymbol{\nabla}\boldsymbol{\psi}\rangle, \quad \text{with } 0 < \alpha \leq 2. \qquad (\text{weighted } H^1\text{-metric}), \tag{4.3.5}$$

Note that in (4.3.5) the choice $\alpha = 0$ corresponds to a full H^1-gradient flow, which is not appropriate since the director does not belong to $H^1(\Omega)$ in general (e.g., in the presence of defects). On the other hand, if $\alpha = 2$, the resulting metric is equivalent to the L^2-metric in (4.3.4). In addition, both (4.3.1) and (4.3.3) are linear SPD systems in the unknowns $\boldsymbol{t}_h^{i,\ell}$ and s_h^{i+1}.

Although in most of our numerical experiments we will set $\tau_{\boldsymbol{n}} = \tau_s$, we observed

that in some situations the flexibility of choosing of different time-step sizes in (4.3.1) and (4.3.3) is decisive in order to move defects in numerical simulations.

In the following proposition, we prove well-posedness and an energy-decreasing property of Algorithm 4.3.1.

Proposition 4.3.1 (properties of Algorithm 4.3.1)**.** *Algorithm 4.3.1 is well-posed and energy-descresing. Specifically, for all $i \in \mathbb{N}_0$, the following assertions hold:*

(i) *For all $\ell \in \mathbb{N}_0$, (4.3.1) admits a unique solution $\boldsymbol{t}_h^{i,\ell} \in \boldsymbol{\mathcal{K}}_h\big[\boldsymbol{n}_h^{i,\ell}\big]$;*

(ii) *The inner loop terminates in a finite number of iterations, i.e., there exists $\ell \in \mathbb{N}_0$ such that the stopping criterion (4.3.2) is met;*

(iii) *(4.3.3) admits a unique solution $s_h^{i+1} \in V_h$ such that $s_h^{i+1}(z) = g_h(z)$ for all $z \in \mathcal{N}_h \cap \Gamma_D$.*

(iv) *There holds*

$$E^h[s_h^{i+1}, \boldsymbol{n}_h^{i+1}] - E^h[s_h^i, \boldsymbol{n}_h^i] \leq -\left(\tau_s \|d_t s_h^{i+1}\|_{L^2(\Omega)}^2 + \tau_{\boldsymbol{n}} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2\right) - \left(\tau_s^2\, E_1^h[d_t s_h^{i+1}, \boldsymbol{n}_h^{i+1}] + \tau_{\boldsymbol{n}}^2 \sum_{\ell=0}^{\ell_i} E_1^h[s_h^i, \boldsymbol{t}_h^{i,\ell}]\right). \tag{4.3.6}$$

In particular, $E^h[s_h^{i+1}, \boldsymbol{n}_h^{i+1}] \leq E^h[s_h^i, \boldsymbol{n}_h^i]$ and equality holds if and only if $(s_h^{i+1}, \boldsymbol{n}_h^{i+1}) = (s_h^i, \boldsymbol{n}_h^i)$ (equilibrium state).

Proof. Let $i \in \mathbb{N}_0$ and $\ell \in \mathbb{N}_0$. For fixed $s_h^i \in V_h$ (resp., $\boldsymbol{n}_h^{i+1} \in \mathbf{V}_h$), the left-hand side of (4.3.1) (resp., of (4.3.3)) is a coercive and continuous bilinear form on

$\mathbf{V}_{h,D}$ (resp., on $V_{h,D}$). Therefore, the variational problem admits a unique solution $\boldsymbol{t}_h^{i,\ell} \in \boldsymbol{\mathcal{K}}_h[\boldsymbol{n}_h^{i,\ell}]$ (resp., $s_h^{i+1} \in V_h$). This shows part (i) and (iii) of Proposition 4.3.1.

Choosing the test function $\boldsymbol{\phi}_h = \tau_{\boldsymbol{n}}\, \boldsymbol{t}_h^{i,\ell} = \boldsymbol{n}_h^{i,\ell+1} - \boldsymbol{n}_h^{i,\ell} \in \boldsymbol{\mathcal{K}}_h[\boldsymbol{n}_h^{i,\ell}]$ in (4.3.1) yields

$$\tau_{\boldsymbol{n}}\|\boldsymbol{t}_h^{i,\ell}\|_*^2+\kappa\langle \boldsymbol{n}_h^{i,\ell+1}\otimes\nabla s_h^i, (\boldsymbol{n}_h^{i,\ell+1}-\boldsymbol{n}_h^{i,\ell})\otimes\nabla s_h^i\rangle_\Omega+\langle s_h^i\boldsymbol{\nabla}\boldsymbol{n}_h^{i,\ell+1}, s_h^i\boldsymbol{\nabla}(\boldsymbol{n}_h^{i,\ell+1}-\boldsymbol{n}_h^{i,\ell})\rangle_\Omega = 0.$$

Using the identity $2a(a-b) = a^2 - b^2 + (a-b)^2$, valid for all $a, b \in \mathbb{R}$, we obtain the identity

$$\begin{aligned}\tau_{\boldsymbol{n}}\|\boldsymbol{t}_h^{i,\ell}\|_*^2 + \frac{\kappa}{2}\|\boldsymbol{n}_h^{i,\ell+1}\otimes\nabla s_h^i\|_{\boldsymbol{L}^2(\Omega)}^2 - \frac{\kappa}{2}\|\boldsymbol{n}_h^{i,\ell}\otimes\nabla s_h^i\|_{\boldsymbol{L}^2(\Omega)}^2 + \frac{\kappa}{2}\|\tau_{\boldsymbol{n}}\boldsymbol{t}_h^{i,\ell}\otimes\nabla s_h^i\|_{\boldsymbol{L}^2(\Omega)}^2 \\ + \frac{1}{2}\|s_h^i\boldsymbol{\nabla}\boldsymbol{n}_h^{i,\ell+1}\|_{\boldsymbol{L}^2(\Omega)}^2 - \frac{1}{2}\|s_h^i\boldsymbol{\nabla}\boldsymbol{n}_h^{i,\ell}\|_{\boldsymbol{L}^2(\Omega)}^2 + \frac{1}{2}\|s_h^i\boldsymbol{\nabla}(\boldsymbol{n}_h^{i,\ell+1}-\boldsymbol{n}_h^{i,\ell})\|_{\boldsymbol{L}^2(\Omega)}^2 = 0,\end{aligned}$$

which can be rewritten in more compact form as

$$E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell+1}] - E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell}] + \tau_{\boldsymbol{n}}\|\boldsymbol{t}_h^{i,\ell}\|_*^2 + \tau_{\boldsymbol{n}}^2\, E_1^h[s_h^i, \boldsymbol{t}_h^{i,\ell}] = 0. \tag{4.3.7}$$

In particular, $E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell+1}] \leq E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell}]$ is valid. Since $E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell}] \geq 0$ for all $i \in \mathbb{N}_0$, the sequence $\{E_1^h[s_h^i, \boldsymbol{n}_h^{i,\ell}]\}_{\ell\in\mathbb{N}_0}$ is convergent (as it is monotonically decreasing and bounded from below). In particular, it is a Cauchy sequence, which entails that the stopping criterion (4.3.2) is met in a finite number of iterations. This shows part (ii) of the proposition.

Let $\ell_i \in \mathbb{N}_0$ be the smallest integer for which the stopping criterion (4.3.2) is satisfied. Recall that $\boldsymbol{n}_h^{i+1} = \boldsymbol{n}_h^{i,\ell_i+1}$ and $\boldsymbol{n}_h^i = \boldsymbol{n}_h^{i,0}$. Summation of (4.3.7) over

$\ell = 0, \dots, \ell_i$ yields that

$$E_1^h[s_h^i, \boldsymbol{n}_h^{i+1}] - E_1^h[s_h^i, \boldsymbol{n}_h^i] + \tau_{\boldsymbol{n}} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2 + \tau_{\boldsymbol{n}}^2 \sum_{\ell=0}^{\ell_i} E_1^h[s_h^i, \boldsymbol{t}_h^{i,\ell}] = 0. \tag{4.3.8}$$

Choosing the test function $w_h = \tau_s d_t s_h^{i+1} = s_h^{i+1} - s_h^i \in V_{h,D}$ in (4.3.3) and performing the same algebraic computation as above, we arrive at

$$\begin{aligned} &E_1^h[s_h^{i+1}, \boldsymbol{n}_h^{i+1}] - E_1^h[s_h^i, \boldsymbol{n}_h^{i+1}] \\ &\quad + \tau_s \|d_t s_h^{i+1}\|_{L^2(\Omega)}^2 + \tau_s^2 \, E_1^h[d_t s_h^{i+1}, \boldsymbol{n}_h^{i+1}] + \langle \psi_c'(s_h^{i+1}) - \psi_e'(s_h^i), s_h^{i+1} - s_h^i \rangle_\Omega = 0. \end{aligned}$$

