Finite element methods for Monge–Ampère type equations

Ellya Kawecki

The Queen's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity term 2018

To Eloise

Acknowledgements

First and foremost, a great deal of credit is due to my supervisor Endre Süli, to whom I owe so much for his time and attention.

I would also like to thank Florian Wechsung, and the rest of the Firedrake and Fenics community. Without them, I would surely still be finishing this thesis in many years to come.

Furthermore, to Omar Lakkis and Tristan Pryer, who helped me along my first steps into mathematical research.

Finally to the lovely friends I have made during my time in Oxford, and the vast amount of college lunches we have enjoyed.

Abstract

This thesis focuses on the numerical analysis of partial differential equations (PDEs), the main goal being the development and analysis of finite element methods (FEMs) for fully nonlinear elliptic PDEs, particularly Monge–Ampère (MA) and Hamilton–Jacobi–Bellman (HJB) equations.

There are two clear distinctions in the approaches that are undertaken in this thesis: firstly, for the approximation of solutions to the MA problem, we implement and analyse a continuous Galerkin (CG) FEM; secondly, to numerically solve the HJB equation, we employ a discontinuous Galerkin (DG) FEM.

Though the chosen approaches (CG vs. DG) applied to the MA and HJB type equations are distinct, the equations themselves are related. A long-standing result, proven by N. Krylov in 1987, allows one to characterise the MA equation as a HJB equation.

Another important theme of this thesis, motivated by domain assumptions, necessary for the well-posedness of MA type problems, and oblique boundary-value problems is the implementation and analysis of FEMs on domains with curved boundaries. In the case of DG methods, where the consistency of the method plays a key role in obtaining a priori error estimates for the numerical solution, this quantitative consideration requires new techniques to extend the existing DG framework.

The main contributions of this thesis are new results concerning the existence and uniqueness of numerical solutions to CG and DG finite element methods on curved domains, for both fully nonlinear elliptic equations, and linear elliptic equations in nondivergence form, with Dirichlet and oblique derivative boundary conditions, as well as optimal a priori error estimates. Furthermore, we prove several key results from the theory of finite elements in the context of curved finite elements, that do not appear to be available in the literature.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

This thesis focuses on the numerical analysis of elliptic partial differential equations (PDEs), the main goal being the development and analysis of finite element methods (FEMs) for fully nonlinear elliptic PDEs. In particular Monge–Ampère (MA) and Hamilton–Jacobi–Bellman (HJB) equations. The investigation into these types of equations is motivated by applications in differential geometry, optics, meteorology, engineering, finance, and problems of optimal control.

There are two clear distinctions in the approaches that are undertaken here: firstly, for the approximation of solutions to the MA problem, we implement and analyse the continuous Galerkin (CG) FEM, introduced by Lakkis & Pryer [78, 79], which is called the nonvariational finite element method (NVFEM), named after its applicability to elliptic problems in nonvariational (also known as nondivergence) form. That is, PDEs that do not, in general, admit a weak formulation, due to the coefficients lacking sufficient regularity; in general, the coefficients are only assumed to bounded and measurable. Indeed, as the MA problem is fully nonlinear, we employ a well known linearisation technique, called Newton's method, in order to compute finite element approximations of the solution. The use of this technique results in a sequence of nondivergence form elliptic equations, justifying the use of the NVFEM. Secondly, to numerically solve the HJB equation, we employ the discontinuous Galerkin (DG) FEM, introduced by Smears & Süli [111]. This method was also originally introduced for the approximation of nondivergence form elliptic equations. In the case of the HJB equation, the corresponding equation is not only fully nonlinear (as in the case of the MA equation), but the corresponding operator defining the equation is in general not differentiable. Classical Newton's method relies on the differentiability of the operator, and thus cannot be used in this case. However, the so called "semismooth

Newton's method", which only requires the subdifferential of the operator to be defined, can be applied. Akin to the approach for the MA problem, this linearisation technique also results in a sequence of nondivergence form elliptic equations.

Though the chosen approaches (CG vs DG) applied to the MA and HJB type equations are distinct, the equations themselves are related. A longstanding result, proven by N. Krylov in 1987, allows one to characterise the MA equation as a HJB equation. The resulting HJB equation does not fall into the framework originally introduced by Smears & Süli [111], and so our approach requires the application of existing classical regularity estimates for problems of MA type, in order to prove that there is an equivalent representation of this HJB equation, for which the use of the method of Smears & Süli is justified.

It is informative, at this point, to define the underlying PDEs. Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domian. A function $u : \Omega \to \mathbb{R}$ is said to be a solution to the MA equation if

$$\det D^2 u = f(x, u, \nabla u), \tag{1.1.1}$$

for all $x \in \Omega$, where $D^2 u$ and ∇u denote the Hessian and the gradient of u respectively, and $f : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ is a given positive function. A function $u : \Omega \to \mathbb{R}$ is said to be a solution to the HJB equation if

$$\sup_{\alpha \in \Lambda} \{A^\alpha : D^2 u - f^\alpha\} = 0, \tag{1.1.2}$$

where the set Λ, and the collection $\{A^\alpha, f^\alpha\}_{\alpha \in \Lambda}$ is given (see Chapter 3 for further assumptions on the set Λ, and the corresponding collection of functions).

The first boundary condition we consider in conjunction with (1.1.1), and (1.1.2), is the *Dirichlet* boundary condition (where the value of the solution is prescribed on the boundary, $\partial\Omega$ of the domain, Ω).

The second boundary condition, which we couple with (1.1.1), arises in the area of optimal mass transportation. Ironically, this boundary condition is called the "*second* boundary condition", and takes the following form: given a second domain $\Upsilon \subset \mathbb{R}^d$, a function $u : \Omega \to \mathbb{R}^d$ that satisfies (1.1.1) is said to satisfy the second boundary condition if

$$\nabla u(\Omega) = \Upsilon. \tag{1.1.3}$$

Of course one may argue that (1.1.3) is not in fact a boundary condition, since the values of the gradient are prescribed in the *interior* of Ω, rather than on the boundary. In contrary to this observation, it is proven in [104] for a uniformly convex function,

and simply connected domains Ω and Υ, that u satisfies (1.1.1) and (1.1.3) if and only if u satisfies (1.1.1) and

$$\nabla u(\partial\Omega) = \partial\Upsilon. \tag{1.1.4}$$

Though the equivalence is proven based on the weaker assumption that Ω and Υ are simply connected domains, the existence and uniqueness results proven by Urbas in [116] require that both Ω and Υ are uniformly convex (see Chapter 2 for a definition of uniform convexity); note that this implies that the domain is Lipschitz continuous [60]. As such, in terms of PDE theory for the MA equation (in particular, for the results present in Chapter 3), this will often be our assumption, unless stated otherwise.

In the case of optimal mass transportation, matter is being transferred from Ω to Υ (consider for instance a pile of sand, being transported into a hole of some fixed distance from the pile of sand). The density of the matter, and the transported matter are described by given functions $f_1 : \Omega \to \mathbb{R}^+$, and $f_2 : \Upsilon \to \mathbb{R}^+$, respectively (for the same example, one would consider f_1 to model the "height" of the sand pile, and f_2 the "capacity" (or depth) of the hole, translated so that it is uniformly positive, see Figure 1.1). The functions must satisfy

$$\int_\Omega f_1 = \int_\Upsilon f_2, \tag{1.1.5}$$

which represents the preservation of total mass. The objective of the optimal mass transportation problem is to find the map $T : \Omega \to \Upsilon$ that minimises the *quadratic cost functional*

$$C[\mu] = \frac{1}{2}\int_\Omega |x - \mu(x)|^2 f_1, \tag{1.1.6}$$

over all maps $\mu : \Omega \to \Upsilon$ that satisfy the *push-forward* condition:

$$\int_\Omega (\psi \circ \mu) f_1 = \int_\Upsilon \psi f_2, \tag{1.1.7}$$

for all $\psi \in L^1(\Upsilon)$. If μ satisfies (1.1.7), we denote this by $\mu \# f_1 = f_2$, and refer to f_2 as the *push-forward* of f_1 by μ. The optimal map T, is the map satisfying

$$C[T] = \inf_{\mu \# f_1 = f_2} \frac{1}{2}\int_\Omega |x - \mu(x)|^2 f_1. \tag{1.1.8}$$

More information about the optimal mass transportation problem with more general cost functions can be found in [118]. An important result in [118] proven by Y. Brenier states that the map T is given uniquely by the gradient of a convex function u. We

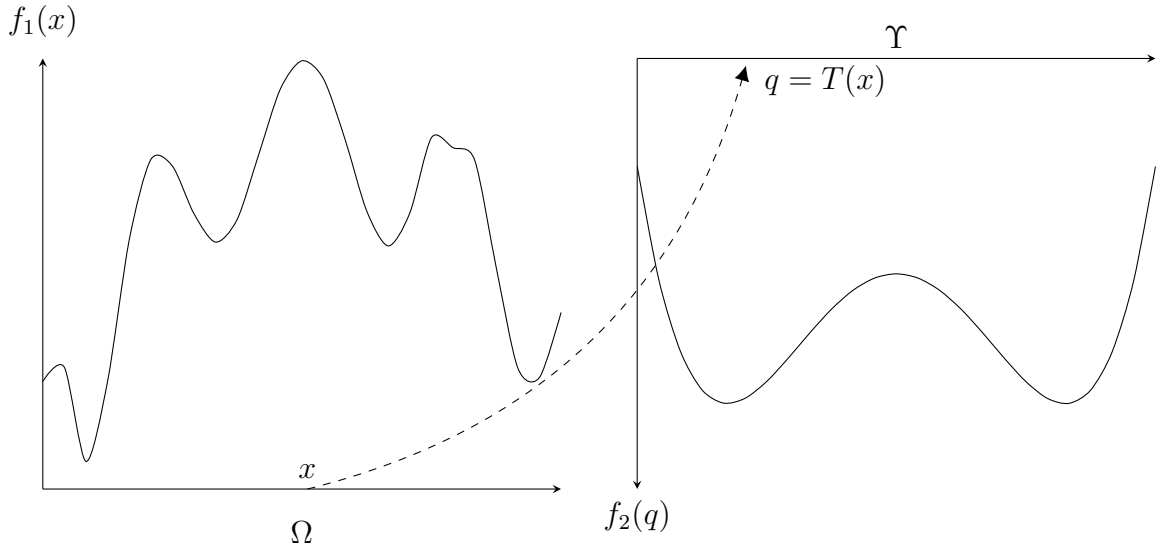


Figure 1.1: Visualisation of an optimal transport problem. The density $f_1 : \Omega \to \mathbb{R}^+$ represents the height of the pile of sand, and the density $f_2 : \Upsilon \to \mathbb{R}^+$ represents the capacity (or depth) of the hole that the sand must be transported into. The objective is to find the map $T : \Omega \to \Upsilon$ that minimises C given by (1.1.6), i.e., T satisfies (1.1.8).

can see that by substituting $T = \nabla u$ in (1.1.7) and applying the change of variables formula, we obtain

$$\int_\Omega (\psi \circ \nabla u) f_1 = \int_\Omega (\psi \circ \nabla u) \det D^2 u f_2(\nabla u),$$

for all $\psi \in L^1(\Upsilon)$, from which we conclude that

$$\det D^2 u - \frac{f_1(x)}{f_2(\nabla u)} = 0. \tag{1.1.9}$$

Notice that this is in fact (1.1.1) with

$$f(x, u, \nabla u) := \frac{f_1(x)}{f_2(\nabla u)}.$$

One form of the MA equation that arises naturally in the context of differential geometry is the problem of prescribed Gaussian curvature. In d-dimensions, this equation is posed as follows: to find $u : \Omega \to \mathbb{R}$ such that

$$\det D^2 u = K(x)(1 + |\nabla u|^2)^{(d+2)/2},$$

where the function K is given. This equation arises due to the fact that the quantity

$$\frac{\det D^2 u}{(1 + |\nabla u|^2)^{(d+2)/2}}$$

is in fact the Gaussian curvature of the graph

$$Z := \{(x, u(x)) : x \in \Omega\}.$$

Thus, reverse engineering this information, given a function $K : \Omega \to \mathbb{R}$, and a function $\phi : \partial\Omega \to \mathbb{R}$, one may ask the question: does there exist a function $u : \Omega \to \mathbb{R}$ that satisfies $u|_{\partial\Omega} = \phi$, for which the Gaussian curvature of the graph Z is equal to K? In doing so, we arrive at the following equation: find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} \det D^2 u(x) &= K(x)(1 + |\nabla u(x)|^2)^{(d+2)/2}, \quad x \in \Omega, \\ u(x) &= \phi(x), \quad x \in \partial\Omega. \end{aligned}$$

It turns out that the problem of prescribed Gaussian curvature also arises in engineering in the field of optics, in the design of freeform reflectors. Given a light source, the goal is to provide a reflective surface that provides a desired light output. Particular examples of illumination problems are the design of car headlights and lampposts. It transpires that such a problem can be modelled as a MA optimal transport problem, with the type of nonlinearity that is present in the problem of prescribed Gaussian curvature:

$$\begin{aligned} \det D^2 u(x) &= K(x)(1 + |\nabla u(x)|^2)^{(d+2)/2}, \quad x \in \Omega, \\ \nabla u(\Omega) &= \Upsilon, \end{aligned}$$

where the domain Ω is represents the light source, and the domain Υ is the area to be illuminated, and the graph of the solution u provides the desired reflective surface. This equation arises due to the law of reflection and conservation of luminous flux, and a corresponding derivation can be found in [104].

A further application of MA optimal transport that has applications in the numerical approximation of PDEs is the area of mesh adaptivity. In such a case, one is provided with a fixed computational domain $\Omega_c \subset \mathbb{R}^d$ and a physical domain $\Upsilon_p \subset \mathbb{R}^d$ where goal is to find a map $T : \Omega_c \to \Upsilon_p$ that satisfies

$$\frac{1}{2}\int_{\Omega_c} |x - T(x)|^2 = \inf_{\mu \# \sigma = \rho} \frac{1}{2}\int_{\Omega_c} |x - \mu(x)|^2,$$

where $\rho > 0$ is a scalar density function, and

$$\sigma = \frac{1}{\Omega_c}\int_{\Upsilon_p} \rho,$$

so that mass is conserved. One then arrives at the MA optimal transport problem:

$$\begin{aligned} \det D^2 u(x) &= \frac{\sigma}{\rho(\nabla u(x))}, \quad x \in \Omega_c, \\ \nabla u(\Omega_c) &= \Upsilon_p. \end{aligned}$$

Upon providing a numerical approximation to the solution of this problem, one obtains an adapted mesh by taking the image of a mesh of Ω_c under the gradient (or a continuous representative of the gradient), ∇u_h, of the numerical solution, u_h (for further details see [26]).

There are several existence and uniqueness results that have been proven over the years for both the MA Dirichlet (MAD) and the MA optimal transport (MAOT) problem. In the case of the MAD problem, L. Caffarelli, L. Nirenberg and J. Spruck [29], N. Krylov [73] and Ivochkina [65] proved the existence, uniqueness and regularity of a uniformly convex solution to (1.1.1), based on the assumption that Ω is uniformly convex, and that Ω, f, and the boundary data, are sufficiently smooth, and that the function f satisfies sufficient growth conditions. The Monge–Ampère equation with the second boundary condition (of which the MAOT problem is a subclass) has also been well analysed. In [116] J. Urbas uses the method of continuity to prove existence, uniqueness and regularity of a solution to the general MA equation (1.1.1), as well as the MAOT problem (1.1.9) with the second boundary condition (1.1.3). This paper consisted of proving necessary estimates to apply the techniques in the author's accompanying paper [117]. In 1997 (close to the time of the publication of [116]), J. Urbas learnt of L. Caffarelli's paper [27], in which L. Caffarelli proves similar results with assumptions weaker than that of the material of J. Urbas' paper, with regards to the smoothness of the domains, with different techniques. These results however were given in the context of (1.1.9), and are not as general as the results given in J. Urbas' papers, which refer to the case (1.1.1), i.e., the source term is more general.

When considering the two-dimensional case ($d = 2$) more is known. In [99] A. V. Pogorelov proved existence and uniqueness of generalised solutions in the sense of Aleksandrov in the case that Ω and Υ are assumed to be bounded and convex. This is a useful result, since in some of our experiments we will consider the case where $\Omega = (-1, 1)^2$. It also reinforces some of the experiments found in [103], where the target domain is not convex. In this case the authors apply the Legendre–Fenchel transform, resulting in an inverse problem in which the source domain is not uniformly convex. In [42] P. Delanoë proved existence and uniqueness of a solution, as well as global regularity and a priori higher regularity results; in this paper the author comments on the difficulty of generalising the result to higher dimensions.

Just as the MA equation poses difficulties analytically due to the nonlinear nature of the problem, it also poses difficulty when formulating and analysing numerical methods for the approximation of solutions. That said, the difficulty has not deterred the advances over the years in both finite difference and finite element approximation of solutions. J-D. Benamou, B. Froese, and A. Oberman [14, 15, 54, 95], have successfully developed and analysed convergent finite difference schemes for the solution of the general MA equation with a Dirichlet boundary condition and the MA problem of optimal mass transport.

