ABSTRACT

Title of dissertation: NON-LINEAR GEOMETRIC PDEs: ALGORITHMS, NUMERICAL ANALYSIS AND COMPUTATION

Dimitrios Ntogkas
Doctor of Philosophy, 2018

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

This dissertation presents the numerical treatment of two geometric non-linear PDEs, the Monge-Ampère Equation and the large bending deformation of plates under an isometry constraint. For the first problem we design a two-scale method and prove rates of convergence in the L^∞ norm, which is an important progress in the numerical analysis of the Monge-Ampère equation and similar equations. For the second problem, we examine the deformation of the mid-plane of the plate and use the fact that it minimizes an energy functional under the isometry constraint. We design a Discontinuous Galerkin method that allows us to construct discrete minimizers of an appropriate energy functional and prove $\Gamma-$convergence to the exact minimizers. We set the theoretical foundation for this method using the problem of a single layer plate and then explore computationally the applicability of the method in more complicated and physically interesting cases using a bilayer model.

NON-LINEAR GEOMETRIC PDEs: ALGORITHMS, NUMERICAL ANALYSIS AND COMPUTATION

by

Dimitrios Ntogkas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Ricardo H. Nochetto, Chair/Advisor
Professor P.S. Krishnaprasad, Dean's Representative
Professor Konstantina Trivisa
Professor Pierre-Emmanuel Jabin
Professor Antoine Mellet

Table of Contents

List of Tables

List of Figures

List of Abbreviations

$\mathcal{E}$	edges of the triangulation
$\mathcal{N}_h$	nodes of triangulation
$\mathcal{N}_h^0$	interior nodes of triangulation
$\mathcal{N}_h^b$	boundary nodes of triangulation
$\mathcal{T}_h$	mesh with mesh-size h
$\mathbb{P}_k$	space of polynomial functions of degree at most k
$\mathbb{Q}_k$	space of polynomial functions of degree at most k in each variable
Σ	skeleton of the triangulation
Ω	domain of the problem
CDG	Continuous-Discontinuous Galerkin
DG	Discontinuous Galerkin
FEM	Finite Element Method
PDE	Partial Differential Equation

Chapter 1: Introduction

There has been an increasing interest in the analysis and the numerical treatment of strongly non-linear partial differential equations (PDEs). This is motivated by the countless applications of such PDEs in Science and Engineering, as well as the mathematical challenges that they offer. In this thesis we are concerned with two very prototypical such problems. The first one is the Monge-Ampère equation, which is a fundamental fully non-linear PDE with applications in optimal transportation, geometry, optics and meteorology. We are interested in this equation because it provides a great starting point for analyzing non-linear PDEs and has a strong geometric flavor, which is a common feature of both problems we tackle. Our work involves the design and implementation of a monotone two-scale method whose discrete solution converges to the viscosity solution of the Monge-Ampère equation under Dirichlet boundary conditions [63]. We also prove rates of convergence in the $L^{\infty}-$norm under various regularity assumptions for the exact solution of the equation, which is an important and timely addition to the literature of numerical analysis for the Monge-Ampère equation [64]. Lastly, we explore a more computational approach that formally increases the accuracy of our two-scale method [62]. We thus modify our discrete operator in order to make it more accurate, but we

lose the property of monotonicity which is critical in achieving convergence. To circumvent this undesirable issue, we employ a filter function that allows us to appropriately combine the original operator with its more accurate counterpart and guarantee convergence of the new method to the viscosity solution. We explore computationally the increased accuracy of this new scheme and examine the role of the filter function.

The second problem of interest is the large bending deformation of plates under a geometrically non-linear isometry constraint. This is motivated by a wide range of applications, including the construction of deformable nano-structures. We first design a numerical method for the simple case of one plate, which we call the single layer, and prove convergence of discrete solutions to exact ones, as well as convergence of our discrete energy functional to the continuous one. Our method improves on a previous approach followed in the literature in aspects that we explore computationally. We then extend this method to account for the physically more interesting case of a bilayer model, i.e. the case where two plates with different mechanical properties are attached together to exploit their different behavior under thermal, electrical or mechanical actuation. The ensuing plate deformations find application in engineering of micro devices (MEMS). We explore this case computationally.

The two problems above have a very strong geometric flavor and are non-linear. Although the mathematical techniques employed in their treatment are not the same, they both require a good physical understanding of the underlying equations and their geometric constraints. Moreover, they are connected through a common application that they share: the Monge-Ampère equation appears as a non-linear

constraint for a special case of what is called the prestrained model, where a plate has no resting configuration and is bound to deform even in the absence of any external forces. This problem is a natural extension of the bilayer model, since, instead of a spontaneous curvature due to gluing plates with different properties, its spontaneous deformation is the result of different properties of the material and it is encoded through the first fundamental form of the deformation of the mid-plane of the plate. The Monge-Ampère equation appears as a constraint when the first fundamental form for the prestrained model is an appropriate bifurcation of the identity matrix [58]. Given our prior work on the Monge-Ampère equation and the specificity of this shared application, we are interested in exploring more general features of the plate models and designing a flexible approach that could be extended to various plate models. This is why we first study the single layer problem and then the bilayer model, as described above. We believe that the techniques developed in this work can be extended to the prestrained model, but we do not pursue this investigation in this thesis.

We now dive a bit more deeply into the background and literature for each of the problems, explain our contribution and provide an outline of the thesis.

1.1 The Monge-Ampère Equation: Chapters 2-3

We consider the Monge-Ampère equation with Dirichlet boundary condition:

$$\begin{cases} \det D^2 u = f & \text{in } \Omega \subset \mathbb{R}^d, \\ u = g & \text{on } \partial\Omega, \end{cases} \tag{1.1.1}$$

where $f \geq 0$ is uniformly continuous, Ω is a uniformly convex domain (not necessarily W^2_∞) and g is a continuous function. We seek a *convex* solution u of (1.1.1), which is critical for (1.1.1) to be elliptic and have a unique viscosity solution [48].

The Monge-Ampère equation has a wide spectrum of applications in optimal mass transport problems [32], geometry [72], nonlinear elasticity [58], optics [36, 47] and meteorology and oceanology [28]. These applications lead to an increasing interest in the investigation of efficient numerical methods. There exist several methods for the Monge-Ampère equation. These include the early work by Oliker and Prussner [70] for space dimension $d = 2$, the vanishing moment methods by Feng and Neilan [39, 40], the penalty method of Brenner, Gudi and Neilan [22], least squares and augmented Lagrangian methods by Dean and Glowinski [29, 30, 46], the finite difference method proposed recently by Benamou, Collino and Mirebeau [13, 61], and a new semi-Lagrangian method by Feng and Jensen [38]. Our work is mostly motivated by the wide-stencil scheme proposed by Froese and Oberman, who proved convergence of the scheme [43]. Awanou [4] proved a linear rate of convergence for classical solutions for the wide-stencil method, when applied to a perturbed Monge-Ampère equation with an extra lower order term δu; the parameter $\delta > 0$ is independent of the mesh and appears in reciprocal form in the rate. In contrast, our analysis hinges on the discrete comparison principle and two discrete barrier functions, which are instrumental in proving convergence to the viscosity solution of (1.1.1). Moreover, our methodology further leads to pointwise error estimates, which we also derive.

Our method hinges on the following formula for the determinant of the semi-

positive Hessian D^2w of a smooth function w [43]:

$$\det D^2w(x) = \min_{(v_1,\dots,v_d)\in\mathbb{S}^\perp} \prod_{j=1}^{d} v_j^T D^2w(x)\ v_j, \tag{1.1.2}$$

where $\mathbb{S}^\perp$ is the set of all $d-$orthonormal bases in $\mathbb{R}^d$. The minimum is achieved for the eigenvectors of D^2w and is equal to $\prod_{j=1}^d \lambda_j$, where $(\lambda_j)_{j=1}^d$ are the corresponding eigenvalues. To discretize (1.1.2) we introduce two scales h and δ. We discretize the domain Ω by a shape regular and quasi-uniform mesh $\mathcal{T}_h$ with spacing h, and construct a space $\mathbb{V}_h$ of continuous piecewise linear functions associated with the mesh $\mathcal{T}_h$. The second coarser scale δ is the length of directions we use to approximate second directional derivatives by central second order differences

$$\nabla^2_\delta w(x;v) := \frac{w(x+\delta v) - 2w(x) + w(x-\delta v)}{\delta^2} \quad \text{and} \quad |v| = 1,$$

for any $w \in C^0(\overline{\Omega})$; this formula will be appropriately modified close to $\partial\Omega$. We denote by u_ε our discrete solution, where $\varepsilon = (h,\delta)$ represents the two scales, and define the discrete Monge-Ampère operator to be

$$T_\varepsilon[u_\varepsilon](x_i) := \min_{(v_1,\dots,v_d)\in\mathbb{S}^\perp} \prod_{j=1}^{d} \nabla^2_\delta u_\varepsilon(x_i;v_j),$$

where x_i is a generic interior node of $\mathcal{T}_h$. This leads to a clear separation of scales, which is a key theoretical advantage over the original wide stencil method of [43]. This also yields continuous dependence of u_ε on data, which along with the discrete comparison principle and the use of some discrete barrier functions give rise to rates of convergence in $L^\infty(\Omega)$ for viscosity solutions of (1.1.1) under some additional regularity requirements. To make the two-scale method practical, we resort to fast

search techniques within [79, 80] to locate points $x_i \pm \delta v_j$, which may not be nodes of $\mathcal{T}_h$ in general.

One of the main tools in this work is the discrete comparison principle that enables us to control the behavior of u_ε and prove its uniform convergence to the unique viscosity solution u of (1.1.1) as $\delta \to 0$ and $h\delta^{-1} \to 0$. It is important to realize, as already observed in [38], that such a convergence is not an immediate consequence of the theory developed by Barles and Souganidis [7]. This theory assumes that the discrete operator is consistent up to the boundary and that the boundary conditions are treated in the viscosity sense; our operator T_ε is only consistent at distance δ from the boundary and our notion of Dirichlet condition is classical. Moreover, the theory of [7] also hinges on a comparison principle for the underlying equation, which in the case of the Monge-Ampère equation (1.1.1) requires that the subsolution and supersolution constructed through the limit supremum and limit infimum of u_ε are convex.

We present two proofs of uniform convergence. The first one relies on regularization of data f, g and Ω and the discrete comparison principle. This approach sets the stage for proving rates of convergence for the 2-scale method. Regularization is a natural device used already for Monge-Ampère by De Philippis and Figalli [31] as a PDE tool and Awanou for numerical purposes [5]. The second approach is along the lines of Barles and Souganidis [7], uses techniques similar to those developed by Feng and Jensen [38], and circumvents the two main issues mentioned above. Controlling the behavior of u_ε in a δ-neighborhood of the boundary $\partial\Omega$ is critical to both approaches. This is achieved via a discrete barrier function; similar constructions

are discussed in [38, 65, 66].

As a next natural step, we prove rates of convergence for the method. We do so for classical solutions with either Hölder or Sobolev regularity and for a special class of viscosity solutions. The first important tool for proving pointwise rates of convergence is the discrete Alexandroff estimate introduced in [65]: if w_h is an arbitrary continuous piecewise linear function, $w_h \geq 0$ on $\partial\Omega_h$, and Γw_h stands for its convex envelope, then

$$\max_{x_i \in \mathcal{N}_h} w_h^-(x_i) \leq C \left(\sum_{x_i \in C_-(w_h)} |\partial \Gamma w_h(x_i)| \right)^{1/d}$$

where Ω_h is our computational domain and $\partial\Gamma w_h$ is the subdifferential of Γw_h, while $C_-(w_h)$ represents the lower contact set of w_h, i.e. the set of interior nodes $x_i \in \mathcal{N}_h^0$ such that $\Gamma w_h(x_i) = w_h(x_i)$; hereafter we write $w_h^-(x_i) := -\min\{w_h(x_i), 0\}$. To control the measure of the subdifferential at each node, we show the following estimate

$$|\partial w_h(x_i)| \leq \delta^d \min_{(v_1,\dots,v_d)\in\mathbb{S}^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j) \quad \forall\, x_i \in \mathcal{N}_h^0,$$

such that the ball centered at x_i and of radius δ is contained in Ω_h. Combining both estimates, we derive the following continuous dependence estimate

$$\max_{\Omega_h} (u_h - w_h)^- \leq C\delta \left(\sum_{x_i \in C_-(u_h - w_h)} \left(T_\varepsilon[u_h](x_i)^{1/d} - T_\varepsilon[w_h](x_i)^{1/d} \right)^d \right)^{1/d}$$

for all continuous piecewise linear functions u_h and w_h such that $T_\varepsilon[u_h](x_i) \geq 0$ and $T_\varepsilon[w_h](x_i) \geq 0$ for all $x_i \in \mathcal{N}_h^0$. This result is instrumental and, combined with operator consistency and a discrete barrier argument close to the boundary,

eventually leads to the following pointwise error estimates

$$\|u_\varepsilon - u\|_{L^\infty(\Omega_h)} \leq C(d, \Omega, f, u)\ h^{\frac{\alpha+k}{\alpha+k+2}}$$

provided $u \in C^{2+k,\alpha}(\overline{\Omega})$ with $0 < \alpha \leq 1$ and $k = 0, 1$, as well as

$$\|u_\varepsilon - u\|_{L^\infty(\Omega_h)} \leq C(d, \Omega, f, u)\ h^{1-\frac{2}{s}}$$

provided $u \in W_p^s(\Omega)$ with $2 + d/p < s \leq 4$ and $p > d$, and δ is suitably chosen in terms of h. We also consider a special case of viscosity solutions with bounded but discontinuous Hessians, and manage to prove a rate of convergence. Since these theorems are proven under the nondegeneracy assumption $f > 0$, we also examine the effect of degeneracy $f \geq 0$ on the rates.

To showcase the performance of our 2-scale method, we present computational experiments for a classical and a degenerate viscosity solution solved with a semi-smooth Newton method. We obtain linear rates for both cases. This indicates a faster convergence than predicted by our theory. However, it is later proven in [59] that the rates are indeed linear for our classical example. We also present an example with unbounded f, which does not fall within our theory, and still observe convergence although with a reduced rate. We observe that, although our method converges linearly, it is not as accurate as the one in [43] in terms of the L^∞ error.

A viable way to ameliorate this is to increase the polynomial degree, which, given the two-scale nature of our scheme, would also require higher order approximation of second directional derivatives. In practice, this pair of corrections corresponds to using quadratic polynomials in order to achieve a better interpolation error and

a more accurate discretization of directional derivatives in order to decrease the truncation error of the operator:

$$\nabla^2_{\delta,a} w(x;v) := \frac{-w(x+\delta v) + 16w(x+\frac{\delta}{2}v) - 30w(x) + 16w(x-\frac{\delta}{2}v) - w(x-\delta v)}{3\delta^2}$$

We observe that this discretization is no longer monotone. As a result a method relying only on this discretization cannot be proven to converge to viscosity solutions of (1.1.1). This is the motivation behind the use of a filtered scheme, along the lines of [44]. The idea of such a scheme hinges on the use of a filter function that combines an accurate and a monotone operator and guarantees that the monotone operator will be used if the accurate operator fails, due to lack of monotonicity. This leads to the notion of "almost monotone" operator that is flexible enough to deliver better accuracy for each fixed mesh size. We note here that, similarly to [44, 69], we require a continuous filter to prove existence and convergence for the scheme, but, in practice, a discontinuous filter is often used [17, 69].

We emphasize that introducing a non-monotone component in our scheme violates a critical tool for our original analysis, the discrete comparison principle, which is instrumental in proving the convergence of the discrete solution to the viscosity solution u of (1.1.1) and deriving max norm error estimates. As a result, one of our first goals is to show that the new operator is "almost" monotone and try to use this fact along with the tools employed so far to obtain again convergence to the viscosity solution under certain assumptions on the filter function.

In order to explore the performance of the accurate operator and the filtered scheme, a substantial part of our presentation is devoted to numerical experiments. We

first verify computationally the increased accuracy of the higher-order operator and present several computational remarks. We then repeat our experiments for the filtered scheme and obtain the anticipated results: the scheme has a smaller error compared to the monotone operator and appears to detect singularities when the solution is not smooth. In fact, we observe that in the case of singularities the filtered scheme performs even better than the accurate operator. We explore this behavior and examine the interplay between the monotone and the accurate operator in the non-smooth case, in order to understand better the role of the chosen filtered function.

The discussion in chapters 2-3 corresponds to the work presented in [62–64].

1.2 Large Bending Deformation of Plates Under Isometry Constraint: Chapters 4-5

Large bending deformation of thin plates is a critical subject for many modern engineering applications due to the extensive use of plate actuators in a variety of systems like thermostats, nano-tubes, micro-robots and micro-capsules [12,54,56,74,76]. From the mathematical viewpoint, there is an increasing interest in the modeling and the numerical treatment of such plates. A rigorous analysis of large bending deformation of plates was conducted in [41], where the geometrically non-linear Kirchhoff model was derived from three dimensional hyperelasticity. Since then, there have been various other interesting models, such as the prestrained model derived in [57].

Previous work on the numerical treatment of large bending deformations includes the single layer problem by Bartels in [8], the bilayer problem by Bartels, Bonito and Nochetto in [11] and the modeling and simulation of thermally actuated bilayer plates by Bartels, Bonito, Muliana and Nochetto in [10]. In all three approaches the model involves minimizing an energy functional that is dominated by the Hessian of the deformation $y : \Omega \to \mathbb{R}^3$ of the mid-plane Ω of the plate. The minimization takes place under Dirichlet boundary conditions for y and ∇y and the isometry constraint

$$\nabla y^T \nabla y = I \text{ a.e.},$$

where I stands for the identity matrix in $\mathbb{R}^2$. The authors use Kirchhoff elements in order to impose the isometry constraint on the nodes of the triangulation and rely on an H^2- gradient flow that allows them to construct solutions of decreasing discrete energy. In [8, 11] it is also proved that the discrete energies Γ-converge to the continuous one. Since for higher order problems a conforming approach would be very costly, the Kirchhoff elements offer a natural non-conforming space for the model that allows the authors to deal with the isometry constraint pointwise.

In Chapter 4 we present a different approach, developed in [19]. We focus on the single layer problem, in order to investigate the applicability of a more flexible method that hinges on a non-conforming space of discontinuous functions and the use of interior penalty terms for the discrete energy, along with a Nitsche approach that enforces the boundary conditions in the limit. To be more precise, given a forcing term $f \in L^2(\Omega)$, we wish to approximate a minimizer $y : \Omega \to \mathbb{R}^3$ of the

continuous energy

$$E[y] = \frac{1}{2}\int_\Omega |D^2 y|^2 - \int_\Omega fy$$

in the set of admissible functions

$$\mathbb{A} = \left\{ y \in H^2(\Omega)^3 \ : \ (\nabla y)^T \nabla y = I \quad \text{a.e. in} \quad \Omega, \quad y = g, \ \nabla y = \Phi_D \quad \text{on} \quad \Gamma_D \right\},$$

where g and Φ_D are the boundary conditions for y and ∇y on a part Γ_D of the boundary. To this end, we define an approximate spae $\mathbb{A}_h^k$ of discontinuous piecewise polynomials of degree $k \geq 2$ that satisfy the approximate isometry constraint

$$\sum_{T \in \mathcal{T}} \left| \int_T (\nabla_h y_h)^T \nabla_h y_h - I \right| \leq Ch, \tag{1.2.1}$$

where $\nabla_h y_h$ stands for the piecewise (or broken) gradient of y_h. We then search for $y_h \in \mathbb{A}_h^k$ that minimizes the discrete energy functional

$$\begin{aligned} E_h[y_h] := \frac{1}{2}\|D_h^2 y_h\|^2_{L^2(\Omega)} - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(\Sigma)} \\ + \frac{\gamma_1}{2}\|h^{-1/2}[\nabla_h y_h]\|^2_{L^2(\Sigma)} + \frac{\gamma_0}{2}\|h^{-3/2}[y_h]\|^2_{L^2(\Sigma)} - \int_\Omega f_h y_h, \end{aligned}$$

where we use again piecewise differentiation, f_h is the Lagrange interpolant of f in a space of polynomials of degree k and Σ stands for the skeleton of the mesh. The above discrete energy is natural in the context of Discontinuous Galerkin (DG) methods, since it involves interior penalty terms and Nitsche terms that allow us to deal with the boundary conditions. This means that the jumps that appear on the energy are defined as

$$[y_h] := \begin{cases} y_h^+ - y_h^-, & e \subset \Omega \\ y_h - g_h, & e \subset \partial\Omega \end{cases}$$

and

$$[\nabla_h y_h] := \begin{cases} \nabla_h y_h^+ - \nabla_h y_h^-, & e \subset \Omega \\ \\ \nabla_h y_h - \Phi_{D,h}, & e \subset \partial\Omega, \end{cases}$$

where g_h and $\Phi_{D,h}$ correspond to discontinuous interpolants of the boundary conditions and $y_h^{\pm}$ are the values of y_h on both sides of the edge e. In order to find such a minimizer, we employ a discrete gradient flow using an appropriate inner product. We show that the gradient flow leads to decreasing discrete energy and guarantees the constraint in (1.2.1). We also prove that the discrete energy E_h Γ-converges to the continuous energy E and that minimizers of E_h converge strongly in $L^2(\Omega)$ to minimizers of E in $\mathbb{A}$.

Our approach is motivated by the flexibility of DG compared to Kirchhoff elements. First of all, Kirchhoff elements require polynomial degree $k = 3$ and suffer from a complicated implementation that involves the introduction and realization of a discrete gradient that maps the gradient of the discrete deformation to another space. Since this is not implemented in finite element libraries, the above difficulty hinders the impact of the method in the engineering community. A DG approach circumvents this issue, since the DG framework is implemented in most finite element libraries. Moreover, Kirchhoff elements are not amenable to adaptively refined partitions, while the DG approach can be potentially extended to graded meshes. Lastly, it was observed numerically that the isometry constraint imposed with Kirchhoff elements at each vertex is very rigid at the expense of efficiency in the approximation. On the contrary, our experiments with DG indicate that this approach is more accurate and that it also adjusts better to the geometry of the

problem.

From a theoretical stand-point, in order to prove $\Gamma-$ convergence of our method, we need to create a recovery sequence for a function $y \in H^2(\Omega)$. To this end, we also need a regularization of y, in order to use interpolation estimates. Our approach allows us to do this without being forced to preserve the isometry constraint or the boundary conditions after regularization, since we do not use pointwise values for the constraint nor do we enforce the boundary conditions explicitly in our discrete space. This offers a simpler argument than the one required for Kirchhoff elements in [8, 11].

At the end of the chapter, we present numerical simulations for the single layer problem that allow us to compare with the Kirchhoff element approach and assess the flexibility of this new method.

Finally, we explore computationally in Chapter 5 the applicability of our method into more physically realistic cases. To that end, we test our method in a variety of examples for the bilayer problem, that are either inspired by the work in [10, 11] or by engineering applications [2, 53, 75].

The discussion in chapters 4-5 mostly corresponds to the work presented in [19].

1.3 Outline

Chapters 2 and 3 are concerned with the numerical treatment of the Monge-Ampère equation. We present our two-scale method on Chapter 2 and prove the convergence to viscosity solutions, as well as rates of convergence for varying regularity assumptions for the exact solution. We close the chapter by presenting our numerical experiments for the method. On Chapter 3 we motivate and introduce the filtered scheme that improves the accuracy of the two-scale method and prove again convergence to the viscosity solutions. We illustrate the performance of the method and compare with the original two-scale scheme. Chapters 4 and 5 are devoted to the large bending deformation problem with isometry constraint. On Chapter 4 we introduce our DG method for the single layer problem. We prove Γ-convergence of the discrete energy to the continuous one and illustrate the accuracy and geometric flexibility of the method numerically. Lastly, on Chapter 5 we extend the DG method to the bilayer model and investigate computationally how it performs for various examples either from [10, 11] or the engineering literature.

Chapter 2: Two-scale method for the Monge-Ampère Equation: Convergence to the viscosity solution

In this Chapter we introduce our two-scale method for solving the Monge-Ampère equation with Dirichlet boundary conditions. We prove the convergence of our solution to the unique viscosity solution of (1.1.1) and provide rates of convergence for solutions with certain regularity assumptions. We explore these results computationally using a classical and a piecewise smooth viscosity solution, as well as a viscosity solution that is not covered by our analysis.

2.1 Two-Scale Method

We start by introducing our method based on the variational formulation (1.1.2) for the determinant of the Hessian D^2u. We then prove a discrete analogous of the convexity property for solutions of our method. Afterwards, we introduce the full discretization of our method, which uses only a finite subset of the orthonormal basis used in (1.1.2).

2.1.1 Ideal Two-Scale Method

Let $\mathcal{T}_h$ be a shape-regular and quasi-uniform triangulation with mesh size h. We denote by Ω_h the union of elements of $\mathcal{T}_h$ and we call it the computational domain. Let $\mathcal{N}_h$ denote the nodes of $\mathcal{T}_h$, $\mathcal{N}_h^b := \{x_i \in \mathcal{N}_h : x_i \in \partial\Omega_h\}$ be the boundary nodes and $\mathcal{N}_h^0 := \mathcal{N}_h \backslash \mathcal{N}_h^b$ be the interior nodes. We require that $\mathcal{N}_h^b \subset \partial\Omega$, which in view of the convexity of Ω implies that Ω_h is also convex and $\Omega_h \subset \Omega$. We denote by $\mathbb{V}_h$ the space of continuous piecewise linear functions over $\mathcal{T}_h$. We recall the notation $\mathbb{S}^\perp$ for the collection of all d-tuples of orthonormal bases and $\mathbf{v} := (v_1, \ldots, v_d) \in \mathbb{S}^\perp$ for a generic element, whence each component v_i belongs to the unit sphere $\mathbb{S}$ of $\mathbb{R}^d$. For $x_i \in \mathcal{N}_h^0$, we use the formula of centered second differences

$$\nabla_\delta^2 w(x_i; v_j) := \frac{w(x_i + \rho\delta v_j) - 2w(x_i) + w(x_i - \rho\delta v_j)}{\rho^2\delta^2}, \tag{2.1.1}$$

where $0 < \rho \leq 1$ is the biggest number such that both $x_i \pm \rho\delta v_j \in \overline{\Omega}_h$ for all $v_j \in \mathbb{S}$; we stress that ρ need not be computed exactly. This is well defined for any $w \in C^0(\overline{\Omega})$, in particular for $w \in \mathbb{V}_h$.

We seek $u_\varepsilon \in \mathbb{V}_h$ such that $u^\varepsilon(x_i) = g(x_i)$ for $x_i \in \mathcal{N}_h^b$ and for $x_i \in \mathcal{N}_h^0$

$$T_\varepsilon[u_\varepsilon](x_i) := \min_{\mathbf{v}\in\mathbb{S}^\perp} \left(\prod_{j=1}^{d} \nabla_\delta^{2,+} u_\varepsilon(x_i; v_j) - \sum_{j=1}^{d} \nabla_\delta^{2,-} u_\varepsilon(x_i; v_j) \right) = f(x_i), \tag{2.1.2}$$

where from now on we use the notation

$$\nabla_\delta^{2,+} u_\varepsilon(x_i; v_j) = \max\left(\nabla_\delta^2 u_\varepsilon(x_i; v_j), 0\right), \quad \nabla_\delta^{2,-} u_\varepsilon(x_i; v_j) = -\min\left(\nabla_\delta^2 u_\varepsilon(x_i; v_j), 0\right).$$

A similar definition was first proposed by Froese and Oberman in [42, 43] for a finite

difference method. The key idea behind (2.1.2) is to enforce a suitable notion of discrete convexity. To build intuition we explore this concept next.

Definition 2.1.1 (discrete convexity)**.** *We say that $w_h \in \mathbb{V}_h$ is discretely convex if*

$$\nabla_\delta^2 w_h(x_i; v_j) \geq 0 \qquad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}.$$

It is important to realize that this definition does not imply convexity in the usual sense, which is rather tricky to achieve with piecewise polynomials [1, 66, 78]. On the other hand, if $w \in C^0(\overline{\Omega}_h)$ is convex, then its Lagrange interpolant $\mathcal{I}_h w$ satisfies $\mathcal{I}_h w \geq w$, whence $\mathcal{I}_h w$ is discretely convex but not necessarily convex.

Lemma 2.1.2 (discrete convexity)**.** *If $w_h \in \mathbb{V}_h$ satisfies*

$$T_\varepsilon[w_h](x_i) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0,$$

then w_h is discretely convex and as a consequence

$$T_\varepsilon[w_h](x_i) = \min_{\mathbf{v} \in \mathbb{S}^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j), \tag{2.1.3}$$

namely

$$\nabla_\delta^{2,+} w_h(x_i; v_j) = \nabla_\delta^2 w_h(x_i; v_j), \quad \nabla_\delta^{2,-} w_h(x_i; v_j) = 0 \quad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}.$$

Conversely, if w_h is discretely convex, then $T_\varepsilon[w_h](x_i) \geq 0$ for all $x_i \in \mathcal{N}_h^0$.

Proof. We distinguish two cases depending on whether $T_\varepsilon[w_h](x_i) > 0$ or not. Let $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}^\perp$ be a d-tuple that realizes the mininum in the definition of $T_\varepsilon[w_h](x_i)$ and note that

$$\prod_{j=1}^{d} \nabla_\delta^{2,+} w_h(x_i; v_j) \geq 0, \quad \sum_{j=1}^{d} \nabla_\delta^{2,-} w_h(x_i; v_j) \geq 0.$$

Case 1: $T_\varepsilon[w_h](x_i) > 0$. If the difference of these two quantities is positive, then so must be the first one. This implies that each factor $\nabla_\delta^{2,+} w_h(x_i; v_j) > 0$, whence the second term must vanish. This readily yields (2.1.3).

Case 2: $T_\varepsilon[w_h](x_i) = 0$. If instead the difference of the two quantities above vanishes, then there are two possible situations. If the first quantity is strictly positive, then the argument in Case 1 implies that the second quantity vanishes, which is a contradiction. Therefore, the alternative option is that both quantities vanish, whence

$$\nabla_\delta^{2,-} w_h(x_i; v_j) = 0 \quad \forall j \quad \Rightarrow \quad \nabla_\delta^2 w_h(x_i; v_j) \geq 0 \quad \forall j.$$

This again implies that w_h is discretely convex along with (2.1.3). Since the converse is trivial the proof is complete. □

2.1.2 Practical Two-Scale Method

The ideal two-scale method of Section 2.1.1 leads to the notion of discrete convexity and Lemma 2.1.2 but cannot be implemented, because the minimum in (2.1.2) entails infinitely many options for $\mathbf{v} \in \mathbb{S}^\perp$. To render the two-scale method practical, we introduce a finite discretization $\mathbb{S}_\theta$ of the unit sphere $\mathbb{S}$ governed by the parameter θ: given $v \in \mathbb{S}$, there exists $v^\theta \in \mathbb{S}_\theta$ such that

$$|v - v^\theta| \leq \theta.$$

Likewise, we approximate the set $\mathbb{S}^\perp$ of d-orthonormal bases by the finite set $\mathbb{S}_\theta^\perp \subset \mathbb{S}^\perp$: for any $\mathbf{v}^\theta = (v_j^\theta)_{j=1}^d \in \mathbb{S}_\theta^\perp$, $v_j^\theta \in \mathbb{S}_\theta$ and there exists $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}^\perp$ such that $|v_j - v_j^\theta| \leq \theta$ for all $1 \leq j \leq d$ and conversely.

If $\varepsilon := (h, \delta, \theta)$, the practical two-scale method now reads: seek $u_\varepsilon \in \mathbb{V}_h$ such that $u_\varepsilon(x_i) = g(x_i)$ for $x_i \in \mathcal{N}_h^b$ and for $x_i \in \mathcal{N}_h^0$

$$T_\varepsilon[u_\varepsilon](x_i) := \min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \left(\prod_{j=1}^{d} \nabla_\delta^{2,+} u_\varepsilon(x_i; v_j) - \sum_{j=1}^{d} \nabla_\delta^{2,-} u_\varepsilon(x_i; v_j) \right) = f(x_i). \tag{2.1.4}$$

We observe that if we relax Definition 2.1.1 (discrete convexity) to be

$$\nabla_\delta^2 w_h(x_i; v_j) \geq 0 \qquad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}_\theta,$$

then Lemma 2.1.2 (discrete convexity) is still valid and we can take

$$T_\varepsilon[w_h](x_i) = \min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j), \tag{2.1.5}$$

provided $T_\varepsilon[w_h](x_i) \geq 0$, as is the case of u_ε in (2.1.4).

We now show that (2.1.4) is monotone and prove a comparison principle for $f \geq 0$.

Lemma 2.1.3 (monotonicity)**.** *Let $u_h, w_h \in \mathbb{V}_h$ be discretely convex. If $u_h - w_h$ attains a maximum at an interior node $z \in \mathcal{N}_h^0$, then*

$$T_\varepsilon[w_h](z) \geq T_\varepsilon[u_h](z).$$

Proof. If $u_h - w_h$ attains a maximum at $z \in \mathcal{N}_h^0$, then

$$u_h(z) - w_h(z) \geq u_h(x_i) - w_h(x_i) \quad \forall x_i \in \mathcal{N}_h.$$

For suitably chosen $0 < \rho \leq 1$, the points $z \pm \delta\rho v_j \in \overline{\Omega}_h$ satisfy

$$u_h(z) - w_h(z) \geq u_h(z \pm \delta\rho v_j) - w_h(z \pm \delta\rho v_j) \quad \forall v_j \in \mathbb{S}_\theta,$$

because this relation holds at the vertices of the simplices where $z \pm \delta\rho v_j$ belong to and both u_h and w_h are piecewise linear. Hence, (2.1.1) implies

$$\nabla_\delta^2 u_h(z; v_j) \leq \nabla_\delta^2 w_h(z; v_j) \quad \forall v_j \in \mathbb{S}_\theta.$$

Since discrete convexity of u_h and w_h implies (2.1.5), the proof is complete. □

Lemma 2.1.4 (discrete comparison principle)**.** *Let $u_h, w_h \in \mathbb{V}_h$ with $u_h \leq w_h$ on the boundary $\partial\Omega_h$ be such that*

$$T_\varepsilon[u_h](x_i) \geq T_\varepsilon[w_h](x_i) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0.$$

Then, $u_h \leq w_h$ in Ω_h.

Proof. Since $u_h, w_h \in \mathbb{V}_h$, it suffices to prove $u_h(x_i) \leq w_h(x_i)$ for all $x_i \in \mathcal{N}_h^0$. In view of Lemma 2.1.2 (discrete convexity) and (2.1.5), we realize that both u_h and w_h are discretely convex and we can rewrite the operator inequality as follows:

$$\min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 u_h(x_i; v_j) \geq \min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0.$$

The proof splits into two steps according to whether this inequality is strict or not.

Step 1. We first consider the strict inequality

$$\min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 u_h(x_i; v_j) > \min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j) \quad \forall x_i \in \mathcal{N}_h^0.$$

We assume by contradiction that there exists an interior node $x_k \in \mathcal{N}_h^0$ such that

$$u_h(x_k) - w_h(x_k) > 0$$

and

$$u_h(x_k) - w_h(x_k) \geq u_h(x_i) - w_h(x_i) \quad \forall x_i \in \mathcal{N}_h.$$

Reasoning as in Lemma 2.1.3 we obtain $\nabla_\delta^2 u_h(x_k; v_j) \leq \nabla_\delta^2 w_h(x_k; v_j)$ for all $v_j \in \mathbb{S}_\theta$. On the other hand, the original strict inequality at $x_i = x_k$ yields

$$\min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_k; v_j) < \prod_{j=1}^{d} \nabla_\delta^2 u_h(x_k; \overline{v}_j)$$

for all possible directions $\overline{\mathbf{v}} = (\overline{v}_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$. Choosing $\overline{\mathbf{v}}$ to be a d-tuple that realizes the minimum of the left-hand side leads to a contradiction.

Step 2. We now deal with the non-strict inequality. We introduce the quadratic strictly convex function $q(x) = \frac{1}{2}\big(|x|^2 - R^2\big)$, which satisfies $q \leq 0$ on $\overline{\Omega}$ for $R > 0$ sufficiently large and in particular $q \leq 0$ on $\partial\Omega_h$. Its Lagrange interpolant $q_h = \mathcal{I}_h q$ is discretely convex and

$$\nabla_\delta^2 q_h(x_i; v_j) \geq \nabla_\delta^2 q(x_i; v_j) = \partial_{v_j v_j}^2 q(x_i) = 1 \quad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}_\theta.$$

For arbitrary $\alpha > 0$, the function $u_h + \alpha q_h$ satisfies $u_h + \alpha q_h \leq u_h \leq w_h$ on $\partial\Omega_h$ and

$$\nabla_\delta^2 (u_h + \alpha q_h)(x_i; v_j) \geq \nabla_\delta^2 u_h(x_i; v_j) + \alpha > \nabla_\delta^2 w_h(x_i; v_j) \quad \forall x_i \in \mathcal{N}_h^0, \quad \forall v_j \in \mathbb{S}_\theta,$$

whence $T_\varepsilon[u_h + \alpha q_h](x_k) > T_\varepsilon[w_h](x_k)$. Applying Step 1 we deduce

$$u_h + \alpha q_h \leq w_h \quad \forall \alpha > 0.$$

Taking the limit as $\alpha \to 0$ gives the asserted inequality. □

2.2 Existence and Uniqueness

We now prove existence and uniqueness of a discrete solution $u_\varepsilon \in \mathbb{V}_h$ of (2.1.4).

Lemma 2.2.1 (existence, uniqueness and stability)**.** *There exists a unique $u_\varepsilon \in \mathbb{V}_h$ that solves the discrete Monge-Ampère equation (2.1.4). The solution u_ε is stable in the sense that $\|u_\varepsilon\|_{L^\infty(\Omega)}$ does not depend on the parameters $\epsilon = (h, \delta, \theta)$ of the method.*

Proof. Since uniqueness is a trivial consequence of Lemma 2.1.4 (discrete comparison principle), we just have to prove existence. To this end, we construct a monotone sequence of discrete convex functions $\left\{u_h^k\right\}_{k=0}^{\infty}$, starting with the initial iterate $u_h^0 \in \mathbb{V}_h$ that satisfies $u_h^0 = \mathcal{I}_h g$ on $\partial\Omega_h$ and

$$T_\varepsilon[u_h^0](x_i) \geq f(x_i) \quad \forall x_i \in \mathcal{N}_h^0.$$

Step 1 : Existence of u_h^0. We repeat the calculations of Step 2 in Lemma 2.1.4 (discrete comparison principle) for

$$q(x) = \frac{1}{2}\|f\|_{L^\infty(\Omega)}^{1/d}|x|^2$$

to obtain that for $q_h = \mathcal{I}_h q$ and for all $x_i \in \mathcal{N}_h^0$

$$T_\varepsilon[q_h](x_i) \geq \|f\|_{L^\infty(\Omega)} \geq f(x_i).$$

We utilize the stability of $\mathcal{I}_h q$ in $L^\infty(\Omega_h)$ to deduce

$$\|q_h\|_{L^\infty(\Omega_h)} \leq C_R \, \|f\|_{L^\infty(\Omega)}^{1/d},$$

where C_R is a geometric constant that depends on the domain Ω.

We next observe that the set of convex functions w satisfying a continuous Dirichlet boundary condition on a uniformly convex domain is non-empty. The solution $w \in C^0(\overline{\Omega})$ of the homogeneous Dirichlet problem (1.1.1) is one such function [48, Theorem 1.5.2]. Let w be convex and solve (1.1.1) with $f = 0$ and Dirichlet condition $w = g - q$, whence $w_h := \mathcal{I}_h w$ satisfies

$$T_\varepsilon[w_h](x_i) \geq 0 \quad \forall x_i \in \mathcal{N}_h^0.$$

and $w_h = \mathcal{I}_h g - q_h$ on $\mathcal{N}_h^b$. We define the initial iterate to be

$$u_h^0 := w_h + q_h$$

and note that u_h^0 is discretely convex and satisfies the Dirichlet condition $u_h^0 = \mathcal{I}_h g$ on $\partial\Omega_h$. Since all the terms in $T_\varepsilon[u_h^0](x_i)$ are non-negative, we also deduce

$$T_\varepsilon[u_h^0](x_i) = \min_{\mathbf{v}\in\mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \Big(\nabla_\delta^2 w_h(x_i; v_j) + \nabla_\delta^2 q_h(x_i; v_j) \Big) \geq f(x_i) \quad \forall x_i \in \mathcal{N}_h^0.$$

Step 2 : Perron Construction. We proceed by induction. Suppose that we have already a discretely convex function $u_h^k \in \mathbb{V}_h$ that satisfies $u_h^k = \mathcal{I}_h g$ on $\partial\Omega_h$ and

$$T_\varepsilon[u_h^k](x_i) \geq f(x_i) \quad \forall x_i \in \mathcal{N}_h^0. \tag{2.2.1}$$

We now construct $u_h^{k+1} \in \mathbb{V}_h$ such that $u_h^{k+1} \geq u_h^k$ in Ω_h, $u_h^{k+1} = \mathcal{I}_h g$ on $\partial\Omega_h$ and satisfies (2.2.1). We consider all interior nodes in order and construct auxiliary functions $u_h^{k,i-1} \in \mathbb{V}_h$ using the first $i-1$ nodes and starting from $u_h^{k,0} := u_h^k$ as follows. At $x_i \in \mathcal{N}_h^0$ we check whether or not $T^\varepsilon[u_h^{k,i-1}](x_i) > f(x_i)$. If so, we increase the value of $u_h^{k,i-1}(x_i)$ and denote the resulting function by $u_h^{k,i}$, until

$$T_\varepsilon[u_h^{k,i}](x_i) = f(x_i).$$

This is possible because the centered second differences (2.1.2) are strictly decreasing with increasing central value for all directions. Expression (2.1.2) also shows that this process potentially increases the centered second differences at other nodes $x_j \neq x_i$, whence

$$T_\varepsilon[u_h^{k,i}](x_j) \geq T_\varepsilon[u_h^{k,i-1}](x_j) \geq f(x_j) \quad \forall x_j \neq x_i.$$

We repeat this process with the remaining nodes x_j for $i < j \leq N$, and set $u_h^{k+1} := u_h^{k,N}$ to be the last intermediate function. By construction, we clearly obtain

$$T_\varepsilon[u_h^{k+1}](x_i) \geq f(x_i), \quad u_h^{k+1}(x_i) \geq u_h^k(x_i) \quad \forall x_i \in \mathcal{N}_h^0.$$

Our construction preserves the boundary values $u_h^{k+1} = \mathcal{I}_h g$ on $\partial\Omega_h$ and enforces the relation $u_h^{k+1} \geq u_h^k$ in Ω_h because both u_h^{k+1}, u_h^k are piecewise linear functions.

Step 3 : Bounds. If $b_h := \max_{x_i \in \mathcal{N}_h^b} g(x_i)$, then we see that $b_h \in \mathbb{V}_h$ and

$$T_\varepsilon[b_h](x_i) = 0 \leq f(x_i) \leq T_\varepsilon[u_h^k](x_i) \quad \forall x_i \in \mathcal{N}_h^0, \quad \forall k \geq 0.$$

We apply Lemma 2.1.4 (discrete comparison principle) to infer that $u_h^k \leq b_h$ for all $k \geq 0$. On the other hand, since $\|u_h^0\|_{L^\infty(\Omega_h)}$ is bounded uniformly in h and $u_h^0 \leq u_h^k$, we deduce the uniform bound

$$\|u_h^k\|_{L^\infty(\Omega)} \leq \Lambda$$

with $\Lambda > 0$ independent of the discretization parameters h, δ and θ.

Step 4 : Convergence. The sequence $\{u_h^k(x_i)\}_{k=1}^\infty$ is monotone and bounded above for all $x_i \in \mathcal{N}_h^0$, and hence converges. The limit

$$u_\varepsilon(x_i) = \lim_{k\to\infty} u_h^k(x_i) \quad \forall x_i \in \mathcal{N}_h^0$$

defines $u_\varepsilon \in \mathbb{V}_h$ and satisfies $u_\varepsilon = \mathcal{I}_h g$ on $\partial\Omega_h$. It also satisfies the desired equality

$$T_\varepsilon[u_\varepsilon](x_i) = f(x_i) \quad \forall x_i \in \mathcal{N}_h^0,$$

since $T_\varepsilon[u_\varepsilon](x_i) = \lim_{k\to\infty} T_\varepsilon[u_h^k](x_i) \geq f(x_i)$ and if the last inequality were strict, then Step 2 could be applied to improve u_ε. This shows existence of a discrete solution u_ε of (2.1.2) as well as the uniform bound $\|u_\varepsilon\|_{L^\infty(\Omega)} \leq \Lambda$. □

2.3 Consistency

We now quantify the operator consistency error in terms of Hölder regularity of D^2u. We start with the definitions of δ-interior region

$$\Omega_{h,\delta} = \{T \in \mathcal{T}_h \ : \ \text{dist}(x, \partial\Omega_h) \geq \delta \ \forall x \in T\}, \tag{2.3.1}$$

and δ-boundary region

$$\omega_{h,\delta} = \Omega_h \setminus \Omega_{h,\delta}.$$

Moreover, given a node $x_i \in \mathcal{N}_h^0$ we denote by

$$B_i := \cup\{\overline{T} : T \in \mathcal{T}_h, \text{ dist } (x_i, T) \leq \hat{\delta}\} \tag{2.3.2}$$

where $\hat{\delta} := \rho\delta$ with $0 < \rho \leq 1$ is the biggest number so that $x_i \pm \hat{\delta}v_j \in \overline{\Omega}_h$ for all $v_j \in \mathbb{S}_\theta$.

Lemma 2.3.1 (consistency of $\nabla_\delta^2 \mathcal{I}_h u$)**.** *Let $u \in W_\infty^2(B_i)$, $\mathcal{I}_h u$ be its Lagrange interpolant in Ω_h, and B_i be defined in (2.3.2). The following two estimates are then valid:*

(i) For all $x_i \in \mathcal{N}_h^0$ and all $v_j \in \mathbb{S}_\theta$, we have

$$\left|\nabla_\delta^2 \mathcal{I}_h u(x_i; v_j)\right| \leq C|u|_{W_\infty^2(B_i)}.$$

(ii) If in addition $u \in C^{2+k,\alpha}(B_i)$ for $k = 0, 1$ and $\alpha \in (0, 1]$, then for all $x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta}$ and all $v_j \in \mathbb{S}_\theta$, we have

$$\left|\nabla_\delta^2 \mathcal{I}_h u(x_i; v_j) - \frac{\partial^2 u}{\partial v_j^2}(x_i)\right| \leq C\left(|u|_{C^{2+k,\alpha}(B_i)}\delta^{k+\alpha} + |u|_{W_\infty^2(B_i)}\frac{h^2}{\delta^2}\right).$$

In both cases C stands for a constant independent of the two scales h and δ, the parameter θ and u.

Proof. We split the proof into three steps.

Step 1. Let $x_i \in \mathcal{N}_h^0$ and $v_j \in \mathbb{S}_\theta$. Since

$$u(x_i + \hat{\delta} v_j) - u(x_i) = \hat{\delta} \int_0^1 \nabla u(x_i + t\hat{\delta} v_j) \cdot v_j \, dt,$$

definition (2.1.1) yields

$$\nabla_\delta^2 u(x_i; v_j) = \hat{\delta}^{-1} \int_0^1 \Big(\nabla u(x_i + t\hat{\delta} v_j) - \nabla u(x_i - t\hat{\delta} v_j) \Big) \cdot v_j \, dt$$

Adding and subtracting $\nabla u(x_i) \cdot v_j$ inside the integral, we similarly arrive at

$$\nabla_\delta^2 u(x_i; v_j) = \int_0^1 \int_0^1 t \ \Big(D^2 u(x_i + st\hat{\delta} v_j) + D^2 u(x_i - st\hat{\delta} v_j) \Big) : v_j \otimes v_j \, ds \, dt,$$

which implies

$$\left| \nabla_\delta^2 u(x_i; v_j) \right| \leq |u|_{W_\infty^2(B_i)}.$$

Step 2. Let $x_i \in \Omega_{h,\delta}$ and assume that $u \in C^{2,\alpha}(B_i)$. We prove the estimate

$$\left| \nabla_\delta^2 u(x_i; v_j) - \frac{\partial^2 u}{\partial v_j^2}(x_i) \right| \leq C |u|_{C^{2,\alpha}(B_i)} \ \delta^\alpha.$$

Write $\nabla_\delta^2 u(x_i; v_j) = I_1 + I_2$, where

$$I_1 = 2 \int_0^1 \int_0^1 t D^2 u(x_i) : v_j \otimes v_j \, ds \, dt = \frac{\partial^2 u}{\partial v_j^2}(x_i)$$

and

$$I_2 = \int_0^1 \int_0^1 t \left(D^2 u(x_i + st \ \delta v_j) - 2 D^2 u(x_i) + D^2 u(x_i - st \ \delta v_j) \right) : v_j \otimes v_j \, ds \, dt.$$

The fact that $u \in C^{2,\alpha}(B_i)$ gives

$$|D^2u(x_i \pm st\delta v_j) - D^2u(x_i)| \leq C|u|_{C^{2,\alpha}(B_i)}\delta^\alpha,$$

whence

$$I_2 \leq C|u|_{C^{2,\alpha}(B_i)}\delta^\alpha.$$

Combining I_1 and I_2, we deduce the asserted estimate for $u \in C^{2,\alpha}(B_i)$ and $k = 0$. For $u \in C^{3,\alpha}(B_i)$, we exploit the symmetry of I_2 to express the integrand in terms of differences of D^3u at points $x_i \pm stz\,\delta v_j$ for $0 < z < 1$ and thus deduce

$$I_2 \leq C|u|_{C^{3,\alpha}(B_i)}\delta^{1+\alpha}.$$

This implies the estimate for $k = 1$

$$\left|\nabla^2_\delta u(x_i; v_j) - \frac{\partial^2 u}{\partial v_j^2}(x_i)\right| \leq C|u|_{C^{3,\alpha}(B_i)}\delta^{1+\alpha}.$$

Step 3. We now study the effect of interpolation, for which it is known that [21]

$$\|u - \mathcal{I}_h u\|_\infty \leq C\ |u|_{W^2_\infty(B_i)}h^2.$$

Therefore, applying definition (2.1.1), we deduce for $x_i \in \Omega_{h,\delta}$

$$|\nabla^2_\delta(u - \mathcal{I}_h u)(x_i; v_j)| \leq C|u|_{W^2_\infty(B_i)}\frac{h^2}{\delta^2}.$$

This completes the proof of (ii) for $k = 0, 1$. Otherwise, δ must be replaced by $\hat{\delta} = \rho\delta \geq Ch$ with $C > 0$ depending only on shape regularity. Therefore, we see that $h^2\hat{\delta}^{-2} \leq C$, which combined with Step 1 yields the estimate in (i). □

We now extend the consistency analysis to the practical two-scale operator T_ϵ.

Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$)**.** *Let $x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta}$ and B_i be defined as in (2.3.2). If $u \in C^{2+k,\alpha}(B_i)$ is convex with $0 < \alpha \leq 1$ and $k = 0, 1$, and $\mathcal{I}_h u$ is its piecewise linear interpolant, then*

$$\left|\det D^2 u(x_i) - T_\varepsilon[\mathcal{I}_h u](x_i)\right| \leq C_1(d, \Omega, u)\delta^{k+\alpha} + C_2(d, \Omega, u)\left(\frac{h^2}{\delta^2} + \theta^2\right), \quad (2.3.3)$$

where

$$C_1(d, \Omega, u) = C|u|_{C^{2+k,\alpha}(B_i)}|u|^{d-1}_{W^2_\infty(B_i)}, \quad C_2(d, \Omega, u) = C|u|^d_{W^2_\infty(B_i)}.$$

If $x_i \in \mathcal{N}_h^0$ and $u \in W^2_\infty(B_i)$, then (2.3.3) remains valid with $\alpha = k = 0$ and $C^{2+k,\alpha}(B_i)$ replaced by $W^2_\infty(B_i)$.

Proof. We recall that $\mathcal{I}_h u$ is discretely convex, namely $\nabla^2_\delta \mathcal{I}_h u(x_i, v_j) \geq 0$ for all $x_i \in \mathcal{N}_h^0$ and $v_j \in \mathbb{S}_\theta$, because u is convex. Therefore, in view of Lemma 2.1.2 (discrete convexity), the definition of $T_\varepsilon[\mathcal{I}_h u]$ reduces to

$$T_\varepsilon[\mathcal{I}_h u](x_i) = \min_{\mathbf{v} \in \mathbb{S}_\theta^\perp} \prod_{j=1}^d \nabla^2_\delta \mathcal{I}_h u(x_i; v_j).$$

Step 1. Let $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$ be the d-tuple that realizes the above minimum. Applying (1.1.2) to the determinant of the Hessian of u, we see that

$$\det D^2 u(x_i) - T_\varepsilon[\mathcal{I}_h u](x_i) \leq \prod_{j=1}^d \frac{\partial^2 u}{\partial v_j^2}(x_i) - \prod_{j=1}^d \nabla^2_\delta \mathcal{I}_h u(x_i; v_j).$$

We now invoke Lemma 2.3.1 (ii) (consistency of $\nabla^2_\delta \mathcal{I}_h u$) to write

$$\left|\frac{\partial^2 u}{\partial v_j^2}(x_i) - \nabla^2_\delta \mathcal{I}_h u(x_i; v_j)\right| \leq C|u|_{C^{2+k,\alpha}(B_i)}\delta^{k+\alpha} + C|u|_{W^2_\infty(B_i)}\frac{h^2}{\delta^2},$$

where $k = 0, 1$. Given the multiplicative structure above, utilizing Lemma 2.3.1 (i) we deduce

$$\det D^2 u(x_i) - T_\varepsilon[\mathcal{I}_h u](x_i) \leq C_1(d, \Omega, u)\delta^{k+\alpha} + C_2(d, \Omega, u)\frac{h^2}{\delta^2},$$

where C_1 and C_2 are defined above.