Applying [59, Lemma 4.1], which yields the inequality

$$E_2^h[s_h^{i+1}] - E_2^h[s_h^i] \leq \langle \psi_c'(s_h^{i+1}) - \psi_e'(s_h^i), s_h^{i+1} - s_h^i \rangle_\Omega,$$

we obtain

$$\begin{aligned} &E_1^h[s_h^{i+1}, \boldsymbol{n}_h^{i+1}] - E_1^h[s_h^i, \boldsymbol{n}_h^{i+1}] + \tau_s \, \|d_t s_h^{i+1}\|_{L^2(\Omega)}^2 + \tau_s^2 \, E_1^h[d_t s_h^{i+1}, \boldsymbol{n}_h^{i+1}] \\ &+ E_2^h[s_h^{i+1}] - E_2^h[s_h^i] \leq 0. \end{aligned}$$

Adding the latter with (4.3.8), and exploiting cancellation of $E_1^h[s_h^i, \boldsymbol{n}_h^{i+1}]$, we deduce

$$\begin{aligned} &E^h[s_h^{i+1}, \boldsymbol{n}_h^{i+1}] - E^h[s_h^i, \boldsymbol{n}_h^i] \\ &\quad \leq -\tau_s \|d_t s_h^{i+1}\|_{L^2(\Omega)}^2 - \tau_s^2 E_1^h[d_t s_h^{i+1}, \boldsymbol{n}_h^{i+1}] - \tau_{\boldsymbol{n}} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2 - \tau_{\boldsymbol{n}}^2 \sum_{\ell=0}^{\ell_i} E_1^h[s_h^i, \boldsymbol{t}_h^{i,\ell}] \leq 0. \end{aligned}$$

This shows (4.3.6) and concludes the proof. □

Remark 4.3.1 (energy decrease)**.** *The right-hand side of* (4.3.6) *characterizes the energy decrease guaranteed by each step of Algorithm 4.3.1 and comprises two contributions: The term*

$$-\left(\tau_s\|d_t s_h^{i+1}\|_{L^2(\Omega)}^2+\tau_{\boldsymbol{n}}\sum_{\ell=0}^{\ell_i}\|\boldsymbol{t}_h^{i,\ell}\|_*^2\right)$$

is the energy decrease due to the gradient-flow nature of Algorithm 4.3.1. The term

$$-\left(\tau_s^2\,E_1^h[d_t s_h^{i+1},\boldsymbol{n}_h^{i+1}]+\tau_{\boldsymbol{n}}^2\sum_{\ell=0}^{\ell_i}E_1^h[s_h^i,\boldsymbol{t}_h^{i,\ell}]\right)$$

is the numerical dissipation due to the backward Euler methods used for the time discretization.

In practical implementations of Algorithm 4.3.1, the outer loop is terminated when

$$\left|E^h[s_h^{i+1},\boldsymbol{n}_h^{i+1}]-E^h[s_h^i,\boldsymbol{n}_h^i]\right|<\mathsf{tol}. \tag{4.3.9}$$

Since the algorithm fulfills a monotone energy decreasing property (see Proposition 4.3.1(iv)), the stopping criterion is met in a finite number of iterations.

The approximations $\boldsymbol{n}_h^{i+1}$ of the director generated by Algorithm 4.3.1 do not satisfy the unit-length constraint at the vertices of the mesh, as in [59, 60]. However, the following proposition shows that violation of this constraint can be controlled by the time-step size $\tau_{\boldsymbol{n}}$, independently of the number of iterations. Moreover, the uniform boundedness in $\boldsymbol{L}^\infty(\Omega)$ of the sequence can be guaranteed if the discretization parameters are chosen appropriately.

Proposition 4.3.2 (properties of discrete director field)**.** *Let $j \geq 1$. The following holds.*

(i) Suppose that the norm induced by the metric $\langle \cdot, \cdot \rangle_*$ used in (4.3.1) is an upper bound for the L^2-norm, i.e., there exists $C_* > 0$ such that

$$\|\boldsymbol{\phi}_h\|_{\boldsymbol{L}^2(\Omega)} \leq C_* \|\boldsymbol{\phi}_h\|_* \quad \text{for all } \boldsymbol{\phi}_h \in \mathbf{V}_{h,D}. \tag{4.3.10}$$

Then, the approximations generated by Algorithm 4.3.1 satisfy

$$\|I_h\big[|\boldsymbol{n}_h^j|^2 - 1\big]\|_{L^1(\Omega)} \leq C_1 \tau_{\boldsymbol{n}}\, E^h[s_h^0, \boldsymbol{n}_h^0], \tag{4.3.11}$$

where $C_1 > 0$ depends only on C_* and the shape-regularity of $\{\mathcal{T}_h\}$.

(ii) Suppose $\tau_{\boldsymbol{n}}$ fulfills the following CFL-type condition:

$$\begin{aligned} \tau_{\boldsymbol{n}} h_{\min}^{-d} &\leq C^* \quad \text{if } \langle \cdot, \cdot \rangle_* \text{ is chosen as (4.3.4)}, \\ \tau_{\boldsymbol{n}} h_{\min}^{2-d-\alpha} (\log h_{\min}^{-1})^2 &\leq C^*, \quad \text{if } \langle \cdot, \cdot \rangle_* \text{ is chosen as (4.3.5)}, \end{aligned} \tag{4.3.12}$$

where $h_{\min} := \min_{K \in \mathcal{T}_h} h_K$ and $C^* > 0$ is arbitrary. Then, the approximations generated by Algorithm 4.3.1 satisfy

$$\|\boldsymbol{n}_h^j\|_{\boldsymbol{L}^\infty(\Omega)} \leq C_2 (1 + E^h[s_h^0, \boldsymbol{n}_h^0]), \tag{4.3.13}$$

where $C_2 > 0$ is proportional to $C^* > 0$ in (4.3.12) with proportionality constant depending on the shape-regularity of $\{\mathcal{T}_h\}$.

Proof of Proposition 4.3.2. Let $j \geq 1$. Summation of (4.3.6) over $i = 0, \dots, j-1$ yields that

$$
\begin{aligned}
&E^h[s_h^j, \boldsymbol{n}_h^j] - E^h[s_h^0, \boldsymbol{n}_h^0] \\
&\quad \leq -\sum_{i=0}^{j-1} \left(\tau_s \|d_t s_h^{i+1}\|_{L^2(\Omega)}^2 + \tau_s^2 E_1^h[d_t s_h^{i+1}, \boldsymbol{n}_h^{i+1}] + \tau_{\boldsymbol{n}} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2 + \tau_{\boldsymbol{n}}^2 \sum_{\ell=0}^{\ell_i} E_1^h[s_h^i, \boldsymbol{t}_h^{i,\ell}] \right).
\end{aligned}
$$

In particular, omitting some nonnegative terms, it follows that

$$
\tau_{\boldsymbol{n}} \sum_{i=0}^{j-1} \sum_{k=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2 \leq E^h[s_h^0, \boldsymbol{n}_h^0]. \tag{4.3.14}
$$

Moreover, for all $z \in \mathcal{N}_h$, the tangential update $\boldsymbol{t}_h^{i,\ell}(z)$ is perpendicular to $\boldsymbol{n}_h^{i,\ell}(z)$, whence $\boldsymbol{n}_h^{i,\ell+1}(z) = \boldsymbol{n}_h^{i,\ell}(z) + \tau_{\boldsymbol{n}} \boldsymbol{t}_h^{i,\ell}(z)$ satisfies $|\boldsymbol{n}_h^{i,\ell+1}(z)|^2 = |\boldsymbol{n}_h^{i,\ell}(z)|^2 + \tau_{\boldsymbol{n}}^2 |\boldsymbol{t}_h^{i,\ell}(z)|^2$. Iterating in ℓ and i gives

$$
|\boldsymbol{n}_h^j(z)|^2 = |\boldsymbol{n}_h^0(z)|^2 + \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} |\boldsymbol{t}_h^{i,\ell}(z)|^2 = 1 + \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} |\boldsymbol{t}_h^{i,\ell}(z)|^2 \geq 1.
$$

Then, using the equivalence of the L^p-norm of a discrete function with the weighted ℓ^p-norm of the vector collecting its nodal values (see, e.g., [11, Lemma 3.4]), for h_z being the diameter of the nodal patch associated with $z \in \mathcal{N}_h$, we see that

$$
\begin{aligned}
\|I_h[|\boldsymbol{n}_h^j|^2] - 1\|_{L^1(\Omega)} &\lesssim \sum_{z \in \mathcal{N}_h} h_z^d \big(|\boldsymbol{n}_h^j(z)|^2 - 1 \big) \leq \tau_{\boldsymbol{n}}^2 \sum_{z \in \mathcal{N}_h} h_z^d \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} |\boldsymbol{t}_h^{i,\ell}(z)|^2 \\
&\lesssim \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_{\boldsymbol{L}^2(\Omega)}^2.
\end{aligned}
$$

Combining (4.3.10) with (4.3.14) leads to

$$\|I_h[|\boldsymbol{n}_h^j|^2] - 1\|_{L^1(\Omega)} \lesssim C_* \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_*^2 \leq C_* \tau_{\boldsymbol{n}}\, E^h[s_h^0, \boldsymbol{n}_h^0],$$

which turns out to be (4.3.11).