One of the computational difficulties of the MAOT equation - setting aside the nonlinearity of the PDE - is the *second* boundary condition; this is a case where we can see a motivation for the design of an algorithm that follows the PDE theory. Since the definition of the second boundary condition (1.1.3) is not compatible with computations, one must use the representation (1.1.4). This however does not close the argument, since the boundary condition is still nonlinear and only given implicitly. In [116] it was proven that in fact the second boundary condition is strictly oblique, that is, that the outer normal derivative of a uniformly convex (equivalently, the inner normal derivative of a uniformly concave) defining function for Υ, applied to the solution at the boundary, is positive.

To do so the author of [116] considers a uniformly concave defining function $a : \mathbb{R}^d \to \mathbb{R}$ for the target domain, that is,

$$\Upsilon = \{q \in \mathbb{R}^d : a(q) > 0\}.$$

Since one can see that $\partial\Upsilon = \{q \in \mathbb{R}^d : a(q) = 0\}$, the boundary condition can be expressed as follows

$$a(\nabla u(x)) = 0, \quad x \in \partial\Omega.$$

Alternatively, one can consider a convex defining function $b : \mathbb{R}^d \to \mathbb{R}$, for the target domain, i.e., $\Upsilon = \{q \in \mathbb{R}^d : b(q) < 0\}$. One such example is the signed distance function, i.e.,

$$b(q) = \begin{cases} -\operatorname{dist}(q, \partial\Upsilon), & q \in \Upsilon, \\ \operatorname{dist}(q, \partial\Upsilon), & q \in \mathbb{R}^d \setminus \Upsilon. \end{cases}$$

The boundary condition can then be thought of as a Hamilton–Jacobi equation. It is this formulation that J.D. Benamou, B. D. Froese, and A. M Oberman use in [15], to implement the boundary condition. It is similar to the method that we use in Chapter 9 to implement the boundary condition in the CG case. In this case, we apply Newton's method to $b(\nabla u(x)) = 0$, $x \in \partial\Omega$, resulting in a sequence of oblique boundary conditions.

There have also been many advances over the years in the implementation and analysis of finite element methods for the MA equation, for instance, in S. Brenner and M. Neilan [22], and M. Neilan [92], convergence results were proven in dimensions three and two respectively. Chapter 8 of this thesis extends the analysis found in [92] from the case that the source term $f(x, u, \nabla u) := f(x)$, to $f(x, u, \nabla u) := \frac{f_1(x,u)}{f_2(\nabla u)}$. For other two and three-dimensional cases, one should look at the work of G. Awanou and H. Li [3, 4], where the authors also consider the notion of the finite element Hessian; by recasting this formulation as a mixed method, they also make use of the divergence-free nature of the Cofactor matrix, $\mathrm{Cof}(D^2u) := \det(D^2u)(D^2u)^{-1}$, to analyse the linearised MA equation. Another method, called the "vanishing moment method" motivated by viscosity solutions [37] of the MA equation was pioneered by X. Feng and M. Neilan [52] in which they artificially perturb the nonlinear equation by a linear fourth-order term, resulting in a sequence of quasilinear fourth order equations. The authors also produced a further paper [51], applying the vanishing moment method to the MAD problem, i.e., they approximate solutions of

$$-\varepsilon\Delta^2 u_\varepsilon + \det(D^2 u_\varepsilon) = f(x). \tag{1.1.10}$$

In [48], X. Feng proved that u_ε, the solution of (1.1.10) (subject to suitable boundary conditions), converges to the viscosity solution of the MAD problem.

Furthermore, a recent publication [93] by M. Neilan, A. Salgado, and W. Zhang provides an excellent overview of existing numerical methods for nonlinear PDEs, including finite element methods for MA and HJB type equations.

In O. Lakkis and T. Pryer [78, 79], the authors proposed the nonvariational finite element method (NVFEM), which is used to approximate the solutions of linear second-order *nondivergence* form, uniformly elliptic equations. The NVFEM is based on the notion of a finite element Hessian or "Hessian recovery" in which the solution and the finite element Hessian of the solution are computed when approximating solutions to the PDE in *strong* form; for more details see Chapters 4, 8, and 9. This method is also applicable to nonlinear elliptic equations by first applying Newton's method to the nonlinear problem, resulting in a sequence of *nondivergence* form elliptic equations. The solutions are then approximated by applying the NVFEM.

As mentioned previously, our CG approach is to implement this method, allowing for the numerical approximation of solutions to the MAD and MAOT problem (see Chapters 8 and 9, respectively). Our experiments in the latter case suffer from suboptimal convergence rates, due to the piecewise linear approximation of curved

domains. Note that we consider curved domains, due to the uniform convexity assumption present in the existence and uniqueness result of J. Urbas [116], guaranteeing well-posedness of the underlying PDE. We introduce a novel scheme, based on applying a global gradient recovery method via L^2-projection, that produces optimal results under piecewise linear finite element approximations. This recovery method was inspired by the work of A. Naga, O. Zienkiewicz, and Z. Zhang [120, 122].

This motivates another important theme of this thesis: the implementation and analysis of FEMs on domains with curved boundaries. In the case of DG methods, where the consistency of the method plays a key role in obtaining a priori error estimates for the numerical solution, this quantitative consideration requires new techniques to extend the existing DG framework [110, 111].

In general, when triangulating any computational domain, one is required to generate a collection of maps from a reference simplex (a d-dimensional triangle of normalised volume), to the computational domain; the union of the images of the reference simplex over the collection of maps is what we consider to be the triangulation.

There are several methods for generating such a collection of maps, for instance, the most simple is to take a collection of *affine* maps (see Figure 1.2), which is certainly sufficient when the boundary has piecewise zero curvature (take the unit square for example), but if the domain has a curved boundary (e.g. the unit disk), then this affine triangulation is a poor approximation, resulting in suboptimal (quadratic at best) convergence rates for even standard conforming finite element methods [109]. Thankfully, there are more sophisticated methods. Theoretically, one (assuming sufficient piecewise smoothness of the domain) can generate a triangulation that approximates the domain *exactly* [16] (see Figure 1.3), though, in practice, this can be difficult to implement, as it requires the knowledge of the charts of the hyper-surface determined by the boundary. In order to circumvent this difficulty, one can instead take a polynomial approximation of these charts (generally by some form of interpolation), and use these to generate a collection of polynomial maps from the reference simplex to the domain. This will not, in general, result in an exact approximation of the domain. However, one can see that if the polynomial degree of the collection of maps is high enough, then optimal convergence results can be restored. Often one will choose a polynomial degree that is equal to the polynomial degree of the finite element space; this type of approximation is called "isoparametric". We implement and analyse both CG and DG FEMs with curved triangulations, in the sense of [16], in Chapters 5, 6, 7, and 8.

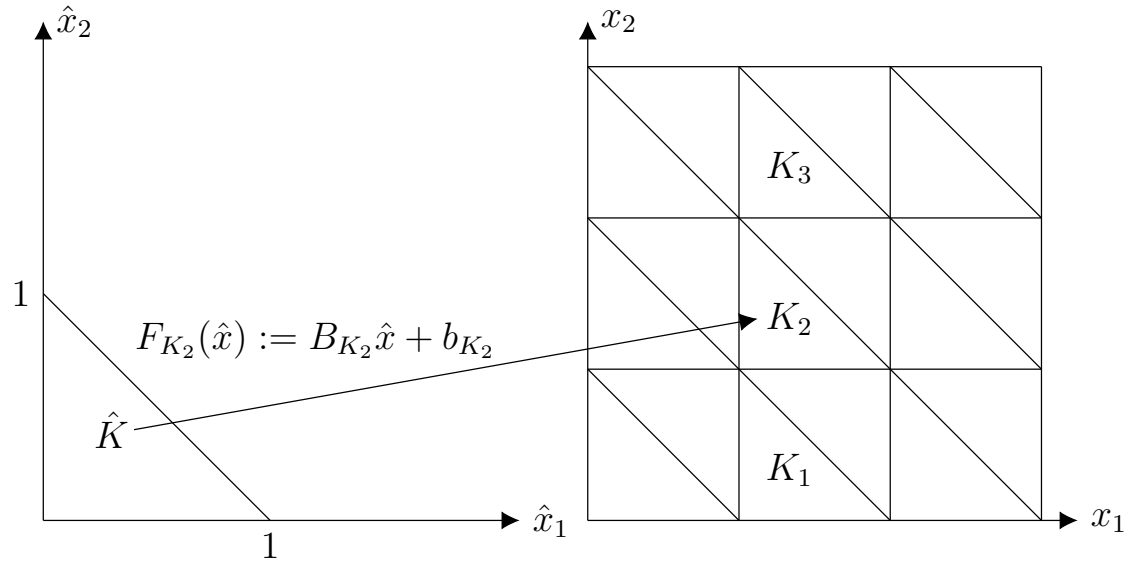


Figure 1.2: An example of the construction of an affine mesh on a square-shaped domain in $\mathbb{R}^2$, employing affine maps from a reference simplex (left) to the domain (right).

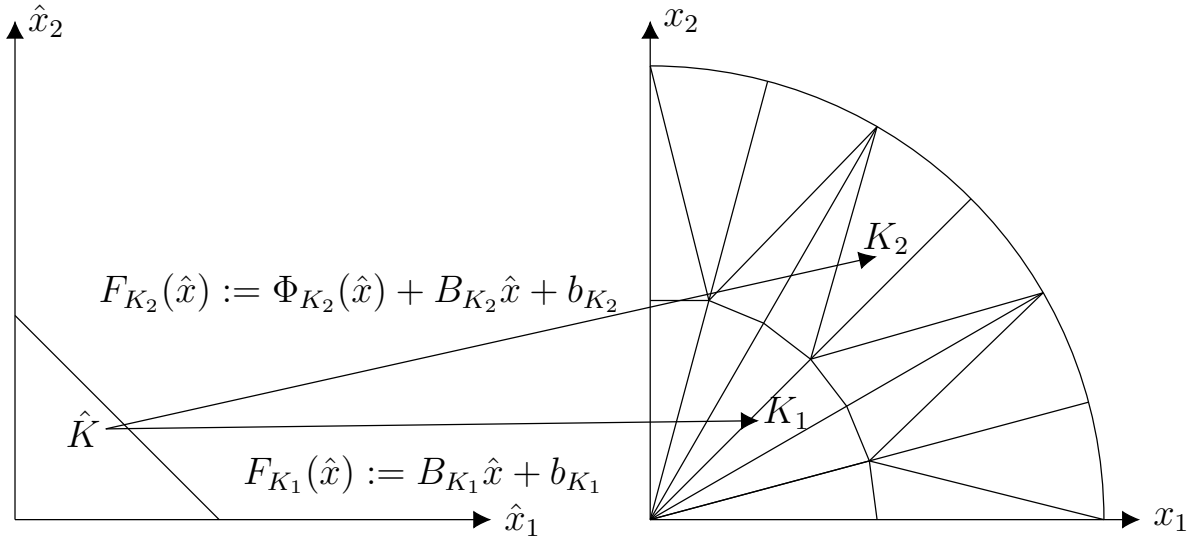


Figure 1.3: An example of the construction of a non-affine mesh on a quarter disk domain (right) in $\mathbb{R}^2$, employing affine maps from a reference simplex (left) to the (interior) triangles with straight edges (e.g. K_1), and non-affine maps from a reference simplex to the (boundary) triangles with curved edges (e.g. K_2).

We now turn to discuss the HJB equation (1.1.2). As mentioned previously, our initial motivation for considering HJB type problems comes from a longstanding result by N. Krylov (see Chapter 3), that characterises the MA equation (1.1.1) as a particular HJB equation, thus motivating the use of the DGFEM introduced by Smears & Süli in [111]. Aside from the applications to the MA equation, the HJB problem (1.1.2) arises in finance, engineering, mean-field games, and problems of optimal control [53, 77].

The problem (1.1.2) and its time dependent counterpart arise directly in the optimal control of stochastic differential equations. In particular, one considers the time evolution of a state vector $X : t \mapsto X_t$, subject to a control process $\alpha(\cdot) : t \mapsto \alpha_t \in \Lambda$, where Λ is a given control set. Here $\alpha(\cdot)$ denotes a Λ-valued function, and α denotes an element of Λ. In what follows, we assume that $(\Omega, \mathbb{F}, \mathbb{P})$ is a probability space, with a filtration $\{\mathbb{F}_t\}$, and a k-dimensional $\{\mathbb{F}_t\}$-Brownian motion $\{W_t\}$ (see [94] for further details). Furthermore, the functions $b : \mathbb{R}^d \times \Lambda \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \times \Lambda \to \mathbb{R}^{d\times k}$, $f, c : \overline{\Omega} \times \Lambda \to \mathbb{R}$, $c \geq 0$. The state vector, X, is assumed to satisfy the following stochastic differential equation for a given $x \in \Omega$:

$$dX_t = b(X_t, \alpha(t))dt + \sigma(X_t, \alpha(t))dW_t \quad \text{for } t > 0, \quad X_0 = x. \tag{1.1.11}$$

We arrive in the setting of bounded domains and Dirichlet boundary conditions by considering indefinite time horizon problems with unbounded terminal times. This means that the control problem is terminated when the state X exits the domain Ω, at which time we charge an "exit cost" (this corresponds directly to the Dirichlet boundary data). We define the $\{\mathbb{F}_t\}$-stopping time as follows

$$\tau := \inf\{s \geq 0 : X_s \notin \Omega\}, \tag{1.1.12}$$

and we set $\tau = \infty$ if $X_t \in \Omega$ for all $t \geq 0$.

Before we can define the optimal control problem, we must define the set of $\mathcal{A}$ of admissible controls (i.e., the set of control processes that we will consider in the optimal control problem). An $\{\mathbb{F}_t\}$-adapted (see [94] for further details), Λ-valued stochastic process $\alpha(\cdot)$ is called an admissible control if: there exists a unique solution Ω of the state equation (1.1.11) corresponding to $\alpha(\cdot)$, and the quantity

$$\mathbb{E}\int_0^\infty |f(X_t, \alpha(t))| e^{-\int_0^t c(X_s,\alpha(s))\,ds}\,dt < \infty,$$

where $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$. We denote by $\mathcal{A}$ the set of *all* admissible controls (note that in applications, there may be more conditions upon

the set of admissible controls). We also note that the definition of the $\{\mathbb{F}_t\}$-stopping time τ is implicitly dependent upon a given $x \in \Omega$ and $\alpha(\cdot) \in \mathcal{A}$.

We then define the cost functional

$$J^{\alpha(\cdot)}(x) := \mathbb{E}\left[\int_0^\tau f(X_t, \alpha(t)) e^{-\int_0^t c(X_t,\alpha(s))\,ds}\,dt + e^{-\int_0^\tau c(X_s,\alpha(s))\,ds}\phi(X_\tau)\mathbb{1}_{\{\tau<\infty\}}\right], \tag{1.1.13}$$

where the function ϕ represents the cost of exiting Ω, and the function $\mathbb{1}_{\{\tau<\infty\}} = 1$ if $\tau < \infty$ and 0 otherwise, and $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$.

We now consider the following stochastic optimal control problem subject to a stochastic differential equation (SDE) given by

$$\min_{\alpha(\cdot)\in\mathcal{A}} J^{\alpha(\cdot)}(x), \quad J^{\alpha(\cdot)}(x) \text{ defined by (1.1.13)}, \tag{1.1.14}$$

$$dX_t = b(X_t, \alpha(t))dt + \sigma(X_t, \alpha(t))dW_t \quad \text{for } t > 0, \quad X_0 = x. \tag{1.1.15}$$

We arrive at a HJB equation by defining the value function:

$$u(x) := -\inf_{\alpha(\cdot)\in\mathcal{A}} J^{\alpha(\cdot)}(x).$$

The value function, $u(x)$, provides the optimal value of the cost functional; in the method of dynamic programming, the value function is then used to calculate the optimal control. One can in fact show that (see Chapter 4 of [53] for further details) the function $u(x)$ satisfies the following HJB equation:

$$\begin{aligned} \sup_{\alpha\in\Lambda}\left\{\frac{1}{2}\sigma^\alpha(\sigma^\alpha)^T : D^2u + b^\alpha\cdot\nabla u - c^\alpha u - f^\alpha\right\} = 0 \quad &\text{in } \Omega, \\ u = \phi \quad &\text{on } \partial\Omega, \end{aligned} \tag{1.1.16}$$

where the functions $\{f^\alpha\}_{\alpha\in\Lambda}$ are defined via $f^\alpha : x \mapsto f(x,\alpha)$ for $\alpha \in \Lambda$ (similarly for $\sigma^\alpha, b^\alpha, c^\alpha$). Note that the diffusion coefficients in (1.1.16) arise due to the stochastic nature of the state equation (1.1.15), following an application of Ito's chain rule.

