Minimal energy configurations of bilayer plates as a Polynomial Optimization Problem

Preetham Mohan* Nung Kwan Yip † Thomas Yu‡

October 15, 2021

Abstract:

We develop a discretization method for solving the minimal energy configuration of bilayer plates based on a mathematical model developed in [22, 6]. Our discretization method employs C^1-spline functions. A highlight of the method involves a trick to handle the nonlinear isometry constraint in such a way that not only numerical integration becomes unnecessary, but also that the final optimization problems are in the form of degree 4 polynomial optimization problems (POP). We develop two different versions of the method, one resulted in a constrained degree 4 POP involving a small tolerance ε, another resulted in an unconstrained degree 4 POP involving a large penalty parameter μ. We develop a mathematical analysis, based on the direct method and techniques in Γ-convergence, to show how ε and μ can be chosen according to the grid size so that the minimizers of the discrete problems converge to that of the continuum variational problem as the grid size goes to zero. We corroborate the theory through a series of computational experiments, and also report an unexpected finding related to the asymmetry of the discretized problems.

Acknowledgments. TY thanks Amir Beck, Thomas Duchamp, Rehka Thomas and Larry Schumaker for helpful discussions. He is supported in part by the National Science Foundation grants DMS 0512673 and DMS 0915068. PM thanks the hospitality of the departments of mathematics at both Drexel University and University of Michigan during his pursuit of the project.

Keywords: Bilayer plates, Splines, Calculus of variation, Γ-convergence, Sparse Polynomial optimization problems, Nonconvex optimization problems

1 Introduction

Bilayer materials appear in the ubiquitous lipid bilayer forming the cell membranes of almost all organisms as well as in many engineering applications [28, 27, 24, 3]. In the latter realm, *bilayer plates* consist of two films of different materials glued together. The films often have two different lattice constants or material properties. Such a mismatch can then lead to internal stress or different mechanical, thermal or electrical responses. Large deformations can form and also be controlled

*Department of Mathematics, University of Michigan. Email: preetham@umich.edu.
†Department of Mathematics, Purdue University. Email: yipn@purdue.edu.
‡Department of Mathematics, Drexel University. Email: yut@drexel.edu.

by external forcings. This phenomena is thus very useful in the design and manufacturing of nano-devices such as nanorolls, microgrippers, and nano-tubes. Hence accurate modelling and efficient numerical simulation of such a mechanism is extremely beneficial.

In [22, 6], a mathematical model for the bending of bilayers plates is derived based on hyperelasticity. The model consists of minimizing the dimensionally reduced elastic energy

$$E[\mathbf{y}] = \frac{1}{2}\iint_{\omega} \|H - Z\|_F^2, \quad H = \text{second fundamental form of } \mathbf{y}, \tag{1.1}$$

within the set of *isometries* $\mathbf{y} : \omega \to \mathbb{R}^3$, i.e. mappings satisfying $d\mathbf{y}^T d\mathbf{y} = \mathbb{I}_{2\times 2}$ in the planar domain $\omega \subset \mathbb{R}^2$, and with prescribed values $\mathbf{y}|_{\partial_D\omega} = \mathbf{y}_D$, $d\mathbf{y}|_{\partial_D\omega} = \Phi_D$ on the Dirichlet portion $\partial_D\omega$ of the boundary $\partial\omega$. The symbol Z refers to a constant 2×2 symmetric matrix and can be viewed as a *spontaneous curvature*, and its presence attributes to the pre-stressed configuration of the plate. The above model can also be deduced, or more accurately, *reduced* from genuine three dimensional elasticity theory [14, 13, 15, 16] and also from atomistic models [23]. On the other hand, the formula in (1.1) can be further simplified on the space of isometries. This will be very useful in the formulation of our numerical method.

In this article, we focus on the computation of the the equilibrium shape of such a bilayer plate. The main challenges are due to the high-order (4-th order) of the equation and the nonlinear isometry constraint. These issues have been investigated in the framework of non-conforming finite elements [5, 4, 6]. Here we employ spline elements which belong to the *energy space*, namely $W^{2,2}(\omega)$, for our current problem. In this sense, our method is *conforming* in FEM parlance. However, this 'conforming' characterization comes with a caveat: spline functions, being piecewise polynomials, cannot be perfect isometries, except in trivial cases. To circumvent this difficulty, we introduce a functional $\mathcal{I}$ that measures the discrepancy of a spline surface $\mathbf{y} = \sum_I \mathbf{c}_I B_I$ from an isometry. (Here and below, B_I is any basis for our spline space, and $\mathbf{c}$ is the coefficient vector of any spline function.) Our final discretized version of (1.1) takes the simple form of

$$\text{(i) an '}\varepsilon\text{-problem':} \quad \min_{\mathbf{c}} \mathcal{E}(\mathbf{c}) \quad \text{s.t. } \mathcal{I}(\mathbf{c}) \le \varepsilon, \tag{1.2}$$

where $\varepsilon > 0$ is a suitably chosen small tolerance, or

$$\text{(ii) a '}\mu\text{-problem':} \quad \min_{\mathbf{c}} \mathcal{E}(\mathbf{c}) + \mu\,\mathcal{I}(\mathbf{c}), \tag{1.3}$$

for an appropriately chosen large penalization parameter μ.

Our use of C^1-spline elements (which are in $W^{2,2}$) is in contrast to the approach advocated in [6], which *deliberately avoids* C^1-elements. Arguably, the use of spline elements is not harder, if not easier, than the use of non-conforming Kirchhoff elements employed in [6], for both the design and the mathematical analysis of the algorithm.[1]

The next section documents the core technical contributions of this paper.

1.1 Contributions

Our core discovery is that, in dealing with the aforementioned difficulty in handling the isometry constraint, we can commit a sort of 'variational crime' in the design of the discretization method

[1] For general domains and higher accuracy orders, it remains to see if the use of non C^1-elements result in algorithms that are easier to *implement* than the spline counterparts. For more comparison of various C^1-spline elements with the merely C^0-Kirchhoff elements, see Section 2.5.

without 'getting into trouble' in the analysis of Γ-convergence. Let us elaborate: the second fundamental form H in (1.1) involves terms such as $(d\mathbf{y}^T d\mathbf{y})^{-1}$ and division by $\|\partial_1\mathbf{y} \times \partial_2\mathbf{y}\|$, which can be ignored if $\mathbf{y}$ were a perfect isometry. As spline functions $\mathbf{y}$ are in most cases not isometries, an honest computation of the energy functional $E[\mathbf{y}]$ would require a numerical integration scheme coupled with the evaluation of the relatively complicated integrand at quadrature points.[2] However, since we constrain $\mathcal{I}[\mathbf{y}]$ to be small anyway, we may commit the misdemeanor of computing $E[\mathbf{y}]$ as if $(d\mathbf{y}^T d\mathbf{y})^{-1}$ is the identity and $\|\partial_1\mathbf{y} \times \partial_2\mathbf{y}\|$ equals 1.

A pleasant surprise arises: the resulting energy functional, called $\mathcal{E}$ instead of E due to the 'misdemeanor', becomes a degree 6 polynomial function in the control coefficients $\mathbf{c}$ of $\mathbf{y}$. With one more similar trick, we can further replace E by a degree 3 polynomial function. This degree 3 polynomial function, denoted by $\mathcal{E}_3$, will be derived in Section 2.3.

It is immediate to see from our definition of $\mathcal{I}$ that it is also a polynomial – but degree 4 – in the control variables $\mathbf{c}$. As a result, both formulations (1.2) and (1.3) lead to polynomial optimization problems. These high-dimensional polynomials are also sparse in the sense that they involve only $O(N)$ monomials, instead of $O(N^D)$ monomials in 'dense' degree D polynomials in N variables. This opens us to the possibility of using the many proposed convexification methods for POPs; see, for example [31, 30, 2, 7] and the references therein. (This direction, however, is not pursued in the current paper for reasons articulated in Section 5.1.)

Not only would our formulations of $\mathcal{I}(\mathbf{c})$ and $\mathcal{E}_3(\mathbf{c})$ rid us from the need of troublesome numerical integrations, but also that the coefficients of the polynomials $\mathcal{I}(\mathbf{c})$ and $\mathcal{E}_3(\mathbf{c})$ can be *pre-computed in closed-forms*. See Sections 2.2 and 2.3. This significantly speeds up the overall optimization process, as those coefficients need not be re-computed when $\mathbf{c}$ is updated repeatedly in the course of optimization.

We further prove the convergence of the methods within the framework of Γ-convergence of functionals when ε and μ are chosen appropriately based on the grid size h associated with the spline spaces. To this end, we need to combine ideas from the theory of classical Γ-convergence with the approximation theory of splines. The theoretical results also prove to be useful as they suggest concrete applicable choices of ε and μ based on h that we can easily use in our simulations.

We carry out a series of numerical computations in Section 5. Besides corroborating the theory, our computations reveal a few properties of our discretization method. In particular, we found cases where the global minimizer of the discrete problem appears to be asymmetric, while the solution of the continuous problem is symmetric. Such a symmetry breaking is also observed in related geometric variational problems, see for example [10, Section 5.3].

1.2 Organization

In Section 2, we first consider the case when ω is a rectangle and a bilayer plate is approximated by tensor product splines. The key contribution is a numerical method, as well as its analysis, for the variational problem (1.1) in the form of a degree 4 sparse *polynomial optimization problem* (POP). The assumption of ω being a rectangle and the use of tensor product spline are only for concreteness and simplicity, as neither of them is essential for the sparse POP formulation. In Section 2.5, we highlight the basic ingredients leading to the sparse POP formulation and outline

[2] This is the case in our earlier work [10], in which the Willmore energy of a subdivision surface has to be computed using numerical integration.

algorithms for dealing with more general domains. In Section 4, we establish existence and Γ-convergence results for two different versions of the sparse POP method. The Γ-convergence results imply that the discrete minimizers from our proposed POPs converge, as the discretization goes to zero, to a minimizer of the continuum variational bilayer problem (1.1). In Section 5, we present computational results of our methods. In Section 6, we provide a summary and some future perspectives.

2 Formulation as Sparse Polynomial Optimization Problems (POP)

In this section, we develop numerical methods for solving the variational problem

$$\begin{aligned} &\min_{\mathbf{y}} E[\mathbf{y}] \\ \text{s.t.}\quad &d\mathbf{y}^T d\mathbf{y} = \mathbb{I}_{2\times 2} \text{ on } \omega, \\ &\mathbf{y}|_{\partial_D\omega}(u,v) = \big[u,v,0\big]^T, \quad \frac{\partial \mathbf{y}}{\partial \hat{\mathbf{n}}}\Big|_{\partial_D\omega}(u,v) = \hat{\mathbf{n}}. \end{aligned} \tag{2.1}$$

In the above, ω is a subset of $\mathbb{R}^2$. We shall assume the $\mathbf{y} : \omega \to \mathbb{R}^3$ above lies in the Sobolev space $W^{2,2}(\omega;\mathbb{R}^3)$; see Section 4. In the boundary condition, $\hat{\mathbf{n}}$ is the unit normal vector pointing towards the interior of ω, this means the plate $\mathbf{y}$ is clamped at the part $\partial_D\omega$ of the boundary of ω. We shall refer to this as a *clamped boundary condition*. Since both E and the isometry condition are invariant under rigid motions, the clamped boundary condition has the effect of avoiding an obvious non-uniqueness of solution in the problem.

In this section, we focus on the case of ω being a rectangle.[3] More specifically, let

$$\omega := [0,A]\times[0,B].$$

For simplicity we assume

$$\partial_D\omega = \{(0,v) : 0 \le v \le B\}. \tag{2.2}$$

The clamped boundary condition in (2.1) can then be expressed as

$$\mathbf{y}(0,v) = \big[0,v,0\big]^T, \quad \frac{\partial \mathbf{y}}{\partial u}(0,v) = \big[1,0,0\big]^T, \quad 0 \le v \le B. \tag{2.3}$$

The proposed numerical methods will be recorded in Section 2.4. The next three subsections develop the basic ingredients.

2.1 Tensor product splines

For a rectangular domain, we use tensor product uniform knot splines to approximate the bilayer plate $\mathbf{y}$. Consider the space S_m^d defined by

$$S_m^d := \{f : [0,m] \to \mathbb{R} : f \in C^{d-1},\ \ f|_{[i,i+1]} \text{ is a polynomial of degree} \le d,\ i = 0,1,\dots,m-1\}.$$

[3] For our theoretical development in the next section, we fall back to a general domain satisfying (2.2).

It is easy to see that $\dim S_m^d = m+d$. We are mostly interested in the case of $d = 2, 3$. The classical result of Curry and Schoenberg [25] asserts that there is a localized basis for S_m^d, called B-splines, denoted here by

$$\begin{aligned} \phi_{-1},\, \phi_0,\, \phi_1, \ldots, \phi_m \text{ for } d = 2, \text{ or} \\ \phi_{-1},\, \phi_0,\, \phi_1, \ldots, \phi_m,\, \phi_{m+1} \text{ for } d = 3. \end{aligned} \tag{2.4}$$

In the uniform knot case considered here, we can choose ϕ_i to be

$$\phi_i(x) = B(x - i),$$

where

$$B = \begin{cases} 1_{[0,1]} * 1_{[0,1]} * 1_{[0,1]}(\cdot + 1), & d = 2 \\ 1_{[0,1]} * 1_{[0,1]} * 1_{[0,1]} * 1_{[0,1]}(\cdot + 2), & d = 3. \end{cases}$$

Note that ϕ_i is symmetric about $i+c$ for $c = 0$ when d is odd and $c = 1/2$ when d is even. Moreover, we have

$$\sum_i \phi_i(x) \equiv 1, \quad \sum_i (i + c)\phi_i(x) \equiv x, \quad x \in [0, m]. \tag{2.5}$$

Denote by $\tau : (0, m) \times (0, n) \to \omega$ the linear map $\tau(x, y) = (h_1 x, h_2 y)$ where

$$h_1 := A/m, \quad h_2 := B/n$$

are the grid sizes in the two dimensions. With ω fixed, write

$$\mathbf{h} = (h_1, h_2), \quad h = \max\{h_1, h_2\}, \quad B_{i,j}^{\mathbf{h}}(u, v) := \phi_i(u/h_1)\phi_j(v/h_2) = ((\phi_i\phi_j) \circ \tau^{-1})(u, v)$$

and

$$\mathcal{S}_{\mathbf{h}} := \mathcal{S}_{\mathbf{h},d} := \Big\{ \mathbf{y} : \omega \to \mathbb{R}^3 \mid \mathbf{y} = \sum_i \sum_j \mathbf{c}_{i,j} B_{i,j}^{\mathbf{h}}, \;\; \mathbf{c}_{i,j} \in \mathbb{R}^3 \Big\}. \tag{2.6}$$

We call $\mathbf{c} = (\mathbf{c}_{i,j})$ the control points or control coefficients. Note that $\mathcal{S}_{\mathbf{h},d} \subset C^{d-1,1}(\omega; \mathbb{R}^3)$ for any $d \geq 0$ and

$$\mathcal{S}_{\mathbf{h},d} \subset C^{d-1,1}(\omega; \mathbb{R}^3) \subset W^{2,2}(\omega; \mathbb{R}^3) \quad \text{when } d \geq 2. \tag{2.7}$$

We now state what the boundary condition (2.3) means to the control coefficients when $d = 3$ and $d = 2$:

- When $\mathbf{y} \in \mathcal{S}_{\mathbf{h},3}$, the first condition of (2.3) becomes

$$\mathbf{y} \circ \tau(0, h_2^{-1} v) = \sum_{i=-1}^{1} \sum_{j=-1}^{n+1} \mathbf{c}_{i,j} \phi_i(0) \phi_j(h_2^{-1} v) = \big[0, v, 0\big]^T$$

  which simplifies to

$$\frac{1}{6} \sum_{j=-1}^{n+1} \phi_j(y)[\mathbf{c}_{-1,j} + 4\mathbf{c}_{0,j} + \mathbf{c}_{1,j}] = \big[0, h_2 y, 0\big]^T, \tag{2.8}$$

whereas the second condition of (2.3) becomes

$$(\frac{\partial \mathbf{y}}{\partial u} \circ \tau)(0, h_2^{-1} v) = \sum_{i=-1}^{1} \sum_{j=-1}^{n+1} \mathbf{c}_{i,j} \phi_i'(0) \phi_j(h_2^{-1} v) = \left[h_1, 0, 0\right]^T$$

which simplifies to

$$\frac{1}{2} \sum_{j=-1}^{n+1} \phi_j(y)[-\mathbf{c}_{-1,j} + \mathbf{c}_{1,j}] = \left[h_1, 0, 0\right]^T. \tag{2.9}$$

Hence (2.3) holds if and only if

$$\mathbf{c}_{-1,j} = -2\mathbf{c}_{0,j} + \left[-h_1,\ 3jh_2,\ 0\right]^T,\ \mathbf{c}_{1,j} = -2\mathbf{c}_{0,j} + \left[h_1,\ 3jh_2,\ 0\right]^T, \quad -1 \le j \le n+1. \tag{2.10}$$

- When $\mathbf{y} \in \mathcal{S}_{\mathbf{h},2}$, a similar calculation shows that the clamp condition (2.3) holds if and only if

$$\mathbf{c}_{-1,j} = \left[-h_1/2,\ (j+1/2)\, h_2,\ 0\right]^T, \quad \mathbf{c}_{0,j} = \left[h_1/2,\ (j+1/2)\, h_2,\ 0\right]^T, \quad -1 \le j \le n. \tag{2.11}$$

2.2 Isometry functional $\mathcal{I}$

Note that the spline space contains the following trivial isometry: if we set $\mathbf{c} = \mathbf{c}^\flat$ with

$$\mathbf{c}^\flat_{i,j} := \begin{bmatrix} (i+c)A/m \\ (j+c)B/n \\ 0 \end{bmatrix},\ c = \begin{cases} 0, & (\text{for } d \text{ odd}), \\ \frac{1}{2}, & (\text{for } d \text{ odd}), \end{cases} \quad \text{then} \quad \mathbf{y}(u,v) = \begin{bmatrix} u \\ v \\ 0 \end{bmatrix} =: \mathbf{y}^\flat \tag{2.12}$$

in virtue of (2.5). We refer to this as the *flat plate* and denote it by $\mathbf{y}^\flat$. Of course, we can also apply an arbitrary rigid motion to the above control data to get other trivial isometries.

However, note that the spline space does not contain any isometries other than the trivial ones. Because of this, we define a functional $\mathcal{I}$ which measures the deviation of a given $\mathbf{y}$ from isometry based on the following observation:

$$\mathcal{I} : W^{2,2}(\omega) \longrightarrow \mathbb{R}_+, \quad \mathcal{I}[\mathbf{y}] := \iint_\omega \left\| d\mathbf{y}^T d\mathbf{y} - \mathbb{I}_{2\times 2} \right\|_F^2 du\, dv. \tag{2.13}$$

We often identify $\mathcal{S}_{\mathbf{h},d}$ with the control points $\mathbf{c} \in \mathbb{R}^{(m+d)\times(n+d)\times 3}$ and write $\mathcal{I}(\mathbf{c})$ instead of $\mathcal{I}[\mathbf{y}]$; as such, we have $\mathcal{I} : \mathbb{R}^{(m+d)\times(n+d)\times 3} \to \mathbb{R}$.