It remains to estimate $\|\boldsymbol{n}_h^j\|_{\boldsymbol{L}^\infty(\Omega)}$. Let us consider first the case of the weighted H^1-metric (4.3.5). Using a global inverse estimate (see, e.g., [11, Remark 3.8]) and the Poincaré inequality, we obtain that

$$\begin{aligned}
\|\boldsymbol{n}_h^j\|_{\boldsymbol{L}^\infty(\Omega)}^2 &= \max_{z\in\mathcal{N}_h} |\boldsymbol{n}_h^j(z)|^2 \leq 1 + \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \max_{z\in\mathcal{N}_h} |\boldsymbol{t}_h^{i,\ell}(z)|^2 \\
&\lesssim 1 + \tau_{\boldsymbol{n}}^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_{\boldsymbol{L}^\infty(\Omega)}^2 \lesssim 1 + \tau_{\boldsymbol{n}}^2\, h_{\min}^{2-d} (\log h_{\min}^{-1})^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{t}_h^{i,\ell}\|_{\boldsymbol{H}^1(\Omega)}^2 \\
&\lesssim 1 + \tau_{\boldsymbol{n}}^2\, h_{\min}^{2-d-\alpha} h^\alpha (\log h_{\min}^{-1})^2 \sum_{i=0}^{j-1} \sum_{\ell=0}^{\ell_i} \|\boldsymbol{\nabla} \boldsymbol{t}_h^{i,\ell}\|_{\boldsymbol{L}^2(\Omega)}^2 \\
&\leq 1 + \tau_{\boldsymbol{n}}\, h_{\min}^{2-d-\alpha} (\log h_{\min}^{-1})^2 E^h[s_h^0, \boldsymbol{n}_h^0].
\end{aligned}$$

Therefore, (4.3.13) is satisfied if $\tau_{\boldsymbol{n}}\, h_{\min}^{2-d-\alpha} |\log h_{\min}|^2 \leq C^*$ with C^* arbitrary. For the L^2-metric (4.3.4), the result follows analogously, provided that $\tau_{\boldsymbol{n}}\, h_{\min}^{-d} \leq C^*$. □

To conclude this section, we discuss the structure of Algorithm 4.3.1 with special emphasis on its nested structure and distinct roles of $\tau_{\boldsymbol{n}}$ and τ_s. Obviously, $\tau_{\boldsymbol{n}}$ controls the violation of the unit length constraint according to (4.3.11), but the roles of subiterations in (4.3.1) and τ_s in (4.3.3) is more subtle and deserves further elaboration. The presence of defects is associated with values $s_h^i(x_j)$ close to zero at nodes x_j, which in turn act as weights in the equation (4.3.1) for the tangential

updates $\boldsymbol{t}_h^{i,\ell}$ of the director field $\boldsymbol{n}_h^{i,\ell}$. The fast decrease to zero of $s_h^i(x_j)$, relative to the growth of $\nabla \boldsymbol{n}_h^i$ in its vicinity, impedes further changes of $\boldsymbol{n}_h^i(x_j)$ because they are not energetically favorable: the defect is thus pinned at the same location x_j for many interations. Experiments with Algorithm 4.3.1 reveal defect pinning if $\tau_{\boldsymbol{n}} = \tau_s$ and one step of (4.3.1) per step of (4.3.3) is utilized. The subiterations within the inner loop (4.3.1) allow $\boldsymbol{n}_h^{i,\ell}$ to adjust to the current value of s_h^i. This mimics an approximate optimization step but with unit length and max norm control dictated by Proposition 4.3.2. In contrast, full optimization has been proposed in [59, 60, 77] instead of (4.3.1), followed by nodal projection onto the unit sphere, whereas one step of a weighted gradient flow (4.3.1) has been advocated in [26] for the Q-tensor model. On the other hand, since τ_s penalizes changes of s_h^i, smaller values of τ_s relative to $\tau_{\boldsymbol{n}}$ delay changes of s_h^i in favor of changes of $\boldsymbol{n}_h^i$. This does not fix the stiff character of (4.3.1), studied in [31], but does remove defect pinning. Several numerical experiments in Section 4.4 document this finding.

4.4 Numerical experiments

In this section, we present a series of numerical experiments that explore the accuracy of Algorithm 4.3.1 and its ability to approximate rather complex defects of nematic LCs in 2d and 3d. In both cases, these results complement the theory and extend it.

We have implemented Algorithm 4.3.1 within the high performance multiphysics finite element software Netgen/NGSolve [69]. To solve the constrained vari-

ational problem (4.3.1), we adopt a saddle point approach. The ensuing linear systems are solved using the built-in conjugate gradient solver of Netgen/NGSolve, while the visualization relies on ParaView [6].

All pictures below obey the following rules. The vector field depicts the director $\boldsymbol{n}$, whereas the color scale refers to the degree of orientation s. Blue regions indicate areas with values of s close to zero, which signify the occurrence of defects, while the red ones indicate regions with largest values of s ($s \approx 0.75$ in our simulations), where the director encodes the local orientation of the LC molecules. We generate unstructured, generally non-weakly acute, meshes within Netgen with desirable meshsize h_0 but the effective maximum size h of tetrahedra in 3d may only satisfy $h \approx h_0$. We will specify h_0 when dealing with unstructured 3d meshes.

We stress that, unlike FEMs proposed in previous works [59, 60], the energy-decreasing property of Algorithm 4.3.1 does rely on meshes being weakly acute (cf. Proposition 4.3.1). Except for simple 3d geometries, such meshes are hard, to impossible, to construct. This is the case of the cylinder domain in Section 4.4.4 and the Saturn ring configurations in Section 4.4.6, for which mesh flexibility is of fundamental importance to capture topologically complicated defects.

Throughout this section, we consider the double well potential

$$\psi(s) = c_{\mathrm{dw}}(\psi_c(s) - \psi_e(s))$$

with

$$\psi_c(s) := 63s^2, \quad \psi_e(s) := -16s^4 + \frac{64}{3}s^3 + 57s^2 - 0.5625, \tag{4.4.1}$$

where $c_{\mathrm{dw}} \geq 0$. Note that, for $c_{\mathrm{dw}} > 0$, ψ has a local minimum at $s = 0$ and a global minimum at $s = \hat{s} := 0.750025$ such that $\psi(\hat{s}) = 0$. Moreover, in view of Proposition 4.3.2 (properties of discrete director field), we measure the violation of the unit-length constraint in terms of the quantity

$$\mathrm{err}_{\boldsymbol{n}} := \|I_h\big[|\boldsymbol{n}_h^N|^2 - 1\big]\|_{L^1(\Omega)}, \tag{4.4.2}$$

where $\boldsymbol{n}_h^N$ denotes the final approximation of the director field generated by Algorithm 4.3.1. Furthermore, unless otherwise specified, we choose the L^2-metric (4.3.4) in (4.3.1), and we set the tolerance $\mathsf{tol} = 10^{-6}$ in both (4.3.2) and (4.3.9).

4.4.1 Lagrange multipliers

Note that in each step of the step (i) of Algorithm 4.3.1 (the inner loop), it requires to solve $\boldsymbol{t}_h^{i,\ell}$ in the admissible set $\boldsymbol{\mathcal{K}}_h(\boldsymbol{n}_h^{i,\ell})$. We realize this node-wise constraint by the method of Lagrange multiplier. Indeed, if we rewrite (4.3.1) as

$$\langle \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h \rangle_* + \tau_{\boldsymbol{n}} a_h(s_h^i; \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h) = -a_h(s_h^i; \boldsymbol{n}_h^{i,\ell}, \boldsymbol{\phi}_h),$$

then in each step of the inner loop, one needs to solve $\boldsymbol{t}_h^{i,\ell} \in \mathbf{V}_{h,D}$ and $\lambda_h^{\ell+1} \in V_{h,D}$ such that

$$\begin{aligned} \langle \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h \rangle_* + \tau_{\boldsymbol{n}} a_h(s_h^i; \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h) + b_h(\boldsymbol{n}_h^{i,\ell}; \lambda_h^{\ell+1}, \boldsymbol{\phi}_h) &= -a_h(s_h^i; \boldsymbol{n}_h^{i,\ell}, \boldsymbol{\phi}_h), \\ b_h(\boldsymbol{n}_h^{i,\ell}; \mu_h, \boldsymbol{t}_h^{i,\ell}) &= 0, \end{aligned} \tag{4.4.3}$$

for any $\boldsymbol{\phi}_h \in \mathbf{V}_{h,D}$ and $\mu_h \in V_{h,D}$. Here,

$$a_h(s_h^i; \boldsymbol{t}_h^{i,\ell}, \boldsymbol{\phi}_h) := \kappa \langle \boldsymbol{t}_h^{i,\ell} \otimes \nabla s_h^i, \boldsymbol{\phi}_h \otimes \nabla s_h^i \rangle + \langle s_h^i \nabla \boldsymbol{t}_h^{i,\ell}, s_h^i \nabla \boldsymbol{\phi}_h \rangle, \tag{4.4.4}$$

and

$$b_h(\boldsymbol{n}_h^{i,\ell}; \lambda_h^{\ell+1}, \boldsymbol{\phi}_h) := \int_\Omega I_h[\lambda_h^{\ell+1}(\boldsymbol{n}_h^{i,\ell} \cdot \boldsymbol{\phi}_h)]. \tag{4.4.5}$$

Note that the bilinear form $a_h(s_h^i; \cdot, \cdot)$ accounts for the variation of energy E_1^h with respect to $\boldsymbol{n}_h$, and the bilinear form $b_h(\boldsymbol{n}_h^{i,\ell}; \cdot, \cdot)$ encodes the constraint. We use a conjugate gradient solver to solve the saddle point system (4.4.3).