Furthermore, one obtains a time dependent HJB equation of the following form for $u : Q \to \mathbb{R}$, $Q := \Omega \times [0,T]$, for some $T \in (0,\infty)$

$$\begin{aligned} \sup_{\alpha\in\Lambda}\left\{\partial_t u - \frac{1}{2}\sigma^\alpha(\sigma^\alpha)^T : D^2u + b^\alpha\cdot\nabla u - c^\alpha u - f^\alpha\right\} = 0 \quad &\text{in } Q, \\ u|_{\{t=0\}} = \phi, \end{aligned} \tag{1.1.17}$$

by considering an indefinite time horizon problem with a bound on the terminal time. In this case, the functions b, σ, f, c, ϕ now have an extra variable of dependence, i.e., $b = b(t, X_t, \alpha(t))$, $\phi = \phi(t, X_t)$ (similarly for σ, f, c), as does the cost functional

$J^{\alpha(\cdot)} = J^{\alpha(\cdot)}(x,t)$, and value function $u = u(x,t)$ (see Chapter 4 of [53] for further details). At a time $s \in [0,T)$ the state vector now evolves for $t \in (s,T)$ according to the SDE

$$dX_t = b(t, X_t, \alpha(t))dt + \sigma(t, X_t, \alpha(t))dW_t \quad \text{for } s < t < T, \quad X_s = x.$$

Moreover, the $\{\mathbb{F}_t\}$-stopping time is defined as follows

$$\tau = \min\{T, \inf\{t \geq s : (t, X_t) \notin Q\}\},$$

i.e., we control up until the smaller of the time T, and the time that the pair (t, X_t) exits Q (technically this leads to terminal condition, and a backwards in time HJB equation, but after a change of variables, one obtains (1.1.17)).

The abstract formulation of the stochastic optimal control problem (1.1.14)–(1.1.15) encompasses many applications, particularly in finance and economics, where the state vector X may, for example, represent a portfolio or economic agent, respectively.

Let us first discuss Merton's portfolio problem with transaction charge [40]. This problem is an extension of the original Merton's portfolio problem with no transaction charge, originally introduced in [91]. The problem we consider models the portfolio of an investor consisting of two assets being traded (with the assumption that trades may be made instantaneously). The first asset is a risk-free bond, with price X_t^1, and the second is a risky asset, stock, with price X_t^2, $t \geq 0$. The investor starts off with an initial endowment $X_0^1 = x$ and $X_0^2 = y$. Then, X_t^1 and X_t^2 satisfy the following SDEs (note that the equation for X_t^1 is deterministic):

$$dX_t^1 = (rX_t^1 - c(t))dt, \quad t > 0, \quad X_0^1 = x, \tag{1.1.18}$$

$$dX_t^2 = \mu X_t^2 dt + \sigma X_t^2 dW_t, \quad t > 0, \quad X_0^2 = y. \tag{1.1.19}$$

where μ is the drift rate and σ is the volatility of the stock, r is the interest rate of the bond, $c(t)$ the rate of consumption, and W_t is a Brownian motion. At any given time, the investor must decide how much (wealth) to consume and how much to invest in stock markets, with the goal of maximising utility from terminal bond and stock. We denote by $\ell(t)$ the rate of transfer from bond to stock, and $m(t)$ the rate of transfer from stock to bond. Furthermore, the investor pays fractions $\lambda_1 \in (0,1)$ and $\lambda_2 \in (0,1)$ of the amount transacted, on purchase and sale of stock, respectively. All such charges are paid from stock. In such a scenario, (1.1.18)–(1.1.19) becomes:

$$dX_t^1 = (rX_t^1 - c(t))dt - (1+\lambda_1)\ell(t)dt + (1-\lambda_2)m(t)dt, \ t > 0, \ X_0^1 = x, \tag{1.1.20}$$

$$dX_t^2 = \mu X_t^2 dt + \sigma X_t^2 dW_t + (\ell(t) - m(t))dt, \ t > 0, \ X_0^2 = y. \tag{1.1.21}$$

We then define the solvency region:

$$\mathscr{S}_{\lambda_1,\lambda_2} := \{(x,y) \in \mathbb{R}^2 : x + (1+\lambda_1)y \geq \text{ and } x + (1-\lambda_2)y \geq 0\},$$

which is where the state vector $X_t := (X_t^1, X_t^2)^T$, representing the investor's holdings, is constrained to lie. For $x, y \in \mathscr{S}_{\lambda_1,\lambda_2}$, we define the admissible set

$$\mathcal{A}_{xy} := \{(\ell(\cdot), m(\cdot), c(\cdot)) \text{ nonnegative } \{\mathbb{F}_t\}\text{-adapted, such that the solutions } (X_t^1, X_t^2) \\ \text{of (1.1.20)–(1.1.21) satisfy } (X_t^1, X_t^2) \in \mathscr{S}_{\lambda_1,\lambda_2}\, \forall t \geq 0\}.$$

The goal is then to maximise the following utility functional

$$J^{\ell(\cdot),m(\cdot),c(\cdot))}(x,y) := \mathbb{E}\int_0^\infty e^{-\rho t} U(c(t))\, dt, \tag{1.1.22}$$

where ρ is the discount factor[1] (modelling the notion that wealth is more valuable today than it is in the future), and an example of a utility function $U(c(t))$ is given by

$$\frac{c(t)^p}{p}, \tag{1.1.23}$$

for some $p \in (-\infty, 1) \setminus \{0\}$. For each p, this utility function belong to the so-called HARA (hyperbolic absolute risk aversion) class.

The value function

$$u(x,y) := \sup_{(\ell(\cdot),m(\cdot),c(\cdot))\in\mathcal{A}_{xy}} J^{\ell(\cdot),m(\cdot),c(\cdot))}(x,y),$$

with $J^{\ell(\cdot),m(\cdot),c(\cdot))}$ defined by (1.1.22), with utility function (1.1.23), satisfies the following HJB equation:

$$\sup_{(\ell,m,c)\geq 0} \left\{ \frac{1}{2}\sigma^2 y^2 \partial_{xx}^2 u + rx\partial_x u + \mu y \partial_y u - \rho u + \ell(-(1+\lambda_1)\partial_x u + \partial_y u) \right. \\ \left. + m((1-\lambda_2)\partial_x u - \partial_y u) + \frac{c^p}{p} \right\} = 0 \quad \text{in } \mathscr{S}_{\lambda_1,\lambda_2}. \tag{1.1.24}$$

We now discuss how HJB equations arise in the context of economics, in mean field games, where an economic agent is represented by their level of consumer goods, and stock of capital. The state vector $X_t = (X_t^1, X_t^2)$ represents the level of consumer goods, X_t^1, and the stock of capital, X_t^2. The model we consider was introduced in [58], and is based on economic considerations. An agent must decide a strategy on

[1]The reader should note that a condition upon ρ, r, μ, p and σ must be satisfied in order to avoid the possibility of arbitrarily large utility, see Condition A of [40] for further details.

consumption and investment levels, $c(t)$ and $i(t)$, respectively, which provides us with the set of controls. The dynamics of the state vector is governed by the following SDEs evolving over $t \in (0,T)$ for some $T \in (0,\infty)$

$$dX_t^1 = (-c(t) - p(t)i(t) + F(X_t^2, p(t))dt + \sigma_1 dW_t^1, \quad t \in (0,T), \quad X_0^1 = x, \tag{1.1.25}$$

$$dX_t^2 = (g(X_t^2, p(t)) + i(t))dt + \sigma_2 dW_t^2, \quad t \in (0,T), \quad X_0^2 = y, \tag{1.1.26}$$

where W_t^1 and W_t^2 are adapted Brownian motions, and σ_1, σ_2 are given constants. The macroeconomic variable $p(t)$ represents the price level. The function F is of the following form

$$F(X^1, p) = F_1(X^1, p) + pF_2(X^1, p),$$

where F_1 is the amount of consumer goods and F_2 is the amount of capital goods (goods that are used to produce other goods). Thus, the production function F gives the total value (measured in consumer goods) of consumer and capital goods of an economic agent with capital X^1 at price level p. Finally, the function g represents the depreciation of capital stock, and depends upon the capital X^1 and price level p. Their strategy is governed by their desire to maximise a utility (corresponding to a function $U = U(c(t), i(t), X_t^1, X_t^2, p(t)))$, which is assumed to encompass their preferences with regards to stocks of consumer goods and capital, and the price level of the economy. The agent seeks to maximise the utility functional

$$J^{c(\cdot),i(\cdot)}(x,y,t) := \mathbb{E}^{x,y} \int_t^T e^{-\rho(s-t)} U(c(s), i(s), X_s^1, X_s^2, p(s))\, ds,$$

where $\mathbb{E}^{x,y}$ is the expectation corresponding to $X_0^1 = x, X_0^2 = y$, and ρ is the discount factor. We maximise $J^{c(\cdot),i(\cdot)}$ over a subset $\mathcal{A}$ of the set of progressively measurable controls $c(\cdot), i(\cdot)$. Defining the value function

$$u(x,y,t) := \sup_{(c(\cdot),i(\cdot)) \in \mathcal{A}} J^{c(\cdot),i(\cdot)}(x,y,t),$$

we arrive at the following (backwards) HJB equation:

$$\partial_t u + \frac{\sigma_1^2}{2}\partial_{xx} u + \frac{\sigma_2^2}{2}\partial_{yy} u + H(x,y,p,\nabla u) - \rho u = 0, \quad (x,y,t) \in \mathbb{R}^2 \times (0,T), \tag{1.1.27}$$

$$u(x,y,T) = 0, \quad (x,y) \in \mathbb{R}^2, \tag{1.1.28}$$

where

$$H(x,y,p,q) := \sup_{c,i} \{(-c - pi + F, g + i) \cdot q + U\} \tag{1.1.29}$$

(note that the variable q represents the ∇u component of H, and that the values of c, i, in the supremum in (1.1.29) depends upon the set $\mathcal{A}$ of admissible controls).

We arrive at a mean field game by considering many economic agents, each of whom maximises their individual utility, by taking into account the behaviour of the entire population. For the full derivation of the mean field game for the current model, we refer the reader to [58], and for literature on mean field games in general [31, 61, 1, 81]. Denoting by μ a probability density that represents the distribution of the entire population of agents, the mean field game equation is the following (backward) HJB equation coupled with a (forward) Fokker–Planck equation:

$$\begin{aligned}\partial_t u + \frac{\sigma_1^2}{2}\partial_{xx} u + \frac{\sigma_2^2}{2}\partial_{yy} u + H(x, y, p, \nabla u) - \rho u = 0, \quad (x, y, t) \in \mathbb{R}^2 \times (0, T),\\ \partial_t \mu + \nabla \cdot (\mu D_q H) = \frac{\sigma_1^2}{2}\partial_{xx}^2 \mu + \frac{\sigma_2^2}{2}\partial_{yy}^2 \mu,\\ u(x, y, T) = 0,\ \mu(x, y, 0) = \mu_0(x, y), \quad (x, y) \in \mathbb{R}^2,\end{aligned}$$

where μ_0 is a given initial distribution. Note that the above problem may come with the additional constraint that μ satisfies $\mu > 0$ and $\int_{\mathbb{R}^2} \mu = 1$ for all $t \in (0, T)$, i.e., that $\mu(t)$ is a positive probability density for all $t \in (0, T)$.

There have been many advances in the construction and analysis of numerical methods for elliptic and parabolic HJB equations. One of the earliest, and most natural approaches for approximating such problems comes in the form of finite difference methods (FDMs) [8, 11, 10, 9, 18, 19, 41, 50, 76, 98, 101], for which, an important abstract framework for the convergence of monotone schemes to *viscosity* solutions [37] of (possibly degenerate) fully nonlinear second-order elliptic and parabolic equations (which encapsulates the aforementioned HJB equations) was developed by G. Barles, and P. Souganidis in [11]. This abstract framework has lead to the development and analysis of other schemes, in particular, nonmonotone schemes (which lead to higher orders of convergence) [18, 98].

However, monotone, wide stencil schemes impose restrictions upon the set of problems that the FDMs can capture, in general. This has lead to the development of methods that bear a closer resemblance to FEMs, namely *semi-Lagrangian* methods, which incorporate interpolation into finite element spaces, allowing for the use of unstructured meshes [41, 50].

In particular, [50] also utilises the result due to N. Krylov (which we utilise in Chapter 7), characterising the MA equation as a HJB equation. In this more flexible semi-Lagrangian framework, the authors prove convergence of the scheme to viscosity solutions of the corresponding HJB equation.

The field of FEMs for HJB problems has seen much fewer developments. As mentioned previously, the difficulty of designing such schemes is due to the lack

of a weak formulation for such problems. As such, the results of this thesis (in Chapters 5–7) serve to expand upon the existing finite element framework. The first advancement (to the author's knowledge) in this direction is given by [68], in which the authors analyse the convergence of monotone piecewise-linear finite element methods in an abstract setting (much in the spirit of G. Barles and P. Sougandinis, except the method does not exhibit the same consistency property as the monotone FDM, see [68]). A short follow-up paper [67] provides numerical experiments, validating the theoretical results of the previous paper, numerically capturing viscosity solutions of HJB equations.

In recent years, the development of FEMs for elliptic problems in nondivergence form, which do not rely upon a weak formulation of the underlying PDE, due to the consideration of *strong* solutions to the PDE, has seen a few advancements [78, 79, 110]. The framework of [110] lead naturally to the DGFEM introduced in [111] in the elliptic setting, and [112] in the parabolic setting. This method is applicable to HJB equations with cofficients that satisfy the Cordes condition. In this thesis, we further develop this method in the elliptic setting, allowing for more complicated geometries, and oblique boundary conditions.

We now find ourselves at the forefront of the development of finite element methods for MA and HJB type equations, and so, it seems pertinent to discuss the layout of this thesis. One should note that Chapters 3 to 9 all contain original research.

The thesis is laid out as follows: Chapter 2 introduces our notational conventions and the definition of relevant function spaces in a continuous setting; Chapter 3 provides existence and uniqueness results for MA and HJB type equations available in the current literature, as well as new results proven as part of this thesis. In Chapter 4 we give all of the necessary definitions, assumptions, and estimates in a discrete setting, which we will refer back to when defining the various finite element methods in the later chapters. This has required the development of several key tools from standard finite element theory to that of curved domains; in particular trace and inverse estimates, discrete Poincaré–Friedrichs' inequalities, and optimal interpolation estimates with fractional Sobolev regularity. Chapter 5 provides a DGFEM for linear elliptic equations with Dirichlet boundary conditions, on curved domains; this constitutes original research, extending the framework of [110]. Chapter 6 extends the framework of Chapter 5 from the Dirichlet boundary condition to the oblique boundary condition in two dimensions; this constitutes original research, and the content has been submitted for publication [71]. In Chapter 7 we provide and analyse a DGFEM for the

approximation of solutions to HJB type equations with Dirichlet and oblique boundary conditions; we utilise the results proven in Chapter 3, producing a DGFEM for the approximation of solutions to MA equations with Dirichlet boundary conditions. Chapter 8 describes our CGFEM for solving the MA equation, as well as an extension of the two-dimensional analysis found in [92] to the case $f(x, u, \nabla u) := \frac{f_1(x,u)}{f_2(\nabla u)}$ (as opposed to $f(x, u, \nabla u) = f(x)$); Chapter 9 explains the extension of the CGFEM in Chapter 8 to the MAOT problem. The work in this particular chapter constitutes a collaborative effort with O. Lakkis, and T. Pryer [72]. We conclude the thesis with a discussion on what has been achieved, and comment on future avenues of research.