Step 2. We now choose $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}^\perp$ to be the d-tuple that realizes the minimum in (1.1.2) for $\det D^2u(x_i)$. We can then write

$$T_\varepsilon[\mathcal{I}_h u](x_i) - \det D^2u(x_i) \le I_1 + I_2$$

where

$$I_1 = \prod_{j=1}^d \nabla_\delta^2 \mathcal{I}_h u(x_i; \hat{v}_j) - \prod_{j=1}^d \frac{\partial^2 u}{\partial \hat{v}_j^2}(x_i), \qquad I_2 = \prod_{j=1}^d \frac{\partial^2 u}{\partial \hat{v}_j^2}(x_i) - \prod_{j=1}^d \frac{\partial^2 u}{\partial v_j^2}(x_i),$$

and $\hat{\mathbf{v}} = (\hat{v}_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$ is an approximation of $\mathbf{v}$ satisfying $|v_j - \hat{v}_j| \le \theta$ for all $1 \le j \le d$. The first term I_1 obeys a similar estimate to Step 1. For the second term I_2 we notice that $\hat{v}_j = v_j + w_j$ with $|w_j| \le \theta$, whence

$$\frac{\partial^2 u}{\partial \hat{v}_j^2}(x_i) = \hat{v}_j^T D^2u(x_i)\hat{v}_j = \frac{\partial^2 u}{\partial v_j^2}(x_i) + 2w_j^T D^2u(x_i)v_j + w_j^T D^2u(x_i)w_j.$$

Using that $\hat{v}_j = v_j + w_j$, we observe that

$$1 = |\hat{v}_j|^2 = |v_j|^2 + 2v_j \cdot w_j + |w_j|^2 \qquad \Rightarrow \qquad |v_j \cdot w_j| = \frac{1}{2}|w_j|^2 \le \frac{1}{2}\theta^2.$$

Since $D^2u(x_i)v_j = \lambda_j v_j$, we thus obtain

$$\left| \frac{\partial^2 u}{\partial \hat{v}_j^2}(x_i) - \frac{\partial^2 u}{\partial v_j^2}(x_i) \right| \le C\ \theta^2\ |u|_{W^2_\infty(B_i)}$$

as well as

$$I_2 \le C\ \theta^2\ |u|^d_{W^2_\infty(B_i)}.$$

This proves (2.3.3).

The remaining statement for $u \in W^2_\infty(B_i)$ is a simple consequence of Lemma 2.3.1 (i) and the above 2-step argument. □

Remark 2.3.3 (regularity)**.** *We give sufficient conditions for the regularity of u in Lemma 2.3.2: if $0 < f_0 \leq f(x) \leq f_1$ for all $x \in \Omega$ and $f \in C^{\alpha}(\overline{\Omega}), g \in C^3(\overline{\Omega})$, and $\partial\Omega \in C^3$, then $u \in C^{2,\alpha}(\overline{\Omega})$ [77, Theorem 1.1]. In such a case, there exist $0 < \lambda \leq \Lambda < \infty$ depending on f, g and Ω such that [32, Theorem 2.10]*

$$\lambda I \leq D^2 u(x) \leq \Lambda I \quad \forall x \in \Omega.$$

Since $\left|\frac{\partial^2 u}{\partial v_j^2}(x_i)\right| \leq \Lambda$, the constants C_1 and C_2 in Lemma 2.3.2 could also be written

$$C_1(d, \Omega, u) = C\Lambda^{d-1}|u|_{C^{2+k,\alpha}(B_i)}, \quad C_2(d, \Omega, u) = C\Lambda^{d-1}|u|_{W^2_{\infty}(B_i)}.$$

Remark 2.3.4 (optimal consistency error)**.** *We observe that for a smooth function the above lemma provides us with a consistency error of order h, after equating δ^2, θ^2 and $\frac{h^2}{\delta^2}$ and taking $\delta = h^{1/2}$.*

2.4 Convergence

Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) shows interior consistency at distance δ to $\partial\Omega_h$ for $u \in C^2(\overline{\Omega})$; hence the Barles-Souganidis theory [7] does not apply directly, as stated in [38]. We compensate with the fact that $\mathcal{I}_h u - u_\varepsilon$ vanishes on $\partial\Omega_h$ and cannot grow faster than $C\delta$ at distance δ to $\partial\Omega_h$. We make this statement rigorous via a barrier argument similar to those in [38,65,66]. To handle the behavior of $u - u_\varepsilon$ inside Ω_h we utilize Lemma 2.1.4 (discrete comparison principle) and Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$). In both cases we need the solution to be $C^2(\overline{\Omega})$, which may in general be false for the viscosity solution and thus requires a regularization argument involving data (f, g, Ω). We discuss these topics in this section and give a variation of the Barles-Souganidis approach as well.

2.4.1 Barrier Functions

We now introduce two discrete barrier functions, one to deal with the boundary behavior and the other one to handle the interior behavior. We employ them again in Section 2.7 to prove rates of convergence under certain regularity assumptions.

Lemma 2.4.1 (discrete boundary barrier)**.** *Let Ω be uniformly convex and $E > 0$ be arbitrary. For each node $z \in \mathcal{N}_h^0$ with $\operatorname{dist}(z, \partial\Omega_h) \leq \delta$, there exists a function $p_h \in \mathbb{V}_h$ such that $T_\varepsilon[p_h](x_i) \geq E$ for all $x_i \in \mathcal{N}_h^0$, $p_h \leq 0$ on $\partial\Omega_h$ and*

$$|p_h(z)| \leq CE^{1/d}\delta$$

with C depending on Ω.

Proof. Take $z_1 \in \partial\Omega_h$ such that $|z - z_1| = \operatorname{dist}(z, \partial\Omega_h) \leq \delta$. Upon extending the segment joining z and z_1, we find $z_2 \in \partial\Omega$ that satisfies the upper bound $|z_2 - z_1| \leq C_1 h$ because Ω is uniformly convex and thus Lipschitz but not necessarily W_∞^2. This implies that for $z_0 \in \partial\Omega$ such that $|z - z_0| = \operatorname{dist}(z, \partial\Omega)$, we have $|z - z_0| \leq |z - z_2| \leq \delta + C_1 h \leq C_2\delta$. We now make a change of coordinates so that z_0 becomes the origin and $z = (0, \ldots, 0, |z - z_0|)$. Since Ω is uniformly convex, it lies inside the ball

$$x_1^2 + x_2^2 + \ldots + x_{d-1}^2 + (x_d - R)^2 \leq R^2,$$

where the radius R depends on $\partial\Omega$ which is not necessarily W_∞^2. Under this coordinate system, let $p(x)$ be the quadratic polynomial

$$p(x) = \frac{E^{1/d}}{2}\left(x_1^2 + x_2^2 + \ldots + x_{d-1}^2 + (x_d - R)^2 - R^2\right)$$

and $p_h = \mathcal{I}_h p$ be its piecewise linear Lagrange interpolant in Ω_h. We note that $p \leq 0$ on $\overline{\Omega}$ yields $p_h \leq 0$ on $\partial\Omega_h$. Since p is convex and $\mathcal{I}_h p \geq p$, we infer that

$$T_\varepsilon[p_h](x_i) \geq T_\varepsilon[p](x_i) = E \quad \forall\ x_i \in \mathcal{N}_h^0,$$

where the last equality is a consequence of p being quadratic and

$$\nabla_\delta^2 p(x_i; v_j) = \partial_{v_j v_j}^2 p(x_i) = E^{1/d} \quad \forall\ v_j \in \mathbb{S}_\theta.$$

Moreover, since $|z - z_0| \leq C_2\delta$, we deduce $|p_h(z)| \leq C_\Omega E^{1/d}\delta$, as asserted. □

The following barrier function q_h and corresponding statement have already been used in the proof of Lemma 2.1.4 (discrete comparison principle).

Lemma 2.4.2 (discrete interior barrier)**.** *Let Ω be contained in the ball $B(x_0, R)$ of center x_0 and radius R. If $q(x) := \frac{1}{2}\big(|x - x_0|^2 - R^2\big)$, then its interpolant $q_h := \mathcal{I}_h q \in \mathbb{V}_h$ satisfies*

$$T_\varepsilon[q_h](x_i) \geq 1 \quad \forall x_i \in \mathcal{N}_h^0, \qquad q_h(x_i) \leq 0 \quad \forall x_i \in \mathcal{N}_h^b.$$

2.4.2 Approximation by Smooth Problems

For data f, g uniformly continuous in Ω, $f \geq 0$, and Ω uniformly convex, the regularity $u \in C^2(\overline{\Omega})$ which would yield small interior consistency error is not guaranteed. We thus embark on a regularization procedure similar to that used by DePhilippis-Figalli [31] and Awanou [5]. We start with a result about continuous dependence on data for viscosity solutions.

Lemma 2.4.3 (continuous dependence on data)**.** *Given $f_1, f_2 \in C(\overline{\Omega})$, $f_1, f_2 \geq 0$, and $g_1, g_2 \in C(\partial\Omega)$, let $u_1, u_2 \in C(\overline{\Omega})$ be the corresponding convex viscosity solutions of (1.1.1). Then there exists a constant C depending on Ω such that*

$$\|u_1 - u_2\|_{L^\infty(\Omega)} \leq C\|f_1 - f_2\|^{1/d}_{L^\infty(\Omega)} + \|g_1 - g_2\|_{L^\infty(\partial\Omega)}.$$

Proof. Let $q \leq 0$ be the barrier function of Lemma 2.4.2 (discrete interior barrier) and $F := \|f_1 - f_2\|^{1/d}_{L^\infty(\Omega)}$, $G := \|g_1 - g_2\|_{L^\infty(\partial\Omega)}$. We consider the auxiliary function

$$u_1^- := u_2 + Fq - G,$$

which is a convex viscosity subsolution of (1.1.1) with data (f_1, g_1). To prove this, let $\phi \in C^2(\Omega)$ and $x_0 \in \Omega$ be a point where $u_1^- - \phi$ attains a maximum. This implies that $u_2 - \big(\phi - Fq + G\big)$ attains also a maximum at x_0. Since u_2 is a viscosity subsolution of (1.1.1), and $D^2q(x_0) = I$ is the identity matrix, we deduce

$$\det\big(D^2\phi(x_0) - FI\big) \geq f_2(x_0) \geq 0.$$

Formula (1.1.2) for two symmetric positive semi-definite matrices A, B easily implies

$$\det(A + B) \geq \det(A) + \det(B).$$

Using this expression for $A = D^2\phi(x_0) - FI$ and $B = FI$ we obtain

$$\det(D^2\phi(x_0)) \geq f_2(x_0) + F^d = f_2(x_0) + \|f_1 - f_2\|_{L^\infty(\Omega)} \geq f_1(x_0).$$

In addition, since $q \leq 0$ in Ω, the function u_1^- satisfies on $\partial\Omega$

$$u_1^- \leq u_2 - G = g_2 - \|g_1 - g_2\|_{L^\infty(\partial\Omega)} \leq g_1.$$

These two properties of u_1^- imply that u_1^- is a viscosity subsolution of (1.1.1) with data (f_1, g_1). Since u_1^- is also convex, the comparison principle for (1.1.1) gives

$$u_1^- \leq u_1 \quad \Rightarrow \quad u_2 - u_1 \leq -Fq + G.$$

We similarly prove the reverse inequality, thus obtaining the desired estimate. □

We stress the monotonicity estimate

$$f_1 \geq f_2 \geq 0, \quad g_1 \leq g_2 \qquad \Rightarrow \qquad u_1 \leq u_2,$$

which is a consequence of u_1 being a convex subsolution of (1.1.1) with data (f_2, g_2).

Using the above result, we now show that we can approximate a viscosity solution u of (1.1.1) by regular (classical) solutions u_n.

Lemma 2.4.4 (approximation of viscosity solutions by smooth solutions)**.** *Let Ω be uniformly convex, f, g be uniformly continuous in Ω, $f \geq 0$, and u be the viscosity solution of (1.1.1) with data (f, g, Ω). Then, there exist a decreasing sequence of uniformly convex and smooth domains Ω_n converging to Ω in the sense that the Hausdorff distance $dist_H(\Omega_n, \Omega) \to 0$, a decreasing sequence of smooth functions $f_n > 0$ such that $f_n \to f$ uniformly in Ω, a sequence of smooth functions g_n such that $g_n \to g$ uniformly in Ω, and a sequence of smooth classical solutions u_n of (1.1.1) with data (f_n, g_n, Ω_n) such that $u_n \to u$ uniformly in Ω as $n \to \infty$.*

Proof. We prove the result in four steps.

Step 1: Domain Approximation. According to [16] there is a sequence of smooth and uniformly convex domains $\widetilde{\Omega}_n \subset \Omega$ that increase to Ω in the sense that

the Hausdorff distance $\text{dist}_H(\widetilde{\Omega}_n, \Omega) \to 0$. Since Ω is convex, it is star-shaped with respect to any of its points. Let's assume that the origin is contained in Ω and dilate the domains $\widetilde{\Omega}_n$ so that the ensuing domains Ω_n satisfy:

$$\Omega \subset \Omega_n \subset \Omega_m \quad m \leq n; \qquad \text{dist}_H(\Omega_n, \Omega) \to 0 \quad n \to \infty.$$

The domains Ω_n inherit the regularity of $\widetilde{\Omega}_n$ as well as their uniform convexity. Given $\delta_n \to 0$ as $n \to \infty$, to be chosen later in Step 4, we relabel Ω_n to be an approximate smooth domain so that $\text{dist}_H(\Omega_n, \Omega) \leq \delta_n$.

Step 2: Data Regularization. Let $\widetilde{\Omega}$ be an auxiliary domain such that $\Omega_n \subset \widetilde{\Omega}$ for all n. We now construct a sequence (f_n, g_n) of smooth functions defined in $\widetilde{\Omega}$ that converge uniformly in Ω to (f, g). We first extend (f, g) to $\widetilde{\Omega}$ and let $\sigma(t)$ be the modulus of continuity in $\widetilde{\Omega}$ for both (f, g) [34, Theorem 2.1.8.]:

$$|f(x) - f(y)|, |g(x) - g(y)| \leq \sigma(|x - y|) \quad \forall x, y \in \widetilde{\Omega}.$$

Let $\rho < \text{dist}_H(\Omega_n, \widetilde{\Omega})$ and let $\phi_\rho \geq 0$ be a standard smooth mollifier function with support in $B(0, \rho)$. We have for $f_\rho = f * \phi_\rho$ that

$$|f_\rho(x) - f(x)| = \left| \int_{\widetilde{\Omega}} (f(x - y) - f(x)) \, \phi_\rho(y) \, dy \right| \leq \sigma(\rho) \quad \forall x \in \Omega_n$$

because ϕ_ρ integrates to one. This implies that

$$\widetilde{f}_\rho(x) := f_\rho(x) + 2\sigma(\rho) \geq f(x) - \sigma(\rho) + 2\sigma(\rho) = f(x) + \sigma(\rho) > 0 \quad \forall x \in \Omega_n.$$

We now take $\rho_1 \leq \rho_2$ and observe that for all $x \in \Omega_n$

$$\begin{aligned} \widetilde{f}_{\rho_1}(x) - \widetilde{f}_{\rho_2}(x) &= \big(f_{\rho_1}(x) + 2\sigma(\rho_1)\big) - \big(f_{\rho_2}(x) + 2\sigma(\rho_2)\big) \\ &\leq f(x) + \sigma(\rho_1) + 2\sigma(\rho_1) - f(x) + \sigma(\rho_2) - 2\sigma(\rho_2) = 3\sigma(\rho_1) - \sigma(\rho_2) \leq 0, \end{aligned}$$

if $\sigma(\rho_1) \leq \frac{\sigma(\rho_2)}{3}$. We thus choose ρ_n such that $\sigma_n = \sigma(\rho_n) = 4^{-n}$ and define $f_n := \widetilde{f}_{\rho_n}$, which is a strictly positive and decreasing sequence of functions satisfying the error estimate

$$\sigma_n \leq f_n(x) - f(x) \leq 3\sigma_n \quad \forall\, x \in \Omega_n. \tag{2.4.1}$$

Similarly, we regularize g by convolution $g_\rho = g * \phi_\rho$ and define $g_n := g_{\rho_n}$ to obtain

$$\|g - g_n\|_{L^\infty(\Omega_n)} \leq \sigma_n. \tag{2.4.2}$$

Step 3: Boundary Behavior. Let u_n be the smooth classical solution of (1.1.1) with data (f_n, g_n, Ω_n), which satisfies $u_n \in C^{2,\alpha}(\Omega_n)$ with norms depending on n but uniform α; this is possible because (f_n, g_n, Ω_n) are smooth, Ω_n is uniformly convex, and $f_n > 0$ [24] [77, Theorem 1.1].

We now compare g and u_n at $z \in \partial\Omega$ without invoking any regularity of u_n but rather using a barrier argument. We start with g: if $y \in \partial\Omega_n$ is the closest point to z, then $|z - y| \leq \delta_n$ and

$$|g(z) - g(y)| \leq \sigma(|z - y|) \leq \sigma(\delta_n).$$

On the other hand, we know that

$$|g(y) - g_n(y)| \leq \sigma(\rho_n).$$

Let p be the quadratic barrier function introduced in the proof of Lemma 2.4.1, but now associated with Ω_n and $y \in \partial\Omega_n$. We consider the (lower) barrier function

$$b_n^-(x) := p(x) + g_n(y) + \nabla g_n(y)(x - y),$$

which satisfies

$$\det D^2 b_n^- = \det D^2 p \geq f_n \quad \text{in } \Omega_n$$

for $E > \|f\|_{L^\infty(\tilde{\Omega})}$ because b_n^- is a linear correction of p. We assert that $b_n^- \leq g_n$ on $\partial\Omega_n$ provided E also satisfies $E \geq C\|g_n\|_{W^2_\infty(\Omega_n)}$ where C depends on the uniform convexity of Ω. If this is true, then applying the comparison principle [48, Theorem 1.4.6] to the smooth functions b_n^- and u_n with data (f_n, g_n, Ω_n) yields

$$b_n^-(x) \leq u_n(x) \quad \forall\, x \in \Omega_n.$$

Taking $x = z$ and making use of the definition of b_n^- results in

$$-CE^{1/d}|z - y| + g_n(y) + \nabla g_n(y)(z - y) \leq u_n(z),$$

whence

$$u_n(z) - g_n(y) \geq -C_n|y - z| \geq -C_n\delta_n.$$

Similarly, upon letting $b_n^+(x) := -p(x) + g_n(y) + \nabla g_n(y)(x - y)$ be an upper barrier function, the preceding argument also shows

$$u_n(z) - g_n(y) \leq C_n|y - z| \leq C_n\delta_n,$$

whence the triangle inequality implies that for all $z \in \partial\Omega$

$$|g(z) - u_n(z)| \leq \sigma(\delta_n) + \sigma(\rho_n) + C_n\delta_n, \tag{2.4.3}$$

where the constant C_n depends on g_n but is independent of u_n. It remains to show

$$b_n^-(x) \leq g_n(x) \quad \forall\, x \in \partial\Omega_n.$$

We first observe that $b_n^-(y) = g_n(y)$ and the tangental gradients $\nabla_{\partial\Omega} b_n^-(y) = \nabla_{\partial\Omega} g_n(y)$ by construction, but g_n grows quadratically away from y on $\partial\Omega_n$ whereas p is just negative on $\partial\Omega_n$. To quantify the last statement, we let $y = 0$ for simplicity

and resort to the uniform convexity of Ω (and thus to that of every Ω_n) to deduce the existence of two balls B_R and B_r tangent to Ω_n at $0 \in \partial\Omega_n$ and so that

$$\Omega_n \subset B_r \subset B_R;$$

hence $r < R$. Note that $0 \in \partial B_r, \partial B_R$ and the centers of these balls are $(0, \dots, 0, r)$ and $(0, \dots, 0, R)$, respectively. We denote $x' = (x_i)_{i=1}^{d-1}$ and note that $x = (x', x_d) \in \partial B_r$ satisfy $|x'|^2 + (x_d - r)^2 = r^2$, whence

$$x_d\Big(1 - \frac{x_d}{2r}\Big) = \frac{1}{2r}|x'|^2 \quad \Rightarrow \quad \frac{1}{2r}|x'|^2 \le x_d \le \frac{1}{r}|x'|^2,$$

provided $x_d \le r$. This in turn implies for $1 < \xi < \frac{R}{r}$ fixed

$$p(x) \le p\Big(x', \frac{1}{2r}|x'|^2\Big) = \frac{E^{1/d}}{2}\Big(1 - \frac{R}{r} + \frac{1}{4r^2}|x'|^2\Big)|x'|^2 \le \frac{E^{1/d}}{2}(1-\xi)|x'|^2 < 0,$$

provided $|x'|^2 \le C_1 := 4r^2\big(\frac{R}{r} - \xi\big)$ and R is used in the definition of p. Since $|x'|^2 \le r^2$ and $x_d^2 \le \frac{|x'|^4}{r^2} \le |x'|^2$, we have that $|x|^2 = |x'|^2 + x_d^2 \le 2|x'|^2$ and we deduce

$$|x'|^2 \le C_1 \quad \Rightarrow \quad p(x) \le -E^{1/d}\frac{(\xi - 1)}{4}|x|^2 = -E^{1/d}C_2|x|^2.$$

On the other hand, for $x \in \partial B_r$ with $|x'| > C_1$ we infer that the distance from x to ∂B_R is strictly positive whence

$$p(x) \le -C_3|x|^2.$$

Since both constants C_2, C_3 depend only on r, R, we see that p grows quadratically on ∂B_r with a constant independent of n, and thus on $\Omega_n \subset B_r$. To compare b_n^- with g_n, we recall that g_n is a smooth function for Taylor formula to give

$$\Big|g_n(x) - g_n(0) - \nabla g_n(0)x\Big| \le \frac{1}{2}|g_n|_{W^2_\infty(\Omega_n)}|x|^2 \qquad \forall\, x \in \Omega_n.$$

We finally choose the factor E in b_n^- proportional to $|g_n|_{W^2_\infty(\Omega_n)}$ and realize that $b_n^-(x) \le g_n(x)$ for all $x \in \partial\Omega_n$ as asserted.

Step 4: Uniform Convergence. We view both u and u_n as viscosity solutions of (1.1.1), the former with data (f, g, Ω) and the latter with data (f_n, u_n, Ω). Applying Lemma 2.4.3 (continuous dependence on data), along with (2.4.1) and (2.4.3), we obtain

$$\begin{aligned}\|u_n - u\|_{L^\infty(\Omega)} &\le C\|f_n - f\|^{1/d}_{L^\infty(\Omega)} + \|u_n - g\|_{L^\infty(\partial\Omega)} \\ &\le C\sigma(\rho_n)^{1/d} + \sigma(\rho_n) + \sigma(\delta_n) + C_n\delta_n.\end{aligned}$$

Given an arbitrary number β we first choose ρ_n so that $C\sigma(\rho_n)^{1/d} + \sigma_n(\rho_n) \le \frac{\beta}{2}$. This choice determines the regularity of g_n, namely its W^2_∞ and $C^{2,\alpha}$ norms in $\widetilde{\Omega}$. Since C_n is proportional to $|g_n|_{W^2_\infty(\Omega_n)}$, we finally select δ_n so that $\sigma(\delta_n) + C_n\delta_n \le \frac{\beta}{2}$. This shows the desired uniform convergence of u_n to u in Ω. □

2.4.3 Uniform Convergence: Regularization Approach

In this section we combine Lemma 2.1.4 (discrete comparison principle), Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$), Lemma 2.4.1 (discrete boundary barrier), Lemma 2.4.2 (discrete interior barrier), and Lemma 2.4.4 (approximation of viscosity solutions by smooth solutions) to prove uniform convergence of u_ε to u in Ω.

Since u_ε is defined in the computational domain Ω_h, and $\Omega_h \subset \Omega$, we extend u_ε to Ω as follows. Given $x \in \Omega \setminus \Omega_h$ let $z \in \partial\Omega_h$ be the closest point to x, which is unique because Ω_h is convex, and let

$$u_\varepsilon(x) := u_\varepsilon(z) = \mathcal{I}_h g(z) \quad \forall\, x \in \Omega \setminus \Omega_h. \tag{2.4.4}$$

Theorem 2.4.5 (uniform convergence)**.** *Let Ω be uniformly convex, $f, g \in C(\overline{\Omega})$ and $f \geq 0$ in Ω. The discrete solution u_ε of (2.1.2) and (2.4.4) converges uniformly to the unique viscosity solution $u \in C(\overline{\Omega})$ of (1.1.1) as $\varepsilon = (h, \delta, \theta) \to 0$ and $\frac{h}{\delta} \to 0$.*

Proof. We first split

$$\|u - u_\varepsilon\|_{L^\infty(\Omega)} \leq \|u - u_\varepsilon\|_{L^\infty(\Omega_h)} + \|u - u_\varepsilon\|_{L^\infty(\Omega \setminus \Omega_h)}$$

and then employ the triangle inequality to write

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \leq \|u - u_n\|_{L^\infty(\Omega_h)} + \|u_n - \mathcal{I}_h u_n\|_{L^\infty(\Omega_h)} + \|\mathcal{I}_h u_n - u_\varepsilon\|_{L^\infty(\Omega_h)}.$$

Next, we recall that Lemma 2.4.4 yields $\|u - u_n\|_{L^\infty(\Omega_h)} \leq \|u - u_n\|_{L^\infty(\Omega)} \to 0$ as $n \to \infty$. In addition, polynomial interpolation theory gives

$$\|u_n - \mathcal{I}_h u_n\|_{L^\infty(\Omega_h)} \leq C|u_n|_{W^2_\infty(\Omega)} h^2 \to 0,$$

as $h \to 0$ for n fixed. On the other hand, (2.4.4) yields

$$|u(x) - u_\varepsilon(x)| = |u(x) - u_\varepsilon(z)| \leq |u(x) - u(z)| + |u(z) - u_\varepsilon(z)| \quad \forall\, x \in \Omega \setminus \Omega_h$$

where $z \in \partial\Omega_h$. If τ is the modulus of continuity of $u \in C(\overline{\Omega})$, we have

$$\|u - u_\varepsilon\|_{L^\infty(\Omega \setminus \Omega_h)} \leq \tau\big(\mathrm{dist}_H(\Omega, \Omega_h)\big) + \|u - u_\varepsilon\|_{L^\infty(\Omega_h)}$$

Since $\mathrm{dist}_H(\Omega, \Omega_h) \to 0$, as $h \to 0$, the proof reduces to showing that $\|\mathcal{I}_h u_n - u_\varepsilon\|_{L^\infty(\Omega_h)}$ can be made arbitrarily small. We do this in three steps.

Step 1: Boundary Estimate. Let p_h be the function of Lemma 2.4.1 (discrete boundary barrier) with constant $E_{n,1} := C|u_n|^d_{W^2_\infty(\Omega)} + 3\sigma_n$, where $C|u_n|^d_{W^2_\infty(\Omega)}$ is the consistency error (2.3.3) from Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) with u_n in place

of u and $3\sigma_n$ is a bound (2.4.1) for $\|f - f_n\|_{L^\infty(\Omega)}$. Since both u_ε and p_h are discretely convex, we have

$$T_\varepsilon[u_\varepsilon + p_h](x_i) \geq T_\varepsilon[u_\varepsilon](x_i) + T_\varepsilon[p_h](x_i) \geq f(x_i) + E_{n,1} \geq T_\varepsilon[\mathcal{I}_h u_n](x_i)$$

for all $x_i \in \mathcal{N}_h^0$. Moreover, since (2.4.3) holds for all $z \in \partial\Omega$ and $\mathcal{N}_h^b \subset \partial\Omega$, linear interpolation implies that $\mathcal{I}_h u_n \geq \mathcal{I}_h g - \xi_n = u_\varepsilon - \xi_n$ on $\partial\Omega_h$ for all h, where $\xi_n := \sigma(\rho_n) + \sigma(\delta_n) + C_n\delta_n$ and $\delta_n \geq \operatorname{dist}_H(\Omega_n, \Omega)$, whence $u_\varepsilon + p_h - \xi_n \leq \mathcal{I}_h u_n$ on $\partial\Omega_h$. Consequently, for all $z \in \mathcal{N}_h^0$ such that $\operatorname{dist}(z, \partial\Omega) \leq 2\delta$, Lemma 2.1.4 (discrete comparison principle) yields

$$u_\varepsilon(z) - CE_{n,1}^{1/d}\delta - \xi_n \leq \mathcal{I}_h u_n(z).$$

A similar argument with $u_\varepsilon - p_h + \xi_n$ gives rise to the reverse estimate.

Step 2: Interior Estimate. We resort to the function q_h of Lemma 2.4.2 (discrete interior barrier) to construct a discrete lower barrier b_ε^- as follows: let

$$E_{n,2} := C|u_n|_{C^{2+\alpha}(\overline{\Omega})}|u_n|_{W^2_\infty(\Omega)}^{d-1}\delta^\alpha + C|u_n|^d_{W^2_\infty(\Omega)}\left(\frac{h^2}{\delta^2} + \theta^2\right) + 3\sigma_n$$

and

$$b_\varepsilon^- := u_\varepsilon + E_{n,2}^{1/d} q_h - CE_{n,1}^{1/d}\delta - \xi_n.$$

Since $q_h \leq 0$, Step 1 guarantees that $b_\varepsilon^- \leq \mathcal{I}_h u_n$ on $\partial\Omega_{h,\delta}$, where $\Omega_{h,\delta}$ is defined in (2.3.1). Applying Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) for u_n instead of u implies

$$T_\varepsilon[b_\varepsilon^-](x_i) \geq T_\varepsilon[u_\varepsilon](x_i) + E_{n,2} = f(x_i) + E_{n,2} \geq T_\varepsilon[\mathcal{I}_h u_n](x_i) \quad \forall\, x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta}$$

where we have used that both u_ε and q_h are discretely convex as well as (2.4.1). Lemma 2.1.4 (discrete comparison principle) yields

$$b_\varepsilon^- = u_\varepsilon + E_{n,2}^{1/d} q_h - CE_{n,1}^{1/d}\delta - \xi_n \leq \mathcal{I}_h u_n \quad \text{in} \quad \Omega_{h,\delta}.$$

A similar argument with $b_\varepsilon^+ := u_\varepsilon - E_{n,2}^{1/d} q_h + C E_{n,1}^{1/d}\delta + \xi_n$ results in $b_\varepsilon^+ \geq \mathcal{I}_h u_n$.

Combining these estimates with those of Step 1, we end up with

$$\|u_\varepsilon - \mathcal{I}_h u_n\|_{L^\infty(\Omega_h)} \leq C E_{n,1}^{1/d}\delta + C E_{n,2}^{1/d} + \xi_n. \tag{2.4.5}$$

Step 3: Uniform convergence in Ω. We finally proceed as in Step 4 of the proof of Lemma 2.4.4 (approximation of viscosity solutions by smooth solutions). Given an arbitrary number $\beta > 0$, we choose ρ_n so that $\sigma(\rho_n) \leq \frac{\beta}{3}$. This dictates the regularity of g_n hidden in the constant C_n of ξ_n, as well as that of u_n, and allows us to select δ_n so that $\sigma(\delta_n) + C_n\delta_n \leq \frac{\beta}{3}$; hence $\xi_n \leq \frac{2\beta}{3}$. We next take $\delta, \frac{h}{\delta}$ and θ small enough, depending on u_n, so that the first two terms of (2.4.5) are $\leq \frac{\beta}{3}$ and thus $\|u_\varepsilon - \mathcal{I}_h u_n\|_{L^\infty(\Omega_h)} \leq \beta$. This completes the proof. □

2.4.4 Uniform Convergence: Barles-Souganidis Approach

In this section we adapt the approach of [7] to our setting. Since (2.4.4) extends the definition of discrete solution u_ε to Ω, we let the limit supremum and limit infimum of u_ε be

$$u^*(x) = \limsup_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} u_\varepsilon(z), \qquad u_*(x) = \liminf_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} u_\varepsilon(z),$$

and observe that u^* is upper semi-continuous and u_* is lower semi-continuous. We show that they attain the Dirichlet boundary condition pointwise. Moreover, they are viscosity subsolution and supersolution of (1.1.1), respectively. An essential difficulty associated with (1.1.1), already mentioned in [38], is that viscosity sub and supersolutions of (1.1.1) must be convex for the comparison principle to be

applicable. Since u_ε is only discretely convex, it is not obvious that u^* and u_* are convex.

To circumvent this issue we proceed as in [38]: we let $\partial^{2,+}_{v_jv_j}u := \max\left(\partial^2_{v_jv_j}u, 0\right)$, $\partial^{2,-}_{v_jv_j}u := -\min\left(\partial^2_{v_jv_j}u, 0\right)$, introduce the continuous version of our ideal discrete operator in (2.1.2)

$$T[u] := \min_{\mathbf{v}=(v_j)_{j=1}^d \in \mathbb{S}^\perp} \left(\prod_{j=1}^d \partial^{2,+}_{v_jv_j}u - \sum_{j=1}^d \partial^{2,-}_{v_jv_j}u \right),$$

and show that u is a convex viscosity solution of (1.1.1) if and only if u is a viscosity solution of the Dirichlet problem

$$T[u] = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{2.4.6}$$

for which we do not require convexity because it is built-in the operator.

Lemma 2.4.6 (equivalence of viscosity solutions)**.** *If $f \in C(\Omega)$ satisfies $f \geq 0$, and $u \in C(\overline{\Omega})$, then u is a viscosity solution of (2.4.6) if and only if u is a convex viscosity solution of (1.1.1).*

Proof. Since u is uniformly continuous in Ω the notion of Dirichlet condition is classical in both cases. We thus verify the equation in the viscosity sense.

Step 1: Necessity. We rely on the notion of convexity of a function $v \in C(\Omega)$ in the viscosity sense: for test function $\phi \in C^2(\Omega)$ that touches v from above at a point $x \in \Omega$ the smallest eigenvalue $\lambda_1[D^2\phi](x)$ of $D^2\phi$ at x satisfies

$$\lambda_1[D^2\phi](x) \geq 0.$$

It is proven in [68] that a continuous function v is convex if and only if it is convex

in the viscosity sense. We show that a viscosity solution u of (2.4.6) is convex in the viscosity sense and use this equivalence to deduce convexity of u.

We observe that u being a viscosity solution of (2.4.6) implies that for $\phi \in C^2(\Omega)$ touching u from above at $x \in \Omega$, we have

$$T[\phi](x) \geq f(x) \geq 0$$

We argue as in Lemma 2.1.2: if there is a direction $v_j \in \mathbb{S}$ for which $\frac{\partial^2 \phi}{\partial v_j^2}(x) < 0$, then $T[\phi](x) < 0$ which contradicts the preceding statement. Therefore

$$\frac{\partial^2 \phi}{\partial v_j^2}(x) \geq 0 \quad \forall\, v_j \in \mathbb{S} \qquad \Rightarrow \qquad \lambda_1[D^2\phi](x) \geq 0.$$

This proves that u is convex as well as

$$\det D^2\phi(x) = T[\phi](x) = \min_{\mathbf{v}=(v_j)_{j=1}^d \in \mathbb{S}^\perp} \prod_{j=1}^d \partial^2_{v_j v_j}\phi(x) \geq f(x)$$

according to (1.1.2). This implies that u is a convex subsolution of (1.1.1).

To prove that u is also a supersolution of (1.1.1), we recall that the definition of viscosity solutions for (1.1.1) uses convex test functions $\phi \in C^2(\Omega)$ [48]; hence $\det D^2\phi = T[\phi]$. Consequently, if $u - \phi$ attains a minimum at $x \in \Omega$, then

$$\det D^2\phi(x) = T[\phi](x) \leq f(x)$$

whence u is a supersolution of (1.1.1).

Step 2: Sufficiency. Let's assume now that u is a convex viscosity solution of (1.1.1), and $\phi \in C^2(\Omega)$ is a test function that touches u at $x_0 \in \Omega$. Inspired by [48, Remark 1.3.3], we decompose $\phi = q + r$ into a quadratic q and a remainder r

$$q(x) = \phi(x_0) + D\phi(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T D^2\phi(x_0)(x - x_0), \quad r(x) = o(|x - x_0|^2);$$

hence $D^2\phi(x_0) = D^2q(x_0)$. If $q^{\pm}(x) := q(x) \pm \sigma|x - x_0|^2$, we then observe that $q^+ \geq \phi$ and $q^- \leq \phi$ in a suitable neighborhood of x_0 provided $\sigma > 0$. We take advantage of $q^{\pm}$ being quadratic to realize that $q^{\pm}$ is convex if and only if $D^2q^{\pm}(x_0) \geq 0$.

If $u - \phi$ attains a local max at x_0, so does $u - q^+$ and $D^2\phi(x_0) \geq 0$ because u is convex. Therefore, the quadratic q^+ is convex and must satisfy

$$\det D^2q^+(x_0) = \det(D^2q(x_0) + 2\sigma I) \geq f(x_0),$$

because u is a viscosity solution of (1.1.1). Take the limit $\sigma \downarrow 0$ to find out that $T[\phi](x_0) = \det D^2\phi(x_0) \geq f(x_0)$ whence u is a viscosity subsolution of (2.4.6).

On the other hand, if $u - \phi$ attains a local min at x_0, so does $u - q^-$. We have now two possible cases. If all the eigenvalues of $D^2\phi(x_0)$ are strictly positive, then q^- is a convex quadratic for σ sufficiently small. This in turn implies

$$\det D^2q^-(x_0) = \det(D^2q(x_0) - 2\sigma I) \leq f(x_0),$$

as u is a viscosity solution of (1.1.1); hence $T[\phi](x_0) = \det D^2\phi(x_0) \leq f(x_0)$ upon letting $\sigma \downarrow 0$. If any eigenvalue of $D^2\phi(x_0)$ is non-positive, then $T[\phi](x_0) \leq 0$ by definition and $T[\phi](x_0) \leq f(x_0)$ because $f \geq 0$. We thus deduce that u is a viscosity supersolution of (2.4.6), whence a viscosity solution of (2.4.6), as asserted. □

We are now ready to prove the convergence of our discrete solution u_ε to the viscosity solution u of (1.1.1).

Theorem 2.4.7 (uniform convergence)**.** *Let Ω be uniformly convex, $f \in C(\Omega) \cap L^\infty(\Omega)$ satisfy $f \geq 0$, and $g \in C(\partial\Omega)$. The discrete solution u_ε of (2.1.2) converges*

uniformly to the unique viscosity solution $u \in C(\overline{\Omega})$ *of* (1.1.1) *as* $\varepsilon = (h, \delta, \theta) \to 0$ *and* $\frac{h}{\delta} \to 0$.

Proof. In view of Lemma 2.4.6 (equivalence of viscosity solutions), we prove that u_ε converges to the viscosity solution of (2.4.6). To this end, we have to deal with a test function $\phi \in C^2(\Omega)$ and its Lagrange interpolant $\phi_h = \mathcal{I}_h \phi$. Without loss of generality we may assume $\phi \in C^{2,\alpha}(\Omega)$. We split the proof into five steps.

Step 1: Consistency. We have the following alternative to (2.3.3)

$$\big|T[\phi](x_0) - T_\varepsilon[\phi_h](x_i)\big| \leq C_1(\phi)\Big(\delta^\alpha + |x_0 - x_i|^\alpha\Big) + C_2(\phi)\Big(\frac{h^2}{\delta^2} + \theta^2\Big),$$

where the constants C_1, C_2 are defined in Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u](x_i)$) and depend on $|\phi|_{C^{2,\alpha}(B_i)}$ and $|\phi|_{W^2_\infty(B_i)}$ with B_i defined in (2.3.2), and $x_0 \in \Omega, x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta}$. The proof of this inequality proceeds along the lines of those of Lemmas 2.3.1 and 2.3.2, except that now we need to deal with the functions $s \mapsto \max(s, 0)$ and $s \mapsto \min(s, 0)$ in the definitions of both T and T_ε because ϕ may not be convex. We exploit that these functions are Lipschitz with constant 1 to write

$$\big|\nabla_\delta^{2,+}\phi_h(x_i; v_j) - \partial^{2,+}_{v_j v_j}\phi(x_0)\big| \lesssim |\phi|_{C^{2,\alpha}(B_i)}\Big(\delta^\alpha + |x_0 - x_i|^\alpha\Big) + |u|_{W^2_\infty(B_i)}\frac{h^2}{\delta^2},$$

together with a similar bound for the operators $\nabla_\delta^{2,-}$ and $\partial^{2,-}_{v_j v_j}$.

Step 2: Subsolutions. We show that u^* is a viscosity subsolution of (2.4.6); likewise u_* is a viscosity supersolution. This hinges on monotonicity and consistency [7]. We must show that if $u^* - \phi$ attains a local maximum at $x_0 \in \Omega$, we have

$$T[\phi](x_0) \geq f(x_0);$$

note that $u^* - \phi$ is upper semi-continuous and the local maximum is well defined. Without loss of generality, we may assume that $u^* - \phi$ attains a strict global maximum at $x_0 \in \Omega$ [52, Remark in p.31], and $x_0 \in \Omega_h$ for h sufficiently small. Let u_ε and z_h be a sequence of functions and nodes such that

$$\lim_{\varepsilon, \frac{h}{\delta} \to 0, z_h \to x_0} u_\varepsilon(z_h) = u^*(x_0).$$

Let $x_h \in \mathcal{N}_h$ be a sequence of nodes so that $u_\varepsilon - \phi_h$ attains a maximum at x_h. We claim that $x_h \to x_0$ as $h \to 0$. If not, then there exists a subsequence $x_h \to y_0$ such that $y_0 \neq x_0$. Since $(u_\varepsilon - \phi_h)(x_h) \geq (u_\varepsilon - \phi_h)(z_h)$, passing to the limit we obtain

$$(u^* - \phi)(y_0) \geq \limsup_{\varepsilon, \frac{h}{\delta} \to 0} (u_\varepsilon - \phi_h)(x_h) \geq \lim_{\varepsilon, \frac{h}{\delta} \to 0, z_h \to x_0} (u_\varepsilon - \phi_h)(z_h) = (u^* - \phi)(x_0).$$

This contradicts the fact that $u^* - \phi$ attains a strict maximum at x_0. Exploiting the fact that $u_\varepsilon - \phi_h$ attains a maximum at x_h, Lemma 2.1.3 (monotonicity) yields

$$T_\varepsilon[\phi_h](x_h) \geq T_\varepsilon[u_\varepsilon](x_h) = f(x_h).$$

Since $f \in C(\Omega)$, to prove $T[\phi](x_0) \geq f(x_0)$ we only need to show that as $\varepsilon, \frac{h}{\delta} \to 0$

$$T_\varepsilon[\phi_h](x_h) \to T[\phi](x_0).$$

This is a consequence of Step 1 and the fact that $x_h \in \Omega_{h,\delta}$ for δ sufficiently small, because $x_0 \in \Omega$, $x_h \to x_0$ and the sequence of $\Omega_h \uparrow \Omega$ is non-decreasing.

Step 3: Boundary Behavior. We now prove that $u^* = u_* = g$ on $\partial\Omega$ via a barrier argument similar to those in [38, 65, 66]; we proceed as in [38]. This is essential to apply the comparison principle for operator T to relate u_*, u^* and u in Step 4.

Let p_k be the quadratic function in the proof of Lemma 2.4.1 (discrete boundary barrier) associated with an arbitrary boundary point $x \in \partial\Omega$ (the origin in the construction of p_k) and with constant $E = k$. We recall that $p_k(x) = 0$ and $p_k(z) \leq 0$ for all $z \in \partial\Omega$ can be made arbitrarily large for $k \to \infty$ by virtue of the uniform convexity of Ω. A simple consequence is that the sequence of points $x_k \in \partial\Omega$ where $g + p_k$ (resp. $g - p_k$) attains a maximum (resp. a minimum) over $\partial\Omega$ converges to x.

We now observe that taking $w_h \equiv 0$ in Lemma 2.1.3 (monotonicity) implies the following maximum principle: if a discretely convex function u_h satisfies $T_\varepsilon[u_h](x_i) > 0$ for all $x_i \in \mathcal{N}_h^0$, then u_h attains a maximum over $\overline{\Omega}_h$ on $\mathcal{N}_h^b \subset \partial\Omega$. Apply this to $T_\varepsilon[u_\varepsilon + \mathcal{I}_h p_k] > 0$ to deduce that $u_\varepsilon + \mathcal{I}_h p_k$ attains its maximum on $\mathcal{N}_h^b$. In view of (2.4.4), we may assume $z \in \Omega_h$ in $u^*(x) = \limsup_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} u_\varepsilon(z)$. Consequently,

$$
\begin{aligned}
u^*(x) &\leq \limsup_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} \big(u_\varepsilon(z) + \mathcal{I}_h p_k(z)\big) - \liminf_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} \mathcal{I}_h p_k(z) \\
&\leq \limsup_{\varepsilon, \frac{h}{\delta} \to 0} \max_{z \in \mathcal{N}_h^b} \big(g + p_k\big)(z) - p_k(x) \leq g(x_k) + p_k(x_k) \leq g(x_k),
\end{aligned}
$$

because $\max_{\mathcal{N}_h^b} g + p_k \leq \max_{\partial\Omega} g + p_k$, whence taking $k \to \infty$ yields $u^*(x) \leq g(x)$.

On the other hand, since $T_\varepsilon[\mathcal{I}_h p_k](x_i) > T_\varepsilon[u_\varepsilon](x_i)$ for all $x_i \in \mathcal{N}_h^0$ and k big enough, Lemma 2.1.3 implies that $u_\varepsilon - \mathcal{I}_h p_k$ attains a minimum on $\mathcal{N}_h^b$. Therefore, arguing as before

$$
u_*(x) \geq \liminf_{\varepsilon, \frac{h}{\delta} \to 0} \min_{z \in \mathcal{N}_h^b} \big(g - p_k\big)(z) + p_k(x) \geq g(x_k) - p_k(x_k) \geq g(x_k),
$$

whence $u_*(x) \geq g(x)$. This in turn gives $u^* \leq g \leq u_* \leq u^*$ on $\partial\Omega$ as asserted.

Step 4: Comparison. To prove that $u^* = u_*$ in $\overline{\Omega}$ we use the following com-

parison principle for (2.4.6): if v^- is a subsolution and is upper semi-continuous in $\overline{\Omega}$, v^+ is a supersolution and is lower semi-continuous in $\overline{\Omega}$, and $v^- \leq v^+$ on $\partial\Omega$, then $v^- \leq v^+$ on $\overline{\Omega}$. This result falls under the umbrella of [27, Theorem 3.3]. It hinges on an argument mentioned in [27, Section 5.C] that is briefly described for a more general form of the Monge-Ampère operator in [52, V.3]. Both operators in (1.1.1) and (2.4.6) satisfy the requirements posed in [52]. We apply this comparison principle to $v^- = u^*$ and $v^+ = u_*$, which satisfy the assumptions in view of Steps 2 and 3, to obtain $u^* \leq u_*$ in $\overline{\Omega}$. Since $u^* \geq u_*$ by definition, this results in $u^* = u_*$ in $\overline{\Omega}$.

Step 5: Uniform Convergence. Step 4 implies the pointwise limit

$$u(x) = \lim_{\varepsilon \to 0, z \to x} u_\varepsilon(z) \quad \forall\, x \in \overline{\Omega}.$$

To see that this gives rise to uniform convergence we argue by contradiction. We assume that for every ε there exist a point $x_\varepsilon \in \overline{\Omega}$ such that $|u(x_\varepsilon) - u_\varepsilon(x_\varepsilon)| \geq \sigma$, for some $\sigma > 0$. Since $\overline{\Omega}$ is compact, there exists a subsequence (not relabeled) $x_\varepsilon \to x_0 \in \overline{\Omega}$. Computing the limit $\varepsilon \to 0$ in the last inequality yields the contradiction $|u(x_0) - u(x_0)| \geq \sigma$. This concludes the proof. □

2.5 Discrete Alexandroff Estimate

In this section, we review several concepts related to convexity as well as the discrete Alexandroff estimate of [65]. This is the first step into our analysis for the rates of convergence for the two-scale method. We first recall several definitions.

Definition 2.5.1 (subdifferential)**.**

(*i*) *The subdifferential of a function w at a point $x_0 \in \Omega_h$ is the set*

$$\partial w(x_0) := \left\{ p \in \mathbb{R}^d : \ w(x) \geq w(x_0) + p \cdot (x - x_0), \quad \forall x \in \Omega_h \right\}.$$

(*ii*) *The subdifferential of a function w on a set $E \subset \Omega_h$ is $\partial u(E) := \cup_{x \in E} \partial w(x)$.*

Definition 2.5.2 (convex envelope and discrete lower contact set)**.**

(*i*) *The convex envelope Γu of a function w is defined to be*

$$\Gamma w(x) := \sup_L \{ L(x), \ L(y) \leq w(y) \text{ for all } y \in \Omega_h \text{ and } L \text{ is affine} \}.$$

(*ii*) *The discrete lower contact set $C_-(w_h)$ of a function $w_h \in \mathbb{V}_h$ is the set of nodes where the function coincides with its convex envelope, i.e.*

$$\mathcal{C}_-(w_h) := \left\{ x_i \in \mathcal{N}_h^0 : \Gamma w_h(x_i) = w_h(x_i) \right\}.$$

Remark 2.5.3 (w_h dominates Γw_h)**.** *Since $w_h \geq \Gamma w_h$, at a contact node $x_i \in \mathcal{C}_-(w_h)$ we have*

$$\nabla_\delta^2 \Gamma w_h(x_i; v_j) \leq \nabla_\delta^2 w_h(x_i; v_j)(x_i) \qquad \forall v_j \in \mathbb{S}_\theta.$$

Remark 2.5.4 (minima of w_h and Γw_h)**.** *A consequence of Definition 2.5.2 (convex envelope and discrete lower contact set) is that the minima of $w_h \in \mathbb{V}_h$ and Γw_h are attained at the same contact nodes and are equal.*

We can now present the discrete Alexandroff estimate from [65], which states that the minimum of a discrete function is controlled by the measure of the subdifferential of its convex envelope in the discrete contact set.

Proposition 2.5.5 (discrete Alexandroff estimate [65]). *Let v_h be a continuous piecewise linear function that satisfies $v_h \geq 0$ on $\partial\Omega_h$. Then,*

$$\max_{x_i \in \mathcal{N}_h^0} v_h(x_i)^- \leq C \left(\sum_{x_i \in \mathcal{C}_-(v_h)} |\partial\Gamma v_h(x_i)| \right)^{1/d}$$

where $C = C(d, \Omega)$ depends only on the dimension d and the domain Ω.

2.6 Continuous Dependence on Data

We now derive the continuous dependence of the discrete solution on data in Section 2.6.3, which is essential to prove rates of convergence. To this end, we first prove a stability estimate in the max norm in Section 2.6.1 and the concavity of the discrete operator in Section 2.6.2.

2.6.1 Stability of the Two-Scale Method.

We start with some geometric estimates. The first and second lemmas connect the discrete Alexandroff estimate with the two-scale method. They allow us to estimate the measure of the subdifferential of a discrete function w_h in terms of our discrete operator $T_\varepsilon[w_h]$, defined in (2.1.2).

Lemma 2.6.1 (subdifferential vs hyper-rectangle). *Let $w \in C^0(\overline{\Omega}_h)$ be convex and $x_i \in \mathcal{N}_h^0$ be so that $x_i \pm \hat{\delta} v \in \overline{\Omega}_h$ for all $v \in \mathbb{S}_\theta$ with $\hat{\delta} \leq \delta$. If $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$ and*

$$\alpha_{i,j}^\pm := \frac{w(x_i \pm \hat{\delta} v_j) - w(x_i)}{\hat{\delta}} \quad \forall\, 1 \leq j \leq d,$$

then

$$\partial w(x_i) \subset \left\{ p \in \mathbb{R}^d : \ \alpha_{i,j}^- \leq p \cdot v_j \leq \alpha_{i,j}^+ \ 1 \leq j \leq d \right\}.$$

Proof. Take $p \in \partial w(x_i)$ and write

$$w(x) \geq w(x_i) + p \cdot (x - x_i) \quad \forall x \in \overline{\Omega}_h.$$

Consequently, for any $1 \leq j \leq d$ we infer that

$$w(x_i + \hat{\delta} v_j) \geq w(x_i) + \hat{\delta}\ p \cdot v_j, \quad w(x_i - \hat{\delta} v_j) \geq w(x_i) - \hat{\delta}\ p \cdot v_j,$$

or equivalently

$$\frac{w(x_i) - w(x_i - \hat{\delta} v_j)}{\hat{\delta}} \leq p \cdot v_j \leq \frac{w(x_i + \hat{\delta} v_j) - w(x_i)}{\hat{\delta}}.$$

This implies that p belongs to the desired set. □

Lemma 2.6.2 (hyper-rectangle volume)**.** *For d-tuple* $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$ *the volume of the set*

$$K = \left\{ p \in \mathbb{R}^d : \ a_j \leq p \cdot v_j \leq b_j, \quad j = 1, \ldots, d \right\}$$

is given by

$$|K| = \prod_{j=1}^{d} (b_i - a_i).$$

Proof. Let $V = [v_1, \cdots, v_d] \in \mathbb{R}^{d \times d}$ be the orthogonal matrix whose columns are the elements of $\mathbf{v}$; hence $v_j = V e_j$ where $\{e_j\}_{j=1}^d$ is the canonical basis in $\mathbb{R}^d$. We now seek a more convenient representation of K

$$\begin{aligned} K &= \left\{ p \in \mathbb{R}^d : \ a_j \leq p \cdot (V e_j) \leq b_j, \quad j = 1, \ldots, d \right\} \\ &= V^{-T} \left\{ x \in \mathbb{R}^d : \ a_j \leq x \cdot e_j \leq b_j, \quad j = 1, \ldots, d \right\} = V^{-T} \widetilde{K}, \end{aligned}$$

whence

$$|K| = |\det V^{-T}|\ |\widetilde{K}| = |\widetilde{K}| = \prod_{j=1}^{d} (b_j - a_j),$$

because $\widetilde{K}$ is an orthogonal hyper-rectangle. □

Combining Lemmas 2.6.1 and 2.6.2 we get the following corollary.

Corollary 2.6.3 (subdifferential vs discrete operator)**.** *For every $x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta}$ and a convex function w we have that*

$$|\partial w(x_i)| \leq \left(\min_{\mathbf{v} \in \mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w(x_i; v_j) \right) \delta^d.$$

Lemma 2.6.4 (stability)**.** *If $w_h \in \mathbb{V}_h$ is $w_h \geq 0$ on $\partial\Omega_h$, then*

$$\max_{x_i \in \mathcal{N}_h^0} w_h(x_i)^- \leq C\delta \left(\sum_{x_i \in \mathcal{C}_-(w_h)} T_\varepsilon[w_h](x_i) \right)^{1/d}.$$

Proof. Since the function $w_h \geq 0$ on $\partial\Omega_h$, we invoke Proposition 2.5.5 (discrete Alexandroff estimate) for w_h to obtain

$$\max_{x_i \in \mathcal{N}_h^0} w_h(x_i)^- \leq C \left(\sum_{x_i \in \mathcal{C}_-(w_h)} |\partial \Gamma w_h(x_i)| \right)^{1/d}$$

Applying Corollary 2.6.3 (subdifferential vs discrete operator) to the convex function $\Gamma w_h(x_i)$ at a contact point $x_i \in \mathcal{C}_-(w_h)$ and recalling Remark 2.5.3, we have

$$|\partial \Gamma w_h(x_i)| \leq \delta^d \min_{\mathbf{v} \in \mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 \Gamma w_h(x_i; v_j) \leq \delta^d \min_{\mathbf{v} \in \mathbb{S}_\theta^\perp} \prod_{j=1}^{d} \nabla_\delta^2 w_h(x_i; v_j) = \delta^d T_\varepsilon[w_h](x_i),$$

where the last equality follows from Lemma 2.1.2 (discrete convexity). □

2.6.2 Concavity of the Discrete Operator

We recall concavity properties of $(\det A)^{1/d}$ for symmetric positive semi-definite matrices A and extend them to T_ε. The results can be traced back to [55, 60], but we present them here for completeness.