We now derive a formula for $\mathcal{I}(\mathbf{c})$. Write

$$\overline{\mathbf{y}} = \mathbf{y} \circ \tau : [0, m] \times [0, n] \to \mathbb{R}^3,$$

where τ is the linear change of variable as in (2.6), and

$$s_{i,j} := [i, i+1] \times [j, j+1].$$

Then

$$\begin{aligned}\mathcal{I}(\mathbf{c}) &= \iint_{\omega} \left\| d\mathbf{y}^T d\mathbf{y} - \mathbb{I}_{2\times 2} \right\|_F^2 du\, dv = \det\tau \iint_{[0,m]\times[0,n]} \left\| \tau^{-T} d\overline{\mathbf{y}}^T d\overline{\mathbf{y}}\, \tau^{-1} - \mathbb{I}_{2\times 2} \right\|_F^2 dx\, dy \\ &= h_1 h_2 \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \iint_{s_{i,j}} \sum_{p,q=1,2} \left[h_p^{-1} h_q^{-1} \langle \partial_p \overline{\mathbf{y}}, \partial_q \overline{\mathbf{y}} \rangle - \delta_{p,q} \right]^2 dx\, dy \\ &= h_1 h_2 \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} \Bigg\{ \sum_{p,q=1,2} h_p^{-2} h_q^{-2} \iint_{s_{i,j}} \langle \partial_p \overline{\mathbf{y}}, \partial_q \overline{\mathbf{y}} \rangle^2\, dx\, dy - \\ &\qquad\qquad 2 \sum_{p=1,2} h_p^{-2} \iint_{s_{i,j}} \langle \partial_p \overline{\mathbf{y}}, \partial_p \overline{\mathbf{y}} \rangle\, dx\, dy + 2 \Bigg\}.\end{aligned} \tag{2.14}$$

We now exploit the shift-invariant and tensor product structure of the spline space. The spline function $\overline{\mathbf{y}}$ restricted to any square $[i, i+1] \times [j, j+1]$ is determined by the control points $\mathbf{c}_{i',j'}$ for $i' \in i + \mathcal{N}(d)$, $j' \in j + \mathcal{N}(d)$, where $\mathcal{N}(d)$ is a fixed set of indices dependent only on the degree d of the spline space. We have

$$\mathcal{N}(2) = \{-1, 0, 1\}, \quad \mathcal{N}(3) = \{-1, 0, 1, 2\}.$$

Now we can express the first integral in the last line for $\mathcal{I}$ above as

$$\begin{aligned}\iint_{s_{i,j}} \langle \partial_p \overline{\mathbf{y}}, \partial_q \overline{\mathbf{y}} \rangle^2\, dx\, dy = \int_i^{i+1}\int_j^{j+1} \Bigg\langle & \sum_{\substack{i_1 \in i+\mathcal{N}(d) \\ j_1 \in j+\mathcal{N}(d)}} \mathbf{c}_{i_1,j_1} B^{(\epsilon(p))}(x - i_1) B^{(1-\epsilon(p))}(y - j_1), \\ & \sum_{\substack{i_2 \in i+\mathcal{N}(d) \\ j_2 \in j+\mathcal{N}(d)}} \mathbf{c}_{i_2,j_2} B^{(\epsilon(q))}(x - i_2) B^{(1-\epsilon(q))}(y - j_2) \Bigg\rangle^2 dy\, dx,\end{aligned}$$

where

$$\epsilon : \{1, 2\} \to \{0, 1\}, \quad \epsilon(1) = 1, \quad \epsilon(2) = 0. \tag{2.15}$$

If we further define

$$T^4_{k_1,k_2,k_3,k_4,\epsilon_1,\epsilon_2} := \int_0^1 B^{(\epsilon_1)}(t - k_1) B^{(\epsilon_2)}(t - k_2) B^{(\epsilon_1)}(t - k_3) B^{(\epsilon_2)}(t - k_4)\, dt \tag{2.16}$$

for $k_1, \ldots, k_4 \in \mathcal{N}(d)$, $\epsilon_1, \epsilon_2 = 0, 1$, then

$$\begin{aligned}&\iint_{s_{i,j}} \langle \partial_p \overline{\mathbf{y}}, \partial_q \overline{\mathbf{y}} \rangle^2\, dx\, dy \\ &= \sum_{\substack{k_1,\ldots,k_4 \in \mathcal{N}(d) \\ l_1,\ldots,l_4 \in \mathcal{N}(d)}} \langle \mathbf{c}_{J_1}, \mathbf{c}_{J_2} \rangle \langle \mathbf{c}_{J_3}, \mathbf{c}_{J_4} \rangle T^4_{k_1,k_2,k_3,k_4,\epsilon(p),\epsilon(q)} T^4_{l_1,l_2,l_3,l_4,1-\epsilon(p),1-\epsilon(q)}, \quad J_\ell = (i, j) + (k_\ell, l_\ell).\end{aligned} \tag{2.17}$$

This in particular shows that the first integral in (2.14) is a degree 4 homogeneous polynomial in the variables of $\mathbf{c}$. Similarly, the second integral in the last line of (2.14) is a degree 2 homogeneous polynomial in $\mathbf{c}$, and can be expressed as

$$\iint\limits_{s_{i,j}} \langle \partial_p \overline{\mathbf{y}}, \partial_p \overline{\mathbf{y}} \rangle \, dx\, dy = \sum_{\substack{k_1,k_2 \in \mathcal{N}(d) \\ l_1,l_2 \in \mathcal{N}(d)}} \langle \mathbf{c}_{J_1}, \mathbf{c}_{J_2} \rangle T^2_{k_1,k_2,\epsilon(p)} T^2_{l_1,l_2,1-\epsilon(p)}, \quad J_\ell = (i,j) + (k_\ell, l_\ell), \tag{2.18}$$

where

$$T^2_{k_1,k_2,\epsilon} := \int_0^1 B^{(\epsilon)}(t-k_1) B^{(\epsilon)}(t-k_2)\, dt. \tag{2.19}$$

Putting everything together, we can write the isometry functional as:

$$\begin{aligned} \mathcal{I}(\mathbf{c}) &= \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \mathcal{I}_{\text{loc}}(P_{i,j}(\mathbf{c})), \\ \mathcal{I}_{\text{loc}}(\mathbf{x}) &= \sum_{I_1,I_2,I_3,I_4 \in \mathcal{N}(d)^2} \alpha^4_{I_1,I_2,I_3,I_4} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle \langle \mathbf{x}_{I_3}, \mathbf{x}_{I_4} \rangle + \sum_{I_1,I_2 \in \mathcal{N}(d)^2} \alpha^2_{I_1,I_2} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle + 2h_1h_2, \end{aligned} \tag{2.20}$$

where

$$P_{i,j} : \mathbb{R}^{(m+|\mathcal{N}(d)|-1)\times(n+|\mathcal{N}(d)|-1)\times 3} \to \mathbb{R}^{|\mathcal{N}(d)|\times|\mathcal{N}(d)|\times 3}$$

maps the global control vertices, $\mathbf{c}$, to the local control vertices, denoted by $\mathbf{x}$, that contribute to the face indexed by (i,j), and

$$\begin{aligned} \alpha^4_{I_1,I_2,I_3,I_4} = \alpha^4_{(k_1,l_1),(k_2,l_2),(k_3,l_3),(k_4,l_4)} &:= \sum_{\epsilon_1,\epsilon_2=0,1} h_1^{1-2(\epsilon_1+\epsilon_2)} h_2^{1-2(2-\epsilon_1-\epsilon_2)} T^4_{k_1,k_2,k_3,k_4,\epsilon_1,\epsilon_2} T^4_{l_1,l_2,l_3,l_4,1-\epsilon_1,1-\epsilon_2}, \\ \alpha^2_{I_1,I_2} = \alpha^2_{(k_1,l_1),(k_2,l_2)} &:= -2 \sum_{\epsilon=0,1} h_1^{1-2\epsilon} h_2^{1-2(1-\epsilon)} T^2_{k_1,k_2,\epsilon} T^2_{l_1,l_2,1-\epsilon}. \end{aligned} \tag{2.21}$$

We can speed up the computation of $\mathcal{I}_{\text{loc}}$ as follows: Let $\leq$ be any ordering imposed on $\mathcal{N}(d) \times \mathcal{N}(d)$, then set

$$\mathcal{I}_2 := \{(I_1, I_2) : I_1, I_2 \in \mathcal{N}(d) \times \mathcal{N}(d),\ I_1 \leq I_2\}.$$

Next, let $\preceq$ be any ordering on $\mathcal{I}_2$, and set

$$\mathcal{I}_4 := \{(I_1,I_2,I_3,I_4) : (I_1,I_2), (I_3,I_4) \in \mathcal{I}_2,\ (I_1,I_2) \preceq (I_3,I_4)\}.$$

We have

$$\begin{aligned} \mathcal{I}_{\text{loc}}(\mathbf{x}) &= \sum_{(I_1,I_2,I_3,I_4) \in \mathcal{I}_2 \times \mathcal{I}_2} \tilde{\beta}^4_{I_1,I_2,I_3,I_4} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle \langle \mathbf{x}_{I_3}, \mathbf{x}_{I_4} \rangle + \sum_{(I_1,I_2) \in \mathcal{I}_2} \beta^2_{I_1,I_2} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle + 2h_1h_2 \\ &= \sum_{(I_1,I_2,I_3,I_4) \in \mathcal{I}_4} \beta^4_{I_1,I_2,I_3,I_4} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle \langle \mathbf{x}_{I_3}, \mathbf{x}_{I_4} \rangle + \sum_{(I_1,I_2) \in \mathcal{I}_2} \beta^2_{I_1,I_2} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle + 2h_1h_2, \end{aligned} \tag{2.22}$$

$$\text{where}\quad \beta^2_{I_1,I_2} = (2-\delta_{I_1,I_2})\alpha^2_{I_1,I_2},$$
$$\tilde\beta^4_{I_1,I_2,I_3,I_4} = \begin{cases} \alpha^4_{I_1,I_2,I_3,I_4} + \alpha^4_{I_2,I_1,I_3,I_4} + \alpha^4_{I_1,I_2,I_4,I_3} + \alpha^4_{I_2,I_1,I_4,I_3} & \text{if } I_1<I_2,\ I_3<I_4 \\ \alpha^4_{I_1,I_2,I_3,I_4} + \alpha^4_{I_1,I_2,I_4,I_3} & \text{if } I_1=I_2,\ I_3<I_4 \\ \alpha^4_{I_1,I_2,I_3,I_4} + \alpha^4_{I_2,I_1,I_3,I_4} & \text{if } I_1<I_2,\ I_3=I_4 \\ \alpha^4_{I_1,I_2,I_3,I_4} & \text{if } I_1=I_2,\ I_3=I_4 \end{cases},$$
$$\text{and}\quad \beta^4_{I_1,I_2,I_3,I_4} = (2-\delta_{(I_1,I_2),(I_3,I_4)})\tilde\beta^4_{I_1,I_2,I_3,I_4}. \tag{2.23}$$

By (2.20) and the chain rule, the gradient and Hessian[4] of $\mathcal{I}$ can then be computed as:

$$\nabla\mathcal{I}(\mathbf{c}) = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} P_{i,j}^T\cdot\nabla\mathcal{I}_{\rm loc}(P_{i,j}(\mathbf{c})),\quad \mathrm{Hess}\,\mathcal{I}(\mathbf{c}) = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} P_{i,j}^T\cdot\mathrm{Hess}\,\mathcal{I}_{\rm loc}(P_{i,j}(\mathbf{c}))\cdot P_{i,j}. \tag{2.24}$$

By (2.22),

$$\begin{aligned}\frac{\partial\mathcal{I}_{\rm loc}}{\partial\mathbf{x}_I} = \sum_{(I_1,I_2,I_3,I_4)\in\mathcal{I}_4}\beta^4_{I_1,I_2,I_3,I_4}\Big[&(\delta_{I,I_1}\mathbf{x}_{I_2}+\delta_{I,I_2}\mathbf{x}_{I_1})\langle\mathbf{x}_{I_3},\mathbf{x}_{I_4}\rangle + \\ &(\delta_{I,I_3}\mathbf{x}_{I_4}+\delta_{I,I_4}\mathbf{x}_{I_3})\langle\mathbf{x}_{I_1},\mathbf{x}_{I_2}\rangle\Big] + \sum_{(I_1,I_2)\in\mathcal{I}_2}\beta^2_{I_1,I_2}\Big[\delta_{I,I_1}\mathbf{x}_{I_2}+\delta_{I,I_2}\mathbf{x}_{I_1}\Big].\end{aligned} \tag{2.25}$$

Here and below, for any functional $\mathcal{F}$ with variables grouped as $(x_I)_I$, we write

$$\frac{\partial\mathcal{F}}{\partial\mathbf{x}_I} := \Big[\frac{\partial\mathcal{F}}{\partial(\mathbf{x}_I)_1},\frac{\partial\mathcal{F}}{\partial(\mathbf{x}_I)_2},\frac{\partial\mathcal{F}}{\partial(\mathbf{x}_I)_3}\Big]^T,\quad \frac{\partial}{\partial\mathbf{x}_J}\frac{\partial\mathcal{F}}{\partial\mathbf{x}_I} := \Big[\frac{\partial^2\mathcal{F}}{\partial(\mathbf{x}_J)_j\partial(\mathbf{x}_I)_i}\Big]_{1\le i,j,\le 3}.$$

We refer to the latter as the (I,J)-th block of the Hessian of $\mathcal{F}$. Note that $\frac{\partial}{\partial\mathbf{x}_J}\frac{\partial\mathcal{F}}{\partial\mathbf{x}_I} = [\frac{\partial}{\partial\mathbf{x}_I}\frac{\partial\mathcal{F}}{\partial\mathbf{x}_J}]^T$.

By (2.25), we have that

$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}_J}\frac{\partial\mathcal{I}_{\rm loc}}{\partial\mathbf{x}_I} = \sum_{(I_1,I_2,I_3,I_4)\in\mathcal{I}_4}\beta^4_{I_1,I_2,I_3,I_4}\Big[&\\ &(\delta_{I,I_1}\delta_{J,I_2}+\delta_{I,I_2}\delta_{J,I_1})\langle\mathbf{x}_{I_3},\mathbf{x}_{I_4}\rangle\mathbb{I}_{3\times3} + (\delta_{I,I_1}\mathbf{x}_{I_2}+\delta_{I,I_2}\mathbf{x}_{I_1})(\delta_{J,I_3}\mathbf{x}_{I_4}+\delta_{J,I_4}\mathbf{x}_{I_3})^T + \\ &(\delta_{I,I_3}\delta_{J,I_4}+\delta_{I,I_4}\delta_{J,I_3})\langle\mathbf{x}_{I_1},\mathbf{x}_{I_2}\rangle\mathbb{I}_{3\times3} + (\delta_{I,I_3}\mathbf{x}_{I_4}+\delta_{I,I_4}\mathbf{x}_{I_3})(\delta_{J,I_1}\mathbf{x}_{I_2}+\delta_{J,I_2}\mathbf{x}_{I_1})^T\Big] \\ + \sum_{(I_1,I_2)\in\mathcal{I}_2}\beta^2_{I_1,I_2}\Big[\delta_{I,I_1}\delta_{J,I_2}+\delta_{I,I_2}\delta_{J,I_1}\Big]&\mathbb{I}_{3\times3}.\end{aligned} \tag{2.26}$$

Remark 2.1 *We state here three useful properties of the isometry functional $\mathcal{I}$:*

(i) $\mathcal{I}(\mathbf{c})$ is a quartic polynomial in the control variables $\mathbf{c}$.

[4]Since we use a quasi-Newton method, the Hessian evaluation is not needed by the optimization solver. The Hessian computation, however, is used for checking if a stationary point is a local minimizer or a saddle point. See Section 5.5.

(ii) No numerical integration is needed for the computation of $\mathcal{I}(\mathbf{c})$ as all integrands involved are polynomials in the spatial variables x, y.

(iii) The $\beta^2_{\cdots}$ and $\beta^4_{\cdots}$ coefficients, as well as the index sets $\mathfrak{I}_2$ and $\mathfrak{I}_4$, are independent of the control data and thus can be pre-computed. (This is very useful for optimization, where $\mathcal{I}$ has to be evaluated many times.)

Note that the polynomial property in (i) is irrelevant to that in (ii); in particular, the fact that $\mathcal{I}(\mathbf{c})$ is quartic in $\mathbf{c}$ holds regardless of the polynomial degree d in the underlying spline space $\mathcal{S}_{\mathbf{h},d}$. In fact, property (i) holds even if we replace the spline space by any linear approximation space. See Section 2.5.

2.3 Functional E and $\mathcal{E}_3$

We write the energy functional E (1.1) here again as:

$$E[\mathbf{y}] = \frac{1}{2}\iint_\omega \|H - Z\|_F^2 \, du\, dv, \quad H = (d\mathbf{y}^T d\mathbf{y})^{-1}\left[\left\langle \partial_p\partial_q\mathbf{y}, \frac{\partial_1\mathbf{y}\times\partial_2\mathbf{y}}{\|\partial_1\mathbf{y}\times\partial_2\mathbf{y}\|}\right\rangle\right]_{p,q=1,2}.$$

When restricted to the spline space $\mathcal{S}$, it does not enjoy either property (i) or (ii) in **Remark 2.1** above, as we cannot dispense with the terms $(d\mathbf{y}^T d\mathbf{y})^{-1}$ and $\|\partial_1\mathbf{y}\times\partial_2\mathbf{y}\|$ when $\mathbf{y}$ is *not* an isometry. As noted earlier, a spline function is almost never an (exact) isometry. Therefore an honest computation of E would require many numerical integrations. In order to facilitate efficient computation, we shall replace E by another energy functional that agrees with E when applied to a perfect isometry. And then we will adjoint our minimization problem with a near-isometry constraint or a penalization term using the functional $\mathcal{I}$ defined in (2.13).