1.2 List of notation in order of appearance

Notation	Reference	Name/ Description
Ω	(1.1.3)	General domain
Υ	(1.1.3)	Optimal transport target domain
$\mathbb{R}^{d\times d}_{\mathrm{Sym}}$, $\mathrm{SO}(2)$	(2.1.1), (2.1.2)	Symmetric and rotation matrices
$GL(\mathbb{R}^d)$	(2.1.3)	Invertible matrices
$A : B$	(2.1.5)	Frobenius inner product
$\overline{K}, K^\circ$	Definition 2.1.4	Closure and interior of the set K
$\mathbb{S}^d$	(2.1.6)	Unit sphere in $\mathbb{R}^{d+1}$
X	(2.1.7)	Krylov control set
X_ξ	(2.1.8)	Subset of Krylov control set
$C^k(K)$	Definition 2.2.2	Space of k times differentiable functions
$C^k(\overline{K})$	Definition 2.2.2	Space of $C^k(K)$ functions with uniformly continuous k-th derivative
$C^{k,\alpha}(\overline{K})$	Definition 2.2.6	$C^k(\overline{K})$ functions with α-Hölder continuous k-th derivative
$C^\infty_c(\overline{K})$	Definition 2.2.4	Smooth functions with compact support
$\|\cdot\|_{C^k(\overline{K})}$,	(2.2.5)	$C^k(\overline{K})$ norms
$\|\cdot\|_{C^{k,\alpha}(\overline{K})}$,	(2.2.5)	$C^{k,\alpha}(\overline{K})$ norms
$\|\cdot\|_{C^k(\overline{K}}$	(2.2.6)	$C^k(\overline{K})$ semi-norms
$\|\cdot\|_{C^{k,\alpha}(\overline{K})}$	(2.2.6)	$C^{k,\alpha}(\overline{K})$ semi-norms
$R^{f,x}$	Definition 2.2.10	Quadratic remainder term
$L^p(K)$	(2.3.1)–(2.3.2)	Lebesgue spaces
$\|\cdot\|_{p,K}$	(2.3.3)–(2.3.4)	Lebesgue space norms
$\langle\cdot,\cdot\rangle_K$	(2.3.5), (4.3.3)	L^2 Inner product
$\langle\cdot\rangle_K$	(2.3.6)	Integral
$L^p_0(K)$	(2.3.7)	Zero integral Lebesgue spaces
Du, ∇u	(2.3.8), (2.3.9)	Derivative and gradient
D^2u	(2.3.10)	Hessian
$W^{m,p}(K)$	(2.3.11)	Sobolev spaces
$H^s(K)$	(2.3.11), (2.3.16)	(Hilbert) Sobolev spaces
$\|\cdot\|_{W^{m,p}(K)}$	(2.3.12)	Sobolev space norms
$\|\cdot\|_{W^{m,p}(K)}$	(2.3.13)	Sobolev space seminorms
$\langle\cdot,\cdot\rangle_{H^m(K)}$	(2.3.14)	(Hilbert) Sobolev space inner products
$\|\cdot\|_{H^r(K)}$	(2.3.15)	Fractional Sobolev space seminorms
τ	Definition 2.3.6	Trace operator
$W^{m,p}_0(K)$	(2.3.21)	Sobolev spaces with trivial trace
$\langle\cdot\|\cdot\rangle$	Definition 2.4.1	Duality pairing
$\langle D^2v\|\phi\rangle$	Definition 2.4.2	Generalised Hessian duality pairing

Notation	Reference	Name/ Description
Λ	(3.3.4)	General HJB control set
$F[\cdot]$	(3.3.7)	HJB operator
$F_\gamma[\cdot]$	(3.3.17)	Renormalised HJB operator
γ, γ^α	(3.3.15), (3.3.16)	Renormalisation functions
β	(3.3.42)	Oblique vector
$H^2_{\beta,0}(K)$	(3.3.45)	Oblique derivative H^2- subspaces
Θ	Definition 3.3.24	Oblique angle
n_K	-	Unit outward normal vector to ∂K
$\mathcal{W}$	Definition 2.6.6	Weingarten map
$\mathcal{H}_{\partial\Omega}$	Definition 2.6.8	Mean curvature
$\mathscr{T}_h$	-	Triangulation
p	-	Finite element space polynomial degree
$[\![\cdot]\!]$	(4.3.1)	Jump operator
$\langle\!\langle\cdot\rangle\!\rangle$	(4.3.2)	Average operator
$\mathscr{E}_h^i, \mathscr{E}_h^b$	Definition 4.3.1	Set of interior faces, set of boundary faces
$\mathscr{E}_h^{i,b}, \mathscr{V}_h^b$	Definition 4.3.1	Set of faces, set of boundary vertices
F_K	(4.4.1)	Map from $\hat{K} \to K$
$K^*, \hat{K}, K$	Definition 4.4.1, 4.4.26	Various d-simplices
$\Phi_K, \tilde{F}_K$	(4.4.1), (4.4.2)	Nonaffine & affine part of F_K
$\tilde{B}_K, \tilde{b}_K$	(4.4.2)	Scaling & translation part of $\tilde{F}_K$
C_K, c	(4.4.3), (4.4.21)	Affine invariant constants
h_K, h	Definition 4.4.7	Local mesh size, global mesh size
$\tilde{h}_F$	(4.4.4)	Local mesh size associated to a face
$C_{\mathcal{T}}$	(4.4.5)	Mesh condition constant
$c_\ell, c_{-\ell}$	(4.4.11), (4.4.12),	Scaling argument constants
σ	(4.4.20)	Shape-regularity constant
$\rho(\cdot)$	(4.4.25)	Diameter of inscribed ball
$\mathscr{T}_h^c, \mathscr{T}_h^f$	Definition 4.4.22	Curved and flat partitions
$\mathscr{E}_h^{b,c}, \mathscr{E}_h^{b,f}$	Definition 4.4.23	Curved and flat face partitions
$v^*, \hat{v}, v$	Definition 4.4.26	Functions associated to $K^*, \hat{K}, K$
λ_i	Definition 4.5.1	Barycentric coordinates
$(\hat{K}, \hat{P}_K, \hat{\Sigma}_K)$	Definition 4.5.2	Reference straight Lagrange FE
(K, P_K, Σ_K)	Definition 4.5.3	Curved Lagrange FE
$\hat{\mu}, \mu$	(4.5.3), (4.5.5)	Degrees of freedom
$V_{h,p}, V_{h,p,0}$	Definition 4.5.4	DG finite element spaces
$\mathbb{V}_{h,p}, \mathring{\mathbb{V}}_{h,p}$	Definition 4.5.5	CG finite element spaces
$W^{s,r}(\Omega; \mathscr{T}_h)$	(4.5.9)	Broken Sobolev space
$\|\cdot\|_{W^{s,r}(\Omega;\mathscr{T}_h)}$	(4.5.10)	Broken Sobolev norm
$\|\cdot\|_{W^{s,r}_*(K)}$	(4.5.11)	Curved Sobolev seminorm
$\|\cdot\|_{W^{s,r}_*(\Omega;\mathscr{T}_h)}$	(4.5.12)	Curved Broken Sobolev seminorm
π_h	Definition 4.5.12	Classical Lagrange interpolation operator
$\lesssim$	Definition 4.5.16	Less than or equal to up to a constant

Notation	Reference	Name/ Description
$\approx$	Definition 4.5.16	Equal up to a constant
$\mathcal{P}$	Definition 4.5.15	Finite element projection operator
Π_h	(4.5.19)	Local Lagrange interpolation operator
$E(r,i)$	(4.6.14)	Chain rule index set
$\nabla_{\mathbf{T}}, \mathrm{div}_{\mathbf{T}}$	(4.10.1)	Tangential gradient and divergence
$\mathbf{H}_h$	Definition 4.11.1	Finite element Hessian
$\mathcal{L}$	(4.11.4)	Lift operator
μ_F, η_F	Sections 5, 6, and 7	Jump stabilisation parameters
u_h	Sections 5–8	Finite element solution
$\|\cdot\|_{h,\theta}$	(5.6.1)	DG norms on $V_{h,p}$
$\|(\cdot,\cdot)\|_{h,\theta}$	(6.8.1)	DG norms on $V_{h,p,0} \times V_{h,0}$
ℓ_F, σ_F	Sections 6 and 7	Oblique jump stabilisation parameters
$B^{\mathcal{D}}_{h,*}, B^{\mathcal{O}}_{h,*}$	(5.4.1), (6.6.1)	Dirichlet & oblique FE bilinear forms
$J^{\mathcal{D}}_h, J^{\mathcal{O}}_h$	(5.4.2), (6.6.2)	D & O FE bilinear jump forms
$B^{\mathcal{D}}_{h,\theta}, B^{\mathcal{O}}_{h,\theta}$	(5.4.3), (6.6.3)	D & O FE bilinear forms
$A^{\mathcal{D}}_h, A^{\mathcal{O}}_h$	(5.4.4), (6.6.4)	Linear D & O operators
$\mathscr{A}^{\mathcal{O}}_h, \mathscr{A}^{\mathcal{D}}_h$	(7.3.1)	Nonlinear D & O operators
$F^{\mathrm{MA}}_h[\cdot]$	(8.2.1)	Discrete MA operator
$\|\cdot\|_h$	(8.2.2)	CG FE space norm
$\|\cdot\|_{-1,h}$	(8.2.2)	CG FE space dual norm
$V, \mathring{V}$	(8.3.1)	Superspaces of $\mathbb{V}_{h,p}$ and $\mathring{\mathbb{V}}_{h,p}$
$F^{\mathrm{MA}}[\cdot]$	(8.3.2)	Continuous MA operator
L_u, R	(8.3.4), (8.3.5)	Linearisation and remainder of F^{MA}
$L_{u,h}, R$	-	Restriction of L_u and R to $\mathbb{V}_{h,p}$
M, M_h	(8.3.6), (8.3.7)	Fixed point operators
u_*	(8.3.14)	Elliptic projection of u
$\mathbb{B}_\rho(u_*)$	(8.3.16)	Ball of radius ρ about u_*
U	Section 9	Finite element solution
$\mathbb{Z}_{h,p}, \mathbb{W}_{h,p}$	(9.4.3), (9.4.4)	CG finite element recovery spaces
$\mathscr{F}$	(9.5.1)	General operator
$\lambda_\flat, \lambda_\sharp$	(9.5.3)	Uniform lower and upper ellipticity constants
$\mathscr{C}$	(9.5.21)	Cone of convex functions
$b, \mathscr{B}[\cdot]$	(9.5.18), (9.5.19)	Distance function and operator
$\mathbf{G}_h$	(9.5.27)	Gradient recovery operator
$\mathscr{H}$	(9.6.1)	Generalised Hessian with gradient recovery
$\widetilde{\mathbf{H}}_h$	(9.6.2)	Finite element Hessian with gradient recovery

Chapter 2

Notation, function spaces, and calculus

2.1 Matrices and sets

2.1.1 Matrices

For a matrix $A \in \mathbb{R}^{m\times n}$, $m, n \in \mathbb{N}$, we denote by A^T, the transpose of A. We denote

$$\mathbb{R}^{d\times d}_{\mathrm{Sym}} := \{A \in \mathbb{R}^{d\times d} : A^T = A\}, \tag{2.1.1}$$

and

$$\mathrm{SO}(2) := \left\{ \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} : \phi \in [0, 2\pi] \right\}, \tag{2.1.2}$$

the set of all 2×2 rotation matrices.

We also denote

$$GL(\mathbb{R}^d) := \{A \in \mathbb{R}^{d\times d} : \det A \neq 0\}, \tag{2.1.3}$$

i.e., the set of invertible $d \times d$ matrices.

For a matrix $A \in \mathbb{R}^{d\times d}$, we denote by $\mathrm{Tr}(A)$ the trace of the matrix to be the sum of the diagonal entries, i.e.,

$$\mathrm{Tr}(A) = \sum_{i=1}^{d} A_{ii}. \tag{2.1.4}$$

Definition 2.1.1 (Positive (semi)definiteness) *For a matrix $A \in \mathbb{R}^{d\times d}$ we write $A > 0$ ($A \geq 0$), and say that A is positive definite (positive semidefinite) if there exists $\mu > 0$ ($\mu \geq 0$) such that*

$$x^T Ax \geq \mu|x|^2 \quad \forall x \in \mathbb{R}^d.$$

Definition 2.1.2 (Frobenius inner product) *For two matrices $A, B \in \mathbb{R}^{d\times d}$, we define the Frobenius inner product of the two matrices as follows:*

$$A : B := \sum_{i,j=1}^{d} A_{ij}B_{ij} = \operatorname{Tr} A^T B = \operatorname{Tr} AB^T. \tag{2.1.5}$$

2.1.2 Sets

Definition 2.1.3 (Multi-index notation) *For a multi-index $\alpha \in \mathbb{N}_0^d$, we denote by*

$$\alpha! := \prod_{j=1}^{d} \alpha_j! \quad \textit{and} \quad |\alpha| := \sum_{j=1}^{d} \alpha_j.$$

Furthermore, for $x \in \mathbb{R}^d$, we denote

$$x^\alpha = \prod_{j=1}^{d} x_j^{\alpha_j}.$$

Definition 2.1.4 (Closure and interior) *For a given set $K \subset \mathbb{R}^{m\times n}$, $m, n \in \mathbb{N}$, we denote $\overline{K}$ to be the closure of K, and K° to be the interior of K.*

In the later chapters we will often consider unit vector-valued functions, and so it seems appropriate to define the set

$$\mathbb{S}^d := \{x \in \mathbb{R}^{d+1} : |x| = 1\}. \tag{2.1.6}$$

As mentioned in the introduction, we shall make use (in particular in Chapters 3 and 7) of a characterisation of the Monge–Ampére (MA) equation as a Hamilton–Jacobi–Bellman (HJB) equation. This characterisation was proven by Krylov [74], and requires the definition of the following control set.

Definition 2.1.5 (Krylov control set) *We define the Krylov control set, X, as follows:*

$$X := \{W \in \mathbb{R}^{d\times d}_{\mathrm{Sym}} : W \geq 0, \operatorname{Tr} W = 1\}. \tag{2.1.7}$$

We also define the subset X_ξ of X for $0 < \xi \leq 1/d^d$, as follows:

$$X_\xi := \{W \in X : \det W \geq \xi\}. \tag{2.1.8}$$

Definition 2.1.6 (Convex set) *Let $U \subset \mathbb{R}^d$. The set U is convex, if, for any $x, y \in U$, the open line segment connecting x and y, $(x, y) := \{(1-t)x+ty : t \in (0, 1)\}$ is contained in $\overline{U}$.*

Definition 2.1.7 (Uniformly convex set) *Let $U \subset \mathbb{R}^d$. The set U is* uniformly convex, *if, for any $x, y \in U$, the open line segment connecting x and y, $(x, y) := \{(1-t)x + ty : t \in (0,1)\}$ is contained in U°.*

2.2 Calculus: derivatives, classical function spaces, Taylor approximation error, and convexity

Definition 2.2.1 *Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $m, n \in \mathbb{N}$ be a continuous function. Given $x \in \mathbb{R}^{m\times n}$, we define the derivative, $Df(x)$, of f at the point x to be the element of $\mathbb{R}^{m\times n}$ that satisfies*

$$Df(x) : y = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon y) - f(x)}{\varepsilon}$$

for all $y \in \mathbb{R}^{m\times n}$. If $U \subset \mathbb{R}^{m\times n}$ is open and $Df(x)$ exists for all $x \in U$, then we say that f is differentiable in U. If the map $\mathbb{R}^{m\times n} \ni x \mapsto Df(x) \in \mathbb{R}^{m\times n}$ is continuous in U, then we say that f is continuously differentiable in U.

Definition 2.2.2 *A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is said to be k times continuously differentiable in $U \subset \mathbb{R}^{m\times n}$, open, if its k-th order derivative, defined iteratively by*

$$D^k f(x) := D(D^{k-1} f)(x)$$

is continuous in U.

Definition 2.2.3 (Spaces of k-times differentiable functions) *Let $U \subset \mathbb{R}^{m\times n}$, be open, $m, n \in \mathbb{N}$. We denote by $C^k(U)$, and $C^k(\overline{U})$, the following spaces of functions:*

$$\begin{aligned} C^k(U) &:= \{f : U \to \mathbb{R} \ : \ f \text{ is } k\text{-times continuously differentiable in } U\}, \\ C^k(\overline{U}) &:= \{f \in C^k(U) : \exists V \subset \mathbb{R}^{m\times n}, \text{open}, \ U \subset V^\circ, \tilde{f} \in C^k(V) : \tilde{f}|_U = f\}. \end{aligned} \tag{2.2.1}$$

Definition 2.2.4 (Hölder continuity) *A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is said to be α-Hölder continuous, with $\alpha \in (0, 1]$, in $U \subset \mathbb{R}^{m\times n}$, open, if there exists a constant $C \geq 0$ such that*

$$|f(x) - f(y)| \leq C|x - y|^\alpha, \tag{2.2.2}$$

for all $x, y \in U$.

Definition 2.2.5 (Lipschitz continuity) *A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is said to be Lipschitz continuous in $U \subset \mathbb{R}^{m\times n}$, open, if it is α-Hölder continuous with $\alpha = 1$. We denote the space of Lipschitz continuous functions by $C^{0,1}(\overline{U})$.*

Definition 2.2.6 ($C^{k,\alpha}$-function space) *Let $U \subset \mathbb{R}^{m\times n}$ be open, $m,n \in \mathbb{N}$, for $k \in \mathbb{N}_0$, $\alpha \in (0,1]$, we denote*

$$C^{k,\alpha}(\overline{U}) := \{f \in C^k(\overline{U}) : D^k f \text{ satisfies (2.2.2)}\}. \tag{2.2.3}$$

Definition 2.2.7 (Functions with compact support) *For $U \subset \mathbb{R}^{m\times n}$, open, we define the space $C_c^k(U)$, $k \in \mathbb{N}_0$, to be the set of k-times continuously differentiable functions with compact support. Furthermore, we define the space*

$$C_c^\infty(U) := \bigcap_{k=1}^{\infty} C_c^k(U). \tag{2.2.4}$$

Definition 2.2.8 *For $U \subset \mathbb{R}^{m\times n}$, and $k \in \mathbb{N}_0$, $\alpha \in (0,1]$, we define the following norms on $C^k(\overline{U})$, and $C^{k,\alpha}(\overline{U})$*

$$\begin{aligned} \|f\|_{C^k(\overline{U})} &:= \max_{1\le j\le k} \max_{x\in\overline{U}} |D^j f|, \\ \|f\|_{C^{k,\alpha}(\overline{U})} &:= \|f\|_{C^k(\overline{U})} + \sup_{x,y\in\overline{U}:x\neq y} \frac{|D^k f(x) - D^k f(y)|}{|x-y|^\alpha}, \end{aligned} \tag{2.2.5}$$

and semi-norms

$$\begin{aligned} |f|_{C^k(\overline{U})} &:= \max_{x\in\overline{U}} |D^k f|, \\ |f|_{C^{k,\alpha}(\overline{U})} &:= |f|_{C^k(\overline{U})} + \sup_{x,y\in\overline{U}:x\neq y} \frac{|D^k f(x) - D^k f(y)|}{|x-y|^\alpha}. \end{aligned} \tag{2.2.6}$$

The analysis present in the later chapters relies on the fact that a $C^k(\overline{\Omega}) \cap C^{k+1}(\Omega)$ function behaves like a $(k+1)$-th degree polynomial locally. That is, we can expand the function as a $(k+1)$-th order polynomial close to a given point. Furthermore, we are able to control the error arising in this expansion to our benefit.