Lemma 2.6.5 (concavity of determinant)**.** *The following two statements are valid.*

(*i*) *For every symmetric positive semi-definite (SPSD) matrix A we have that*

$$(\det A)^{1/d} = \frac{1}{d} \inf \left\{ \mathrm{tr}(AB) \;\middle|\; B \text{ is SPD and } \det B = 1 \right\}$$

(*ii*) *The function $A \mapsto (\det A)^{1/d}$ is concave on SPSD matrices.*

Proof. We proceed in three steps.

Step 1: Proof of (i) for A invertible. Let B be SPD with $\det B = 1$. Then $B^{1/2}$ is well defined, $\det(B^{1/2}) = 1$ and we obtain

$$\det A = \det(B^{1/2} A B^{1/2}).$$

Let P be an orthogonal matrix that converts $B^{1/2}AB^{1/2}$ into a diagonal matrix D, namely $D = PB^{1/2}AB^{1/2}P^T$. Applying the geometric mean inequality yields

$$\det(B^{1/2}AB^{1/2})^{1/d} = (\det D)^{1/d} \leq \frac{1}{d}\mathrm{tr}D = \frac{1}{d}\mathrm{tr}(B^{1/2}AB^{1/2}) = \frac{1}{d}\mathrm{tr}(AB),$$

where we have used the invariance of the trace under cyclic permutations of the factor to write the last two equalities. This shows that

$$(\det A)^{1/d} \leq \frac{1}{d} \inf \left\{ \mathrm{tr}(AB) \;\middle|\; B \text{ is SPD and } \det B = 1 \right\}$$

This inequality is actually equality provided A is invertible. In fact, we can take $B = (\det A)^{1/d} A^{-1}$, which is SPD and $\det B = 1$. This proves (i) for A nonsingular.

Step 2: Proof of (i) for A singular. Given the singular value decomposition of A

$$A = \sum_{i=1}^{d} \lambda_i v_i \otimes v_i, \quad \lambda_1 \geq \cdots \lambda_k > \lambda_{k+1} = \cdots = \lambda_d = 0,$$

with orthogonal vectors $(v_i)_{i=1}^d$, we can assume that $k > 0$ for otherwise $A = 0$ and the assertion is trivial. Given a parameter $\sigma > 0$, let B be defined by

$$B := \sum_{i=1}^{k} \sigma v_i \otimes v_i + \sum_{i=k+1}^{d} \sigma^{-\beta} v_i \otimes v_i$$

and $\beta = k/(d-k)$ because then $\det B = \sigma^k \sigma^{-\beta(d-k)} = 1$. Therefore,

$$AB = \sigma \sum_{i=1}^{k} \lambda_i v_i \otimes v_i \quad \Rightarrow \quad \operatorname{tr}(AB) = \sigma \sum_{i=1}^{k} \lambda_i \to 0 \quad \text{as } \sigma \to 0,$$

which proves (i) for A singular since B is SPD.

Step 3: Proof of (ii). Let A and B be SPSD matrices and $0 \le \lambda \le 1$. Then $\lambda A + (1-\lambda)B$ is also SPSD and we can apply (i) to

$$\begin{aligned}
(\det[\lambda A + (1-\lambda)B])^{1/d} &= \frac{1}{d} \inf\left\{ \operatorname{tr}[(\lambda A + (1-\lambda)B)C] \,\middle|\, C \text{ is SPD and } \det C = 1 \right\} \\
&\ge \frac{\lambda}{d} \inf\left\{ \operatorname{tr}(AC) \,\middle|\, C \text{ is SPD and } \det C = 1 \right\} \\
&\quad + \frac{1-\lambda}{d} \inf\left\{ \operatorname{tr}(BC) \,\middle|\, C \text{ is SPD and } \det C = 1 \right\} \\
&= \lambda (\det A)^{1/d} + (1-\lambda)(\det B)^{1/d}.
\end{aligned}$$

This completes the proof. □

Upon relabeling $\widehat{A} = \lambda A$ and $\widehat{B} = (1-\lambda)B$, which are still SPSD, we can write Lemma 2.6.5 (ii) as follows:

$$(\det \widehat{A})^{1/d} + (\det \widehat{B})^{1/d} \le \left(\det(\widehat{A} + \widehat{B}) \right)^{1/d}. \tag{2.6.1}$$

We now show that our discrete operator $T_\varepsilon[\cdot]$ possesses a similar property.

Corollary 2.6.6 (concavity of discrete operator)**.** *Given two functions $u_h, w_h \in \mathbb{V}_h$, we have*

$$\left(T_\epsilon[u_h](x_i)\right)^{1/d} + \left(T_\epsilon[w_h](x_i)\right)^{1/d} \le \left(T_\epsilon[u_h + w_h](x_i)\right)^{1/d},$$

for all nodes $x_i \in \mathcal{N}_h^0$ *such that* $\nabla_\delta^2 u_h(x_i; v_j) \geq 0,\ \nabla_\delta^2 w_h(x_i; v_j) \geq 0$ *for all* $v_j \in \mathbb{S}_\theta$.

Proof. We argue in two steps.

Step 1. For $a = (a_j)_{j=1}^d \in \mathbb{R}^d$ with $a_j \geq 0,\ j = 1, \dots, d$ we consider the function

$$f(a) := \left(\prod_{j=1}^d a_j \right)^{1/d},$$

which can be conceived as the determinant of a diagonal (and thus symmetric) positive semi-definite matrix with diagonal elements $(a_j)_{j=1}^d$, i.e.

$$f(a) = \big(\det \operatorname{diag}\{a_1, \dots, a_d\} \big)^{1/d}.$$

Applying (2.6.1) to $\widehat{A} = \operatorname{diag}\{a_1, \dots, a_d\}, \widehat{B} = \operatorname{diag}\{b_1, \dots, b_d\}$ with $a = (a_j)_{j=1}^d, b = (b_j)_{j=1}^d \geq 0$ componentise, we deduce

$$f(a) + f(b) \leq f(a + b).$$

Step 2. We now apply this formula to the discrete operator. Since both u_h, w_h are discretely convex at $x_i \in \mathcal{N}_h^0$, so is $u_h + w_h$, and we can apply Lemma 2.1.2 (discrete convexity) to write

$$T_\varepsilon[u_h + w_h](x_i) = \prod_{j=1}^d \nabla_\delta^2 [u_h + w_h](x_i; v_j)$$

for a suitable $\mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}_\theta^\perp$. Making use again of (2.1.3), this time for u_h and w_h and for the specific set of directions $\mathbf{v}$ just found, we obtain

$$\begin{aligned} \big(T_\varepsilon[u_h](x_i)\big)^{\frac{1}{d}} + \big(T_\varepsilon[w_h](x_i)\big)^{\frac{1}{d}} &\leq \left(\prod_{j=1}^d \nabla_\delta^2 u_h(x_i; v_j) \right)^{\frac{1}{d}} + \left(\prod_{j=1}^d \nabla_\delta^2 w_h(x_i; v_j) \right)^{\frac{1}{d}} \\ &\leq \left(\prod_{j=1}^d \nabla_\delta^2 u_h(x_i; v_j) + \nabla_\delta^2 w_h(x_i; v_j) \right)^{\frac{1}{d}} = \big(T_\varepsilon[u_h + w_h](x_i)\big)^{\frac{1}{d}}, \end{aligned}$$

where the second inequality is given by Step 1 for $a = (\nabla^2_\delta u_h(x_i; v_j))_{j=1}^d$ and $b = (\nabla^2_\delta w_h(x_i; v_j))_{j=1}^d$. This is the asserted estimate. □

2.6.3 Continuous Dependence of the Two-Scale Method on Data

We are now ready to prove the continuous dependence of discrete solutions on data. This will be instrumental later for deriving rates of convergence for the two-scale method.

Proposition 2.6.7 (continuous dependence on data)**.** *Given two functions $u_h, w_h \in \mathbb{V}_h$ such that $u_h \geq w_h$ on $\partial\Omega_h$ and*

$$T_\varepsilon[u_h](x_i) = f_1(x_i) \geq 0 \quad \text{and} \quad T_\varepsilon[w_h](x_i) = f_2(x_i) \geq 0$$

at all interior nodes $x_i \in \mathcal{N}_h^0$, we have that

$$\max_{\Omega_h}(u_h - w_h)^- \leq C\ \delta \left(\sum_{x_i \in \mathcal{C}_-(u_h - w_h)} \left(f_1(x_i)^{1/d} - f_2(x_i)^{1/d} \right)^d \right)^{1/d}.$$

Proof. Since $u_h - w_h \in \mathbb{V}_h$ and $u_h - w_h \geq 0$ on $\partial\Omega_h$, Lemma 2.6.4 (stability) yields

$$\max_{x_i \in \mathcal{N}_h^0} (u_h - w_h)(x_i)^- \leq C\delta \left(\sum_{x_i \in \mathcal{C}_-(u_h - w_h)} T_\epsilon[u_h - w_h](x_i) \right)^{1/d}.$$

Since $x_i \in \mathcal{C}_-(u_h - w_h)$, we have that $\nabla^2_\delta(u_h - w_h)(x_i; v_j) \geq 0$, whence

$$\nabla^2_\delta u_h(x_i; v_j) \geq \nabla^2_\delta w_h(x_i; v_j) \geq 0 \quad \forall v_j \in \mathbb{S}_\theta,$$

where we have made use of Lemma 2.1.2 (discrete convexity). Invoking Corollary 2.6.6 (concavity of discrete operator) for $u_h - w_h$ and w_h, we deduce

$$\left(T_\epsilon[u_h - w_h](x_i) \right)^{1/d} \leq \left(T_\epsilon[u_h](x_i) \right)^{1/d} - \left(T_\epsilon[w_h](x_i) \right)^{1/d},$$

whence

$$\max_{x_i \in \mathcal{N}_h^0} (u_h - w_h)(x_i)^- \leq C\delta \left(\sum_{x_i \in \mathcal{C}_-(u_h - w_h)} \left(T_\epsilon[u_h](x_i)^{1/d} - T_\epsilon[w_h](x_i)^{1/d} \right)^d \right)^{1/d}$$
$$= C\delta \left(\sum_{x_i \in \mathcal{C}_-(u_h - w_h)} \left(f_1(x_i)^{1/d} - f_2(x_i)^{1/d} \right)^d \right)^{1/d}.$$

This completes the proof. □

2.7 Rates of Convergence

We now combine the preceding estimates to prove pointwise convergence rates for solutions with continuous Hessians, and either Hölder or Sobolev regularity, and later for a special case of viscosity solutions with discontinuous Hessians; these results require the nondegeneracy assumption $f \geq f_0 > 0$. We also deal with the degenerate case $f \geq 0$ and derive error estimates of reduced order. Our estimates do not require h small and are stated over the computational domain $\Omega_h \subset \Omega$. The two barrier functions introduced in Section 2.4.1 are critical for this analysis.

2.7.1 Error Estimates for Solutions with Hölder Regularity

We first deal with classical solutions u of (1.1.1) of class $C^{2+k,\alpha}(\overline{\Omega})$, with $k = 0, 1$ and $0 < \alpha \leq 1$, and derive pointwise error estimates. We proceed as follows. We first use Lemma 2.4.1 (discrete boundary barrier) to control $u_\varepsilon - \mathcal{I}_h u$ in the δ-neighborhood $\omega_{h,\delta}$ of $\partial\Omega_h$, where the consistency error of $T_\varepsilon[\mathcal{I}_h u]$ is of order one according to Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$). In the δ-interior region $\Omega_{h,\delta}$ we combine the interior consistency error of $T_\varepsilon[\mathcal{I}_h u]$ from Lemma 2.3.2 and Propo-

sition 2.6.7 (continuous dependence on data). Judicious choices of δ and θ in terms of h conclude the argument.

Theorem 2.7.1 (rates of convergence for classical solutions)**.** *Let $f(x) \geq f_0 > 0$ for all $x \in \Omega$. Let u be the classical solution of (1.1.1) and u_ε be the discrete solution of (2.1.2). If $u \in C^{2,\alpha}(\overline{\Omega})$ for $0 < \alpha \leq 1$ and*

$$\delta = R_0(u)\ h^{\frac{2}{2+\alpha}}, \quad \theta = R_0(u)^{-1}\ h^{\frac{\alpha}{2+\alpha}}$$

with $R_0(u) = |u|^{\frac{1}{2+\alpha}}_{W^2_\infty(\Omega)}\ |u|^{-\frac{1}{2+\alpha}}_{C^{2,\alpha}(\overline{\Omega})}$, then

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \leq C(\Omega, d, f_0)\Big(|u|^{\frac{1}{2+\alpha}}_{C^{2,\alpha}(\overline{\Omega})}\ |u|^{d-\frac{1}{2+\alpha}}_{W^2_\infty(\Omega)} + \big(1 + R_0(u)\big)\ |u|_{W^2_\infty(\Omega)}\Big)\ h^{\frac{\alpha}{2+\alpha}}.$$

Otherwise, if $u \in C^{3,\alpha}(\overline{\Omega})$ for $0 < \alpha \leq 1$ and

$$\delta = R_1(u)\ h^{\frac{2}{3+\alpha}}, \quad \theta = R_1(u)^{-1}\ h^{\frac{1+\alpha}{3+\alpha}}$$

with $R_1(u) := |u|^{\frac{1}{3+\alpha}}_{W^2_\infty(\Omega)}|u|^{-\frac{1}{3+\alpha}}_{C^{3,\alpha}(\overline{\Omega})}$, then

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \leq C(\Omega, d, f_0)\Big(|u|^{\frac{1}{3+\alpha}}_{C^{3,\alpha}(\overline{\Omega})}\ |u|^{d-\frac{1}{3+\alpha}}_{W^2_\infty(\Omega)} + \big(1 + R_1(u)\big)\ |u|_{W^2_\infty(\Omega)}\Big)\ h^{\frac{1+\alpha}{3+\alpha}}.$$

Proof. If $R_k(u) := |u|^{\frac{1}{2+k+\alpha}}_{W^2_\infty(\Omega)}|u|^{-\frac{1}{2+k+\alpha}}_{C^{2+k,\alpha}(\overline{\Omega})}$, $k = 0, 1$, we prove below the estimate

$$\max_{\Omega_h}(u_\varepsilon - \mathcal{I}_h u) \lesssim \Big(\big(1 + R_k(u)|u|_{W^2_\infty(\Omega)}\big) + |u|^{\frac{1}{2+k+\alpha}}_{C^{2+k,\alpha}(\overline{\Omega})}\ |u|^{d-\frac{1}{2+k+\alpha}}_{W^2_\infty(\Omega)}\Big)\ h^{\frac{k+\alpha}{2+k+\alpha}}$$

with a hidden constant depending on Ω, d, f_0. We proceed in three steps. The estimates for $\max_{\Omega_h}(\mathcal{I}_h u - u_\varepsilon)$ are similar and thus omitted. Adding the interpolation error $\|u - \mathcal{I}_h u\|_{L^\infty(\Omega_h)} \leq Ch^2|u|_{W^2_\infty(\Omega)}$ [21] readily gives the asserted estimates because $\frac{k+\alpha}{2+k+\alpha} \leq \frac{1}{2}$ for $k = 0, 1$ and $0 < \alpha \leq 1$.

Step 1: Boundary estimate. We show that for $z \in \mathcal{N}_h^0$ so that $\operatorname{dist}(z, \partial\Omega_h) \leq \delta$, we have:

$$u_\varepsilon(z) - \mathcal{I}_h u(z) \leq C|u|_{W^2_\infty(\Omega)}\delta.$$

Given the function p_h of Lemma 2.4.1 (discrete boundary barrier), for z fixed, we examine the behavior of $u_\varepsilon + p_h$. For any interior node $x_i \in \mathcal{N}_h^0$, we have

$$\begin{aligned}\prod_{j=1}^{d} \nabla^2_\delta(u_\varepsilon + p_h)(x_i; v_j) &= \prod_{j=1}^{d}(\nabla^2_\delta u_\varepsilon(x_i; v_j) + \nabla^2_\delta p_h(x_i; v_j)) \\ &\geq \prod_{j=1}^{d} \nabla^2_\delta u_\varepsilon(x_i; v_j) + \prod_{j=1}^{d} \nabla^2_\delta p_h(x_i; v_j) \quad \forall \mathbf{v} = (v_j)_{j=1}^d \in \mathbb{S}^\perp_\theta,\end{aligned}$$

because $\nabla^2_\delta u_\varepsilon(x_i; v_j) \geq 0$ and $\nabla^2_\delta p_h(x_i; v_j) \geq 0$. We apply Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) to obtain

$$\begin{aligned}T_\varepsilon[u_\varepsilon + p_h](x_i) &\geq T_\varepsilon[u_\varepsilon](x_i) + T_\varepsilon[p_h](x_i) \\ &\geq f(x_i) + E \\ &\geq T_\varepsilon[\mathcal{I}_h u](x_i) - C|u|^d_{W^2_\infty(\Omega)} + E \geq T_\varepsilon[\mathcal{I}_h u](x_i),\end{aligned}$$

provided $E \geq C|u|^d_{W^2_\infty(\Omega)}$. Since $\mathcal{I}_h u = u_\varepsilon$ and $p_h \leq 0$ on $\partial\Omega_h$, we deduce from Lemma 2.1.4 (discrete comparison principle) that

$$u_\varepsilon(z) + p_h(z) \leq \mathcal{I}_h u(z),$$

whence,

$$u_\varepsilon(z) - \mathcal{I}_h u(z) \leq C|u|_{W^2_\infty(\Omega)}\delta.$$

Step 2: Interior estimate. We show that for all $x_i \in \mathcal{N}_h^0$ so that $\operatorname{dist}(x_i, \partial\Omega_h) \geq \delta$ we have:

$$T_\varepsilon[u_\varepsilon](x_i) - T_\varepsilon[\mathcal{I}_h u](x_i) \leq C_1(u)\delta^{\alpha+k} + C_2(u)\left(\frac{h^2}{\delta^2} + \theta^2\right)$$

with $k = 0, 1$ and

$$C_1(u) = C|u|_{C^{2+k,\alpha}(\overline{\Omega})}\ |u|^{d-1}_{W^2_\infty(\Omega)}, \quad C_2(u) = C|u|^d_{W^2_\infty(\Omega)}$$

dictated by Lemma 2.3.2. Step 1 guarantees that

$$u_\varepsilon - \mathcal{I}_h u \leq C|u|_{W^2_\infty(\Omega)}\delta \quad \text{on } \partial\Omega_{h,\delta},$$

where $\Omega_{h,\delta}$ is defined in (2.3.1). Let $d_\varepsilon := \mathcal{I}_h u - u_\varepsilon + C|u|_{W^2_\infty(\Omega)}\delta$ and note that $d_\varepsilon \geq 0$ on $\partial\Omega_{h,\delta}$. We then apply Proposition 2.6.7 (continuous dependence on data) to d_ε in $\Omega_{h,\delta}$, in conjunction with Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$), to obtain

$$\max_{\Omega_{h,\delta}} d_\varepsilon^- \lesssim \delta \left(\sum_{x_i \in \mathcal{C}_-(d_\varepsilon)} \left((f(x_i) + e)^{1/d} - f(x_i)^{1/d} \right)^d \right)^{1/d}$$

with $e := C_1(u)\delta^{\alpha+k} + C_2(u)\left(\frac{h^2}{\delta^2} + \theta^2\right)$. We now use that the function $t \mapsto t^{1/d}$ is concave with derivative $\frac{1}{d}t^{1/d-1}$ and $f(x_i) \geq f_0 > 0$ to estimate

$$(f(x_i) + e)^{1/d} - f(x_i)^{1/d} \leq \frac{e}{d f_0^{\frac{d-1}{d}}},$$

whence

$$\max_{\Omega_{h,\delta}} d_\varepsilon^- \lesssim \delta \left(\sum_{x_i \in \mathcal{C}_-(d_\varepsilon)} \left(C_1(u)\delta^{\alpha+k} + C_2(u)\left(\frac{h^2}{\delta^2} + \theta^2\right)\right)^d \right)^{1/d}.$$

Since the cardinality of $\mathcal{C}_-(d_\varepsilon)$ is bounded by that of $\mathcal{N}_h$, which in turn is bounded by Ch^{-d} with C depending on shape regularity, we end up with

$$\max_{\Omega_h} (u_\varepsilon - \mathcal{I}_h u) \lesssim |u|_{W^2_\infty(\Omega)}\delta + \frac{\delta}{h}\left(C_1(u)\delta^{\alpha+k} + C_2(u)\left(\frac{h^2}{\delta^2} + \theta^2\right)\right).$$

Step 3: Choice of δ and θ. To find an optimal choice of δ and θ in terms of h, we minimize the right-hand side of the preceding estimate. We first set $\theta^2 = \frac{h^2}{\delta^2}$ and

equate the last two terms

$$C_1(u)\delta^{k+\alpha} = C_2(u)\frac{h^2}{\delta^2} \quad \Longrightarrow \quad \delta = R_k(u)h^{\frac{2}{2+k+\alpha}}.$$

Writing again $C_1(u)$ and $C_2(u)$ in terms of $|u|_{C^{2+k,\alpha}(\overline{\Omega})}$ and $|u|_{W^2_\infty(\Omega)}$, we thus obtain

$$\max_{\Omega_h}(u_\varepsilon - \mathcal{I}_h u) \lesssim R_k(u)|u|_{W^2_\infty(\Omega)}\ h^{\frac{2}{2+k+\alpha}} + |u|^{\frac{1}{2+k+\alpha}}_{C^{2+k,\alpha}(\overline{\Omega})}\ |u|^{d-\frac{1}{2+k+\alpha}}_{W^2_\infty(\Omega)}\ h^{\frac{k+\alpha}{2+k+\alpha}}.$$

Finally, the desired estimate follows immediately because $k+\alpha \le 2$.

□

2.7.2 Error Estimates for Solutions with Sobolev Regularity

We now derive error estimates for solutions $u \in W^s_p(\Omega)$ with $s > 2 + \frac{d}{p}$ so that $W^s_p(\Omega) \subset C^2(\overline{\Omega})$. We exploit the structure of the estimate of Proposition 2.6.7 (continuous dependence on data) which shows that its right-hand side accumulates in l^d rather than l^∞.

Theorem 2.7.2 (convergence rate for W^s_p solutions)**.** *Let $f \ge f_0 > 0$ in Ω and let the viscosity solution u of (1.1.1) be of class $W^s_p(\Omega)$ with $\frac{d}{p} < s-2-k \le 1,\ k = 0, 1$. If u_ε is the discrete solution of (2.1.2) and*

$$\delta = R(u)\ h^{\frac{2}{s}}, \quad \theta = R(u)^{-1}\ h^{1-\frac{2}{s}},$$

with $R(u) := |u|^{\frac{1}{s}}_{W^2_\infty(\Omega)}|u|^{-\frac{1}{s}}_{W^s_p(\Omega)}$, then

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \le C(d,\Omega,f_0)\Big(|u|^{\frac{1}{s}}_{W^s_p(\Omega)}\ |u|^{d-\frac{1}{s}}_{W^2_\infty(\Omega)} + \big(1+R(u)\big)|u|_{W^2_\infty(\Omega)}\Big)h^{1-\frac{2}{s}}.$$

Proof. We proceed as in Theorem 2.7.1 to show an upper bound for $u_\varepsilon - \mathcal{I}_h u$. The boundary estimate of Step 1 remains intact, namely

$$u_\varepsilon(z) - \mathcal{I}_h u(z) \le C\ |u|_{W^2_\infty(\Omega)}\ \delta$$

for all $z \in \mathcal{N}_h^0$ such that $\operatorname{dist}(z, \partial\Omega_h) \leq \delta$. On the other hand, Step 2 yields

$$\max_{\Omega_{h,\delta}}(u_\varepsilon - \mathcal{I}_h u) \lesssim \delta |u|_{W^2_\infty(\Omega)} + \delta \left(\sum_{x_i \in \mathcal{N}_h^0} C_1(u)^d \delta^{(k+\alpha)d} + C_2(u)^d \left(\frac{h^2}{\delta^2} + \theta^2 \right)^d \right)^{1/d},$$

where $C_1(u)$ and $C_2(u)$ are defined in Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) and $0 < \alpha = s - 2 - k - \frac{d}{p} \leq 1$ corresponds to the Sobolev embedding $W_p^s(B_i) \subset C^{2+k,\alpha}(B_i)$. In the following calculations we resort to the Sobolev inequality [45, Theorem 2.9]

$$|u|_{C^{2+k,\alpha}(B_i)} \leq C |u|_{W_p^s(B_i)},$$

involving only semi-norms. We stress that $C > 0$ depends on the Lipschitz constant of B_i but not on its size. The latter is due to the fact that the Sobolev numbers of $W_p^{s-2-k}(B_i)$ and $C^{0,\alpha}(B_i)$ coincide: $0 < s - k - 2 - d/p = \alpha \leq 1$. We refer to [45, Theorem 2.9] for a proof for $0 < s < 1$. We now use the Hölder inequality with exponent $\frac{p}{d} > 1$ to obtain

$$\begin{aligned}\left(\sum_{x_i \in \mathcal{N}_h^0} C_1(u)^d \right)^{\frac{1}{d}} &\lesssim \left(\sum_{x_i \in \mathcal{N}_h^0} |u|^d_{W_p^s(B_i)} |u|^{d(d-1)}_{W^2_\infty(B_i)} \right)^{\frac{1}{d}} \\ &\lesssim \left(\sum_{x_i \in \mathcal{N}_h^0} |u|^{d\frac{p}{d}}_{W_p^s(B_i)} \right)^{\frac{1}{d}\frac{d}{p}} \left(\sum_{x_i \in \mathcal{N}_h^0} |u|^{d(d-1)\frac{p}{p-d}}_{W^2_\infty(B_i)} \right)^{\frac{1}{d}\frac{p-d}{p}}.\end{aligned}$$

Since the cardinality of the set of balls B_i containing an arbitrarily given $x \in \Omega$ is proportional to $\left(\frac{\delta}{h}\right)^d$, while the cardinality of $\mathcal{N}_h^0$ is proportional to h^{-d}, we get

$$\begin{aligned}\left(\sum_{x_i \in \mathcal{N}_h^0} C_1(u)^d \right)^{\frac{1}{d}} &\lesssim \left(\frac{\delta}{h} \right)^{\frac{d}{p}} |u|_{W_p^s(\Omega)} \left(h^{-d} |u|^{\frac{d(d-1)p}{p-d}}_{W^2_\infty(\Omega)} \right)^{\frac{p-d}{pd}} \\ &\lesssim \frac{\delta^{\frac{d}{p}}}{h} |u|_{W_p^s(\Omega)} |u|^{d-1}_{W^2_\infty(\Omega)}.\end{aligned}$$

Exploiting that $\alpha + k + \frac{d}{p} + 1 = s - 1$, we readily arrive at

$$\delta \left(\sum_{x_i \in \mathcal{N}_h^0} C_1(u)^d \, \delta^{(k+\alpha)d} \right)^{\frac{1}{d}} \lesssim \frac{\delta^{s-1}}{h} |u|_{W_p^s(\Omega)} |u|^{d-1}_{W^2_\infty(\Omega)}.$$

In addition, we have

$$\left(\sum_{x_i \in \mathcal{N}_h^0} C_2(u)^d \right)^{\frac{1}{d}} \lesssim |u|^d_{W^2_\infty(\Omega)} \, \frac{1}{h},$$

whence

$$\delta \left(\sum_{x_i \in \mathcal{N}_h^0} C_2(u)^d \left(\frac{h^2}{\delta^2} + \theta^2 \right)^d \right)^{\frac{1}{d}} \lesssim |u|^d_{W^2_\infty(\Omega)} \, \frac{\delta}{h} \left(\frac{h^2}{\delta^2} + \theta^2 \right).$$

Collecting the previous estimates, we end up with

$$\max_{\Omega_h} (u_\varepsilon - \mathcal{I}_h u) \lesssim \delta |u|_{W^2_\infty(\Omega)} + |u|^{d-1}_{W^2_\infty(\Omega)} \frac{\delta}{h} \left(|u|_{W^s_p(\Omega)} \delta^{s-2} + |u|_{W^2_\infty(\Omega)} \left(\frac{h^2}{\delta^2} + \theta^2 \right) \right).$$

To find an optimal relation among h, δ and θ, we first choose $\theta^2 = \frac{h^2}{\delta^2}$ and next equate the two terms in the second summand to obtain

$$\delta = R(u) \; h^{\frac{2}{s}}, \quad \theta = R(u)^{-1} \; h^{1-\frac{2}{s}},$$

whence

$$\max_{\Omega_h} (u_\varepsilon - \mathcal{I}_h u) \lesssim R(u) |u|_{W^2_\infty(\Omega)} h^{\frac{2}{s}} + |u|^{\frac{1}{s}}_{W^s_p(\Omega)} \; |u|^{d-\frac{1}{s}}_{W^2_\infty(\Omega)} h^{1-\frac{2}{s}}.$$

Adding the interpolation error estimate $\|u - \mathcal{I}_h u\|_{L^\infty(\Omega)} \lesssim h^2 |u|_{W^2_\infty(\Omega)}$, and using that $2 > \frac{2}{s} \geq 1 - \frac{2}{s}$ for $2 < s \leq 4$, leads to the asserted estimate. □

The error estimate of Theorem 2.7.2 (convergence rate for W^s_p-solutions) is of order $\frac{1}{2}$ for $s = 4$ and $u \in W^4_p(\Omega)$ with $p > d$. This rate requires much weaker regularity than the corresponding error estimate in Theorem 2.7.1, namely $u \in C^{3,1}(\overline{\Omega}) = W^4_\infty(\Omega)$. In both cases, the relation between δ and h is $\delta \approx h^{\frac{1}{2}}$.

2.7.3 Error Estimates for Piecewise Smooth Solutions

We now derive pointwise rates of convergence for a larger class of solutions than in Section 2.7.2. These are viscosity solutions which are piecewise W^s_p but have dis-

continuous Hessians across a Lipschitz $(d-1)$-dimensional manifold $\mathcal{S}$. Since $T_\varepsilon[\mathcal{I}_h u]$ has a consistency error of order one in a δ-region around $\mathcal{S}$, due to the discontinuity of $D^2 u$, we exploit the fact that the measure of this region is proportional to $\delta|\mathcal{S}|$. We are thus able to adapt the argument of Theorem 2.7.2 (convergence rate for W_p^s solutions), and accumulate such consistency error in l^d, at the expense of an extra additive term of order $h^{-1}\delta^{1+\frac{1}{d}}$. This term is responsible for a reduced convergence rate when $u \in W_p^s(\Omega \setminus \mathcal{S})$, $s > 2 + \frac{1}{d}$.

Theorem 2.7.3 (convergence rate for piecewise smooth solutions)**.** *Let $\mathcal{S}$ denote a $(d-1)$-dimensional Lipschitz manifold that divides Ω into two disjoint subdomains Ω_1, Ω_2 so that $S = \overline{\Omega}_1 \cap \overline{\Omega}_2$. Let $f \geq f_0 > 0$ in Ω and let $u \in W_p^s(\Omega_i) \cap W_\infty^2(\Omega)$, for $i = 1, 2$ and $\frac{d}{p} < s - 2 - k \leq 1, k = 0, 1$, be the viscosity solution of (1.1.1). If u_ε denotes the discrete solution of (2.1.2), then for $\beta = \min\{s, 2 + \frac{1}{d}\}$ we have*

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \leq C(d, \Omega, f_0) \left(R(u)^{-1} |u|_{W_\infty^2(\Omega)}^d + \big(1 + R(u)\big) |u|_{W_\infty^2(\Omega)} \right) h^{1-\frac{2}{\beta}},$$

with $R(u) = \left(\frac{|u|_{W_\infty^2(\Omega)}}{|u|_{W_p^s(\Omega\setminus\mathcal{S})} + |u|_{W_\infty^2(\Omega)}} \right)^{\frac{1}{\beta}}$ *and* $|u|_{W_p^s(\Omega\setminus\mathcal{S})} := \max_i |u|_{W_p^s(\Omega_i)}$, *provided*

$$\delta = R(u)\ h^{\frac{2}{\beta}}, \quad \theta = R(u)^{-1}\ h^{1-\frac{2}{\beta}}.$$

Proof. We proceed as in Theorems 2.7.1 and 2.7.2. The boundary layer estimate relies on the regularity $u \in W_\infty^2(\Omega)$ which is still valid, whence for all $x \in \Omega_h$ such that $\operatorname{dist}(x, \partial\Omega_h) \leq \delta$ we obtain

$$u_\varepsilon(x) - \mathcal{I}_h u(x) \leq C |u|_{W_\infty^2(\Omega)} \delta.$$

Consider now the internal layer

$$\mathcal{S}_h^\delta := \{x \in \Omega_h : \ \operatorname{dist}(x, \mathcal{S}) \leq \delta\},$$

which is the region affected by the discontinuity of the Hessian D^2u. Recall the auxiliary function $d_\varepsilon = \mathcal{I}_h u - u_\varepsilon + C|u|_{W^2_\infty(\Omega)}\delta$ of Theorem 2.7.1 (rates of convergence for classical solutions) and split the contact set $\mathcal{C}^\delta_-(d_\varepsilon) := \mathcal{C}_-(d_\varepsilon) \cap \Omega_{h,\delta}$ as follows:

$$\mathcal{S}^\delta_{h,1} := \mathcal{C}^\delta_-(d_\varepsilon) \cap \mathcal{S}^\delta_h, \quad \mathcal{S}^\delta_{h,2} := \mathcal{C}^\delta_-(d_\varepsilon) \setminus \mathcal{S}^\delta_h.$$

An argument similar to Step 2 (interior estimate) of Theorem 2.7.1, based on combining Proposition 2.6.7 (continuous dependence on data) and Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) with assumption $f \geq f_0 > 0$, yields

$$\begin{aligned}\max_{\Omega_{h,\delta}} d^-_\varepsilon \lesssim \delta &\left(\sum_{x_i \in \mathcal{S}^\delta_{h,1}} C_2(u)^d \right)^{1/d} \\ &+ \delta \left(\sum_{x_i \in \mathcal{S}^\delta_{h,2}} C_1(u)^d \delta^{(k+\alpha)d} + C_2(u)^d \left(\frac{h^2}{\delta^2} + \theta^2 \right)^d \right)^{1/d} =: I_1 + I_2,\end{aligned}$$

because the consistency error in $\mathcal{S}^\delta_{h,1}$ is bounded by $C_2(u) = C|u|^d_{W^2_\infty(B_i)}$. As in Theorem 2.7.2 (convergence rate for W^s_p solutions), $C_1(u)$ satisfies

$$C_1(u) \lesssim |u|_{W^s_p(B_i)} |u|^{d-1}_{W^2_\infty(B_i)}.$$

Since the number of nodes $x_i \in \mathcal{S}^\delta_{h,1}$ is bounded by $C|\mathcal{S}|\delta h^{-d}$, we deduce

$$I_1 \lesssim \delta \left(\sum_{x_i \in \mathcal{S}^\delta_{h,1}} C_2(u)^d \right)^{1/d} \lesssim |u|^d_{W^2_\infty(\Omega)} \frac{\delta^{1+\frac{1}{d}}}{h}.$$

For I_2 we distinguish whether x_i belongs to Ω_1 or Ω_2 and accumulate $C_1(u)$ in ℓ^p, exactly as in Theorem 2.7.2, to obtain

$$I_2 \lesssim |u|^{d-1}_{W^2_\infty(\Omega)} \left(|u|_{W^s_p(\Omega\setminus\mathcal{S})} \frac{\delta^{s-1}}{h} + |u|_{W^2_\infty(\Omega)} \frac{\delta}{h} \left(\frac{h^2}{\delta^2} + \theta^2 \right) \right).$$

Collecting the previous estimates and using the definition of β yields

$$\max_{\Omega_h}(u_\varepsilon - \mathcal{I}_h u) \lesssim |u|_{W^2_\infty(\Omega)}\delta + |u|^{d-1}_{W^2_\infty(\Omega)}\frac{\delta}{h}\left(\left(|u|_{W^s_p(\Omega\setminus\mathcal{S})} + |u|_{W^2_\infty(\Omega)}\right)\delta^{\beta-2} + |u|_{W^2_\infty(\Omega)}\left(\frac{h^2}{\delta^2}+\theta^2\right)\right).$$

We finally realize that this estimate is similar to that in the proof of Theorem 2.7.2 except for the middle term on the right-hand side. Therefore, we proceed as in Theorem 2.7.2 to find the relation between δ, θ and h, add the estimate $\|u - \mathcal{I}_h u\|_{L^\infty(\Omega)} \lesssim h^2|u|_{W^2_\infty(\Omega)}$, and eventually derive the asserted error estimate. □

2.7.4 Error Estimates for Piecewise Smooth Solutions with Degenerate f

We observe that in all three preceding theorems we assume that $f \geq f_0 > 0$. This is an important assumption in the proofs, since it allows us to use the concavity of $t \mapsto t^{1/d}$ and Proposition 2.6.7 (continuous dependence on data) to obtain

$$(f(x_i)+e)^{1/d} - f(x_i)^{1/d} \leq \frac{e}{d f_0^{\frac{d-1}{d}}}, \tag{2.7.1}$$

where e is related to the consistency of the operator in Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$). We see that this is only possible if $f_0 > 0$. If we allow f to touch zero, then (2.7.1) reduces to

$$(f(x_i)+e)^{1/d} - f(x_i)^{1/d} \leq e^{1/d}, \tag{2.7.2}$$

with equality for $f(x_i) = 0$. This leads to a rate of order $\left(\frac{\delta}{h}\right)^{1-\frac{2}{d}} \geq 1$ for $d \geq 2$. To circumvent this obstruction, we use Lemma 2.4.2 (interior barrier function) which

allows us to introduce an extra parameter $\sigma > 0$ that compensates for the lack of lower bound $f_0 > 0$ and yields pointwise error estimates of reduced order.

Theorem 2.7.4 (degenerate forcing $f \geq 0$)**.** *Let $\mathcal{S}$ denote a $(d-1)$-dimensional Lipschitz manifold that divides Ω into two disjoint subdomains Ω_1, Ω_2 such that $\mathcal{S} = \overline{\Omega}_1 \cap \overline{\Omega}_2$. Let $f \geq 0$ in Ω and let $u \in W_p^s(\Omega_i) \cap W_\infty^2(\Omega)$, for $i = 1, 2$ and $\frac{d}{p} < s - 2 - k \leq 1, k = 0, 1$, be the viscosity solution of (1.1.1). If u_ε denotes the discrete solution of (2.1.2), then for $\beta = \min\{s, 2 + \frac{1}{d}\}$ we have*

$$\|u - u_\varepsilon\|_{L^\infty(\Omega_h)} \leq C(d, \Omega)|u|_{W_\infty^2(\Omega)}\Big(1 + R(u) + R(u)^{-\frac{1}{d}}\Big)\ h^{\frac{1}{d}\left(1-\frac{2}{\beta}\right)}$$

with $R(u) = \left(\frac{|u|_{W_\infty^2(\Omega)}}{|u|_{W_p^s(\Omega\setminus\mathcal{S})} + |u|_{W_\infty^2(\Omega)}}\right)^{\frac{1}{\beta}}$ *and* $|u|_{W_p^s(\Omega\setminus\mathcal{S})} := \max_i |u|_{W_p^s(\Omega_i)}$*, provided*

$$\delta = R(u)\ h^{\frac{2}{\beta}}, \quad \theta = R(u)^{-1}\ h^{1-\frac{2}{\beta}}.$$

Proof. We employ the interior barrier function q_h of Lemma 2.4.2 scaled by a parameter $\sigma > 0$ to control $u_\varepsilon - \mathcal{I}_h u$ and $\mathcal{I}_h u - u_\varepsilon$ in two steps. The parameter σ allows us to mimic the calculation in (2.7.1). In the third step we choose σ optimally with respect to the scales of our scheme.

Step 1: Upper bound for $u_\varepsilon - \mathcal{I}_h u$. We let $w_h := u_\varepsilon + \sigma q_h$ and $v_h := \mathcal{I}_h u + C|u|_{W_\infty^2(\Omega)}\delta$, observe that $T_\varepsilon[w_h](x_i) \geq f(x_i) + \sigma^d$, and proceed as in Step 1 of Theorem 2.7.1 to show $w_h(z) \leq v_h(z)$ for all $z \in \mathcal{N}_h^0$ such that $\operatorname{dist}(z, \partial\Omega_h) \leq \delta$.

We now focus on $\Omega_{h,\delta}$ and define the auxiliary function $d_\varepsilon := v_h - w_h$ and contact set $\mathcal{C}_-^\delta(d_\varepsilon) := \mathcal{C}_-(d_\varepsilon) \cap \Omega_{h,\delta}$. Since the previous argument guarantees that $d_\varepsilon \geq 0$ on $\partial\Omega_{h,\delta}$, Proposition 2.6.7 (continuous dependence on data) gives

$$\max_{\Omega_{h,\delta}} d_\varepsilon^- \lesssim \delta \left(\sum_{x_i \in \mathcal{C}_-^\delta(d_\varepsilon)} \Big(\big(T_\varepsilon[v_h](x_i)\big)^{1/d} - \big(T_\varepsilon[w_h](x_i)\big)^{1/d} \Big)^d \right)^{1/d}.$$

If e_i is the local consistency error given in Lemma 2.3.2, we further note that

$$T_\varepsilon[v_h](x_i) \le f(x_i) + e_i, \quad T_\varepsilon[w_h](x_i) \ge T_\varepsilon[u_\varepsilon](x_i) + T_\varepsilon[\sigma q_h](x_i) \ge f(x_i) + \sigma^d$$

for all $x_i \in \mathcal{N}_h^0$, whence

$$\max_{\Omega_{h,\delta}} d_\varepsilon^- \lesssim \delta \left(\sum_{x_i \in \mathcal{C}_-^\delta(d_\varepsilon)} \left(\big(f(x_i) + e_i\big)^{1/d} - \big(f(x_i) + \sigma^d\big)^{1/d} \right)^d \right)^{1/d}.$$

We now observe that $e_i \ge \sigma^d$ for all $x_i \in \mathcal{C}_-^\delta(d_\varepsilon)$ because all terms in the above sum are non-negative. If there is no such x_i, then the above bound implies that $d_\varepsilon^- = 0$ and $w_h \le v_h$, whence $u_\varepsilon - \mathcal{I}_h u \lesssim \sigma + |u|_{W^2_\infty(\Omega)}\delta$. Otherwise, the above observation combined with (2.7.1) and $f(x_i) \ge 0$ implies

$$\begin{aligned} &\big(f(x_i) + e_i\big)^{1/d} - \big(f(x_i) + \sigma^d\big)^{1/d} \\ &\qquad = \big(f(x_i) + \sigma^d + (e_i - \sigma^d)\big)^{1/d} - \big(f(x_i) + \sigma^d\big)^{1/d} \\ &\qquad \le \frac{e_i - \sigma^d}{d\sigma^{d\frac{d-1}{d}}} \le d^{-1}\sigma^{1-d} e_i. \end{aligned}$$

We next proceed exactly as in Theorem 2.7.3 (convergence rate for piecewise smooth solutions) to derive an upper bound for d_ε^-, but with the additional factor σ^{1-d}. Employing the definition of d_ε, we thereby obtain

$$u_\varepsilon - \mathcal{I}_h u \lesssim \sigma + |u|_{W^2_\infty(\Omega)}\delta + \sigma^{1-d}\,\frac{\delta}{h}\left(C_1(u)\delta^{s-2} + C_2(u)\,\left(\delta^{1/d} + \frac{h^2}{\delta^2} + \theta^2\right)\right),$$

where $C_1(u) = C|u|_{W^s_p(\Omega\setminus\mathcal{S})}|u|^{d-1}_{W^2_\infty(\Omega)}$ and $C_2(u) = C|u|^d_{W^2_\infty(\Omega)}$.

Step 2: Lower bound for $u_\varepsilon - \mathcal{I}_h u$. To prove the reverse inequality, we proceed as in Step 1, except that this time we define $v_h := u_\varepsilon + C|u|_{W^2_\infty(\Omega)}\delta$ and $w_h := \mathcal{I}_h u + \sigma q_h$. An argument similar to Step 1 yields $w_h \le v_h$ in $\omega_{h,\delta}$. Moreover, recalling Lemma 2.3.2 (consistency of $T_\varepsilon[\mathcal{I}_h u]$) we have for all $x_i \in \mathcal{N}_h^0$

$$T_\varepsilon[v_h](x_i) = f(x_i) \le T_\varepsilon[\mathcal{I}_h u](x_i) + e_i, \quad T_\varepsilon[w_h](x_i) \ge T_\varepsilon[\mathcal{I}_h u](x_i) + \sigma^d,$$

where e_i is a local bound for the consistency error. Combining this with Proposition 2.6.7 (continuous dependence on data) in $\Omega_{h,\delta}$ gives

$$\max_{\Omega_{h,\delta}} d_\varepsilon^- \lesssim \delta \left(\sum_{x_i \in \mathcal{C}_-^\delta(d_\varepsilon^-)} \left(\left(T_\varepsilon[\mathcal{I}_h u](x_i) + e_i\right)^{\frac{1}{d}} - \left(T_\varepsilon[\mathcal{I}_h u](x_i) + \sigma^d\right)^{\frac{1}{d}} \right)^d \right)^{\frac{1}{d}},$$

Since $\mathcal{I}_h u$ is discretely convex, we apply Lemma 2.1.2 (discrete convexity) to deduce $T_\varepsilon[\mathcal{I}_h u](x_i) \geq 0$ and next argue as in Step 1 to obtain

$$\mathcal{I}_h u - u_\varepsilon \lesssim \sigma + |u|_{W^2_\infty(\Omega)}\delta + \sigma^{1-d}\,\frac{\delta}{h}\left(C_1(u)\delta^{s-2} + C_2(u)\,\left(\delta^{1/d} + \frac{h^2}{\delta^2} + \theta^2\right)\right).$$

Step 3: Choice of δ, θ and σ. Since $\|u - \mathcal{I}_h u\|_{L^\infty(\Omega_h)} \leq C|u|_{W^2_\infty(\Omega)}h^2$, combining Steps 1 and 2 yields

$$\begin{aligned}\|u_\varepsilon - u\|_{L^\infty(\Omega_h)} &\lesssim \sigma + |u|_{W^2_\infty(\Omega)}(\delta + h^2) \\ &\quad + \sigma^{1-d}\,\frac{\delta}{h}\left(C_1(u)\delta^{s-2} + C_2(u)\,\left(\delta^{1/d} + \frac{h^2}{\delta^2} + \theta^2\right)\right).\end{aligned}$$

We now minimize the right-hand side upon choosing δ, θ and σ suitably with respect to h. We first recall the definition of β and choose δ and θ as in Theorem 2.7.3. At this stage it only remains to find σ upon solving

$$\sigma = C_2(u)\sigma^{1-d}\frac{h}{\delta} = C\sigma^{1-d}|u|^d_{W^2_\infty(\Omega)}R(u)^{-1}\; h^{1-\frac{2}{\beta}},$$

which leads to

$$\sigma = |u|_{W^2_\infty(\Omega)}R(u)^{-\frac{1}{d}}\; h^{\frac{1}{d}\left(1-\frac{2}{\beta}\right)}.$$

Since $\beta > 2$ we get $h^2 + \delta \leq \left(1 + R(u)^{-\frac{1}{\beta}}\right)h^{\frac{2}{\beta}}$ and

$$\|u_\varepsilon - u\|_{L^\infty(\Omega_h)} \lesssim |u|_{W^2_\infty(\Omega)}\left(1 + R(u)\right)h^{\frac{2}{\beta}} + |u|_{W^2_\infty(\Omega)}R(u)^{-\frac{1}{d}}\; h^{\frac{1}{d}\left(1-\frac{2}{\beta}\right)}.$$

This yields the asserted estimate and finishes the proof. □

Theorem 2.7.4 is an extension of Theorem 2.7.3 to the degenerate case $f \geq 0$, but the same techniques and estimates extend as well to Theorems 2.7.1 and 2.7.2.

2.8 Numerical Experiments

We present three examples in the square domain $\Omega = \Omega_h = [0,1]^2$. The fact that Ω is not uniformly convex does not affect the existence of our discrete solution u_ε, as the Dirichlet datum g is the trace of a convex function; however this is beyond the assumptions of the convergence theory. We implement the two-scale method within the MATLAB software FELICITY [79, 80]. We first consider two examples with smooth Hessian and with discontinuous Hessian, and observe linear experimental rates of convergence with respect to h, which are better than predicted by the theory in Sections 2.7.1 (error estimates for solutions with Hölder regularity) and 2.7.4 (error estimates for piecewise smooth solutions with degenerate f). The third example entails an unbounded right hand side f and is not guaranteed to converge by theory. We still observe convergence experimentally.

2.8.1 Semi-Smooth Newton Method

We solve the nonlinear algebraic equation (2.1.2) via a damped semi-smooth Newton iteration. Let $\mathbf{z} := (z_h(x_i))_{i=1}^N \in \mathbb{R}^N$ stand for the vector of nodal values of a generic $z_h \in \mathbb{V}_h$; thus N is the cardinality of $\mathcal{N}_h$. If $\mathbf{u}_n = (u_\varepsilon^n(x_i))_{i=1}^N$, $\mathbf{DT}_\varepsilon[\mathbf{u}_n]$ is the Jacobian matrix of the nonlinear map $\mathbf{T}_\varepsilon : \mathbb{R}^N \to \mathbb{R}^N$ at $\mathbf{u}_n$, and $\mathbf{f}_n = (f(x_i))_{i=1}^N$,

then a Newton increment is given by

$$\mathbf{DT}_\varepsilon[\mathbf{u}_n]\ \mathbf{w}_n = \mathbf{f}_n - \mathbf{T}_\varepsilon[\mathbf{u}_n]$$

and the n-th Newton step by $\mathbf{u}_{n+1} = \mathbf{u}_n + \tau \mathbf{w}_n$, where the damping parameter $\tau \in (0, 1]$, which might depend on n, satisfies

$$\|f - T_\varepsilon[u_\varepsilon^n + \tau w_n]\|_{L^2(\Omega)} < \|f - T_\varepsilon[u_\varepsilon^n]\|_{L^2(\Omega)}.$$

We now explain the construction of $\mathbf{DT}_\varepsilon[\mathbf{u}_n]$. Evaluating $\nabla_\delta^2 z_h(x_i; v_j)$, in view of (2.1.1), requires knowing z_h at $x_i^\pm = x_i \pm \rho\delta v_j$, which are not necessarily nodes of $\mathcal{N}_h$. Since $x_i^\pm$ belong to two simplices of $\mathcal{T}_h$, and $z_h \in \mathbb{V}_h$, the values $z_h(x_i^\pm)$ can be determined in terms of the barycentric coordinates of $x_i^\pm$. Therefore, if we define $\boldsymbol{\nabla}_\delta^2 \mathbf{z}(i; v_j) := \nabla_\delta^2 z_h(x_i; v_j)$, then we realize that this operator involves $2(d+1)+1$ components of $\mathbf{z}$ and is thus sparse. We likewise define $\boldsymbol{\nabla}_\delta^{2,+}$ and $\boldsymbol{\nabla}_\delta^{2,-}$ to be the component-wise versions of $\nabla_\delta^{2,+}$ and $\nabla_\delta^{2,-}$. The operator $\mathbf{T}_\varepsilon$ reads

$$\mathbf{T}_\varepsilon[\mathbf{z}](i) := \min_{\mathbf{v}=(v_j)_{j=1}^d \in \mathbb{S}_\theta^\perp} \left(\prod_{j=1}^d \boldsymbol{\nabla}_\delta^{2,+} \mathbf{z}(i; v_j) - \sum_{j=1}^d \boldsymbol{\nabla}_\delta^{2,-} \mathbf{z}(i; v_j) \right) = f(x_i), \qquad (2.8.1)$$

according to (2.1.4).

Let now $\mathbf{v}^i = (v_j^i)_{j=1}^d \in \mathbb{S}_\theta^\perp$ be a set of directions that realize the minimum of $\mathbf{T}_\varepsilon[\mathbf{u}_n](i)$ and denote $\mathbf{V} := (\mathbf{v}^i)_{i=1}^N \in \mathbb{R}^{d\times d\times N}$, the collection of the minimizing d-tuples $\mathbf{v}^i$ for all $i = 1, \dots N$. Combining the above notations, we denote the matrix that contains the j-th minimizing directions for each node by $\mathbf{V}_j \in \mathbb{R}^{d\times N}$. This allows us to display our Jacobian in a vectorized form, using the notation $\boldsymbol{\nabla}_\delta^2 \mathbf{u}_n(\mathbf{V}_j) := (\nabla_\delta^2 \mathbf{u}_n(i; v_j^i))_{i=1}^N$, since (2.8.1) gives for $\mathbf{z} = \mathbf{u}_n$:

$$\mathbf{T}_\varepsilon[\mathbf{u}_n] = \bigodot_{j=1}^d \boldsymbol{\nabla}_\delta^{2,+} \mathbf{u}_n(\mathbf{V}_j) - \sum_{j=1}^d \boldsymbol{\nabla}_\delta^{2,-} \mathbf{u}_n(\mathbf{V}_j),$$

where $\odot$ stands for the component-wise multiplication of vectors. Using Danskin's Theorem [15] and the product rule, we can then obtain $\mathbf{DT}_\varepsilon[\mathbf{u}_n]\mathbf{w}_n$. For that, we need to differentiate $\boldsymbol{\nabla}_\delta^{2,+}\mathbf{u}_n(\mathbf{V}_j)$. We observe that for each component, $\boldsymbol{\nabla}_\delta^{2,+}\mathbf{u}_n(i;v_j^i) = \max\{\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(i;v_j^i),0\}$ is not differentiable at $\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(i;v_j^i)=0$. As a result, we use the so-called slant derivative in the direction $\mathbf{w}_n$ [26, 49], in order to compute:

$$\mathbf{D}[\boldsymbol{\nabla}_\delta^{2,+}\mathbf{u}_n(\mathbf{V}_j)]\mathbf{w}_n = \mathbf{H}^+[\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(\mathbf{V}_j)]\odot\boldsymbol{\nabla}_\delta^2\mathbf{w}_n(\mathbf{V}_j),$$

each component of which is equal to

$$(\mathbf{D}[\boldsymbol{\nabla}_\delta^{2,+}\mathbf{u}_n(\mathbf{V}_j)]\mathbf{w}_n)_i = \begin{cases} \boldsymbol{\nabla}_\delta^2\mathbf{w}_n(i;v_j^i) & \text{if } \boldsymbol{\nabla}_\delta^2\mathbf{u}_n(i;v_j^i)>0 \\ 0 & \text{if } \boldsymbol{\nabla}_\delta^2\mathbf{u}_n(i;v_j^i)\le 0. \end{cases}$$

Here $\mathbf{H}^+$ is the operator that assigns 1 to a strictly positive component and zero otherwise. Similarly, $\mathbf{D}[\boldsymbol{\nabla}_\delta^{2,-}\mathbf{u}_n(\mathbf{V}_j)]\mathbf{w}_n = \mathbf{H}^-[\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(\mathbf{V}_j)]\odot\boldsymbol{\nabla}_\delta^2\mathbf{w}_n(\mathbf{V}_j)$ where $\mathbf{H}^-$ assigns -1 to a non-positive component and 0 otherwise. We are now ready to employ Danskin's Theorem [15] and the product rule to obtain similarly to [43]:

$$\mathbf{DT}_\varepsilon[\mathbf{u}_n]\mathbf{w}_n = \sum_{j=1}^{d}\boldsymbol{\nabla}_\delta^2\mathbf{w}_n(\mathbf{V}_j)\bigodot\Big(\mathbf{H}^+[\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(\mathbf{V}_j)]\bigodot_{k\neq j}\boldsymbol{\nabla}_\delta^{2,+}\mathbf{u}_n(\mathbf{V}_k)-\mathbf{H}^-[\boldsymbol{\nabla}_\delta^2\mathbf{u}_n(\mathbf{V}_j)]\Big),$$

The presence of both operators $\mathbf{H}^+$ and $\mathbf{H}^-$ enforces discrete convexity, and their definition at zero yields non-singular Jacobians computationally. This flexibility in choosing $\mathbf{H}^+$ and $\mathbf{H}^-$ with vanishing argument is consistent with the definition of the slant derivative for the max and min functions [49].