More precisely, note that for any isometry $\mathbf{y}$, we have $d\mathbf{y}^T d\mathbf{y} = \mathbb{I}$, $\|\partial_1\mathbf{y}\times\partial_2\mathbf{y}\| = 1$, and

$$\langle \partial_p\partial_q\mathbf{y}, \partial_1\mathbf{y}\times\partial_2\mathbf{y}\rangle^2 = \|\partial_p\partial_q\mathbf{y}\|_2^2. \tag{2.27}$$

Hence we will consider the following functional:

$$\mathcal{E}_3[\mathbf{y}] := \frac{1}{2}\sum_{p,q=1,2}\iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2 - 2\langle\partial_p\partial_q\mathbf{y}, \partial_1\mathbf{y}\times\partial_2\mathbf{y}\rangle Z_{p,q} + Z_{p,q}^2 \, du\, dv$$

so that

$$E[\mathbf{y}] = \mathcal{E}_3[\mathbf{y}] \text{ when } \mathbf{y} \text{ is an isometry.} \tag{2.28}$$

We now derive an efficient algorithm for computing $\mathcal{E}_3$, its gradient and Hessian.

$$\begin{aligned}
\mathcal{E}_3(\mathbf{c}) &= \frac{1}{2}\sum_{p,q=1,2}\iint_{\omega} \|\partial_p\partial_q\mathbf{y}\|_2^2 - 2\langle\partial_p\partial_q\mathbf{y}, \partial_1\mathbf{y}\times\partial_2\mathbf{y}\rangle Z_{p,q} + Z_{p,q}^2\,du\,dv \\
&= \frac{h_1h_2}{2}\sum_{p,q=1,2}\iint_{[0,m]\times[0,n]} h_p^{-2}h_q^{-2}\|\partial_p\partial_q\overline{\mathbf{y}}\|_2^2 - 2h_p^{-1}h_q^{-1}h_1^{-1}h_2^{-1}\langle\partial_p\partial_q\overline{\mathbf{y}}, \partial_1\overline{\mathbf{y}}\times\partial_2\overline{\mathbf{y}}\rangle Z_{p,q} + Z_{p,q}^2\,dx\,dy \\
&= \frac{h_1h_2}{2}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{p,q=1,2}\Bigg\{h_p^{-2}h_q^{-2}\iint_{s_{i,j}}\langle\partial_p\partial_q\overline{\mathbf{y}}, \partial_p\partial_q\overline{\mathbf{y}}\rangle dx\,dy - \\
&\qquad 2h_p^{-1}h_q^{-1}h_1^{-1}h_2^{-1}Z_{p,q}\iint_{s_{i,j}}\langle\partial_p\partial_q\overline{\mathbf{y}}, \partial_1\overline{\mathbf{y}}\times\partial_2\overline{\mathbf{y}}\rangle dx\,dy + Z_{p,q}^2\Bigg\}
\end{aligned} \tag{2.29}$$

Similar to (2.17), we have

$$\begin{aligned}
\iint_{s_{i,j}}\langle\partial_p\partial_q\overline{\mathbf{y}}, \partial_1\overline{\mathbf{y}}\times\partial_2\overline{\mathbf{y}}\rangle dx\,dy &= \sum_{\substack{k_1,k_2,k_3\in\mathcal{N}(d)\\ l_1,l_2,l_3\in\mathcal{N}(d)}}\langle\mathbf{c}_{J_1}, \mathbf{c}_{J_2}\times\mathbf{c}_{J_3}\rangle T^3_{k_1,k_2,k_3,\epsilon(p),\epsilon(q)}T^3_{l_1,l_3,l_2,1-\epsilon(p),1-\epsilon(q)}, \\
\iint_{s_{i,j}}\langle\partial_p\partial_q\overline{\mathbf{y}}, \partial_p\partial_q\overline{\mathbf{y}}\rangle dx\,dy &= \sum_{\substack{k_1,k_2\in\mathcal{N}(d)\\ l_1,l_2\in\mathcal{N}(d)}}\langle\mathbf{c}_{J_1}, \mathbf{c}_{J_2}\rangle T^2_{k_1,k_2,\epsilon(p)+\epsilon(q)}T^2_{l_1,l_2,2-\epsilon(p)-\epsilon(q)},
\end{aligned} \tag{2.30}$$

where $J_\ell = (i,j)+(k_\ell,l_\ell)$, $\ell = 1,\dots,3$,

$$T^3_{k_1,k_2,k_3,\epsilon_1,\epsilon_2} = \int_0^1 B^{(\epsilon_1+\epsilon_2)}(t-k_1)B'(t-k_2)B(t-k_3)\,dt, \tag{2.31}$$

and $T^2_{k_1,k_2,\eta}$, $\eta\in\{0,1,2\}$, is defined exactly as in (2.19) (only that the ϵ in (2.19) is now allowed to go up to 2.) We then have

$$\begin{aligned}
\mathcal{E}_3(\mathbf{c}) &= \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\mathcal{E}_{3,\mathrm{loc}}(P_{i,j}(\mathbf{c})), \\
\mathcal{E}_{3,\mathrm{loc}}(\mathbf{x}) &= \sum_{I_1,I_2\in\mathcal{N}(d)^2}\widetilde{\alpha}^2_{I_1,I_2}\langle\mathbf{x}_{I_1}, \mathbf{x}_{I_2}\rangle - \sum_{I_1,I_2,I_3\in\mathcal{N}(d)^2}\alpha^3_{I_1,I_2,I_3}\langle\mathbf{x}_{I_1}, \mathbf{x}_{I_2}\times\mathbf{x}_{I_3}\rangle + \frac{h_1h_2}{2}\|Z\|_F^2,
\end{aligned} \tag{2.32}$$

where

$$\begin{aligned}
\widetilde{\alpha}^2_{(k_1,l_1),(k_2,l_2)} &:= \frac{1}{2}\sum_{\epsilon_1,\epsilon_2=0,1}h_1^{1-2(\epsilon_1+\epsilon_2)}h_2^{1-2(2-\epsilon_1-\epsilon_2)}T^2_{k_1,k_2,\epsilon_1+\epsilon_2}T^2_{l_1,l_2,2-\epsilon_1-\epsilon_2} \\
\alpha^3_{(k_1,l_1),(k_2,l_2),(k_3,l_3)} &:= \sum_{\epsilon_1,\epsilon_2=0,1}h_1^{-(\epsilon_1+\epsilon_2)}h_2^{-(2-\epsilon_1-\epsilon_2)}Z_{2-\epsilon_1,2-\epsilon_2}T^3_{k_1,k_2,k_3,\epsilon_1,\epsilon_2}T^3_{l_1,l_3,l_2,1-\epsilon_1,1-\epsilon_2}.
\end{aligned} \tag{2.33}$$

The degree 3 term of $\mathcal{E}_{3,\mathrm{loc}}(\mathbf{x})$ can be compressed using the invariance of a scalar triple product under a circular shift of the arguments and that a scalar triple product vanishes when the arguments are not distinct: Let $\leq$ be an ordering on $\mathcal{N}(d) \times \mathcal{N}(d)$ and

$$\begin{aligned}
\mathcal{I}_2 &:= \{(I_1, I_2) : I_1, I_2 \in \mathcal{N}(d) \times \mathcal{N}(d),\ I_1 \leq I_2\}, \\
\mathcal{I}_3 &:= \{(I_1, I_2, I_3) : I_1, I_2, I_3 \in \mathcal{N}(d) \times \mathcal{N}(d),\ I_1 \leq I_2 \leq I_3\}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathcal{E}_{3,\mathrm{loc}}(\mathbf{x}) = &- \sum_{(I_1,I_2,I_3)\in\mathcal{I}_3} \beta^3_{I_1,I_2,I_3} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \times \mathbf{x}_{I_3} \rangle + \sum_{(I_1,I_2)\in\mathcal{I}_2} \gamma^2_{I_1,I_2} \langle \mathbf{x}_{I_1}, \mathbf{x}_{I_2} \rangle + \frac{h_1 h_2}{2} \|Z\|_F^2 \\
\text{where} \quad \beta^3_{I_1,I_2,I_3} &:= \sum_{\pi\in S_3} (-1)^{\mathrm{sgn}(\pi)} \alpha^3_{I_{\pi(1)},I_{\pi(2)},I_{\pi(3)}}, \quad \gamma^2_{I_1,I_2} := (2 - \delta_{I_1,I_2}) \widetilde{\alpha}^2_{I_1,I_2}.
\end{aligned} \tag{2.34}$$

Besides the reduction in the number of summands, the coefficients $\beta^3_{\cdots}$ and $\gamma^2_{\cdots}$, as well as the index sets $\mathcal{I}_3$ and $\mathcal{I}_2$, are independent of the control data and so can be pre-computed. (The same comment was made about the isometry constraint functional $\mathcal{I}$ earlier.)

Similar to (2.24), the gradient and Hessian of $\mathcal{E}_3$ can be expressed as:

$$\begin{aligned}
\nabla \mathcal{E}_3(\mathbf{c}) &= \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} P_{i,j}^T \cdot \nabla \mathcal{E}_{3,\mathrm{loc}}(P_{i,j}(\mathbf{c})) \\
\mathrm{Hess}\, \mathcal{E}_3(\mathbf{c}) &= \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} P_{i,j}^T \cdot \mathrm{Hess}\, \mathcal{E}_{3,\mathrm{loc}}(P_{i,j}(\mathbf{c})) \cdot P_{i,j}.
\end{aligned} \tag{2.35}$$

The above expressions show how to assemble the global gradient vector and Hessian matrix from the local ones. By (2.34), the local gradient can be derived as follows:

$$\begin{aligned}
\frac{\partial \mathcal{E}_{3,\mathrm{loc}}}{\partial \mathbf{x}_I} = &- \sum_{(I_1,I_2,I_3)\in\mathcal{I}_3} \beta^3_{I_1,I_2,I_3} \Big[\delta_{I,I_1} (\mathbf{x}_{I_2} \times \mathbf{x}_{I_3}) + \delta_{I,I_2} (\mathbf{x}_{I_3} \times \mathbf{x}_{I_1}) + \delta_{I,I_3} (\mathbf{x}_{I_1} \times \mathbf{x}_{I_2}) \Big] + \\
&\sum_{(I_1,I_2)\in\mathcal{I}_2} \gamma^2_{I_1,I_2} \Big[\delta_{I,I_1} \mathbf{x}_{I_2} + \delta_{I,I_2} \mathbf{x}_{I_1} \Big].
\end{aligned} \tag{2.36}$$

By (2.36), and the formula

$$\frac{\partial}{\partial \mathbf{x}_I} (\mathbf{x}_J \times \mathbf{x}_K) = \delta_{I,K} [\mathbf{x}_J]_\times - \delta_{I,J} [\mathbf{x}_K]_\times, \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_\times := \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix},$$

the (I,J)-th block of $\operatorname{Hess}\mathcal{E}_{3,\mathrm{loc}}$ is given by:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{x}_J}\frac{\partial \mathcal{E}_{3,\mathrm{loc}}}{\partial \mathbf{x}_I} = &-\sum_{(I_1,I_2,I_3)\in\mathcal{I}_3}\beta^3_{I_1,I_2,I_3}\Big[\delta_{I,I_1}\frac{\partial}{\partial \mathbf{x}_J}(\mathbf{x}_{I_2}\times\mathbf{x}_{I_3})+\delta_{I,I_2}\frac{\partial}{\partial \mathbf{x}_J}(\mathbf{x}_{I_3}\times\mathbf{x}_{I_1})+\\ &\qquad \delta_{I,I_3}\frac{\partial}{\partial \mathbf{x}_J}(\mathbf{x}_{I_1}\times\mathbf{x}_{I_2})\Big]+\sum_{(I_1,I_2)\in\mathcal{I}_2}\gamma^2_{I_1,I_2}\Big[\delta_{I,I_1}\frac{\partial}{\partial \mathbf{x}_J}\mathbf{x}_{I_2}+\delta_{I,I_2}\frac{\partial}{\partial \mathbf{x}_J}\mathbf{x}_{I_1}\Big]\\ = &\sum_{(I_1,I_2,I_3)\in\mathcal{I}_3}\beta^3_{I_1,I_2,I_3}\Big[(\delta_{I,I_2}\delta_{J,I_3}-\delta_{I,I_3}\delta_{J,I_2})[x_{I_1}]_\times+(\delta_{I,I_3}\delta_{J,I_1}-\delta_{I,I_1}\delta_{J,I_3})[x_{I_2}]_\times\\ &\qquad +(\delta_{I,I_1}\delta_{J,I_2}-\delta_{I,I_2}\delta_{J,I_1})[x_{I_3}]_\times\Big]+\sum_{(I_1,I_2)\in\mathcal{I}_2}\gamma^2_{I_1,I_2}\Big[\delta_{I,I_1}\delta_{J,I_2}+\delta_{J,I_1}\delta_{I,I_2}\Big]\mathbb{I}_{3\times 3}.\end{aligned}$$

Note how the skew-symmetry of $[x_I]_\times$ implies $\frac{\partial}{\partial \mathbf{x}_J}\frac{\partial \mathcal{E}_{3,\mathrm{loc}}}{\partial \mathbf{x}_I}=[\frac{\partial}{\partial \mathbf{x}_I}\frac{\partial \mathcal{E}_{3,\mathrm{loc}}}{\partial \mathbf{x}_J}]^T$, as it should.

With the above formulation, $\mathcal{E}_3$, being a degree 3 polynomial in the control variables, shares similar properties (i)–(iii) stated in **Remark 2.1** for the functional $\mathcal{I}$.

2.4 Sparse POPs

Our proposed numerical method for (2.1) when ω is a rectangle is to solve either

$$\begin{aligned}&\min_{\mathbf{c}}\ \mathcal{E}_3(\mathbf{c})\\ \text{s.t.}\quad &\mathcal{I}(\mathbf{c})\le\varepsilon\\ &\mathbf{c}\text{ satisfies (2.10) when } d=3 \text{ or (2.11) when } d=2.\end{aligned}\tag{2.37}$$

or

$$\begin{aligned}&\min_{\mathbf{c}}\ \mathcal{E}_3(\mathbf{c})+\mu\mathcal{I}(\mathbf{c})\\ \text{s.t.}\quad &\mathbf{c}\text{ satisfies (2.10) when } d=3 \text{ or (2.11) when } d=2.\end{aligned}\tag{2.38}$$

The developments in the last three subsections imply that (2.37) is a polynomial optimization problem: recall that $\mathcal{E}_3(\mathbf{c})$ and $\mathcal{I}(\mathbf{c})$ are polynomials in $\mathbf{c}$ of degree 3 and degree 4 respectively, the boundary condition (2.10) or (2.11) is a linear condition on $\mathbf{c}$.

Remark 2.2 *The linear boundary condition reduces the dimensionality of* $\mathbf{c}$ *by* $2(n+d)$. *In subsequent writing, we assume that such a dimensionality reduction has been applied and continue to write* $\mathbf{c}$ *as the reduced variable vector, and we conveniently dispense with the boundary condition. Note that* $E_3(\mathbf{c})$ *and* $\mathcal{I}(\mathbf{c})$ *are still polynomials of degree 3 and 4, respectively, in the reduced variables.*

What is also crucial for computation is that $\mathcal{E}_3(\mathbf{c})$ and $\mathcal{I}(\mathbf{c})$ are *sparse* polynomials: by (2.20) and (2.32), both $\mathcal{E}_3(\mathbf{c})$ and $\mathcal{I}(\mathbf{c})$ have only $O(mn)$ non-zero monomials. Recall that a general degree D polynomial in N variables has up to $O(N^D)$ monomials. The fact that $\mathcal{E}_3(\mathbf{c})$ and $\mathcal{I}(\mathbf{c})$ have only $O(mn)$ non-zero monomials is attributed to the local supports of the B-splines (2.4).

2.5 Spline spaces in general domains

The discussion in the previous section implies that the sparse POP formulations are a consequence of the use of a linear approximation space with a local basis. Splines with local bases for general polygonal domain is an extensively studied subject. Consequently, we may consider the following extensions of our sparse POP method, for successively more general domains.

- Still for the case of a rectangular domain, replace uniform knot tensor product splines by non-uniform knot ones.
- For a polygonal domain with only sides parallel to the coordinate axis (the type of domains considered in [6]), tessellate it with rectangles and use Sibson splines [12].
- For a general polygonal domain, tessellate it with an arbitrary triangulation and use Powell-Sabin splines [20].

For a review of Powell-Sabin splines and many other constructions of splines on triangulations, see [17]. For a review of Sibson splines, see [12]. It is interesting to note that while Sibson splines on rectangles and Powell-Sabin's splines on triangles are quite similar in spirit, the former is not applicable to arbitrary quadrangulations in the same way that Powell-Sabin's splines can be applied to arbitrary triangulations.

These methods result in C^1-finite element spaces with nodal basis functions supported at the 1-ring around the vertex associated with the basis function. Of course, there are higher order spline constructions on arbitrary triangulations, at the cost of larger local supports.

We also point out that the quadrilateral Kirchhoff elements employed in [6] are very similar to Sibson splines: they share exactly the same degrees of freedom and 1-ring local supports. The quadrilateral Kirchhoff elements are C^0 piecewise bi-degree 3 polynomials, whereas Sibson splines are C^1-piecewise total degree 2 polynomials.

All the aforementioned splines, except the Kirchhoff elements, are regular enough to be in the Sobolev space $W^{2,2}(\omega)$ which is the natural habitat for the variation problem (2.1). In this sense, our proposed numerical methods based on such spline approximations are *conforming*.

The small local supports make certain inner-products, similar to (2.16) and (2.31), relatively easy to compute. As in the case of rectangular domains, such inner-products can be pre-computed before the optimization procedure.

3 Preparation for the theory

For the theoretical development, we shall make the blanket assumption that ω is a polygonal domain with a portion of the boundary $\partial\omega$ being straight. In addition, there is a family of finite-dimensional linear spaces $S_h \subset W^{2,2}(\omega;\mathbb{R})$, $h > 0$. An element $\mathbf{y}_h \in [S_h]^3$ is intended to approximate a solution $\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3)$ of (2.1). More specific properties about the approximation spaces S_h will be given in Section 3.3. We do wish to point out that the current setting, in particular the geometry of the domain and the boundary conditions, can certainly be further generalized but our present results can already illustrate the key ideas of the approach we are taking.

3.1 Boundary condition in $W^{2,2}$

We first clarify the meaning of the boundary condition (2.3) for $\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3)$. More generally, consider

$$\mathbf{y}(0,v) = \mathbf{f}, \text{ and } \frac{\partial \mathbf{y}}{\partial u}(0,v) = \mathbf{g}, \tag{3.1}$$

where $\mathbf{f}$ and $\mathbf{g}$ are 'regular enough' – see below – $\mathbb{R}^3$-valued functions defined on

$$E := \{(0,v) : 0 < v < B\} \subset \partial\omega. \tag{3.2}$$

Since

$$W^{2,2}(\omega) \hookrightarrow C^{0,\lambda}(\overline{\omega}), \quad \forall\, 0 < \lambda < 1,$$

by Sobolev's embedding theorem, the first boundary condition in (3.1) can be interpreted in the classical pointwise sense, provided that $\mathbf{f}$ is sufficiently regular. For our purpose, it suffices that $\mathbf{f} \in C^1(\bar{E})$.

By the trace theorem [11, 1], the trace operator $T : C^\infty(\bar{\omega}) \to L^p(E)$, $T(f) := f|_E$, has a unique continuous extension to $T : W^{1,p}(\omega) \to L^p(E)$. Therefore, for $\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3)$, the left-hand side of the second boundary condition in (3.1) should be interpreted as the trace of $\frac{\partial \mathbf{y}}{\partial u}$, and $\mathbf{g}$ should be in the image of the trace operator $T : W^{1,2}(\omega;\mathbb{R}^3) \to L^2(E;\mathbb{R}^3)$. The characterization of this image is quite technical, see [1, Theorem 7.39]. For our purpose here, it suffices to have $\mathbf{g} \in C(\bar{E})$.

The right-hand sides of the clamped boundary condition (2.3), whose components are constant and linear functions, are not only regular enough, but also can be satisfied *exactly* by any reasonable spline function. This property will be exploited to simplify the analysis; see next subsection. For a general pair of $\mathbf{f}$ and $\mathbf{g}$, an approximant $\mathbf{y}_h \in [S_h]^3$ is only expected to satisfy (3.1) approximately for small h.

3.2 Continuity of functionals in $W^{2,2}$

Next we state and prove some basic and useful properties of the functionals. Let

$$\mathrm{Iso}^{2,2} := \mathrm{Iso}^{2,2}(\omega;\mathbb{R}^3) := \{\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3) : d\mathbf{y}^T d\mathbf{y} = \mathbb{I}_{2\times 2} \text{ a.e.}\}.$$

Note that for $\mathbf{y} \in W^{2,2}$, $\mathbf{y} \in \mathrm{Iso}^{2,2}$ if and only if $\mathcal{I}(\mathbf{y}) = 0$.