4.4.2 Point defect in 2D

In striking contrast with the Oseen–Frank model, the Ericksen model allows point defects to have finite energy in 2D: the blow-up of $|\boldsymbol{\nabla n}|$ near a defect is compensated by infinitesimal values of s for the energy $E[s, \boldsymbol{n}]$ (1.3.1) to stay bounded. We examine this basic mechanism with simulations of a point defect in 2D and study the influence of the discretization parameters on the performance of Algorithm 4.3.1.

We consider the unit square $\Omega = (0, 1)^2$, and set $\kappa = 2$ in (1.3.1) as well as $c_{\mathrm{dw}} = 0.1(0.3)^{-2}$ in (4.4.1). We impose Dirichlet boundary conditions for s and $\boldsymbol{n}$ on $\partial\Omega$, namely

$$g = \hat{s} \quad \text{and} \quad \boldsymbol{q} = \boldsymbol{r}/g = \frac{(x - 0.5, y - 0.5)}{|(x - 0.5, y - 0.5)|} \quad \text{on } \partial\Omega. \tag{4.4.6}$$

To initialize Algorithm 4.3.1, we consider a constant degree of orientation $s_h^0 = \hat{s}$

in Ω and a director $\boldsymbol{n}_h^0$ exhibiting an off-center point defect located at $(0.24, 0.24)$. Due to the imposed boundary conditions and for symmetry reasons, we expect that an energy-decreasing dynamics moves the defect to the center of the square; see Figure 4.2.

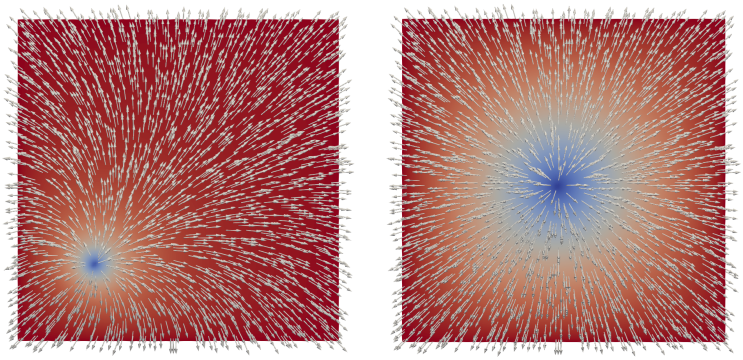

Figure 4.2: Point defect experiment of Section 4.4.2: Plot of the approximation $(s_h^1, \boldsymbol{n}_h^1)$ after the first iteration (left) and of the final approximation $(s_h^N, \boldsymbol{n}_h^N)$ (right). The gradient flow algorithm moves the defect to the center of the domain.

In our first experiment, we consider a uniform mesh $\mathcal{T}_h$ of the unit square consisting of 2048 right triangles. The resulting mesh size is $h = \sqrt{2}\,2^{-5}$. Moreover, we set $\tau_{\boldsymbol{n}} = \tau_s = 0.1$ and compare the results obtained for different choices of the metric $\langle \cdot, \cdot \rangle_*$ in (4.3.1); cf. (4.3.4)–(4.3.5). Table 4.1 displays the outputs for each run. On the one hand, we observe that using the L^2-metric leads to the fastest dynamics in terms of both number of iterations and CPU time. On the other hand, the violation of the unit-length constraint is smaller for the weighted H^1-metrics. For smaller values of α in the weighted H^1-metric, Algorithm 4.3.1 terminates with a configuration exhibiting defect pinning at an off-center location. The expected equilibrium state, depicted in Figure 4.2(b), can be restored when reducing the time-step size τ_s.

metric	$E^h[s_h^N, \boldsymbol{n}_h^N]$	N	$\min(s_h^N)$	$\mathrm{err}_{\boldsymbol{n}}$	CPU time (s)
L^2	2.984	60	0.0757	0.0404	64.83
weighted H^1, $\alpha = 2.0$	2.944	67	0.0750	0.0370	98.65
weighted H^1, $\alpha = 1.9$	2.938	65	0.0754	0.0362	111.69
weighted H^1, $\alpha = 1.8$	2.932	67	0.0755	0.0353	130.17
weighted H^1, $\alpha = 1.7$	2.926	80	0.0760	0.0342	154.92

Table 4.1: Point defect experiment of Section 4.4.2: Final outputs of Algorithm 4.3.1 for different choices of metric $\langle \cdot, \cdot \rangle_*$, namely value of the energy $E^h[s_h^N, \boldsymbol{n}_h^N]$ for the equilibrium state, total number of iterations N, smallest value of the final s_h^N, error in the unit-length constraint in (4.4.2), and the CPU time.

In our second set of experiments, we investigate the effect of mesh refinements and changes of time steps on the results. To this end, we first repeat the simulation using three uniform meshes with $h = \sqrt{2}\,2^{-5-\ell}$ ($\ell = 0, 1, 2$); we set $\tau_{\boldsymbol{n}} = 0.1\,2^{-2\ell}$, in agreement with the first CFL condition in (4.3.12) for $d = 2$. We collect the results of computations in Table 4.2 (upper), and observe that both $\min(s_h^N)$ and $\mathrm{err}_{\boldsymbol{n}}$ decrease about linearly with h whereas the energy $E^h[s_h^N, \boldsymbol{n}_h^N]$ also decreases and converges asymptotically. We next consider a fixed mesh with $h = \sqrt{2}\,2^{-5}$ and study the decay of $\mathrm{err}_{\boldsymbol{n}}$ in (4.4.2) as the time-step size $\tau_{\boldsymbol{n}}$ decreases; see Table 4.2 (lower). In this third set of experiments, we let $\tau_{\boldsymbol{n}} = (0.1)2^{-5-\ell}$ ($\ell = 0, 1, 2$), and $\mathsf{tol} = 10^{-5}\tau_{\boldsymbol{n}}$ in both (4.3.2) and (4.3.9). The computational results in Table 4.2 (lower) confirm the first-order convergence with respect to $\tau_{\boldsymbol{n}}$ established in Proposition 4.3.2; see (4.3.11) that bounds $\mathrm{err}_{\boldsymbol{n}}$ in terms of $\tau_{\boldsymbol{n}} E^h[s_h^0, \boldsymbol{n}_h^0]$. This explains the behavior of $\mathrm{err}_{\boldsymbol{n}}$ in Table 4.2 (upper) upon refinement, which increases $E^h[s_h^0, \boldsymbol{n}_h^0]$ because $\boldsymbol{n}_h^0$ has a point defect while s_h^0 is constant and does not compensate the blow up of $\nabla \boldsymbol{n}_h^0$.

h	$E^h[s_h^N, \boldsymbol{n}_h^N]$	N	$\min(s_h^N)$	$\mathrm{err}_{\boldsymbol{n}}$	CPU time (s)
$\sqrt{2}\,2^{-5}$	2.984	60	0.0757	0.0404	64.83
$\sqrt{2}\,2^{-6}$	2.940	61	0.0422	0.0232	592.23
$\sqrt{2}\,2^{-7}$	2.939	133	0.0289	0.0100	7919.25

$\tau_{\boldsymbol{n}}$	$\mathrm{err}_{\boldsymbol{n}}$
$(0.1)2^{-5}$	0.00610
$(0.1)2^{-6}$	0.00346
$(0.1)2^{-7}$	0.001927

Table 4.2: Point defect experiment of Section 4.4.2: Final outputs of Algorithm 4.3.1 for different uniform meshes with meshsize h and time steps $\tau_{\boldsymbol{n}} = Ch^2$ (upper) and different time step sizes $\tau_{\boldsymbol{n}}$ with fixed meshsize $h = \sqrt{2}\,2^{-5}$ (lower).