We initialize the Newton iteration with $\mathbf{u}_0$ corresponding to the Galerkin solution in $\mathbb{V}_h$ to the auxiliary problem $\Delta u_0 = (d!f)^{1/d}$ in Ω and $u_0 = g$ on $\partial\Omega$, as

proposed in [43], but only for the coarser mesh $h = 2^{-5}$. For all subsequent refinements we interpolate the discrete solution in the previous coarse mesh and use it as initial guess. This greatly improves the residual error and leads to minimal or no damping.

2.8.2 Accuracy

We examine the performance of our two-scale method mainly with two examples, with smooth and discontinuous Hessians; a third example entails an unbounded f. For the first two examples we choose $\delta = h^\alpha$ and $\theta = h^\beta$ for appropriate $\alpha, \beta > 0$ which yield provable rates of convergence according to Section 2.7. We stress that smaller values of θ lead to similar convergence rates but affect the sparsity pattern of the matrix in the semi-smooth Newton iteration because the number of search directions within $\mathbb{S}_\theta$ increases. We thus choose θ consistent with the theory. The computation of ρ in (2.1.1) is exact, because Ω_h is a square, although need not be in general. We stop the Newton iterations when

$$\|f - T_\varepsilon[u_\varepsilon^{n+1}]\|_{L^2(\Omega)} < 10^{-8}\|f - T_\varepsilon[u_\varepsilon^0]\|_{L^2(\Omega)}.$$

Smooth Hessian: We choose the solution u and forcing f to be

$$u(x) = e^{|x|^2/2}, \quad f(x) = (1 + |x|^2)e^{|x|^2} \quad \forall x \in \Omega.$$

We choose $\delta, \theta \approx h^{1/2}$ on the basis of Theorem 2.7.1, and report the results in Table 2.1(a) and Figure 2.1(a). We observe linear experimental convergence rates with respect to h, thus better than predicted in Theorem 2.7.1. They are however

justified for this case in [59]. The number $P = 4(D-1)$ stands for the number of points $x_i \pm \delta v_j$ used in the evaluation of the operator T_ε at each interior node $x_i \in \mathcal{N}_h^0$ and for D directions v_j in a quarter circle dictated by θ.

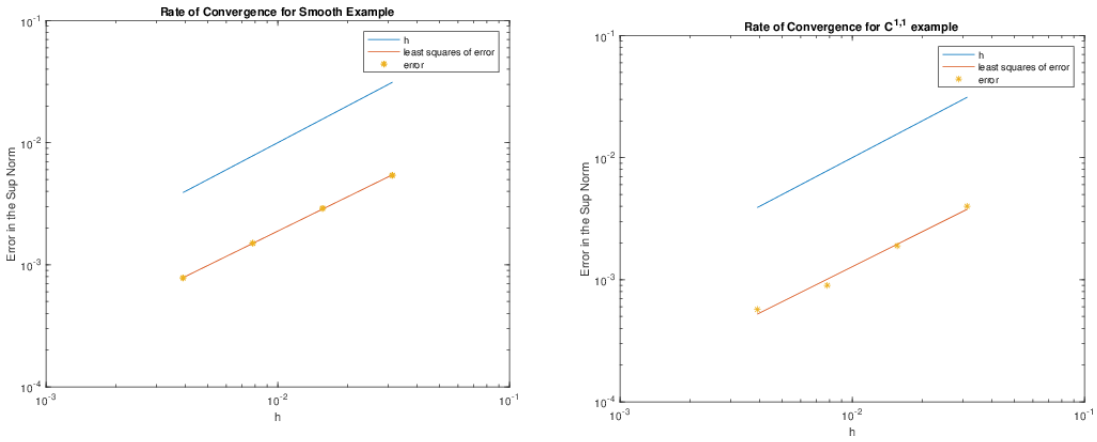


Figure 2.1: Experimental rates of convergence: the order is about 1 in terms of h for both the smooth Hessian with $\delta, \theta \approx h^{1/2}$ (left) and discontinuous Hessian $\delta \approx h^{4/5}, \theta \approx h^{2/5}$ (right).

Discontinuous Hessian: We choose the solution u and forcing function f to be

$$u(x) = \frac{1}{2}\left(\max(|x - x_0| - 0.2, 0)\right)^2, \quad f(x) = \max\left(1 - \frac{0.2}{|x - x_0|}, 0\right) \quad \forall x \in \Omega,$$

where $x_0 = (0.5, 0.5)$. Since $f = 0$ in the ball centered at x_0 of radius 0.2, this example is degenerate elliptic. We choose $\delta = O(h^{4/5})$ and $\theta = O(h^{2/5})$ on the basis of Theorem 2.7.4 and observe experimentally again a linear decay rate in h, which is better than predicted. This time Theorem 2.7.4 suggests a larger θ, but we choose a smaller θ without compromising the sparsity pattern of the Newton matrix. As illustrated on Table 2.1, despite its degeneracy and lack of global regularity, this example does not exhibit any problematic behavior compared to the smooth case.

Degrees of freedom	P: number of points	$L_\infty-$error	Newton steps
N= 1089, $h = 2^{-5}$	16	5.4 10^{-3}	8
N=4225, $h = 2^{-6}$	24	2.8 10^{-3}	7
N=16641, $h = 2^{-7}$	36	1.5 10^{-3}	7
N= 66049, $h = 2^{-8}$	52	7.8 10^{-4}	8

Degrees of freedom	P: number of points	$L_\infty-$error	Newton steps
N= 1089, $h = 2^{-5}$	20	4.0 10^{-3}	10
N=4225, $h = 2^{-6}$	28	1.9 10^{-3}	9
N=16641, $h = 2^{-7}$	36	9.0 10^{-4}	9
N= 66049, $h = 2^{-8}$	48	5.7 10^{-4}	9

Table 2.1: Smooth Hessian with $\delta, \theta \approx h^{1/2}$ (top), Discontinuous Hessian with $\delta \approx h^{4/5}, \theta \approx h^{2/5}$ (bottom). The convergence rate is about linear in h for both cases (see Figure 2.1), whereas the number of Newton steps seem insensitive to the dimension N of the nonlinear system.

We next explore the behavior of the operator T_ε in terms of the sign of the truncation error $E_\varepsilon[u_\varepsilon] := f - T_\varepsilon[u_\varepsilon]$. In Figure 2.2 (left) we split the interior nodes $\mathcal{N}_h^0$ into three sets, using the threshold $\texttt{eps} \approx 10^{-16}$ close to the machine precision of MATLAB: blue nodes x_i (34% of $\mathcal{N}_h^0$) correspond to $E_\varepsilon[u_\varepsilon](x_i) < -\texttt{eps}$; yellow nodes x_i (34% of $\mathcal{N}_h^0$) correspond to $E_\varepsilon[u_\varepsilon](x_i) > \texttt{eps}$; and magenta nodes x_i (32% of $\mathcal{N}_h^0$) correspond to $|E_\varepsilon[u_\varepsilon](x_i)| \leq \texttt{eps}$. Moreover, Figure 2.2 (left) displays with dashed lines the circle of discontinuity $|x - x_0| = 0.2$ and the two circles that are $\delta-$away from it. We point out that all points between the outer and inner circle are affected by the singularity, but they are mostly magenta nodes.

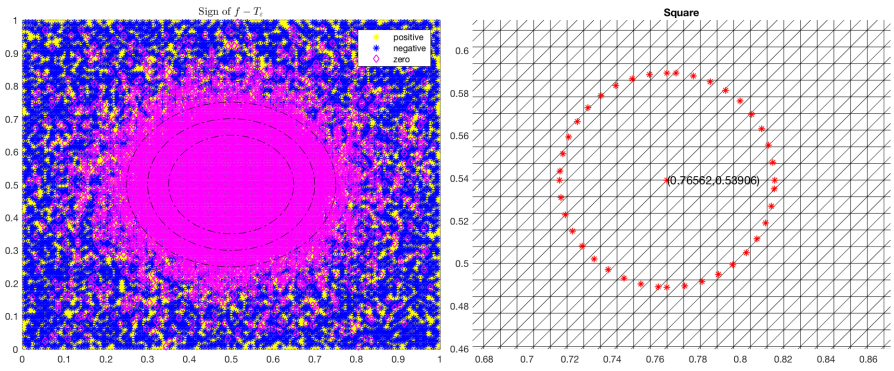


Figure 2.2: (left) Sign of the truncation error $E_\varepsilon[u_\varepsilon] = f - T_\varepsilon[u_\varepsilon]$ at nodes $x_i \in \mathcal{N}_h^0$ for the example with discontinuous Hessian and $h = 2^{-7}$: blue node x_i if $E_\varepsilon[u_\varepsilon](x_i) < -\texttt{eps}$, yellow node x_i if $E_\varepsilon[u_\varepsilon](x_i) > \texttt{eps}$, and magenta node x_i if $|E_\varepsilon[u_\varepsilon](x_i)| \leq \texttt{eps}$, where $\texttt{eps} \approx 10^{-6}$. We observe that the region $|x_i - x_0| < 0.2$, where $f \equiv 0$, is magenta. (right) Set of directions in $\mathbb{S}_\theta$ centered at a node $x_i \in \mathcal{N}_h^0$ and scaled by δ for the same example; note that $\delta/h \approx 7$.

Lastly, we use the same example to provide some insight on the two-scale nature of our method. In Figure 2.2 (right) we display a node $x_i = (0.7656, 0.5391)$ within a zoomed mesh, and the thirty-six directions v_j in $\mathbb{S}_\theta$ scaled by δ which are used for the calculation of $T_\varepsilon[u_\varepsilon](x_i)$ for mesh size $h = 2^{-7}$. We see that for this specific instance, $\delta/h \approx 7$, and that most points $x_i \pm \delta v_j$ are not nodes. We employ a fast search routine within FELICITY to locate such points [79, 80].

Unbounded f: We finally present computational results for an example that does not fall within our theory because the right hand side f is not uniformly bounded. More precisely, we consider the following f, which becomes unbounded near the corner $(1, 1)$ of Ω, and the corresponding exact solution u, which is twice differentiable in Ω but possesses an unbounded gradient near $(1, 1)$ [43]:

$$u(x) = -\sqrt{2 - |x|^2}, \quad f(x) = 2(2 - |x|^2)^{-2} \quad \forall x \in \Omega.$$

Table 2.2 shows that our method converges as the meshsize h decreases, but with a reduced rate and at the cost of an increased number of Newton iterations. We choose δ and θ similarly to the smooth Hessian case, but without any theoretical justification. We note that now we do not follow the approach of interpolating the coarse solution to the finer mesh, because $u \notin W^2_\infty(\Omega)$. Instead, we use the initial guess that corresponds to $\Delta u_0 = (d!f)^{1/d}$, which introduces more damping, say $\tau < 1$, in the Newton method.

Degrees of freedom	P: number of points	$L_\infty-$error	Newton steps
N= 1089, $h = 2^{-5}$	16	8.3 10^{-3}	8
N=4225, $h = 2^{-6}$	24	5.0 10^{-3}	15
N=16641, $h = 2^{-7}$	36	3.3 10^{-3}	18
N= 66049, $h = 2^{-8}$	52	2.0 10^{-3}	50

Table 2.2: Unbounded f. We observe that the method converges, but with a rate slower than linear and at the cost of increasing number of Newton iterations with each refinement.

Computational Performance: The process of locating the triangle of the mesh containing $x_i \pm \delta v_j$ and computing the barycentric coordinates is a rather small percentage of the total computing time. For instance, for $h = 2^{-7}$ and the smooth Hessian, this represents just 3% (< 3 sec) of the total computation time (90 sec). Due to the reduced sparsity pattern of the Newton matrices, the most time demanding task of the method is solving the linear systems, for which we use Matlab's backslash operator. This takes 42.7% of the total time. This computation is performed on an Intel 2.2 GHz i7 CPU, 16 GB RAM using Matlab R2017b.

Chapter 3: Convergent Two-Scale Filtered Scheme

In this chapter we introduce a more accurate version of our method in Chapter 2. We improve the accuracy by modifying our method using quadratic polynomials and a higher order approximation of the directional derivatives. This leads to what we call the *accurate operator* that we present in Section 3.1. This operator improves the accuracy of the method formally, but lacks the important property of monotonicity, hence cannot be proven to converge to the viscosity solution. This leads us to the main contribution of this chapter, which is the filtered scheme, where we combine the accurate and the monotone operators in order to have improved accuracy and provable convergence. We revise the numerical examples of Section 2.8 and discuss the performance of the new method.

3.1 Two-Scale Accurate Operator

In Remark 2.3.4 we saw that the smallest consistency error we can achieve from Lemma 2.3.2 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$) for the monotone operator $T_{\varepsilon,m}$ is of order h. We now explore a modification $T_{\varepsilon,a}$ to the operator $T_{\varepsilon,m}$ that allows us to achieve a better consistency error. We call $T_{\varepsilon,a}$ the accurate operator and we show that it has better formal approximation properties. However, we see that the operator is

not monotone and, as a consequence, the convergence results from Theorems 2.4.5 and 2.4.7 do not carry over.

3.1.1 Definition

We define the space $\mathbb{V}_h^2$ of continuous, piecewise quadratic functions. This allows us to get a smaller interpolation error in our approximation. In order to maximize the effect of this change, we introduce the method in the context of Isoparametric Finite Elements. We assume that our domain Ω is piecewise uniformly convex and piecewise $C^{1,1}$, so that we can guarantee the existence of invertible and quadratic maps that transform the master element into elements with curved sides connecting boundary nodes $\mathcal{N}_h^b$ [25]. We call the resulting mesh $\mathcal{T}_h^2$. Associated with $\mathcal{T}_h^2$ and $\mathbb{V}_h^2$, we define the piecewise quadratic interpolant $\mathcal{I}_h^2$. We also employ a more accurate approximation of the second directional derivatives that relies on five, rather than three, point stencils. This means that for our discrete operator and $u_\varepsilon \in \mathbb{V}_h^2$ we use this time

$$\nabla^2_{\delta,a} u_\varepsilon(x_i; v_j) := \frac{-u_\varepsilon(x_i + \delta v_j) + 16u_\varepsilon(x_i + \frac{\delta}{2} v_j) - 30u_\varepsilon(x_i) + 16u_\varepsilon(x_i - \frac{\delta}{2} v_j) - u_\varepsilon(x_i - \delta v_j)}{3\delta^2}, \tag{3.1.1}$$

where $x_i \in \mathcal{N}_h^0$ and $v_j \in \mathbb{S}_\theta$. The accurate scheme then becomes: We seek $u_\varepsilon \in \mathbb{V}_h^2$ such that $u_\varepsilon(x_i) = g(x_i)$ for $x_i \in \mathcal{N}_h^b$ and for $x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,a}[u_\varepsilon](x_i) := \min_{\mathbf{v} \in \mathbb{S}_\theta^\perp} \left(\prod_{j=1}^{d} \nabla^{2,+}_{\delta,a} u_\varepsilon(x_i; v_j) - \sum_{j=1}^{d} \nabla^{2,-}_{\delta,a} u_\varepsilon(x_i; v_j) \right) = f(x_i), \tag{3.1.2}$$

where again, $\varepsilon := (h, \delta, \theta)$ corresponds to the scales of $T_{\varepsilon,a}$.

3.1.2 Consistency Properties

We now present a consistency lemma to showcase the improved accuracy of the operator. We only sketch the proof of the result, since it mostly follows along the lines of Lemma 2.3.1 (consistency of $\nabla^2_\delta \mathcal{I}_h u$) . Note that our use of Isoparametric Finite Elements does not affect the following results, because in the interior of the domain the elements of $\mathcal{T}_h$ are straight and the δ-domain, $\Omega_{h,\delta}$, does not contain curved boundary cells.

Lemma 3.1.1 (consistency of $\nabla^2_{\delta,a}\mathcal{I}^2_h u$)**.** *Let $u \in W^3_\infty(B_i)$, $\mathcal{I}^2_h u$ be its Lagrange interpolant in $\mathbb{V}^2_h$, and B_i be defined in (2.3.2). The following two estimates are then valid:*

(i) For all $x_i \in \mathcal{N}^0_h$ and all $v_j \in \mathbb{S}_\theta$, we have

$$\left|\nabla^2_{\delta,a}\mathcal{I}^2_h u(x_i; v_j)\right| \leq C\ |u|_{W^2_\infty(B_i)}.$$

(ii) If in addition $u \in C^{2+k,\alpha}(B_i)$ for $k = 0, \ldots, 3$ and $\alpha \in (0, 1]$, then for all $x_i \in \mathcal{N}^0_h \cap \Omega_{h,\delta}$ and all $v_j \in \mathbb{S}_\theta$, we have

$$\left|\nabla^2_{\delta,a}\mathcal{I}^2_h u(x_i; v_j) - \frac{\partial^2 u}{\partial v_j^2}(x_i)\right| \leq C\left(|u|_{C^{2+k,\alpha}(B_i)}\delta^{k+\alpha} + |u|_{W^3_\infty(B_i)}\frac{h^3}{\delta^2}\right).$$

In both cases C stands for a constant independent of the two scales h and δ, the parameter θ and the function u.

Proof. (i) We rewrite (3.1.1) as

$$\nabla^2_{\delta,a}u(x; v) = \frac{4}{3}\frac{u(x + \frac{\delta}{2}) - 2u(x) + u(x - \frac{\delta}{2})}{\left(\frac{\delta}{2}\right)^2} - \frac{1}{3}\frac{u(x + \delta v) - 2u(x) + u(x - \delta v)}{\delta^2}$$

and then proceed as in Lemma 2.3.1 to obtain

$$\nabla^2_{\delta,a} u(x_i; v) \leq C|u|_{W^2_\infty(B_i)}.$$

We then incorporate the interpolation error [21]

$$\|u - \mathcal{I}^2_h u\|_{L^\infty(\Omega_h)} \leq C\ |u|_{W^3_\infty(B_i)} h^3,$$

along with the relation $h \leq \delta$ to complete the proof of (i).

(ii) This part follows from the same calculations as in Lemma 2.3.1 applied to the terms in (i). The higher approximation order is the result of cancellation of higher order derivatives.

□

Lemma 3.1.1 implies the following result, which is similar to Lemma 2.3.2.

Lemma 3.1.2 (consistency of $T_{\varepsilon,a}[\mathcal{I}^2_h u]$)**.** *Let $x_i \in \mathcal{N}^0_h \cap \Omega_{h,\delta}$ and B_i be defined as in (2.3.2). If $u \in C^{2+k,\alpha}(B_i)$ with $0 < \alpha \leq 1$ and $k = 0, \ldots, 3$ is convex, and $\mathcal{I}^2_h u$ is its piecewise quadratic interpolant, then*

$$\left|\det D^2 u(x_i) - T_{\varepsilon,a}[\mathcal{I}^2_h u](x_i)\right| \leq C_1(d, \Omega, u)\delta^{k+\alpha} + C_2(d, \Omega, u)\frac{h^3}{\delta^2} + C_3(d, \Omega, u)\theta^2, \tag{3.1.3}$$

where

$$\begin{aligned} &C_1(d, \Omega, u) = C|u|_{C^{2+k,\alpha}(B_i)}|u|^{d-1}_{W^2_\infty(B_i)}, \quad C_2 = C|u|_{W^3_\infty(B_i)}|u|^{d-1}_{W^2_\infty(B_i)} \\ &C_3 = C|u|^d_{W^2_\infty(B_i)}. \end{aligned} \tag{3.1.4}$$

If $x_i \in \mathcal{N}^0_h$ and $u \in W^2_\infty(B_i)$, then (3.1.3) remains valid with $\alpha = k = 0$ and $C^{2+k,\alpha}(B_i)$ replaced by $W^2_\infty(B_i)$.

Proof. The proof is again the same as in Lemma 2.3.2, since we use Lemma 3.1.1 (consistency of $\nabla^2_{\delta,a}\mathcal{I}^2_h u$) along with the bound $|\lambda_j| \leq |u|_{W^2_\infty(B_i)}$ on the eigenvalues λ_j of D^2u. □

Remark 3.1.3 (optimal consistency error)**.** *We observe that for a smooth function $u \in C^{5,1}(\overline{\Omega})$ the above lemma now provides us with a consistency error of order h^2 instead of only order h, which corresponds again to the optimal choice $\delta = h^{1/2}$.*

Remark 3.1.4 (lack of monotonicity)**.** *The definition of $\nabla^2_{\delta,a}$ in (3.1.1) implies that $T_{\varepsilon,a}$ is not monotone, since a change in u at a certain x_j could affect the second difference at another node x_i in two possible ways. It can either decrease or increase it, depending on whether it affects the behavior of $-u(x_i \pm \delta v)$ or $u(x_i \pm \frac{\delta}{2} v)$, respectively. We also note that $T_{\varepsilon,a}$ is defined on a space of piecewise quadratic functions, which means that the behavior at the nodes does not translate monotonically to the behavior inside simplices.*

The above remarks indicate that despite the improved accuracy of this operator, it is not possible to prove that a sequence of discrete solutions of (3.1.2) converges to the unique viscosity solution of (1.1.1).

3.2 Filtered Scheme

The above observations are the main motivation for designing a scheme that can exhibit a more accurate behavior than the monotone scheme, while also providing some flexibility that allows for convergence results. In this section we design such a scheme. We call it two-scale filtered scheme, because it involves two scales

as in Chapter 2, as well as a filter, in a similar fashion to [44]; see [17] for Hamilton-Jacobi equations. We use a filter scale to enforce an "almost monotone" structure on the scheme. At the end of this section we prove the convergence of the solution of the filtered scheme to the viscosity solution of (1.1.1).

3.2.1 General Definition of the Operator

We start by introducing the scheme using only some properties of the filter function, but without specifying its explicit form. We then discuss in Section 3.2.2 the precise definition of the filter. The idea of a filter function is to allow the accurate operator to take over the monotone operator, as long as it does not exhibit any undesirable behavior due to lack of monotonicity. We this in mind, we create the filter function in a way that controls the relative behavior of the accurate and monotone operators through a filter scale. To achieve this, we require our filter F to be a compactly supported and continuous (hence uniformly bounded) function that equals to the identity in part of its support. This will allow us to exploit the above idea of using the difference of the two operators as the decisive factor for the behavior of the filtered operator. We also introduce a filter scale $\tau = \tau(\varepsilon)$, where $\varepsilon = (h_a, h_m, \delta_a, \delta_m, \theta_a, \theta_m)$ and depends on the scales h_a, δ_a, θ_a and h_m, δ_m, θ_m of the accurate and monotone operators, respectively. Note that, since in each case we choose δ and θ to be functions of h, we can simply write $\tau = \tau(h_a, h_m)$. For the rest of the analysis it is important to observe that $\tau \to 0$ as $h_a, h_m \to 0$.

We now define the filter operator $T_{\varepsilon,f}$ and the discrete solution $\mathbf{U}_\varepsilon$. We first in-

troduce some notation that allows us to work with both operators from Sections 2.1.2 and 3.1. We call $\mathcal{T}_h^1$ a mesh of size h that corresponds to linear elements in the space $\mathbb{V}_h^1$. Similarly, we call $\mathcal{T}_{2h}^2$ a mesh of size $2h$ for isoparametric quadratic elements in $\mathbb{V}_{2h}^2$. We observe that for two successive triangulations $\mathcal{T}_{2h}^1$ and $\mathcal{T}_h^1$, the corresponding polygonal domains satisfy $\Omega_{2h} \subset \Omega_h$. We also recall that $\mathcal{N}_h^b \subset \partial\Omega$. This implies that the nodal values of $\mathbb{V}_{2h}^2$ in $\mathcal{T}_{2h}^2$ coincide with the nodal values of $\mathbb{V}_h^1$ in $\mathcal{T}_h^1$. We illustrate this in Figure 3.1 for the two-dimensional case and explain it briefly here: the statement is clear for cells in the interior of the domain, where the isoparametric map from $\mathcal{T}_{2h}^1$ to $\mathcal{T}_{2h}^2$ is the identity map and the midpoints of the interior triangles of $\mathcal{T}_{2h}^1$ that are used as nodal values for $\mathcal{T}_{2h}^2$ are the new vertices of the refined triangles of $\mathcal{T}_h^1$. On the other hand, boundary cells of $\mathcal{T}_{2h}^1$ are mapped to cells with curved boundary faces (edges for $d = 2$), in order to construct $\mathcal{T}_{2h}^2$. The quadratic space $\mathbb{V}_{2h}^2$ defined over $\mathcal{T}_{2h}^2$ involves nodal values on the midpoints of all faces and for the curved ones, the points used for these nodal values lie on the boundary. When the triangulation is refined, these points coincide with the new boundary vertices of $\mathcal{T}_h^1$. Consequently, the nodal values of $\mathbb{V}_h^1$ over $\mathcal{T}_h$ and of $\mathbb{V}_{2h}^2$ over $\mathcal{T}_{2h}^2$ correspond to the exact same physical coordinates.

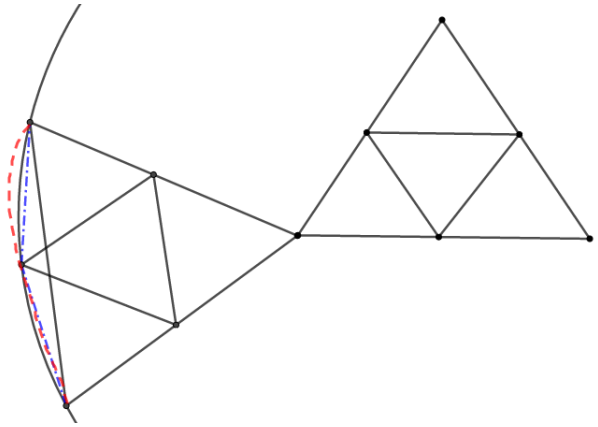


Figure 3.1: Illustration of the refinement close and away from the boundary for $d = 2$. In the interior we create four new triangles upon each refinement, by connecting the midpoints of the coarser cell. The space $\mathbb{V}_h^1$ corresponding to $\mathcal{T}_h^1$ and defined over these four triangles shares the same nodal values as the space $\mathbb{V}_{2h}^2$ corresponding to $\mathcal{T}_{2h}^2$ and defined over the original triangle. On boundary elements, the boundary point that is used in the construction and as a nodal value of the isoparametric element in $\mathcal{T}_{2h}^2$ for $\mathbb{V}_{2h}^2$ becomes the new boundary node for $\Omega_h \supset \Omega_{2h}$ that is used for the four new triangles of $\mathcal{T}_h^1$. The blue "$-\cdot-\cdot-$" dashed line corresponds to the new edges introduced after the refinement. The red "$---$" dashed line illustrates the curved edge of the curved element that is isoparametric to the original triangle.

We now exploit the above structure in order to introduce our discrete solution $\mathbf{U}_\varepsilon$. It is a grid function, in the sense that we think of it as a vector $\{\mathbf{U}_\varepsilon^i\}_i$ of nodal values of a piecewise linear function over $\mathcal{T}_h^1$, which, as explained above, are also the nodal values of a piecewise quadratic function over $\mathcal{T}_{2h}^2$. This allows us to define two functions $u_\varepsilon^1 \in \mathbb{V}_h^1$ over $\mathcal{T}_h^1$ and $u_\varepsilon^2 \in \mathbb{V}_{2h}^2$ over $\mathcal{T}_{2h}^2$ using $\mathbf{U}_\varepsilon$ as nodal values. We can now compare the monotone and accurate operator applied to u_ε^1 and u_ε^2, respectively. We seek $\mathbf{U}_\varepsilon$ such that $\mathbf{U}_\varepsilon = g(x_i)$ for all $x_i \in \mathcal{N}_h^b$ and for all $x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) := T_{\varepsilon,m}[u_\varepsilon^1](x_i) + \tau F\left(\frac{T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)}{\tau}\right) = f(x_i). \quad (3.2.1)$$

This definition of the filtered operator is twofold:

(i) *Minimize the risk:* As we explained in Section 3.1, the accurate operator (3.1.2) exhibits a smaller consistency error than the monotone operator (2.1.2) (see Lemma 2.3.2 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$) and Lemma 3.1.2 (consistency of $T_{\varepsilon,a}[\mathcal{I}_h^2 u]$)), but the improved accuracy comes at the cost of the lack of monotonicity. Therefore, given the importance of monotonicity for the convergence of schemes in the context of viscosity solutions, the scheme in Section 3.1 cannot be guaranteed to converge. Especially in cases where the right-hand side f degenerates, hence when we cannot expect to have a classical solution of (1.1.1), this lack of monotonicity could in principle lead to the failure of the accurate operator. This is why we use a filter that examines how much the two operators $T_{\varepsilon,m}$ and $T_{\varepsilon,a}$ differ. We will see in Section 3.2.2 that if their difference is in absolute value bigger than τ, the monotone operator is employed in order to guarantee that nothing goes wrong and that our scheme converges. We note here that this is a rather simple approach that could lead to using the monotone operator even in cases where the accurate operator is much better. We explore and discuss this further in Section 3.4.

(ii) *Almost monotonicity:* The main feature of F is that it is uniformly bounded and in particular, $|F(s)| \leq 1$ for all $s \in \mathbb{R}$. This leads to Lemma 3.2.1 (almost monotonicity), which, in turn, is critical to prove existence of a discrete solutions and its convergence. Being able to prove such results is an important component of this work, hence the definition of the filter and of the filter oper-

ator (3.2.1) are closely aligned with this goal. In Section 3.2.2 we also take into account the possible degeneracy of the right-hand side f and provide explicit definitions of F.

For the rest of the presentation, we assume that $h_a = 2h_m = 2h$ and use the notation $h \to 0$ for our convergence result. However, we keep the notation δ_m and δ_a, because these quantities do not always have the same expression with respect to h. Moreover, in order to simplify our presentation, we introduce the notation

$$A[\mathbf{U}_\varepsilon](x_i) := T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i). \tag{3.2.2}$$

This quantity is often used in the following sections as the argument of F, similarly to (3.2.1).

We now use the uniform bound of F and (3.2.1) to show that the filtered operator satisfies the following notion of "almost monotonicity".

Lemma 3.2.1 (almost monotonicity)**.** *For two grid functions $\mathbf{V}_h, \mathbf{W}_h$ such that $\mathbf{V}_h - \mathbf{W}_h$ attains a maximum at an interior node $x_i \in \mathcal{N}_h^0$, we have that*

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) + 2\tau \geq T_{\varepsilon,f}[\mathbf{V}_h](x_i).$$

Proof. Let $v_h^1, w_h^1 \in \mathbb{V}_h^1$ and $v_h^2, w_h^2 \in \mathbb{V}_{2h}^2$ be the corresponding functions with the same nodal values as $\mathbf{V}_h$ and $\mathbf{W}_h$, respectively. We use the definition of the filtered operator to write

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) = T_{\varepsilon,m}[w_h^1](x_i) + k_i\ \tau, \quad T_{\varepsilon,f}[\mathbf{V}_h](x_i) = T_{\varepsilon,m}[v_h^1](x_i) + \tilde{k}_i\ \tau$$

where $|k_i|, |\tilde{k}_i| \leq 1$, since $|F(s)| \leq 1$ for all s. We now employ [63, Lemma 2.3]

(monotonicity) for $T_{\varepsilon,m}$ and w_h^1, v_h^1 to deduce that

$$T_{\varepsilon,m}[w_h^1](x_i) \geq T_{\varepsilon,m}[v_h^1](x_i).$$

Combining these expressions, we obtain

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) + (\tilde{k}_i - k_i)\tau \geq T_{\varepsilon,f}[\mathbf{V}_h](x_i),$$

whence

$$T_{\varepsilon,f}[\mathbf{W}_h](x_i) + 2\tau \geq T_{\varepsilon,f}[\mathbf{V}_h](x_i).$$

This concludes the proof. □

3.2.2 Definition of Filter Function

As our filter function we can use a continuous and uniformly bounded function F_σ that is equal to the identity for $|s| \leq 1$ and then continuously transitions to zero:

$$F_\sigma(s) := \begin{cases} s, & |s| \leq 1 \\ 0, & |s| \geq 1+\sigma \\ -\frac{1}{\sigma}\, s + \frac{1+\sigma}{\sigma}, & 1 < s < 1+\sigma \\ -\frac{1}{\sigma}\, s - \frac{1+\sigma}{\sigma}, & -1-\sigma < s < -1 \end{cases} \tag{3.2.3}$$

The parameter σ encodes the continuous transition of F_σ to zero. Its choice and use is further discussed in Section 3.4.

We see that (3.2.3) satisfies the properties mentioned in Section 3.2.1, i.e. it is uniformly bounded by one and allows the accurate operator to take effect only if it does not differ much from the monotone operator. However, given our analysis in Chapter 2, one important feature of $T_{\varepsilon,m}$ is that it can guarantee the discrete

convexity of the discrete solutions. This may not be guaranteed with (3.2.3) in cases where the right hand side f touches zero. Although this is possible to deal with asymptotically, it is also desirable to mimic the properties of the continuous problem, which justifies preserving the discrete convexity of the discrete solutions. We explain now why this property may not be guaranteed if we use (3.2.3) and suggest a way to ameliorate it.

We observe that for every grid function $\mathbf{U}_h$ with corresponding $u_h^1 \in \mathbb{V}_h^1$ and $u_h^2 \in \mathbb{V}_{2h}^2$ and for all $x_i \in \mathcal{N}_h^0$ there exists $|k_i| \leq 1$ that depends on the values of $T_{\varepsilon,a}[u_h^2](x_i), T_{\varepsilon,m}[u_h^1](x_i)$ such that

$$T_{\varepsilon,f}[u_h](x_i) = \begin{cases} T_{\varepsilon,m}[u_h^1](x_i), & |A[\mathbf{U}_h](x_i)| \geq (1+\sigma)\tau \\ T_{\varepsilon,m}[u_h^1](x_i) - |k_i|\tau, & -(1+\sigma)\tau < A[\mathbf{U}_h](x_i) \leq 0 \\ T_{\varepsilon,m}[u_h^1](x_i) + |k_i|\tau, & 0 < A[\mathbf{U}_h](x_i) < (1+\sigma)\tau \end{cases} \tag{3.2.4}$$

Consider now the case when there exists $x_i \in \mathcal{N}_h^0$ such that $f(x_i) = 0$. If we want to preserve the discrete convexity of the function u_ε^1 that corresponds to the discrete solution $\mathbf{U}_\varepsilon$ and that is guaranteed by the monotone operator whenever $f \geq 0$, we need to make sure that $T_{\varepsilon,m}[u_\varepsilon^1](x_i) \geq 0$. This is also used in Lemma 3.2.3 (existence and stability), in order to prove the existence of a discrete solution using results from Chapter 2. The above calculation shows that for $x_i \in \mathcal{N}_h^0$ we end up with

$$T_{\varepsilon,m}[u_\varepsilon^1](x_i) = \begin{cases} f(x_i) = 0, & |A[\mathbf{U}_\varepsilon](x_i)| \geq (1+\sigma)\tau \\ f(x_i) + |k_i|\tau \geq 0, & -(1+\sigma)\tau < A[\mathbf{U}_\varepsilon](x_i) \leq 0 \\ f(x_i) - |k_i|\tau \leq 0, & 0 < A[\mathbf{U}_\varepsilon](x_i)(x_i) < (1+\sigma)\tau \end{cases} \tag{3.2.5}$$

This means that, in order to preserve $T_{\varepsilon,m}[u_\varepsilon^1](x_i) \geq 0$ for all $x_i \in \mathcal{N}_h^0$, we need to modify our filter in order to exclude the third case, where we may end up with a negative value for the monotone operator, when $k_i \neq 0$. To this end, we introduce a non-symmetric modification of the filter function:

$$\tilde{F}_\sigma(s) := \begin{cases} s, & -1 \leq s \leq 0 \\ -\frac{1}{\sigma}\, s - \frac{1+\sigma}{\sigma}, & -1-\sigma \leq s \leq -1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.6}$$

Remark 3.2.2 (choice of filter function)**.** *This filter allows us to avoid the third case where $T_{\varepsilon,m}[u_\varepsilon^1] \leq 0$, which, in turn, allows us in Lemma 3.2.3 (existence and stability) to construct a discrete convex solution for our operator even when f touches zero. From now on we make the convention that (3.2.3) is used in (3.2.1) whenever $f > 0$ everywhere and (3.2.6) is used whenever we have a right-hand side that touches zero. We emphasize that this is a decision that we make once depending on the right-hand side and it does not depend on x. In Lemma 3.2.3 (existence and stability) we show that for all $f \geq f_0 > 0$ we can choose $\tau \leq f_0$ and prove the existence of a discretely convex solution of (3.2.1) using (3.2.3) as our filter. This is not a big restriction, since in practice the lower bound of f is expected to be of order 1, while τ is some positive power of h. On the other hand, choosing*

the non-symmetric filter $\tilde{F}_\sigma$ may worsen the symmetry of the resulting system, while excluding parts of the domain where $T_{\varepsilon,a}$ could have still been used, i.e. whenever $f \geq \tau$. Consequently, we employ it only when it is necessary. In Section 3.4 we test our method for a smooth example with strictly positive f using (3.2.3) everywhere, as well as a $C^{1,1}$ example where f is zero in a big part of the domain, in which case we employ (3.2.6) everywhere. We also explore briefly a space-dependent choice of filter.

3.2.3 Existence of Discrete Solution

In this section we prove that (3.2.1) has a unique discrete solution $\mathbf{U}_\varepsilon$. This hinges on Lemma 3.2.1 (almost monotonicity) and the results in Chapter 2. We exploit the existence proven in Lemma 2.2.1 (existence, uniqueness and stability) combined with a fixed point argument that was used in [44].

Lemma 3.2.3 (existence and stability)**.** *There exists a grid function $\mathbf{U}_\varepsilon$ that is the solution of the filtered scheme (3.2.1) with appropriate choice of filter function and is stable in the sense that $\|\mathbf{U}_\varepsilon\|_{l^\infty(\Omega)}$ does not depend on the parameters involved in ϵ.*

Proof. We first show that for any grid function $\mathbf{U}_h$ with corresponding functions $u_h^1 \in \mathbb{V}_h^1$ and $u_h^2 \in \mathbb{V}_{2h}^2$, we can find a grid function $\mathbf{Y}_h(\mathbf{U}_h)$ with corresponding functions $y_h^1(\mathbf{U}_h) \in \mathbb{V}_h^1$ and $y_h^2(\mathbf{U}_h) \in \mathbb{V}_{2h}^2$ such that for all $x_i \in \mathcal{N}_h^0$

$$T_{\varepsilon,m}[y_h^1(\mathbf{U}_h)](x_i) + \tau F_\sigma(A[\mathbf{U}_h](x_i)) = f(x_i). \tag{3.2.7}$$

We focus on grid functions with corresponding u_h^1 that is discretely convex and that

satisfy the boundary conditions on $\mathcal{N}_h^b$. Note that the Lagrange interpolant in $\mathbb{V}_h^1$ of any convex function is discretely convex and [48, Theorem 1.6.2] guarantees that there exists a convex function satisfying the boundary conditions. Hence the set of such functions is non-empty for all δ_m. We prove the result in two steps.

Step 1. We prove that for every grid function $\mathbf{U}_h$ such that the corresponding function $u_h^1 \in \mathbb{V}_h^1$ is discretely convex, we can find $\mathbf{Y}_h(\mathbf{U}_h)$ such that (3.2.7) holds and the corresponding $y_h^1(\mathbf{U}_h) \in \mathbb{V}_h^1$ is discretely convex. To this end we distinguish two cases with respect to whether f is strictly positive or not.

(i) $\boldsymbol{f > 0}$ **:** We use the filter function F_σ from (3.2.3) with $\sigma > 0$ and recall (3.2.5). Since $f > 0$ and uniformly continuous in Ω, there exists $f_0 > 0$ such that $f(x) \geq f_0 > 0$ for all $x \in \Omega$. Then, choosing $\tau \leq f_0$, (3.2.7) becomes

$$T_{\varepsilon,m}[y_h^1(\mathbf{U}_h)](x_i) = f(x_i) - \tau F_\sigma\left(\frac{A[\mathbf{U}_h](x_i)}{\tau}\right)$$
$$= \begin{cases} f(x_i) \geq 0, & |A[\mathbf{U}_h](x_i)| \geq (1+\sigma)\tau \\ f(x_i) + |k_i|\tau > 0, & -(1+\sigma)\tau < A[\mathbf{U}_h](x_i) \leq 0 \\ f(x_i) - |k_i|\tau \geq 0, & 0 < A[\mathbf{U}_h](x_i) < (1+\sigma)\tau \end{cases}$$

for $|k_i| \leq 1$. We now employ the existence result from Lemma 2.2.1 (existence, uniqueness and stability) to conclude that there exists a unique solution $y_h^1(\mathbf{U}_h) \in \mathbb{V}_h^1$ to the above problem with $\|y_h^1(\mathbf{U}_h)\|_{L^\infty} \leq \Lambda$, where Λ depends on $\|g\|_{L^\infty}$ and $\|f\|_{L^\infty}$, since $\tau \leq f_0 \leq \|f\|_{L^\infty}$. The existence is guaranteed by the continuity of f and F, since we can interpolate $F_\sigma(A[\mathbf{U}_h])$ as a continuous piecewise linear function in the whole domain. Using the nodal values of $y_h^1(\mathbf{U}_h)$, we thus obtain $\mathbf{Y}_h(\mathbf{U}_h)$.

(ii) $\boldsymbol{f \geq 0}$ **:** In this case, motivated by (3.2.5), we employ the filter function in (3.2.6) with $\sigma > 0$ to obtain

$$T_{\varepsilon,m}[y_h^1(\mathbf{U}_h)](x_i) = f(x_i) - \tau \tilde{F}_\sigma \left(\frac{A[\mathbf{U}_h](x_i)}{\tau} \right)$$

$$= \begin{cases} f(x_i) + |k_i|\tau \geq 0, & -(1+\sigma)\tau < A[\mathbf{U}_h](x_i) \leq 0 \\ f(x_i) \geq 0, & \text{otherwise} \end{cases}$$

where $|k_i| \leq 1$. We can now use Lemma 2.2.1 (existence, uniqueness and stability) and conclude the existence of a unique $y_h^1(\mathbf{U}_h) \in \mathbb{V}_h^1$ with a uniform bound in L^∞, hence existence of the grid function $\mathbf{Y}_h(\mathbf{U}_h)$.

Step 2. In view of Step 1, we can restrict ourselves to the compact, convex set of grid functions $\mathbf{U}_h$ with corresponding $u_h^1 \in \mathbb{V}_h^1$ that is discretely convex and that satisfy the boundary conditions, as well as $\|\mathbf{U}_h\|_{l^\infty(\Omega)} \leq \Lambda$. Then, Step 1 implies that for all such $\mathbf{U}_h$ we can find $\mathbf{Y}_h(\mathbf{U}_h)$ solution of (3.2.7) such that $\|\mathbf{Y}_h(\mathbf{U}_h)\|_{l^\infty(\Omega)} \leq \Lambda$. Then, similarly to [44], by Brouwer's fixed point theorem, $\mathbf{Y}_h$ has a fixed point in that set, whence there exists $\mathbf{U}_\varepsilon$, uniformly bounded in the max norm independent of the parameters in ε and such that $\mathbf{U}_\varepsilon = g(x_i)$ for all $x_i \in \mathcal{N}_h^b$ and for all $x_i \in \mathcal{N}_h^0$:

$$T_{\varepsilon,m}[\mathbf{U}_\varepsilon](x_i) + \tau F_\sigma \left(\frac{A[\mathbf{U}_\varepsilon](x_i)}{\tau} \right) = f(x_i).$$

This concludes the proof. □

Remark 3.2.4 (non-uniqueness)**.** *We emphasize that the above proof does not guarantee the existence of a unique solution to* (3.2.1)*, since in principle we can have more than one fixed points for* (3.2.7)*. However, we can use the discrete comparison*

principle introduced in [63] to show that two different solutions of (3.2.1) are very close to each other in the L^∞ norm.

Lemma 3.2.5 (control of the lack of uniqueness)**.** *Let $\mathbf{U}_\varepsilon, \mathbf{V}_\varepsilon$ be two discrete solutions of (3.2.1). Then,*

$$\|\mathbf{U}_\varepsilon - \mathbf{V}_\varepsilon\|_{l^\infty(\Omega)} \leq C\ \tau^{1/d},$$

where C depends only on the dimension d and Ω.

Proof. The result is an immediate consequence of the discrete comparison principle introduced in Lemma 2.1.4 (discrete comparison principle) and the fact that by construction, $T_{\varepsilon,m}[u^1_\varepsilon](x_i) \geq 0$ and $T_{\varepsilon,m}[v^1_\varepsilon](x_i) \geq 0$ for all $x_i \in \mathcal{N}^0_h$, where $u^1_\varepsilon, v^1_\varepsilon \in \mathbb{V}^1_h$ are defined as usual. To show this, we use the discrete barrier $q_h = \mathcal{I}^1_h(|x - x_0|^2 - R^2) \in \mathbb{V}^1_h$ that was introduced in Lemma 2.4.2, where x_0 and $R > 0$ are such that $\Omega \subset B_R(x_0)$. Since $T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = T_{\varepsilon,f}[\mathbf{V}_\varepsilon](x_i) = f(x_i)$ for all $x_i \in \mathcal{N}^0_h$, we have that

$$T_{\varepsilon,m}[u^1_\varepsilon](x_i) \leq T_{\varepsilon,m}[v^1_\varepsilon](x_i) + 2\tau \leq T_{\varepsilon,m}\Big[v^1_\varepsilon + \frac{(2\tau)^{1/d}}{2} q_h\Big](x_i)$$

and

$$T_{\varepsilon,m}[v^1_\varepsilon](x_i) \leq T_{\varepsilon,m}[u^1_\varepsilon](x_i) + 2\tau \leq T_{\varepsilon,m}\big[u^1_\varepsilon + \frac{(2\tau)^{1/d}}{2} q_h\big](x_i)$$

for all $x_i \in \mathcal{N}^0_h$. Applying Lemma 2.1.4 (discrete comparison principle), we obtain that

$$|u^1_\varepsilon - v^1_\varepsilon| \leq \Big|\frac{(2\tau)^{1/d}}{2} q_h\Big| \leq C\tau^{1/d},$$

which concludes the proof. □

3.2.4 Convergence

We now prove the convergence of the solution u_ε of (3.2.1) to the unique viscosity solution of (1.1.1). To this end, we follow the proof of convergence of Theorem 2.4.7 (uniform convergence) which in turn modifies that of [7] to account for the Dirichlet boundary conditions and the lack of operator consistency near $\partial\Omega_h$. In addition, we exploit the almost monotone nature of the scheme, as in [44], to further adjust the proof of [7] and derive uniform convergence of u_ε to u in Ω. In both Theorem 2.4.7 and the current result, we need to employ a discrete barrier argument, in order to control the behavior of the discrete solution close to the boundary. We recall here for convenience the discrete barrier function introduced in Lemma 2.4.1.

Lemma 3.2.6 (discrete boundary barrier)**.** *Let Ω be uniformly convex and $E > 0$ be arbitrary. For each node $z \in \mathcal{N}_h^0$ with $\operatorname{dist}(z, \partial\Omega_h) \leq \delta$, there exists a function $p_h \in \mathbb{V}_h^1$ such that $T_{\varepsilon,m}[p_h](x_i) \geq E$ for all $x_i \in \mathcal{N}_h^0$, $p_h \leq 0$ on $\partial\Omega_h$ and*

$$|p_h(z)| \leq CE^{1/d}\delta,$$

with C depending on the curvature of the boundary.

Before proceeding further, we recall the following continuous version of the Monge-Ampère operator

$$T[u] := \min_{\mathbf{v}=(v_j)_{j=1}^d \in \mathbb{S}^\perp} \left(\prod_{j=1}^d \partial_{v_j v_j}^{2,+} u - \sum_{j=1}^d \partial_{v_j v_j}^{2,-} u \right),$$

and the equivalence of the viscosity solutions of (1.1.1) and

$$T[u] = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega,$$

as proven in Lemma 2.4.6 (equivalence of viscosity solutions). We can now prove the uniform convergence of the discrete solution of (3.2.1) in Ω. Since $\mathbf{U}_\varepsilon$ and u^1_ε are defined in the computational domain Ω_h, and $\Omega_h \subset \Omega$, we extend u^1_ε to Ω as follows. Given $x \in \Omega \setminus \Omega_h$ let $z \in \partial\Omega_h$ be the closest point to x, which is unique because Ω_h is convex, and let

$$u^1_\varepsilon(x) := u^1_\varepsilon(z) = \mathcal{I}^1_h g(z) \quad \forall\, x \in \Omega \setminus \Omega_h. \tag{3.2.8}$$

This is needed in order to control the behavior of u^1_ε close to the boundary using tools from Theorem 2.4.7 (uniform convergence).

We also introduce the limit supremum and the limit infimum of the discrete solution, namely

$$u^*(x) = \limsup_{\varepsilon\to 0, z\to x} u^1_\varepsilon(z), \qquad u_*(x) = \liminf_{\varepsilon\to 0, z\to x} u^1_\varepsilon(z) \quad \forall x \in \Omega.$$

In the above definition we also require (but omit from the notation) that for $h_m = h$ and $h_a = 2h$, we have that $\frac{h_m}{\delta_m}, \frac{h_a}{\delta_a} \to 0$ as $h \to 0$. We observe that u^* is upper semi-continuous and u_* is lower semi-continuous. Since the proof follows closely the one in Theorem 2.4.7, we emphasize only the parts of it that are different for the filtered operator. In the following calculations we do not rely on the precise definition of the filter function F and only assume that it satisfies Lemma 3.2.1 (almost monotonicity).

Theorem 3.2.7 (uniform convergence)**.** *Let Ω be uniformly convex, $f \in C(\Omega) \cap L^\infty(\Omega)$ satisfy $f \geq 0$, and $g \in C(\partial\Omega)$. The discrete solution $\mathbf{U}_\varepsilon$ of (3.2.1) converges uniformly to the unique viscosity solution $u \in C(\overline{\Omega})$ of (1.1.1) as $\varepsilon = \varepsilon(h) \to 0$.*

Proof. In view of Lemma 2.4.6 (equivalence of viscosity solutions), we prove that $\mathbf{U}_\varepsilon$ converges to the viscosity solution of (2.4.6). To this end, we have to deal with a test function $\phi \in C^2(\Omega)$ and the respective grid function $\mathbf{\Phi}_h$ with corresponding piecewise polynomial functions $\phi_h^1 = \mathcal{I}_h^1\phi \in \mathbb{V}_h^1$ and the $\phi_h^2 = \mathcal{I}_h^2\phi \in \mathbb{V}_{2h}^2$. Without loss of generality we may assume $\phi \in C^{2,\alpha}(\Omega)$. We split the proof into five steps.

Step 1: Consistency. We have the following consistency estimate for the operator T in (2.4.6):

$$\left|T[\phi](x_0) - T_{\varepsilon,f}[\mathbf{\Phi}_h](x_i)\right| \leq C_1(\phi)\Big(\delta_m^\alpha + |x_0 - x_i|^\alpha\Big) + C_2(\phi)\Big(\frac{h^2}{\delta_m^2} + \theta_m^2\Big) + \tau,$$

where the constants C_1, C_2 are defined in Lemma 2.3.2 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u](x_i)$) and depend on $|\phi|_{C^{2,\alpha}(B_i)}$ and $|\phi|_{W_\infty^2(B_i)}$ with B_i defined in (2.3.2), and $x_0 \in \Omega, x_i \in \mathcal{N}_h^0 \cap \Omega_{h,\delta_m}$. This is an immediate consequence of Lemma 2.3.2, the Lipschitz continuity of the min and max functions and the fact that $|T_{\varepsilon,f}[\mathbf{\Phi}_h] - T_{\varepsilon,m}[\phi_h^1]| \leq \tau$.

Step 2: Subsolutions. We show that u^* is a viscosity subsolution of (2.4.6); likewise u_* is a viscosity supersolution. This hinges on monotonicity and consistency [7]. In our case, we employ the monotonicity of (2.1.2) and the fact that the operator in (3.2.1) is almost monotone, in the sense described in Lemma 3.2.1 (almost monotonicity). We must show that if $u^* - \phi$ attains a local maximum at $x_0 \in \Omega$, we have

$$T[\phi](x_0) \geq f(x_0);$$

note that $u^* - \phi$ is upper semi-continuous and the local maximum is well defined. Without loss of generality, we may assume that $u^* - \phi$ attains a strict global maximum at x_0 [52, Remark in p.31] and $x_0 \in \Omega_h$ for h sufficiently small. Let $x_h \in \mathcal{N}_h$

be a sequence of nodes so that $\mathbf{U}_\varepsilon - \mathbf{\Phi}_h$ attains a maximum at x_h. Similarly to Theorem 2.4.7 (uniform convergence), we claim that $x_h \to x_0$ as $h \to 0$. Exploiting the fact that $\mathbf{U}_\varepsilon - \mathbf{\Phi}_h$ attains a maximum at x_h, Lemma 3.2.1 (almost monotonicity) yields

$$T_{\varepsilon,f}[\mathbf{\Phi}_h](x_h) + 2\tau \geq T_{\varepsilon,f}[u_\varepsilon](x_h) = f(x_h),$$

where $\tau \to 0$ as $\varepsilon \to 0$. Since $f \in C(\Omega)$, to prove $T[\phi](x_0) \geq f(x_0)$ we only need to show that as $\varepsilon \to 0$

$$T_{\varepsilon,f}[\mathbf{\Phi}_h](x_h) \to T[\phi](x_0).$$

This is a consequence of Step 1 and the fact that $x_h \in \Omega_{h,\delta_m}$ for δ_m sufficiently small, because $x_0 \in \Omega$, $x_h \to x_0$ and the sequence of $\Omega_h \uparrow \Omega$ is non-decreasing.

Step 3: Boundary Behavior. We now prove that $u^* = u_* = g$ on $\partial\Omega$ via a barrier argument similar to that in Theorem 2.4.7 (uniform convergence). This is essential in order to apply the comparison principle for operator T to relate u_*, u^* and u in Step 4.