We recall here the functional $\mathcal{E}_3$:

$$\mathcal{E}_3(\mathbf{y}) \;=\; \frac{1}{2}\sum_{p,q=1,2}\iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2 - \sum_{p,q=1,2}\iint_\omega \langle \partial_p\partial_q\mathbf{y}, \partial_1\mathbf{y}\times\partial_2\mathbf{y}\rangle Z_{p,q} + \frac{1}{2}\sum_{p,q=1,2} Z_{p,q}^2\,\mathrm{area}(\omega)$$

We introduce

$$\mathcal{E}_{3,\mathrm{I}}[\mathbf{y}] = \frac{1}{2}\sum_{p,q=1,2}\iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2 \quad \text{and} \quad \mathcal{E}_{3,\mathrm{II}}[\mathbf{y}] = \sum_{p,q=1,2}\iint_\omega \langle \partial_p\partial_q\mathbf{y}, \partial_1\mathbf{y}\times\partial_2\mathbf{y}\rangle Z_{p,q}. \tag{3.3}$$

It is clear that $\mathcal{E}_{3,\mathrm{I}}$ is well-defined and continuous on $W^{2,2}$. In fact, $\mathcal{E}_{3,\mathrm{I}} : W^{2,2} \to \mathbb{R}$ is also weakly lower-semi-continuous:

Proposition 3.1 (Weak LSC of $\mathcal{E}_{3,\mathrm{I}}$.) *$\mathcal{E}_{3,\mathrm{I}} : W^{2,2} \to \mathbb{R}$ is weakly lower semi-continuous, i.e. if $\mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}(\omega;\mathbb{R}^3)$, then*

$$\liminf_{n\to\infty} \mathcal{E}_{3,\mathrm{I}}[\mathbf{y}_n] \geq \mathcal{E}_{3,\mathrm{I}}[\mathbf{y}]. \tag{3.4}$$

Proof: Let $\mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}$. The proof follows from a classical trick exploiting convexity: Since $\xi^2 \geq \xi_0^2 + 2\xi_0(\xi - \xi_0)$, we have $\|\partial_p\partial_q\mathbf{y}_n\|_2^2 \geq \|\partial_p\partial_q\mathbf{y}\|_2^2 + 2\,\langle \partial_p\partial_q\mathbf{y}, \partial_p\partial_q\mathbf{y}_n - \partial_p\partial_q\mathbf{y}\rangle$. Hence

$$\liminf_{n\to\infty} \iint_\omega \|\partial_p\partial_q\mathbf{y}_n\|_2^2 \geq \iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2 + 2 \underbrace{\liminf_{n\to\infty} \iint_\omega \langle \partial_p\partial_q\mathbf{y}, \partial_p\partial_q\mathbf{y}_n - \partial_p\partial_q\mathbf{y}\rangle_{\mathbb{R}^3}}_{=\lim_n \langle \partial_p\partial_q\mathbf{y}, \partial_p\partial_q\mathbf{y}_n - \partial_p\partial_q\mathbf{y}\rangle_{L^2(\omega;\mathbb{R}^3)} = 0, \text{ by weak convergence}} = \iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2,$$

as claimed. ∎

It can also be shown that $\mathcal{E}_{3,\mathrm{I}}$ is *not* weakly continuous. On the other hand, we have:

Proposition 3.2 (Weak Continuity of $\mathcal{E}_{3,\mathrm{II}}$.) *$\mathcal{E}_{3,\mathrm{II}}$ is well-defined on $W^{2,2}$ and $\mathcal{E}_{3,\mathrm{II}} : W^{2,2} \to \mathbb{R}$ is weakly continuous, i.e. if $\mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}(\omega;\mathbb{R}^3)$, then*

$$\lim_{n\to\infty} \mathcal{E}_{3,\mathrm{II}}[\mathbf{y}_n] = \mathcal{E}_{3,\mathrm{II}}[\mathbf{y}]. \tag{3.5}$$

The proof of Proposition 3.2 uses the following standard results:

(I) Applying the Rellich-Kondrachov Compactness Theorem [1] to our 2-dimensional domain ω, we have

$$W^{2,2}(\omega) \subset\subset W^{1,q}(\omega), \quad \forall\, 1 \leq q < \infty. \tag{3.6}$$

(II) If $f_n \rightharpoonup f$ in L^2 and $g_n \to g$ in L^2, then $\langle f_n, g_n\rangle_{L^2} \to \langle f, g\rangle_{L^2}$.

(III) If $f_n \rightharpoonup f$ in $W^{2,2}(\omega)$, then $f_n \to f$ in $W^{1,q}(\omega)$ for any $q \in [1,\infty)$.

Proof of Proposition 3.2. Let $\mathbf{y} \in W^{2,2}$. By (3.6) with $q = 4$, $\partial_1\mathbf{y}$, $\partial_2\mathbf{y} \in L^4$, so $\partial_1\mathbf{y} \times \partial_2\mathbf{y} \in L^2$. As $\partial_p\partial_q\mathbf{y} \in L^2$, $\mathcal{E}_{3,\mathrm{II}}[\mathbf{y}]$ is finite by the Cauchy-Schwarz inequality. Let $\mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}$. By (III) above, $(\mathbf{y}_n)$ converges strongly to $\mathbf{y}$ in $W^{1,4}$. This in turn implies that $\partial_1\mathbf{y}_n \times \partial_2\mathbf{y}_n \to \partial_1\mathbf{y} \times \partial_2\mathbf{y}$ in L^2. By assumption, we have $\partial_p\partial_q\mathbf{y}_n \rightharpoonup \partial_p\partial_q\mathbf{y}$ in L^2, so (3.5) follows from (II) above. ∎

Since $\mathcal{E}_3$ is the sum of a weakly LSC functional and a weakly continuous one, it is weakly LSC. At the same time, it is strongly continuous. To summarize:

Proposition 3.3 (Continuity of $\mathcal{E}_3$) *$\mathcal{E}_3 : W^{2,2} \to \mathbb{R}$ is continuous and weakly lower semi-continuous.*

Proposition 3.4 *Let $\mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}$. If $\lim_{n\to\infty} \mathcal{E}_3[\mathbf{y}_n] = \mathcal{E}_3[\mathbf{y}]$, then $\lim_{n\to\infty} \|\mathbf{y}_n\|_{W^{2,2}} = \|\mathbf{y}\|_{W^{2,2}}$ and also $\mathbf{y}_n \to \mathbf{y}$ in $W^{2,2}$.*

Proof: Recall that $\mathcal{E}_3 = \mathcal{E}_{3,\mathrm{I}} + \mathcal{E}_{3,\mathrm{II}} + \text{constant}$. Moreover, by Proposition 3.2, $\lim_{n\to\infty} \mathcal{E}_{3,\mathrm{II}}[\mathbf{y}_n] = \mathcal{E}_{3,\mathrm{II}}[\mathbf{y}]$. So by the assumption of the proposition we also have $\lim_{n\to\infty} \mathcal{E}_{3,\mathrm{I}}[\mathbf{y}_n] = \mathcal{E}_{3,\mathrm{I}}[\mathbf{y}]$. Note that $\mathcal{E}_{3,\mathrm{I}}$ is the semi-norm $|\mathbf{y}|_{W^{2,2}}$. Next, recall that weak convergence in $W^{2,2}$ implies strong convergence in $W^{1,2}$, so $\|\mathbf{y}_n\|_{W^{1,2}} \to \|\mathbf{y}\|_{W^{1,2}}$. Altogether, we have $\|\mathbf{y}_n\|_{W^{2,2}} = \|\mathbf{y}_n\|_{W^{1,2}} + |\mathbf{y}|_{W^{2,2}} \to \|\mathbf{y}\|_{W^{2,2}}$. The proof is then completed by the general fact that 'weak convergence + norm convergence ⇒ strong convergence'. ∎

We shall also need the following fact which follows easily from the form of $\mathcal{I}$:

Proposition 3.5 (Continuity of $\mathcal{I}$) *$\mathcal{I} : W^{1,4} \to \mathbb{R}$ is well-defined and continuous.*

3.3 Approximation properties of $S_h(\omega)$

Here we state the specific properties of the discrete space S_h needed in our analysis. We assume that the spline space S_h is at least C^1 and reproduces quadratics, i.e. $\Pi_2 \subset S_h \subset C^1(\omega)$ where Π_2 is the set of polynomials with total degree no more than 2. In particular, the flat plate

$$\mathbf{y}^\flat : \omega \to \mathbb{R}^3, \quad \mathbf{y}^\flat(u,v) = [u,v,0]^T, \tag{3.7}$$

belongs to $[S_h]^3$ for any $h > 0$. Moreover, we assume there is an operator $Q_h : W^{2,2} \to S_h$ such that

[A1] For every $y \in W^{2,2}$,

$$\|\nabla(y - Q_h y)\|_{L^2(\omega)} \leq Ch|y|_{W^{2,2}(\omega)}. \tag{3.8}$$

[A2] For every $y \in C^1(\omega)$, $\lim_{h\to 0} \|\nabla(y - Q_h y)\|_{L^\infty(\omega)} = 0$. Moreover, there exists a $d \geq 1$ such that for every $y \in C^{d+1}(\omega)$

$$\|\nabla(y - Q_h y)\|_{L^\infty(\omega)} \leq Ch^d|y|_{C^{d+1}(\omega)}. \tag{3.9}$$

[A3] If $\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3)$ satisfies the clamped boundary condition (2.3), then so does $Q_h\mathbf{y} \in [S_h]^3$.

These properties are established in the spline literature [25, 19, 17]. The largest integer $d+1$ for which [A2] holds is the approximation order of the spline space. For the tensor product spline space $\mathcal{S}_{\mathbf{h},d}$ considered in Section 5, $h = |\mathbf{h}| = \max\{h_1, h_2\}$ and the approximation order is $d+1$. For 6-split Powell-Sabin splines, [A2] holds for $d = 2$. In the spline literature, different approaches for choosing the linear operator Q_h are proposed, and they all satisfy the fundamental quasi-interpolation property $Q_h\pi = \pi$ for all $\pi \in \Pi_d$. Moreover, for most spline spaces (including tensor product and Powell-Sabin splines), Q_h can be chosen to satisfy [A3]; this is directly related to the fact that the functions involved in the clamped boundary condition (2.3) are either constant or linear.

Now we present a result, pertaining to our isometry function $\mathcal{I}$, that illustrates how these approximation properties enter our analysis in the next section.

Theorem 3.6 *Assume S_h satisfies the said approximation properties [A1, A2, A3] above. Then we have the following statements.*

1. *There is a constant $C > 0$, independent of h, such that for any $\mathbf{y} \in \mathrm{Iso}^{2,2}(\omega)$,*

$$\mathcal{I}[Q_h\mathbf{y}] \leq Ch^2|\mathbf{y}|^2_{W^{2,2}(\omega)}. \tag{3.10}$$

2. *There is a constant $C > 0$, independent of h, such that for any $\mathbf{y} \in C^{d+1} \cap \mathrm{Iso}^{2,2}(\omega)$,*

$$\mathcal{I}[Q_h\mathbf{y}] \leq Ch^{2d}|\mathbf{y}|^2_{C^{d+1}(\omega)}. \tag{3.11}$$

Proof: For any $\mathbf{y} \in \mathrm{Iso}^{2,2}(\omega)$, write $\mathbf{y}_h := Q_h\mathbf{y}$. Since $d\mathbf{y}^T d\mathbf{y} = \mathbb{I}_{2\times 2}$, we have

$$\begin{aligned}
\mathcal{I}[\mathbf{y}_h] &= \iint_\omega \|d\mathbf{y}_h^T d\mathbf{y}_h - \mathbb{I}_{2\times 2}\|_F^2 = \iint_\omega \|d\mathbf{y}_h^T d\mathbf{y}_h - d\mathbf{y}^T d\mathbf{y}\|_F^2 \\
&= \iint_\omega \big\|(d\mathbf{y}_h^T - d\mathbf{y}^T)d\mathbf{y}_h + d\mathbf{y}^T(d\mathbf{y}_h - d\mathbf{y})\big\|_F^2 \\
&\le 2\iint_\omega \big\|(d\mathbf{y}_h^T - d\mathbf{y}^T)d\mathbf{y}_h\big\|_F^2 + \big\|d\mathbf{y}^T(d\mathbf{y}_h - d\mathbf{y})\big\|_F^2 \\
&\le 2\iint_\omega \left(\|d\mathbf{y}_h\|_2^2 + \|d\mathbf{y}\|_2^2\right)\cdot \big\|d\mathbf{y}_h - d\mathbf{y}\big\|_F^2 \qquad (3.12)
\end{aligned}$$

(where in the last step we have used the fact that $\|AB\|_F \le \|A\|_2\|B\|_F$ with $\|\cdot\|_F^2$ being the square of the Frobenius norm, sum of squares of all the matrix elements and $\|\cdot\|_2$ being the operator or spectral norm in L^2.)

For (3.10), since $\mathbf{y} \in \mathrm{Iso}^{2,2}$, i.e. $d\mathbf{y}(\cdot)^T d\mathbf{y}(\cdot) = \mathbb{I}_{2\times 2}$, we have that $\|d\mathbf{y}(\cdot)\|_2 = 1$. Hence by (3.12),

$$\begin{aligned}
\mathcal{I}[\mathbf{y}_h] &\le 2\left(\max_\omega \|d\mathbf{y}_h(\cdot)\|_2^2 + \max_\omega \|d\mathbf{y}(\cdot)\|_2^2\right)\iint_\omega \big\|d\mathbf{y}_h - d\mathbf{y}\big\|_F^2 \\
&\le C'\big(\big\|d\mathbf{y}_h\big\|_{L^\infty}^2 + \big\|d\mathbf{y}\big\|_{L^\infty}^2\big)\big\|d\mathbf{y}_h - d\mathbf{y}\big\|_{L^2}^2.
\end{aligned}$$

By [A1] and the first part of [A2],

$$\begin{aligned}
\mathcal{I}[\mathbf{y}_h] &\le C'\Big[\big(\big\|d\mathbf{y}\big\|_{L^\infty} + \big\|d\mathbf{y}_h - d\mathbf{y}\big\|_{L^\infty}\big)^2 + \big\|d\mathbf{y}\big\|_{L^\infty}^2\Big]\big\|d\mathbf{y}_h - d\mathbf{y}\big\|_{L^2}^2 \\
&\le C'[(1+o(1))^2 + 1^2]\,[Ch|\mathbf{y}|_{W^{2,2}(\omega)}]^2 \\
&\le C''h^2|\mathbf{y}|_{W^{2,2}(\omega)}^2.
\end{aligned}$$

For (3.11), similar to above, but now utilizing (3.9),

$$\begin{aligned}
\mathcal{I}[\mathbf{y}_h] &\le 2\iint_\omega \left(\|d\mathbf{y}_h\|_2^2 + \|d\mathbf{y}\|_2^2\right)\cdot \big\|d\mathbf{y}_h - d\mathbf{y}\big\|_F^2 \\
&\le C'\left(\|d\mathbf{y}_h\|_{L^2}^2 + \|d\mathbf{y}\|_{L^2}^2\right)\big\|d\mathbf{y}_h - d\mathbf{y}\big\|_{L^\infty}^2 \\
&\le C'\cdot O(1)\cdot Ch^{2d}|\mathbf{y}|_{C^{d+1}(\omega)}^2 \\
&\le C''h^{2d}|\mathbf{y}|_{C^{d+1}(\omega)}^2.
\end{aligned}$$

(In the second to last step, as in the proof of (3.10), we have used the fact that $\mathbf{y}$ is an isometry so that $d\mathbf{y}_h$ is close to $d\mathbf{y}$ in L^2.) ∎

3.4 Γ-convergence

We recall the fundamental concept of Γ-convergence, which our main results are based on. Let $\mathcal{X}$ be a first countable topological space. A sequence of functionals $\mathcal{F}_n : \mathcal{X} \to \bar{\mathbb{R}}$ ($:= \mathbb{R}\cup\{\infty\}$) is said to Γ-converge to $\mathcal{F} : \mathcal{X} \to \bar{\mathbb{R}}$ as $n \to \infty$ if the following two statements hold:

1. **(Lower bound)** For any sequence $x_n \in \mathcal{X}$ that converges to $x \in \mathcal{X}$, we have

$$\liminf_{n\to\infty} \mathcal{F}_n(x_n) \ge \mathcal{F}(x).$$

2. **(Upper bound/Recovery sequence)** For any $x \in \mathcal{X}$, there are $x_n \in \mathcal{X}$ such that $\lim_{n\to\infty} x_n = x$ and
$$\limsup_{n\to\infty} \mathcal{F}_n(x_n) \le \mathcal{F}(x).$$
(Combined with the lower bound, the latter inequality is equivalent to $\lim_{n\to\infty} \mathcal{F}_n(x_n) = \mathcal{F}(x)$.)

The fundamental fact about Γ-convergence is the following:

Theorem 3.7 (Minimizers converge to minimizers) *If $\mathcal{F}_n$ Γ-converges to $\mathcal{F}$, and x_n is a minimizer for $\mathcal{F}_n$, then every cluster point x^* of the sequence x_n is a minimizer of $\mathcal{F}$; moreover, $\lim_{n\to\infty} \mathcal{F}_n(x_n)$ exists and equals $\mathcal{F}(x^*)$.*

Note that this theorem says nothing about the possession of minimizers by $\mathcal{F}_n$, nor does it say anything about the existence of cluster point in the (presumably existent) sequence x_n.[5] Some coercivity and compactness properties of the functionals and the space $\mathcal{X}$ are needed for establishing these existences.

The topological space $\mathcal{X}$ relevant to us is the Sobolev space $W^{2,2}(\omega)$ equipped with its weak topology, which is first countable but not metrizable. Also, in the next section, when we have a family of functionals $\mathcal{F}_h$, parameterized by a continuous $h \in (0, h_0]$, by '$\mathcal{F}_h$ Γ-converges to $\mathcal{F}$ as $h \to 0$' we mean $\mathcal{F}_{h_n}$ Γ-converges to $\mathcal{F}$ (in the sense above) for every sequence h_n satisfying $\lim_{n\to\infty} h_n = 0$.

4 Existence of Minimizer and Γ-Convergence

For notational convenience, we write
$$\mathring{\mathcal{S}}_h := \{\mathbf{y} \in [S_h]^3 : \mathbf{y} \text{ satisfies (2.3)}\}$$
$$\mathring{W}^{2,2} := \{\mathbf{y} \in W^{2,2}(\omega;\mathbb{R}^3) : \mathbf{y} \text{ satisfies (2.3)}\}.$$

Our bilayer plate variational problem can be written as
$$\min_{\mathbf{y}\in M} E[\mathbf{y}], \quad M := \mathring{W}^{2,2} \cap \mathrm{Iso}^{2,2}. \tag{4.1}$$

We consider its finite-dimensional counterparts with constaint and penalization:
$$\min_{\mathbf{y}\in M_{h,\varepsilon}} \mathcal{E}_3[\mathbf{y}], \quad M_{h,\varepsilon} := \left\{\mathbf{y} \in \mathring{\mathcal{S}}_h : \mathcal{I}[\mathbf{y}] \le \varepsilon\right\} \tag{4.2}$$

and
$$\min_{\mathbf{y}\in \mathring{\mathcal{S}}_h} \mathcal{E}_3[\mathbf{y}] + \mu\,\mathcal{I}[\mathbf{y}]. \tag{4.3}$$

In this section, we establish the following results.

[5] However, if we assume that a minimizer x_n of $\mathcal{F}_n$ exists for every n, and that the sequence x_n has a cluster point x^*, then Theorem 3.7 does guarantee that $\mathcal{F}$ has a minimizer, namely x^*.