4.4.3 Plane defect in 3D

We simulate a plane defect in the unit cube $\Omega = (0,1)^3$ located at $\{z = 0.5\}$, according to [76, Section 6.4]. We set $\kappa = 0.2$ in (1.3.1) and $c_{\mathrm{dw}} = 0$ in (4.4.1). We impose Dirichlet boundary conditions on the top and bottom faces Γ_D of the cube

$$g = \hat{s},\ \boldsymbol{q} = (1,0,0) \text{ on } \partial\Omega \cap \{z = 0\}, g = \hat{s},\ \boldsymbol{q} = (0,1,0) \text{ on } \partial\Omega \cap \{z = 1\}.$$

The exact solution is $\boldsymbol{n}(z) = (1,0,0)$ for $z < 0.5$ and $\boldsymbol{n}(z) = (0,1,0)$ for $z > 0.5$, while $s(z) = 0$ on $z = 0.5$ and linear on $(0, 0.5) \cup (0.5, 1)$ [76, Section 6.4]. Our numerical results are consistent with reproduce those in [59, Section 5.3]. To initialize Algorithm 4.3.1, we set $s_h^0 = \hat{s}$ and $\boldsymbol{n}_h^0$ to be a regularized point defect away from the center of the cube. Figure 4.3 displays the three components of $\boldsymbol{n}_h^k$ and s_h^k evaluated along the vertical line $(0.5, 0.5, z)$ for iterations $k = 1, 31, 79$ computed on a uniform mesh with $h = \sqrt{3}\,0.05$ and $\tau_{\boldsymbol{n}} = \tau_s = 0.01$.

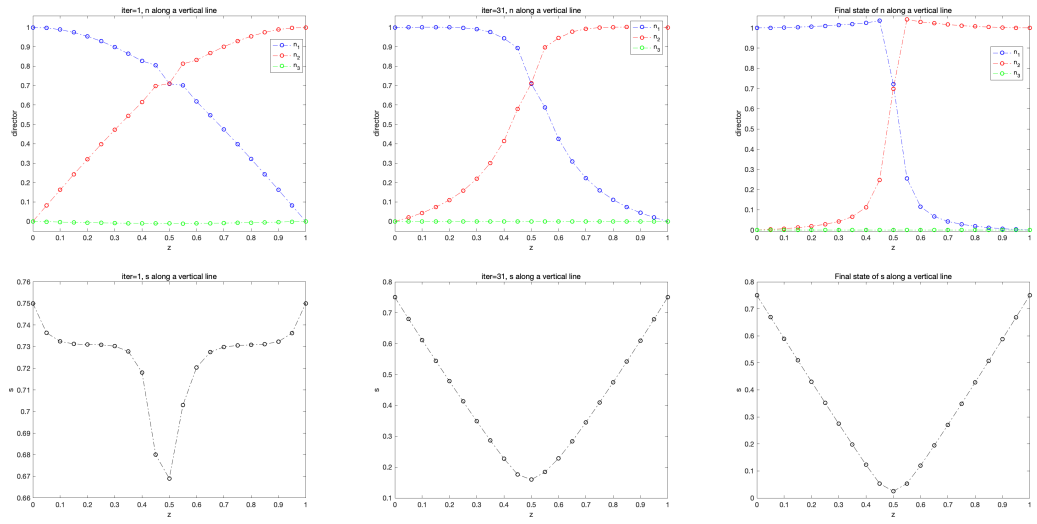

Figure 4.3: Plane defect of Section 4.4.3: Plots of the three components of $\boldsymbol{n}_h^k$ (first row) and plots of s_h^k (second row) for iterations $k = 1, 31, 79$. In the final configuration ($k = 79$), the energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 0.247$, $\min(s_h^N) = 0.0101$, and $\mathrm{err}_{\boldsymbol{n}} = 0.0556$. Moreover, there is a transition layer between about $z = 0.4$ and $z = 0.6$, and s_h is almost linear in $(0, 0.4)$ and $(0.6, 1)$.

4.4.4 Effect of κ on equilibria

The value of constant $\kappa > 0$ in (1.3.1) plays a crucial role in the formation of defects. For large values of κ, the dominant term in $E_1[s, \boldsymbol{n}]$ is $\int_\Omega \kappa |\nabla s|^2$ that prevents variations of s. Typically s tends to be close to a (usually positive) constant and the model behaves much like the simpler Oseen–Frank model, where defects are less likely to occur (and no defects with finite energy beyond point defects are allowed). On the other hand, for small values of κ, the energy is dominated by $\int_\Omega s^2 |\boldsymbol{\nabla} \boldsymbol{n}|^2$, which allows s to become zero to compensate large gradients of $\boldsymbol{n}$, and defects are then more likely to occur. In this section, we investigate this dichotomy numerically.

We consider a cylindrical domain Ω in 3D with lateral boundary Γ_D

$$\Omega = \{(x, y, z) \in \mathbb{R}^3 : (x - 0.5)^2 + (y - 0.5)^2 < 0.5^2,\ 0 < z < 1\},$$

$$\Gamma_D = \{(x, y, z) \in \mathbb{R}^3 : (x - 0.5)^2 + (y - 0.5)^2 = 0.5^2,\ 0 < z < 1\},$$

and impose the Dirichlet conditions on Γ_D

$$g = \hat{s} \quad \text{and} \quad \boldsymbol{q} = \boldsymbol{r}/g = \frac{(x - 0.5, y - 0.5, 0)}{|(x - 0.5, y - 0.5, 0)|}, \tag{4.4.7}$$

The top and bottom faces of Ω are treated as free boundaries and the double well potential ψ is neglected, i.e., $c_{\mathrm{dw}} = 0$ in (4.4.1). The analysis in [76, Section 6.5] predicts that minimizers of the energy exhibit a line defect along the central axis of the cylinder if κ is sufficiently small, whereas they are smooth (no defects) if κ is sufficiently large.

Figure 4.4 displays the final configurations obtained for $\kappa = 0.2$ and $\kappa = 2$. To discretize Ω, we consider an unstructured mesh generated by Netgen with $h_0 = 0.05$. For both values of κ, we set $\hat{s}$ as initial condition of the degree of orientation. For $\kappa = 0.2$, we set $\tau_{\boldsymbol{n}} = 0.1$ and $\tau_s = 10^{-3}$ and take as initial condition for the director field an off-center point defect located at the slice $z = 0.5$. For $\kappa = 2$, we set $\tau_{\boldsymbol{n}} = \tau_s = 0.01$ and initialize $\boldsymbol{n}_h^0$ as an off-center point defect located at the slice $z = 0.25$. These computational results are consistent with those in [59] and confirm the predicted effect of κ [76, Section 6.5].

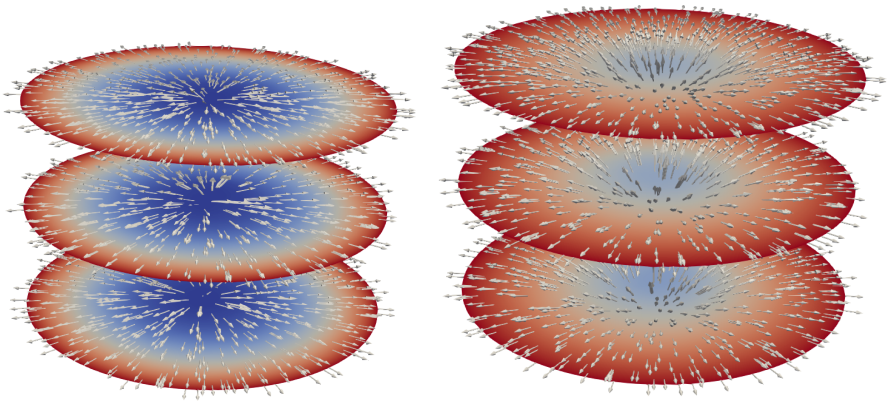

Figure 4.4: Effect of κ in Section 4.4.4: Equilibria for $\kappa = 0.2$ (left) and $\kappa = 2$ (right). Both pictures show s_h^N and $\boldsymbol{n}_h^N$ on the slices $z = 0.2, 0.5, 0.8$. If $\kappa = 0.2$, the final configuration exhibits a line defect along the central axis of the cylinder; the final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 0.806$, $\min(s_h^N) = -7.33 \times 10^{-4}$, $\mathrm{err}_{\boldsymbol{n}} = 0.0778$, and the total number of iterations N is 226. If $\kappa = 2$, the z-component of the director is not zero. This behavior is usually referred to as *fluting effect* or *escape to the third dimension* [76]. Moreover, the degree of orientation is bounded well away from zero; the final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 2.635$, $\min(s_h^N) = 0.224$, $\mathrm{err}_{\boldsymbol{n}} = 0.044$, and the total number of iterations N is 17.

4.4.5 Propeller defect

In this section, we investigate a new defect discovered in [59, Section 5.4]. We consider a setup similar to the one discussed in Section 4.4.4, except that the domain is the unit cube $\Omega = (0,1)^3$, and we again set $c_{\mathrm{dw}} = 0$ in (4.4.1). The top and bottom faces of the cube are treated as free boundary, while the same strong anchoring conditions as in (4.4.7) are imposed on the vertical faces Γ_D of the cube (lateral boundary). The initial conditions are $s_h^0 = \hat{s}$ for the degree of orientation and an off-center point defect located on the slice $z = 0.5$ for the director. The domain is discretized using an unstructured mesh generated by Netgen with $h_0 = 0.025$, and we set $\tau_{\boldsymbol{n}} = 0.02$. We consider the values $\kappa = 2$ and $\kappa = 0.1$. For $\kappa = 2$ and $\tau_s = 0.2$, the computational results agree with those of Section 4.4.4: the equilibrium state is

smooth and is characterized by a nonzero z-component (fluting effect).