Let p_k be the quadratic function in the proof of Lemma 2.4.1 (discrete boundary barrier) associated with an arbitrary boundary point $x \in \partial\Omega$ (the origin in the construction of p_k) and with constant $E = k$. We recall that $p_k(x) = 0$ and $p_k(z) \leq 0$ for all $z \in \partial\Omega$ can be made arbitrarily large for $k \to \infty$ by virtue of the uniform convexity of Ω. A simple consequence is that the sequence of points $x_k \in \partial\Omega$ where $g + p_k$ (resp. $g - p_k$) attains a maximum (resp. a minimum) over $\partial\Omega$ converges to x as $k \to \infty$.

Lemma 2.1.3 (monotonicity) implies the following maximum principle for the

monotone operator: if a discretely convex function $u_h^1 \in \mathbb{V}_h^1$ satisfies $T_{\varepsilon,m}[u_h^1](x_i) > 0$ for all $x_i \in \mathcal{N}_h^0$, then u_h^1 attains a maximum over Ω_h on $\mathcal{N}_h^b \subset \partial\Omega$. We now see that, since $T_{\varepsilon,m}[u_\varepsilon^1] \geq 0$, we have that for k big enough and all $x_i \in \mathcal{N}_h^0$

$$\begin{aligned} T_{\varepsilon,m}[u_\varepsilon^1 + \mathcal{I}_h^1 p_k](x_i) &\geq T_{\varepsilon,m}[u_\varepsilon^1] + T_{\varepsilon,m}[\mathcal{I}_h^1 p_k](x_i) \\ &\geq T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) - \tau + E \\ &= f(x_i) - \tau + E > 0, \end{aligned}$$

whence $u_\varepsilon^1 + \mathcal{I}_h^1 p_k$ attains its maximum on $\mathcal{N}_h^b$. In view of (3.2.8), we may assume $z \in \Omega_h$ in the limit $u^*(x) = \limsup_{\varepsilon, \frac{h}{\delta} \to 0, z \to x} u_\varepsilon^1(z)$. Consequently,

$$\begin{aligned} u^*(x) &\leq \limsup_{\varepsilon \to 0, z \to x} \big(u_\varepsilon^1(z) + \mathcal{I}_h^1 p_k(z)\big) - \liminf_{\varepsilon \to 0, z \to x} \mathcal{I}_h^1 p_k(z) \\ &\leq \limsup_{\varepsilon \to 0} \max_{z \in \mathcal{N}_h^b} \big(g + p_k\big)(z) - p_k(x) \leq g(x_k) + p_k(x_k) \leq g(x_k), \end{aligned}$$

because $\max_{\mathcal{N}_h^b} g + p_k \leq \max_{\partial\Omega} g + p_k$. Hence taking $k \to \infty$ yields $u^*(x) \leq g(x)$.

On the other hand, since $T_{\varepsilon,m}[\mathcal{I}_h^1 p_k](x_i) > T_{\varepsilon,m}[u_\varepsilon^1](x_i)$ for all $x_i \in \mathcal{N}_h^0$ and k big enough, Lemma 2.1.3 (monotonicity) implies that $u_\varepsilon^1 - \mathcal{I}_h^1 p_k$ attains a minimum on $\mathcal{N}_h^b$. Therefore, arguing as before

$$u_*(x) \geq \liminf_{\varepsilon \to 0} \min_{z \in \partial\Omega} \big(g - p_k\big)(z) + p_k(x) \geq g(x_k) - p_k(x_k) \geq g(x_k),$$

whence $u_*(x) \geq g(x)$. This in turn gives $u^* \leq g \leq u_* \leq u^*$ on $\partial\Omega$ as asserted.

Step 4: Comparison. To prove that $u^* = u_*$ in $\overline{\Omega}$ we use the comparison principle for (2.4.6) as in Theorem 2.4.7 (uniform convergence). Since u^* and u_* are a subsolution and supersolution respectively of (2.4.6) and they agree on the boundary, we can deduce that $u^* \leq u_*$ in $\overline{\Omega}$. Combining with $u^* \geq u_*$, by definition, this results in $u^* = u_*$ in $\overline{\Omega}$.

Step 5: Uniform Convergence. This is identical to Theorem 2.4.7 and concludes the proof. □

3.3 Numerical Examples: Accurate Scheme

In this section we illustrate the improved performance of the accurate operator. We present in Table 3.1 the L^∞ error of the solution of the accurate operator for the following two examples of Section 2.8:

Smooth Hessian:

$$u(x) = e^{|x|^2/2}, \quad f(x) = (1 + |x|^2)e^{|x|^2} \quad \forall x \in \Omega.$$

Discontinuous Hessian:

$$u(x) = \frac{1}{2}\left(\max(|x - x_0| - 0.2, 0)\right)^2, \quad f(x) = \max\left(1 - \frac{0.2}{|x - x_0|}, 0\right) \quad \forall x \in \Omega.$$

We compare with the respective errors for the monotone operator for the same number of degrees of freedom, N. We note that in the comparison we employ the notation h for the mesh-size used for the monotone operator, hence the respective size of the accurate operator for the same number of degrees of freedom is $2h$. The computational domain is $\Omega = \Omega_h = [0, 1]^2$ and for the second example $x_0 = (0.5, 0.5)$ is the center of the domain. We note here that, since our computational domain is not curved and it is the same as Ω, the map of the Isoparametric Finite Elements is the identity. In a general case, where Ω would be uniformly convex instead, this would no longer be the case. Despite the generality with which we introduce our

method, we choose to present our experiments for this domain, in order to easily compare them with the results in Section 2.8.

DoFs	P: # of points	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	Newton steps
N= 4225, $h = 2^{-6}$	56	2.8 10^{-3}	1.91 10^{-4}	6
N=16641, $h = 2^{-7}$	88	1.5 10^{-3}	8.60 10^{-5}	5
N=66049, $h = 2^{-8}$	144	7.8 10^{-4}	4.17 10^{-5}	5
N= 263169, $h = 2^{-9}$	224	6.4 10^{-4}	2.42 10^{-5}	7

DoFs	P: # of points	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	Newton steps
N= 4225, $h = 2^{-6}$	40	1.9 10^{-3}	2.48 10^{-4}	9
N=16641, $h = 2^{-7}$	56	9.0 10^{-4}	1.51 10^{-4}	12
N=66049, $h = 2^{-8}$	72	5.7 10^{-4}	8.34 10^{-5}	14
N= 263169, $h = 2^{-9}$	96	3.83 10^{-4}	5.31 10^{-5}	20

Table 3.1: L^∞ error for monotone and accurate operators for the smooth Hessian (top) and the discontinuous Hessian (bottom). P : number of points $x_i \pm \delta v_j$ used for $T_{\varepsilon,a}$ at each $x_i \in \mathcal{N}_h^0$; $P = 8(D-1)$, where D is the number of directions v_j in a quarter circle, as determined by the value of θ. The operator has an accuracy of one to two orders higher compared to $T_{\varepsilon,m}$. The number of Newton iterations corresponds to the accurate operator.

We observe that the accurate operator exhibits a significant improvement beyond one order of magnitude in the smooth case. Although there is no theoretical result to support this, it can be explained by the regularity of this example and the higher order of the consistency error in Lemma 3.1.2 (consistency of $T_{\varepsilon,a}[\mathcal{I}_h^2 u]$). At this point, it is worth noting that the accuracy improvement for the monotone operator for the smooth example exhibits saturation in the last refinement. Upon examining where the error is bigger, we observe that it appears on the boundary

layer that arises from the definition of $T_{\varepsilon,m}$. As shown in Theorem 2.7.1 (rates of convergence for classical solutions), this error does not follow the consistency of the solution, but is bounded through a barrier argument by $C\ \|u\|_{W^2_\infty(\Omega)}\ \delta_m$. This is true for the above case too.

As far as the piecewise smooth example is concerned, we observe again an increase in accuracy of $T_{\varepsilon,a}$ over $T_{\varepsilon,m}$, despite the fact that the predicted error in Theorem 2.7.4 (degenerate forcing $f \geq 0$) for this example is determined by the dimension of the problem rather than the regularity of the solution. In contrast to Table 2.1, this time we notice an increase in the number of Newton iterations with each refinement. This may be attributed to the lack of monotonicity of $T_{\varepsilon,a}$.

3.3.1 Computational Remarks

At this point we explain some implementation features of the accurate method, that are also relevant for the filtered scheme.

Sparsity: By comparing the approximations of second directional derivatives that we use in the monotone operator (2.1.1) and accurate operator (3.1.1), we immediately observe that the latter requires almost double the amount of points in the evaluation of $\nabla^2_{\delta,a}\mathbf{U}_\varepsilon$ for each point $x_i \in \mathcal{N}^0_h$ and direction $v_j \in \mathbb{S}_\theta$. This results in a sparsity pattern with a wider bandwidth and more non-zero elements, since for each extra point $x_i \pm \frac{\delta}{2}v_j$, we need to use the degrees of freedom of the simplex where $x_i \pm \frac{\delta}{2}v_j$ belong. To illustrate this, we present in Figure 3.2 the sparsity of the gradient matrix for both the monotone and the accurate operator for the same

amount of degrees of freedom, in particular $h = 2^{-6}$ and $N = 4225$.

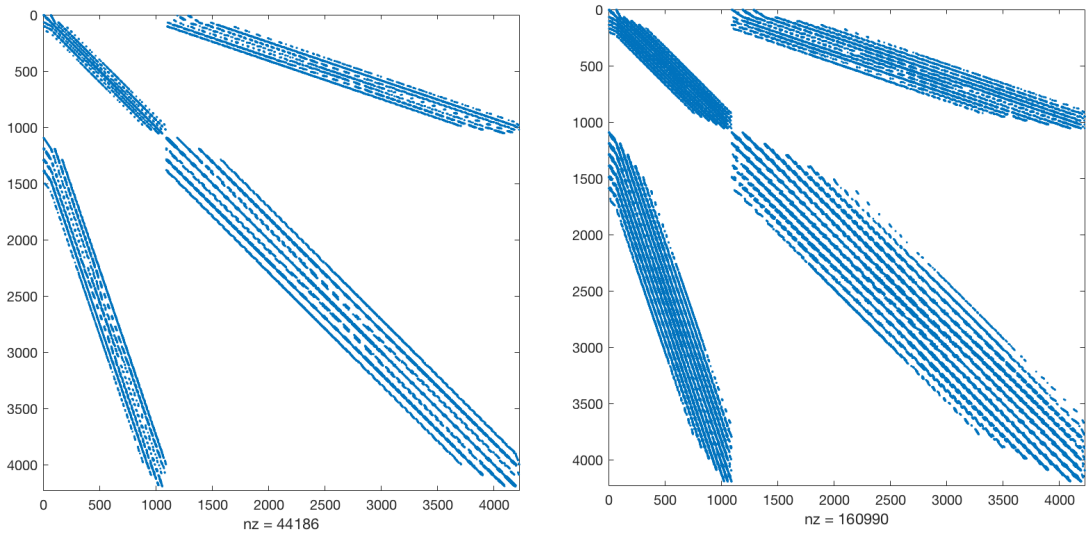


Figure 3.2: Sparsity of the Jacobian for $C^{1,1}$ example: a) monotone operator, 44, 186 non-zero elements (left), b) accurate operator, 160, 990 non-zero elements (right).

Solver: The original operator $T_{\varepsilon,m}$ was implemented for the results in Section 2.8 using a direct solver for the linear system, i.e. Matlab's backslash operator. However, for very fine meshes and cases where we have many directions an iterative method like conjugate gradient is more appropriate. As illustrated in Figure 3.2 and discussed earlier, the accurate operator's sparsity pattern includes a wide bandwidth of non-zero elements. This makes a direct solver a less favorable option for very fine meshes. Choosing a small tolerance for the conjugate gradient method seems to provide similar results as the direct solver and hence not affect the computationally observed errors. The design of suitable preconditioners is important, but remains an open issue.

3.4 Numerical Examples: Filtered Scheme

We now combine the monotone and accurate operator into the filtered scheme in (3.2.1). We explore the accuracy of the new operator and determine the parts of the domain where the monotone operator is enforced instead of the accurate one. We call this the *active set* of the filtered scheme. Before doing that, we first discuss briefly the choice of σ in F_σ and fix a value for our implementation.

3.4.1 Choice of σ in F_σ:

We observe that if $\sigma = 0$ in (3.2.3), we obtain the discontinuous function

$$F_0(x) := \begin{cases} x, & |x| \leq 1 \\ 0, & |x| > 1, \end{cases}$$

while for any choice $\sigma > 0$ the filter function F_σ is continuous. We recall that continuity of F_σ is only used in Lemma 3.2.3 (existence and stability) in order to apply Lemma 2.2.1 (existence, uniqueness and stability) that requires continuity for the right hand side of the equation. We note, however, that our results do not require any extra restriction on the value of σ. This means that we can choose σ as small as we want and then perform our asymptotic analysis (numerical experiments) for $h \to 0$; in practice, we take $\sigma = 10^{-4}$. This leads to a very small support for the two last cases in (3.2.3), which in practice never shows up. This is desirable to us, since our goal is to give full control to $T_{\varepsilon,a}$ in smooth parts of the solution, while we want to employ the monotone operator otherwise. However, the two cases that correspond to $s \in [1, 1+\sigma] \cup [-1-\sigma, -1]$ in (3.2.3) result in a linear combination

of the two operators. For example, if $\sigma = 1$ and $-2\tau \leq A[\mathbf{U}_\varepsilon](x_i) \leq -\tau$, we obtain

$$\begin{aligned} T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) &= T_{\varepsilon,m}[u_\varepsilon^1](x_i) - (T_{\varepsilon,a}[u_\varepsilon^2](x_i) - T_{\varepsilon,m}[u_\varepsilon^1](x_i)) - 2\tau \\ &= 2\ T_{\varepsilon,m}[u_\varepsilon^1](x_i) - T_{\varepsilon,a}[u_\varepsilon^2](x_i) - 2\tau \end{aligned}$$

A form like the above for $\sigma = 1$ would result in the respective component $\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i)$ of the Jacobian of the method to be calculated using

$$2\ \nabla T_{\varepsilon,m}[u_\varepsilon^1](x_i) - \nabla T_{\varepsilon,a}[u_\varepsilon^2](x_i),$$

where $\nabla T_{\varepsilon,m}$ and $\nabla T_{\varepsilon,a}$ are associated with the Jacobian matrix of the monotone and accurate operators at x_i. This could create problems with the implementation of the semi-smooth Newton iteration we employ (see Section 2.8.1). This was also observed in [44], where the authors acknowledge the prospects of a poorly conditioned or ill-posed system and thereby decide to approximate the above component of the Jacobian by

$$\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) \approx 2\ \nabla T_{\varepsilon,m}[u_\varepsilon^1](x_i).$$

We prefer, instead, to avoid these issues altogether by choosing a rather small value of σ. This retains the existence and convergence results from the theory in Section 3.2, while allowing for an implementation that resembles the one corresponding to the discontinuous filter F_0 for which

$$\nabla T_{\varepsilon,f}[\mathbf{U}_\varepsilon](x_i) = \begin{cases} \nabla T_{\varepsilon,a}[u_\varepsilon^2](x_i), & |A[\mathbf{U}_\varepsilon](x_i)| \leq \tau \\ \nabla T_{\varepsilon,m}[u_\varepsilon^1](x_i), & \text{otherwise.} \end{cases}$$

We stress here that this notation corresponds to the process followed in order to construct the component of the Jacobian of the system that corresponds to x_i,

using the Jacobian matrices of the monotone and accurate operator. This process is still in the context of semi-smooth Newton, as in Section 2.8.1, since we treat the non-differentiability of F_σ at $s = \pm\tau$ and $s = \pm(1+\sigma)\tau$ similarly to the max and min functions in Section 2.8.1.

3.4.2 Numerical Experiments:

Before presenting our results in detail, using the same examples as in Section 3.3, we make two general observations.

- **Choice of filter scale τ:** We stress that the behavior of the scheme depends greatly on the choice of the filtered scale τ. A bigger τ allows the accurate operator to take control, while the monotone operator guarantees convergence when the accurate operator has a very large consistency error. On the other hand, smaller values of τ lead to the presence of the active set, where the monotone operator becomes dominant. Since we measure the error in the L^∞ norm, a big active set could prevent us from achieving better accuracy than with $T_{\varepsilon,m}$. Although there is no clear recipe for choosing τ, the definition (3.2.1) of $T_{\varepsilon,f}$ indicates that, in order for the accurate operator to be active at a point x_i, we need $|(T_{\varepsilon,a}[u_\varepsilon^2] - T_{\varepsilon,m}[u_\varepsilon^1])(x_i)| \leq \tau$. Consequently, τ has to be greater than the truncation error of the monotone operator, since, otherwise, we may have an active set even in parts of the domain where the solution is smooth. We follow this approach to generate Tables 3.2 and 3.3 and Figure 3.6.

- **Boundary Layer:** We observe in our experiments that the active set contains

points near the boundary $\partial\Omega$. This can be explained by the fact that the monotone operator is of order one near the boundary, where $\delta_m \approx h_m = h$ (see Lemma 2.3.2 (consistency of $T_{\varepsilon,m}[\mathcal{I}_h^1 u]$)) while the accurate operator becomes of order $\approx \frac{h^3}{h^2} \approx h$ (see Lemma 3.1.2 (consistency of $T_{\varepsilon,a}[\mathcal{I}_h^2 u]$)), hence the two operators may behave very differently.

We now document the performance of the filtered scheme for the two examples and investigate the effect of τ in the size and location of the active set. As in Section 3.3, the performance of the scheme is easily compared with the monotone and accurate operators, since $\Omega_h = \Omega$ for all $h > 0$.

Smooth Hessian: We start by illustrating the error estimates for the smooth example for $\tau = 6e^2 h$ in Table 3.2. This choice is motivated by the theoretical truncation error of $T_{\varepsilon,m}$ for the corresponding choice of δ_m. For this example we use the original, symmetric, filter F_σ with $\sigma = 10^{-4}$. This falls under the existence and convergence results of Section 3.2, since $f(x) = (1 + |x|^2)e^{|x|^2} \geq e > \tau$ in Ω for all values of h that are used in Table 3.2. We observe a small active set, with relative size around $1\% - 2\%$ for all refinements.

h	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	$T_{\varepsilon,f}$	$T_{\varepsilon,f}$ Newton	Active Set
$h = 2^{-5}$	$5.4\ 10^{-3}$	$5.16\ 10^{-4}$	$1.01\ 10^{-3}$	6	17
$h = 2^{-6}$	$2.8\ 10^{-3}$	$1.91\ 10^{-4}$	$3.16\ 10^{-4}$	6	57
$h = 2^{-7}$	$1.5\ 10^{-3}$	$8.60\ 10^{-5}$	$1.20\ 10^{-4}$	6	214
$h = 2^{-8}$	$7.8\ 10^{-4}$	$4.17\ 10^{-5}$	$5.00\ 10^{-5}$	7	918
$h = 2^{-9}$	$6.4\ 10^{-4}$	$2.42\ 10^{-5}$	$1.99\ 10^{-5}$	8	5035

Table 3.2: Smooth Hessian, $\tau = 6e^2h$. $T_{\varepsilon,f}$ has up to an order smaller L^∞ error than $T_{\varepsilon,m}$. Active Set documents the number of points where the lower-accuracy $T_{\varepsilon,m}$ is used. The amount of Newton iterations for the filtered scheme is also presented.

The active set for the above choice of τ and $h = 2^{-8}$ is displayed in Figure 3.3. We observe the aforementioned boundary layer, especially close to the upper-right corner, most likely due to lack of symmetry of u and its derivatives.

Figure 3.3: Active set for smooth example: $\tau = 6e^2h_m$. The active set is concentrated on the upper corner and, less, on the lower and left side. We illustrate the upper corner and lower side in a zoom-in.

Discontinuous Hessian: For the $C^{1,1}$ example we choose $\tau = 0.62\ h^{2/5}$. This is motivated by the theoretical consistency error of the monotone operator for the choice $\delta_m = 2.44\ h$. We use the non-symmetric filter $\tilde{F}$ in (3.2.6). We observe that

the active set is located on the circle of discontinuity of the Hessian and near the boundary. We also see that the error in the L^∞ norm is slightly better than in Table 3.1. We observe computationally that coarser choices of τ, as for example $\tau = O(h^{1/5})$, correspond to an empty active-set. We note that, in contrast to the smooth example, we now observe a gradual increase of the relative size of the active set. This is more prominent in the last three refinements, where this relative size increases from around 1.5% to 2.9% and then to 4.1%.

h	$T_{\varepsilon,m}$	$T_{\varepsilon,a}$	$T_{\varepsilon,f}$	$T_{\varepsilon,f}$ Newton	Active Set
$h = 2^{-5}$	$4.0\ 10^{-3}$	$5.67\ 10^{-4}$	$5.50\ 10^{-4}$	5	22
$h = 2^{-6}$	$1.9\ 10^{-3}$	$2.48\ 10^{-4}$	$2.48\ 10^{-4}$	8	8
$h = 2^{-7}$	$9.0\ 10^{-4}$	$1.51\ 10^{-4}$	$1.40\ 10^{-4}$	12	248
$h = 2^{-8}$	$5.7\ 10^{-4}$	$8.34\ 10^{-5}$	$7.58\ 10^{-5}$	14	1904
$h = 2^{-9}$	$3.83\ 10^{-4}$	$5.31\ 10^{-5}$	$4.89\ 10^{-5}$	16	10825

Table 3.3: Discontinuous Hessian, $\tau = 0.62\ h^{2/5}$. $T_{\varepsilon,f}$ has the lowest L^∞ error out of the three. It's one order smaller than $T_{\varepsilon,m}$ and slightly better than $T_{\varepsilon,a}$. The latter is explained in Figure 3.4.

In order to explain the smaller errors of Table 3.3 with respect to Table 3.1, we present in Figure 3.4 a contour plot with $|u_\varepsilon - u|$ for the accurate scheme and $h_a = 2h = 2^{-6}$. In the same figure we illustrate the corresponding active set associated with the filtered scheme and observe that the circle of discontinuity of the Hessian, which dominates the active set, is precisely the part of the domain where the accurate operator exhibits the biggest error. This explains why using the monotone operator at these points increases the accuracy and provides experimental

justification of the filtered scheme.

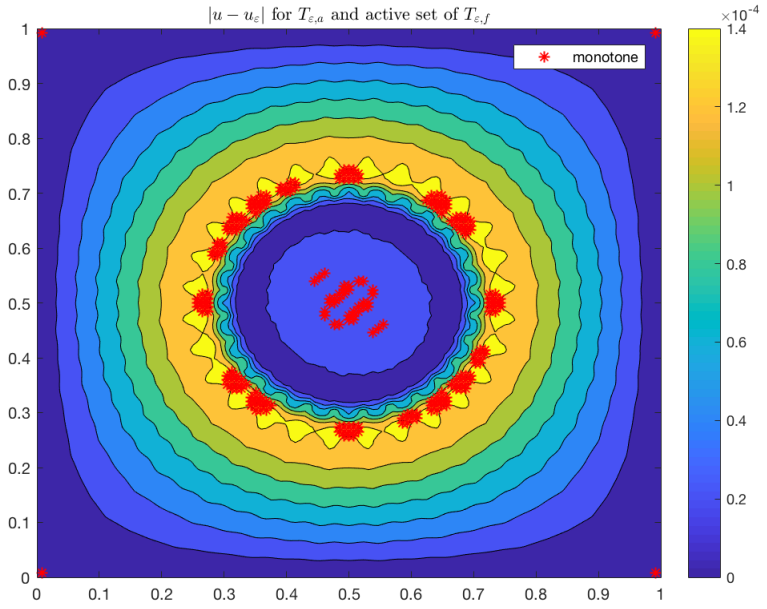


Figure 3.4: Error $|u_\varepsilon - u|$ of $T_{\varepsilon,a}$ for $h_a = 2h = 2^{-6}$ and respective active set for filtered scheme. We observe that many points on the circle of discontinuity of the Hessian are in the active set and correspond to the highest absolute error for $T_{\varepsilon,a}$.

We now explore further the behavior of the active set. We illustrate it on Figure 3.5 (left) for the final iteration that corresponds to $h = 2^{-6}$. We observe that it includes the area δ_m- away of the circle of discontinuity, the center of the domain, where $f = 0$ and the problem degenerates, and a small boundary layer at the corners of the domain.

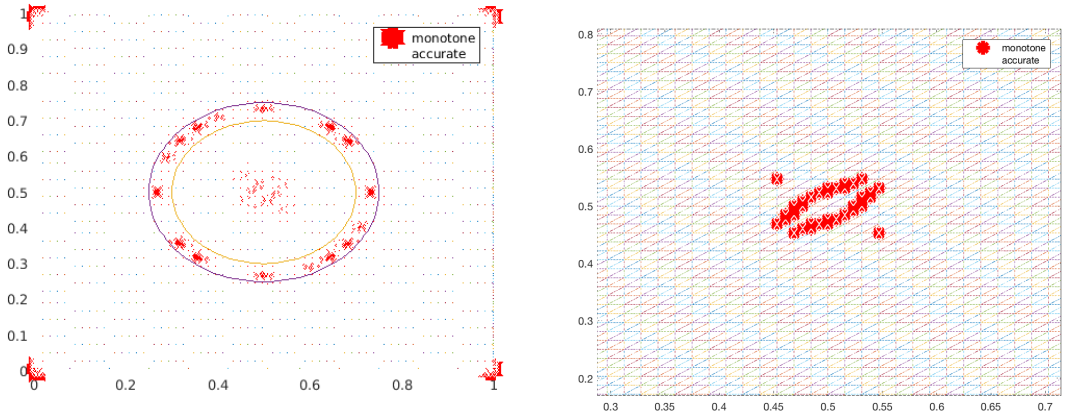


Figure 3.5: (a) Active set for discontinuous Hessian: $h = 2^{-6}$, $\tau = 0.62\ h^{2/5}$ and $\tilde{F}_\sigma$ (left). The circles correspond to $\{|x - x_0| = 0.2\}$ and $\{|x - x_0| = 0.2 + \delta_m\}$. (b) Active set for a space-dependent filter function (right).

That the active set is focused around the circle of discontinuity of the Hessian is a desirable property of the filtered scheme, whereas points near the corners can be explained by the disparate consistency errors of the operators near the boundary. In order to understand the presence of active points within $\{|x - x_0| \leq 0.2\}$, we experiment with a space-dependent filter function: we restrict the use of the accurate operator only for those nodes $x_i \in \mathcal{N}_h^0$ such that $f(x_i) \leq \tau$. This function helps to shed some light on the behavior of the two operators. We present a zoomed-in illustration of the active set that corresponds to the same parameters as in Figure 3.5, but using this space-dependent filter on the right part of Figure 3.5. We observe that the active set is now only restricted to the center of the circle. By examining the values of the two operators in the active set, we realize that the monotone operator is of order 10^{-18}, while the accurate operator is of order 10^{-8}, which explains

why this active set remains present. Since $f = 0$ in $\{|x - x_0| \leq 0.2\}$, we want to choose the operator whose value is closer to zero. This is why both (3.2.6) and its space-dependent version force us to use the monotone operator whenever the accurate operator is "positive enough" in this region. For (3.2.6) this includes parts of the Hessian discontinuity and the center of the circle, while for the space-dependent version of the filter, this is only enforced in the center of the domain. We see that, although at first the definition of (3.2.6) may raise concerns about restricting the accurate operator too much, it actually allows us to employ the monotone operator precisely on the parts of the domain where it is needed. This is portrayed in Figure 3.4 which illustrates that the active set introduced by (3.2.6) involves the points of lowest accuracy of $T_{\varepsilon,a}$. This is why in Figure 3.6 we observe smaller errors for the filtered operator than for the accurate.

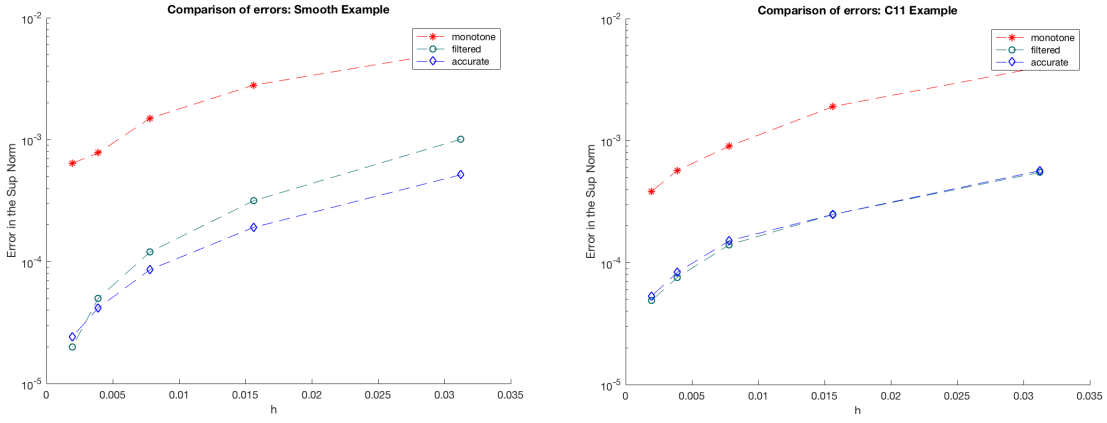


Figure 3.6: Error for monotone, accurate and filtered operator: a) Smooth example, for filtered scheme: $\tau = 4e^2 h$ (left), b) $C^{1,1}$ example, for filtered scheme: $\tau = 0.41\ h^{2/5}$ (right).

Conclusions: We see that in both examples the choice of the filtered scale

τ is not a trivial task and it can have a dramatic effect on the outcome. Since the accurate operator performs better close to the boundary, it is expected that for many choices of τ there will be a boundary layer, where the monotone operator will be active. Unfortunately, this is due to the bad behavior of the monotone operator near the boundary and not to the filter capturing a singularity. This can be explained by the simplicity of the filter functions being used, because they compare only the values of the two operators and do not take into account the respective value of the right hand side.

A crucial observation is that the convergence result allows for a great deal of flexibility in the choice of τ. We can always choose τ to be relatively big with respect to ε, but still satisfy $\tau \to 0$ as $\varepsilon \to 0$. This results in the accurate operator being the one that is always active, hence allowing us to fully exploit the higher accuracy it offers, without sacrificing the convergence result. We do not present a table for this case, because it corresponds exactly to the results of Section 3.3. Instead we present two comparative figures for the errors due to the monotone, the accurate and the filtered scheme, which correspond to the results from Table 3.2 and Table 3.3. We see that the filtered scheme is much more efficient than the monotone one even with the presence of boundary layers. It even outperforms the accurate scheme in the non-smooth case.

3.4.3 Choice of Solver:

The discussion in Section 3.3.1 indicates that we need to take into account the sparsity of the resulting gradient matrix and possibly choose between a direct and an iterative solver. This can be treated differently based on the example at hand. For instance, when we have a strictly positive right hand side, we may try to approach the choice of δ_m and θ_m and even of δ_a and θ_a too (with some modifications) based on Theorem 2.7.1 (rates of convergence for classical solutions). On the other hand, a right hand side that touches zero may yield a choice of scales of the order described in Theorem 2.7.4 (degenerate forcing $f \geq 0$), even if we do not know the exact values of the ensuing constants. The first choice is more likely to lead to the use of more directions, hence a bigger bandwidth for the gradient matrix. Moreover, if f is smooth enough that we are able to assume that the solution u is $C^2(\Omega)$ and strictly convex, then the gradient matrix is more likely to be positive definite, hence, in conjunction with the reduced sparsity, it may be preferable to use an iterative solver like the conjugate gradient. On the other hand, other than the trivial case of constants, whenever f touches zero we cannot expect strict convexity and global regularity for u. In that case, given also that Theorem 2.7.4 (degenerate forcing $f \geq 0$) suggests a bigger choice of θ_m which leads to a sparser matrix, it may be safer to choose a direct solver.

3.4.4 Implementation Challenges:

The above exposition shows that the filtered scheme provides the desirable combination of provable convergence and increased accuracy and adds to the family of similar approaches, such as [44]. However, this improvement is not without challenges. In particular, implementing the filtered scheme requires special care in the following three aspects.

(i) Similarly to $T_{\varepsilon,m}$, for each node $x_i \in \mathcal{N}_h^0$ we need to locate the appropriate simplex where $x_i \pm \frac{\delta}{2} v_j$ and $x_i \pm \delta v_j$ belong to, in order to calculate the second differences for the monotone and accurate operator. This is a process that now needs to take place for both operators. We employ the searching methods of FELICITY to achieve this in minimal time, as in Section 2.8.

(ii) In Section 2.8.1 we also mention that we choose an initial guess that corresponds to a solution of an appropriate Laplacian problem for the coarsest mesh and then at each subsequent refinement, we use the interpolant of the previous mesh as the initial guess. This is a very efficient choice that we want to preserve for both the accurate and the filtered scheme. To achieve this, we need to interpolate a quadratic function on a finer mesh. We thus create the new mesh at the end of each iteration and use the searching and interpolation capabilities of FELICITY, in order to associate each node of the finer mesh with a simplex of the coarser mesh and interpolate the solution there.

(iii) In the definition of the filter operator in (3.2.1) we use $u_\varepsilon^1 \in \mathbb{V}_h^1$ and $u_\varepsilon^2 \in \mathbb{V}_{2h}^2$

that correspond to a piecewise linear function in $\mathbb{V}_h^1$ and a piecewise quadratic function in $\mathbb{V}_{2h}^2$, each defined on a mesh $\mathcal{T}_h^1$ and $\mathcal{T}_{2h}^2$, respectively, and each involving the same number, N, of degrees of freedom. From an implementation viewpoint, when we construct the meshes and the respective finite element spaces, these degrees of freedom do not necessarily have the same numbering and in practice, they often do not. Consequently, in order to compare $T_{\varepsilon,m}[u_\varepsilon^1]$ and $T_{\varepsilon,a}[u_\varepsilon^2]$ at each nodal point, we need to create a communication between the two meshes and the degrees of freedom. However, taking advantage of the efficiency of Matlab for vectorized quantities, creating a map between the degrees of freedom for the two meshes takes minimal time.

3.5 Conclusions

In this Chapter we introduce two methods to solve the Monge-Ampère equation (1.1.1). The first is an accurate scheme that hinges on quadratic interpolation and a higher-order approximation of the directional derivatives in (1.1.2). It exhibits errors in the L^∞ norm of one to two orders of magnitude lower than the monotone operator introduced in Chapter 2. However, higher order accuracy comes at the cost of monotonicity and prevents us from proving convergence results for this operator. The second method circumvents this issue by combining the monotone and the accurate operators into a filtered scheme. This yields convergence to the viscosity solution using the stability and monotonicity of the monotone operator. We employ two filter functions that allow us to detect parts of the domain where the

accurate operator could under-perform due to lack of regularity of the solution, as for example in the degenerate example that we present. We explore the two methods computationally and illustrate the increased performance of both by comparing them with the numerical results from Section 2.8.2. Lastly, we investigate the effect of the filter function and the filter scale and discuss some computational challenges of the method.

Chapter 4: DG Approach to Large Bending Deformations with Isometry Constraint

In this Chapter we introduce a Discontinuous-Galerkin (DG) approach to find minimizers of the single layer functional

$$E[y] = \frac{1}{2}\int_\Omega |D^2 y|^2 - \int_\Omega fy \tag{4.0.1}$$

in the set of admissible functions under the non-linear isometry constraint

$$\nabla y^T \nabla y = I \text{ a.e.,} \tag{4.0.2}$$

We recall that this implies that we are searching for minimizers in the space

$$\mathbb{A} = \left\{ y \in H^2(\Omega)^3 \ : \ (\nabla y)^T \nabla y = I \quad \text{a.e. in} \quad \Omega, \quad y = g, \ \nabla y = \Phi_D \quad \text{on} \quad \Gamma_D \right\}. \tag{4.0.3}$$

We prove the $\Gamma-$convergence of our discrete energy functional E_h, defined later in (4.2.9) in Section 4.2, to the continuous energy E. We also show that absolute minimizers of E_h converge to absolute minimizers of (4.0.1).

4.1 Derivation of the Energy Functional

Before we introduce our discretization of the problem, we motivate our analysis by briefly illustrating how (4.0.1) can be heuristically derived from three-dimensional

hyperelasticity. To this end, we follow an approach similar to [9, 11]. In three dimensions, the pure bending of a plate with thickness s can be characterized by an energy functional that scales like s^3:

$$\tilde{E}_s[u_s] = \int_{\Omega_s} W(\nabla u_s)\, dx - \int_{\Omega_s} \tilde{f}_s u_s\, dx \;\approx s^3,$$

where $\Omega_s = \Omega \times (-s/2, s/2)$, u_s is the three dimensional deformation of the plate and $\tilde{f}_s$ is a forcing term that scales like s^2, i.e. we can write $\tilde{f}_s(x', x_3) = s^2 f_s(x', x_3)$. We also call $f(x') := \frac{1}{s}\int_{-s/2}^{s/2} f_s(x', x_3)\, dx_3$. We now introduce the rescaled functional

$$E_s[u_s] = s^{-3} \int_{\Omega_s} W(\nabla u_s)\, dx,$$

where for simplicity we omit the forcing term and examine it later. In order to simplify the exposition, we assume that for all $F \in \mathbb{R}^{3\times 3}$

$$W(F) = \operatorname{dist}^2(F, \mathrm{SO}(3)),$$

where $\mathrm{SO}(3) = \left\{F \in \mathbb{R}^{3\times 3} \;:\; F^T F = I,\ \det F = 1\right\}$. It is discussed in [9, 11] that W can have a more general form that we do not consider here. It is also shown in [9, 11] that for F in a neighborhood of SO(3) we have

$$W(F) \approx \frac{1}{4}|F^T F - I|^2,$$

which we use without a proof. In order to derive the limiting functional for $s \to 0$, we employ the Kirchhoff assumption and write

$$u_s = u(x', x_3) = y(x') + x_3 b(x'),$$

where x' corresponds to the coordinates (x_1, x_2) and b is normal to the surface parametrized by y. This means that y is the deformation of the middle plane Ω of

the plate and that segments normal to Ω are mapped to straight segments that are normal to the deformed surface. Using the notation ∇' for differentation along x' we have that

$$\nabla u = [\nabla' y, b] + [x_3 \nabla' b, 0].$$

Combining everything, we see that

$$E_s[u] \approx \frac{s^{-3}}{4} \int_{\Omega_s} |(\nabla u)^T \nabla u - I|^2 \, dx.$$

Using the notation

$$g_s = s^{-1}((\nabla' y)^T \nabla y - I), \quad H = -(\nabla' y)^T \nabla' b, \quad k = (\nabla' b)^T b,$$

we can rewrite this as

$$E_s[u] \approx \frac{s^{-3}}{4} \int_{\Omega_s} \left| \begin{bmatrix} sg_s - 2x_3 H + x_3^2 k & (\nabla' b)^T b \\ b^T(\nabla' b) & |b|^2 - 1 \end{bmatrix} \right|^2 dx.$$

To guarantee that the above quantity is bounded as $s \to 0$, we need to force $|b|^2 = 1$, which implies that $b^T \nabla' b = 0$ and

$$E_s \approx \frac{s^{-3}}{4} \int_{\Omega_s} |sg_s - 2x_3 H + x_3^2 k|^2 \, dx.$$

We see now that, in order for this integral to be bounded as $s \to 0$, we need to have $g_s = 0$, in which case, by integrating over x_3 and ignoring higher order terms, we obtain

$$E_s = \frac{1}{12} \int_{\Omega} |H|^2 \, dx'.$$

We emphasize that $g_s = 0$ is precisely the pointwise isometry constraint $(\nabla' y)^T \nabla' y = I$, where from now on we will no longer use the notation ∇', since all functions will

be defined in $\Omega \subset \mathbb{R}^2$. Since y is an isometry, the above functional is up to a constant equal to the Hessian term in (4.0.1). To see this, we observe that by the definition of the isometry we have that

$$\partial_i y \cdot \partial_j y = \delta_{ij}, \quad i, j = 1, 2,$$

hence

$$\partial_k y \cdot \partial_{ij} y = 0 \Longrightarrow \partial_{ij} y \parallel \nu = \partial_1 y \times \partial_2 y, \quad i, j, k = 1, 2$$

which, in turn, combined with the definition of the second fundamental form, leads to

$$h_{ij} = \partial_{ij} y \cdot \nu = \partial_{ij} y.$$

Before we conclude, we also incorporate the forcing term. Rescaling it by s^{-3}, similarly to E_s, we can see that

$$\begin{aligned} s^{-3} \int_{\Omega_s} \tilde{f}_s(x', x_3) \cdot u_s(x', x_3)\, dx &= s^{-1} \int_{-t/2}^{t/2} \int_{\Omega} f_s(x', x_3) \cdot y(x')\, dx' \\ &\quad + s^{-1} \int_{-t/2}^{t/2} \int_{\Omega} x_3\ f_s(x', x_3) \cdot b(x')\, dx'\, dx_3 \\ &\xrightarrow{s \to 0} \int_{\Omega} f(x') \cdot y(x')\, dx'. \end{aligned}$$

We note here that a similar analysis was presented in [11] for the bilayer model. Chapter 5 is devoted to extending our method to this model.

4.2 Discretization of the Energy and Key Properties

In this section we derive our discrete energy, motivated by the biharmonic problem with the same boundary conditions. We introduce the discrete space and

provide standard estimates for the functions that lie inside it. We use the estimates to prove the coercivity of the discrete energy.

4.2.1 Motivation

Our goal is to find an isometry $y \in \mathbb{A}$ that minimizes (4.0.1). Our approach hinges on a non-conforming approximation of the space of minimizers, for which we use a space of discontinuous polynomial functions. In order to motivate our choice of discrete energy, let $D_h^2 y_h$ be the piecewise Hessian of a discrete function y_h and consider first the discrete energy

$$\tilde{E}_h[y_h] := \frac{1}{2}\int_\Omega |D_h^2 y_h|^2.$$

This energy does not involve any information about the discontinuities of the first and higher derivatives of y_h, nor the boundary conditions. As a result, it does not seem possible to prove that if there exists $y \in H^2(\Omega)^3$ such that $y_h \to y$ in $L^2(\Omega)$, then

$$\tilde{E}_h[y_h] \to E[y],$$

nor to use an argument that involves the weak convergence of $D_h^2 y_h$ to $D^2 y$ to prove the lim-inf condition

$$E[y] \leq \liminf_{h\to 0} \tilde{E}_h[y_h]$$

that is required for Γ convergence. This means that we need to employ some additional terms for the discrete energy. It also implies that we need to define an appropriate discrete Hessian that will be able to capture these discontinuities and allow for the desired convergence results in a way that naturally connects with our

discrete energy. To this end, we observe that if we did not have an isometry constraint and if we had zero boundary conditions, our minimization problem would be equivalent to the weak formulation of the biharmonic problem in H_0^2. As a result, we use the biharmonic problem with our boundary conditions and integration by parts, in order to derive an appropriate formulation for our discrete problem:

$$\begin{cases} \Delta^2 y = f, \quad \text{in } \Omega, \\ \quad y = g, \quad \text{on } \partial\Omega, \\ \nabla y = \Phi_D, \quad \text{on } \partial\Omega. \end{cases} \tag{4.2.1}$$

In order to obtain an appropriate discrete energy, we first derive a consistent bilinear form a_h, which also corresponds to the derivative of the discrete energy along a variation of the discrete solution. By consistency we mean that for the discrete bilinear form $a_h(\cdot, v_h)$ and for y the strong solution of the bilaplacian, we want to have

$$a_h(y, v_h) = \int_\Omega f v_h \tag{4.2.2}$$

for all v_h in the discrete space.

4.2.2 Derivation

In order to derive the scheme, we start with the bilaplacian of y and test with a discontinuous function ϕ. Before we proceed, we extend the boundary conditions g and Φ_D to the whole domain (without relabeling) and approximate them by discontinuous interpolants g_h and $\Phi_{D,h}$. We define the average of a discontinuous function

ϕ along an edge e to be

$$\{\phi\} := \begin{cases} \frac{\phi^+ + \phi^-}{2}, & e \subset \Omega \\ \\ \phi, & e \subset \partial\Omega \end{cases}$$

where $\phi^+ = \phi|_{T_1}$ and $\phi^- = \phi|_{T_2}$ for T_1, T_2 the two elements that share e. We also define the jumps of the test function ϕ_h as

$$[\phi_h] := \begin{cases} \phi_h^+ - \phi_h^-, & e \subset \Omega \\ \\ \phi_h, & e \subset \partial\Omega \end{cases} \tag{4.2.3}$$

and

$$[\nabla_h \phi_h] := \begin{cases} \nabla_h \phi_h^+ - \nabla_h \phi_h^-, & e \subset \Omega \\ \\ \nabla_h \phi_h, & e \subset \partial\Omega, \end{cases} \tag{4.2.4}$$

We proceed now with the derivation of the scheme. We assume for simplicity that both y and ϕ are scalar, since we can then apply the process component-wise for $y \in \mathbb{R}^3$. Although we do not introduce the discrete space yet, we employ a mesh $\mathcal{T}$. The mesh could made of triangular or quadrilateral elements. We call $\mathcal{E}^0$ and $\mathcal{E}^b$ the interior and boundary edges and Σ^i and Σ^b the interior and boundary skeleton, respectively. We assume that ϕ is differentiable in each cell as required in the derivation. Integration by parts yields:

$$\begin{aligned} (\Delta^2 y, \phi)_{L^2(T)} &= \sum_{i,j} \int_T \partial_{iijj} y \ \phi = \sum_{i,j} \left(- \int_T \partial_{iij} y \ \partial_j \phi + \int_{\partial T} \partial_{iij} y \ \nu^j \ \phi \right) \\ &= \sum_{i,j} \left(\int_T \partial_{ij} y \ \partial_{ij} \phi - \int_{\partial T} \partial_{ij} y \ \nu^i \ \partial_j \phi + \int_{\partial T} \partial_{iij} y \ \nu^j \ \phi \right) \\ &= \int_T D^2 y \ : \ D^2 \phi - \int_{\partial T} \partial_\nu (\nabla y) \ \cdot \ \nabla \phi + \int_{\partial T} \partial_\nu (\Delta y) \ \phi. \end{aligned}$$

We sum over all elements to obtain

$$(\Delta^2 y, \phi)_{L^2(\Omega)} = \sum_{T\in\mathcal{T}} \int_T D^2 y : D^2\phi - \sum_{e\in\mathcal{E}^0} \int_e [\partial_\nu(\nabla y)\nabla\phi] - \sum_{e\in\mathcal{E}^b} \int_e \partial_\nu(\nabla y)\nabla\phi$$
$$+ \sum_{e\in\mathcal{E}^0} \int_e [\partial_\nu(\Delta y)\phi] + \sum_{e\in\mathcal{E}^b} \int_e \partial_\nu(\Delta y)\phi.$$

We now employ the following equality for the interior edges

$$a^+b^+ - a^-b^- = \frac{a^+ + a^-}{2}(b^+ - b^-) + (a^+ - a^-)\frac{b^+ + b^-}{2},$$

which leads to

$$(\Delta^2 y, \phi)_{L^2(\Omega)} = \sum_{T\in\mathcal{T}} \int_T D^2 y : D^2\phi - \sum_{e\in\mathcal{E}^0} \left(\int_e \{\partial_\nu \nabla y\} [\nabla\phi] + \int_e [\partial_\nu \nabla y] \{\nabla\phi\} \right)$$
$$- \sum_{e\in\mathcal{E}^b} \int_e \partial_\nu(\nabla y)\nabla\phi + \sum_{e\in\mathcal{E}^0} \left(\int_e \{\partial_\nu(\Delta y)\} [\phi] + \int_e [\partial_\nu(\Delta y)] \{\phi\} \right)$$
$$+ \sum_{e\in\mathcal{E}^b} \int_e \partial_\nu(\Delta y)\phi.$$

Owing to the regularity of y and the fact that $\Delta^2 y = \Delta(\Delta y) = f \in L^2(\Omega)$, we have that $[\partial_\nu(\nabla y)] = 0,\ [\partial_\nu(\Delta y)] = 0$ on $\mathcal{E}^0$, hence, the above expression becomes:

$$(\Delta^2 y, \phi)_{L^2(\Omega)} = \sum_{T\in\mathcal{T}} \int_T D^2 y : D^2\phi - \sum_{e\in\mathcal{E}} \int_e \{\partial_\nu \nabla y\} [\nabla\phi] + \sum_{e\in\mathcal{E}} \int_e \{\partial_\nu(\Delta y)\} [\phi].$$

This implies that if we define the bilinear form to be

$$a_h(y,\phi) = (D_h^2 y, D_h^2\phi)_{L^2(\Omega)} - (\{\partial_\nu \nabla y\}, [\nabla\phi])_{L^2(\Sigma)} + (\{\partial_\nu(\Delta y)\}, [\phi])_{L^2(\Sigma)},$$

then it is consistent in the sense of (4.2.2) for all discontinuous ϕ. We observe, however, that this form is neither symmetric nor coercive with respect to i.e. the broken H^2-norm

$$\|D_h^2\phi_h\|^2_{L^2(\Omega)} = \sum_{T\in\mathcal{T}} \|D^2\phi_h\|^2_{L^2(T)}.$$

To fix this, while still preserving the consistency of the form, we add terms that are zero for the strong solution y. In order to do that in a way consistent with the Nitsche approach, we rewrite the boundary conditions for y as:

$$[y] := y - g = 0, \quad [\nabla y] := \nabla y - \Phi_D = 0, \quad \text{on} \quad \partial\Omega \tag{4.2.5}$$

As a first step, in order to symmetrize the bilinear form, we add a term that corresponds to

$$(\{\partial_\nu \nabla \phi\}, [\nabla y])_{L^2(\Sigma)} .$$

In fact, this term is zero on the interior edges of the triangulation, since the gradient of the exact solution has no jump inside Ω due to the H^2 regularity. Given a boundary edge e that corresponds to an element T and using (4.2.5), the above term becomes

$$(\{\partial_\nu \nabla \phi\}, [\nabla y])_{L^2(e)} = \ (\partial_\nu \nabla \phi|_T, \nabla y|_T - \Phi_D)_{L^2(e)}$$

Similarly, we add the term

$$(\{\partial_\nu \Delta \phi\}, [y])_{L^2(\Sigma)} ,$$

which is also equal to zero for the exact solution y. Lastly, we add two stabilization terms that contribute to the stability of the scheme. They are both equal to zero for the exact solution and are denoted by S_0 and S_1

$$S_0(y, \phi) := \frac{\gamma_0}{h^3} \left([y], [\phi]\right)_{L^2(\Sigma)}$$

and

$$S_1(y, \phi) := \frac{\gamma_1}{h} \left([\nabla y], [\nabla \phi]\right)_{L^2(\Sigma)} ,$$

where γ_0, γ_1 are constants with values that are determined later. Since we only added terms that are zero for the exact solution, we obtain the following symmetric and consistent bilinear form in the sense of (4.2.2):

$$\begin{aligned} a_h(y,\phi) := & (D_h^2 y, D_h^2 \phi)_{L^2(\Omega)} \\ & - (\{\partial_\nu \nabla y_h\}, [\nabla \phi])_{L^2(\Sigma)} - (\{\partial_\nu \nabla \phi\}, [\nabla y])_{L^2(\Sigma)} \\ & + (\{\partial_\nu \Delta y\}, [\phi])_{L^2(\Sigma)} + (\{\partial_\nu \Delta \phi\}, [y])_{L^2(\Sigma)} \\ & + S_0(y,\phi) + S_1(y,\phi). \end{aligned} \tag{4.2.6}$$

4.2.3 Discretization of $E[y]$

We now introduce a discrete space of piecewise polynomial functions of degree $k \geq 2$ that are not required to be continuous in $\overline{\Omega}$. This space is not conforming, since such functions are not in $H^2(\Omega)$, or even in $H^1(\Omega)$. We define the space

$$\mathbb{V}_h^k := \left\{ v_h \in L^2(\Omega) \ : \ v_h|_T \in \mathbb{P}_k(T) \ \forall \ T \in \mathcal{T} \right\}.$$

The bilinear form (4.2.6) can be now written as:

$$\begin{aligned} a_h(y_h, v_h) := & (D_h^2 y_h, D_h^2 v_h)_{L^2(\Omega)} \\ & - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h v_h])_{L^2(\Sigma)} - (\{\partial_\nu \nabla_h v_h\}, [\nabla_h y_h])_{L^2(\Sigma)} \\ & + (\{\partial_\nu \Delta_h y_h\}, [v_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h v_h\}, [y_h])_{L^2(\Sigma)} \\ & + \frac{\gamma_1}{h}([\nabla_h y_h], [\nabla_h v_h])_{L^2(\Sigma)} + \frac{\gamma_0}{h^3}([y_h], [v_h])_{L^2(\Sigma)}, \end{aligned} \tag{4.2.7}$$

where we follow the notation $\nabla_h, \Delta_h, D_h^2$ for piecewise differentiation. In arguments restricted to single elements we can use the actual derivatives instead. We emphasize that the boundary jump terms for the discrete solution y_h involve the Nitsche

penalization

$$[y_h] := y_h - g_h, \quad [\nabla y_h] := \nabla y_h - \Phi_{D,h}, \quad \text{on} \quad \Sigma^b, \tag{4.2.8}$$

while for v_h we have

$$[v_h] := v_h, \quad [\nabla v_h] := \nabla v_h, \quad \text{on} \quad \Sigma^b.$$

We can now define the discrete energy for functions $y_h \in \left[\mathbb{V}_h^k\right]^3$ as follows:

$$\begin{aligned} E_h[y_h] := \frac{1}{2}\|D_h^2 y_h\|^2_{L^2(\Omega)} - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(\Sigma)} \\ + \frac{\gamma_1}{2}\|h^{-1/2}[\nabla_h y_h]\|^2_{L^2(\Sigma)} + \frac{\gamma_0}{2}\|h^{-3/2}[y_h]\|^2_{L^2(\Sigma)} - \int_\Omega f_h y_h, \end{aligned} \tag{4.2.9}$$

By using the notation $S_i[y_h] := S_i[y_h, y_h],\ i = 0, 1$, we can rewrite this as

$$\begin{aligned} E_h[y_h] := \frac{1}{2}\|D_h^2 y_h\|^2_{L^2(\Omega)} - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} \\ + (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(\Sigma)} + S_0[y_h] + S_1[y_h] - \int_\Omega f y. \end{aligned} \tag{4.2.10}$$

We observe that by taking the derivative of this energy along a variation $v_h \in \left[\mathbb{V}_h^k\right]^3$, we obtain (4.2.7). We also define the associated energy norm to be

$$|||y_h|||_h^2 := \|D_h^2 y_h\|^2_{L^2(\Omega)} + \|h^{-1/2}\left[\nabla_h y_h\right]\|^2_{L^2(\Sigma)} + \|h^{-3/2}[y_h]\|^2_{L^2(\Sigma)}. \tag{4.2.11}$$

In order for (4.2.9) to be meaningful with respect to the original minimization problem in (4.0.1)-(4.0.3), we define a subspace of $\left[\mathbb{V}_h^k\right]^3$ that is the discrete analogous of (4.0.3) and involves a discrete isometry constraint:

$$\mathbb{A}_h^k := \left\{ y_h \in \left[\mathbb{V}_h^k\right]^3, \quad \sum_{T\in\mathcal{T}} \left| \int_T \left((\nabla_h y_h)^T \nabla_h y_h - I \right) \right| \leq ch \quad \right\} \tag{4.2.12}$$

where c is a uniform constant, an appropriate choice of which we discuss in Section 4.5. We now use $\mathbb{A}$ and $\mathbb{A}_h^k$ to define the continuous and discrete energies again

as

$$E[y] := \begin{cases} \displaystyle\int_\Omega |D^2 y|^2 - \int_\Omega fy, & y \in \mathbb{A} \\ +\infty, & y \notin \mathbb{A}, \end{cases}$$

and

$$E_h[y_h] := \begin{cases} \dfrac{1}{2}\|D_h^2 y_h\|^2_{L^2(\Omega)} - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(\Sigma)} \\ + S_0[y_h] + S_1[y_h] - \displaystyle\int_\Omega f_h y_h, \quad y_h \in \mathbb{A}_h^k \\ +\infty, \quad y_h \notin \mathbb{A}_h^k, \end{cases}$$

to emphasize the constraints in each space.