Theorem 4.1 *The minimization problem (4.1) has a minimizer. For any $\varepsilon > 0$, (4.2) has a minimizer. For any large enough $\mu > 0$, (4.3) has a minimizer.*

We introduce the following "new functionals" purely for the convenience of formulating the Γ-convergence result:

$$\tilde{E}[\mathbf{y}] := \begin{cases} E[\mathbf{y}] & \text{if } \mathbf{y} \in M, \\ \infty & \text{otherwise} \end{cases}, \tag{4.4}$$

$$\tilde{\mathcal{E}}_{h,\varepsilon}[\mathbf{y}] := \begin{cases} \mathcal{E}_3[\mathbf{y}] & \text{if } \mathbf{y} \in M_{h,\varepsilon} \\ \infty & \text{otherwise} \end{cases}, \tag{4.5}$$

$$\tilde{E}_d[\mathbf{y}] := \begin{cases} E[\mathbf{y}] & \text{if } \mathbf{y} \in M \cap \{\mathbf{y} \in C^{d+1}\} \\ \infty & \text{otherwise} \end{cases}, \tag{4.6}$$

$$\tilde{\mathcal{E}}^{h,\mu}[\mathbf{y}] := \begin{cases} \mathcal{E}_3[\mathbf{y}] + \mu\mathcal{I}[\mathbf{y}] & \text{if } \mathbf{y} \in M_{h,\varepsilon} \\ \infty & \text{otherwise} \end{cases}. \tag{4.7}$$

In the above, d is the integer in condition [A2] and the second part of Theorem 3.6 (i.e. the approximation order of $\mathcal{S}_h$ is $d+1$.)

Theorem 4.2 *1. Let ε_h be such that*

$$\lim_{h\to 0} \varepsilon_h = 0 \ \text{ and } \ \lim_{h\to 0} \frac{h^2}{\varepsilon_h} = 0, \tag{4.8}$$

then $\tilde{\mathcal{E}}_{h,\varepsilon_h}$ Γ-converges to $\tilde{E}$ as $h \to 0$.

2. Let ε_h be such that

$$\lim_{h\to 0} \varepsilon_h = 0 \ \text{ and } \ \lim_{h\to 0} \frac{h^{2d}}{\varepsilon_h} = 0, \tag{4.9}$$

then $\tilde{\mathcal{E}}_{h,\varepsilon_h}$ Γ-converges to $\tilde{E}_d$ as $h \to 0$.

Theorem 4.3 *1. If μ_h is such that*

$$\lim_{h\to 0} \mu_h = \infty, \ \text{ and } \ \lim_{h\to 0} \mu_h h^2 = 0, \tag{4.10}$$

then $\tilde{\mathcal{E}}_3^{h,\mu_h}$ Γ-converges to $\tilde{E}$.

2. If μ_h is such that

$$\lim_{h\to 0} \mu_h = \infty, \ \text{ and } \ \lim_{h\to 0} \mu_h h^{2d} = 0, \tag{4.11}$$

then $\tilde{\mathcal{E}}_3^{h,\mu_h}$ Γ-converges to $\tilde{E}_d$.

The above results will be proved in the next few sections.

Our final theoretical result which shows that discrete minimizers converge to continous ones is mostly a consequence of Theorem 4.2 and 4.3. The main missing ingredient to be filled is a compactness property for any solution $\mathbf{y}_h^*$ of (4.2) and (4.3), respectively.

Theorem 4.4 *Let*

1. $\mathbf{y}_h^* = \mathbf{y}_{h,\varepsilon_h}^*$ *be a solution of* (4.2) *with* ε_h *chosen according to* (4.8) *or*
2. $\mathbf{y}_h^* = \mathbf{y}_{h,\mu_h}^*$ *be a solution of* (4.3) *with* μ_h *chosen according to* (4.10),

then

(i) $\{\mathbf{y}_h^*\}$ *has a cluster point* $\mathbf{y}^*$ *in the weak* $W^{2,2}$ *topology.*

(ii) Any cluster point $\mathbf{y}^*$ *of* $\mathbf{y}_h^*$ *is a solution of* (4.1)*, and* $\lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h^*] = \mathcal{E}_3[\mathbf{y}^*] = E[\mathbf{y}^*] = \min_{\mathbf{y}\in M} E[\mathbf{y}]$.

(iii) The weak $W^{2,2}$ *convergence of (a subsequence of)* $\mathbf{y}_h^*$ *to* $\mathbf{y}^*$ *can be improved to strong convergence.*

If we assume additionally that any solution of (4.1) *is* C^{d+1} *smooth with its* C^{d+1}*-norm bounded by some constant* $K' > 0$*, then the same conclusions (i)-(iii) hold if we set either*

3. $\mathbf{y}_h^* = \mathbf{y}_{h,\varepsilon_h}^*$ *to be a solution of* (4.2) *with* ε_h *chosen according to* (4.9)*, or*
4. $\mathbf{y}_h^* = \mathbf{y}_{h,\mu_h}^*$ *to be a solution of* (4.3) *with* μ_h *chosen according to* (4.11).

Proof: We first collect the coercivity results for the functionals E, $\mathcal{E}_3$ and $\mathcal{E}_3 + \mu\mathcal{I}$:

$$\begin{aligned}
\|\mathbf{y}\|_{W^{2,2}} &\leq C_1 E(\mathbf{y}) + C_2, \quad \forall\, \mathbf{y} \in M;\\
\|\mathbf{y}_h\|_{W^{2,2}} &\leq C_1 \mathcal{E}_3(\mathbf{y}_h) + C_2, \quad \forall\, \mathbf{y}_h \in M_{h,\varepsilon} \text{ and } h > 0,\\
\|\mathbf{y}_h\|_{W^{2,2}} &\leq C_1(\mathcal{E}_3[\mathbf{y}_h] + \mu\mathcal{I}[\mathbf{y}_h]) + C_2, \quad \forall\, \mathbf{y}_h \in \mathring{\mathcal{S}}_h \text{ and } \mu \text{ large enough.}
\end{aligned}$$

The first and third come from (4.13) and (4.19) while the second is an easy adaptation of (4.13) due to the fact that $\|\partial_1\mathbf{y} \times \partial_2\mathbf{y}\|^2_{L^2(\omega)} \leq A\mathcal{I}(\mathbf{y}) + B$.

Using the flat plate $\mathbf{y}^\flat$ (3.7) as a candidate, the following uniform apriori bound also holds for some constant $K^\flat$:

$$\|\mathbf{y}^*\|_{W^{2,2}},\ \|\mathbf{y}_{h,\varepsilon_h}^*\|_{W^{2,2}},\ \|\mathbf{y}_{h,\mu_h}^*\|_{W^{2,2}} \leq K^\flat. \tag{4.12}$$

Then (i) follows from weak compactness of bounded balls in $W^{2,2}$.

Statement (ii) is a consequence of Γ-convergence, Theorem 3.7.

Statement (iii) is then a direct consequence of Proposition 3.4 and the second statement in (ii).

■

4.1 Proof of Theorem 4.1

4.1.1 Existence of minimizer for (4.1)

We use the direct method in calculus of variations; see, for example, [29, Theorem 1.2]. This part of the theorem follows after we show:

(i) **Coercivity of** E**:** $E(\mathbf{y}) \to \infty$ as $\|\mathbf{y}\|_{W^{2,2}} \to \infty$, $\mathbf{y} \in M$. Or more precisely, there exist C_1, C_2 such that

$$\|\mathbf{y}\|_{W^{2,2}} \leq C_1 E(\mathbf{y}) + C_2, \quad \forall\, \mathbf{y} \in M. \tag{4.13}$$

(ii) **Weak closedness of M:** If $M \ni \mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}$, then $\mathbf{y} \in M$.

(iii) **Weak LSC of E:** If $M \ni \mathbf{y}_n \rightharpoonup \mathbf{y}$ in $W^{2,2}$, then $\liminf_{n\to\infty} E(\mathbf{y}_n) \geq E(\mathbf{y})$.

Proof of (i). When $\mathbf{y} \in W^{2,2}$ satisfies the boundary condition (2.3), then, together with the Poincaré inequality, we have boundedness for $\|\mathbf{y}\|_{W^{1,2}} = \iint_\omega \|\mathbf{y}\|_2^2 + \sum_{p=1,2} \| \iint_\omega \partial_p \mathbf{y}\|_2^2$. Therefore

$$\|\mathbf{y}\|_{W^{2,2}} = \|\mathbf{y}\|_{W^{1,2}} + \sum_{p,q=1,2} \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2 \leq C'' + \sum_{p,q=1,2} \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2, \tag{4.14}$$

When $\mathbf{y} \in \mathrm{Iso}^{2,2}$,

$$\begin{aligned} E(\mathbf{y}) = \mathcal{E}_3(\mathbf{y}) &= \frac{1}{2} \sum_{p,q=1,2} \left\{ \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2 - 2 \iint_\omega \langle \partial_p\partial_q \mathbf{y}, \partial_1 \mathbf{y} \times \partial_2 \mathbf{y}\rangle Z_{p,q} + Z_{p,q}^2 \,\mathrm{area}(\omega) \right\} \\ &\geq c \sum_{p,q=1,2} \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2 + C, \quad \text{for some } c > 0 \text{ and constant } C. \end{aligned} \tag{4.15}$$

In the above, we have used the fact that $2|ab| \leq \sigma a^2 + b^2/\sigma$ for any $\sigma > 0$ and (consequently)

$$\begin{aligned} 2 \sum_{p,q=1,2} \left| \iint_\omega \langle \partial_p\partial_q \mathbf{y}, \partial_1 \mathbf{y} \times \partial_2 \mathbf{y}\rangle Z_{p,q} \right| &\leq \sum_{p,q=1,2} \left\{ \sigma Z_{p,q} \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2 + \frac{Z_{p,q}}{\sigma} \iint_\omega \underbrace{\|\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}\|_2^2}_{=1 \text{ a.e.}} \right\} \\ &= \epsilon \sum_{p,q=1,2} \iint_\omega \|\partial_p\partial_q \mathbf{y}\|_2^2 + C' \end{aligned} \tag{4.16}$$

for some small $\epsilon > 0$ and constant C'.

The desired coercivity estimate (4.13) follows by combining (4.14) and (4.15).

Proof of (ii). First, we consider a $\mathbf{q} \in M$ – this is possible as $M \neq \emptyset$. Then any $\mathbf{y} \in M$ can be decomposed as $\mathbf{y} = \tilde{\mathbf{y}} + \mathbf{q}$ where $\tilde{\mathbf{y}}$ satisfies:

$$\tilde{\mathbf{y}}(0, v) = 0, \quad \frac{\partial \tilde{\mathbf{y}}}{\partial u}(0, v) = 0, \tag{4.17}$$

$$d\tilde{\mathbf{y}}^T d\tilde{\mathbf{y}} + d\mathbf{q}^T d\tilde{\mathbf{y}} + d\tilde{\mathbf{y}}^T d\mathbf{q} = 0. \tag{4.18}$$

We define $\tilde{M}_0 = \{\tilde{\mathbf{y}} \in W^{2,2} : \tilde{\mathbf{y}} \text{ satisfies (4.17)}\}$ and $\tilde{M}_1 = \{\tilde{\mathbf{y}} \in W^{2,2} : \tilde{\mathbf{y}} \text{ satisfies (4.18)}\}$. Then $M = \tilde{M}_0 \cap \tilde{M}_1 + \mathbf{q}$.

Since $\tilde{M}_0$ is the closure under the $W^{2,2}$-norm of the space

$$\tilde{C}_0^\infty(\omega) = \left\{ \tilde{\mathbf{f}} \in C^\infty(\bar{\omega}) : \tilde{\mathbf{f}}(0, v) = 0, \ \frac{\partial \tilde{\mathbf{f}}}{\partial u}(0, v) = 0 \right\},$$

it is a Hilbert space itself. As any Hilbert space is reflexive, $\tilde{M}_0$ is closed under weak convergence.

Second, by Sobolev embedding again, $W^{2,2}(\omega)$ is compactly embedded in $W^{1,4}(\omega)$. Hence (4.18) is preserved under weak convergence of $W^{2,2}$.

Together, we have shown that $\tilde{M}_0 \cap \tilde{M}_1$ is closed under weak convergence of $W^{2,2}$. Then the same holds for M.

Proof of (iii). This follows simply from the weak LSC of $\mathcal{E}_3$ (Proposition 3.3) restricted to M, using again the fact that $E|_M = \mathcal{E}_3|_M$.

With the above established, we then consider a minimizing sequence for E. By (i), the sequence has a uniform bound in $W^{2,2}$. By (ii), the sequence has a weak $W^{2,2}$ limit $\mathbf{y}_*$ in M. Statement (iii) ensures that $\mathbf{y}_*$ is a minimizer. ∎

4.1.2 Existence of minimizer for (4.2)

Since $\mathcal{E}_3(\mathbf{c})$ is s polynomial function in $\mathbf{c}$, it is also continuous. The constraint set $M_{h,\varepsilon}$, viewed as a subset of $\mathbb{R}^{\dim \mathcal{S}_h}$, is

$$\mathbb{M} := \Big\{ \mathbf{c} \in \mathbb{R}^{\dim \mathcal{S}_h} : \mathcal{I}(\mathbf{c}) \leq \varepsilon \text{ and } \mathbf{c} \text{ satisfies the affine condition determined by (2.3)} \Big\}.$$

To prove existence of minimizer, it suffices to show that $\mathbb{M}$ is a non-empty compact subset of $\mathbb{R}^{\dim \mathcal{S}_h}$. Note that $\mathbb{M}$ is non-empty, as $M_{h,\varepsilon}$ always contains the flat plate $\mathbf{y}^\flat$. It remains to show that $\mathbb{M}$ is closed and bounded.

Since $\mathcal{I} : \mathbb{R}^{\dim \mathcal{S}_h} \to [0, \infty)$ is a polynomial function, it is also continuous, therefore $\mathcal{I}^{-1}([0, \varepsilon])$ is closed. Next, recall that the boundary conditions (2.3) impose a set of affine conditions on $\mathbf{c}$. We conclude that $\mathbb{M}$ is closed as it is the intersection of a closed subset and an affine subspace of $\mathbb{R}^{\dim \mathcal{S}_h}$.

It remains to be shown that $\mathbb{M}$ is bounded. Since $\mathcal{S}_h$ is a finite-dimensional susbspace of $L^2(\omega; \mathbb{R}^3)$, $\mathbb{M}$ is bounded in $\mathbb{R}^{\dim \mathcal{S}_h}$ if and only if $M_{h,\varepsilon}$ is bounded in L^2. For the latter, note that an upper bound for $\iint_\omega \|d\mathbf{y}^T d\mathbf{y} - \mathbb{I}_{2\times 2}\|_F^2$ – given by the $\mathcal{I}(\mathbf{c}) \leq \varepsilon$ assumption – gives an upper bound for $\sum_{p=1,2} \iint_\omega \|\partial_p \mathbf{y}\|_2^2$. As in the previous section, we can then invoke a Poincaré inequality and the boundary condition to obtain a uniform upper bound for $\|\mathbf{y}\|_{L^2}$. ∎

4.1.3 Existence of minimizer for (4.3)

We first establish a coercivity estimate, similar to (4.13), but of the form (for large enough μ)

$$\|\mathbf{y}\|_{W^{2,2}} \leq C_1(\mathcal{E}_3[\mathbf{y}] + \mu \mathcal{I}[\mathbf{y}]) + C_2. \tag{4.19}$$

Reusing the basic trick in (4.16), but without assuming that $\mathbf{y}$ is an isometry, we have

$$\begin{aligned} \mathcal{E}_3[\mathbf{y}] &= \frac{1}{2} \sum_{p,q=1,2} \iint_\omega \|\partial_p \partial_q \mathbf{y}\|_2^2 - 2\langle \partial_p \partial_q \mathbf{y}, \partial_1 \mathbf{y} \times \partial_2 \mathbf{y} \rangle Z_{p,q} + Z_{p,q}^2 \, du\, dv. \\ &\geq \frac{1}{2} \sum_{p,q=1,2} \iint_\omega \|\partial_p \partial_q \mathbf{y}\|_2^2 - \frac{1}{\mu} Z_{p,q} \|\partial_p \partial_q \mathbf{y}\|_2^2 - \mu Z_{p,q} \|\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}\|^2 \, du\, dv + C'. \end{aligned}$$

Note that

$$\iint_\omega \|\partial_1 \mathbf{y} \times \partial_2 \mathbf{y}\|^2 \, du\, dv \leq A\mathcal{I}(\mathbf{y}) + B. \tag{4.20}$$

Hence for μ large enough, it holds that

$$\begin{aligned}\mathcal{E}_3[\mathbf{y}] + \mu\mathcal{I}(y) &\geq \frac{1}{2}\sum_{p,q=1,2}\iint_\omega \|\partial_p\partial_q\mathbf{y}\|_2^2 - \frac{1}{\mu}Z_{p,q}\|\partial_p\partial_q\mathbf{y}\|_2^2 + C' \\ &\geq C\sum_{p,q=1,2}\iint_\omega \|\partial_p\partial_q\mathbf{y}\|^2\,du\,dv - D \\ &= C|\mathbf{y}|_{W^{2,2}} - D.\end{aligned}$$

So (4.19), with $\|\mathbf{y}\|_{W^{2,2}}$ ($= \|\mathbf{y}\|_{W^{1,2}} + |\mathbf{y}|_{W^{2,2}}$) replaced by the semi-norm $|\mathbf{y}|_{W^{2,2}}$, holds for any large enough μ. Finally, the imposed boundary condition (2.3) guarantees, in virtue of the Poincaré inequality, that $\|\mathbf{y}\|_{W^{1,2}}$ is bounded (recall (4.14)), and the desired coercivity estimate follows.

Since the boundary condition (2.3) is equivalent to an affine condition on the spline control variables $\mathbf{c}$, $\mathring{\mathcal{S}}_h$ can be identified with an affine subspace of $\mathbb{R}^{\dim \mathcal{S}_h}$, and $\mathcal{E}_3 + \mu\mathcal{I}|_{\mathring{\mathcal{S}}_h}$ is a degree 4 polynomial function, hence continuous, in $\dim \mathring{\mathcal{S}}_h$ (> 0) variables. By norm equivalence in finite-dimensions and (4.19), this polynomial function is also coercive for large μ. We therefore conclude that a minimizer must exist for $\mathcal{E}_3 + \mu\mathcal{I} : \mathring{\mathcal{S}}_h \to \mathbb{R}$. ∎

Remark 4.5 The mere fact that (4.2) and (4.3) are POPs does not warrant existence of a minimizer. In fact, even an unconstrained POP of which the objective polynomial is bounded below may not possess a minimizer. For instance, $\min\ (1 - x_1x_2)^2 + x_1^2$ has no minimizer in $\mathbb{R}^2$. Note that this degree 4 polynomial objective function lacks the coercivity property.

4.2 Proof of Theorem 4.2

It suffices to show that,

Lower bound: If $\mathbf{y}_h \rightharpoonup_{W^{2,2}} \mathbf{y}$, then

$$\liminf_{h\to 0} \tilde{\mathcal{E}}_{h,\varepsilon_h}[\mathbf{y}_h] \geq \tilde{E}(\mathbf{y}).$$

Upper bound/Recovery sequence: For any $\mathbf{y} \in \mathcal{X}$, there are $\mathbf{y}_h \in \mathcal{X}$ such that $\mathbf{y}_h \rightharpoonup \mathbf{y}$ and

$$\lim_{h\to 0} \tilde{\mathcal{E}}_{h,\varepsilon_h}[\mathbf{y}_h] = \tilde{E}(\mathbf{y}).$$

First we consider Part 1 of the Theorem for $\mathcal{X} = M$.