Theorem 2.2.9 (Taylor approximation) *Let $f \in C^{k,1}(\overline{U}) \cap C^{k+1}(U)$, $k \in \mathbb{N}$, with $U \subset \mathbb{R}^{m\times n}$, $m,n \in \mathbb{N}$. Then, for a fixed $a \in U$, we may express f as follows:*

$$f(x) = \sum_{|\alpha|\le k} \frac{D^\alpha f(a)}{\alpha!}(x-a)^\alpha + \sum_{|\beta|=k+1} R_\beta(x)(x-a)^\beta,$$

where

$$R_\beta(x) = \frac{|\beta|}{\beta!} \int_0^1 (1-s)^{|\beta|-1} D^\beta f(a + s(x-a))\, ds.$$

Proof: Firstly, we note that by Theorem A.5 in [113], if a function $g \in C([a,b];\mathbb{R})$ such that the derivatives of g of order up to and including k are defined and continuous

on $[a,b]$, $D^k g$ is differentiable on the open interval (a,b), and $D^{k+1}g$ is integrable on (a,b). Then, for each $t \in [a,b]$, we have

$$g(t) = \sum_{j=0}^{k} \frac{D^j g(a)}{k!} + \int_a^t \frac{(t-s)^k}{k!} D^{k+1} g(s)\, ds.$$

Defining $g(t) := f(a + t(x-a))$, noting that $f(x) = g(1)$ and $f(a) = g(0)$, applying the above, we obtain the desired result. $\square$

Definition 2.2.10 (Quadratic remainder) *Let $f \in C^{k,1}(\overline{U}) \cap C^{k+1}(U)$, for some $U \subset \mathbb{R}^{m\times n}$, open, $m,n \in \mathbb{N}$ and some $k \geq 2$. We define the quadratic remainder function, $R^{f,x} : U \to \mathbb{R}$, of f, associated to a point $x \in U$ as follows*

$$\begin{aligned} R^{f,x}(y) :=& f(x+y) - f(x) - Df(x)\cdot y \\ =& \sum_{|\alpha|\leq 1} \frac{D^\alpha f(x)}{\alpha!} y^\alpha - f(x) - Df(x)\cdot y + \sum_{|\beta|=2} R_\beta(x+y) y^\beta \\ =& \sum_{|\beta|=2} R_\beta(x+y) y^\beta. \end{aligned} \tag{2.2.7}$$

Lemma 2.2.11 (Difference of two quadratic remainders estimate) *Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $m,n \in \mathbb{N}$, satisfy $f \in C^{2,1}(\overline{U}) \cap C^3(U)$, where $U \subset \mathbb{R}^{m\times n}$ is open. Then, for a given $x \in U$, and for any $y, z \in U$ we have*

$$|R^{f,x}(y) - R^{f,x}(z)| \leq C(m,n)\|f\|_{C^{2,1}(\overline{U})} \left(\sum_{1\leq|\beta|\leq 2} |y^\beta| + |z^\beta| \right) |y - z|.$$

Proof: Firstly, if $y = z$, then $|R^{f,x}(y) - R^{f,x}(z)| = 0$, and the result is trivial. Assume that $y \neq z$, then

$$\begin{aligned} |R^{f,x}(y) - R^{f,x}(z)| &= |\sum_{|\beta|=2} R_\beta(x+y)y^\beta - R_\beta(x+z)z^\beta| \\ &\leq \frac{1}{2} \sum_{|\beta|=2} |(R_\beta(x+y) + R_\beta(x+z))(y^\beta - z^\beta)| \\ &\quad + \frac{1}{2} \sum_{|\beta|=2} |(y^\beta + z^\beta)(R_\beta(x+y) - R_\beta(x+z))| \\ &=: \frac{1}{2}(I_1 + I_2). \end{aligned}$$

Now,

$$\begin{aligned}
I_2 &\leq \sum_{|\beta|=2} (|y^\beta| + |z^\beta|) \frac{|\beta|}{\beta!} \int_0^1 (1-t)^{|\beta|-1} |D^\beta f(x+ty) - D^\beta f(x+tz)|\, dt \\
&\leq \sum_{|\beta|=2} (|y^\beta| + |z^\beta|) \frac{2}{\beta!} \int_0^1 (1-t) \|D^2 f\|_{C^{0,1}(\overline{U})} |(x+ty) - (x+tz)|\, dt \\
&= \sum_{|\beta|=2} (|y^\beta| + |z^\beta|) \frac{2}{\beta!} \int_0^1 (1-t)t \|D^2 f\|_{C^{0,1}(\overline{U})}\, dt |y-z| \\
&= \sum_{|\beta|=2} \frac{1}{3\beta!} (|y^\beta| + |z^\beta|) \|D^2 f\|_{C^{0,1}(\overline{U})} |y-z| \\
&\leq \frac{1}{3} \|D^2 f\|_{C^{0,1}(\overline{U})} \sum_{|\beta|=2} (|y^\beta| + |z^\beta|) |y-z|.
\end{aligned}$$

Now we turn to I_1. Firstly we note that for any $(i,j) \in \{1,\dots,m\} \times \{1,\dots,n\}$ the following holds:

$$\begin{aligned}
|y-z| &= \left(\sum_{k,l=1}^{m,n} |y_l^k - z_l^k|^2 \right)^{1/2} \\
&\geq \left(|y_j^i - z_j^i|^2 \right)^{1/2} = |y_j^i - z_j^i|.
\end{aligned} \tag{2.2.8}$$

It then follows that for any $|\beta| = 2$

$$\frac{|y^\beta - z^\beta|}{|y-z|} = \frac{|y_{l_1}^{k_1} y_{l_2}^{k_2} - z_{l_1}^{k_1} z_{l_2}^{k_2}|}{|y-z|}, \tag{2.2.9}$$

for some $(k_1,l_1),(k_2,l_2) \in \{1,\dots,m\} \times \{1,\dots,n\}$. Moreover,

$$\begin{aligned}
|y_{l_1}^{k_1} y_{l_2}^{k_2} - z_{l_1}^{k_1} z_{l_2}^{k_2}| &\leq \frac{1}{2} |(y_{l_1}^{k_1} + z_{l_1}^{k_1})(y_{l_2}^{k_2} - z_{l_2}^{k_2})| + \frac{1}{2} |(y_{l_2}^{k_2} + z_{l_2}^{k_2})(y_{l_1}^{k_1} - z_{l_1}^{k_1})| \\
&\leq \frac{1}{2} (|y_{l_1}^{k_1}| + |z_{l_1}^{k_1}|) |y_{l_2}^{k_2} - z_{l_2}^{k_2}| + \frac{1}{2} (|y_{l_2}^{k_2}| + |z_{l_2}^{k_2}|) |y_{l_1}^{k_1} - z_{l_1}^{k_1}| \\
&\leq \frac{1}{2} (|y_{l_1}^{k_1}| + |z_{l_1}^{k_1}| + |y_{l_2}^{k_2}| + |z_{l_2}^{k_2}|) |y-z|.
\end{aligned} \tag{2.2.10}$$

Applying (2.2.10) to (2.2.9), we obtain

$$\frac{|y^\beta - z^\beta|}{|y-z|} \leq \frac{1}{2} (|y_{l_1}^{k_1}| + |z_{l_1}^{k_1}| + |y_{l_2}^{k_2}| + |z_{l_2}^{k_2}|).$$

Thus

$$\sum_{|\beta|=2} \frac{|y^\beta - z^\beta|}{|y-z|} \leq \frac{1}{2} \sum_{|\beta|=2} \left(\sum_{|\beta|=1} |y^\beta| + |z^\beta| \right) \leq \frac{\widetilde{C}(m,n)}{2} \sum_{|\beta|=1} |y^\beta| + |z^\beta|,$$

i.e.,

$$\sum_{|\beta|=2} |y^\beta - z^\beta| \leq \frac{\widetilde{C}(m,n)}{2} \left(\sum_{|\beta|=1} |y^\beta| + |z^\beta| \right) |y-z|.$$

It now follows that

$$\begin{aligned}
I_1 &\leq \sum_{|\beta|=2} |R_\beta(x+y) + R_\beta(x+z)||y^\beta - z^\beta| \\
&\leq \|f\|_{C^2(\overline{U})} \sum_{|\beta|=2} \frac{|\beta|}{\beta!} \int_0^1 (1-t)^{|\beta|-1}\, dt |y^\beta - z^\beta| \\
&= \|f\|_{C^2(\overline{U})} \sum_{|\beta|=2} \frac{2}{\beta!} \int_0^1 (1-t)\, dt |y^\beta - z^\beta| \\
&\leq \|f\|_{C^2(\overline{U})} \sum_{|\beta|=2} |y^\beta - z^\beta| \\
&\leq \frac{\widetilde{C}(m,n)}{2} \|f\|_{C^2(\overline{U})} \left(\sum_{|\beta|=1} |y^\beta| + |z^\beta| \right) |y-z|.
\end{aligned}$$

We use our estimates for I_1 and I_2 to obtain

$$\begin{aligned}
&|R^{f,x}(y) - R^{f,x}(z)| \\
&\quad \leq \frac{1}{2} \left(\frac{\widetilde{C}(m,n)\|f\|_{C^2(\overline{U})}}{2} \sum_{|\beta|=1} |y^\beta| + |z^\beta| + \frac{1}{3} \|D^2 f\|_{C^{0,1}(\overline{U})} \sum_{|\beta|=2} |y^\beta| + |z^\beta| \right) |y-z| \\
&\quad \leq \frac{\widetilde{C}(m,n)+1}{4} \|f\|_{C^{2,1}(\overline{U})} \left(\sum_{1\leq|\beta|\leq 2} |y^\beta| + |z^\beta| \right) |y-z|,
\end{aligned}$$

which is the desired estimate with $C(m,n) := (\widetilde{C}(m,n)+1)/4$. $\square$

Corollary 2.2.12 *Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, $m,n \in \mathbb{N}$, satisfy $f \in C^{2,1}(\overline{U}) \cap C^3(U)$, where $U \subset \mathbb{R}^{m\times n}$ is a bounded open set that contains the ball $B_r := \{x \in \mathbb{R}^{m\times n} : |x| \leq r\}$, $r \in [0,1)$. Then, for a given $x \in U$, and for any $y, z \in B_r$ we have*

$$\begin{aligned}
|R^{f,x}(y) - R^{f,x}(z)| &\leq C(m,n)\|f\|_{C^{2,1}(\overline{U})}(|y| + |z|)|y-z| \\
&\leq 2C(m,n)r\|f\|_{C^{2,1}(\overline{U})}|y-z|.
\end{aligned}$$

Proof: Applying Lemma 2.2.11 yields

$$|R^{f,x}(y) - R^{f,x}(z)| \leq C(m,n)\|f\|_{C^{2,1}(\overline{U})} \left(\sum_{1\leq|\beta|\leq 2} |y^\beta| + |z^\beta| \right) |y-z|. \qquad (2.2.11)$$

Since $y \in B_r$ and $r \in [0,1)$, it follows that for $|\beta| = 2$, $|y^\beta| \leq |y^{\beta'}|$ for some $|\beta'| = 1$. Similarly for $z \in B_r$. Thus, by the equivalence of norms of finite-dimensional spaces, we obtain

$$\begin{aligned}\sum_{1\leq|\beta|\leq 2} |y^\beta| + |z^\beta| &\leq C(m,n) \sum_{|\beta|=1} |y^\beta| + |z^\beta| \\ &\leq C(m,n)(|y| + |z|).\end{aligned}$$

Applying the above to (2.2.11), and noting that $|y|, |z| \leq r$, we obtain the desired estimates. □

Corollary 2.2.13 *Under the assumptions of Lemma 2.2.11, assume further that U contains the ball B_r, $r \in [0,1)$; then, the following holds for a given $x \in U$, and any $y \in B_r$:*

$$|R^{f,x}(y)| \leq C(m,n)\|f\|_{C^{2,1}(\overline{U})}|y|^2.$$

Proof: Notice that

$$|R^{f,x}(y)| = |R^{f,x}(y) - R^{f,x}(0)|.$$

Applying Corollary 2.2.12 with $z = 0$ yields the desired results. □

Definition 2.2.14 (Convex function) *Let $U \subset \mathbb{R}^d$ be open, and let $f \in C^2(U)$. The function f is called convex if its Hessian, D^2u is positive semidefinite for all $x \in U$. That is,*

$$\xi^T D^2 f(x) \xi \geq 0 \quad \forall \xi \in \mathbb{R}^d, \forall x \in U.$$

Definition 2.2.15 (Uniformly convex function) *Let $U \subset \mathbb{R}^d$ be open, and let $f \in C^2(U)$. The function f is called* uniformly *convex if its Hessian, D^2u is uniformly positive definite. That is, there exists a constant $\mu > 0$ such that*

$$\xi^T D^2 f(x) \xi \geq \mu|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \forall x \in U.$$

2.3 Function spaces

2.3.1 Lebesgue spaces

Let $K \subset \mathbb{R}^d$ be an open, bounded domain. We define the following Lebesgue spaces for $1 \leq p < \infty$,

$$L^p(K) = \left\{ v : K \to \mathbb{R} : \int_K |v|^p < \infty \right\}, \tag{2.3.1}$$

and for $p = \infty$,

$$L^\infty(K) = \{ v : K \to \mathbb{R} : \exists M \in \mathbb{R}^+ : |v(x)| \leq M \quad a.e.\ x \in K \}, \tag{2.3.2}$$

where the integration in (2.3.1) and the "a.e." in (2.3.2) are considered with respect to the Lebesgue measure. We equip the spaces $L^p(K)$, $1 \leq p < \infty$, and $L^\infty(K)$ with the following norms

$$\|v\|_{p,K} = \left(\int_K |v|^p \right)^{\frac{1}{p}}, \tag{2.3.3}$$

$$\|v\|_{\infty,K} = \inf\{M \in \mathbb{R}^+ : |v(x)| \leq M \quad a.e.\ x \in K\}, \tag{2.3.4}$$

respectively. Note that the pair $(L^p(K), \|\cdot\|_{p,K})$ forms a Banach space for any $1 \leq p \leq \infty$.

In the case that $p = 2$, we may equip $L^2(K)$ with the inner product

$$\langle u, v \rangle_K = \int_K uv. \tag{2.3.5}$$

In this case, the pair $(L^2(K), \langle \cdot, \cdot \rangle_K)$ forms a Hilbert space.

In the case that $p = 1$, for $m, n \in \mathbb{N}$, we consider the linear functional $\langle \cdot \rangle_K : L^1(K; \mathbb{R}^{m \times n}) \to \mathbb{R}^{m \times n}$ given by

$$\langle u \rangle_K := \langle u, 1 \rangle_K = \int_K u. \tag{2.3.6}$$

We also define the following subsets of $L^p(K)$ for $1 < p < \infty$:

$$L^p_0(K) := \left\{ v \in L^p(K) : \int_K v = 0 \right\}, \tag{2.3.7}$$

and equip them with the respective L^p-norms given by (2.3.3).

We define

$$Du := \left[\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d} \right], \tag{2.3.8}$$

to be the derivative of a function $u : K \to \mathbb{R}$, and we define the gradient of u, ∇u to be the derivative's transpose, i.e.,

$$\nabla u = (Du)^T. \tag{2.3.9}$$

For second order derivatives, we denote by $D^2 u$, the Hessian of u, i.e., the $d \times d$ matrix of second order partial derivatives of u;

$$[D^2 u]^i_j = \frac{\partial^2 u}{\partial x_i \partial x_j}. \tag{2.3.10}$$

2.3.2 Sobolev spaces

Definition 2.3.1 (Weak derivative) *If $u, v \in L^1(K)$, with $K \subset \mathbb{R}^d$, open, we say that v is the α^{th}-weak derivative of u, where $\alpha \in \mathbb{N}_0^d$ is a multi-index, if*

$$\int_K u\, D^\alpha \varphi = (-1)^{|\alpha|} \int_K v\, \varphi,$$

for all $\varphi \in C_c^\infty(K)$. Here the (classical) derivative $D^\alpha \varphi$ is defined as

$$D^\alpha \varphi = \frac{\partial^{|\alpha|}\varphi}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}.$$

In the case that u has an α^{th}-weak derivative, we denote it by $D^\alpha u$.

For $1 \le p \le \infty$, $m \in \mathbb{N}_0$ we define the standard Sobolev spaces (see [46]):

$$\begin{aligned} W^{m,p}(K) &:= \{v \in L^p(K) : D^\alpha v \in L^p(K),\ \forall \alpha : |\alpha| \le m\}, \\ H^m(K) &:= W^{m,2}(K). \end{aligned} \tag{2.3.11}$$

Here $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ is a multi-index, $|\alpha| = \sum_{i=1}^d \alpha_i$, and the derivatives, D^α, are understood in the weak sense. We endow $W^{m,p}(K)$ with the following norms:

$$\begin{aligned} \|v\|_{W^{m,p}(K)} &= \left(\sum_{|\alpha| \le m} \|D^\alpha v\|_{p,K}^p \right)^{\frac{1}{p}},\ \text{if } 1 \le p < \infty, \\ \|v\|_{W^{m,\infty}(K)} &= \max_{|\alpha| \le m} \|D^\alpha v\|_{\infty,K},\ \text{if } p = \infty, \end{aligned} \tag{2.3.12}$$

and semi norms:

$$\begin{aligned} |v|_{W^{m,p}(K)} &= \left(\sum_{|\alpha| = m} \|D^\alpha v\|_{p,K}^p \right)^{\frac{1}{p}},\ \text{if } 1 \le p < \infty, \\ |v|_{W^{m,\infty}(K)} &= \max_{|\alpha| = m} \|D^\alpha v\|_{\infty,K},\ \text{if } p = \infty. \end{aligned} \tag{2.3.13}$$

Note that the pair $(W^{m,p}(K), \|\cdot\|_{W^{m,p}(K)})$ is a Banach space. We can also equip the space $H^m(K)$ with the following inner product

$$\langle u, v \rangle_{H^m(K)} = \sum_{|\alpha| \le m} \langle D^\alpha u, D^\alpha v \rangle_K. \tag{2.3.14}$$

Note that the pair $(H^m(K), \langle \cdot, \cdot \rangle_{H^m(K)})$ forms a Hilbert space.