4.2.4 Coercivity of E_h

We now prove that (4.2.9) is both coercive and continuous in $\left[\mathbb{V}_h^k\right]^3$ for γ_0, γ_1 large enough. Since the following result holds for all $y_h \in \left[\mathbb{V}_h^k\right]^3$, we do not require that $y_h \in \mathbb{A}_h^k$. We first prove some standard estimates that allow us to control the $L^2(\Omega)$ and piecewise $H^1(\Omega)$ norms of y_h by $|||y_h|||_h$. To this end, we employ the notation

$$|||w_h|||^2_{H_h^1(\Omega)} = \|\nabla_h w_h\|^2_{L^2(\Omega)} + \|h^{-1/2}\,[w_h]\,\|^2_{L^2(\Sigma)}, \tag{4.2.13}$$

which corresponds to an appropriate H^1 semi-norm that combines the piece-wise gradient with scaled discontinuities along the edges. We apply this norm component-wise to the gradients $\nabla_h y_h$ of functions $y_h \in \left[\mathbb{V}_h^k\right]^3$. In case of triangles, a natural extension of our notation is $\nabla_h y_h \in \left[\mathbb{V}_h^{k-1}\right]^{3\times 2}$, where

$$\mathbb{V}_h^{k-1} := \left\{ v_h \in L^2(\Omega) \ : \ v_h|_T \in \mathbb{P}_{k-1}(T) \ \forall\, T \in \mathcal{T} \right\}$$

For quadrilaterals $\nabla_h y_h \in \left[\mathbb{V}_h^{k-1}\right]^{3\times 2}$, since differentiations does not necessarily change the order of the polynomial. Before we proceed, we recall the Clement interpolant constructed in [18]. We introduce it on a scalar context, but apply it component-wise whenever necessary. The interpolant allows us to construct a continuous version of both y_h and $\nabla_h y_h$, that is critical to prove the Γ-convergence result in Section 4.4. In order to construct it, we consider for each element $T \in \mathcal{T}$ the star $\omega(T) \subset \mathcal{T}$, which satisfies

$$\operatorname{diam}(\omega(T)) \leq C\ h_T$$

with $C > 0$ depending only on the shape regularity of the triangulation and $\sigma(T) = \Sigma \cap \omega(T)$, the part of the skeleton of $\mathcal{T}$ interior to $\omega(T)$.

Definition 4.2.1. *(Smoothing interpolant) Let z be a node of $\mathcal{T}$ and denote by $\omega_z = supp\ \psi_z$ the set of elements in $\mathcal{T}$ that lie in the support of the nodal basis for z (star of the node). This basis is with respect to a space $\mathbb{V}_h^l(\omega(z)) \cap H_0^1(\Omega)$ for $l \geq 1$, where the notation $\mathbb{V}_h^l(\omega(z))$ stands for functions supported in $\omega(z)$. The Clement interpolant $\Pi_h : \mathbb{V}_h^l \to \mathbb{V}_h^l \cap H_0^1(\Omega)$ is then defined by*

$$\Pi_h w_h := \sum_{z \in \mathcal{N}_h} \Pi_h^z w_h(z)\ \psi_z,$$

where for each z, $\Pi_h^z w_h \in \mathbb{V}_h^l(\omega(z)) \cap H_0^1(\Omega)$ is the solution of:

$$\int_{\omega(z)} (w_h - \Pi_h^z w_h) W = 0 \quad \forall\ W \in \mathbb{V}_h^l(\omega(z)) \cap H_0^1(\Omega).$$

We recall the following approximation property for the interpolant, that was proven in [18].

Lemma 4.2.2. *(estimates for smoothing interpolant) Let Π_h be the interpolant of Definition 4.2.1. Then for $|\alpha| = 0, 1$ and all $T \in \mathcal{T}$*

$$\|D^\alpha(w_h - \Pi_h w_h)\|_{L^2(T)} \leq C \left\| h^{\frac{1-2|\alpha|}{2}} [w_h] \right\|_{L^2(\sigma(T))} \quad \forall\ w_h \in \mathbb{V}_h^l.$$

We can now prove error bounds for the L^2 norm of $y_h \in \left[\mathbb{V}_h^k\right]^3$.

Lemma 4.2.3 (Poincaré type inequality for $|||\cdot|||_h$)**.** *Let $y_h \in \left[\mathbb{V}_h^k\right]^3$ and let $g_h, \Phi_{D,h}$ be the functions employed in the boundary jumps (could be zero). Let also $|||\cdot|||_h$ and $|||\cdot|||_{H_h^1(\Omega)}$ be as in (4.2.11) and (4.2.13) The following estimates are valid:*

(i)

$$|||\nabla_h y_h - \Phi_{D,h}|||_{H_h^1(\Omega)} \leq |||\nabla_h y_h|||_{H_h^1(\Omega)} + |||\Phi_{D,h}|||_{H_h^1(\Omega)} \leq |||y_h|||_h + |||\Phi_{D,h}|||_{H_h^1(\Omega)}$$

(ii) (Poincaré type Inequality)

$$\|\nabla_h y_h - \Phi_{D,h}\|_{L^2(\Omega)} \leq C\ |||\nabla_h y_h - \Phi_{D,h}|||_{H_h^1(\Omega)}$$

or

$$\|\nabla_h y_h\|_{L^2(\Omega)} \leq C\ \left(|||y_h|||_h + |||\Phi_{D,h}|||_{H_h^1(\Omega)} \right).$$

(iii)

$$\|y_h\|_{L^2(\Omega)} + \|\nabla_h y_h\|_{L^2(\Omega)} \leq C\ \left(|||y_h|||_h + |||g_h|||_{H_h^1(\Omega)} + |||\Phi_{D,h}|||_{H_h^1(\Omega)} \right)$$

Proof. (i) Immediate from triangle inequality and the definition of $|||\cdot|||_{H_h^1(\Omega)}$.

(ii) We employ the interpolant of Definition 4.2.1. Let $w_h = \nabla_h y_h - \Phi_{D,h} \in \left[\mathbb{V}_h^{k-1}\right]^{3\times 2}$ and $\Pi_h w_h \in \left[\mathbb{V}_h^{k-1}\right]^{3\times 2} \cap [H_0^1(\Omega)]^{3\times 2}$. Then, Lemma 4.2.2 implies that

$$\|\Pi_h w_h - w_h\|_{L^2(\Omega)} \leq C\ \|h^{1/2}[w_h]\|_{L^2(\Sigma)}$$

and

$$\begin{aligned}\|\nabla(\Pi_h w_h)\|^2_{L^2(\Omega)} &= \sum_{T\in\mathcal{T}} \|\nabla(\Pi_h w_h)\|^2_{L^2(T)} \\ &\leq 2\sum_{T\in\mathcal{T}} \|\nabla(\Pi_h w_h - w_h)\|^2_{L^2(T)} + \|\nabla w_h\|^2_{L^2(T)} \\ &\leq C\left(\sum_{T\in\mathcal{T}} \|\nabla w_h\|^2_{L^2(T)} + \sum_{e\in\mathcal{E}} \|h^{-1/2}[w_h]\|^2_{L^2(e)}\right) \\ &= C\ |||w_h|||^2_{H^1_h(\Omega)}\end{aligned}$$

We combine the above estimates with the Poincaré Inequality for $\Pi_h w_h \in [H^1_0(\Omega)]^{3\times 2}$ to obtain

$$\begin{aligned}\|w_h\|_{L^2(\Omega)} &\leq \|\Pi_h w_h - w_h\|_{L^2(\Omega)} + \|\Pi_h w_h\|_{L^2(\Omega)} \\ &\leq C\left(\ \|h^{1/2}[w_h]\|_{L^2(\Sigma)} + \|\nabla\Pi_h w_h\|_{L^2(\Omega)}\right) \\ &\leq C|||w_h|||_{H^1_h(\Omega)}\end{aligned}$$

The desired inequality now follows from part (i).

(iii) We apply the same technique as in (ii) for $z_h = y_h - g_h$ to obtain

$$\|z_h\|_{L^2(\Omega)} \leq C|||z_h|||_{H^1_h(\Omega)},$$

where we also use that for $h < 1$

$$\|h^{-1/2}[y_h]\|_{L^2(\Sigma)} \leq \|h^{-3/2}[y_h]\|_{L^2(\Sigma)}$$

Combining with (i) and (ii) we obtain the desired inequality.

□

We can now use this lemma to prove the coercivity of the discrete energy with respect to the discrete energy norm $|||\cdot|||_h$.

Lemma 4.2.4 (Coercivity of E_h)**.** *Let $g_h, \Phi_{D,h}$ piecewise polynomials with uniformly bounded H_h^1-norms and let $y_h \in \left[\mathbb{V}_h^k\right]^3$. The following implication is true:*

$$E_h[y_h] \leq C \implies \quad \textit{there exists} \quad C \quad \textit{such that} \quad |||y_h|||_h \leq C.$$

Proof. Recall that

$$\begin{aligned} E_h[y_h] := \frac{1}{2}\|D_h^2 y_h\|_{L^2(\Omega)}^2 - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(\Sigma)} \\ + S_0[y_h] + S_1[y_h] - \int_\Omega f_h y_h, \end{aligned}$$

We first employ Young's inequality and the trace inequality for discrete spaces to get

$$\begin{aligned} -(\{\partial_\nu \nabla_h y_h\}, [\nabla_h y_h])_{L^2(\Sigma)} &\geq -\frac{1}{2\rho}\|h^{1/2}\{\partial_\nu \nabla_h y_h\}\|_{L^2(\Sigma)}^2 - \frac{\rho}{2}\|h^{-1/2}\left[\nabla_h y_h\right]\|_{L^2(\Sigma)}^2 \\ &\geq -\frac{C}{\rho}\|D_h^2 y_h\|_{L^2(\Omega)}^2 - \frac{\rho}{2}\|h^{-1/2}\left[\nabla_h y_h\right]\|_{L^2(\Sigma)}^2. \end{aligned}$$

This means that for appropriate choice of ρ and for γ_1 big enough, the above term can be absorbed into $\|D_h^2 y_h\|_{L^2(\Omega)}$ and S_1. We now consider the third derivative term and use the inverse inequality for $\mathbb{V}_h^k$, along with the discrete trace inequality. More precisely, we have that for $e \in \mathcal{E}$

$$\begin{aligned} \|\{\partial_\nu \Delta_h y_h\}\|_{L^2(e)} &\leq \|D^3 y_h\|_{L^2(e)} \leq C h^{-1/2}\|D^3 y_h\|_{L^2(\omega(e))} \\ &\leq C h^{-1/2}\ h^{-1}\|D^2 y_h\|_{L^2(\omega(e))} \leq C h^{-3/2}\|D^2 y_h\|_{L^2(\omega(e))}, \end{aligned}$$

where $\omega(e)$ corresponds to the union of the two elements containing e. As a result,

$$\begin{aligned} (\{\partial_\nu \Delta_h y_h\}, [y_h])_{L^2(e)} &= (h^{3/2}\{\partial_\nu \Delta_h y_h\}, h^{-3/2}\left[y_h\right])_{L^2(e)} \\ &\leq \|h^{3/2}\{\partial_\nu \Delta_h y_h\}\|_{L^2(e)}\|h^{-3/2}\left[y_h\right]\|_{L^2(e)} \\ &\leq C\|D_h^2 y_h\|_{L^2(\omega(e))}\|h^{-3/2}\left[y_h\right]\|_{L^2(e)} \end{aligned}$$

Using Young's Inequality again, we can choose γ_0 big enough to absorb the above term in $\|D_h^2 y_h\|_{L^2(\Omega)}$ and S_0. To conclude, we use Lemma 4.2.3 (iii) to obtain a bound for the forcing term

$$\int_\Omega f_h y_h \leq \|f\|_{L^2(\Omega)}\|y_h\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}\left(|||y_h|||_h + \|g_h\|_{H_h^1(\Omega)} + \|\Phi_{D,h}\|_{H_h^1(\Omega)}\right).$$

We combine the preceding statements to deduce the assertion. □

4.3 Discrete Hessian

In this section we provide a suitable definition of a discrete Hessian $H_h[y_h]$ for $y_h \in \left[\mathbb{V}_h^k\right]^3$, as motivated in Section 4.2. We first introduce two lifting operators that allow us to deal with the inherent discontinuities that appear in our discrete energy. We prove a very important property for $H_h[y_h]$ that is critical in Section 4.4 for the limit-infimum argument, i.e.

$$\frac{1}{2}\|H_h[y_h]\|_{L^2(\Omega)}^2 - (f_h, y_h)_{L^2(\Omega)} \leq E_h[y_h], \tag{4.3.1}$$

where we emphasize that the inequality is proven with constant 1. Lastly, we show that for any $y_h \in \left[\mathbb{V}_h^k\right]^3$ that converges in $[L^2(\Omega)]^3$ to a $y \in [H^2(\Omega)]^3$, we have the weak convergence of $H_h[y_h]$, i.e.

$$H_h[y_h] \rightharpoonup D^2 y \quad \text{in} \quad [L^2(\Omega)]^3. \tag{4.3.2}$$

4.3.1 Definition and L^2 bound of $H_h[y_h]$

Motivated by Section 4.2, we define two lifting operators that allow us to deal with the discontinuities of high derivatives of $y_h \in \left[\mathbb{V}_h^k\right]^3$. We then use these oper-

ators to define $H_h[y_h]$ and derive a crude estimate for $\|H_h[y_h]\|_{L^2(\Omega)}$. For simplicity we define the operators assuming that y_h is scalar. Then we can apply everything component-wise for $y_h \in \left[\mathbb{V}_h^k\right]^3$.

To motivate the definition of the lifting operators, we look at two of the terms in (4.2.9). We first consider the term

$$-(\{\partial_\nu(\nabla_h y_h)\}, [\nabla_h y_h])_{L^2(\Sigma)},$$

For all $e \in \mathcal{E}$, we let $r_e^{l_1} : [L^2(e)]^d \to [\mathbb{V}_h^{l_1}]^{d\times d},\ l_1 \geq 0$ be the lifting operator defined as follows: for all $\phi \in [L^2(e)]^d$:

$$\int_\Omega r_e^{l_1}(\phi) : \tau_h = \int_e \{\tau_h\}\, \nu_e \cdot \phi \quad \forall \tau_h \in [\mathbb{V}_h^{l_1}]^{d\times d}. \tag{4.3.3}$$

The second lifting operator corresponds to the term

$$(\{\partial_\nu(\Delta_h y_h)\}, [y_h])_{L^2(\Sigma)}.$$

For all $e \in \mathcal{E}$, we denote $b_e^{l_2} : L^2(e) \to [\mathbb{V}_h^{l_2}]^{d\times d},\ l_2 \geq 0$ to be the lifting operator defined as follows: for all $\phi \in L^2(e)$:

$$\int_\Omega b_e^{l_2}(\phi) : \tau_h = \int_e \{\text{div } \tau_h\} \cdot \nu_e\ \phi \quad \forall \tau_h \in [\mathbb{V}_h^{l_2}]^{d\times d}. \tag{4.3.4}$$

Our goal is to use that

$$\text{div}(\nabla_h(\nabla_h y_h)) = \nabla_h(\Delta_h y_h),$$

in order to apply the lifting operators for $\tau_h = \nabla_h(\nabla_h y_h)$. We fix $l_1 = k$ and $l_2 = k+1$ and omit l_1, l_2 from our notation for the rest of the paper.

Remark 4.3.1. *(choice of l_1, l_2) We choose l_1, l_2 in a way that guarantees that (4.3.3) and (4.3.4) are well defined and the results in Lemma 4.3.3 (L^2-bounds of lifting operators) hold true, regardless of the choice of triangular or quadrilateral elements for the mesh. For instance, in the proof of part (ii) of Lemma 4.3.3 we use τ_h of the same polynomial degree as $\nabla_h y_h$, which means that for quadrilateral elements we must use $l_1 = k$. For l_2 the same choice of $l_2 = k$ would suffice for the definition of (4.3.4) to make sense. However, the construction we employ in the proof of part (iv) requires $\tau_h \in \mathbb{V}_h^{k+1}$. We wish to investigate whether we can relax this requirement using a different construction, but we fix $l_2 = k+1$ for completeness. We also note that parts (ii) and (iv) do not contribute to our $\Gamma-$convergence results, but they are rather important in order to show that $S_0, S_1 \to 0$ after we establish the $\Gamma-$convergence of E_h to E.*

We can now define the discrete Hessian.

Definition 4.3.2 (discrete Hessian)**.** *Let $y_h \in \left[\mathbb{V}_h^k\right]^3$ and r_e, b_e as in (4.3.3)-(4.3.4). We define the discrete Hessian operator $H_h : [\mathbb{V}_h^k]^3 \to [\mathbb{V}_h^{k+1}]^{3\times d\times d}$ as*

$$\begin{aligned} H_h[y_h] &:= \nabla_h(\nabla_h y_h) - R_h([\nabla_h y_h]) + B_h([y_h]) \\ &:= D_h^2 y_h - \sum_{e\in\mathcal{E}} r_e([\nabla_h y_h]) + \sum_{e\in\mathcal{E}} b_e([y_h]), \end{aligned} \tag{4.3.5}$$

where $\nabla_h\nabla_h \equiv D_h^2$ and R_h and B_h are applied component-wise to y_h.

We now prove some estimates about the lifting operators that are instrumental in order to prove (4.3.1) and (4.3.2), which in turn are critical for the Γ-convergence of E_h in Section 4.4. We prove the result for one component, but it immediately

extends to $\left[\mathbb{V}_h^k\right]^3$. The lemma is motivated by a similar approach for second order operators in [35].

Lemma 4.3.3 (L^2-bounds of lifting operators)**.** *Let r_e, b_e be as in (4.3.3)-(4.3.4). The following estimates are true for all $e \in \mathcal{E}$:*

(i) For all $y_h \in \mathbb{V}_h^k$

$$\|r_e([\nabla_h y_h])\|_{L^2(\Omega)} \leq Ch^{-1/2} \|[\nabla_h y_h]\|_{L^2(e)}$$

(ii) and conversely,

$$h^{-1/2} \|[\nabla_h y_h]\|_{L^2(e)} \leq C\|r_e([\nabla_h y_h])\|_{L^2(\Omega)}$$

(iii) Similarly, for all $y_h \in \mathbb{V}_h^k$

$$\|b_e([y_h])\|_{L^2(\Omega)} \leq Ch^{-3/2} \|[y_h]\|_{L^2(e)}$$

(iv) and conversely,

$$h^{-3/2} \|[y_h]\|_{L^2(e)} \leq C\|b_e([y_h])\|_{L^2(\Omega)}$$

Proof. (i) We set $\phi_h = [\nabla_h y_h]$ and $\tau_h = r_e([\nabla_h y_h])$ in (4.3.3) and use the discrete trace inequality to obtain

$$\begin{aligned}
\|r_e([\nabla_h y_h])\|^2_{L^2(\Omega)} &= \int_\Omega r_e([\nabla_h y_h]) : r_e([\nabla_h y_h]) = \int_e \{r_e([\nabla_h y_h])\}\ \nu_e \cdot [\nabla_h y_h] \\
&\leq \|\{r_e([\nabla_h y_h])\ \}\|_{L^2(e)} \|[\nabla_h y_h]\|_{L^2(e)} \\
&\leq Ch^{-1/2}\|r_e([\nabla_h y_h])\|_{L^2(\Omega)}\|[\nabla_h y_h]\|_{L^2(e)}
\end{aligned}$$

This implies that

$$\|r_e([\nabla_h y_h])\|_{L^2(\Omega)} \leq Ch^{-1/2} \|[\nabla_h y_h]\|_{L^2(e)}$$

(ii) For the converse, let $e \in \mathcal{E}$ and define for any function $\phi_h \in \mathbb{P}_k(e)$ its extension $P_e(\phi_h) \in \mathbb{P}_k(\omega(e))$ that is constant along lines orthogonal to e. Then, by the equivalence of discrete norms and a scaling argument, we can see that

$$\|P_e(\phi_h)\|_{L^2(\omega(e))} \leq Ch^{1/2}\|\phi_h\|_{L^2(e)} \tag{4.3.6}$$

We now choose $\tau_h = P_e([\nabla_h y_h]\nu_e^T)$, which is zero outside the elements of $\omega(e)$, and $\phi = [\nabla_h y_h]$ in (4.3.3). We use $[\nabla_h y_h]\nu_e^T$ in order to obtain a discrete function in $\mathbb{R}^{d\times d}$. This is a rank-1 matrix, where each column is a multiple of $[\nabla_h y_h]$. Using this, we have:

$$\begin{aligned}
\int_\Omega r_e([\nabla_h y_h]) : P_e([\nabla_h y_h](\nu_e)^T) &= \int_e \left\{P_e([\nabla_h y_h]\nu_e^T)\right\} \nu_e \cdot [\nabla_h y_h] \\
&= \int_e [\nabla_h y_h]\nu_e^T \nu_e \cdot [\nabla_h y_h] \\
&= \int_e |\nu_e|^2 \ [\nabla_h y_h] \cdot [\nabla_h y_h] \\
&= \int_e |[\nabla_h y_h]|^2,
\end{aligned}$$

whence,

$$\begin{aligned}
\int_e |[\nabla_h y_h]|^2 &= \int_\Omega r_e([\nabla_h y_h]) : P_e([\nabla_h y_h]\nu_e^T) \\
&\leq \|r_e([\nabla_h y_h])\|_{L^2(\Omega)} \|P_e([\nabla_h y_h]\nu_e)\|_{L^2(\omega(e))},
\end{aligned}$$

which, combined with (4.3.6), gives

$$\int_e |[\nabla_h y_h]|^2 \leq C \ h^{1/2} \ \|r_e([\nabla_h y_h])\|_{L^2(\Omega)} \ \|[\nabla_h y_h]\|_{L^2(e)},$$

which in turns leads to the desired estimate.

(iii) We now use similar techniques to prove the bound for b_e. This time we set

$\tau_h = b_e([y_h])$ in (4.3.4) to obtain

$$\begin{aligned}\|b_e([y_h])\|^2_{L^2(\Omega)} &= \int_\Omega b_e([y_h]) : b_e([y_h]) = \int_e \{\mathrm{div}(b_e([y_h]))\} \cdot \nu_e \ [y_h] \\ &\leq \|\{\mathrm{div}(b_e([y_h]))\}\|_{L^2(e)} \, \|[y_h]\|_{L^2(e)} \\ &\leq Ch^{-1/2} \|\mathrm{div}(b_e([y_h]))\|_{L^2(\omega(e))} \|[y_h]\|_{L^2(e)} \\ &\leq Ch^{-3/2} \|b_e([y_h])\|_{L^2(\omega(e))} \|[y_h]\|_{L^2(e)},\end{aligned}$$

whence,

$$\|b_e([y_h])\|_{L^2(\Omega)} \leq Ch^{-3/2} \|[y_h]\|_{L^2(e)}.$$

(iv) For the converse, we need for each edge e to construct τ_h such that $\{\mathrm{div}\tau_h\} \cdot \nu_e = [y_h]$ on e. To achieve this, we create $\tau_h \in \mathbb{V}_h^{k+1}$ so that $\mathrm{div}\tau_h \cdot \nu_e = P_e([y_h])$ on each element that shares the edge e (if it is an interior edge), where P_e is the extension defined in (ii). We need P_e in order to construct τ_h by controlling its divergence on each element. Then, the problem amounts to constructing a matrix with a given divergence. For the unit vector ν_e we pick $m \in [1, d]$ such that the m-th component ν_e^m of ν_e is non-zero. For any fixed d, we can always find an element ν_e^m of the unit vector ν_e such that $|\nu_e^m| \geq \frac{1}{\sqrt{d}}$ and we choose the biggest such component, to make sure that it is not close to zero. Then, in order to solve $\mathrm{div}\tau_h \cdot \nu_e = P_e([y_h])$ on each element, we set $(\tau_h)_{i,j} = 0$ for all $i, j \neq m$ and let $\sigma_m := (\tau_h)_{m,m} \neq 0$ be so that

$$\mathrm{div}\tau_h \cdot \nu_e = \frac{\partial \sigma_m}{\partial x_m} \nu_e^m = P_e([y_h]),$$

which is simple to solve for polynomial functions, since it implies that

$$\sigma_m = \int_{x_m|_T} \frac{P_e([y_h])}{\nu_e^m} dx_m \in \mathbb{V}_h^{k+1}. \tag{4.3.7}$$

We call the result of the above construction $\tau_h([y_h])$ and we plug it in (4.3.4) to get:

$$\int_\Omega b_e([y_h]) : \tau_h([y_h]) = \int_e \{\mathrm{div}\tau_h([y_h])\} \cdot \nu_e \ [y_h] = \int_e \{P_e([y_h])\} \ [y_h] = \int_e |[y_h]|^2.$$

We now have that

$$\int_e |[y_h]|^2 = \int_\Omega b_e([y_h]) : \tau_h([y_h]) \leq \|b_e([y_h])\|_{L^2(\Omega)} \|\tau_h([y_h])\|_{L^2(\omega(e))},$$

where, owing to (4.3.7),

$$\|\tau_h(y_h)\|_{L^2(\omega(e))} \leq C \ h \ \|P_e([y_h])\|_{L^2(\omega(e))}.$$

Combining with (4.3.6), we obtain:

$$\begin{aligned}
\int_e |[y_h]|^2 &= \int_\Omega b_e([y_h]) : \tau_h([y_h]) \\
&\leq \|b_e([y_h])\|_{L^2(\Omega)} \|\tau_h([y_h])\|_{L^2(\omega(e))} \\
&\leq C \ h \ \|b_e([y_h])\|_{L^2(\Omega)} \ \|P_e([y_h])\|_{L^2(\omega(e))} \\
&\leq h^{3/2} \|b_e([y_h])\|_{L^2(\Omega)} \ \|[y_h]\|_{L^2(e)}.
\end{aligned}$$

This completes the proof.

□

We have the following immediate corollary

Corollary 4.3.4 (L^2 bound of discrete Hessian)**.** *For any $y_h \in \left[\mathbb{V}_h^k\right]^3$, the following bound holds:*

$$\|H_h[y_h]\|_{L^2(\Omega)} \leq C|||y_h|||_h.$$

Proof. This is an immediate consequence of Definition 4.3.2 (discrete Hessian), Lemma 4.3.3 (L^2-bounds of lifting operators) and the observation that

$$\|R_h([\nabla_h y_h])\|^2_{L^2(\Omega)} \leq 4 \sum_{e \in \mathcal{E}} \|r_e([\nabla_h y_h])\|^2_{L^2(\Omega)}$$

and

$$\|B_h([y_h])\|^2_{L^2(\Omega)} \leq 4 \sum_{e \in \mathcal{E}} \|b_e([y_h])\|^2_{L^2(\Omega)}.$$

The factor 4 corresponds to a mesh of quadrilaterals (four sides), whereas a triangular mesh would only require a factor of 3. □

4.3.2 Upper bound for $\|H_h[y_h]\|_{L^2(\Omega)}$

We now prove an immediate consequence of Definition 4.3.2 (discrete Hessian) and Lemma 4.3.3 (L^2-bounds of lifting operators). We separate this result from the previous, in order to emphasize its importance for the rest of our analysis.

Proposition 4.3.5 (L^2-bound of H_h with respect to E_h)**.** *Let $y_h \in \left[\mathbb{V}_h^k\right]^3$ and $H_h[y_h]$ be the discrete Hessian of y_h defined in (4.3.5). Then for $E_h[y_h]$ defined in (4.2.9) we have*

$$\frac{1}{2}\|H_h[y_h]\|^2_{L^2(\Omega)} - (f_h, y_h)_{L^2(\Omega)} \leq E_h[y_h].$$

Proof. We prove the result by expanding $\|H_h[y_h]\|^2_{L^2(\Omega)}$

$$\begin{aligned}
\frac{1}{2}\int_\Omega |H_h(y_h)|^2 &= \frac{1}{2}\int_\Omega |D_h^2 y_h - R_h([\nabla_h y_h]) + B_h([y_h])|^2 \\
&= \frac{1}{2}\int_\Omega |D_h^2 y_h|^2 - \int_\Omega D_h^2 y_h : R_h([\nabla_h y_h]) + \int_\Omega D_h^2 y_h : B_h([y_h]) \\
&+ \frac{1}{2}\int_\Omega |B_h([y_h]) - R_h([\nabla_h y_h])|^2 \\
\vdots \\
&= \frac{1}{2}\int_\Omega |D_h^2 y_h|^2 - \sum_{e\in\mathcal{E}}\int_e \{\partial_\nu \nabla_h y_h\} : [\nabla_h y_h] + \sum_{e\in\mathcal{E}}\int_e \{\partial_\nu \Delta_h y_h\}\cdot [y_h] \\
&+ \frac{1}{2}\int_\Omega |B_h([y_h]) - R_h([\nabla_h y_h])|^2 \\
&= E_h[y_h] + \int_\Omega f_h y_h + \mathcal{I},
\end{aligned}$$

where

$$\mathcal{I} = \frac{1}{2}\int_\Omega |B_h([y_h]) - R_h([\nabla_h y_h])|^2 - \frac{\gamma_1}{2}\sum_{e\in\mathcal{E}}\frac{1}{h}\int_e |[\nabla_h y_h]|^2 - \frac{\gamma_0}{2}\sum_{e\in\mathcal{E}}\frac{1}{h^3}\int_e |[y_h]|^2.$$

This means that we only have to show that

$$-\sum_{e\in\mathcal{E}}\frac{\gamma_1}{h}\int_e |[\nabla_h y_h]|^2 - \sum_{e\in\mathcal{E}}\frac{\gamma_0}{h^3}\int_e |[y_h]|^2 + \int_\Omega |B_h([y_h]) - R_h([\nabla_h y_h])|^2 \le 0.$$

This is an immediate consequence of (i) and (iii) of Lemma 4.3.3 (L^2-bounds for lifting operators) and the proof of Corollary 4.3.4 (L^2 bound of discrete Hessian) for γ_0, γ_1 big enough. This concludes the proof. □

4.3.3 Weak convergence of H_h

We now prove the weak convergence of H_h that is critical later in order to prove a lim-inf inequality for E_h. We prove it under a convergence assumption for y_h, but we see in Section 4.4 that we can incorporate it into a much stronger result. Before we proceed, we introduce one last notation. We denote by $P_h^k\phi$ a discrete

approximation of a function ϕ into the space of polynomials of degree k. We can choose a conforming interpolant or a discontinuous approximation as in [71]. Then we have that for all $\phi \in C_0^\infty(\Omega)$, owing to the regularity of ϕ,

$$|||P_h^k \phi - \phi|||_h \to 0.$$

Lemma 4.3.6 (weak convergence of H_h)**.** *Let* $\{y_h\}_h \subset \left[\mathbb{V}_h^k\right]^3$ *satisfy* $|||y_h|||_h \leq C$ *for all* h*. We also assume that* y_h *converges to a function* $y \in [H^2(\Omega)]^3$ *in* $[L^2(\Omega)]^3$*. Then,*

$$H_h[y_h] \rightharpoonup D^2 y \quad in \quad [L^2(\Omega)]^{3\times 2\times 2}.$$

Proof. Let $\phi \in C_0^\infty(\mathbb{R}^2)^{3\times 2\times 2}$. We denote $P_h\phi \equiv P_h^k\phi$ and use definition 4.3.2 (discrete Hessian) and integration by parts to create terms that we can control.

$$\begin{aligned}
\int_\Omega H_h[y_h] : \phi &= \int_\Omega \nabla_h \nabla_h y_h : \phi - \int_\Omega R_h([\nabla_h y_h]) : \phi + \int_\Omega B_h([y_h]) : \phi \\
&= -\int_\Omega \nabla_h y_h : \mathrm{div}\phi + \sum_{e\in\mathcal{E}} \int_e \phi\nu_e : [\nabla_h y_h] \\
&\quad - \int_\Omega R_h([\nabla_h y_h]) : (\phi - P_h\phi) - \int_\Omega R_h([\nabla_h y_h]) : P_h\phi \\
&\quad + \int_\Omega B_h([y_h]) : \phi \\
&= -\int_\Omega \nabla_h y_h : \mathrm{div}\phi - \int_\Omega R_h([\nabla_h y_h]) : (\phi - P_h\phi) \\
&\quad + \sum_{e\in\mathcal{E}} \left(-\int_e \{P_h\phi\}\, \nu_e : [\nabla_h y_h] + \int_e \{\phi\}\, \nu_e : [\nabla_h y_h] \right) \\
&\quad + \int_\Omega B_h([y_h]) : \phi.
\end{aligned}$$

We perform one more integration by parts and use definition (4.3.4) in order to treat

the B_h operator:

$$\begin{aligned}\int_\Omega H_h[y_h] : \phi &= \int_\Omega y_h \cdot \mathrm{div}(\mathrm{div}\phi) - \sum_{e\in\mathcal{E}} \int_e \mathrm{div}\phi \ \nu_e \cdot [y_h] \\ &- \quad \int_\Omega R_h([\nabla_h y_h]) : (\phi - P_h\phi) + \sum_{e\in\mathcal{E}} \int_e \{\phi - P_h\phi\} \, \nu_e : [\nabla_h y_h] \\ &+ \int_\Omega B_h([y_h]) : (\phi - P_h\phi) + \int_\Omega B_h([y_h]) : P_h\phi \\ &= \int_\Omega y_h \cdot \mathrm{div}(\mathrm{div}\phi) - \int_\Omega R_h([\nabla_h y_h]) : (\phi - P_h\phi) \\ &+ \sum_{e\in\mathcal{E}} \int_e \{\phi - P_h\phi\} \, \nu_e : [\nabla_h y_h] \\ &+ \int_\Omega B_h([y_h]) : (\phi - P_h\phi) + \sum_{e\in\mathcal{E}} \int_e \{\mathrm{div}(P_h\phi - \phi)\} \ \nu_e \cdot [y_h] \\ &= T_1 + T_2 + T_3 + T_4 + T_5.\end{aligned}$$

We now look at each of those terms separately. The first one converges to

$$\int_\Omega y \cdot \mathrm{div}(\mathrm{div}\phi) = -\int_\Omega \nabla y : \mathrm{div}\phi = \int_\Omega D^2 y : \phi$$

due to the L^2 convergence of y_h, the zero boundary conditions of ϕ and the fact that $y \in [H^2(\Omega)]^3$. For the second term we use the uniform bound of $|||y_h|||_h$ and Corollary 4.3.4 (L^2-bound of discrete Hessian) to get

$$|T_2| \leq \|R_h([\nabla_h y_h])\|_{L^2(\Omega)} \|P_h\phi - \phi\|_{L^2(\Omega)} \leq C \ \|P_h\phi - \phi\|_{L^2(\Omega)} \to 0.$$

We now show that T_3 converges to zero too. For $e \in \mathcal{E}$, we utilize the scaled trace inequality to write

$$\|P_h\phi - \phi\|_{L^2(e)} \leq C \ \left(h^{-1/2}\|P_h\phi - \phi\|_{L^2(\omega(e))} + h^{1/2}\|\nabla_h(P_h\phi - \phi)\|_{L^2(\omega(e))}\right),$$

which in view of

$$\sum_{e\in\mathcal{E}} h^{-1} \ \|[\nabla_h y_h]\|^2_{L^2(e)} \leq C$$

implies that

$$
\begin{aligned}
|T_3| &\leq \sum_{e\in\mathcal{E}} \|P_h\phi-\phi\|_{L^2(e)}\|[\nabla_h y_h]\|_{L^2(e)} \\
&= \sum_{e\in\mathcal{E}} h^{1/2}\ \|P_h\phi-\phi\|_{L^2(e)}\ h^{-1/2}\ \|[\nabla_h y_h]\|_{L^2(e)} \\
&\leq \sum_{e\in\mathcal{E}} C\ \left(\|P_h\phi-\phi\|_{L^2(\omega(e))}+h\|\nabla_h(P_h\phi-\phi)\|_{L^2(\omega(e))}\right)\ h^{-1/2}\ \|[\nabla_h y_h]\|_{L^2(e)} \\
&\leq C\left(\sum_{T\in\mathcal{T}} \|P_h\phi-\phi\|^2_{L^2(T)}+h^2\|\nabla_h(P_h\phi-\phi)\|^2_{L^2(\omega(e))}\right)^{1/2} \\
&\qquad \left(\sum_{e\in\mathcal{E}} h^{-1}\ \|[\nabla_h y_h]\|^2_{L^2(e)}\right)^{1/2} \to 0.
\end{aligned}
$$

Similarly, we see that due to the uniform bound of $|||y_h|||_h$:

$$
|T_4| \leq \|B_h([y_h])\|_{L^2(\Omega)}\|P_h\phi-\phi\|_{L^2(\Omega)} \leq C\ \|P_h\phi-\phi\|_{L^2(\Omega)} \to 0.
$$

Lastly,

$$
\begin{aligned}
|T_5| &\leq \sum_{e\in\mathcal{E}} \|\left\{\mathrm{div}(P_h\phi-\phi)\right\}\|_{L^2(e)}\|[y_h]\|_{L^2(e)} \\
&= \sum_{e\in\mathcal{E}} h^{3/2}\ \|\left\{\mathrm{div}(P_h\phi-\phi)\right\}\|_{L^2(e)}\ h^{-3/2}\ \|[y_h]\|_{L^2(e)},
\end{aligned}
$$

where we use again the scaled trace inequality

$$
\|\left\{\mathrm{div}(P_h\phi-\phi)\right\}\|_{L^2(e)} \leq C\left(h^{-1/2}\|\nabla_h(P_h\phi-\phi)\|_{L^2(\omega(e))}+h^{1/2}\|D^2_h(P_h\phi-\phi)\|_{L^2(\omega(e))}\right)
$$

in conjunction with

$$
\sum_{e\in\mathcal{E}} h^{-3}\|[y_h]\|^2_{L^2(e)} \leq C
$$

to obtain

$$
\begin{aligned}
|T_5| &\leq C\left(\sum_{T\in\mathcal{T}} h^2\|\nabla_h(P_h\phi-\phi)\|^2_{L^2(T)}+h^4\|D^2_h(P_h\phi-\phi)\|^2_{L^2(T)}\right)^{1/2} \\
&\qquad \left(\sum_{e\in\mathcal{E}} h^{-3}\|[y_h]\|^2_{L^2(e)}\right)^{1/2} \to 0.
\end{aligned}
$$

This concludes the proof of the weak convergence of $H_h[y_h]$ to D^2y. □

4.4 Γ-Convergence of E_h

In this section we prove the Γ-convergence of E_h to E. To this end, we prove a lim-inf inequality for E_h using a compactness argument for $y_h \in \mathbb{A}_h^k$ with uniformly bounded energies, combined with results from Section 4.3. We also prove a lim-sup inequality using a simple regularization argument and interpolation, in order to construct a recovery sequence for any $y \in \mathbb{A}$. We combine these results to prove the Γ-convergence of E_h and deduce the convergence of global minimizers of E_h in $\mathbb{A}_h^k$ to a global minimizer of E in $\mathbb{A}$. Lastly, we combine these results to prove that the stabilization terms $S_0[y_h]$ and $S_1[y_h]$ for such minimizers converge to zero and that the convergence of $H_h[y_h]$ in Lemma 4.3.6 is strong in $[L^2(\Omega)]^3$.

4.4.1 Equi-coercivity and compactness in L^2

In Section 4.2 we prove in Lemma 4.2.4 (coercivity of E_h) that for $y_h \in \left[\mathbb{V}_h^k\right]^3$ with $E_h[y_h] \leq C$ we have that $|||y_h|||_h \leq C$. We also show in Lemma 4.2.3 (estimates for smoothing interpolant) that we can control the L^2 and piecewise H^1 norm of such a y_h too. We now use these results along with the Clement interpolant from Definition 4.2.1 and Lemma 4.3.6 (weak convergence of H_h) to prove a compactness result for a sequence of y_h with uniformly bounded energies.

Proposition 4.4.1 (compactness in $L^2(\Omega)$)**.** *Let* $\{y_h\} \subset \left[\mathbb{V}_h^k\right]^3$ *be a sequence such that for all* $h > 0$

$$E_h[y_h] \leq C.$$

Then we can find a subsequence (without relabeling) and a function $y \in H^2(\Omega)^3$ such that

$$y_h \to y \quad in \quad L^2(\Omega)^3$$

and $y = g$, $\nabla y = \Phi_D$ on the boundary. The function y also satisfies the isometry constraint

$$\nabla y^T \nabla y = I \quad a.e.$$

Lastly, the discrete Hessian $H_h[y_h]$ converges to $D^2 y$ weakly in $[L^2(\Omega)]^3$, i.e.

$$H_h[y_h] \xrightarrow{L^2(\Omega)} D^2 y$$

Proof. We prove the result in several steps.

Step 1. (weak convergence in L^2) We first use the equi-coercivity of $\{y_h\}$, i.e. $E_h[y_h] \leq C$ and Lemma 4.2.4 (coercivity of E_h) to conclude that $|||y_h|||_h \leq C$ for all h. We now employ Lemma 4.2.3 (Poincaré type inequality for $|||\cdot|||_h$) to obtain $\|y_h\|_{L^2(\Omega)} \leq C'$ for all h. This immediately implies that a subsequence y_h (which we do not relabel) converges weakly to some $y \in [L^2(\Omega)]^3$,

$$y_h \xrightarrow{L^2(\Omega)} y$$

Step 2. (strong convergence and H^1 regularity of y) We now recall the smoothing interpolant from Definition 4.2.1 and use it to prove that the convergence is strong and that $y \in [H^1(\Omega)]^3$. To that end, we set

$$z_h = \Pi_h(y_h - g_h) \ \in \left[\mathbb{V}_h^k\right]^3 \cap [H_0^1(\Omega)]^3,$$

where the interpolant is taken component-wise, and we show that z_h converges weakly in $[H^1(\Omega)]^3$ to a $z \in [H^1(\Omega)]^3$. Note that this immediately implies that it

converges strongly to z in $[L^2(\Omega)]^3$. Since $z_h \in [H_0^1(\Omega)]^3$, we can write

$$|z_h|^2_{H^1(\Omega)} = \sum_{T\in\mathcal{T}} |z_h|^2_{H^1(T)},$$

which allows us to consider each element separately:

$$\begin{aligned}
|z_h|^2_{H^1(T)} &\leq 2\left(|\Pi_h(y_h - g_h) - (y_h - g_h)|^2_{H^1(T)} + |y_h - g_h|^2_{H^1(T)}\right) \\
&\preceq \|h^{-1/2}\,[y_h]\,\|^2_{L^2(\sigma(T))} + |y_h - g_h|^2_{H^1(T)} + \|h^{-1/2}\,[g_h]\,\|^2_{L^2(\sigma(T))}
\end{aligned}$$

where for the second inequality we split the $[y_h - g_h]$ term and $\sigma(T)$ stands again for the interior skeleton of the star of T. Summing over all elements, we obtain

$$|z_h|^2_{H^1(\Omega)} \preceq \sum_{T\in\mathcal{T}} |\nabla_h y_h|^2_{L^2(T)} + \sum_{e\in\mathcal{E}} \|h^{-1/2}\,[y_h]\,\|^2_{L^2(e)} + \|\nabla_h g_h\|^2_{L^2(\Omega)} + \sum_{e\in\mathcal{E}} \|h^{-1/2}\,[g_h]\,\|^2_{L^2(e)}.$$

We conclude that $|z_h|_{H^1(\Omega)}$ is uniformly bounded, since for $h < 1$ the first two terms are bounded by $|||y_h|||_h$ and the other two terms are also uniformly bounded due to the regularity of g. As a consequence z_h converges weakly to some $z \in [H_0^1(\Omega)]^3$ and strongly to z in $[L^2(\Omega)]^3$. The uniform bound $|||y_h|||_h \leq C$ implies that

$$\sum_{e\in\Sigma} \|[y_h]\|^2_{L^2(e)} \leq Ch^3.$$

As a result, using Lemma 4.2.2 (estimates for smoothing interpolant) for z_h, we obtain

$$\begin{aligned}
\|(y_h - g_h) - z\|^2_{L^2(\Omega)} &= \sum_{T\in\mathcal{T}} \|(y_h - g_h) - z\|^2_{L^2(T)} \\
&\leq 2\sum_{T\in\mathcal{T}} \left(\|(y_h - g_h) - z_h\|^2_{L^2(T)} + \|z_h - z\|^2_{L^2(T)}\right) \\
&\leq 2\sum_{T\in\mathcal{T}} \left(Ch\ \|[y_h]\|^2_{L^2(\sigma(T))} + \|z_h - z\|^2_{L^2(T)} + Ch\ \|[g_h]\|^2_{L^2(\sigma(T))}\right) \\
&\leq C'h^4 + \|z_h - z\|^2_{L^2(\Omega)} + Ch\ \|[g_h]\|^2_{L^2(\Sigma)} \to 0.
\end{aligned}$$

Since $g_h \to g$ in $[L^2(\Omega)]^3$, this implies that y_h converges in $L^2(\Omega)$ to $z+g$. Since the weak $L^2(\Omega)$ limit of y_h is y and it's unique, we conclude that

$$y_h \to y \quad \text{in} \quad [L^2(\Omega)]^3$$

and that $y \in [H^1(\Omega)]^3$ with $y = g$ on $\partial\Omega$.

Step 3. ($L^2(\Omega)$ convergence of $\nabla_h y_h$) We can now repeat the above process for

$$Z_h = \Pi_h(\nabla_h y_h - \Phi_{D,h}) \ \in \left[\mathbb{V}_h^k\right]^3 \cap [H_0^1(\Omega)]^{3\times 2},$$

and show that Z_h converges weakly in $[H_0^1(\Omega)]^{3\times 2}$ and strongly in $[L^2(\Omega)]^{3\times 2}$ to a $Z \in [H_0^1(\Omega)]^{3\times 2}$. Considering again each element separately, we have

$$|Z_h|^2_{H^1(T)} \preceq \|h^{-1/2}\left[\nabla_h y_h\right]\|^2_{L^2(\sigma(T))} + |\nabla_h y_h - \Phi_{D,h}|^2_{H^1(T)} + \|h^{-1/2}\left[\Phi_{D,h}\right]\|^2_{L^2(\sigma(T))}$$

Summing over all elements, we obtain

$$|Z_h|^2_{H^1(\Omega)} \preceq \sum_{T\in\mathcal{T}} |D_h^2 y_h|^2_{L^2(T)} + \sum_{e\in\mathcal{E}} \|h^{-1/2}\left[\nabla_h y_h\right]\|^2_{L^2(e))} + \|\Phi_{D,h}\|^2_{H_h^1(\Omega)},$$

hence $|Z_h|_{H^1(\Omega)}$ is uniformly bounded and thus converges strongly to Z in $[L^2(\Omega)]^{3\times 2}$. Using the uniform bound $|||y_h|||_h \leq C$ yileds

$$\sum_{e\in\Sigma} \|[\nabla_h y_h]\|^2_{L^2(e)} \leq Ch$$

and proceeding similarly to Step 2 and invoking Lemma 4.2.2 (estimates for smoothing interpolant) for Z_h we obtain

$$\begin{aligned}\|(\nabla_h y_h - \Phi_{D,h}) - Z\|^2_{L^2(\Omega)} &= \sum_{T\in\mathcal{T}} \|(\nabla_h y_h - \Phi_{D,h}) - Z\|^2_{L^2(T)} \\ &\leq C'h^2 + \|Z_h - Z\|^2_{L^2(\Omega)} + Ch\ \|[\Phi_{D,h}]\|^2_{L^2(\Sigma)} \to 0.\end{aligned}$$

Since $\Phi_{D,h} \to \Phi_D$ in $[L^2(\Omega)]^{3\times 2}$, this implies that $\nabla_h y_h$ converges in $L^2(\Omega)$ to $Z+\Phi_D$.

Step 4. ($H^2(\Omega)$ regularity of y and remaining boundary conditions) In Step 2 we show that $y_h \to y$ in $[L^2(\Omega)]^3$ and in Step 3 we show that $\nabla_h y_h \to Z + \Phi_D$ in $[L^2(\Omega)]^{3\times 2}$ and that $Z + \Phi_D \in [H^1(\Omega)]^{3\times 2}$. It remains to show that $\nabla y = Z + \Phi_D$, which establishes that $y \in [H^2(\Omega)]^3$ and that $\nabla y = \Phi_D$ on $\partial\Omega$. To this end, we use that the $L^2(\Omega)$ convergence of $\nabla_h y_h$ implies convergence in the sense of distribution, i.e.

$$\nabla_h y_h \rightharpoonup Z + \Phi_D \quad \text{in} \quad \mathcal{D}'.$$

We can now use integration by parts to deduce that the distributional derivative Dy_h of y_h is given by

$$-\int_\Omega y_h \cdot \mathrm{div}\phi = \int_\Omega \nabla_h y_h \ \phi - \int_\Sigma [y_h]\nu^T \ \phi$$

for all $\phi \in [C_0^\infty(\Omega)]^{3\times 2}$. We now employ the fact that

$$|||y_h|||_h \le C \Longrightarrow \|h^{-1/2}[y_h]\|_{L^2(\Sigma)} \le Ch \to 0.$$

to conclude that

$$\left|\int_\Sigma [y_h]\nu^T \ \phi\right| \le C \ \|h^{-1/2}[y_h]\|_{L^2(\Sigma)} \ \|h^{1/2}\phi\|_{L^2(\Sigma)} \le C \ |\Omega|^{1/2} \ \|h^{-1/2}[y_h]\|_{L^2(\Sigma)} \ \|\phi\|_{L^\infty(\Omega)} \to 0$$

for all $\phi \in C_0^\infty(\Omega)^{3\times 2}$, where we used that

$$\left(\sum_{e\in\mathcal{E}} \|h^{1/2}\phi\|^2_{L^2(e)}\right)^{1/2} \le \|\phi\|_{L^\infty(\Omega)} \left(\sum_{e\in\mathcal{E}} h^2\right)^{1/2} \le C|\Omega|^{1/2}\|\phi\|_{L^\infty(\Omega)}.$$

This means that in fact the distributional limits of Dy_h and $\nabla_h y_h$ coincide, since

$$\langle Dy, \phi\rangle = -\int_\Omega y\cdot\mathrm{div}\phi \xleftarrow{\mathcal{D}'} -\int_\Omega y_h\cdot\mathrm{div}\phi = \int_\Omega \nabla_h y_h \ \phi - \int_\Sigma [y_h]\nu^T \ \phi \xrightarrow{\mathcal{D}'} \int_\Omega (Z+\Phi_D) \ \phi + 0,$$

where we used the uniqueness of the limit for $\nabla_h y_h$. We can now conclude that $y \in [H^2(\Omega)]^3$ and $\nabla y = \Phi_D$ on $\partial\Omega$. We also see that $\nabla_h y_h \to \nabla y$ in $[L^2(\Omega)]^{3\times 2}$.

Step 5. (isometry constraint) We next show that y satisfies the isometry constraint a.e. in Ω upon proving that

$$\|\nabla y^T \nabla y - I\|_{L^1(\Omega)} = 0.$$

To this end, we need to show that the uniform bound on the piecewise Hessian of y_h along with the discrete isometry constraint implies that

$$\sum_{T\in\mathcal{T}} \|\nabla_h y_h^T \nabla_h y_h - I\|_{L^1(T)} \leq Ch.$$

Then the result follows since

$$\begin{aligned}
\|\nabla y^T \nabla y - I\|_{L^1(\Omega)} &= \sum_{T\in\mathcal{T}} \|\nabla y^T \nabla y - I\|_{L^1(T)} \\
&= \sum_{T\in\mathcal{T}} \|(\nabla y^T \nabla y - \nabla_h y_h^T \nabla_h y_h) + (\nabla_h y_h^T \nabla_h y_h - I)\|_{L^1(T)} \\
&\leq \sum_{T\in\mathcal{T}} \left(\|\nabla y^T (\nabla y - \nabla_h y_h)\|_{L^1(T)} + \|(\nabla_h y_h^T - \nabla y^T)\nabla_h y_h\|_{L^1(T)}\right) \\
&\quad + \sum_{T\in\mathcal{T}} \|\nabla_h y_h^T \nabla_h y_h - I\|_{L^1(T)},
\end{aligned}$$

which implies that

$$\begin{aligned}
\|\nabla y^T \nabla y - I\|_{L^1(\Omega)} &\leq \left(\sum_{T\in\mathcal{T}} \|\nabla y\|_{L^2(T)}^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|\nabla_h (y - y_h)\|_{L^2(T)}^2\right)^{1/2} \\
&\quad + \left(\sum_{T\in\mathcal{T}} \|\nabla y_h\|_{L^2(T)}^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|\nabla_h (y - y_h)\|_{L^2(T)}^2\right)^{1/2} \\
&\quad + \sum_{T\in\mathcal{T}} \|\nabla_h y_h^T \nabla_h y_h - I\|_{L^1(T)} \\
&\leq \left(\|\nabla y\|_{L^2(\Omega)} + \|\nabla_h y_h\|_{L^2(\Omega)}\right) \|\nabla_h y_h - \nabla y\|_{L^2(\Omega)} + Ch \to 0,
\end{aligned}$$

where we use the strong convergence of $\nabla_h y_h$ to ∇y in $[L^2(\Omega)]^{3\times 2}$ and the uniform bound of $\nabla_h y_h$ and ∇y in $[L^2(\Omega)]^{3\times 2}$. We now prove our claim for the $L^1(\Omega)$ bound of the isometry defect. To this end, we use (4.4.1), i.e.

$$\sum_{T\in\mathcal{T}} \left| \int_T \nabla_h y_h^T \nabla_h y_h - I \right| \leq ch. \tag{4.4.1}$$

This implies that on each element the integral of the isometry constraint is controlled by h. We combine this with the fact that we have control of the broken Hessian $D_h^2 y_h$ and thus of possible oscillations of the isometry constraint in each element. To see this, let

$$J_T := \int_T \nabla_h y_h^T \nabla_h y_h - I \quad \text{and} \quad A_T := \frac{J_T}{|T|}$$

and observe that

$$\|\nabla_h y_h^T \nabla_h y_h - I\|_{L^1(T)} \leq \|(\nabla_h y_h^T \nabla_h y_h - I) - A_T\|_{L^1(T)} + \|A_T\|_{L^1(T)}.$$

For the second term, we see that

$$\|A_T\|_{L^1(T)} = \int_T \frac{|J_T|}{|T|} = |J_T|.$$

For the first term we can now apply Poincaré inequality in $L^1(T)$ to obtain

$$\begin{aligned} \|(\nabla_h y_h^T \nabla_h y_h - I) - A_T\|_{L^1(T)} &\leq C\ h\ \|\nabla_h(\nabla_h y_h^T \nabla_h y_h)\|_{L^1(T)} \\ &\leq C\ h\ \|D_h^2 y_h\|_{L^2(T)} \|\nabla_h y_h\|_{L^2(T)}. \end{aligned}$$

We sum over all elements and use the uniform bound on $E_h[y_h]$, Lemma 4.2.3

(Poincaré type inequality for $|||\cdot|||_h$) and (4.4.1) to conclude that

$$\begin{aligned}\sum_{T\in\mathcal{T}} \|\nabla_h y_h^T \nabla_h y_h - I\|_{L^1(T)} &\leq \sum_{T\in\mathcal{T}} \left(\|(\nabla_h y_h^T \nabla_h y_h - I) - A_T\|_{L^1(T)} + \|A_T\|_{L^1(T)}\right) \\ &\leq C\ h\ \sum_{T\in\mathcal{T}} \|D_h^2 y_h\|_{L^2(T)} \|\nabla_h y_h\|_{L^2(T)} + \sum_{T\in\mathcal{T}} |J_T| \\ &\leq C\ h\ \left(\sum_{T\in\mathcal{T}} \|D_h^2 y_h\|_{L^2(T)}^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|\nabla_h y_h\|_{L^2(T)}^2\right)^{1/2} \\ &+ \sum_{T\in\mathcal{T}} |J_T| \leq C\ h.\end{aligned}$$

Step 6. (weak convergence of $H_h[y_h]$) This step follows immediately, since $y_h \to y$ in $[L^2(\Omega)]^3$ and $y \in [H^2(\Omega)]^3$, whence Lemma 4.3.6 (weak convergence of H_h) implies that

$$H_h[y_h] \rightharpoonup D^2 y \quad \text{in} \quad [L^2(\Omega)]^{3\times 2}.$$

This concludes the proof. □

4.4.2 Lim-inf property of E_h

The goal of Section 4.4 is to prove the Γ-convergence of E_h to E and more precisely, to show that every cluster point y of a sequence $\{y_h\}$ of global minimizers of E_h is a global minimizer of E in $\mathbb{A}$. To this end, we need to prove a lim-inf and a lim-sup property. The lim-inf property follows from the preceding results, but we prove it here separately to highlight it. Instead of looking at a general $y_h \to y$ in $[L^2(\Omega)]^3$, we focus on a sequence of $\{y_h\}_h$ with $E_h[y_h] \leq C$ for all h, since otherwise the lim-inf inequality is trivial. Moreover, the uniform bound on y_h implies the existence of a cluster point y due to Proposition 4.4.1 (compactness in $L^2(\Omega)$).