For $\kappa = 0.1$, the final configuration reported in [59, Section 5.4, Figure 5] consists of two plane defects intersecting at the vertical symmetry axis of the cube, the so-called propeller defect. Whether this was a numerical artifact due to the inherent symmetries of the structured uniform weakly acute meshes used in [59] for simulation was an intriguing open question that we now answer. Owing to the flexibility of our approach regarding meshes, we repeated the experiment using an unstructured non-symmetric mesh with $\tau_s = 10^{-4}$. Our computational results confirm the emergence of the propeller defect in Figure 4.5, which in turn displays the director field $\boldsymbol{n}_h^k$ at iterations $k = 0, 1, 2766$ with colors indicating the size of s_h^k.

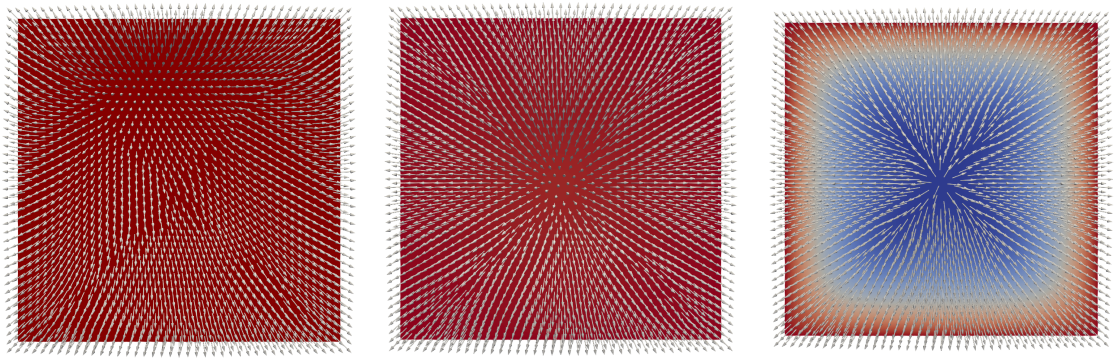

Figure 4.5: Propeller defect of Section 4.4.5: Evolution of the order parameters on the top face of the cube ($z = 1$). Plots of the initial state $(s_h^0, \boldsymbol{n}_h^0)$ (left), of the intermediate approximation $(s_h^1, \boldsymbol{n}_h^1)$ obtained after the first iteration (middle), and of the equilibrium state $(s_h^N, \boldsymbol{n}_h^N)$ after 2766 iterations (right). In the initial state, due to the off-center point defect at $z = 0.5$, there is a corresponding region on the slice for $z = 1$ where $\boldsymbol{n}$ is aligned with z-direction. After the first iteration, in which $\boldsymbol{n}$ is minimized for fixed $s = \hat{s}$, by symmetry the defect has moved to the center on $z = 0.5$. Correspondingly, on the top surface of the cube, the region where $\boldsymbol{n}$ is aligned with the z-axis has moved to the center. The final state is a propeller defect consisting of a planar X-like configuration extruded in the z-direction. The final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 0.592$, $\min(s_h^N) = -1.575 \times 10^{-4}$, $\mathrm{err}_{\boldsymbol{n}} = 0.0265$, and the total number of iterations N is 2766.

4.4.6 Colloidal effects in nematic LCs

Colloidal particles suspended in a nematic LC can induce interesting topological defects and distortions [44, 74]. One prominent example is the so-called *Saturn ring defect*, a director configuration characterized by a circular ring singularity surrounding a spherical particle and located around its equator. Such defects are typically nonorientable and captured within the Landau - de Gennes Q-tensor model [25, 26], but the Ericksen model yields similar orientable defects under suitable boundary conditions [60, 77]. We confirm the ability of Algorithm 4.3.1 to produce similar configurations.

In this section, we exploit the flexibility of Algorithm 4.3.1 regarding meshes, together with the built-in Constructive Solid Geometry (CSG) approach of Netgen/NGSolve, to explore numerically the formation of Saturn-ring-like defects induced by nonspherical or multiple particles.

4.4.6.1 One ellipsoidal particle

Let $\Omega_c = (0,1)^3$ be the unit cube and let $\Omega_s \subset \Omega_c$ be an ellipsoid centered at $(0.5, 0.5, 0.5)$ with axes parallel to the coordinate axes and semiaxis lengths equal to 0.3 (x-direction), 0.075 (y-direction), and 0.075 (z-direction); Ω_s has an aspect ratio $1:4$. The computational domain is then $\Omega := \Omega_c \setminus \overline{\Omega_s}$. We set $\kappa = 1$ in (1.3.1) as well as $c_{\mathrm{dw}} = 0.2$ in (4.4.1). On $\partial\Omega = \partial\Omega_c \cup \partial\Omega_s$, we impose strong anchoring

conditions

$$g = \hat{s} \text{ on } \partial\Omega, \quad \boldsymbol{q} = \boldsymbol{r}/g = \boldsymbol{\nu} \text{ on } \partial\Omega_s, \quad \text{and} \quad \boldsymbol{q} = \boldsymbol{r}/g = \boldsymbol{n}_{sr} \text{ on } \partial\Omega_c, \tag{4.4.8}$$

where $\boldsymbol{\nu} : \partial\Omega_s \to \mathbb{S}^{d-1}$ denotes the outward-pointing unit normal vector of Ω_s and $\boldsymbol{n}_{sr} : \partial\Omega_s \to \mathbb{S}^{d-1}$ smoothly interpolates between the constant values $(0, 0, -1)$ on the bottom face and $(0, 0, 1)$ on the top face of the cube. (see [60, Figure 11]). These boundary conditions are essential in order to induce the defect. The initial conditions for Algorithm 4.3.1 are given by

$$s_h^0 = \hat{s} \text{ in } \Omega \quad \text{and} \quad \boldsymbol{n}_h^0(z) = \begin{cases} (0,0,1) & z \in \Omega \text{ and } z_3 \geq 0.5, \\ (0,0,-1) & z \in \Omega \text{ and } z_3 < 0.5, \\ \boldsymbol{q}(z) & z \in \partial\Omega, \end{cases} \tag{4.4.9}$$

for $z = (z_1, z_2, z_3) \in \mathcal{N}_h$. Figure 4.6 displays cuts of the final configuration obtained using Algorithm 4.3.1 with an unstructured mesh with $h_0 = 0.05$ and time-step sizes $\tau_{\boldsymbol{n}} = \tau_s = 0.01$.

4.4.6.2 Multiple spherical particles

We conclude this section with two novel and challenging simulations involving multiple spherical colloidal particles. In both cases, the domain has the form $\Omega := \Omega_c \setminus \overline{\Omega_s}$, where $\Omega_c \subset \mathbb{R}^3$ denotes a simply connected domain (representing the LC container), whereas $\Omega_s \subset \Omega_c$ denotes the region occupied by spherical colloidal

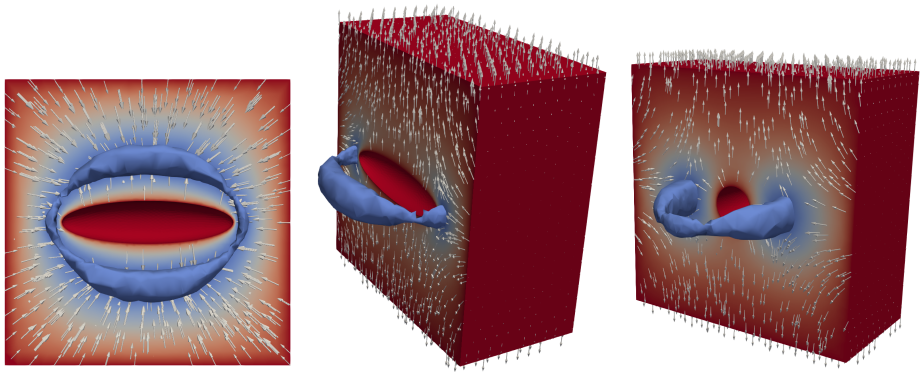

Figure 4.6: Saturn ring experiment of Section 4.4.6.1. Three different perspectives of the Saturn ring defect around an ellipsoidal particle: slice $z = 0.5$ (left), a 3D view clipped at $y = 0.5$ (middle), and a 3D view clipped at $x = 0.5$ (right). The blue ring surrounding the particle, the iso-surface for $s = 0.15$, provides a good approximation of the defect. We stress that neither the distance between the defect and the particle nor the defect diameter are constant, which is a consequence of the anisotropic shape of the particle. The final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 7.263$, $\min(s_h^N) = 0.0128$, $\text{err}_{\boldsymbol{n}} = 0.145$, and the total number of iterations N is 33.

particles. We set $\kappa = 1$ in (1.3.1) and $c_{\text{dw}} = 0.2$ in (4.4.1). Moreover, boundary and initial conditions are suitable extensions to the multiple particle case of (4.4.8) and (4.4.9) considered in Section 4.4.6.1.