Definition 2.3.2 (Non integer Sobolev space) *For $0 < r < 1$, we define the Sobolev space $H^r(\Omega)$ as follows*

$$H^r(\Omega) = \{v \in L^2(\Omega) : |v|_{H^r(\Omega)} < \infty\},$$

where

$$|v|^2_{H^r(\Omega)} := \int_\Omega \int_\Omega \frac{|v(x_1) - v(x_2)|^2}{|x_1 - x_2|^{d+2r}}. \tag{2.3.15}$$

Definition 2.3.3 (Higher order Non integer Sobolev space) *For $s \in \mathbb{R} \setminus \mathbb{N}$, decomposing $s = \ell + r$ with $\ell \in \mathbb{N}$, $0 < r < 1$, we define the Sobolev space $H^s(\Omega)$ as follows*

$$H^s(\Omega) = \{v \in H^\ell(\Omega) : D^\alpha v \in H^r(\Omega), \forall \alpha : |\alpha| = l\}. \tag{2.3.16}$$

Lemma 2.3.4 (Non integer H^r-multipliers) *Assume that $u \in H^r(\Omega)$, $0 < r < 1$ and $\psi \in C^{0,1}(\overline{\Omega})$. Then, there exists a constant C depending only on d and r, such that*

$$|u\,\psi|_{H^r(\Omega)} \leq \sqrt{2}\|\psi\|_{L^\infty(\Omega)}|u|_{H^r(\Omega)} + \sqrt{2}C(d,r)\sqrt{1+\mathrm{diam}(\Omega)^2}|\psi|_{C^{0,1}(\overline{\Omega})}\|u\|_{L^2(\Omega)}. \tag{2.3.17}$$

Proof: We see that

$$\begin{aligned} |u\,\psi|^2_{H^r(\Omega)} &= \int_\Omega \int_\Omega \frac{|u(x_1)\psi(x_1) - u(x_2)\psi(x_2)|^2}{|x_1 - x_2|^{d+2r}} \\ &\leq 2\int_\Omega \int_\Omega \frac{|u(x_1)\psi(x_1) - u(x_2)\psi(x_1)|^2}{|x_1 - x_2|^{d+2r}} + 2\int_\Omega \int_\Omega \frac{|u(x_2)\psi(x_1) - u(x_2)\psi(x_2)|^2}{|x_1 - x_2|^{d+2r}}. \end{aligned} \tag{2.3.18}$$

It then follows that

$$\begin{aligned} \int_\Omega \int_\Omega \frac{|u(x_1)\psi(x_1) - u(x_2)\psi(x_1)|^2}{|x_1 - x_2|^{d+2r}} &\leq \|\psi\|^2_{L^\infty(\Omega)} \int_\Omega \int_\Omega \frac{|u(x_1) - u(x_2)|^2}{|x_1 - x_2|^{d+2r}} \\ &= \|\psi\|^2_{L^\infty(\Omega)}|u|^2_{H^r(\Omega)}. \end{aligned} \tag{2.3.19}$$

Furthermore,

$$\begin{aligned} \int_\Omega \int_\Omega \frac{|u(x_2)\psi(x_1) - u(x_2)\psi(x_2)|^2}{|x_1 - x_2|^{d+2r}} &= \int_\Omega \int_{\Omega\cap\{|x_1-x_2|\leq 1\}} \frac{|u(x_2)|^2|\psi(x_1) - \psi(x_2)|^2}{|x_1 - x_2|^{d+2r}} \\ &\quad + \int_\Omega \int_{\Omega\cap\{|x_1-x_2|\geq 1\}} \frac{|u(x_2)|^2|\psi(x_1) - \psi(x_2)|^2}{|x_1 - x_2|^{d+2r}} \\ &\leq |\psi|^2_{C^{0,1}(\overline{\Omega})} \int_\Omega \int_{\Omega\cap\{|x_1-x_2|\leq 1\}} \frac{|u(x_2)|^2}{|x_1 - x_2|^{d+2(r-1)}} \\ &\quad + |\psi|^2_{C^{0,1}(\overline{\Omega})} \int_\Omega \int_{\Omega\cap\{|x_1-x_2|\geq 1\}} \frac{|u(x_2)|^2|x_1 - x_2|^2}{|x_1 - x_2|^{d+2r}} \end{aligned}$$

Noting that for $x_1, x_2 \in \Omega \cap \{|x_1 - x_2| \geq 1\}$, $|x_1 - x_2| \leq \text{diam}(\Omega)$, and applying the change of variables $z = x_2 - x_1$, and we obtain

$$\begin{aligned}\int_\Omega \int_\Omega \frac{|u(x_2)\psi(x_1) - u(x_2)\psi(x_2)|^2}{|x_1 - x_2|^{d+2r}} &\leq |\psi|^2_{C^{0,1}(\overline{\Omega})} \int_\Omega \int_{|z|\leq 1} \frac{1}{|z|^{d+2(r-1)}} |u(x_2)|^2 \\ &\quad + \text{diam}(\Omega)^2 |\psi|^2_{C^{0,1}(\overline{\Omega})} \int_\Omega \int_{|z|\geq 1} \frac{1}{|z|^{d+2r}} |u(x_2)|^2 \\ &\leq C(d,r)(1 + \text{diam}(\Omega))|\psi|^2_{C^{0,1}(\overline{\Omega})} \|u\|^2_{L^2(\Omega)}. \end{aligned} \tag{2.3.20}$$

Applying (2.3.19) and (2.3.20) to (2.3.18), we obtain

$$|u\,\psi|^2_{H^r(\Omega)} \leq 2\|\psi\|^2_{L^\infty(\Omega)} |u|^2_{H^r(\Omega)} + 2C(d,r)(1 + \text{diam}(\Omega)^2)|\psi|_{C^{0,1}(\overline{\Omega})} \|u\|^2_{L^2(\Omega)}.$$

Taking square roots above, we obtain (2.3.17). □

2.3.2.1 Traces

We first state the following theorem from [60].

Theorem 2.3.5 *Let Ω be a bounded open subset of $\mathbb{R}^d$ with a $C^{k,1}$ boundary, with $k \in \mathbb{N}_0$. Assume that $\frac{1}{2} < s \leq k+1$, and $s - \frac{1}{2}$ is not an integer. Then, the mapping*

$$u \mapsto u|_{\partial\Omega},$$

which is defined for $u \in C^{k,1}(\overline{\Omega})$, has a unique continuous extension as an operator from $H^s(\Omega)$ to $H^{s-1/2}(\partial\Omega)$.

Definition 2.3.6 (Trace operator) *Under the assumptions of Theorem 2.3.5 on k, s, and Ω, we define $\tau : H^s(\Omega) \to H^{s-\frac{1}{2}}(\partial\Omega)$ to be the unique extension of the operator $C^{k,1}(\overline{\Omega}) \ni u \mapsto u|_{\partial\Omega} \in C^{k,1}(\partial\Omega)$, from $H^s(\Omega)$ to $H^{s-\frac{1}{2}}(\partial\Omega)$.*

Corollary 2.3.7 *Assume that $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. Then, the trace operator, τ, maps $H^2(\Omega)$ to $H^{1/2}(\partial\Omega)$.*

Proof: Under the given assumptions, by Theorem 2.3.5, it follows that $\tau : H^1(\Omega) \to H^{1/2}(\partial\Omega)$. Since $H^2(\Omega) \subset H^1(\Omega)$, the statement holds. □

Remark 2.3.8 *Corollary 2.3.7 emphasises the fact that the trace of a $H^2(\Omega)$ function is well-defined for Lipschitz continuous domains. The higher domain boundary-regularity assumption of Theorem 2.3.5 is required in order to prove the corresponding higher regularity of traces of Sobolev functions.*

Sobolev spaces with trivial trace: We define the space

$$W_0^{m,p}(\Omega) := \{v \in W^{m,p}(\Omega) : \tau(D^\alpha v)|_{\partial\Omega} = 0,\ \forall \alpha : |\alpha| \leq m-1\}. \tag{2.3.21}$$

Density of smooth functions in Sobolev spaces: For $1 \leq p < \infty$, $k \in \mathbb{N}$ the spaces $W^{m,p}(\Omega)$ and $W_0^{m,p}(\Omega)$ can be equivalently defined in the following way:

$$W^{m,p}(\Omega) := \overline{C^\infty(\Omega)}^{\|\cdot\|_{W^{m,p}(\Omega)}},$$

and

$$W_0^{m,p}(\Omega) := \overline{C_c^\infty(\Omega)}^{\|\cdot\|_{W^{m,p}(\Omega)}}.$$

2.4 Generalised Hessian

Definition 2.4.1 (Duality pairing) *Let H be a normed space, with dual space, H'. For $f \in H'$, we denote its action on functions $v \in H$, by the following duality pairing*

$$\langle f|v\rangle_{H\times H'} := f(v) \quad \forall v \in H.$$

The continuous Galerkin FEMs introduced in Chapters 8 and 9 require the notion of the so-called "finite element Hessian", which in turn relies on the definition of the "Generalised Hessian". Looking first at a function $v \in C^2(\Omega)\cap C^1(\overline{\Omega})$, an application of integration by parts shows us that the Hessian of v, D^2v, satisfies (the $d \times d$ set of equations)

$$\langle D^2 v, \varphi\rangle_\Omega = -\langle \nabla v\, D\varphi\rangle_\Omega + \langle \nabla v\, n_{\partial\Omega}^T \varphi\rangle_{\partial\Omega} \quad \forall \varphi \in H^1(\Omega),$$

where the integrals, $\langle\cdot\rangle_\Omega$ and $\langle\cdot\rangle_{\partial\Omega}$, on the right hand side are considered in the sense of (2.3.6), and $n_{\partial\Omega}$ is the unit outward normal to $\partial\Omega$. We now generalise this to a given function $v \in H^1(\Omega)$ with $\nabla v n_{\partial\Omega}^T|_{\partial\Omega} \in \left(H^{1/2}(\partial\Omega)'\right)^{d\times d}$.

Definition 2.4.2 (Generalised Hessian) *Let $v \in H^1(\Omega)$, with $[\nabla v\, n_{\partial\Omega}^T]_j^i|_{\partial\Omega} \in H^{1/2}(\partial\Omega)'$, $i,j = 1,\dots d$. We define the generalised Hessian of v, also denoted by D^2v to be the element of $(H^1(\Omega)')^{d\times d}$ that satisfies the duality pairing*

$$\langle [D^2 v]_j^i|\varphi\rangle := -\left\langle \frac{\partial v}{\partial x_i}, \frac{\partial \varphi}{\partial x_j}\right\rangle_\Omega + \langle [\nabla v\, n_{\partial\Omega}^T]_j^i|\varphi\rangle_{(H^{1/2}(\partial\Omega))'\times H^{1/2}(\partial\Omega)} \quad \forall \varphi \in H^1(\Omega), \tag{2.4.1}$$

and all $i,j = 1,\dots,d$, where $n_{\partial\Omega}$ is the unit outward normal to $\partial\Omega$.

Remark 2.4.3 *Let $v \in H^1(\Omega)$, with $\nabla v\, n_{\partial\Omega}^T|_{\partial\Omega} \in \left(H^{1/2}(\partial\Omega)'\right)^{d\times d}$; then, the generalised Hessian of v, D^2v, is symmetric. This also implies the symmetry of the finite element Hessian, which we will define in Chapter 4.*

2.5 Domain type definitions and assumptions

Definition 2.5.1 (C^k domain) *A domain $\Omega \subset \mathbb{R}^d$ called C^k for $k \in \mathbb{N}$, if, for any $x \in \partial\Omega$, there exists an open neighbourhood V of x in $\mathbb{R}^d$ and an orthogonal coordinate system $(y_1, \ldots, y_d)$, such that*

$$V = \{(y_1, \ldots, y_d) : -a_j < y_j < a_j, 1 \leq j \leq d\};$$

as well as a uniformly C^k function φ defined on $V' = \{(y_1, \ldots, y_{d-1}) : -a_j < y_j < a_j, 1 \leq j \leq d-1\}$ and such that

$$\begin{aligned}
&|\varphi(y')| \leq a_d/2 \text{ for every } y' = (y_1, \ldots, y_{d-1}) \in V',\\
&\Omega \cap V = \{y = (y', y_d) \in V : y_d < \varphi(y')\},\\
&\Gamma \cap V = \{y = (y', y_d) \in V : y_d = \varphi(y')\}.
\end{aligned}$$

Remark 2.5.2 *The analysis of the numerical methods present in Chapters 5, and 7 only requires the domain, Ω, to be sufficiently piecewise regular, with piecewise non-negative curvature. This means that there are cases where the numerical methods admit the existence and uniqueness of a solution, but the assumptions of the existence and uniqueness Theorems 3.3.16 and 3.3.29, for the underlying PDEs, are not satisfied. This motivates the following two definitions.*

Definition 2.5.3 (Piecewise C^k domain) *A domain $\Omega \subset \mathbb{R}^d$ is piecewise C^k for $k \in \mathbb{N}$, if we may express the boundary of Ω, $\partial\Omega$, as a finite union*

$$\partial\Omega = \bigcup_{n=1}^{N} \overline{\Gamma_n}, \tag{2.5.1}$$

where each $\Gamma_n \subset \mathbb{R}^d$ is of zero d-dimensional Lebesgue measure, and admits a local representation as the graph of a uniformly C^k function. That is, for each n, and at each $x \in \Gamma_n$ there exists an open neighbourhood V_n of x in $\mathbb{R}^d$ and an orthogonal coordinate system $(y_1^n, \ldots, y_d^n)$, such that

$$V_n = \{(y_1^n, \ldots, y_d^n) : -a_j^n < y_j^n < a_j^n, 1 \leq j \leq d\};$$

as well as a uniformly C^k function φ_n defined on $V_n' = \{(y_1^n, \ldots, y_{d-1}^n) : -a_j^n < y_j^n < a_j^n, 1 \leq j \leq d-1\}$ and such that

$$\begin{aligned}
&|\varphi_n(y^{n\prime})| \leq a_d^n/2 \text{ for every } y^{n\prime} = (y_1^n, \ldots, y_{d-1}^n) \in V_{n'},\\
&\Omega \cap V = \{y^n = (y^{n\prime}, y_d^n) \in V : y_d^n < \varphi_n(y^{n\prime})\},\\
&\Gamma_n \cap V = \{y^n = (y^{n\prime}, y_d^n) \in V : y_d^n = \varphi_n(y^{n\prime})\}.
\end{aligned}$$

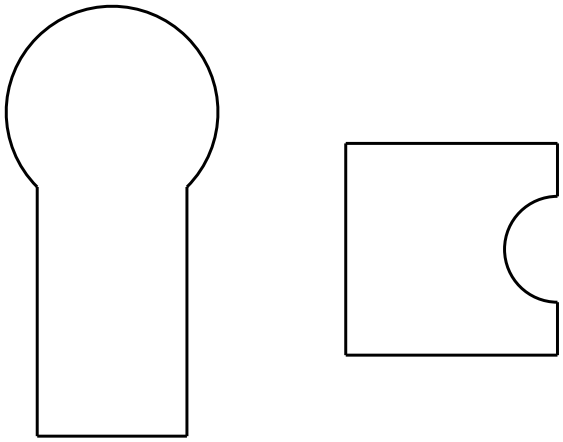


Figure 2.1: Examples of the "key-hole" shaped domain (left) given by (2.5.2), and a domain with a boundary portion of strictly negative curvature (right) given by (2.5.3).

Remark 2.5.4 *Definition 2.5.3 tells us that for a given Γ_n and $x \in \Gamma_n$, that, in a neighbourhood of x, the domain Ω is "below" the graph of φ_n, and Γ_n is the graph of φ_n. From this point of view, we may characterise a notion of piecewise convexity, for which the function φ_n is also assumed to be* concave.

Definition 2.5.5 (Piecewise C^k and piecewise convex domain) *A domain $\Omega \subset \mathbb{R}^d$, is a piecewise C^k and piecewise convex domain, with $k \geq 2$, if Ω is Lipschitz continuous, and the boundary of Ω, $\partial\Omega$, is given by a finite union of the form (2.5.1), and each $\Gamma_n \subset \mathbb{R}^d$, can be represented as the graph of a uniformly C^k,* concave *function φ_n, in the local coordinates.*

Remark 2.5.6 *Note that if Ω is a piecewise C^k and piecewise convex domain, with $k \geq 2$, it is not in necessarily convex. For example, consider the "key-hole shaped" domain*

$$\Omega = \{x^2 + y^2 < 1 : y \geq 1/\sqrt{2}\} \cup [-1/\sqrt{2}, 1/\sqrt{2}] \times [-3, 1/\sqrt{2}]. \tag{2.5.2}$$

See Figure (2.1).