Lemma 4.4.2 (lim-inf property)**.** *Let* $\{y_h\}$ *satisfy* $y_h \in \mathbb{A}_h^k$ *and* $E_h[y_h] \leq C$ *for all* h*. Then, there exists* $y \in \mathbb{A}$ *such that* $y_h \to y$ *in* $[L^2(\Omega)]^3$ *and*

$$E[y] \leq \liminf_{h\to 0} E_h[y_h].$$

Proof. Since $y_h \in \mathbb{A}_h^k$ and $E_h[y_h] \leq C$, we invoke Proposition 4.4.1 (compactness in $L^2(\Omega)$) to show that there exists a $y \in [H^2(\Omega)]^3$ that satisfies the boundary conditions and the isometry constraint as well as

$$y_h \to y \quad \text{in} \quad [L^2(\Omega)]^3$$

and

$$H_h[y_h] \xrightarrow{L^2(\Omega)} D^2 y.$$

We use the lower-semicontinuity of the L^2-norm and the fact that $(f_h, y_h)_{L^2(\Omega)} \to (f, y)_{L^2(\Omega)}$ to write

$$\frac{1}{2}\int_\Omega |D^2 y|^2 - \int_\Omega fy \leq \frac{1}{2}\liminf_{h\to 0}\int_\Omega |H_h(y_h)|^2 - \int_\Omega f_h y_h.$$

We then employ Proposition 4.3.5 (L^2-bound of H_h with respect to E_h), i.e.

$$\frac{1}{2}\int_\Omega |H_h(y_h)|^2 - \int_\Omega f_h y_h \leq E_h[y_h],$$

and combine the two inequalities to deduce that

$$E[y] \leq \liminf_{h\to 0} E_h[y_h].$$

Since $y \in \mathbb{A}$, this concludes the proof. □

4.4.3 Lim-sup property of E_h

We now prove the second inequality that is necessary for the Γ-convergence of E_h. Since we are interested in minimizers of E, we prove the existence of a recovery sequence in $[L^2(\Omega)]^3$ for a function $y \in \mathbb{A}$. Therefore, we use a simple regularization argument followed by Lagrange interpolation. However, we emphasize that due to the way we force the discrete isometry constraint in (4.4.1) and the Nitsche approach for the boundary conditions, we do not need to preserve the isometry constraint or the boundary conditions of y during our regularization argument.

Lemma 4.4.3 (lim-sup property)**.** *For any $y \in \mathbb{A}$, there exists a recovery sequence $\{y_h\}_h$ with $y_h \in \mathbb{A}_h^k$ for each h such that*

$$y_h \to y \quad in \quad [L^2(\Omega)]^3$$

and

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

Proof. We prove the result in several steps.

Step 1. (recovery sequence) We start by constructing a recovery sequence $\{y_h\}_h$ for y. Without loss of generality, we assume that there exists $\Lambda > 0$ such that $\|y\|_{H^2(\Omega)} \leq \Lambda$. We use the standard Lagrange interpolant of $\mathbb{V}_h^k$, i.e. $y_h = \mathcal{I}_h y$. We first show that $y_h \in \mathbb{A}_h^k$ for each h, whence $E_h[y_h]$ is finite. This means that we need to show that y_h satisfies the linearized isometry constraint (4.4.1). To this end, we use that

$$\|\nabla y_h\|_{L^2(T)} \leq \|\nabla y\|_{L^2(T)} + \|\nabla(y_h - y)\|_{L^2(T)} \leq \|\nabla y\|_{L^2(T)} + C_I h \|D^2 y\|_{L^2(T)},$$

where C_I is a local interpolation constant for $\mathcal{I}_h$.

$$\begin{aligned}
\sum_{T\in\mathcal{T}} \|(\nabla_h y_h)^T \nabla_h y_h - I\|_{L^1(T)} &\leq \sum_{T\in\mathcal{T}} \big(\|(\nabla y)^T \nabla y - I\|_{L^1(T)} + \|\nabla y_h\|_{L^2(T)} \|\nabla(y - y_h)\|_{L^2(T)} \\
&\quad + \|\nabla y\|_{L^2(T)} \|\nabla(y - y_h)\|_{L^2(T)} \big) \\
&\leq 2 \sum_{T\in\mathcal{T}} \|\nabla y\|_{L^2(T)} \; C_I h \; \|D^2 y\|_{L^2(T)} + C_I h^2 \sum_{T\in\mathcal{T}} \|D^2 y\|^2_{L^2(T)} \\
&\leq 2C_I \; h \left(\sum_{T\in\mathcal{T}} \|\nabla y\|^2_{L^2(T)} \right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|D^2 y\|^2_{L^2(T)} \right)^{1/2} \\
&\quad + C_I h^2 \sum_{T\in\mathcal{T}} \|D^2 y\|^2_{L^2(T)} \\
&\leq 2C_I \Lambda^2 \; (h + h^2).
\end{aligned}$$

We assume that Λ is such that $c \geq 2C_I\Lambda^2$. Ignoring higher order terms, this choice implies that

$$\sum_{T\in\mathcal{T}} \left| \int_T (\nabla_h y_h)^T \nabla_h y_h - I \right| \leq \sum_{T\in\mathcal{T}} \|(\nabla_h y_h)^T \nabla_h y_h - I\|_{L^1(T)} \leq ch.$$

Step 2. (convergence of $E_h[y_h]$) In Step 1 we verified that $y_h \in \mathbb{A}^k_h$, whence $E_h[y_h]$ is finite. It now remains to show that

$$E_h[y_h] - \left(\frac{1}{2} \int_\Omega |D^2 y|^2 - \int_\Omega fy \right) \to 0.$$

To this end, we prove the convergence of $\|D^2_h y_h\|_{L^2(\Omega)}$ and $S_1[y_h]$, since the proof for $S_0[y_h]$ is similar to that of $S_1[y_h]$ and the convergence of the remaining terms can be deduced using the estimates in Lemma 4.2.4 (coercivity of E_h). For all subsequent proofs we assume that $y \in [H^3(\Omega)]^3$ and deal with the $[H^2(\Omega)]^3$ case in Step 3. We start with the piecewise Hessian $D^2_h y_h$. We first present briefly the H^2-stability of the Lagrange interpolant, using the Bramble-Hilbert lemma along with inverse

inequalities. To this end, let l be a linear function and consider

$$\|D^2 y_h\|_{L^2(T)} = \|D^2 \mathcal{I}_h(y-l)\|_{L^2(T)} \leq Ch^{-2}\|\mathcal{I}_h(y-l)\|_{L^2(T)} \leq Ch^{-1}\|\mathcal{I}_h(y-l)\|_{L^\infty(T)}$$
$$\leq Ch^{-1}\|y-l\|_{L^\infty(T)} \leq C\|D^2 y\|_{L^2(T)}.$$

We can now see that

$$\begin{aligned}\sum_{T\in\mathcal{T}} |D^2 y_h|^2 - |D^2 y|^2 &= \sum_{T\in\mathcal{T}} (D^2 y_h - D^2 y)(D^2 y_h + D^2 y) \\ &\leq C_I C \sum_{T\in\mathcal{T}} h \, \|D^3 y\|_{L^2(T)} \, \|D^2 y\|_{L^2(T)} \\ &\leq C_I C \, h \left(\sum_{T\in\mathcal{T}} \|D^3 y\|^2_{L^2(T)}\right)^{1/2} \left(\sum_{T\in\mathcal{T}} \|D^2 y\|^2_{L^2(T)}\right)^{1/2} \\ &= C_I C \, h \, \|D^3 y\|_{L^2(\Omega)} \, \|D^2 y\|_{L^2(\Omega)} \to 0.\end{aligned}$$

We now examine the jump term $S_1[y_h]$. We use the scaled trace inequality

$$\begin{aligned}\|\nabla_h(y_h - y)\|^2_{L^2(e)} &\leq Ch^{-1}\|\nabla(y_h - y)\|^2_{L^2(\omega(e))} + h\|D^2(y_h - y)\|^2_{L^2(\omega(e))} \\ &\leq Ch^3 \|D^3 y\|^2_{L^2(\omega(e))}\end{aligned}$$

and the fact that $[\nabla y] = 0$ everywhere due to the $H^2(\Omega)$ regularity of y and (4.2.5). This means that

$$\begin{aligned}\sum_{e\in\mathcal{E}} \frac{\gamma_1}{h} \int_e [\nabla_h y_h]^2 &= \sum_{e\in\mathcal{E}} \frac{\gamma_1}{h} \int_e [\nabla_h (y_h - y)]^2 \\ &\leq C \sum_{T\in\mathcal{T}} \frac{\gamma_1}{h} \, h^3 \, \|D^3 y\|^2_{L^2(T)} \\ &\leq C \, h^2 \|D^3 y\|^2_{L^2(\Omega)} \to 0.\end{aligned}$$

Step 3. (regularization of y) In the previous step we assume that $y \in [H^3(\Omega)]^3$ in order to use appropriate interpolation estimates. We now show how we deal with the H^2 regularity of y. We first employ a smooth approximation y^ε of y. Let $\varepsilon \to 0$ be a mollifying parameter, ϕ_ε be a standard mollifier and $y^\varepsilon = y * \phi_\varepsilon$ be a regularization

of y. Since $y \in [H^2(\Omega)]^3$, we have that $y^\varepsilon \to y$ in $[H^2(\Omega)]^3$. We let y_h^ε be the Lagrange interpolant of y^ε and showcase our claim by adjusting the calculations for the convergence of $S_1[y_h]$ from the previous step. We have that

$$\sum_{e\in\mathcal{E}} \frac{\gamma_1}{h}\int_e [\nabla_h y_h]^2 = \sum_{e\in\mathcal{E}} \frac{\gamma_1}{h}\int_e [\nabla_h (y_h - y)]^2$$
$$\leq \sum_{T\in\mathcal{T}} C\frac{\gamma_1}{h}\, h\|D_h^2(y_h - y)\|_{L^2(T)}^2 \leq C\sum_{T\in\mathcal{T}} \|D_h^2(y_h - y)\|_{L^2(T)}^2.$$

We now use y^ε and y_h^ε to make the above estimate as small as possible:

$$\|D^2(y - y_h)\|_{L^2(T)}^2 \preceq \|D_h^2(y - y^\varepsilon)\|_{L^2(T)}^2 + \|D_h^2(y^\varepsilon - y_h^\varepsilon)\|_{L^2(T)}^2 + \|D_h^2(y_h^\varepsilon - y_h)\|_{L^2(T)}^2$$
$$\preceq \|D_h^2(y - y^\varepsilon)\|_{L^2(T)}^2 + h\,\|y^\varepsilon\|_{H^3(T)}^2 + C\|D_h^2(y - y^\varepsilon)\|_{L^2(T)}^2$$
$$\leq (1 + C)\|y - y^\varepsilon\|_{H^2(T)}^2 + h\,\|y^\varepsilon\|_{H^3(T)}^2,$$

whence

$$\sum_{T\in\mathcal{T}} \|D_h^2(y_h - y)\|_{L^2(T)}^2 \preceq (1 + C)\|y - y^\varepsilon\|_{H^2(\Omega)}^2 + h\,\|y^\varepsilon\|_{H^3(\Omega)}^2.$$

This means that for every $\delta > 0$ we can pick ε small enough so that $\|y - y^\varepsilon\|_{H^2(\Omega)} < \delta$. This leads to a certain value for $\|y^\varepsilon\|_{H^3(\Omega)}$. We can then choose h small enough so that the second term is also less than δ, hence making the sum as small as we want. The same argument applies for all calculations of Step 2. This concludes the proof. □

4.4.4 Γ-Convergence of E_h

We now combine the preceding results to prove the Γ-convergence of E_h to E. We show that cluster points of global minimizers of E_h are global minimizers of E, without assuming that such minimizers exist. We then use this to prove a much stronger version of Proposition 4.4.1 (compactness in $L^2(\Omega)$).

Theorem 4.4.4 (convergence of global minimizers)**.** *Let* $\{y_h\}_h$ *be a sequence of almost global minimizers of* E_h*, namely*

$$E_h[y_h] \leq \inf_{w_h \in \mathbb{A}_h} E_h[w_h] + \epsilon_h \leq C,$$

where $\epsilon_h \to 0$ *as* $h \to 0$ *and* C *is a constant independent of* h*. Then,* $\{y_h\}_h$ *is precompact in* $[L^2(\Omega)]^3$ *and every cluster point* y *of* y_h *is an absolute minimizer of* E*, namely*

$$E[y] = \inf_{w \in \mathbb{A}} E[w].$$

Moreover, up to a subsequence,

$$\lim_{h \to 0} E_h[y_h] = E[y].$$

Proof. We first invoke Lemma 4.4.2 (lim-inf property) to deduce that there exists a $y \in \mathbb{A}$ such that (a subsequence) $y_h \to y$ in $[L^2(\Omega)]^3$ and

$$E[y] \leq \liminf_{h \to 0} E_h[y_h].$$

To show that y is a global minimizer of E, we let $\eta > 0$ and consider $w \in \mathbb{A}$ such that

$$E[w] \leq \inf_{z \in \mathbb{A}} E[z] + \eta/2.$$

Lemma 4.4.3 (lim-sup property) implies that there exists a recovery sequence of $w_h \in \mathbb{A}_h^k$ such that $w_h \to w$ in $[L^2(\Omega)]^3$ and

$$E_h[w_h] \leq E[w] + \eta/2.$$

We now use these two properties and the fact that y_h is a global minimizer of E_h

and $w_h \in \mathbb{A}_h^k$ to deduce that

$$E[y] \leq \liminf_{h\to 0} E_h[y_h] \leq \liminf_{h\to 0}(E_h[w_h] + \epsilon_h) \leq \limsup_{h\to 0} E_h[w_h]$$

$$\leq E[w] + \eta/2 \leq \inf_{z\in\mathbb{A}} E[z] + \eta.$$

Since the above holds for all $\eta > 0$, it implies that y is an absolute minimizer of E. By choosing $w = y$ in the lim-inf and lim-sup property we also immediately infer that

$$\lim_{h\to 0} E_h[y_h] = E[y].$$

This concludes the proof. □

4.4.5 Strong Convergence of $H_h[y_h]$ and $S_i[y_h]$

We now use the previous result to show the strong convergence of the scaled jump terms to zero and of the discrete Hessian $H_h[y_h]$ of y_h as in Theorem 4.4.4 (convergence of absolute minimizers).

Corollary 4.4.5. *(strong convergence of $H_h[y_h]$ and $S_i[y_h]$) Let y_h be a sequence of almost minimizers as in Theorem 4.4.4 (convergence of absolute minimizers) and y be the limiting minimizer of E. Then, Proposition 4.4.1 (compactness in $L^2(\Omega)$) can be strengthened by the following statements:*

(i) $H_h[y_h] \to D^2 y$ *strongly in* $[L^2(\Omega)]^{3\times 2\times 2}$.

(ii) $S_0[y_h],\ S_1[y_h] \to 0$.

(iii) $D_h^2(y_h) \to D^2 y$ *strongly in* $[L^2(\Omega)]^{3\times 2\times 2}$.

Proof. (i) The previous theorem combined with Proposition 4.3.5 (L^2-bound of H_h with respect to E_h) implies that

$$\begin{aligned}\limsup_{h\to 0}\frac{1}{2}\|H_h[y_h]\|^2_{L^2(\Omega)} &\leq \limsup_{h\to 0}\left(E_h[y_h]+\int_\Omega f_h y_h\right)\\ &\leq E[y]+\int_\Omega fy=\frac{1}{2}\int_\Omega |D^2 y|^2,\end{aligned}$$

hence, combining with the weak convergence of $H_h[y_h]$ to $D^2 y$ and the lower semi-continuity of the L^2-norm, we obtain

$$\|H_h[y_h]\|_{L^2(\Omega)}\to\|D^2 y\|_{L^2(\Omega)}.$$

(ii) In Theorem 4.4.4 (convergence of absolute minimizers) we prove that

$$E_h[y_h]\to E[y]=\frac{1}{2}\int_\Omega |D^2 y|^2-\int_\Omega fy.$$

Invoking Step (i) we infer that

$$E_h[y_h]+\frac{1}{2}\|H_h[y_h]\|^2_{L^2(\Omega)}-\int_\Omega f_h y_h\to E[y]-\frac{1}{2}\int_\Omega |D^2 y|^2+\int_\Omega fy=0.$$

Owing to the proof in Proposition 4.3.5 (L^2-bound of H_h with respect to E_h), this implies that

$$\frac{1}{2}\|R_h([\nabla_h y_h])-B_h([y_h])\|^2_{L^2(\Omega)}-S_1[y_h]-S_0[y_h]\to 0.$$

In view of Lemma 4.3.3 (L^2-bounds of lifting operators) we immediately arrive at

$$S_0[y_h]\approx B_h([y_h])\to 0,\quad S_1[y_h]\approx R_h([\nabla_h y_h])\to 0.$$

(iii) This statement follows directly from definition (4.3.5) of $H_h[y_h]$ and Steps (i) and (ii).

□

4.5 Gradient Flow

We now define a gradient flow that allows us to construct discrete solutions y_h starting from an initial guess y_h^0. Given y_h^n, we employ the flow so that in each step $n+1$, we reduce the energy $E_h[y_h^{n+1}] < E_h[y_h^n]$. We do this by searching for a $y_h^{n+1} = y_h^n + \delta y_h^{n+1}$, such that

$$\delta y_h^{n+1} \in \mathcal{F}_h^n[y_h^n] := \left\{ \phi_h \in \left[\mathbb{V}_h^k\right]^3,\ \int_T \left((\nabla_h \phi_h)^T \nabla_h y_h^n + (\nabla_h y_h^n)^T \nabla_h \phi_h \right) = 0 \right\}$$

and that satisfies

$$(\delta y_h^{n+1}, \phi_h)_h + \Delta t\ a_h(y_h^n + \delta y_h^{n+1}, \phi_h) = \Delta t\ (f_h, \phi_h)_{L^2(\mathcal{T})}, \tag{4.5.1}$$

for every $\phi_h \in \mathcal{F}_h^n[y_h^n]$, where

$$\begin{aligned}(\delta y_h^{n+1}, \phi_h)_h := (D_h^2 \delta y_h^{n+1}, D_h^2 \phi_h)_{L^2(\Omega)} + h^{-1/2}([\nabla_h \delta y_h^{n+1}], [\nabla_h \phi_h])_{L^2(\Sigma)} \\ + h^{-3/2}([\delta y_h^{n+1}], [\phi_h])_{L^2(\Sigma)}.\end{aligned}$$

We also recall that

$$\begin{aligned}a_h(y_h, v_h) := (D_h^2 y_h, D_h^2 v_h)_{L^2(\Omega)} \\ - (\{\partial_\nu \nabla_h y_h\}, [\nabla_h v_h])_{L^2(\Sigma)} - (\{\partial_\nu \nabla_h v_h\}, [\nabla_h y_h])_{L^2(\Sigma)} \\ + (\{\partial_\nu \Delta_h y_h\}, [v_h])_{L^2(\Sigma)} + (\{\partial_\nu \Delta_h v_h\}, [y_h])_{L^2(\Sigma)} \\ + \frac{\gamma_1}{h}([\nabla_h y_h], [\nabla_h v_h])_{L^2(\Sigma)} + \frac{\gamma_0}{h^3}([y_h], [v_h])_{L^2(\Sigma)}.\end{aligned}$$

Remark 4.5.1 (linearized isometry constraints)**.** *The isometry constraint is linearized at each step and imposed weakly, namely in mean over each element T*

$$\int_T (\nabla_h \delta y_h^{n+1})^T \nabla_h y_h^n + (\nabla_h y_h^n)^T \nabla_h \delta y_h^{n+1} = 0. \tag{4.5.2}$$

Notice that $0 \in \mathcal{F}_h^n[y_h^n]$ whence, $\mathcal{F}_h^n[y_h^n] \neq \emptyset$ and the Lax-Milgram theorem implies the existence and uniqueness of a solution to each iteration of the gradient flow. We now show that the gradient flow results in a decrease of E_h.

Lemma 4.5.2 (energy decay)**.** *Let y_h^n be an iterate of the gradient flow and assume that the solution δy_h^{n+1} of (4.5.1) is non-zero. Then, for $y_h^{n+1} = y_h^n + \delta y_h^{n+1}$ we have that*

$$E_h[y_h^{n+1}] < E_h[y_h^n].$$

If y_h^0 is the initial guess for the flow and y_h^N the final iterate, then we also have that

$$\left(\frac{1}{2} + \frac{1}{\Delta t}\right) \sum_{n=0}^{N-1} |||\delta y_h^{n+1}|||_h^2 + E_h[y_h^N] = E_h[y_h^0] \tag{4.5.3}$$

Proof. We set $\phi = \delta y_h^{n+1}$ in (4.5.1) to obtain

$$|||\delta y_h^{n+1}|||_h^2 + \Delta t \; a_h(y_h^{n+1}, \delta y_h^{n+1}) = \Delta t (f_h, \delta y_h^{n+1}).$$

We split $a_h(y_h^{n+1}, \delta y_h^{n+1}) = \frac{1}{2} a_h(y_h^{n+1}, \delta y_h^{n+1}) + \frac{1}{2} a_h(y_h^{n+1}, \delta y_h^{n+1})$ and use simple algebraic calculations to re-write the above equality as

$$|||\delta y_h^{n+1}|||_h^2 + \frac{\Delta t}{2} a_h(y_h^{n+1}, y_h^{n+1}) - \frac{\Delta t}{2} a_h(y_h^n, y_h^n) + \frac{\Delta t}{2} a_h(\delta y_h^{n+1}, \delta y_h^{n+1}) = \Delta t (f_h, \delta y_h^{n+1}).$$

We recall that $E_h[y_h] = \frac{1}{2} a_h(y_h, y_h) - (f_h, y_h)_{L^2(\Omega)}$ to deduce that

$$\left(1 + \frac{\Delta t}{2}\right) |||\delta y_h^{n+1}|||_h^2 + \Delta t \; E_h[y_h^{n+1}] - \Delta t \; E_h[y_h^n] = 0,$$

whence

$$E_h[y_h^{n+1}] < E_h[y_h^n].$$

We sum over n and divide by Δt to obtain (4.5.3). □

We now show that the gradient flow guarantees the discrete isometry constraint (4.4.1) for appropriate choice of Δt .

Lemma 4.5.3 (discrete isometry constraint)**.** *Let y_h^0 be the initial guess for (4.5.1) that is also an isometry, i.e.*

$$(\nabla_h y_h^0)^T \nabla_h y_h^0 = I \quad a.e.\ in \quad \Omega$$

Then, the last iterate y_h^N satisfies the discrete isometry constraint in (4.4.1) if $\Delta t \ \leq h$ and

$$E_h[y_h^0] \leq c.$$

In this case, the discrete isometry defect decreases linearly with Δt

$$\sum_{T\in\mathcal{T}} \left| \int_T \left((\nabla_h y_h^N)^T \nabla_h y_h^N - I \right) \right| \preceq \Delta t \ E_h[y_h^0].$$

Proof. The result relies on the linearized isometry constraint (4.5.2) for δy_h^{n+1}

$$\int_T \left((\nabla_h \delta y_h^{n+1})^T \nabla_h y_h^n + (\nabla_h y_h^n)^T \nabla_h \delta y_h^{n+1} \right) = 0.$$

We start by showing that the integral isometry increases in each iteration of the flow. In fact, we have

$$\begin{aligned}
\int_T (\nabla_h y_h^{n+1})^T \nabla_h y_h^{n+1} &= \int_T \left(\nabla_h (y_h^{n+1} - y_h^n) \right)^T \nabla_h y_h^{n+1} + \int_T (\nabla_h y_h^n)^T \nabla_h y_h^{n+1} \\
&= \int_T \left(\nabla_h (y_h^{n+1} - y_h^n) \right)^T \left(\nabla_h (y_h^{n+1} - y_h^n) \right) \\
&+ \int_T \left(\nabla_h (y_h^{n+1} - y_h^n) \right)^T \nabla_h y_h^n + \int_T (\nabla_h y_h^n)^T \nabla_h y_h^{n+1} \\
&= \int_T \left(\nabla_h (y_h^{n+1} - y_h^n) \right)^T \left(\nabla_h (y_h^{n+1} - y_h^n) \right) \\
&- \int_T (\nabla_h y_h^n)^T \nabla_h (y_h^{n+1} - y_h^n) + \int_T (\nabla_h y_h^n)^T \nabla_h y_h^{n+1}.
\end{aligned}$$

The last equality is due to (4.5.2). The above calculations imply that

$$\int_T (\nabla_h y_h^{n+1})^T \nabla_h y_h^{n+1} = \int_T \left(\nabla_h (y_h^{n+1} - y_h^n)\right)^T \left(\nabla_h (y_h^{n+1} - y_h^n)\right) + \int_T (\nabla_h y_h^n)^T \nabla_h y_h^n.$$

Summing for $n = 0$ to $n = N - 1$ and using that y_h^0 is an isometry, we obtain

$$\begin{aligned}\left| \int_T \left((\nabla_h y_h^{N+1})^T \nabla_h y_h^{N+1} - I\right) \right| &= \left| \sum_{n=0}^{N} \int_T \left(\nabla_h (y_h^{n+1} - y_h^n)\right)^T \left(\nabla_h (y_h^{n+1} - y_h^n)\right) \right| \\ &\leq \sum_{n=0}^{N} \|\nabla_h \delta y^{n+1}\|^2_{L^2(T)}.\end{aligned}$$

We now combine this with Lemma 4.2.3 (Poincaré type inequality for $|||\cdot|||_h$) for $\delta y_h^{n+1} \in \left[\mathbb{V}_h^k\right]^3$ and the energy decay relation (4.5.3) to conclude that

$$\sum_{T \in \mathcal{T}} \left| \int_T \left((\nabla_h y_h^N)^T \nabla_h y_h^N - I\right) \right| \preceq \Delta t \; E_h[y_h^0].$$

This is the asserted estimate. □

Remark 4.5.4 (choice of c in discrete isometry constraint)**.** *Lemma 4.5.3 (discrete isometry constraint) implies that the choice of c in (4.4.1) relies on the initial guess y_h^0 of the flow. By choosing*

$$c \geq E_h[y_h^0], \tag{4.5.4}$$

we guarantee that all solutions y_h of the flow belong in $\mathbb{A}_h^k$. The choice of c appears also in Lemma 4.4.3 (lim-sup property), where we assume that for $\|y\|_{H^2(\Omega)} \leq \Lambda$ and C_I the interpolation constant for the Lagrange interpolant in $\mathbb{V}_h^k$,

$$c \geq 2C_I \Lambda^2. \tag{4.5.5}$$

This does not violate the generality of the results in Theorem 4.4.4 (convergence of absolute minimizers), since we can take $\Lambda = \|w\|_{H^2(\Omega)}$ for w as in the proof of Theorem 4.4.4 and choose c such that it satisfies both (4.5.4) and (4.5.5).

Remark 4.5.5 (monotonicity of the constraint)**.** *We observe that if we start with an initial guess such that for all* $T \in \mathcal{T}_h$

$$\int_T ((\nabla_h y_h^0)^T \nabla_h y_h^0 - I) \geq 0,$$

then the above calculations imply that for all n *and all* $T \in \mathcal{T}_h$

$$\int_T (\nabla_h y_h^{n+1})^T \nabla_h y_h^{n+1} \geq \int_T (\nabla_h y_h^n)^T \nabla_h y_h^n \geq ... \geq I|T|.$$

4.6 Solvability and comparison between DG, CDG and Kirchhoff approach

In this section we provide some information about our approach to obtain the discrete solution of (4.5.1) under the linearized isometry constraint (4.5.2) for each gradient flow step. We also give some details about the implementation of our method. Since this section is tied to our implementation, we refer to our space as $\mathbb{V}_h := [\mathbb{V}_h^2]^3$, as we use polynomials of degree 2 for all our computations.

4.6.1 Lagrange Multipliers

In order to find δy_h^{n+1} that satisfies the linearized isometry constraint (4.5.2), we employ Lagrange multipliers. More specifically, our implementation amounts to finding $(\delta y_h^{n+1}, \lambda_h^{n+1}) \in \mathbb{V}_h \times \Lambda_h$ such that

$$\begin{aligned} (1+\Delta t)\ a_h(\delta y_h^{n+1}, v_h) + \Delta t\ b_h(\lambda_h^{n+1}, v_h) &= \Delta t\ (f, v_h) - \Delta t\ a_h(y_h^n, v_h) \quad \forall v_h \in \mathbb{V}_h \\ b_h(\mu_h, \delta y_h^{n+1}) &= 0 \quad \forall \mu_h \in \Lambda_h, \end{aligned} \tag{4.6.1}$$

where

$$\Lambda_h := \left\{\lambda : \Omega \to \mathbb{R}^{2\times 2},\ \lambda^T = \lambda,\ \lambda|_T \in \mathbb{P}_0(T)\ \forall T \in \mathcal{T}_h\right\}$$

and

$$b_h(\lambda_h, v_h) := \sum_{T\in\mathcal{T}_h} \int_T \lambda_h : \left(\nabla v_h^T \nabla y_h^n + (\nabla y_h^n)^T \nabla v_h\right). \tag{4.6.2}$$

This is a saddle point problem. Since the multipliers μ_h are piecewise constant, for any $T \in \mathcal{T}_h$ we can choose μ_h such that $\mu_h|_T \neq 0$ and zero everywhere else, whence $b(\mu_h, \delta y_h^{n+1}) = 0$ implies

$$\mu_h : \int_T (\nabla \delta y_h^{n+1})^T \nabla y_h^n + (\nabla y_h^n)^T \nabla \delta y_h^{n+1} = 0,$$

which is (4.5.2) in disguise. In fact, by choosing $\mu_h|_T$ to be

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

we can conclude that

$$\int_T (\nabla \delta y_h^{n+1})^T \nabla y_h^n + (\nabla y_h^n)^T \nabla \delta y_h^{n+1} = 0$$

which is the desired isometry constraint (4.5.2). We later explore in Section 4.6.3 the solvability of this saddle problem.

4.6.2 Implementation

We now provide some specific details about the implementation of our method. As explained in Section 1.2, our goal for introducing this method is to provide a flexible and efficient method for which we can prove Γ-convergence results as in [8,11], while providing a simpler implementation, without the restrictions arising

from the Kirchhoff elements. The motivation for this is to extend the method so that we can reproduce the results obtained in [8,10,11] as well as to apply the method to cases motivated by experiments from the engineering community [2,53,75]. In order to set the foundation for this process and to profit from designing a more flexible and efficient method, we try to enhance the implementation speed and scalability of the method. We thus implement our method in parallel, using the built-in capabilities of the dealii library [3,6]. Moreover, we choose to exploit the structure of our problem, in order to use an iterative solver that relies on a Schur complement approach. More precisely:

- **Schur Complement:** Let A be the square matrix that corresponds to the discrete bilinear form a_h in (4.2.7). The coercivity of a_h in Lemma 4.2.4 (coercivity of E_h) implies that A is positive definite. We also let B_n be the matrix corresponding to the form b_h in (4.6.2), where the subscript n corresponds to the dependence of the matrix on y_h^n. Then if we call $\boldsymbol{\delta Y}_h^{n+1}$ and $\boldsymbol{\Lambda}_h^{n+1}$ the vectors that correspond to the solution $(\delta y_h^{n+1}, \lambda_h^{n+1})$ of (4.6.1), the resulting linear system is given by

$$\begin{bmatrix} A & B_n^T \\ B_n & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta Y}_h^{n+1} \\ \boldsymbol{\Lambda}_h^{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_n \\ \mathbf{0} \end{bmatrix}, \tag{4.6.3}$$

  where $\mathbf{f}_n$ corresponds to the right-hand side of (4.6.1) and also depends on y_h^n. The solution of this system can be given by first solving

$$(B_n A^{-1} B_n^T)\boldsymbol{\Lambda}_h^{n+1} = B_n A^{-1} \mathbf{f}_n \tag{4.6.4}$$

for $\mathbf{\Lambda}_h^{n+1}$ and then obtaining $\boldsymbol{\delta Y}_h^{n+1}$ through

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

- **Solver:** By examining the terms that comprise (4.2.7), we observe that A does not vary with each step of the fixed point iteration. This means that we can use a direct solver to obtain the LU decomposition of A only once and then use it each time it appears in the Schur complement matrix $B_n A^{-1} B_n^T$. We can now solve (4.6.4) using a conjugate gradient method, which allows us to only use matrix-vector products, without the need of explicitly creating and storing the full matrix.

- **Parallelization and Scalability:** The above process is efficient in parallel, since we assemble and decompose A only once at the beginning of the gradient flow and then B_n and $\mathbf{f}_n$, which depend on the current solution y_h^n, can be quickly assembled at each iteration. Using the conjugate gradient solver is also more efficient when running in parallel, than trying to solve the whole system using a direct solver. We note, however, that for millions of degrees of freedom the initialization of the LU decomposition of A through a direct solver may not be efficient, despite being performed only once. In that case, a possibly preconditioned iterative solver may be more appropriate, but we have not pursued this option for the current work.

- **Solvability:** At this point, we want to emphasize that the efficiency of the conjugate gradient method depends on the properties of B_n. For instance, if

B_n is full rank, the Schur complement matrix is positive definite and the system is guaranteed to have a unique solution $(\boldsymbol{\delta Y}_h^{n+1}, \boldsymbol{\Lambda}_h^{n+1})$. Another important factor is the condition number of the Schur complement matrix. In Section 4.6.3 we explore this issue. Later in Section 4.7 we provide some computational results that guide us into what we believe to be the optimal approach out of the options we considered.

4.6.3 Existence of Discrete Solution

In this section we explore the solvability of (4.6.1) and of the resulting system in (4.6.3). Since (4.6.1) is a saddle point problem, the natural approach would be to first examine the continuous problem and try to prove the existence and uniqueness of a solution, using an inf-sup condition for appropriate choice of spaces. If we call $\mathbb{V}$ the continuous space of deformations and $\mathbb{Q}$ the corresponding space for the Lagrange multipliers, the importance of an inf-sup condition can be also illustrated by the following calculation that relates the Schur complement of (4.6.3) to an inf-sup condition. We introduce the continuous version of the bilinear forms in (4.6.1), i.e. the continuous bilinear form $a : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$

$$a(y, v) = \int_\Omega D^2 y : D^2 v$$

and the continuous bilinear form $b : \mathbb{V} \times \mathbb{Q} \to \mathbb{R}$

$$b(y, q) = \int_\Omega q : (\nabla y^T \nabla u + \nabla u^T \nabla y),$$

where $u \in \mathbb{V}$ is fixed. As a continuous analogue of (4.6.3), we can then define the operators $A : \mathbb{V} \to \mathbb{V}$, $B : \mathbb{V} \to \mathbb{Q}$ and $B^T : \mathbb{Q} \to \mathbb{V}$ such that

$$< Av, w >_{\mathbb{V}} = < v, Aw >_{\mathbb{V}} = a(v, w) \quad \forall v, w \in \mathbb{V}$$

and

$$< Bv, q >_{\mathbb{Q}} = < v, B^T q >_{\mathbb{V}} = b(v, q) \quad \forall v \in \mathbb{V},\ q \in \mathbb{Q},$$

as well as the analogue of the Schur complement operator $S = BA^{-1}B^T : \mathbb{Q} \to \mathbb{Q}$. We can now see how S being non-singular amounts to an inf-sup argument for a and b:

$$\begin{aligned}
\langle Sq, q \rangle_{\mathbb{Q}} &= \langle BA^{-1}B^T q, q \rangle_{\mathbb{Q}} = \langle A^{-1/2}B^T q, A^{-1/2}B^T q \rangle_{\mathbb{V}} \\
&= \sup_{w \in \mathbb{V}} \frac{\langle w, A^{-1/2}B^T q \rangle_{\mathbb{V}}^2}{\langle w, w \rangle_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{\langle v, B^T q \rangle_{\mathbb{V}}^2}{\langle v, Av \rangle_{\mathbb{V}}} \\
&= \sup_{v \in \mathbb{V}} \frac{b(v, q)^2}{a(v, v)} \geq \frac{\beta^2}{\alpha} \|q\|_{\mathbb{Q}}^2,
\end{aligned}$$

where the last inequality is only possible if $a(v, v) \leq \alpha \|v\|_{\mathbb{V}}^2$ and $b(v, q) \geq \beta \|v\|_{\mathbb{V}} \|q\|_{\mathbb{Q}}$. We see that, given the good properties of the bilinear form a (on the discrete level this is the continuity of a_h implied by the calculations in Lemma 4.2.4 (coercivity of E_h)), the non-singularity of S amounts to the inf-sup condition for the bilinear form b.

In the above analysis we have intentionally not specified the exact spaces for $\mathbb{V}$ and $\mathbb{Q}$. In terms of the existing literature, we are aware of some related work in [50] that would allow us to prove an inf-sup condition for b for appropriate choice of these spaces. However, this work relies on stronger regularity assumptions for the fixed function $u \in \mathbb{V}$, which corresponds to the current surface and which in our case is only assumed to be in $H^2(\Omega)$. This is a restriction that we cannot overcome

in our analysis, since the regularity of u is determined by the gradient flow (4.5.1). As a result, we cannot prove an inf-sup condition and a uniform bound for β at the continuous level.

Since we are not able to prove an inf-sup condition that would allow us to deduce the existence and stability of our discrete solution, we need to rely on our computations, in order to explore the solvability of our problem numerically. This brings us to Section 4.7, where on the one hand we compare two DG-based approaches to the Galerkin one and on the other hand, we compare the solvability of these DG-based approaches in terms of the numerical behavior of their Schur complement.

4.7 Comparison of DG, CDG and Kirchhoff Approaches

We now explore the difference between three different approaches, the fully DG scheme, a continuous-discontinuous scheme (CDG) and the Kirchhoff elements of [8]. The CDG scheme is the same as the one we introduce in Section 4.2 with the only difference that the discrete functions are now assumed to be continuous. This leads to the lower order jumps $[y_h]$ being zero in the interior of the domain. However, since ∇y_h are not assumed to be continuous, this is still a non-conforming approach. Our motivation for this comparison is two-fold. On one hand, we want to shed some light on our decision to pursue a fully DG approach instead of the continuous-discontinuous scheme; a choice that might otherwise seem arbitrary. Our second goal is to illustrate the advantages of the DG approach with respect to the

Kirchhoff elements. The latter is investigated further in Section 4.8. For now we use Example 4.8.1 (vertical load), where we examine five different aspects of the computations: the convergence of the discrete energy and the isometry defect, the number of gradient flow iterations, the deformation along the diagonal of the domain and the average number of Schur iterations needed in each gradient flow step. We present the example here too for convenience.

Example 4.7.1. *Let* $\Omega = (0,4) \times (0,4)$ *and* $\Gamma_D = \{0\} \times [0,4] \cup [0,4] \times \{0\}$ *be the part of the domain where we want to enforce the boundary conditions*

$$g = (x_1, x_2, 0) \quad \textit{and} \quad \Phi_D = [I_2, 0]^T.$$

We apply a vertical force of magnitude 2.5×10^{-2}.

Before we present our results, it is worth explaining our choice of example and metrics. We recall from Lemma 4.5.3 (discrete isometry constraint) that the isometry defect can be expected to converge with order Δt. This was exploited in [8] to present computational rates for the defect, using $\Delta t = h$. Using the same domain, boundary conditions and forcing, as well as the same choice of Δt, allows us to compare the three approaches in terms of computational convergence. Another reason for choosing this example is that it provides the opportunity to compare the geometric flexibility of each method. More precisely, from a physical viewpoint, one would expect that there will be no deformation in the region $\{x \leq 4 - y\}$ for an exact isometry, due to the clamped boundary conditions. This was explored in [8] by observing the deformation along the diagonal $\{x = 4 - y\}$. We examine this in our two DG-based approaches too, in order to showcase their improved geometric

flexibility. It can be seen from Table 4.1 below that the continuous-discontinuous method is the most efficient with respect to these metrics, as well as the number of gradient flow iterations, while both DG approaches outperform the Kirchhoff elements. However, in order to appropriately compare the two DG approaches, we also need to consider the number of Schur complement iterations in each gradient flow step, since this is the main factor affecting their speed. We recall from Section 4.6.2 that we employ the conjugate gradient method, in order to solve the system that corresponds to the Schur complement and obtain the Lagrange multipliers in (4.6.4). Since we have not been able to prove an inf-sup condition that guarantees a stable solution for this system on the continuous or discrete level, we are using the numerical experiments to help us understand how this issue translates in practice for the two methods.

We first observe that the fully DG approach is much faster than CDG in this respect, since the conjugate gradient method for DG requires much fewer iterations than the CDG approach. In fact, the latter often requires even more iterations than the largest dimension of the underlying matrix, which is a strong indication that the continuous-discontinuous galerkin method (CDG) may not satisfy a discrete inf-sup condition. Heuristically, the difference between the two methods could be related to the element-wise nature of the fully DG approach. More precisely, in order to guarantee that the Schur complement system has a unique solution, we need the matrix B_n in (4.6.3) that contains the contribution of the Lagrange multipliers to be full rank. In the fully DG approach this amounts to showing that for given y_h^n

(previous solution) and for all $T \in \mathcal{T}_h$

$$\int_T \lambda_h^{n+1} : \left(\nabla v_h^T \nabla y_h^n + (\nabla y_h^n)^T \nabla v_h\right) = 0 \quad \forall v_h \in \mathbb{V}_h \Longrightarrow \lambda_h^{n+1} = 0 \text{ on } T.$$

On the other hand, the CDG approach does not allow such an element-wise simplification, since the continuity of y across interelement boundaries leads to an interplay between different cells.

We now present several simulations, that allow us to compare the three methods based on the described metrics. We start with Table 4.1, where we exhibit the discrete energy denoted by E_h, the discrete isometry defect, denoted by I_h, and the number of gradient flow iterations, # GF, for all three methods and three different mesh sizes, $h = 2^{-l},\ l = 2, 3, 4$. We observe that the CDG method outperforms the other two with respect to all three metrics. Moreover, the fully discontinuous approach also outperforms the Kirchhoff elements method, especially in terms of the magnitude and the speed of convergence of the discrete energy and the discrete isometry defect. The number of gradient flow iterations for the fully DG and the Kirchhoff method seems to behave similarly.

h	E_h			I_h			# GF		
	Kirch.	CDG	DG	Kirch.	CDG	DG	Kirch.	CDG	DG
2^{-2}	-1.01e-2	-5.23e-3	-7.53e-3	8.67e-3	4.61e-3	7.10e-3	22	11	13
2^{-3}	-9.82e-3	-3.29e-3	-5.76e-3	7.12e-3	1.53e-3	2.89e-3	40	18	28
2^{-4}	-9.04e-3	-1.91e-3	-4.26e-3	5.14e-3	4.52e-4	1.09e-3	71	33	76

Table 4.1: Comparison of discrete energy, E_h, discrete isometry defect, I_h, and number of gradient flow iterations, # GF, for the Kirchhoff, the continuous-discontinuous (CDG) and the fully discontinuous (DG) method.

Geometric Flexibility. We next present three figures that illustrate the geometric flexibility of each method, measured by the artificial deformation created along the diagonal $\{x + y = 4\}$. Smaller deformation implies a more consistent behavior with the physics of the problem. We first attach Figure 3 from [8]. It presents the diagonal deformation for meshes of size $2^{-l}, l = 2, \ldots, 6$. We then present two figures that correspond to the CDG and the fully DG approach for mesh sizes $h = 2^{-l}, l = 2, 3, 4$. We observe that, similarly to the results in Table 4.1, the CDG approach performs the best. In fact, both the DG approaches outperform the Kirchhoff elements significantly. For instance, we observe that for $h \leq 2^{-2}$ and $h \leq 2^{-3}$ for the CDG and the fully DG method, respectively, the diagonal deformation is less than 0.02, which is the smallest value observed in Figure 4.1 and corresponds to $h = 2^{-6}$.

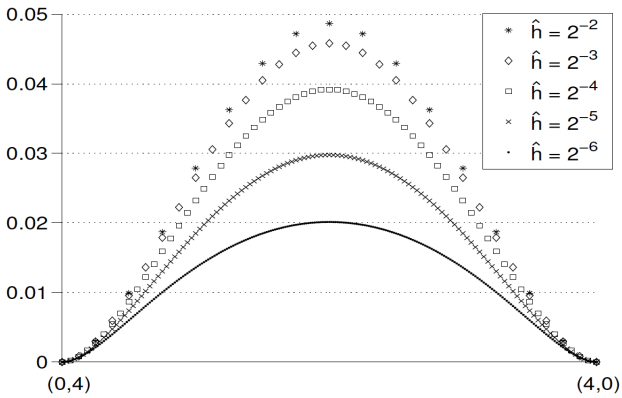


Figure 4.1: Deformation along the diagonal $x + y = 4$ for Example 4.8.1 in [8], using the Kirchhoff elements.

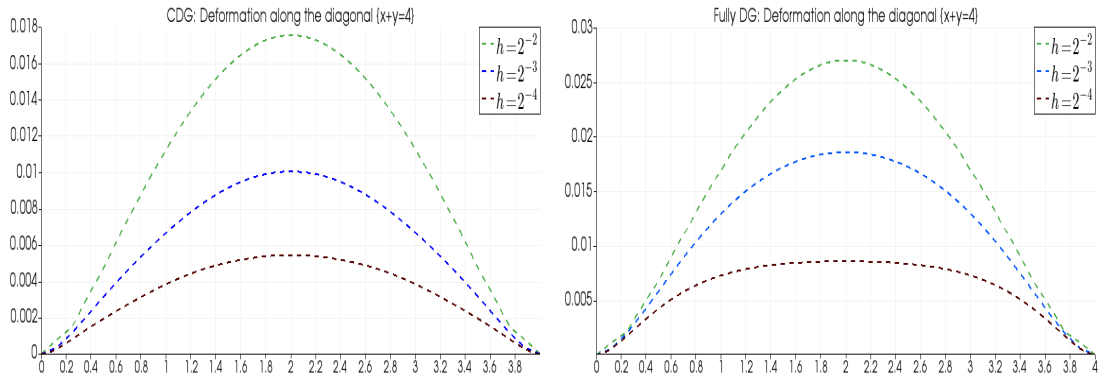


Figure 4.2: Deformation along the diagonal $x + y = 4$ for Example 4.8.1 (vertical load), using the CDG (left) and fully DG (right) approaches for the same geometry and f as in [8]. We observe smaller deformations than the ones for Kirchhoff elements. The smallest deformation in Figure 4.1 for $h = 2^{-6}$ is achieved with $h = 2^{-2}$ and $h = 2^{-3}$ by the CDG and DG methods, respectively.

The previous simulations indicate a clear computational advantage of the two DG-flavored approaches with respect to the Kirchhoff elements. They outperform the Kirchhoff element approach both in terms of convergence of E_h and I_h as well as geometric flexibility, while they are simpler to implement in standard libraries and require polynomial degree 2 rather than 3.

Conjugate Gradient Iterations: We now compare the two methods with respect to the last metric, the number of conjugate gradient iterations per gradient flow step. We illustrate the results on Table 4.2. Note that the Schur complement is a square matrix with dimension equal to the number of degrees of freedom for the Lagrange multipliers. We observe that the amount of conjugate gradient iterations per gradient flow step for the continuous-discontinuous approach exceeds by a multiplicative factor the dimension of the Schur complement. On the other hand, the

respective number of iterations for the fully DG method is not only much smaller, but is also a decreasing fraction of the dimension of the Schur complement. This behavior of the CDG method makes an asymptotic comparison very hard, since the method becomes much slower with each refinement. This is why we have restricted ourselves in the cases $h = 2^{-l}, l = 2, 3, 4$. This clear advantage of the fully DG approach is also the reason why it is our method of choice. In Section 4.8 we explore more the method and present more asymptotic results for it.

h	CDG		DG	
	S. C. dim	# Iterations	S. C. dim	# Iterations
2^{-2}	768	3,900	768	350
2^{-3}	3072	11,200	3072	470
2^{-4}	12,288	25, 000	12,288	1220

Table 4.2: Comparison of the number of conjugate gradient iterations for the continuous-discontinuous (CDG) and the fully discontinuous (DG) method. S.C. dim stands for the dimension of the Schur complement matrix.

4.8 Numerical Experiments

In this section we explore and compare the performance of our method with that of the Kirchhoff elements. We are interested in the speed and accuracy of the method, as well as its ability to capture the physics and the geometry of the problems appropriately. We observe that our Discontinuous Galerkin approach seems to be more flexible with a comparable or better speed. We provide specific examples in the rest of the section. We emphasize that we use polynomial degree $k = 2$ for all examples.

4.8.1 Vertical load on a square domain

This is a simple example of bending due to vertical load. We use the same configuration as in [8] in order to provide an accurate comparison of the two methods. We deal with a square domain with two of the non-parallel sides being clamped.

Example 4.8.1. *Let* $\Omega = (0,4) \times (0,4)$ *and* $\Gamma_D = \{0\} \times [0,4] \cup [0,4] \times \{0\}$ *be the part of the boundary where we enforce the boundary conditions*

$$g = (x_1, x_2, 0) \quad \textit{and} \quad \Phi_D = [I_2, 0]^T.$$

We apply a vertical force of magnitude 2.5×10^{-2}.

We first revisit Table 4.1 that illustrates the convergence of the energy E_h and the isometry defect I_h as the mesh size decreases, but this time it is restricted only to our method (fully DG) and includes more refinements. We choose again $\Delta t = h$ and observe that the isometry defect decays super-linearly with Δt , which is better than the linear convergence predicted in Lemma 4.5.3 (discrete isometry constraint). Compared to [8], we obtain a more clear rate, while the defect itself is smaller (up to one order of magnitude). The number of gradient flow iterations is similar for both methods.

No. Cells	DoFs	$h = \Delta t$	E_h	Isometry Defect	Iterations
256	7680	2^{-2}	-7.53e-3	7.10e-3	13
1024	30,720	2^{-3}	-5.76e-3	2.89e-3	28
4096	122,880	2^{-4}	-4.26e-3	1.09e-3	76
16384	491,520	2^{-5}	-3.30e-3	4.17e-4	140

Table 4.3: Number of cells, degrees of freedom, discrete energy, isometry defect and number of gradient flow iterations for square plate, clamped in two sides with vertical forcing. We observe super-linear rates for the isometry defect, while theory predicts linear rates for the case $\Delta t = h$.

We now explore the geometric behavior of our method. More precisely, it was observed in [8] that there was an artificial displacement along the diagonal $x+y = 4$, which does not correspond to the actual physics of the problem. This displacement becomes smaller with smaller mesh size. Our method introduces the same artificial displacement. However, we can see in the following figure that in our case this displacement is smaller, even by one order for the smallest mesh size.

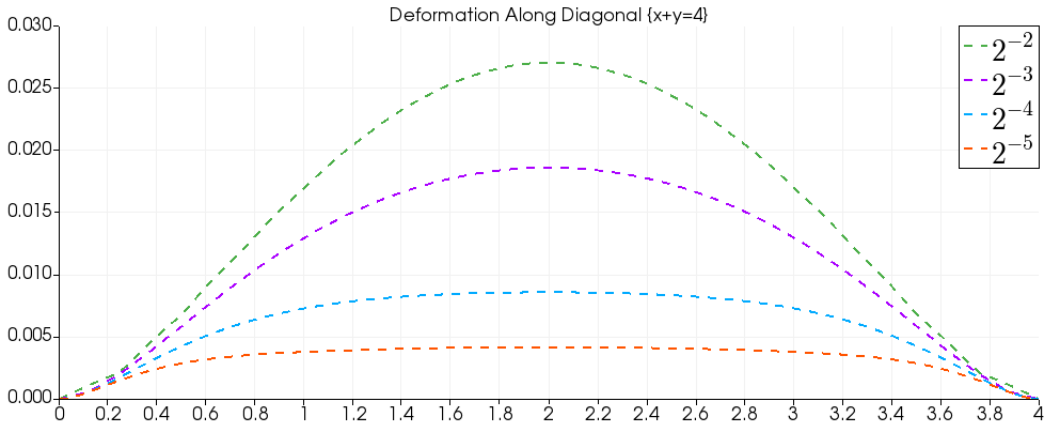


Figure 4.3: Deformation along the diagonal $x + y = 4$. We observe smaller deformations than with the Kirchhoff elements (up to one order of magnitude).