Figure 4.7 shows the equilibrium state corresponding to $\Omega_c = (0, 1)^3$ and a pair of disjoint spherical colloids Ω_s with radii 0.1 and centered at $(0.3, 0.5, 0.5)$ and $(0.7, 0.5, 0.5)$. Algorithm 4.3.1 employs an unstructured mesh with $h_0 = 0.025$ and time-step sizes $\tau_{\boldsymbol{n}} = \tau_s = 0.0025$. A novel *fat figure eight* defect forms.

Figure 4.8 depicts the equilibrium state corresponding to $\Omega_c = (-0.1, 1.1)^3$ and a colloidal region consisting of six spheres. The latter have radii 0.1 and centers located at $(0.2, 0.5, 0.5)$, $(0.8, 0.5, 0.5)$, $(0.5, 0.2, 0.5)$, $(0.5, 0.8, 0.5)$, $(0.5, 0.5, 0.2)$, and $(0.5, 0.5, 0.8)$ distributed symmetrically with respect to the cube center. Algorithm 4.3.1 utilizes an unstructured mesh with $h_0 = 0.025$ and time-step sizes

$\tau_{\boldsymbol{n}} = \tau_s = 0.0025$.

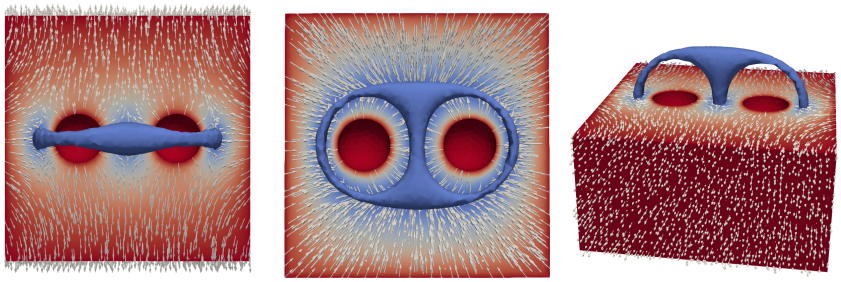

Figure 4.7: Two-particle experiment of Section 4.4.6.2. *Fat figure "8" defect* around two spherical colloids viewed from different perspectives: slice $y = 0.5$ (left), slice $z = 0.5$ (middle), and a 3D view clipped at $y = 0.5$ (right). The blue ring surrounding the particle is the iso-surface for $s = 0.12$, which provides a good approximation of the defect. The final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 7.656$, $\min(s_h^N) = 0.0146$, $\text{err}_{\boldsymbol{n}} = 0.0972$, and the total number of iterations N is 57.

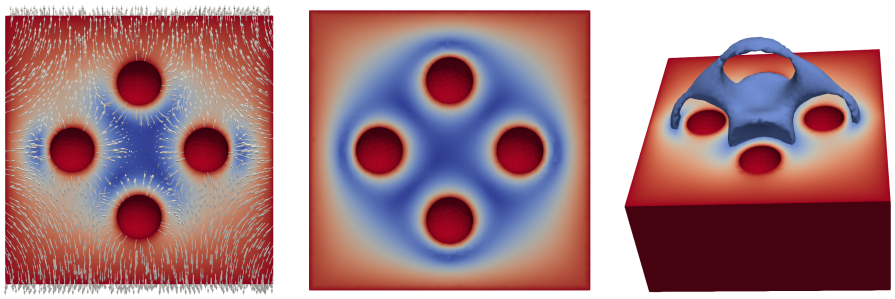

Figure 4.8: Six-particle experiment of Section 4.4.6.2. Defect around six spherical colloids viewed from different perspectives: slice $y = 0.5$ (left), slice $z = 0.5$ (middle), and a 3D view clipped at $y = 0.5$ (right); the slice $x = 0.5$ is similar to $y = 0.5$. The blue ring surrounding the particles (in the right picture) is the iso-surface for $s = 0.15$, which provides a good approximation of the defect. Therefore the defect appears to be a combination of a large Saturn ring defect around particles with center in the plane $z = 0.5$ and a planar X-like configuration rotating with axis $x = 0.5, y = 0.5, -0.1 < z < 1$. The final energy is $E^h[s_h^N, \boldsymbol{n}_h^N] = 14.703$, $\min(s_h^N) = 0.00355$, $\text{err}_{\boldsymbol{n}} = 0.160$, and the total number of iterations N is 72.

Bibliography

[1] P.J. Ackerman, J. Van De Lagemaat, and I.I. Smalyukh. Self-assembly and electrostriction of arrays and chains of hopfion particles in chiral liquid crystals. *Nature communications*, 6(1):1–9, 2015.

[2] S. Alben, B. Balakrisnan, and E. Smela. Edge effects determine the direction of bilayer bending. *Nano letters*, 11(6):2280–2285, 2011.

[3] F. Alouges. A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case. *SIAM J. Numer. Anal.*, 34(5):1708–1726, 1997.

[4] L. Ambrosio. Existence of minimal energy configurations of nematic liquid crystals with variable degree of orientation. *Manuscripta Math.*, 68(2):215–228, 1990.

[5] T. Araki and H. Tanaka. Colloidal aggregation in a nematic liquid crystal: topological arrest of particles by a single-stroke disclination line. *Physical review letters*, 97(12):127801, 2006.

[6] U. Ayachit. *The ParaView Guide: A Parallel Visualization Application*. Kitware, Inc., USA, 2015.

[7] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II – a general purpose object oriented finite element library. *ACM Trans. Math. Softw.*, 33(4):24/1–24/27, 2007.

[8] J.W. Barrett, X. Feng, and A. Prohl. 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(1):175–199, 2006.

[9] S. Bartels. Numerical analysis of a finite element scheme for the approximation of harmonic maps into surfaces. *Mathematics of computation*, 79(271):1263–1301, 2010.

[10] S. Bartels. Finite element approximation of large bending isometries. *Numer. Math.*, 124(3):415–440, 2013.

[11] S. Bartels. *Numerical methods for nonlinear partial differential equations*, volume 47 of *Springer Series in Computational Mathematics*. Springer, 2015.

[12] S. Bartels, A. Bonito, A.H. Muliana, and R.H. Nochetto. Modeling and simulation of thermally actuated bilayer plates. *J. Comput. Phys.*, 354:512–528, 2018.

[13] S. Bartels, A. Bonito, and R.H. Nochetto. Bilayer plates: Model reduction, Γ-convergent finite element approximation, and discrete gradient flow. *Comm. Pure Appl. Math.*, 70(3):547–589, 2017.

[14] S. Bartels and C. Palus. Stable gradient flow discretizations for simulating bilayer plate bending with isometry and obstacle constraints. *arXiv preprint arXiv:2004.00341*, 2020.

[15] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In *Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–109. Antwerpen, Belgium, 1997.

[16] S. Berrone, A. Bonito, R. Stevenson, and M. Verani. An optimal adaptive fictitious domain method. *Mathematics of Computation*, 88(319):2101–2134, 2019.

[17] K. Bhattacharya, M. Lewicka, and M. Schäffner. Plates with incompatible prestrain. *Arch. Rational Mech. Anal.*, 221(1):143–181, 2016.

[18] C. Blanc. Colloidal crystal ordering in a liquid crystal. *Science*, 352(6281):40–41, 2016.

[19] A. Bonito, D. Guignard, R.H. Nochetto, and S. Yang. LDG approximation of large deformations of prestrained plates. *arXiv preprint arXiv:2011.01086*, 2020.

[20] A. Bonito, D. Guignard, R.H. Nochetto, and S. Yang. Numerical analysis of the LDG approach for the approximation of the large deformations of prestrained plates. In preparation.

[21] A. Bonito and R.H. Nochetto. Quasi-optimal convergence rate of an adaptive discontinuous Galerkin method. *SIAM J. Numer. Anal.*, 48(2):734–771, 2010.

[22] A. Bonito, R.H. Nochetto, and D. Ntogkas. DG approach to large bending deformations with isometry constraint. *arXiv preprint arXiv:1912.03812 [math.NA]*, 2019.

[23] A. Bonito, R.H. Nochetto, and D. Ntogkas. Discontinuous Galerkin approach to large bending deformation of a bilayer plate with isometry constraint. *arXiv preprint arXiv:2002.00114*, 2020.

[24] A. Bonito, R.H. Nochetto, and S. Yang. LDG approximation of large deformations of bilayer plates. In preparation.

[25] J. P. Borthagaray and S. W. Walker. Chapter 5 – The Q-tensor model with uniaxial constraint. In A. Bonito and R. H. Nochetto, editors, *Geometric Partial Differential Equations - Part II*, volume 22 of *Handbook of Numerical Analysis*, pages 313–382. Elsevier, 2021.

[26] J.P. Borthagaray, R.H. Nochetto, and S.W. Walker. A structure-preserving FEM for the uniaxially constrained Q-tensor model of nematic liquid crystals. *Numer. Math.*, 145(4):837–881, 2020.

[27] S. Brenner and R. Scott. *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media, 2007.