Proof of lower bound: Let $\mathbf{y}_h \rightharpoonup_{W^{2,2}} \mathbf{y}$. Without loss of generality, suppose $\liminf_{h\to 0} \tilde{\mathcal{E}}_{h,\varepsilon_h}[\mathbf{y}_h] < \infty$, otherwise, there is nothing to prove. Then there is a sequence h_n, such that $\mathbf{y}_{h_n} \in M_{h_n,\varepsilon_{h_n}}$ and hence

$$\tilde{\mathcal{E}}_{h_n,\varepsilon_{h_n}}[\mathbf{y}_n] = \mathcal{E}_3[\mathbf{y}_{h_n}] \text{ and } \liminf_{h\to 0} \tilde{\mathcal{E}}_{h,\varepsilon_h}[\mathbf{y}_h] = \liminf_{n\to\infty} \mathcal{E}_3[\mathbf{y}_{h_n}] < \infty.$$

In particular, $\mathcal{I}[\mathbf{y}_{h_n}] \leq \varepsilon_{h_n}$ for large enough n. Furthermore, since $\mathbf{y}_{h_n} \rightharpoonup \mathbf{y}$ in $W^{2,2}$, $\mathbf{y}_{h_n}$ converges strongly to $\mathbf{y}$ in $W^{1,4}$. So, by the continuity of $\mathcal{I}$ in $W^{1,4}$ (Proposition 3.5) and $\mathcal{I}[\mathbf{y}_n] \leq \varepsilon_{h_n} \to 0$, we have that

$$\mathcal{I}[\mathbf{y}] = \lim_{n\to\infty} \mathcal{I}[\mathbf{y}_{h_n}] = 0.$$

In other words, $\mathbf{y} \in \mathrm{Iso}^{2,2}$. Moreover, $\mathbf{y}_{h_n}$ satisfies the boundary condition (2.3), which is preserved under weak convergence in $W^{2,2}$ (see the proof of weak closedness of M in Section 4.1.1), so $\mathbf{y}$ satisfies (2.3) also. Therefore, $\mathbf{y} \in M$. The lower bound inequality then follows from the LSC of $\mathcal{E}_3$ (Proposition 3.3):

$$\liminf_{h\to 0} \tilde{\mathcal{E}}_{h,\varepsilon_h}[\mathbf{y}_h] = \liminf_{n\to\infty} \mathcal{E}_3[\mathbf{y}_{h_n}] \geq \mathcal{E}_3[\mathbf{y}] \geq \tilde{E}[\mathbf{y}].$$

Proof of upper bound/recovery sequence: Let $\mathbf{y} \in W^{2,2}$ be given. It suffices to assume $\mathbf{y} \in M$, otherwise $\tilde{E}[\mathbf{y}] = \infty$ and the desired upper bound is trivial. For $\mathbf{y} \in M$, we choose the recovery sequence to be $\mathbf{y}_h := Q_h\mathbf{y}$. By property [A1] in Section 3.3, $Q_h\mathbf{y} \rightharpoonup_{W^{2,2}} \mathbf{y}$.

As $\mathbf{y}$ satisfies the clamped boundary condition (2.3), so does $Q_h\mathbf{y}$ (Property [A3], Section 3.3), i.e. $Q_h\mathbf{y} \in \mathring{\mathcal{S}}_h$. By the continuity of $\mathcal{E}_3 : W^{2,2} \to \mathbb{R}$ (Proposition 3.3),

$$\lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}]. \tag{4.21}$$

By the first part of Theorem 3.6, we have

$$\mathcal{I}[Q_h\mathbf{y}] \leq Ch^2\|\mathbf{y}\|^2_{W^{2,2}}$$

Consequently, for h small enough, we have $\mathcal{I}[Q_h\mathbf{y}] = O(h^2) \ll \varepsilon_h$ and hence $\mathcal{E}_{h,\varepsilon_\mathbf{h}}[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}_h]$. Thus we have,

$$\lim_{h\to 0} \mathcal{E}_{h,\varepsilon_\mathbf{h}}[\mathbf{y}_h] = \lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}] = \tilde{E}[\mathbf{y}].$$

For Part 2 of the Theorem, we consider $\mathcal{X} = M \cap C^{d+1}(\omega)$, endowed with the weak-$W^{2,2}$ topology. The proof of the lower bound remains the same. For the proof of the upper bound, the only change is that the recovery sequence is chosen according to [A2] in Section 3.3. Then by the second part of Theorem 3.6, we have

$$\mathcal{I}[Q_h\mathbf{y}] \leq Ch^{2d}\|\mathbf{y}\|^2_{C^{d+1}(\omega)} \ll \varepsilon_h.$$

Hence we similarly have $\mathcal{E}_{h,\varepsilon_\mathbf{h}}[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}_h]$ and that

$$\lim_{h\to 0} \mathcal{E}_{h,\varepsilon_\mathbf{h}}[\mathbf{y}_h] = \lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}] = \tilde{E}_d[\mathbf{y}].$$

■

4.3 Proof of Theorem 4.3

For the first part of the theorem, we consider $\mathcal{X} = M$.
Lower bound. Let $\mathbf{y}_h \rightharpoonup_{W^{2,2}} \mathbf{y}$. If $\mathbf{y} \in M$, then

$$\liminf_{h\to 0} \mathcal{E}^{h,\mu_h}[\mathbf{y}_h] \geq \liminf_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] \geq \mathcal{E}_3[\mathbf{y}] = \tilde{E}[\mathbf{y}]. \tag{4.22}$$

In the above, we have used the weak LSC of $\mathcal{E}_3$ (Corollary 3.3.)

If $\mathbf{y} \notin M$, then $\mathbf{y}$ either fails to satisfy the boundary condition (2.3) or $\mathcal{I}[\mathbf{y}] > 0$. At the same time, $\tilde{E}[\mathbf{y}] = \infty$.

- Suppose $\mathbf{y}$ does not satisfy the boundary condition (2.3). We claim that, for all small enough h, $\mathbf{y}_h$ will not satisfy (2.3) either and hence $\liminf_{h\to 0} \tilde{\mathcal{E}}^{h,\mu_h}[\mathbf{y}_h] = \infty$ and so the lower bound is automatically satisfied. For otherwise, there exists $h_n \to 0$ such that each $\mathbf{y}_{h_n}$ satisfies the boundary condition, then by the preservation of the boundary condition under weak convergence, $\mathbf{y}$ will satisfy the boundary condition also, which causes a contradiction.

- Next, if $\mathcal{I}(\mathbf{y}) > 0$, then $\lim_{h\to 0} \mathcal{I}[\mathbf{y}_h] = \mathcal{I}(\mathbf{y}) > 0$ (as $\mathbf{y}_h \to \mathbf{y}$ in $W^{1,4}$). Combined with the assumption that $\lim_{h\to 0} \mu_h = \infty$, we have $\lim_{h\to 0} \mu_h \mathcal{I}[\mathbf{y}_h] = \infty$. Note that (4.22) implies in particular that $\mathcal{E}_3[\mathbf{y}_h]$ is bounded from below uniformly in h. Hence we also have

$$\lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] + \mu_h \mathcal{I}[\mathbf{y}_h] = \infty,$$

  and $\liminf_{h\to 0} \tilde{\mathcal{E}}^{h,\mu_h}[\mathbf{y}_h] = \infty = \tilde{E}[\mathbf{y}]$.

Upper bound/Recovery sequence. For any $\mathbf{y}$, if $\mathbf{y} \notin M$, then $\tilde{E}[\mathbf{y}] = \infty$, and there is nothing to prove. If $\mathbf{y} \in M$, $\mathbf{y}_h := Q_h\mathbf{y} \in \mathcal{S}_h$ satisfies (2.3), $\lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \mathcal{E}_3[\mathbf{y}] = \tilde{E}(\mathbf{y})$, and $\mathcal{I}[\mathbf{y}_h] = O(h^2)$ by the first part of Theorem 3.6. Consequently, the condition on μ_h gives $\lim_{h\to 0} \mu_h \mathcal{I}[\mathbf{y}_h] = 0$. Hence $\lim_{h\to 0} \tilde{\mathcal{E}}^{h,\mu_h}[\mathbf{y}_h] = \lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \tilde{E}[\mathbf{y}]$.

For Part 2 of the Theorem with $\mathcal{X} = M \cap C^{d+1}(\omega)$, again the only change is in the proof of the upper bound. The use of [A2] in Section 3.3 leads to $\mathcal{I}[\mathbf{y}_h] = O(h^{2d})$ by the second part of Theorem 3.6. Consequently, the condition on μ_h gives $\lim_{h\to 0} \mu_h \mathcal{I}[\mathbf{y}_h] = 0$. Hence $\lim_{h\to 0} \tilde{\mathcal{E}}^{h,\mu_h}[\mathbf{y}_h] = \lim_{h\to 0} \mathcal{E}_3[\mathbf{y}_h] = \tilde{E}_d[\mathbf{y}]$. ∎

5 Algorithmic Considerations & Computational Results

Let us summarize what our numerical solution of the bilayer problem (2.1) so far. We have developed two numerical solutions, (2.37) and (2.38). Both methods are based on uniform knot tensor product B-splines applied to a rectangular domain $\omega = [0, A] \times [0, B]$, and we developed efficient algorithms for computing the relevant functionals $\mathcal{I}$ and $\mathcal{E}_3$ defined in Sections 2.2-2.3. The first method is a constrained polynomial optimization problem with the specification of a tolerance ε, whereas the second is an unconstrained one with the specification of a penalization parameter μ.

The Γ-convergence results from Section 4 say that if ε and μ are chosen according to

$$\varepsilon(\mathbf{h}) = C_\epsilon |\mathbf{h}|^{2d-\epsilon}, \quad \mu(\mathbf{h}) = c_\epsilon |\mathbf{h}|^{-2d+\epsilon} \tag{5.1}$$

for any arbitrarily small $\epsilon > 0$ and any constants C_ϵ, $c_\epsilon > 0$, then 'discrete minimizers converge to the continuum minimizer' as $h \to 0$. (The above requires the assumption that the solution is smooth enough.) When $Z = \frac{1}{r}\mathbb{I}_{2\times 2}$, $h = \max\{h_1, h_2\}$ can be replaced by h_1; see below. In practice, ε (resp. μ) should be chosen to be as small (resp. big) as possible in order for the computed surface to be close to an isometry, but yet not too small (resp. big), lest the resulted surface may be forced to be close to the flat plate which is the only spline surface that satisfies $\mathcal{I} = 0$. We shall show in Section 5.2 that this is indeed the case. The difficult conundrum in the choice of ε and μ may be attributed to the use of a linear spline space for approximating the inherently nonlinear space of isometries.

As a benchmark, when $Z = \frac{1}{r}\mathbb{I}_{2\times 2}$ and $\omega = [0, A] \times [0, B]$, the variational problem (2.1) has the unique minimizer

$$\mathbf{y}(u, v) = [r\cos(u/r - \pi/2),\ v,\ r + r\sin(u/r - \pi/2)]^T; \tag{5.2}$$

see [22]. Its energy is

$$E = \frac{AB}{2r^2}. \tag{5.3}$$

The parameterized surface above is the flat rectangle ω 'rolled up' into a cylinder of radius r and height B.[6] This is an interesting result as the isotropic $Z = \frac{1}{r}\mathbb{I}_{2\times 2}$ encourages the plate to be umbilical, but this is disallowed by the isometry constraint. It is also a representative result, as the fabrication of nanotubes based on bilayer films is a predominant motivation for the study of (2.1).

In general, we need $h = |\mathbf{h}| = \max(h_1, h_2)$ to be small in order to have a good approximation of the solution. But in this special case, it can be shown that minimizers of our proposed discrete problems, (4.2) or (4.3), converge to that of the continuum problem, i.e. (5.2), with any fixed n – in fact just $n = 1$ – and $m \to \infty$, with the proviso that ε and μ are chosen according to (5.1), but with $|\mathbf{h}| = \max(h_1, h_2)$ replaced by h_1. In other words, we only need $h_1 = A/m$, but not $h_2 = B/n$, to be small in order to obtain a good approximation. (This is nothing but a reflection of the fact that the exact solution depends on only one variable.) The proof is an easy adaptation of the arguments based on Γ-convergence in Section 4; the only extra argument needed is the fact that the cylinder (5.2) satisfies $\partial^2\mathbf{y}/\partial v^2 = 0$, which yields, by a result on tensor-product splines [26, Theorem 8], that

$$\left\|\partial_u(\mathbf{y} - Q_{\mathbf{h}}\mathbf{y})\right\|_{L^\infty} \le Ch_1^d\left\|\partial_u^{d+1}\mathbf{y}\right\|_{L^\infty}, \quad \left\|\partial_v(\mathbf{y} - Q_{\mathbf{h}}\mathbf{y})\right\|_{L^\infty} \le Ch_1^{d+1}\left\|\partial_u^{d+1}\partial_v\mathbf{y}\right\|_{L^\infty}, \tag{5.4}$$

provided that the degree of the spline d is at least 1. (Recall from (2.7) that we choose d to be at least 2.)

After discussing our choice of optimization solvers, the numerical results will be presented in Sections 5.2 to 5.8. The following are some further remarks about our numerical experimentation.

1. We will present results for both the ε- and μ-problems. But in comparison and in hindsight, we do find the latter, being an *unconstrained* problem, is relatively easier to implement and runs faster. Hence starting from Section 5.5, we will only report results for the μ-problem.

2. Due to the refinability of spline functions, it seems reasonable and advantageous to iteratively refine the mesh instead of always starting from the same initial data, for example the flat plate. We have tried both but the results do not indicate much difference or reveal any significant phenomena. Hence, due to space constraint, for some of the following sections, we will selectively present only one of the two methods.

3. In the most challenging computations reported in Figures 5 and 6, typical running times on a personal desktop (with an Intel i7 CPU processor at 3.7 GHz) can range from tens of hours to a few days. On a processor in Google cloud, they can range from a couple of hours to one or two days, rougly three times faster.

[6] In fact, if we dispense with the boundary conditions in (2.1), and Z continues to be isotropic, i.e. $Z = \frac{1}{r}\mathbb{I}_{2\times 2}$, then the rectangle ω rolled up into a cylinder of radius r **along any axis** would have the same (minimal) energy $\frac{\text{area}(\omega)}{2r^2}$; and the result is not specific to ω being a rectangle.

5.1 Choice of optimization solvers

Ideally, we would like to use the sparse POP solver [31, 30], as in principle the convexification approach guarantees the finding of a global minimizer. However, the inevitable dimension elevation in the convexification renders the method impractical for our problem, except for very small problem sizes. As a simple illustration of such a dimension elevation, consider the one-dimensional unconstrained POP: $\min_{x\in\mathbb{R}} p(x)$, where $p(x)$ is a degree 4 polynomial. This nonconvex problem is equivalent to the following convex optimization problem:

$$\max_{\gamma,Q} \gamma \quad \text{s.t.} \quad p(x) - \gamma = [1, x, x^2]Q[1, x, x^2]^T, \quad \gamma \in \mathbb{R}, \quad 0 \preceq Q \in \mathbb{R}^{3\times 3}_{\text{sym}}.$$

The univariate 'x' in the original non-convex problem is replaced by the 7 variables in γ and the upper triangular part of Q in the convex semi-definite program. This convexification relies on the fact that an univariate non-negative polynomial is always a sum of squares, an innocent-sounding fact that is no longer true in higher-dimensions. This is where the wisdom of Lasserre's relaxation [18], founded on Putinar's Positivstellensatz [21], comes in. For a general POP, we no longer have a clear-cut equivalent convex reformulation. Instead, the theory promises a *hierarchy* of convex problems of which their solutions, somewhere deep enough in the hierarchy, give the solution of the POP. None of the unknown depth, the dimension elevation, or the practical difficulty in solving high-dimensional semi-definite programs is helpful. With the tools currently available to us, the cost of convexification simply outweighs the benefit.[7]

We resort to traditional solvers based on quasi-Newton methods. However, not all is lost, as the POP structure facilitates:

1. the use of the formulas derived in Section 2.2 and Section 2.3 to compute $\mathcal{I}$, $\nabla\mathcal{I}$, $\mathcal{E}_3$ and $\nabla\mathcal{E}_3$ *without the need of any numerical integration*, and

2. *the ease of implementing exact line search*, as solving $\min_{t\geq 0} p(x + tv)$ amounts to finding the global minimizer of a univariate degree 4 polynomial when p is a (multivariate) degree 4 polynomial.

For the ε-problem, we simply use `fmincon` in the optimization toolbox of Matlab. In solving the μ-problem, we find that in a number of cases our exact line search BFGS solver, denoted by BFGS-exact in later writing, gives more accurate results than Matlab's `fminunc`, which is supposed to implement BFGS with a version of Wolfe line search.

5.2 ε and μ should not be too big or too small

Our first set of computation results confirm the basic premise that ε and μ in (4.2) and (4.3) (respectively) should not be chosen to be too big or too small. We consider a relatively easy case of the problem, namely $A = 2$, $B = 1$, $Z = \mathbb{I}_{2\times 2}$, and apply our methods with small discretization

[7] At the time of the writing of this article, we became aware of the recent article [32] for solving very large semi-definite programs. Potentially, this can be combined with the convexification methods of POPs to efficiently find a near global optimizer, which can then be used an initial guess for a traditional quasi-Newton or gradient descent method. It is an open problem to see how practical such a method is and if such an approach can provably guarantee the finding of a global minimizer.

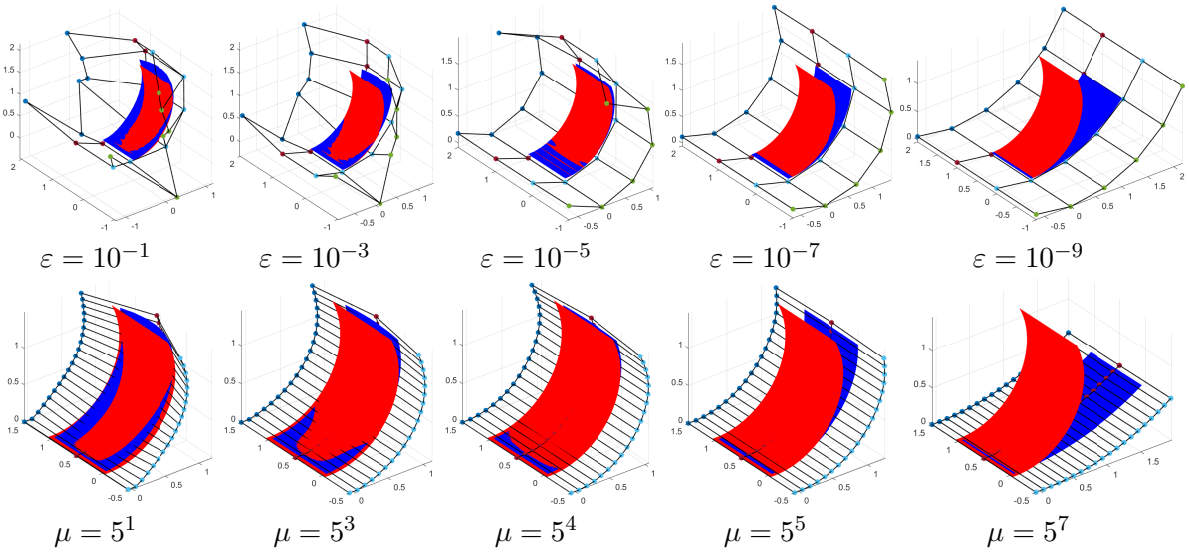


Figure 1: In each panel, the red surface is the exact solution (5.2) of (4.1), with $A = 2$, $B = 1$, $Z = \mathbb{I}_{2\times 2}$. The wireframes depict the control points of the blue B-spline surfaces, which are the computed minimizers. Top row: minimizers of the 'ε-problem' (4.2), with spline degree $d = 3$, grid size $(m, n) = (3, 1)$, and $\varepsilon = 10^{-i}$ for $i = 1, 3, 5, 7, 9$. Bottom row: minimizers of the 'μ-problem' (4.3), with spline degree $d = 2$, grid size $(m, n) = (18, 1)$, and $\mu = 5^i$ for $i = 1, 3, 4, 5, 7$.

parameters. When the problem size is small, it is possible to use the sparse POP solver [31] to verify if our computed solutions are close to a global minimizer.