Remark 2.5.7 *The assertion that ϕ_n is concave, in Definition 2.5.5, excludes piecewise smooth domains with boundary portions of strictly negative curvature, for example the subset of $\mathbb{R}^2$ given by:*

$$Z := ([-2, 0] \times [0, 2]) \setminus \{x_1^2 + x_2^2 < 1/4\}, \tag{2.5.3}$$

for which the corresponding functions φ_n locally describing $\partial Z \cap \{x_1^2 + x_2^2 = 1/4\}$ must be uniformly convex (see figure (2.1)).

2.6 The Weingarten map and curvature

In this thesis, we often consider domains that have a curved boundary. Furthermore, our finite element spaces are defined on triangulations of the domain, where the triangulation provides an *exact* representation of the domain. A consequence of this, is the appearance of curvature dependent terms when applying integration by parts identities. As such, it is pertinent that we define the notion of the second fundamental form and mean curvature. We do this under the assumption that the domain has a C^2 boundary (and so we can provide a notion of curvature *everywhere* on the boundary); we then define a notion of piecewise curvature on domains that are assumed to be Lipschitz continuous and *piecewise* C^2.

Definition 2.6.1 (The outward unit normal vector) *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, with boundary $\partial\Omega$. We denote by $n_{\partial\Omega} : \partial\Omega \to \mathbb{S}^d$, the unit outward normal vector to $\partial\Omega$.*

Remark 2.6.2 *If the domain Ω is only Lipschitz continuous, the normal vector is defined a.e. with respect to the $(d-1)$-dimensional Lebesgue measure. Consider, for example, the unit outward normal to the unit square, $[0,1] \times [0,1]$. In this case, the unit outward normal is not well defined at the corners $(0,0), (1,0), (1,1), (0,1)$ (the collection of which is a set of measure zero with respect to the one-dimensional Lebesgue measure), but*

$$n_{\partial\Omega} = \begin{cases} (1,0)^T \text{ on } \{x = 1, 0 < y < 1\}, \\ (-1,0)^T \text{ on } \{x = 0, 0 < y < 1\}, \\ (0,1)^T \text{ on } \{0 < x < 1, y = 1\}, \\ (0,-1)^T \text{ on } \{0 < x < 1, y = 0\}. \end{cases}$$

Thus, $n_{\partial\Omega}$ is defined a.e. on $\partial\Omega$.

Lemma 2.6.3 (Regularity of the unit normal vector) *Assume that $\Omega \subset \mathbb{R}^d$ is a C^k domain, $k \in \mathbb{N}$. Then, we have that $n_{\partial\Omega} \in C^{k-1}(\partial\Omega; \mathbb{S}^d)$.*

Proof: By definition at any $x \in \partial\Omega$, there exists a neighbourhood V of x, with orthogonal coordinate system $(y_1, \ldots, y_d)$ and a uniformly C^k function $\varphi : V' \to \mathbb{R}$, such that $\partial\Omega \cap V = \{y = (y', y_d) \in V : y_d = \varphi_n(y')\}$. It then follows that $\partial\Omega \cap V = \{y \in V : g(y) = 0\}$, where $g(y) = y_d - \varphi(y')$. Furthermore on $\partial\Omega \cap V$, $n_{\partial\Omega} = \frac{\nabla g}{|\nabla g|} = \frac{(-\nabla_{y'}\varphi, 1)}{\sqrt{1+|\nabla_{y'}\varphi|^2}}$. Then, since φ is uniformly C^k in V', it follows that $n_{\partial\Omega}$ is

C^{k-1} on $\partial\Omega \cap V$. However, this holds for all $x \in \partial\Omega$, and thus $n_{\partial\Omega} \in C^1(\partial\Omega; \mathbb{S}^d)$. $\square$

Corollary 2.6.4 (Piecewise regularity of the unit normal vector) *Assume that $\Omega \subset \mathbb{R}^d$ is a piecewise C^k domain, $k \in \mathbb{N}$, with C^2 portions $\Gamma_1, \ldots, \Gamma_N$, $N \in \mathbb{N}$. Then, we have that $n_{\partial\Omega} \in C^{k-1}(\Gamma_n; \mathbb{S}^d)$, $n = 1, \ldots, N$.*

Proof: This is analogous to the proof of Lemma 2.6.3, utilising the local level set description of each C^k boundary portion Γ_j, $j = 1, \ldots, N$. $\square$

Definition 2.6.5 (Tangential gradient) *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, with boundary $\partial\Omega$. We define the tangential gradient, $\nabla_{\mathbf{T}}$, as follows:*

$$\nabla_{\mathbf{T}} = \nabla - n_{\partial\Omega} \frac{\partial}{\partial n_{\partial\Omega}},$$

where $\frac{\partial}{\partial n_{\partial\Omega}} = n_{\partial\Omega} \cdot \nabla$.

Definition 2.6.6 (Weingarten map) *Let $\Omega \subset \mathbb{R}^d$ be a C^k domain, where the integer $k \geq 2$. We define the Weingarten map, $\mathcal{W} \in C^{k-2}(\partial\Omega; \mathbb{R}^{d\times d})$, to be the tangential gradient of the unit outward normal to $\partial\Omega$, that is*

$$\mathcal{W} := \nabla_{\mathbf{T}} n_{\partial\Omega}^T.$$

If a function $f \in C^k(\partial\Omega)$, $k \in \mathbb{N}$, one can see that for all $x \in \partial\Omega$, $\nabla_{\mathbf{T}} f(x)$ is a tangent vector to $\partial\Omega$ at x, and thus the Weingarten map has a trivial eigenvalue of zero, with corresponding eigenvector, $n_{\partial\Omega}$.

Definition 2.6.7 *We denote by $\kappa_1, \ldots, \kappa_{d-1}$ the* nontrivial *eigenvalues of $\mathcal{W}$ (ordered by increasing magnitude), the principal curvatures of $\partial\Omega$, and we define the constant $\kappa_{\partial\Omega} := \inf_{x\in\partial\Omega} \kappa_1(x)$, to be the minimal principal curvature of $\partial\Omega$.*

Definition 2.6.8 (Mean curvature) *Let $\Omega \subset \mathbb{R}^d$ be a C^k domain, where the integer $k \geq 2$. We define the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} \in C^{k-2}(\partial\Omega)$, to be the trace of the Weingarten map, i.e.,*

$$\mathcal{H}_{\partial\Omega} := \mathrm{Tr}(\mathcal{W}) = \nabla_{\mathbf{T}} \cdot n_{\partial\Omega}. \tag{2.6.1}$$

Definition 2.6.9 (Piecewise Weingarten map) *Let $\Omega \subset \mathbb{R}^d$ be a piecewise C^k domain, $k \in \mathbb{N}$, with C^k portions $\Gamma_1, \ldots, \Gamma_N$, $N \in \mathbb{N}$, where the integer $k \geq 2$. We define the Weingarten map, $\mathcal{W} : \partial\Omega \to \mathbb{R}^{d\times d}$, as follows*

$$\mathcal{W}|_{\Gamma_n} := \nabla_{\mathbf{T}}(n_{\partial\Omega}^T|_{\Gamma_n}), \quad n = 1, \ldots, N.$$

Definition 2.6.10 (Piecewise mean curvature) *Let $\Omega \subset \mathbb{R}^d$ be a piecewise C^k domain, $k \in \mathbb{N}$, with C^k portions $\Gamma_1, \ldots, \Gamma_N$, $N \in \mathbb{N}$, where the integer $k \geq 2$. We define the piecewise mean curvature $\mathcal{H}_{\partial\Omega} : \partial\Omega \to \mathbb{R}$ as follows,*

$$\mathcal{H}_{\partial\Omega}|_{\Gamma_n} := \nabla_{\mathbf{T}} \cdot (n_{\partial\Omega}|_{\Gamma_n}), \quad n = 1, \ldots, N. \tag{2.6.2}$$

Chapter 3

PDEs and PDE analysis

3.1 New contributions and existing results

The original contributions of this chapter involve the analysis of Hamilton–Jacobi–Bellman (HJB) type equations and their links to Monge–Ampère (MA) type equations. Let us first summarise the existing results related HJB type problems that we build upon:

- The existence and uniqueness of strong (H^2-regular) solutions to HJB equations satisfying the Cordes condition (see Section 3.3.1 for a definition of the Cordes condition), with *homogeneous* Dirichlet boundary conditions on convex domains is proven in [111].
- The Miranda–Talenti estimate for subsets of functions of H^2 that satisfy the oblique boundary condition (see Definition 3.3.1 for further details) is proven in [89] (this is a key tool in our analysis of HJB type equations).
- In [74] it is proven that the MA equation is equivalent to a particular HJB equation (see Section 3.4 for further details).

Original contributions of this chapter:

1. We provide a general framework for the wellposedness of HJB type equations in subspaces of H^2 that satisfy the Miranda–Talenti estimate. Furthermore, utilising the Miranda–Talenti estimate proven in [89], this framework provides the existence and uniqueness of solutions to HJB type equations with oblique boundary conditions.
2. We prove the existence and uniqueness of strong (H^2-regular) solutions to HJB equations satisfying the Cordes condition (see Section 3.3.1 for a definition of

the Cordes condition), with *inhomogeneous* Dirichlet boundary conditions on convex domains.

3. We modify the characterisation in [74], characterising the MA equation as a HJB equation that satisfies the Cordes condition. This requires an assumption of uniform convexity and uniform $C^{2,\alpha}$-regularity, $\alpha \in (0,1)$, of the solution of the MA equation.

3.2 Monge–Ampère type equations

In this section we provide existence and uniqueness results for MA type equations from [65] in the Dirichlet case, and [27] in the case of the second boundary condition. We will utilise these results in the later chapters. As outlined in the introduction, the main focus of this thesis is to design and analyse finite element methods (FEMs), for the approximation of solutions to Monge–Ampère (MA) type equations.

The two types of MA equations we consider are determined (and distinguished) by their corresponding boundary conditions. The first boundary condition that we consider is the Dirichlet boundary condition, i.e., the value of the solution is prescribed on the boundary of the domain. The second boundary condition that we consider is a nonlinear gradient constraint, which requires that the gradient of the solution to maps one given domain to another. This leads us to the following boundary-value problems (BVPs): Firstly, given a uniformly convex, open domain $\Omega \subset \mathbb{R}^d$, a function $u : \Omega \to \mathbb{R}^d$ satisfies the Monge–Ampère Dirichlet (MAD) problem, if

$$\begin{cases} \det D^2 u(x) = f(x, u(x), \nabla u(x)), & x \in \Omega, \\ u(x) = \phi(x), & x \in \partial\Omega, \end{cases} \tag{3.2.1}$$

where $f : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^+$, and $\phi : \partial\Omega \to \mathbb{R}$ are given (particular assumptions upon f, ϕ and Ω, that guarantee the well-posedness of (3.2.1) will be given later on in this chapter).

Secondly, given two uniformly convex, open domains $\Omega, \Upsilon \subset \mathbb{R}^d$, a function $u : \Omega \to \mathbb{R}$ satisfies the MA problem with second boundary conditions, if

$$\begin{cases} \det D^2 u(x) = f(x, u(x), \nabla u(x)), & x \in \Omega, \\ \nabla u(\Omega) = \Upsilon, \end{cases} \tag{3.2.2}$$

where $f : \Omega \times \mathbb{R} \times \Upsilon \to \mathbb{R}^+$ (again, particular assumptions upon f, Ω, and Υ that guarantee the well-posedness of (3.2.2) will be given later on in this chapter).

Equations (3.2.1) and (3.2.2) have been the interest of study for many years, in the case of (3.2.1) see, for example [29, 73, 65], and for (3.2.2) see [27, 116]. This literature leads us to the following existence and uniqueness theorems, which we will use later on in order to analyse the finite element methods we propose, where the existence, uniqueness, and regularity properties of solutions come into play. We now state the following existence theorem from [65]. Recall that the minimal principal curvature, $\kappa_{\partial\Omega}$ is defined in Definition 2.6.7.

Theorem 3.2.1 (MAD problem existence theorem) *Assume that $f : \Omega \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^+$ satisfies*

$$f(x, z, q) \leq \zeta(|q|), \tag{3.2.3}$$

for some nondecreasing function $\zeta(t) \geq 1$. Assume that ζ, the boundary data $\phi : \partial\Omega \to \mathbb{R}$, and the minimal principal curvature of $\partial\Omega$, $\kappa_{\partial\Omega}$, satisfy the following:

$$1 < d\kappa_{\partial\Omega}^d \int_0^\infty \frac{t^{d-1}}{\zeta(t + \rho_\phi)}\, dt, \quad \rho_\phi = (1 + \frac{4d}{\kappa_{\partial\Omega}^4})\|\phi\|_{C^2(\partial\Omega)}. \tag{3.2.4}$$

Furthermore, assume that $\partial\Omega \in C^{k+2,\alpha}$, $\phi \in C^{k+2,\alpha}(\partial\Omega)$, $f \in C^{k,\alpha}(\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d)$, $k \in \mathbb{N}_0$, $0 < \alpha < 1$. Furthermore, assume that there exists a constant $\delta > 0$ such that $f \geq \delta$. Then the problem (3.2.1) has a uniformly convex solution in the class $C^{k+2,\alpha}(\overline{\Omega})$.

Remark 3.2.2 *Condition (3.2.4) constitutes a connection between the curvature of $\partial\Omega$, the boundary datum, ϕ, and the growth in the q component of the function f. One particular example that this assumption encompasses is the case that $f \in C^{k,\alpha}(\overline{\Omega})$, $k \in \mathbb{N}_0$, $\alpha \in (0, 1)$, is independent of z and q, i.e., $f := f(x)$ (the verification of (3.2.4) in the homogeneous Dirichlet case, for $d = 2$ is provided in the proof of Theorem 3.5.3). A further example is the $d = 2$ problem of prescribed Gaussian curvature, where $f(x, z, q) := K(x)(1 + |q|^2)^2$, with $K \in C^{k,\alpha}(\overline{\Omega}; \mathbb{R}^+)$ a uniformly positive function, $k \in \mathbb{N}_0$, $\alpha \in (0, 1)$. In this case, we may define $\zeta(t) := K^*(1 + t^2)^2$, where $K^* := \max\{\sup_{x\in\overline{\Omega}} |K(x)|, 1\}$. Firstly, this implies that $\zeta \geq K^* \geq 1$ is nondecreasing, and that $f(x, z, q) \leq \zeta(|q|)$ for all $x, z, q \in \Omega \times \mathbb{R} \times \mathbb{R}^d$.*

Furthermore,

$$\begin{aligned}\int_0^\infty \frac{t}{\zeta(t+\rho_\phi)}\,dt &= \frac{1}{K^*}\int_0^\infty \frac{t+\rho_\phi}{(1+(t+\rho_\phi)^2)^2}\,dt - \frac{\rho_\phi}{K^*}\int_0^\infty \frac{1}{(1+(t+\rho_\phi)^2)^2}\,dt \\ &= \frac{1}{K^*}\int_0^\infty \frac{d}{dt}\left(-\frac{1}{2(1+(t+\rho_\phi)^2)}\right)\,dt \\ &\quad - \frac{\rho_\phi}{K^*}\int_0^\infty \frac{d}{dt}\left(\frac{1}{2}\left(\frac{t+\rho_\phi}{1+(t+\rho_\phi)^2} - \arctan(t+\rho_\phi)\right)\right)\,dt \\ &= \frac{1}{2K^*}\left(-\frac{1}{1+(t+\rho_\phi)^2} - \frac{\rho_\phi t+\rho_\phi^2}{1+(t+\rho_\phi)^2} - \rho_\phi \arctan(\rho_\phi+t)\right)\Bigg|_0^\infty\end{aligned}$$

$$= \frac{1}{2K^*}\left(\frac{1+\rho_\phi^2}{1+\rho_\phi^2} - \rho_\phi\left(\frac{\pi}{2} - \arctan(\rho_\phi)\right)\right) = \frac{1}{2K^*}\left(1 - \rho_\phi\left(\frac{\pi}{2} - \arctan(\rho_\phi)\right)\right). \tag{3.2.5}$$

The equation $1 - x(\pi/2 - \arctan(x)) = 0$ *has no real roots, and since the integrand in (3.2.5) is nonnegative, we deduce that the integral in (3.2.5) is strictly positive for finite values of* ρ_ϕ*. This provides us with a sufficient condition for the existence of a solution to the prescribed Gaussian curvature equation, namely*

$$2\kappa_{\partial\Omega}^2 \int_0^\infty \frac{t}{\zeta(t+\rho_\phi)}dt = \frac{\kappa_{\partial\Omega}^2}{K^*}\left(1 - \rho_\phi\left(\frac{\pi}{2} - \arctan(\rho_\phi)\right)\right) > 1, \tag{3.2.6}$$

where we recall that ρ_ϕ *is defined in (3.2.4).*

Note that in the homogeneous Dirichlet case, $\phi \equiv 0$*, and so* $\rho_\phi = 0$*. It then follows that (3.2.6) simplifies to*

$$\frac{\kappa_{\partial\Omega}^2}{K^*} > 1.$$

We now state existence theorems from [27].