[28] H. Brezis, J.-M. Coron, and E.H. Lieb. Harmonic maps with defects. *Communications in Mathematical Physics*, 107(4):649–705, 1986.

[29] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous finite elements for diffusion problems. *Atti Convegno in onore di F. Brioschi (Milano 1997), Istituto Lombardo, Accademia di Scienze e Lettere*, pages 197–217, 1999.

[30] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, 16(4):365–378, 2000.

[31] S. Carter, A. Rotem, and S.W. Walker. A domain decomposition approach to accelerate simulations of structure preserving nematic liquid crystal models. *Journal of Non-Newtonian Fluid Mechanics*, 283:104335, 2020.

[32] B. Cockburn and C.-W. Shu. The local discontinuous galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35(6):2440–2463, 1988.

[33] P.-G. De Gennes and J. Prost. *The physics of liquid crystals*, volume 83. Oxford university press, 1993.

[34] M.P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, 1976.

[35] E. Efrati, E. Sharon, and R. Kupferman. Elastic theory of unconstrained non-euclidean plates. *J. Mech. Phys. Solids*, 57(4):762–775, 2009.

[36] E. Efrati, E. Sharon, and R. Kupferman. Hyperbolic non-euclidean elastic strips and almost minimal surfaces. *Phys. Rev. E*, 83(4):046602, 2011.

[37] J.L. Ericksen. Liquid crystals with variable degree of orientation. *Archive for Rational Mechanics and Analysis*, 113(2):97–120, 1991.

[38] L.C. Evans and R.F. Gariepy. *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015.

[39] G. Friesecke, R.D. James, and S. Müller. A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity. *C.R. Math.*, 55(11):1461–1506, 2002.

[40] G. Friesecke, R.D. James, and S. Müller. A hierarchy of plate models derived from nonlinear elasticity by gamma-convergence. *Arch. Rational Mech. Anal.*, 180(2):183–236, 2006.

[41] G. Friesecke, S. Müller, and R.D. James. Rigorous derivation of nonlinear plate theory and geometric rigidity. *C.R. Math.*, 334(2):173–178, 2002.

[42] A. Goriely and M. Ben Amar. Differential growth and instability in elastic shells. *Phys. Rev. Lett.*, 94(19):198103, 2005.

[43] M. Gromov. *Partial Differential Relations*, volume 13. Springer-Verlag, Berlin-Heidelberg, 1986.

[44] Y. Gu and N.L. Abbott. Observation of saturn-ring defects around solid microspheres in nematic liquid crystals. *Physical Review Letters*, 85(22):4719, 2000.

[45] J. Guan, H. He, D.J. Hansford, and L.J. Lee. Self-folding of three-dimensional hydrogel microstructures. *The Journal of Physical Chemistry B*, 109(49):23134–23137, 2005.

[46] J. Guan, H. He, L.J. Lee, and D.J. Hansford. Fabrication of particulate reservoir-containing, capsulelike, and self-folding polymer microstructures for drug delivery. *Small*, 3(3):412–418, 2007.

[47] Q. Han and J.-X. Hong. *Isometric embedding of Riemannian manifolds in Euclidean spaces*, volume 13. American Mathematical Soc., 2006.

[48] P. Hornung. Approximating w2, 2 isometric immersions. *Comptes Rendus Mathematique*, 346(3-4):189–192, 2008.

[49] S. Janbaz, R. Hedayati, and A.A. Zadpoor. Programming the shape-shifting of flat soft matter: from self-rolling/self-twisting materials to self-folding origami. *Materials Horizons*, 3(6):536–547, 2016.

[50] D.-H. Kim and J.A. Rogers. Stretchable electronics: materials strategies and devices. *Advanced materials*, 20(24):4887–4892, 2008.

[51] J. Kim, J.A. Hanna, R.C. Hayward, and C.D. Santangelo. Thermally responsive rolling of thin gel strips with discrete variations in swelling. *Soft Matter*, 8(8):2375–2381, 2012.

[52] Y. Klein, E. Efrati, and E. Sharon. Shaping of elastic sheets by prescription of non-euclidean metrics. *Science*, 315(5815):1116–1120, 2007.

[53] M. Lewicka and M.R. Pakzad. Scaling laws for non-euclidean plates and the $W^{2,2}$ isometric immersions of Riemannian metrics. *ESAIM: Contr. Optim. C.A.*, 17(4):1158–1173, 2011.

[54] F.-H. Lin. On nematic liquid crystals with variable degree of orientation. *Comm. Pure Appl. Math.*, 44(4):453–468, 1991.

[55] S.-Y. Lin and M. Luskin. Relaxation methods for liquid crystal problems. *SIAM Journal on Numerical Analysis*, 26(6):1310–1324, 1989.

[56] C.D. Modes, K. Bhattacharya, and M. Warner. Disclination-mediated thermo-optical response in nematic glass sheets. *Phys. Rev. E*, 81(6):060701, 2010.

[57] C.D. Modes, K. Bhattacharya, and M. Warner. Gaussian curvature from flat elastica sheets. *Proc. Royal Soc.*, 467(2128):1121–1140, 2010.

[58] R.H. Nochetto, M. Ruggeri, and S. Yang. Gamma-convergent projection-free finite element methods for nematic liquid crystals: The Ericksen model. *arXiv preprint arXiv:2103.13926*, 2021.

[59] R.H. Nochetto, S.W. Walker, and W. Zhang. A finite element method for nematic liquid crystals with variable degree of orientation. *SIAM J. Numer. Anal.*, 55(3):1357–1386, 2017.

[60] R.H. Nochetto, S.W. Walker, and W. Zhang. The ericksen model of liquid crystals with colloidal and electric effects. *jcp*, 352:568–601, 2018.

[61] D. Ntogkas. *Non-linear geometric PDEs: algorithms, numerical analysis and computation*. PhD thesis, 2018.

[62] D.A. Di Pietro and A. Ern. Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible navier–stokes equations. *Math. Comp.*, 79(271):1303–1330, 2010.

[63] D.A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Mathématiques et Applications. Springer Berlin Heidelberg, 2011.

[64] E.G. Poznyak and E.V. Shikin. Small parameters in the theory of isometric imbeddings of two-dimensional riemannian manifolds in euclidean spaces. *J. Math. Sci.*, 74(3):1078–1116, 1995.

[65] T. Pryer. Discontinuous Galerkin methods for the p-biharmonic equation from a discrete variational perspective. *Electron. Trans. Numer. Anal.*, 41:328 – 349, 2014.

[66] D. Schmaljohann. Thermo-and ph-responsive polymers in drug delivery. *Advanced drug delivery reviews*, 58(15):1655–1670, 2006.

[67] B. Schmidt. Minimal energy configurations of strained multi-layers. *Calc. Var. Partial Differential Equations*, 30(4):477–497, 2007.

[68] B. Schmidt. Plate theory for stressed heterogeneous multilayers of finite bending energy. *J. Math. Pures Appl.*, 88(1):107–122, 2007.

[69] J. Schöberl et al. Netgen/ngsolve, 2017.

[70] R. Schoen, K. Uhlenbeck, et al. A regularity theory for harmonic maps. *Journal of Differential Geometry*, 17(2):307–335, 1982.

[71] E. Sharon and E. Efrati. The mechanics of non-euclidean plates. *Soft Matter*, 6(22):5693–5704, 2010.

[72] B. Simpson, G. Nunnery, R. Tannenbaum, and K. Kalaitzidou. Capture/release ability of thermo-responsive polymer particles. *Journal of Materials Chemistry*, 20(17):3496–3501, 2010.

[73] J.S. Sodhi and I.J. Rao. Modeling the mechanics of light activated shape memory polymers. *International Journal of Engineering Science*, 48(11):1576–1589, 2010.

[74] H. Stark. Director field configurations around a spherical particle in a nematic liquid crystal. *The European Physical Journal B-Condensed Matter and Complex Systems*, 10(2):311–321, 1999.

[75] G. Stoychev, N. Puretskiy, and L. Ionov. Self-folding all-polymer thermoresponsive microcapsules. *Soft Matter*, 7(7):3277–3279, 2011.

[76] E.G. Virga. *Variational theories for liquid crystals*. CRC Press, 2018.

[77] S.W. Walker. A finite element method for the generalized Ericksen model of nematic liquid crystals. *ESAIM Math. Model. Numer. Anal.*, 54(4):1181–1220, 2020.

[78] S.M. Wise, C. Wang, and J.S. Lowengrub. An energy-stable and convergent finite-difference scheme for the phase field crystal equation. *SIAM J. Numer. Anal.*, 47(3):2269–2288, 2009.

[79] Z.L. Wu, M. Moshe, J. Greener, H. Therien-Aubin, Z. Nie, E. Sharon, and E. Kumacheva. Three-dimensional shape transformations of hydrogel sheets induced by small-scale modulation of internal stresses. *Nat. Commun.*, 4:1586, 2013.

[80] A. Yavari. A geometric theory of growth mechanics. *J. Nonlinear Sci.*, 20(6):781–830, 2010.