The top row of Figure 1 shows the computed solutions of the ε-problem (4.2), with spline degree $d = 3$ and grid sizes $m = 3$, $n = 1$, and various values of ε. It is evident that when ε is too big ($\varepsilon = 10^{-1}$ and 10^{-3}), the solution (in blue) is far away from the true solution of (4.1) (in red). In particular, the surface is far from being isometric to the flat plate. In this example, $\varepsilon = 10^{-5}$ gives the best approximation to the true solution (5.2). If ε is too small, the minimizer approaches the flat rectangle.

The same behavior is observed for the μ-problem (4.3). The computed solution is far from an isometry if μ is too small, and close to the flat plate when μ is too big. In the bottom row of Figure 1, it is visually clear that the intermediate $\mu = 5^4$ gives the best approximation to the true solution among the five values of μ.

5.3 Empirical convergence

In this and the following sections we report computational results by choosing ε and μ to be

$$\varepsilon(\mathbf{h}) = |\mathbf{h}|^{2d-1}, \quad \mu(\mathbf{h}) = |\mathbf{h}|^{-2d+1}. \tag{5.5}$$

This is (5.1) with $\epsilon = 1$ and $C_\epsilon = c_\epsilon = 1$. According to our theory, this choice guarantees Γ-convergence.

Figure 2 depicts our computed solutions for the case of $A = 2$, $B = 1$, $Z = 2\mathbb{I}_{2\times 2}$. The grid size is chosen to be $m = 2^i$, $i = 3, 4, 5, 6$ and $n = 1$, and ε or μ is chosen according to (5.5), but with

$|\mathbf{h}| = \max(A/m, B/n)$ replaced by $h_1 = A/m$ – recall the comments around (5.4) for a justification of this choice.

We explore the ε- and μ-problem for both $d = 2$ and $d = 3$. In all four cases, we employ iterative refinement. More precisely, the initial guess of the smallest sized problem, namely $(m, n) = (8, 1)$, is solved with the flat plate as the initial guess. For larger values of m, the control mesh of the resulted minimizing B-spline surface is refined, using the refinability of spline functions, to a control mesh that is double in size in the first direction. Then resulted control mesh is used as the initial guess of the numerical optimization with the next larger problem size.

We shall now make a series of observations based on Figure 2. Along the way, we expose some of the elusive behaviors of the numerical solutions, caused by the confluence of *floating point errors*, an unexpected *asymmetry* property of discretization, a (not so well-known) symmetry-preserving property of standard optimization methods, and the *non-convexity* of the problem.

We begin with a mundane comment: the computational results support our theoretical prediction that the method converges. By inspecting the figures from top to bottom in each column of Figure 2, as the grid size m increases, the computed solution appears to approach the true solution of the continuous problem (5.2), both visually and according to the values of $\mathcal{I}$, $\mathcal{E}_3$ and the L^∞ error. By (5.3), the energy of the cylinder is 4 in this case. As m increases, we indeed observe that $\mathcal{I}$ approaches 0, $\mathcal{E}_3$ approaches the value 4 and the L^∞ error decreases. As expected, degree 3 splines give more accurate results than degree 2 splines.

It happens that in this example, the degree 3 cases are more susceptible to 'symmetry breaking' than degree 2 – compare columns 2 and 4 vs columns 1 and 3, although such a symmetry breaking is also observed in the degree 2 cases – see Section 5.6. We will have more discussion on this phenomenon in Section 5.5.

When ε is too small or μ is too big, our solvers may encounter numerical difficulties due to floating point issues. This can be seen in the bottommost panel of the second and fourth columns: in these two cases, not only that the computed surface with $m = 2^6$ has a higher L^∞ error compared to the one computed with $m = 2^5$, but also that *the solver fails to find a feasible point*. Notice that the $\mathcal{I}$ value of the bottommost panel in the second column is higher than the desired ε. The problem must be caused by floating errors and numerical issues pertaining to the constraint optimization solvers. Both `fmincon()` in Matlab and `SNOPT` fail to find a feasible point in this case. To get an idea of the level of floating errors, consider the control data $\mathbf{c}^\flat$ defined by (2.12) with $(m, n) = (64, 1)$ and $d = 3$. Our Matlab implementation of $\mathcal{I}$ gives $\mathcal{I}(\mathbf{c}^\flat) = -7.22\ldots \times 10^{-11}$, while $\mathcal{I}$ should always be non-negative and $\mathcal{I}(\mathbf{c}^\flat) = 0$.

5.4 Minimizer of the ε-problem is at the boundary

Our computations also show that *the minimizer of the ε-problem is located at the boundary of the constraint set*, i.e. $\mathcal{I}(\mathbf{c}^*) = \varepsilon$ if $\mathbf{c}^*$ is a minimizer of the ε-problem. See the numerical results in the captions reported in the first two columns of Figure 2. We offer an intuitive explanation for the observation. Recall from (2.28) that the functional $\mathcal{E}_3$ only coincides with the energy functional E when applied to an isometry, but a B-spline surface is almost never perfectly flat. Recall also that E is always non-negative, whereas $\mathcal{E}_3$ is a cubic polynomial, meaning that there are directions so that its value goes to $-\infty$. In other words, *E and $\mathcal{E}_3$ can be very different away from the isometries*! When minimizing $\mathcal{E}_3$ over the constraint 'tube' $\mathcal{I}(\mathbf{c}) \leq \varepsilon$, the optimization method pushes $\mathbf{c}$ towards a surface that approximates the solution of the variational problem, but the 'by-and-large correct'

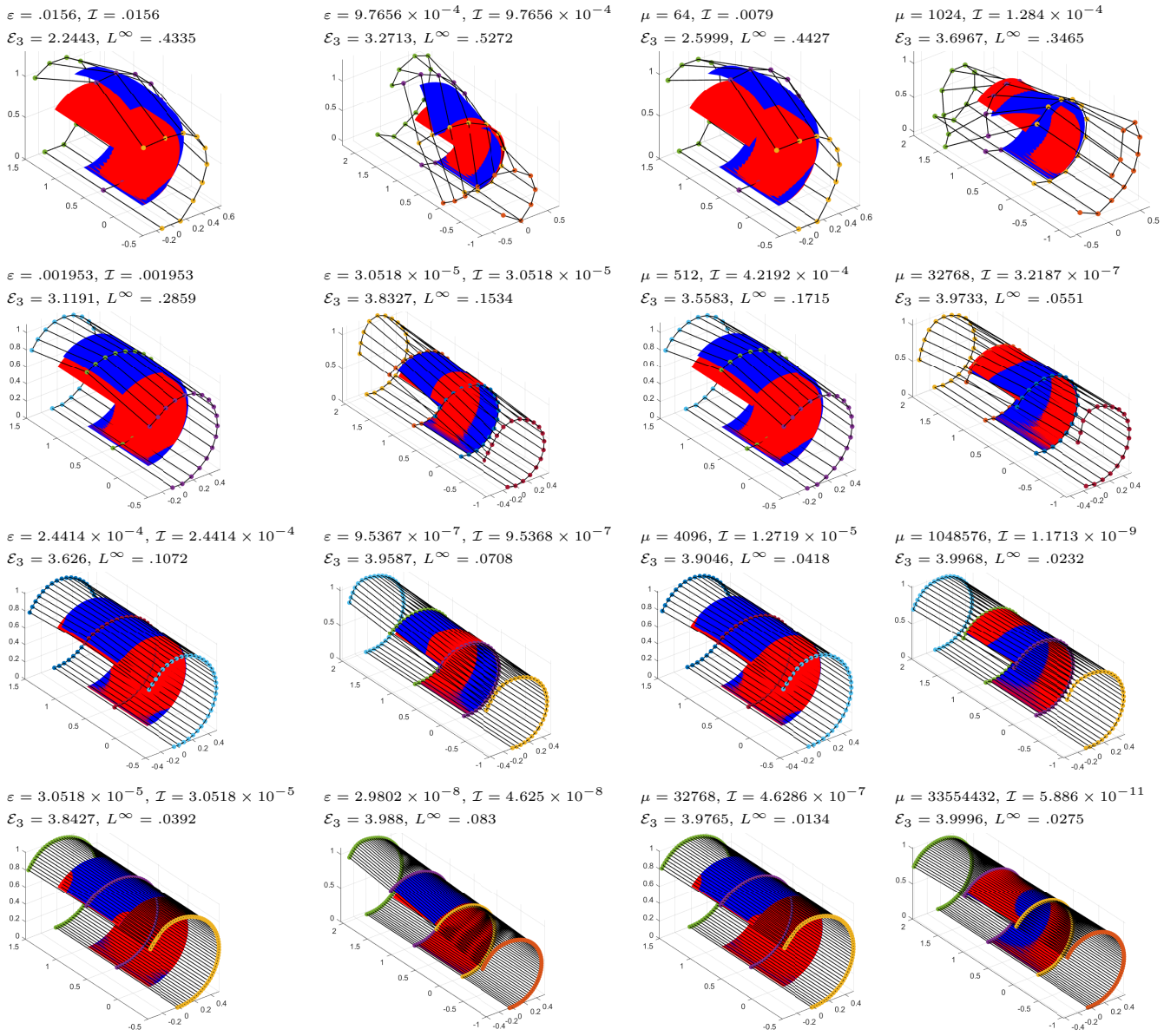


Figure 2: In each panel, the red surface is the exact solution (5.2) of (4.1), with $A = 2$, $B = 1$, $Z = 2\mathbb{I}_{2\times 2}$ and the blue surface is the computed solution. First and second columns: minimizers of the 'ε-problem' (4.2), with spline degree $d = 2$ (first column) and $d = 3$ (second column), third and fourth columns: minimizers of the 'μ-problem' (4.3), with spline degree $d = 2$ (3rd column) and $d = 3$ (4th column). In each column, the grid sizes are chosen to be, from top to bottom, $(m, n) = (2^i, 1)$ for $i = 3, 4, 5, 6$. In each case, ε or μ is chosen according to (5.5), with $|\mathbf{h}|$ replaced by h_1. The value denoted by L^∞ is the error between the computed and exact (cylinder) solutions.

search direction has a 'bad component' that pushes $\mathbf{c}$ towards an $-\infty$ direction of $\mathcal{E}_3$. The latter also means that the method pushes $\mathbf{c}$ towards the boundary of the constraint set.

5.5 Symmetry

The cylinder (5.2) has a reflectional symmetry about the $y = B/2$ plane. Let us first spell out some elementary algebra: If $S \subset \mathbb{R}^3$, then its reflection about the plane $y = B/2$ is given by

$$\text{flip}(S) := \{[x, B - y, z]^T : [x, y, z]^T \in S\}.$$

If S is a parametric surface with parameterization $\mathbf{y} : [0, A] \times [0, B] \to \mathbb{R}^3$, then it is most natural to parameterize $\text{flip}(S)$ by

$$\begin{aligned} \text{flip}(\mathbf{y}) &: [0, A] \times [0, B] \to \mathbb{R}^3, \\ \text{flip}(\mathbf{y})(u, v) &:= [\mathbf{y}_1(u, B - v), B - \mathbf{y}_2(u, B - v), \mathbf{y}_3(u, B - v)]^T. \end{aligned} \tag{5.6}$$

Then we have the following equivalent (reflectional) symmetry property:

$$S = \text{flip}(S) \text{ (as point sets)} \iff \text{flip}(\mathbf{y}) = \mathbf{y} \text{ (as maps)}. \tag{5.7}$$

Also, $\mathbf{y}$ satisfies the clamped boundary condition (2.3) iff $\text{flip}(\mathbf{y})$ satisfies the same condition.

The map in (5.2) clearly satisfies $\text{flip}(\mathbf{y}) = \mathbf{y}$. In this case, should we expect that our numerical solution satisfies the same symmetry? The numerical solutions shown in of Figure 2 suggest that the answer is positive when $d = 2$ (columns 1 and 3), but negative when $d = 3$ (columns 2 and 4). (We also use $\|\mathbf{y} - \text{flip}(\mathbf{y})\|_\infty$ to quantify the degree of symmetry breaking in our exploration.)

We focus on the fourth column, as it is easier to analyze unconstrained optimization methods. We first point out that the BFGS method (used for solving the μ-problem), in *exact arithmetic*, preserves symmetry, i.e. if the initial guess satisfies (5.7), such as the flat plate $\mathbf{y}^\flat$, then all subsequent iterates satisfy the same symmetry. (This symmetry preserving property was explored in our recent work [10].) This holds regardless of whether the (discrete) global minimizer is symmetric or not.

When $d = 2$, the global minimizer appears to be symmetric, while for $d = 3$, the global minimizers appear to be asymmetric. Floating point errors seem to help break the symmetry preservation property in the case when the actual (discrete) global minimizer is asymmetric. We observe also in Figure 2, columns 2 and 4, that the degree of asymmetry vanishes as m grows, as it should in virtue of our Γ-convergence results.

Such a convoluted situation deserves a recap and further explorations. We first summarize:

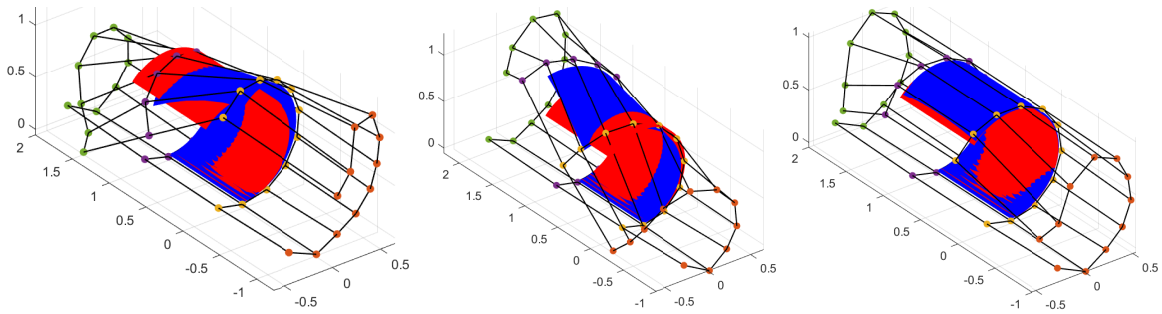
(i) The solution of the variational problem, when $Z = \frac{1}{r}\mathbb{I}_{2\times 2}$, is reflectional symmetric.

(ii) Yet, the solution of the discrete problem may not share the same reflectional symmetry.

(iii) Our optimization algorithm preserves symmetry in exact arithmetic.

(iv) Yet, floating point error helps breaking symmetry.

We provide firmer evidences for (ii) above.



$\mathcal{E}_3 + \mu\mathcal{I} \approx 3.8282$ $\mathcal{E}_3 + \mu\mathcal{I} \approx 3.8282$ $\mathcal{E}_3 + \mu\mathcal{I} \approx 3.8293$

Figure 3: Solutions of the μ-problem with $m = 8$, $n = 1$, $d = 3$, $\mathbb{Z} = 2\mathbb{I}_{2\times 2}$, $\mu = 1024$. First two panels: solutions obtained from our BFGS-exact solver with a Gaussian perturbed flat plate as initial guess. They are flipped versions of each other. Third panel: solution obtained from the same BFGS-exact solver but with a symmetrization step added, and using the flat plate as initial guess.

We explore the case of the μ-problem for $(m, n, d) = (8, 1, 3)$, i.e. the one in the panel on the second row and fourth column of Figure 2. The results are illustrated in Figure 3. Instead of running the BFGS method with the flat plate as initial guess, we perturb every coordinate of the flat plate by a Gaussian random number $\sim N(0, \sigma^2)$ with $\sigma = 0.1$. When solving the same problem in this way, approximately half of the times we obtained the same solution as shown in the panel, and the rest of the times we obtain the flipped version of the same surface. In all trials, the objective value agrees up to 10 significant digits, with $\mathcal{E}_3 + \mu\mathcal{I} \approx 3.828199$. We increase σ to 1 and to 10 even, and the same results are obtained. These are strong evidences that the asymmetric blue surface in the panel, and the flipped version of it, are the unique global minimizers.

In contrast, if we compare the results above with what would be gotten by replacing the BFGS solver with a 'symmetrized counterpart', i.e. in each BFGS step we replace the iterate $\mathbf{c}^k$ by $(\mathbf{c}^k + \text{flip}(\mathbf{c}^k))/2$ in order to remove the asymmetry in $\mathbf{c}^k$ caused by floating point errors.[8] Using either the flat plate or a quasi-interpolator of the exact solution (both reflectional symmetric) as the initial guess, the minimal objective value is $\mathcal{E}_3 + \mu\mathcal{I} \approx 3.829258$ (> 3.828199)! This strongly suggests that the global minimum of the objective, when restricted to the subspace of symmetric control meshes, is strictly greater than the true global minimum.

Note that the symmetric solution in the third panel of Figure 3 is a more accurate approximation to the true solution of the continuum variational problem (the red surface), while being a less accurate approximation to a global minimizer of the discrete problem. (Of course, the use of the 'symmetrized solver' is a cheat and is just for the sake of our argument here.)

Finally, our computations suggest that the symmetric blue surface in the third panel of Figure 3 is likely to be a *saddle point* of the objective functional $\mathcal{E}_3 + \mu\mathcal{I}$, while the blue surfaces in the first and second panels are (local) minimizers of the same functional. These statements are seen by computing the lowest eigenvalue of the Hessian of $\mathcal{E}_3 + \mu\mathcal{I}$, using the formulas developed in Sections 2.2 and 2.3. In the former case, the lowest eigenvalue is found to be negative, hence the solution is a saddle point, while in the latter case, the lowest eigenvalue is found to be positive, hence a local minimizer.

[8] The flip operator in (5.6) descends to an operator for the control vertices of tensor product B-spline surfaces.

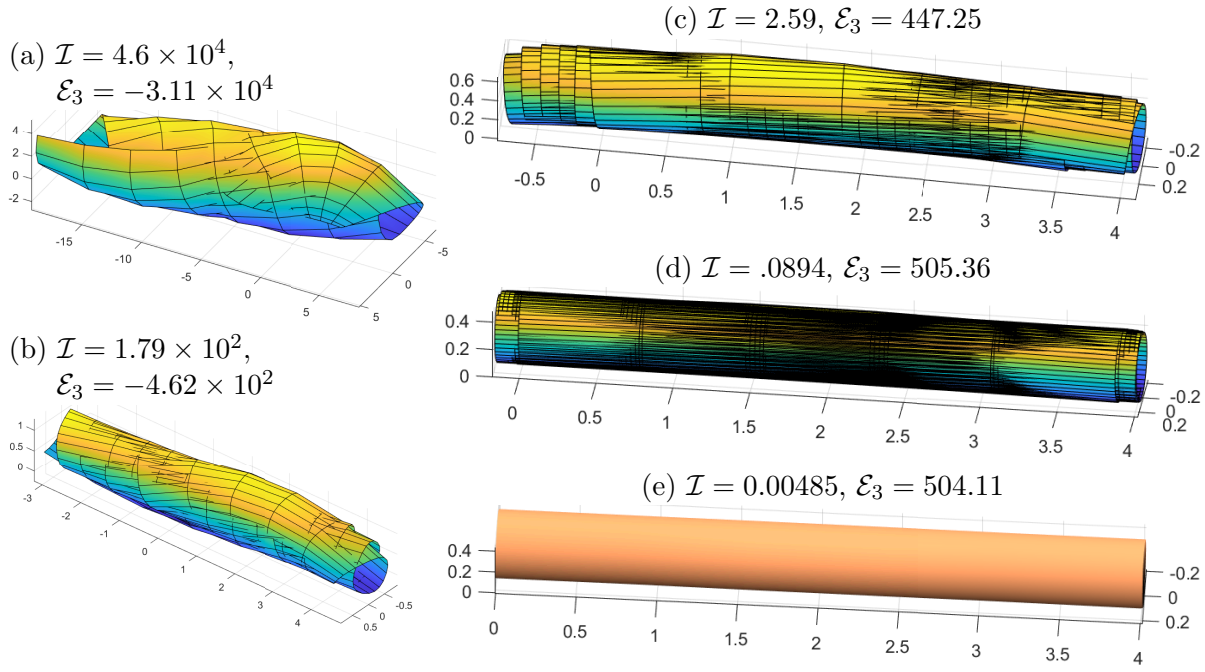


Figure 4: Solutions of the μ-problem in subsection 5.6, with $m = 2^i$, $i = 3, 4, 5, 6, 7$, $n = 1$, $d = 2$. As m increases, $\mathcal{I}$ approaches 0 and $\mathcal{E}_3$ approaches $AB/2r^2 = 500$. (With $m = 256$, the minimizer satisfies $\mathcal{I} = .000305$, $\mathcal{E}_3 = 501.41$, and the surface is visually the same as that in panel (e).)