4.8.2 Obstacle Problem

In the previous example we observe some of the flexibility of the DG approach in terms of capturing the physical behavior of the plates. To explore this further, we introduce an extra element to the deformation: an obstacle. We use a square plate clamped on one side and we exert a vertical force. We require that the plate does not penetrate the obstacle. We choose a simple case where the obstacle is a rigid flat plate at height $z = 0.2$. This example is motivated by [10], where the deformation of the plate is the result of thermal actuation of bilayer hinges connected to the plate. In our case, we use a vertical load instead. From a mathematical viewpoint, this obstacle problem can be treated by introducing the convex set of the constraint

$$K = \left\{ y \in [L^2(\Omega)]^3 \quad | \quad y_3 \leq 0.2 \right\},$$

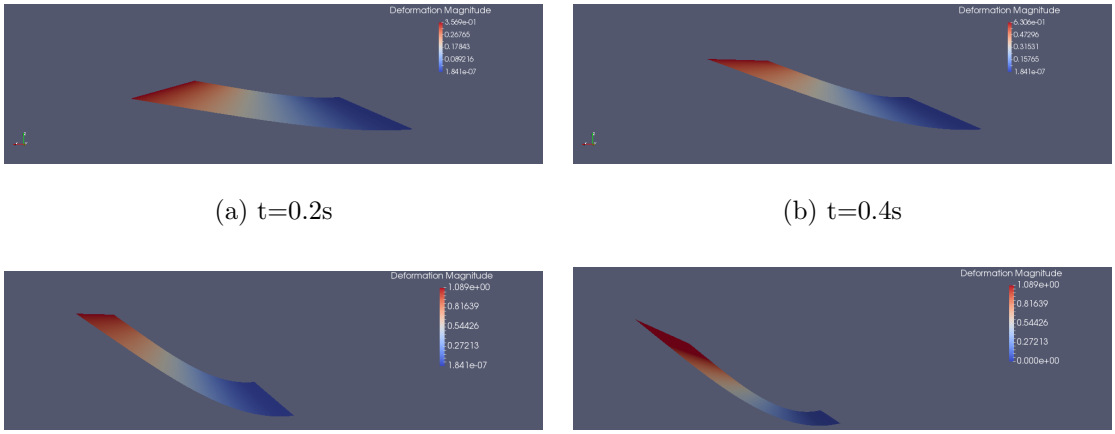
along with a splitting of variables. We introduce another deformation s, $s \approx y$ such that $s \in K$ always and penalize the $L^2(\Omega)$-distance between y and s. If ε is an obstacle penalty parameter, we add the following extra term to the energy

$$\frac{1}{\varepsilon} \| y - s \|^2_{L^2(\Omega)}.$$

At the discrete level, this affects the gradient flow at each step, where we use as s_h the L^2-projection of the previous solution y_h^n in K. We refer to [10] for more details about the variable splitting and the projection.

Example 4.8.2. *Let $\Omega = (-1, 1) \times (-1, 1)$. Let $\Gamma_D = \{-1\} \times [-1, 1]$ be the part of the boundary where we enforce clamped boundary conditions. We apply a vertical force $f = (0, 0, 1)$.*

To describe the performance of our method, we first present the deformation of the plate for three different time steps for the same configuration but without any obstacle. We see that the plate remains clamped on the side while the third component increases and the plate eventually bends.



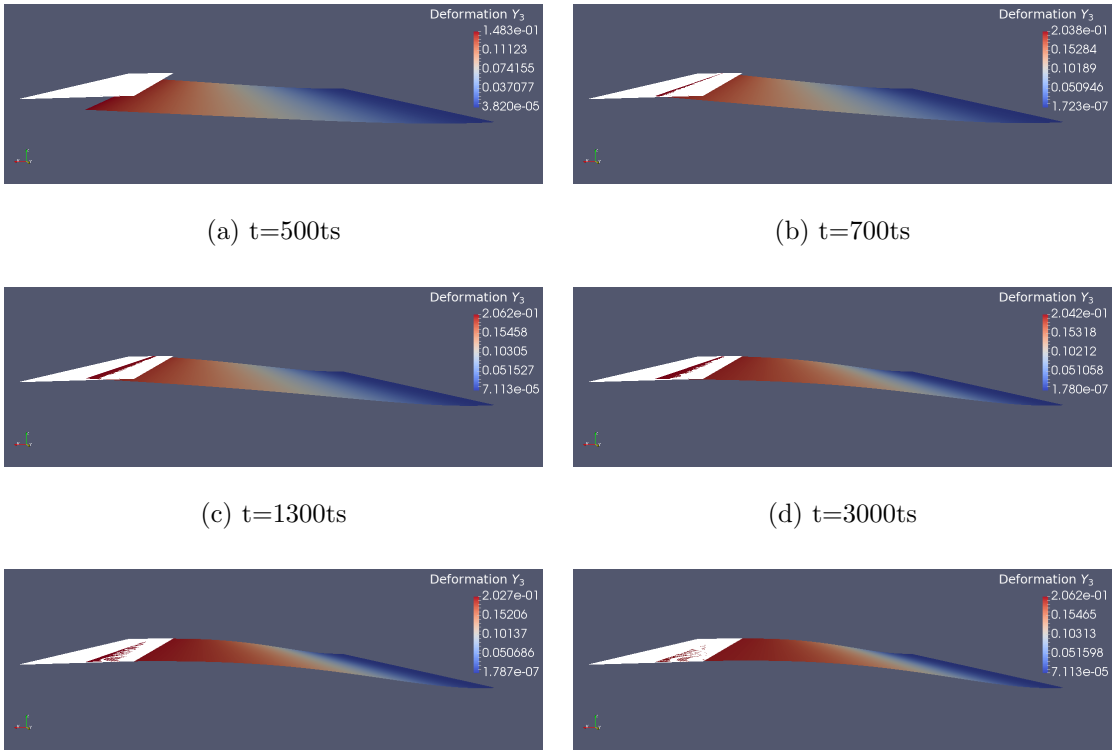
(a) t=0.2s (b) t=0.4s

(c) t=1.0s (d) t=4.0s

Figure 4.4: Deformation of the plate at four different stages: (a) Initial stages (b) Slight Bending (c) Increased Bending, (d) Final state.

From a physical viewpoint we would expect that the obstacle would make no difference at the first steps of the deformation. However, when the plate meets the obstacle, we expect it to stop rising and start bending accordingly so that it complies with the obstacle and the force. We observe that this is indeed the case. The plate rises until it meets the obstacle and then adjusts by bending. However, since we do not force y_h numerically to belong to the set K, but rather penalize its L^2-distance from its projection in K, we see that there is a minor penetration of the obstacle. The plate adjusts to the presence of the obstacle and this crossing decreases. This

behavior can be enhanced by further refining and by choosing the obstacle parameter ε to be smaller. To illustrate the deformation, we provide six different time steps in Figure 4.5. The illustration corresponds to 1024 cells, $\varepsilon = 3 \times 10^{-4}$ and time step $\Delta t \ = 5 \times 10^{-4} s$. The isometry defect at the end of this simulation is only 9.1×10^{-5}.



(a) t=500ts (b) t=700ts

(c) t=1300ts (d) t=3000ts

(e) t=5000ts (f) t=13382ts

Figure 4.5: Deformation of the plate at six different stages: (a) Initial stage, deformation determined only by forcing, (b) First contact with the obstacle, (c) Penetration of the obstacle, (d) Starts adjusting shape to decrease crossing, (e) Stronger bending, decreased crossing, (f) Final stage, crosses obstacle by 0.002 (1% of obstacle height).

4.8.3 Compressive Case

As our last example, we explore another geometrically and physically interesting case, that of compression. We use a rectangular plate with a small vertical load and we impose compressive boundary conditions at two opposite sides.

Example 4.8.3. *Let* $\Omega = (-2, 2) \times (0, 1)$. *Let also* $\Gamma_D = \{-2, 2\} \times [0, 1]$ *be the two sides where we impose the compressive boundary conditions*

$$g = (0.3x_1, x_2, 0) \quad \textit{and} \quad \Phi_D = [I_2, 0]^T.$$

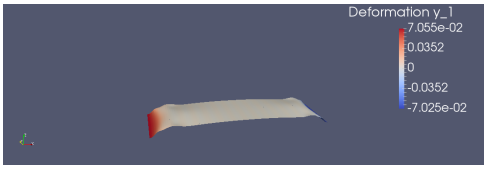
We apply a vertical force of magnitude 10^{-2}.

Given the compressive nature of the problem, the plate could either bend upwards or downwards (buckling), resulting to two deformations which are the reflection of each other with respect to the x-y plane and have the same minimizing energy. This is why we apply a small force, in order to choose one of the two deformations, which in our case corresponds to a positive third component. We note here that we need a stronger force than the one in [8] (10^{-5}). A possible explanation is that our huge initial energy may require a stronger force in order to be guided towards one or the other minimizing states. We start again with a flat configuration as our initial guess. This results in a big initial energy due to the Nitsche boundary terms. In order to prevent this from creating a very abrupt and non-physical deformation in the gradient flow, we employ an approach of quasi-static nature: we do not enforce the boundary conditions from the beginning of the flow, but rather do it gradually. We use a parameter that starts from zero and increases throughout

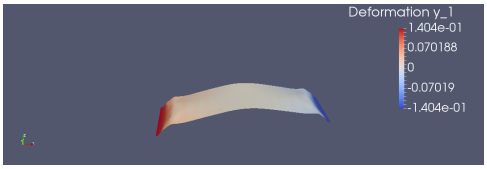
the flow until it reaches the value 1. In order to achieve a gradual adjustment to the boundary conditions, we use

$$(1 - bc\_flow)\ id + bc\_flow\ g_h,$$

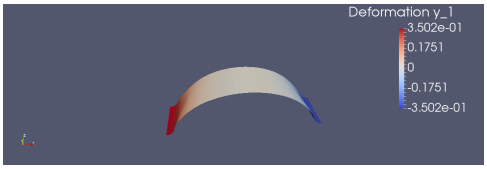
where bc_flow is the above mentioned parameter and id stands for the identity function. As the mesh becomes finer, this parameter needs to grow slower, in order to compensate for the big initial energy and allow for a more natural flow. To illustrate the effect of this parameter, we depict in Figure 4.6 the deformation that corresponds to various stages of the flow. The time step of the gradient flow is $\Delta t = 0.04625s$. The bc_flow parameter increases linearly by 5×10^{-5} in each iteration. We see that the compressive nature of the boundary conditions becomes more obvious after the bc_flow parameter becomes greater than $1/2$ and that the boundary conditions are achieved at the end of the deformation. The isometry defect at the end of the flow is $I_h[y_h] = 5.04\ 10^{-3}$.



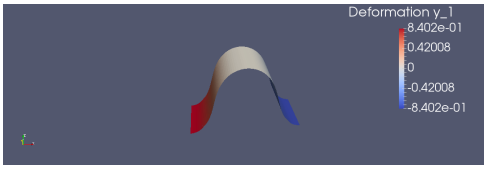
(a) bc_flow=0.05



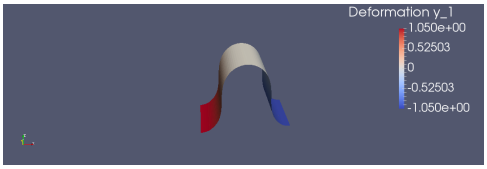
(b) bc_flow=0.1



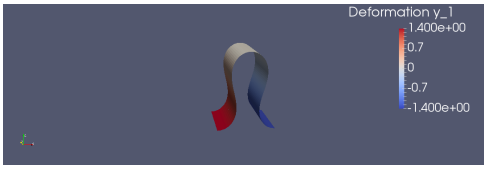
(c) bc_flow=0.25



(d) bc_flow=0.6



(e) bc_flow=0.65



(f) bc_flow=1

Figure 4.6: Deformation of the plate at seven different stages: (a)-(c) Initial stages, slow deformation determined mostly by the forcing, (d)-(e) The compressive boundary conditions become more active and determine the shape of the plate, (f) Final deformation, compressive boundary conditions are achieved.

Chapter 5: Bilayer Model - Computational Extension

In this Chapter we explore computationally the applicability of our DG method from Chapter 4 to more physically interesting cases that can be modeled by attaching together two plates with different mechanical properties. We briefly introduce this model and derive a modification to our method, which allows us to deal with this model. We test the resulting formulation using various examples from [10, 11] and [2, 53, 75].

5.1 Simplified Bilayer Model

We first recall briefly the bilayer model from [11] and use an approach similar to [10] to adjust the method developed in Chapter 4 to the bilayer case. On the continuous level the bilayer problem involves the minimization of the energy functional

$$E[y] = \frac{1}{2}\int_\Omega |H + Z|^2 - \int_\Omega fy, \tag{5.1.1}$$

where y is the deformation of the midplane Ω of a plate constructed by clamping together two layers with different material and mechanical properties. The matrix function Z corresponds to this mismatch between the different materials and encodes the spontaneous curvature of the bilayer plate. In the above energy H stands again

for the second fundamental form of y, while f is a forcing term in $[L^2(\Omega)]^3$, which from now on we assume to be zero, since the plates are inclined to deform due to Z without the contribution of an external force. Similarly to Section 4.1 for the single-layer problem, (5.1.1) can be derived from three dimensional hyperelasticity [11]. In order to discretize the above functional in a way that can be combined with our discretization of the single layer, we follow an approach similar to [10]. As a first step, we repeat some calculations from Section 4.1. Let z_{ij} and h_{ij}, $i,j=1,2$ be the components of Z and H and let y be an isometry. By the definition of an isometry we have that

$$\partial_i y \cdot \partial_j y = \delta_{ij}, \quad i,j=1,2,$$

hence

$$\partial_k y \cdot \partial_{ij} y = 0 \Longrightarrow \partial_{ij} y \parallel \nu = \partial_1 y \times \partial_2 y, \quad i,j,k=1,2$$

which, in turn, combined with the definition of the second fundamental form, leads to

$$h_{ij} = \partial_{ij} y \cdot \nu = \partial_{ij} y.$$

Similarly, using that $H : Z = \sum_{ij} \partial_{ij} y \cdot z_{ij} \nu$ we can rewrite (5.1.1) as

$$E[y] = \frac{1}{2} \int_{\emptyset} |D^2 y|^2 + 2 \sum_{ij} \partial_{ij} y \cdot z_{ij} \nu + |Z|^2.$$

Taking now the Euler-Lagrange equations for this functional we obtain

$$\int_\Omega D^2 y : D^2 v + \int_\Omega \sum_{ij} \partial_{ij} v \cdot z_{ij} \, (\partial_1 y \times \partial_2 y) + \int_\Omega \sum_{ij} \partial_{ij} y \cdot z_{ij} \, (\partial_1 y \times \partial_2 v + \partial_1 v \times \partial_2 y) = 0$$

for all $v \in [H^2(\Omega)]^3$ such that

$$\nabla y^T \nabla v + \nabla v \nabla y^T = 0 \quad \text{in } \Omega. \tag{5.1.2}$$

We now use again the fact that y is an isometry, hence $\partial_{ij} y = h_{ij}\nu$, along with (5.1.2) and the formula $(a \times b) \cdot (c \times d) = (a \cdot c)(b \cdot d)$, to see that

$$\begin{aligned}
\partial_{ij} y \cdot z_{ij}\ (\partial_1 y \times \partial_2 v + \partial_1 v \times \partial_2 y) &= z_{ij}\ h_{ij}\ (\partial_1 y \times \partial_2 y) \cdot (\partial_1 y \times \partial_2 v) \\
&\quad + z_{ij}\ h_{ij}\ (\partial_1 y \times \partial_2 y) \cdot (\partial_1 v \times \partial_2 y) \\
&= z_{ij}\ h_{ij} \underbrace{(\partial_1 y \cdot \partial_1 y)}_{1} \underbrace{(\partial_2 y \cdot \partial_2 v)}_{0} \\
&\quad + z_{ij}\ h_{ij} \underbrace{(\partial_1 y \cdot \partial_1 v)}_{0} \underbrace{(\partial_2 y \cdot \partial_2 y)}_{1} \\
&= 0.
\end{aligned}$$

This means that we can re-write the Euler Lagrange equations as

$$\int_\Omega D^2 y : D^2 v + \int_\Omega \sum_{ij} \partial_{ij} v \cdot z_{ij}\ (\partial_1 y \times \partial_2 y) = 0. \tag{5.1.3}$$

We observe that these equations involve two terms, only the first of which is present in the corresponding Euler-Lagrange equations of the single layer energy functional. We exploit this, along with the good properties of the single layer discretization, in order to come up with the simplest possible discretization of the above equations. Since our discretization in Sections 4.2 and 4.5 involves all the appropriate terms that capture the discontinuities in the discrete space $\mathbb{V}_h^k$ and allow for a good approximation of the Hessians $D^2 y$ and $D^2 v$, we expect that using the same discretization for (5.1.3) can potentially provide an equally good approximation. We thus discretize (5.1.3) by solving the following slight modification of the discrete gradient flow presented in Section 4.5:

$$\begin{aligned}
\big(\delta y_h^{n+1}, v_h\big)_h + \tau\ a_h\big(\delta y_h^{n+1}, v_h\big) &= \tau\ (f_h, v_h)_{L^2(\mathcal{T}_h)} - \tau\ a_h\big(y_h^n, v_h\big) \\
&\quad - \tau \sum_{T \in \mathcal{T}} \int_T \sum_{ij} z_{ij}\ \partial_{ij} v_h \cdot\ (\partial_1 y_h^n \times \partial_2 y_h^n).
\end{aligned}$$

The extra term on the right hand side corresponds to the second term in (5.1.3). We note here that this term is independent of the unknown δy_h^{n+1} because it involves the previous solution y_h^n and is only part of the right hand side of the flow. Moreover, we emphasize that it involves only a piecewise computation of the added quantity without taking into account any possible jumps of $\partial_{ij} y_h^n$ and of $\partial_1 y_h^n \times \partial_2 y_h^n$. We also stress that this discretization follows along the lines of [10], where the extra term was treated explicitly and added to the right hand side and, unlike [11], no procedure is used to guarantee that $|\partial_1 y_h^n \times \partial_2 y_h^n| \geq 1$. Consequently, we do not investigate here any theoretical properties of the discretization, but rather explore its efficiency by testing it over various numerical experiments, similar to the ones presented in [10, 11]. For a rigorous study of the the DG approach to the bilayer model, we would need to employ the discrete Hessian $H_h[y_h]$ constructed in Section 4.3 in order to approximate $\partial_{ij} y_h^n$ and possibly introduce jump terms and normalization of $\partial_1 y_h^n \times \partial_2 y_h^n$. Lastly, we would need to create an implicit version of this flow that would instead use $\partial_1 y_h^{n+1} \times \partial_2 y_h^{n+1}$, in order to prove $\Gamma-$convergence of an appropriate discrete energy $\tilde{E}_h$ to the continuous (5.1.1).

5.2 Numerical Experiments

In this section we explore various examples motivated by the work in [10, 11] and experimental work in [2, 53, 75] in order to understand better the computational behavior of the Discontinuous Galerkin discretization introduced in Chapter 4. We wish to observe whether the simple extension of the method, presented in Section 5.1,

can lead to interesting computational results, that capture the physical properties of the bilayer bending problem. To this end, we test our method in a variety of examples, exploring the effect of different choices of the spontaneous curvature matrix Z, the boundary conditions and the aspect ratio of the plate. All these factors are crucial for any engineering application, since they can be appropriately combined in order to achieve various final configurations for the plate that undergoes large, geometrically nonlinear deformations.

5.2.1 Boundary Conditions

Before we proceed with the specific examples, it is worth discussing briefly our approach for boundary conditions (B.C.). Our experiments fall into two main categories:

- **Dirichlet B.C. on $\Gamma_D \subset \partial\Omega$:** This case is covered by our theory in Section 4.2.4, by employing the jump terms and the respective stabilizations only on the boundary edges that belong to Γ_D. Our discrete energy is coercive in $\Omega \cup \Gamma_D$ and does not see $\partial\Omega \setminus \Gamma_D$. Our analysis for the $\Gamma-$convergence of E_h guarantees that the limiting y satisfies the boundary conditions on Γ_D.

- **B.C. Free:** In this case we do not pose any boundary conditions. Instead, we allow the thin plate to deform without any restrictions. In order to achieve this numerically, we need to aid our linear system in providing a unique discrete solution. This entails removing rotations and constants from the kernel of our system. In the previous case, this would be guaranteed by the boundary conditions, but,

in their absence, we now need to do this manually. To this end, we adjust the discrete energy inner product $(\cdot,\cdot)_h$ used in the gradient flow in Section 4.5. We add an extra L^2 term multiplied by a small parameter that in all our examples is taken to be 10^{-2}. This lower order term allows us to have a non-singular system even in the absence of boundary conditions and prevents us from obtaining solutions that are up to a constant or a rotation.

5.2.2 Clamped Plate, $Z = -I$

We start with the simplest case of a rectangular plate $\Omega = [-5, 5] \times [-2, 2]$, clamped on the side $\{-5\} \times [-2, 2]$, with spontaneous curvature given by $Z = -I$. In [73] it was shown that the minimizer for this case is a cylinder. We present below computations leading to a cylindrical final configuration, which was also the case in [11]. It is worth mentioning here that in our case this stage is achieved without any self-crossing of the plate, which is in general hard to prevent.

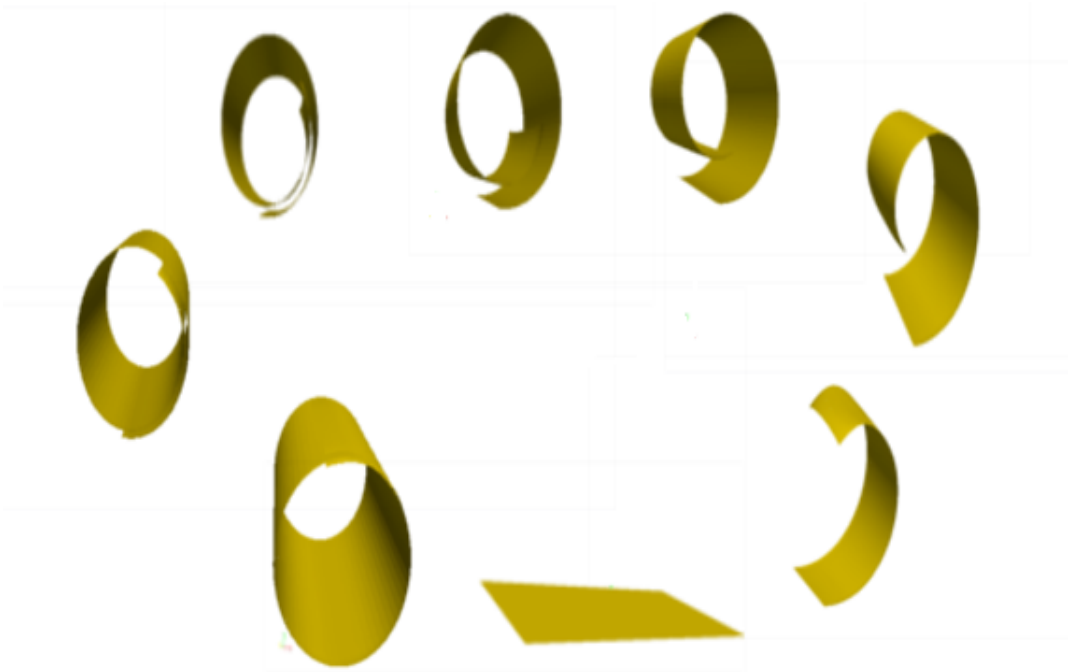


Figure 5.1: Pseudo-evolution (counter-clockwise) towards the equilibrium of a clamped rectangular plate with spontaneous curvature $Z = -I$. The bilayer plate is depicted at times 0.0, 0.5, 1.5, 3.0, 5.0, 8.0, 10.0, 20.0 $\times 10^{-3}$ of the gradient flow. The plate reaches a cylindrical shape asymptotically, which is an absolute minimizer.

5.2.3 Clamped Plate, Anisotropic Curvature

We now explore a different spontaneous curvature matrix. We examine a plate $[-2, 2] \times [-3, 3]$, clamped along the side $[-2, 2] \times \{-3\}$ with the anisotropic spontaneous curvature

$$Z = \begin{bmatrix} -3 & 2 \\ 2 & -3 \end{bmatrix}.$$

This choice of Z corresponds to principal curvatures 5 and 1, which in contrast to the previous example correspond to principal directions that do not align with the coordinate axes, but form an angle of $\pi/4$ with them. This, combined with

the clamped boundary conditions, leads to a different deformation that does not completely assume the cylindrical shape, but rather gradually "rolls" into a conic shape (corkscrew). We illustrate this deformation in Figure 5.2.

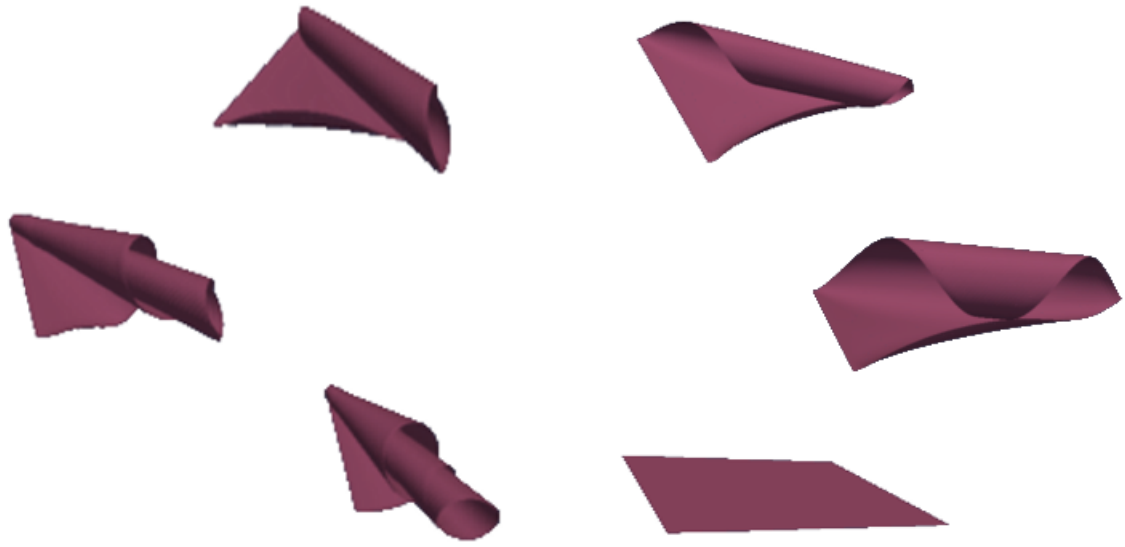


Figure 5.2: Pseudo-evolution (counter-clockwise) towards the equilibrium of a clamped rectangular plate with anisotropic spontaneous curvature and principal directions forming an angle of $\pi/4$ with the coordinate axes. The bilayer plate is depicted (counter-clockwise) at times 0.0, 0.3, 1.0, 10.0, 50.0, 170.0 $\times 10^{-3}$ of the gradient flow.

5.2.4 Clamped Plate, Principal Curvatures of Opposite sign

We consider the same plate, $[-2,2]\times[-3,3]$, clamped along the side $[-2,2]\times\{-3\}$, but with spontaneous curvature

$$Z = \begin{bmatrix} 5 & 0 \\ 0 & -5 \end{bmatrix}.$$

The principal curvatures have now opposite signs. At first each curvature tries to force its own deformation of the plate, but eventually one dominates and the structure assumes a cylindrical configuration, as depicted in Figure 5.3.

Figure 5.3: Pseudo-evolution towards the equilibrium of a clamped rectangular plate with principal curvatures of opposite sign. The bilayer plate is depicted (bottom-left to top-right) at times 0.0, 0.7, 2.5, 7.0, 10.0, 12.0, 14.0, 19.0, 67.0 $\times 10^{-3}$ of the gradient flow.

5.2.5 Plate Clamped in the Middle, Anisotropic Curvature

We now explore an example that was motivated by [2], where a rectangular plate is clamped in the middle and then allowed to deform with anisotropic spontaneous curvature. The plate is given by $[-5, 5] \times [-2, 2]$, clamped along the middle

line $\{0\} \times [-2, 2]$ and

$$Z = \begin{bmatrix} -5 & 0 \\ 0 & -1 \end{bmatrix}.$$

This leads to the formation of two cylinders in each side of the clamped part. In this example our numerical experiments exhibit some self-crossing before the two cylinders separate. Eventually, the cylinders reach a distance from each other, possibly due to the way the clamping is forced numerically (along all elements that contain the middle line of the plate).

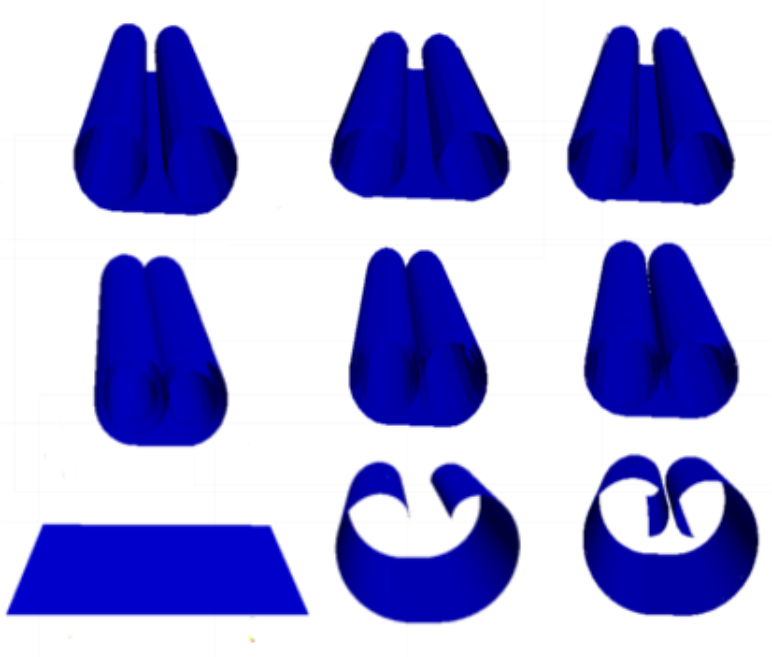


Figure 5.4: Pseudo-evolution towards the equilibrium of a rectangular plate clamped in the middle, with anisotropic spontaneous curvature. The bilayer plate is depicted (bottom-left to top-right) at times 0.0, 0.1, 0.5, 1.5, 2.5, 5.3 $\times 10^{-3}$ of the gradient flow.

5.2.6 Free Plate, Anisotropic Curvature

We now explore a cigar-type configuration motivated by work of a group in mechanical engineering [75]. This time the same plate, $[-5,5]\times[-2,2]$, is completely devoid of boundary conditions and has an anisotropic spontaneous curvature

$$Z = \begin{bmatrix} -3 & 2 \\ 2 & -3 \end{bmatrix}.$$

We observe in Figure 5.5 that the plate deforms similarly to Figure 5.2, but since there is no clamped side, it does so in a symmetric way and eventually assumes a cylindrical configuration (cigar).

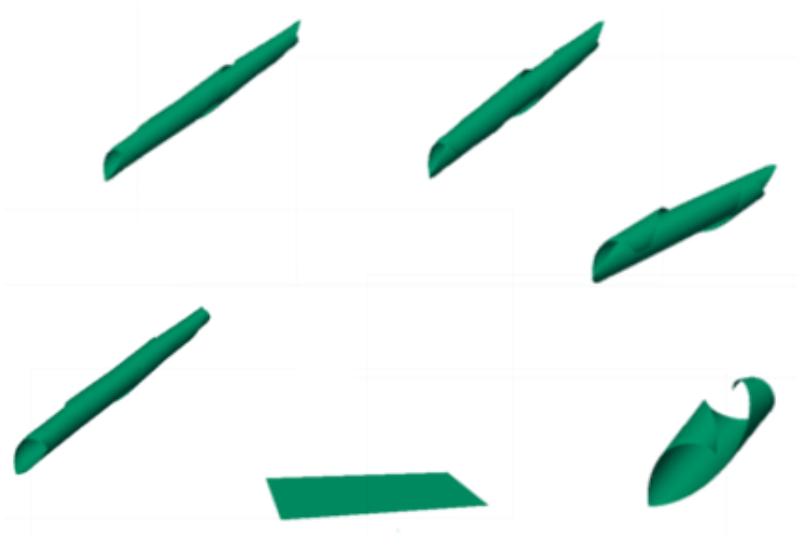


Figure 5.5: Pseudo-evolution towards the equilibrium of a rectangular plate with anisotropic spontaneous curvature and no boundary conditions. The bilayer plate is depicted (counter-clockwise) at times 0.0, 0.2, 0.6, 6.0, 10.0, 12.0, 20.0, 60.0, 150.0 $\times 10^{-3}$ of the gradient flow. At the final three steps presented we observe that the plate assumes a tighter configuration as an effort to become a full cylinder.

5.2.7 Free Plate, Wavy Pattern

In this example we explore wavy patterns attained experimentally in [53]. The authors create various geometric shapes including wave patterns and DNA-like structures (next example) using bilayer and multilayer materials with the goal of then using these shapes for more complex self-folding and self-organizing structures. We present here our numerical recreation of the wave pattern and in the next example a DNA-like helix, constructed similarly to Figure 5.5. In this case the wave pattern is created experimentally by alternating the position of the polymer that acts as the bilayer. More precisely, the plate is split into eight parts of equal size and the bilayer formulation is forced on the upper side of one part and the lower side of the next one. Numerically we implement this by alternating the sign of the spontaneous curvature between each part. The plate has dimensions $[-8, 8] \times [-1, 1]$ and in each part we use $Z = I$ or $Z = -I$. We do not pose any boundary conditions to the plate. The behavior of our simulations is illustrated in Figure 5.6. We note that the gradient flow is much faster than previous examples and, as a consequence, this simulation is depicted for much fewer gradient flow steps: the plate assumes the fourth configuration very fast and afterwards the deformation and the energy change very slowly.

Figure 5.6: Pseudo-evolution towards the equilibrium of a rectangular plate with high aspect ratio and alternating curvature along each of its eight parts. The bilayer plate is depicted (bottom-left to top-right) at times 0.0, 0.1, 1.0, 2.3 $\times 10^{-3}$ of the gradient flow.

5.2.8 Free Plate, DNA

We now present the second example motivated by [53]. In this case, we have a plate with a high aspect ratio and dimensions $[-8, 8] \times [-0.5, 0.5]$ that is free to deform without any boundary conditions, under the effect of a spontaneous curvature given by

$$Z = \begin{bmatrix} -1 & 3/2 \\ 3/2 & -1 \end{bmatrix}.$$

This choice of spontaneous curvature corresponds to principal directions that form an angle of $\pi/4$ with the coordinate axes and to principal curvatures $-1/2$ and $5/2$. We observe numerically that this combination of magnitude for the principal curvatures and of aspect ratio lead to a deformation that resembles the twisting of DNA molecules. A different combination would lead to another deformation of similar nature, but with different visual results in terms of the twisting of the plate.

We illustrate the resulting deformation in Figure 5.7 for various instances of the gradient flow.

Figure 5.7: Pseudo-evolution towards the equilibrium of a rectangular plate with high aspect ratio and anisotropic curvature with principal directions forming an angle of $\pi/4$ with the axes. The bilayer plate is depicted (bottom-left to top-right) at times 0.0, 0.5, 2.0, 5.0, 15.0, 48.0 $\times 10^{-3}$ of the gradient flow.

5.2.9 Computational Investigation of the Effect of Aspect Ratio, Curvature and Numerical Constants

In this section we discuss various physical and numerical aspects that can affect the performance of the method. To this end, we focus on the clamped example of Section 5.2.2 with $Z = -rI$ and we present a numerical illustration of the final configuration for a range of aspect ratios (2.5, 2, 1.5, 1) and curvatures $r = 5, 3, 1$. We compare with the results from [11] and try to get some intuition on the varying effects of these choices, as well as of the penalty constants γ_0 and γ_1. Before we proceed, we need to clarify certain aspects of the following simulations:

- **Mesh refinements:** The results are presented for 4 uniform refinements, i.e.

256 cells, whereas in [11] results of this nature were presented for 5 uniform refinements. However, most of the individual simulations have been tested for both refinement levels in order to verify that the shape of the final configuration is stable with respect to refinements. We intend to continue this investigation further.

- **Steady state:** A natural way to terminate the simulations is to guarantee that upon two gradient flow steps the change in the energy is less than a chosen tolerance. We call the configuration of the plate at the last time step as steady state. However, since this is an ongoing investigation and such comparative simulations are time consuming, we present results that correspond to a total of $80,000$ gradient flow steps. This number was chosen empirically, since the visible change of the shape of the plate after this many iterations is minor and we have observed that in all simulations the energy keeps decreasing very slowly from a point on, without contributing any further to the shape of the final configuration. We acknowledge that not having a steady state configuration prevents us from rigorously drawing certain conclusions that we wish to present later.

- **Numerical Parameters:** Figure 5.8 is an attempt to present a unified picture of the various aspect ratios and curvature that we experimented with for the problem of Section 5.2.2. However, not all of the simulations presented correspond to the same numerical parameters (time step, accuracy for the conjugate gradient solver, penalty constants). We expand on this in details below.

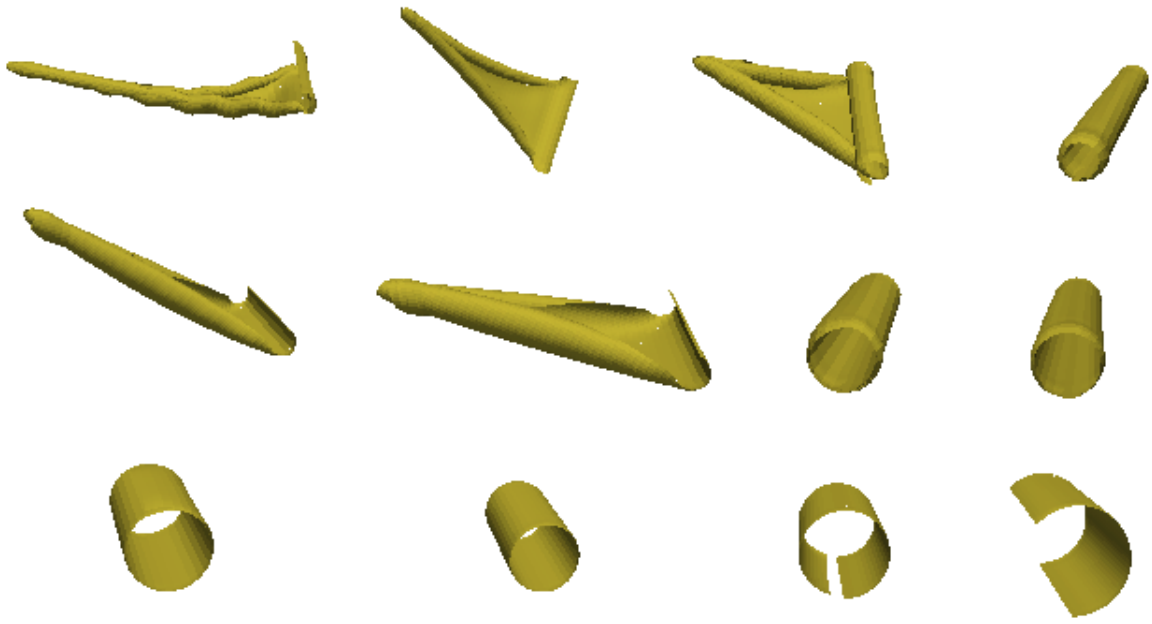


Figure 5.8: Final configuration after $80,000$ gradient flow steps of a plate $[-2a, 2a] \times [-2, 2]$ where $a = 2.5, 2, 1.5, 1$ (left to right) and $Z = -rI$ with $r = 5, 3, 1$ (top to bottom). We observe that in many cases the plate is able to achieve the cylindrical configuration, as expected from the theory in [73]. However, there are also instances of a dog-ear effect, similarly to what was observed in [11].

Using Figure 5.8 as our basis, we discuss now the effect of the numerical parameters used for our method: the time step Δt that determines the speed of the gradient flow, the tolerance level for the conjugate gradient solver for the Schur complement and the choice of the penalty constants γ_0 and γ_1. This discussion is motivated by some discrepancy in the behavior of the method with respect to these parameters, as well as the comparison of the method with the results presented in [11]. We split our exposition into two main aspects of the computations:

1. **Breaking of symmetry:** As we mentioned before, the illustrations of Figure 5.8 do not correspond to the same numerical parameters. To be more precise,

we originally experimented with a fixed time step $\Delta t = 5e-3$ and a tolerance of $1e-8$ for the iterative solver. This resulted into a breaking of symmetry in half of the tested cases: the plate deforms without creating any strong dog-ear effect and ends up crossing itself in an effort to assume a cylindrical configuration. Right before or shortly after this self-crossing occurs, the plate tilts towards one side and after a state of self-crossing, it assumes a shape similar to the clamped anisotropic case of Section 5.2.3. This breaking of symmetry was also observed in [11] and can be the result of either an error due to the tolerance of the solver that has been magnified by a big condition number or due to the time step not being small enough in order to capture the physics of the problem and ensure a small isometry defect. As a matter of fact, changing these two numerical parameters resolved the breaking of symmetry for some of the simulations. However, it was observed that the choice of penalty parameters also contributes to this factor. We explain this more at the next point.

2. **Local minimizers:** A limitation of the theory developed both in Chapter 4 and in [8, 11] is that we cannot guarantee that the gradient flow will converge to a global instead of a local minimizer. In fact, it was observed in [11] that for certain choices of aspect ratio and curvature the gradient flow lead to dog ear effects and the plate was not always able to assume the expected cylindrical configuration. This was attributed to potentially finding a local rather than a global minimizer of E_h. In contrast to [8, 11] where the energy functional

is fixed, we have two very crucial parameters, γ_0 and γ_1 that guarantee the coercivity of our method. It is conceivable that the choice of these parameters can affect the discrete energy in terms of the existence and location of local minimizers. We observe computationally that changing these parameters can on one hand resolve the issue of the breaking of symmetry and on the other hand, it can lead to eliminating the dog ear effect in some cases. However, having a combination of time step, solver tolerance and these two parameters makes it a rigorous investigation of the effect of each of these factors a rather complicated and time consuming task. Moreover, such an analysis can only be done within each problem and what may be found to be an optimal parameter for one problem may (and will probably) not be optimal for a different experiment. As a result, we do not present specific illustrations for different parameters, but rather acknowledge the effect they may have into converging to a local or global minimizer of E_h.

Bibliography

[1] N. E. Aguilera, P. Morin, *On convex functions and the finite element method*, SIAM J. Numer. Anal., 47(4):3139-3157, 2009.

[2] S. Alben, B. Balakrisnan, E. Smela, *Edge effects determine the direction of bilayer bending*, Nano Letters 11, 6, 2280-2285, 2011.

[3] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, D. Wells, *The* `deal.II` *Library, Version 8.5*, Journal of Numerical Mathematics, 25(3):137–146, 2017.

[4] G. Awanou, *Convergence rate of a stable, monotone and consistent scheme for the Monge-Ampère equation*, Symmetry, 8 (18):1-7, 2016.

[5] G. Awanou, *Standard finite element for the numerical resolution of the elliptic Monge-Ampère equation: Aleksandrov solutions*, arXiv:1310.4568v5.

[6] W. Bangerth, R. Hartmann, G. Kanschat, *deal.II – a General Purpose Object Oriented Finite Element Library*, ACM Trans. Math. Softw., 33(4):24/1–24/27, 2017.

[7] G. Barles, P. Souganidis, *Convergence of approximation schemes for linear second order equations*, Asymptotic Anal., 4(3):271-283, 1991.

[8] S. Bartels, *Finite element approximation of large bending isometries*, Numer. Math. 124, 3, 415-440, 2013.

[9] S. Bartels, *Numerical Methods for Nonlinear Partial Differential Equations*, Springer International Publishing, 2015.

[10] S. Bartels, A. Bonito, A. H. Muliana, R. H. Nochetto, *Modeling and simulation of thermally actuated bilayer plates*, Journal of Computational Physics, (354), 512-528, 2018.

[11] S. Bartels, A. Bonito, R. H. Nochetto, *Bilayer Plates: Model Reduction, Γ-Convergent Finite Element Approximation and Discrete Gradient Flow*, Commun. Pure Appl. Math., 2015.

[12] N. Bassik, B. Abebe, K. Laflin, and D. Gracias, *Photolithographically patterned smart hydrogel based bilayer actuators*, Polymer 51, 6093-6098, 2010.

[13] J-D. Benàmou, F. Collino, J-M. Mirebeau, *Monotone and consistent discretization of the Monge-Ampère operator*, arXiv:1409.6694, 2014.

[14] J-D. Benàmou, B. Froese, A. Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, ESAIM: M2AN, Volume 44 (2010), Number 4, pp 737-758.

[15] D. P. Bertsekas, A. Nedic, A. E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.

[16] Z. Blocki, *Smooth exhaustion functions in convex domains*, Proc. Amer. Math. Soc., 125(2), 477-484, 1997.

[17] O. Bokanowski, M. Falcone, S. Sahu, *An efficient filtered scheme for some first order time-dependent Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 38(1):171-195, 2016.

[18] A. Bonito, R. H. Nochetto, *Quasi-Optimal Convergence Rate of an Adaptive Discontinuous Galerkin Method*, SIAM J. Numer. Anal., 48(2), 734-771, 2010.

[19] A. Bonito, R. H. Nochetto, D. Ntogkas, *DG approach to large bending deformations with isometry constraint*, (in preparation), 2018.

[20] S. C. Brenner, K. Wang, J. Zhao,*Poincaré-Friedrichs Inequalities for Piecewise H^2 Functions*, Numer. Funct. Anal. Optim., 25, 463-478, 2004.

[21] S. C. Brenner, R. Scott,*The Mathematical Theory of Finite Element Methods*, Springer, 2008.

[22] S. C. Brenner, T. Gudi, M. Neilan, L-Y. Sung.,*C^0 penalty methods for the fully nonlinear Monge-Ampère equation*, Math. Comp., 80(276):1979-1995, 2011.

[23] A. Buffa, C. Ortner,*Compact embeddings of broken Sobolev spaces and applications*, IMA Numer. Anal., 29, 827-855, 2009.

[24] L.A. Caffarelli, L. Nirenberg, J. Spruck,*The Dirichlet problem for nonlinear second-order elliptic equations I. Monge-Ampère equation*, Comm. Pure Appl. Math., 37 (3): 369-402, 1984.

[25] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Soc. for Industrial and Applied Math., Philadelphia, PA, USA, 2002.

[26] X. Chen, Z. Nashed, L. Qi, *Smoothening methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal. 38 (4):1200-1216, 2000.

[27] M.G. Crandall, H. Ishii, P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27(1):1-67, 1992.

[28] M. Cullen, R. J. Douglas, *Applications of the Monge-Ampre equation and Monge transport problem to meteorology and oceanography*, Contemp. Math., 1998.

[29] E. J. Dean, R. Glowinski, *An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions*, Electron. Trans. Numer. Anal., 22: 71-96, 2006.

[30] E. J. Dean, R. Glowinski, *On the numerical solution of the elliptic Monge-Ampère equation in dimension two: A least-squares approach*, Partial Differential Equations, Comput. Meths. Appl. Sci. 16, Springer, Dordrecht, 43-63, 2008.

[31] G. DePhilippis, A. Figalli, *Second order stability for the Monge-Ampère equation and strong Sobolev convergence of optimal transport maps*, Anal. PDE 6 (4):993-1000, 2013.

[32] G. DePhilippis, A. Figalli, *The Monge-Ampère equation and its link to optimal transportation*, Bull. Amer. Math. Soc. 51 (4):542-580, 2014.

[33] D. A. Di Pietro, A. Ern, *Discrete Functional Analysis Tools for Discontinuous Galerkin Methods with Application to the Incompressible Navier-Stokes Equations*, Math. Comp., 79(271), 1303-1330, 2010.

[34] R. Engelking, *General Topology*, Helderman, Berlin, 1989

[35] A. Ern, J-L Guermond, *Finite element quasi-interpolation and best approximation*, ESAIM: M2AN 51, 1367-1385, 2017.

[36] Z. Feng, B. Froese, R. Liang *Composite method for precise freeform optical beam shaping*, Appl. Opt. 54, 9364-9369, 2015.

[37] X. Feng, R. Glowinski, M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, SIAM Rev., 55(2):205-267, 2013.

[38] X. Feng, M. Jensen, *Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids*, arXiv:1602.04758v2, Sep 2016.

[39] X. Feng, M. Neilan, *Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method*, SIAM J. Numer. Anal., 47(2), 1226-1250, 2009.

[40] X. Feng, M. Neilan, *Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations*, J. Sci. Comput., 38 (1):74-98, 2009.

[41] G. Friesecke, R. D. James, S. Müller *A theorem on geometric rigidity and the derivation of nonlinear plate theory from three-dimensional elasticity*, Comm. Pure Appl. Math. 55, 11, 1461-1506, 2002.

[42] B. Froese, *A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions*, SIAM J. Sci. Comput., 34(3):A1432-A1459, 2012.

[43] B. Froese, A. Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher*, SIAM J. Numer. Anal., 49(4):1692-1714, 2012.

[44] B. Froese and A. Oberman, *Convergent filtered schemes for the Monge-Ampère partial differential equation*, SIAM J. Numer. Anal., 51(1), 423444, 2013.

[45] E. Giusti, 2003, *Direct Methods in the Calculus of Variation*, Singapore: World Scientific.

[46] R. Glowinski, *Numerical methods for fully nonlinear elliptic equations*, Proceedings of the 6th International Congress on Industrial and Applied Mathematics, R. Jeltsch and G. Wanner, eds., ICIAM 07, Invited Lectures, 155–192, 2009.

[47] P. Guan, X.-J. Wang *On a Monge-Ampère equation arising in geometric optics*, J. Differential Geom. 48 (2), 205-223, 1998.

[48] C. Gutiérrez, *The Monge-Ampère Equation*, Birkhäuser, 2001.

[49] M. Hintermüller, K. Ito, K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13(3):865-888, 2003.

[50] P. Hornung, M. Lewicka, M. Reza Pakzad, *Infinitesimal isometries on developable surfaces and asymptotic theories for thin developable shells* , arXiv:1210.0047, 2012.

[51] P. Hornung, *Approximation of flat $W^{2,2}$ isometric immersions by smooth ones*, Arch. Ration. Mech. Anal. 199, 1015- 1067, 2011.

[52] H. Ishii, P. L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Diff. Eqs. 83 (1):26-78, 1990.

[53] S. Janbaz, R. Hedayati, A. A. Zadpoor, *Programming the shape-shifting of flat soft matter: from self-rolling/self-twisting materials to self-folding origami*, Mater. Horiz., 3, 536-547, 2016.

[54] E. Jager, E. Smela, O. Inganäs, *Microfabricating conjugated polymer actuators*, Science 290, 1540-1545, 2000.

[55] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Springer Netherlands, 1987.

[56] J.-N. Kuo, G.-B. Lee, W.-F. Pan, H.-L. Lee, *Shape and thermal effects of metal films on stress-induced bending of micromachined bilayer cantilever*, Japanese Journal of Applied Physics 44, 5R, 3180, 2005.

[57] K. Bhattacharya, M. Lewicka, M. Schaffner, *Plates with incompatible prestrain*, arXiv:1401.1609, 2014.

[58] M. Lewicka, P. Ochoa,M. R. Pakzad. *Variational models for prestrained plates with Monge-Ampère constraint*, Differential Integral Equations, 28, 861–898, 2015.

[59] W. Li, R. H. Nochetto *Optimal pointwise error estimates for two-scale methods for the Monge-Ampère equation*, 2018, (submitted).

[60] P. L. Lions, *Hamilton-Jacobi-Bellman equations and the optimal control of stochastic systems*, Proc. Int. Congress of Math., Warsaw, 1983.

[61] J-M. Mirebeau, *Discretization of the 3D Monge-Ampère operator, between wide stencils and power diagrams*, arXiv:1503.00947, 2014.

[62] R. H. Nochetto, D. Ntogkas *Convergent two-scale filtered scheme for the Monge-Ampère equation* , 2018 (in preparation).

[63] R. H. Nochetto, D. Ntogkas, W. Zhang *Two-scale method for the Monge-Ampère equation: convergence to the viscosity solution* , Math. Comp. (accepted), 2018.

[64] R. H. Nochetto, D. Ntogkas, W. Zhang *Two-scale method for the Monge-Ampère equation: pointwise error estimates*, 2018 (submitted).

[65] R. H. Nochetto, W. Zhang, *Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form*, Found. Comp. Math. (online), 2017.

[66] R. H. Nochetto, W. Zhang, *Pointwise rates of convergence for the Oliker-Prussner method for the Monge-Ampère equation*, arXiv:1611.02786.

[67] A. Oberman, *Convergent difference schemes for nonlinear elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problems*, SIAM J. Numer. Anal, 44 (2):879-895, 2006.

[68] A. Oberman, *The convex envelope is the solution of a nonlinear obstacle problem.* Proc. Amer. Math. Soc. 135:1689-1694, 2007.

[69] A. Oberman, T. Salvador, *Filtered schemes for Hamilton-Jacobi equations: A simple construction of convergent accurate difference schemes*, J. Comput. Phys., 284, pp. 367388, 2015.

[70] V. I. Oliker, L. D. Prussner, *On the numerical solution of the equation* $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - (\partial^2 z/\partial x \partial y)^2 = f$ *and its discretizations*, I. Numer. Math., 54(3):271-293, 1988.

[71] B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*, Society for Industrial and Applied Mathematics, 2008.

[72] E. T. Sawyer, R. L. Wheeden *Regularity of Degenerate MongeAmpre and Prescribed Gaussian Curvature Equations in Two Dimensions*, Potential Anal, 24: 267, 2006.

[73] B. Schmidt, *Minimal energy configurations of strained multi-layers*, Calc. Var. Partial Differential Equations, 30:4,477-497, 2007.

[74] O. Schmidt, K. Eberl, *Thin solid films roll up into nanotubes*, Nature 410, 168, 2001.

[75] B. Simpson, G. Nunnery, R. Tannenbaum, K. Kalaitzidou, *Capture/release ability of thermo-responsive polymer particles*, J. Mater. Chem., 20, 3496-3501, 2010.

[76] E. Smela, O. Inganös, I. Lundström, *Controlled folding of micrometer-size structures*, Science 268, 5218, 17351738, 1995.

[77] N. S. Trudinger, X. J. Wang, *Boundary regularity for the Monge-Ampère and affine maximal surface equations*, Ann. of Math. (2) 167:993-1028, 2008.

[78] G. Wachsmuth, *Conforming approximation of convex functions with the finite element method*, Numer. Math., 1-32, 2017.

[79] S. W. Walker, *FELICITY: A Matlab/C++ Toolbox for Developing Finite Element Methods and Simulation Modeling*, submitted.

[80] S. W. Walker, *FELICITY: Finite ELement Implementation and Computational Interface Tool for You.* http://www.mathworks.com/matlabcentral/fileexchange/31141-felicity.