5.6 A challenging problem: $A = 10$, $B = 4$, $Z = 5\mathbb{I}_{2\times 2}$

We use our method to solve the problem with parameters listed in the subsection heading. With a larger physical size but a much smaller radius ($r = 1/5$), the plate *rolls up a lot more*. As such, this instance of the problem is considered by [6] to be numerically challenging.

Our method is able to handle this situation. In Figure 4, we show the results with spline degree $d = 2$ and grid sizes $(m, n) = (2^i, 1)$, $i = 3, 4, 5, 6, 7$. We use the μ-method, again with μ chosen by (5.5) with $|\mathbf{h}|$ replaced by h_1. As in Figure 2, we use iterative refinement, and the flat plate as the initial guess for the coarsest grid. Unlike the situation in Figure 2, we see clear 'symmetry breaking' – which is necessarily caused by floating point errors – even with $d = 2$. The symmetry breaking manifests itself in the protrusion seen in the rolled-up plate; see panel (c) and (d). The degree of protrusion diminishes as the grid size increases, and is hardly visible in panel (e).

As expected, we obtain similar, but more accurate, results with spline degree $d = 3$.

Given the non-convexity of the problem and that the flat plate is quite far away from the cylinder, it is a pleasant surprise that the method succeeds in finding the global minimizer using the flat plate as an initial guess. However, we confess that in solving this cylinder problem, there are many cases with bigger values of A, B and/or smaller values of $r = 1/\alpha$ for which our method would not converge to the cylinder (5.2). In these cases, the method leads to 'spurious' critical points which do not look like the expected solution at all. This is a clear reflection of the highly non-convex nature of the problem.

5.7 A non-diagonal, anisotropic problem: $A = 6$, $B = 4$, $Z = \left[\begin{smallmatrix} 3 & -2 \\ -2 & 3 \end{smallmatrix}\right]$

In this subsection, we consider the case listed in the heading, following the *corkscrew shape* example in [6, Section 6]. Note that the spontaneous curvature matrix Z is not only non-diagonal, but also anisotropic: the principle curvatures are 5 and 1, and the principle directions are $[1, -1]^T$ and $[1, 1]^T$, respectively. The plate is expected to roll up along a line forming the angle 45° with the coordinate axes.

The left column of Figure 5 shows the numerical solutions of the μ-problem with $(m, n) = (2^i A, 2^i B)$, $i = 1, 2, 3$, $d = 3$, with μ chosen by (5.5). We again use the BFGS-exact solver, and use the flat plate as the initial guess for the smallest problem, and progress to the larger size problems using iterative refinements. We obtain similar results (shown on the right column) for the same μ-problem without using iterative refinements, with $(m, n) = (kA, kB)$, $k = 3, 5, 7, 9$, $d = 3$ and always using the flat plate as the initial guess. While we do not observe clear distinction between the two sets of solutions, iterative refinement is certainly faster.

5.8 Diagonal anisotropic spontaneous curvatures

Notice that when $Z = \left[\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right]$, the cylinder (5.2) with $r = 1/\alpha$, being an isometry with second fundamental form $H = Z$, has zero energy and hence must be the unique minimizer for the continuous problem. In [22], Schmidt established that the same cylinder is a minimizer of the problem when $Z = \left[\begin{smallmatrix} \alpha & 0 \\ 0 & \alpha \end{smallmatrix}\right]$. Given this, it seems natural to conjecture that the same cylinder is a minimizer of the problem when $Z = \left[\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}\right]$ when $0 \le \beta \le \alpha$, with minimal energy

$$E = \frac{1}{2}\int_0^A \int_0^B \left\| \left[\begin{smallmatrix} \alpha & 0 \\ 0 & 0 \end{smallmatrix}\right] - \left[\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}\right] \right\|_F^2 \, du\, dv = \frac{\beta^2 AB}{2}.$$

This same cylinder appears to be a minimizer even when $\beta \in [-\alpha, \alpha]$. See the first row of Figure 6 for an evidence of this conjecture when $A = 6$, $B = 4$ and $Z = \left[\begin{smallmatrix} 5 & 0 \\ 0 & -5 \end{smallmatrix}\right]$. Note that at coarse discretization level, the solution surface (first panel, top row) is far from an isometry and exhibits hyperbolic points at the two ends of the cylinder.

When β is outside $[-\alpha, \alpha]$, the solution surface should, by the same conjecture, have a tendency to roll up like a cylinder, but now in the v-direction. The clamped boundary condition at $u = 0$, however, would disallow this to happen. Hence, the solution is not expected to be a cylinder. The second row of Figure 6 shows our simulation result when $A = 6$, $B = 4$ and $Z = \left[\begin{smallmatrix} 5 & 0 \\ 0 & 6 \end{smallmatrix}\right]$.

6 Summary and Perspectives

The main challenges of the current problem are due to the high order (4-th order) of the equation and also the nonlinear isometry constraint. As demonstrated in this paper, appropriate choice of the approximation spaces which are in particular conforming, renders an efficient and yet simple strategy - a sparse polynomial type optimization algorithm. It is certainly of interest to see how such an approach can be understood in a broader context and also extended to a wider range of problems. Here we provide some points for further exploration.

First, the current Γ-convergence results Theorems 4.2, 4.3, and 4.4 that relate the discrete and continuum problems are qualitative results. It will definitely be of practical interest to investigate the convergence rate of the approximate solution. The error estimate might depend, among other

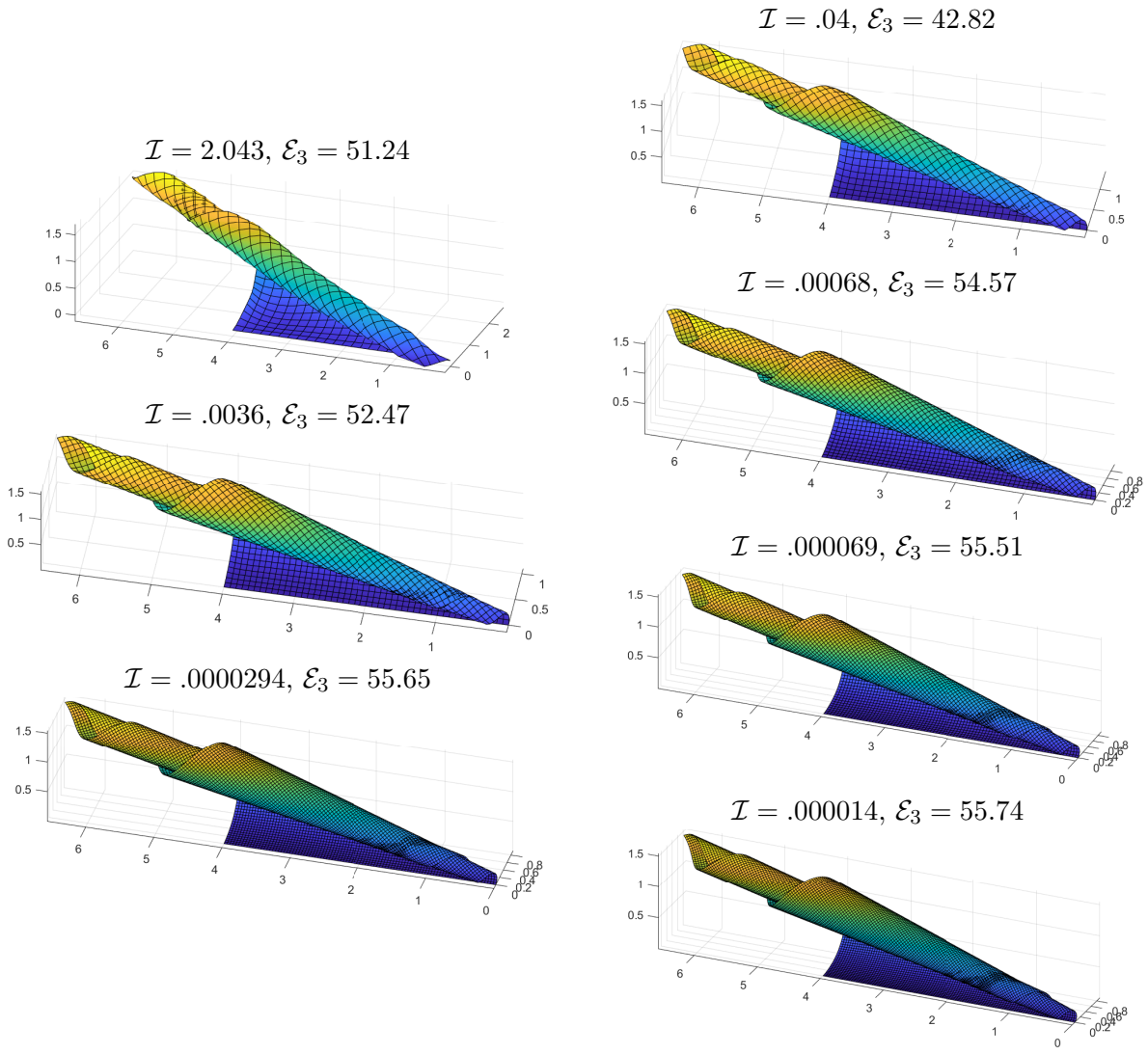


Figure 5: Left column: solutions of the μ-problem in subsection 5.7, with $(m, n) = (2^i A, 2^i B)$, $i = 1, 2, 3$, $d = 3$, using iterative refinements. Right column: solutions of the same μ-problem, with $(m, n) = (kA, kB)$, $k = 3, 5, 7, 9$, $d = 3$ and without using iterative refinements.

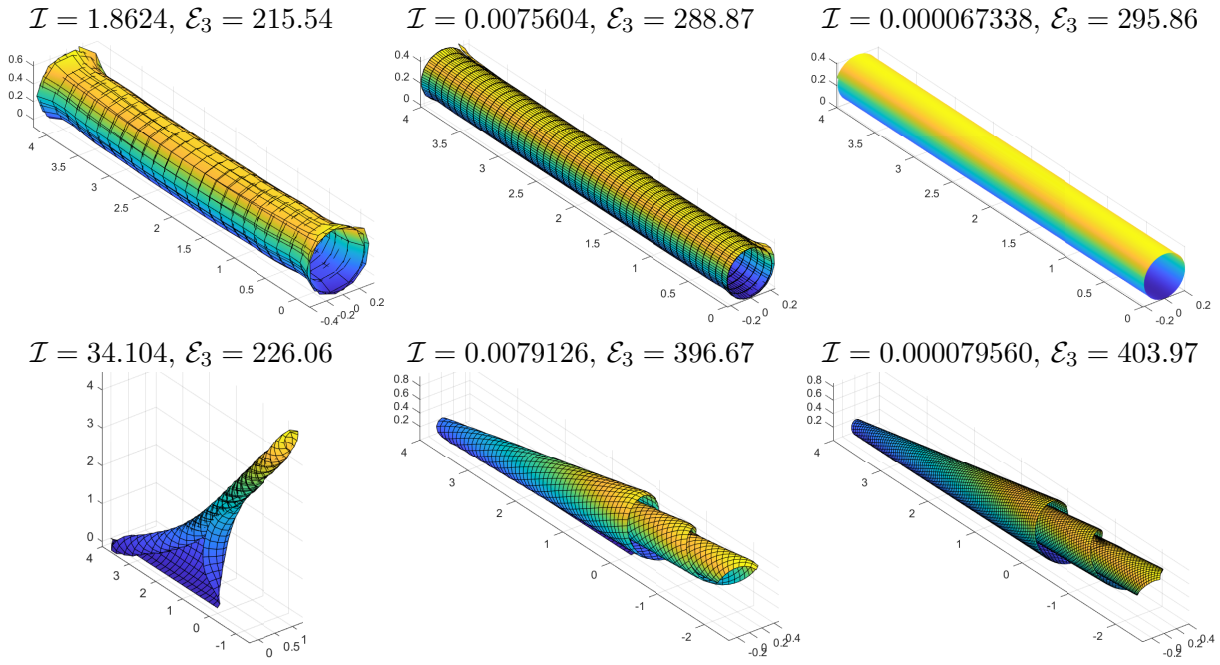


Figure 6: Simulation results for $(A, B) = (6, 4)$, $Z = \left[\begin{smallmatrix} 5 & 0 \\ 0 & \beta \end{smallmatrix}\right]$, $\beta = -5$ (top row) and $\beta = 6$ (bottom row). The solutions are based on the μ-problem, with grid sizes $(m, n) = (2^i A, 2^i B)$, $i = 1, 2, 3$, spline degree $d = 3$, and using iterative refinements.

things, on the approximation space, the discretization method, (higer-order) regularity (beyond $W^{2,2}$) and the non-degeneracy of the solutions. In the current setting, the isometry constraint is relaxed or penalized by the small and large parameters ε and μ. What is the description of the Γ-limit if these parameters fall out of the range depicted by Theorems 4.2 and 4.3? It might also be interesting to provide the next order information in terms of Γ-expansion.

As described in Section 5.5, the discrete solutions are arguably asymmetric even if the continuum solutions are symmetric. How genuine is such a phenomenon? Our Γ-convergence result certainly says that such an asymmetry will disappear as $h \to 0$. Can this also be quantified? In a broader context, what is the relationship of discrete symmetries or invariances in connection to those of continuum problems?

As indicated in Section 2.5, we may replace the tensor-product splines by the Kirchhoff elements in [6] to produce yet another family of POPs, in either the constrained or unconstrained form. However, we believe that the resultant method will fail to converge. Note that a lot of expertise is required to make the non-conforming elements to work [6, 8, 9]. The specific discretization method proposed in this paper bypasses those techniques and, as a by-product, gives rise to the POP formulation not found in other related works.

Our functional E, under the isometry constraint, becomes $\mathcal{E}_3$ which is a simple quadratic form of $D^2\mathbf{y}$. In a broader context, there are lots of rooms to explore about the Γ-limits of functionals that depend on second order or curvature information. A well known example is Willmore functional and its discretization – see for example [10]. Solutions can be non-trivial due to the underlying geometric invariance and topological constraints.

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Pure and Applied Mathematics. Academic Press, 2 edition, 2003.

[2] A. A. Ahmadi and A. Majumdar. DSOS and SDSOS optimization: LP and SOCP-based alternatives to sum of squares optimization. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5, March 2014.

[3] S. Alben, B. Balakrisnan, and E. Smela. Edge effects determine the direction of bilayer bending. *Nano Letters*, 11(6):2280–2285, 2011.

[4] S. Bartels. Approximation of large bending isometries with discrete kirchhoff triangles. *SIAM Journal on Numerical Analysis*, 51(1):516–525, 2013.

[5] S. Bartels. Finite element approximation of large bending isometries. *Numerische Mathematik*, 124(3):415–440, 2013.

[6] S. Bartels, A. Bonito, and R. H. Nochetto. Bilayer plates: Model reduction, Γ-convergent finite element approximation, and discrete gradient flow. *Communications on Pure and Applied Mathematics*, 70(3):547–589, 2017.

[7] G. Blekherman, P. A. Parrilo, and R. Thomas. *Semidefinite Optimization and Convex Algebraic Geometry*. Society for Industrial and Applied Mathematics, USA, 2012.

[8] A. Bonito, R. H. Nochetto, and D. Ntogkas. Discontinuous Galerkin approach to large bending deformation of a bilayer plate with isometry constraint. *J. Comput. Phys.*, 423:109785, 2020.

[9] A. Bonito, R. H. Nochetto, and D. Ntogkas. DG approach to large bending plate deformations with isometry constraint. *Mathematical Models and Methods in Applied Sciences*, 31(01):133–175, 2021.

[10] J. Chen, T. P.-Y. Yu, P. Brogan, R. Kusner, Y. Yang, and A. Zigerelli. Numerical methods for biomembranes: conforming subdivision versus non-conforming PL methods. *Mathematics of Computation*, 90(328):471–516, 2021.

[11] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. AMS, 1998.

[12] G. Farin. Triangular Bernstein-Bézier patches. *Computer Aided Geometric Design*, 3:83–127, 1986.

[13] G. Friesecke, R. D. James, and S. Müller. The Föppl–von Kármán plate theory as a low energy γ-limit of nonlinear elasticity. *Comptes Rendus Mathematique*, 335(2):201–206, 2002.

[14] G. Friesecke, R. D. James, and S. Müller. Rigorous derivation of nonlinear plate theory and geometric rigidity. *Comptes Rendus Mathematique*, 334(2):173–178, 2002.

[15] G. Friesecke, R. D. James, and S. Müller. A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 55(11):1461–1506, 2002.

[16] G. Friesecke, R. D. James, and S. Müller. A hierarchy of plate models derived from nonlinear elasticity by gamma-convergence. *Archive for rational mechanics and analysis*, 180(2):183–236, 2006.

[17] M. J. Lai and L. L. Schumaker. *Spline Functions on Triangulations*. Cambridge University Press, 2007.

[18] J. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[19] T. Lyche and L. L. Schumaker. Local spline approximation methods. *J. Approximation Theory*, 15(4):294–325, 1975.

[20] M. J. D. Powell and M. A. Sabin. Piecewise quadratic approximations on triangles. 3:316–325, 1977.

[21] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[22] B. Schmidt. Minimal energy configurations of strained multi-layers. *Calculus of Variations and Partial Differential Equations*, 30(4):477–497, Dec 2007.

[23] B. Schmidt. On the passage from atomic to continuum theory for thin films. *Archive for rational mechanics and analysis*, 190(1):1–55, 2008.

[24] O. G. Schmidt and K. Eberl. Thin solid films roll up into nanotubes. *Nature*, 410:168, Mar 2001.

[25] L. L. Schumaker. *Spline functions: basic theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2007.

[26] L. L. Schumaker and L. Wang. On Hermite interpolation with polynomial splines on T-meshes. *Journal of Computational and Applied Mathematics*, 240:42–50, 2013.

[27] E. Smela, O. Inganäs, and I. Lundström. Controlled folding of micrometer-size structures. *Science*, 268(5218):1735–1738, 1995.

[28] E. Smela, O. Inganäs, Q. Pei, and I. Lundström. Electrochemical muscles: Micromachining fingers and corkscrews. *Advanced Materials*, 5(9):630–632, 1993.

[29] M. Struwe. *Variational methods: Applications to nonlinear partial differential equations and Hamiltonian systems*. Springer-Verlag, Berlin, fourth edition, 2008.

[30] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17(1):218–242, 2006.

[31] H. Waki, S. Kim, M. Muramatsu M. Kojima, H. Sugimoto, and M. Yamashita. https://sparsepop.sourceforge.io/.

[32] A. Yurtsever, J. A. Tropp, O. Fercoq, M. Udell, and V. Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, Jan 2021.