Theorem 3.2.3 *Assume that* Ω *and* Υ *are uniformly convex domains in* $\mathbb{R}^d$*, with* $\partial\Omega, \partial\Upsilon \in C^{2,1}$*. In addition, assume that* $f \in C^{1,1}(\overline{\Omega} \times \mathbb{R} \times \Upsilon)$ *is a positive function satisfying*

$$f(x,z,q) \to \infty \quad as \quad z \to \infty,$$
$$f(x,z,q) \to 0 \quad as \quad z \to -\infty,$$

uniformly for $(x,q) \in \Omega \times \Upsilon$*. Then, problem (3.2.2) has a convex solution* $u \in C^{3,\alpha} \cap C^{2,\alpha}(\overline{\Omega})$ *for any* $\alpha \in (0,1)$*. The solution is unique if* $\frac{\partial f}{\partial z} > 0$ *in* $\Omega \times \mathbb{R} \times \Upsilon$*.*

Theorem 3.2.4 (MA optimal transport problem existence theorem) *Assume that*

$$f(x,z,q) = \frac{f_1(x)}{f_2(q)},$$

where $f_1 \in C^{1,1}(\overline{\Omega})$ *and* $f_2 \in C^{1,1}(\overline{\Upsilon})$ *are two positive functions satisfying*

$$\int_\Omega f_1 = \int_\Upsilon f_2. \tag{3.2.7}$$

Furthermore, assume that Ω *and* Υ *satisfy the assumptions of Theorem 3.2.3. Then, problem (3.2.2) has a convex solution* $u \in C^{3,\alpha}(\Omega) \cap C^{2,\alpha}(\overline{\Omega})$ *for any* $\alpha \in (0,1)$*. Any two such solutions differ by a constant.*

Remark 3.2.5 *Elliptic regularity theory implies that the solutions obtained in Theorems 3.2.3 and 3.2.4 are more regular if* $\partial\Omega$*,* $\partial\Upsilon$ *and* f, f_1, f_2 *are more regular. In particular, if* $\partial\Omega$*,* $\partial\Upsilon$*, and* f, f_1, f_2 *are* C^∞*, then* $u \in C^\infty(\overline{\Omega})$*.*

3.3 HJB type equations and Miranda–Talenti estimates

In this section, we consider nonlinear equations of the form: find $u : \Omega \to \mathbb{R}$ such that

$$\sup_{\alpha\in\Lambda}\{A^\alpha : D^2u - f^\alpha\} = 0 \quad \text{a.e. in} \quad \Omega, \tag{3.3.1}$$

along with either Dirichlet boundary conditions:

$$u = g \quad \text{on } \partial\Omega, \tag{3.3.2}$$

wher $g : \partial\Omega \to \mathbb{R}$ is given, or oblique boundary conditions:

$$\beta \cdot \nabla u \quad \text{is constant on } \partial\Omega, \tag{3.3.3}$$

where $\beta : \partial\Omega \to \mathbb{S}^{d-1}$ is given. The PDEs given by (3.3.1) coupled with (3.3.2) *or* (3.3.3), are of Hamilton–Jacobi–Bellman (HJB) type. They arise in finance, economics, mean-field games, and problems of optimal control. In Section 3.4 we will see that they also arise when considering problems of MA type.

The index set

$$\Lambda \tag{3.3.4}$$

is called the "control set" (since, in the context of optimal control problems, it indexes the set of controls), and is assumed to be a compact metric space. We also assume that we have the real-valued functions

$$A_{ij} = A_{ji} \in C(\overline{\Omega} \times \Lambda), \quad i,j = 1,\dots d, \text{ and } f \in C(\overline{\Omega} \times \Lambda). \tag{3.3.5}$$

Then, for each $\alpha \in \Lambda$, consider the functions $A_{ij}^{\alpha} : x \mapsto A_{ij}(x,\alpha),\ x \in \overline{\Omega}$, and $f^{\alpha} : x \mapsto f(x,\alpha),\ x \in \overline{\Omega}$, and define the matrix–valued functions $A^{\alpha} = (A_{ij}^{\alpha})$.

For our purposes, Λ indexes a set of nondivergence form elliptic operators L^{α}, defined by

$$L^{\alpha}u := A^{\alpha} : D^2u, \quad \alpha \in \Lambda,$$

where each $A^{\alpha} \in L^{\infty}(\Omega; \mathbb{R}^{d\times d}_{\text{sym}})$, and the resulting family of operators $\{L^{\alpha}\}_{\alpha\in\Lambda}$ is uniformly elliptic, that is: there exist positive constants $\mu_1 \leq \mu_2$ such that

$$\mu_1|\xi|^2 \leq \sum_{i,j=1}^{d} A_{ij}^{\alpha}(x)\xi_i\xi_j \leq \mu_2|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega,\ \forall \alpha \in \Lambda. \tag{3.3.6}$$

Thus, we see that $L^{\alpha} : H^2(\Omega) \to L^2(\Omega)$ for each $\alpha \in \Lambda$. Furthermore, it follows that the nonlinear operator F defined by

$$F[u] := \sup_{\alpha\in\Lambda}\{A^{\alpha} : D^2u - f^{\alpha}\}, \quad u \in H^2(\Omega), \tag{3.3.7}$$

also maps $H^2(\Omega)$ into $L^2(\Omega)$. We may now express the problem (3.3.1) as follows: to find $u : \Omega \to \mathbb{R}$ such that

$$F[u] = 0 \quad \text{a.e. in } \Omega. \tag{3.3.8}$$

By considering suitable subsets H of $H^2(\Omega)$ (corresponding to particular boundary conditions), we are able to deduce under which conditions there may exist a unique u belonging to H that satisfies (3.3.8). Thus, we identify the conditions upon A^{α}, f^{α} for $\alpha \in \Lambda$, the domain, Ω, and the boundary conditions, that lead to the unique solvability of the corresponding boundary-value problems. In particular, we will consider subspaces $H \subset H^2(\Omega)$ that satisfy the following definition.

Definition 3.3.1 *A closed, linear subspace H of $H^2(\Omega)$, satisfies the Miranda–Talenti (MT) estimate property, if*

$$|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)} \quad \forall u \in H, \tag{3.3.9}$$

and

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2(\Omega)} \quad \forall u \in H, \tag{3.3.10}$$

where C is a positive constant independent of u.

Remark 3.3.2 *A particular example of a subset of $H^2(\Omega)$ that satisfies Definition 3.3.1 is the space $H^2(\Omega) \cap H_0^1(\Omega)$,* when *the domain $\Omega \subset \mathbb{R}^d$ is assumed to be convex (for further justification of this, see Theorem 3.3.16). This choice of space corresponds to the homogeneous Dirichlet boundary-value problem.*

Remark 3.3.3 *The fact that the constant on the right-hand side of inequality (3.3.9) is equal to* 1 *is essential to the proof of Theorem 3.3.8. In particular, it is utilised in order to obtain the final estimate of (3.3.21).*

3.3.1 Uniform ellipticity and the Cordes condition

It turns out that for dimensions $d \geq 3$, the uniform ellipticity assumption (3.3.6), alone, does not, in general, imply the unique solvability of the following linear Dirichlet boundary-value problem:

$$\begin{cases} A : D^2 u = 0 & \text{a.e. in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.3.11}$$

in $H^2(\Omega) \cap H_0^1(\Omega)$. That is, we can provide an example of a coefficient matrix $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\text{Sym}})$ that satisfies (3.3.6), for which the corresponding problem (3.3.11) admits a nontrivial solution that belongs to $H^2(\Omega) \cap H_0^1(\Omega)$ (clearly, $u \equiv 0$ satisfies (3.3.11), and belongs to $H^2(\Omega) \cap H_0^1(\Omega)$). Note that (3.3.11) is a special case of the boundary-value problem given by (3.3.1) coupled with (3.3.2), where $g \equiv 0$, the control set Λ is a singleton set, the functions $A_{ij} \in L^\infty(\Omega)$, and, furthermore, $f \equiv 0$.

The example we provide is from [89], pg 18, and is originally due to C. Pucci.

Example 3.3.4 (Nontrivial solution) *In this example we take Ω to be the unit ball in $\mathbb{R}^d$, $d \geq 3$. Let us define*

$$A_{ij}(x) := \delta_{ij} + r_1 \frac{x_i x_j}{|x|^2}, \quad i, j = 1, \ldots, d,$$

where $r_1 = (d + r_2 - 2)/(1 - r_2)$, and $\max\{2 - d/2, 0\} < r_2 < 1$. It is then clear that

$$\sum_{i,j=1}^{d} A_{ij}(x)\xi_i\xi_j = \sum_{i,j=1}^{d} (\delta_{ij} + r_1 \frac{x_i x_j}{|x|^2})\xi_i\xi_j \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega,$$

and that

$$|\xi|^2 \leq \sum_{i,j=1}^{d} (\delta_{ij} + r_1 \frac{x_i x_j}{|x|^2})\xi_i\xi_j \leq (1 + r_1)|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega,$$

i.e., (3.3.6) holds with $\mu_1 = 1$, and $\mu_2 = 1 + r_1$. We now define $u(x) := |x|^{r_2} - 1$, and see that

$$D^2_{ij} u = r_2(r_2 - 2)x_i x_j |x|^{r_2 - 4} + r_2 \delta_{ij} |x|^{r_2 - 2} \in L^2(\Omega), \quad i, j = 1 \ldots, d.$$

It then follows that for a.e. $x \in \Omega$,

$$\begin{aligned}
A : D^2 u &= \sum_{i,j=1}^{d} \left(\delta_{ij} + r_1 \frac{x_i x_j}{|x|^2} \right) (r_2(r_2-2)x_i x_j |x|^{r_2-4} + r_2 \delta_{ij} |x|^{r_2-2}) \\
&= |x|^{r_2-2} \left(\sum_{i=1}^{d} r_2(r_2-2)x_i^2 |x|^{-2} + r_2 + r_1 r_2 x_i^2 |x|^{-2} + \sum_{i,j=1}^{d} r_1 r_2 (r_2-2) x_i^2 x_j^2 |x|^{-4} \right) \\
&= |x|^{r_2-2} \left(r_2(r_2-2) + d r_2 + r_1 r_2 + r_1 r_2 (r_2-2) \right) \\
&= |x|^{r_2-2} (r_2^2 + (d-2) r_2 + r_1 r_2 (r_2-1)) \\
&= |x|^{r_2-2} (r_2^2 + (d-2) r_2 + ((d + r_2 - 2)/(1-r_2))(r_2-1)) \\
&= |x|^{r_2-2} (r_2^2 - r_2^2) = 0.
\end{aligned}$$

Thus, $A : D^2 u = 0$ *a.e. in* Ω, *and* $u \in H^2(\Omega) \cap H_0^1(\Omega)$, *i.e., u satisfies (3.3.11).*

Example 3.3.4 shows that for $d \geq 3$, the uniform ellipticity condition (3.3.6) is not in general sufficient to deduce uniqueness of solutions. This leads us to define the following condition: the family of operators $\{L^\alpha\}_{\alpha \in \Lambda}$ satisfies the *Cordes condition* if there exists $\varepsilon \in (0, 1]$ such that for each $\alpha \in \Lambda$,

$$\frac{|A^\alpha|^2}{(\mathrm{Tr}(A^\alpha))^2} \leq \frac{1}{d - 1 + \varepsilon} \quad \text{in } \overline{\Omega}, \tag{3.3.12}$$

where the norm $|\cdot|$ is the Frobenius norm, which is defined as follows:

$$|A| := \sqrt{A : A}, \quad \forall A \in \mathbb{R}^{m \times n},\ m, n \in \mathbb{N}, \tag{3.3.13}$$

where we recall that the Frobenius inner product above is define by (2.1.5).

Remark 3.3.5 *Condition (3.3.12) can be viewed as a strong ellipticity condition. In particular, let us assume that a positive definite matrix* $A \in \mathbb{R}^{d \times d}_{\mathrm{Sym}}$ *satisfies*

$$\frac{|A|^2}{(\mathrm{Tr}(A))^2} \leq \frac{1}{d - 1 + \varepsilon},$$

for some $\varepsilon \in (0, 1]$. *Noting that* $|I_d|^2 = I_d : I_d = d$, *and* $\mathrm{Tr}(A) = I_d : A$, *where* I_d *is the* $d \times d$ *identity matrix, rearranging the above, and taking square roots yields the following equivalent inequality*

$$\frac{I_d : A}{|I_d||A|} \geq \frac{\sqrt{d - 1 + \varepsilon}}{\sqrt{d}}. \tag{3.3.14}$$

Since I_d *and* A *are symmetric, we may associate them with vectors in* $\mathbb{R}^{d(d+1)/2}$, *and thus the value on the left-hand side of (3.3.14) is the equal to* $\cos\varphi$, *where* φ

is the angle between these two vectors. Furthermore, the right-hand side of (3.3.14) is positive, which means that the angle between the vectors is strictly less than $\pi/2$. Moreover, the right-hand side of (3.3.14) is increasing in the dimension d, and thus the angle is decreasing in d, and so, for higher dimensions one can deduce that the Cordes condition is a stronger assumption.

We may utilise the Cordes condition (3.3.12), by renormalising the family of operators $\{L^\alpha\}_{\alpha\in\Lambda}$. This is achieved by defining the strictly positive function $\gamma : \overline{\Omega}\times\Lambda \to \mathbb{R}^+$ by

$$\gamma(x,\alpha) := \frac{\operatorname{Tr} A^\alpha}{|A^\alpha|^2}, \tag{3.3.15}$$

and, similarly to defining the functions A^α_{ij}, f^α; for each $\alpha\in\Lambda$, we define

$$\gamma^\alpha : x \mapsto \gamma(x,\alpha),\ x\in\overline{\Omega}. \tag{3.3.16}$$

It follows from the continuity assumptions upon the coefficients and from the uniform ellipticity condition (3.3.6) that $\gamma\in C(\overline{\Omega}\times\Lambda)$, as well as the fact that there exists a constant $\gamma_0>0$ such that $\gamma\geq\gamma_0$ on $\overline{\Omega}\times\Lambda$. We then consider the renormalised family of operators $\{\gamma^\alpha L^\alpha\}_{\alpha\in\Lambda}$, which induces a renormalised operator $F_\gamma : H^2(\Omega)\to L^2(\Omega)$, given as follows:

$$F_\gamma[u] := \sup_{\alpha\in\Lambda}\{\gamma^\alpha(A^\alpha : D^2u - f^\alpha)\},\quad u\in H^2(\Omega). \tag{3.3.17}$$

The significance of the Cordes condition (3.3.12), and the choice of renormalisation parameter γ, is motivated as follows: let us take the operator $\gamma^\alpha L^\alpha$, for some $\alpha\in\Lambda$, and a function $u\in H^2(\Omega)$. Then, we see that for a.e. $x\in\Omega$,

$$|\gamma^\alpha L^\alpha u - \Delta u| = |(\gamma^\alpha A^\alpha - I_d) : D^2u| \leq |\gamma^\alpha A^\alpha - I_d||D^2u|.$$

Furthermore,

$$\begin{aligned}
|\gamma^\alpha A^\alpha - I_d|^2 &= (\gamma^\alpha)^2|A^\alpha|^2 - 2\gamma^\alpha\operatorname{Tr}(A^\alpha) + d\\
&= d - \frac{\operatorname{Tr}(A^\alpha)^2}{|A^\alpha|^2} && \text{by the definition (3.3.15) of } \gamma\\
&\leq d - (d-1+\varepsilon) && \text{by the Cordes condition (3.3.12)}\\
&= 1-\varepsilon.
\end{aligned}$$

Thus

$$|\gamma^\alpha L^\alpha u - \Delta u| \leq \sqrt{1-\varepsilon}|D^2u| \quad \text{a.e. in } \Omega, \tag{3.3.18}$$

where $\varepsilon\in(0,1]$, that is, we can control the difference between the operator $\gamma^\alpha L^\alpha$, and the Laplacian, Δ.

Remark 3.3.6 *In the case that $d = 2$, uniform ellipticity implies the Cordes condition (3.3.12). Indeed, let us take $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\mathrm{Sym}})$ that satisfies the uniform ellipticity condition (3.3.6). Then, denoting $\lambda_1, \lambda_2 : \Omega \to \mathbb{R}$ to be lower and upper eigenvalues of the (matrix-valued) function A, respectively, we see that*

$$\begin{aligned}\frac{|A|^2}{(\mathrm{Tr}(A))^2} &= \frac{\lambda_1^2 + \lambda_2^2}{(\lambda_1 + \lambda_2)^2} \\ &= \frac{\lambda_1^2 + \lambda_2^2}{\lambda_1^2 + \lambda_2^2 + 2\lambda_1\lambda_2} \\ &= \frac{1}{1 + 2\lambda_1\lambda_2/(\lambda_1^2 + \lambda_2^2)} \\ &= \frac{1}{(d-1) + 2/(\frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1})}.\end{aligned}$$

Moreover, the function $f(t) := 2/(t + t^{-1})$ is increasing on $(0, 1]$, and since $\lambda_1/\lambda_2 : \Omega \to (0, 1]$, and $\lambda_1/\lambda_2 \geq \mu_1/\mu_2$, we have that $f(\lambda_1/\lambda_2) \geq f(\mu_1/\mu_2)$, and so,

$$\frac{|A|^2}{(\mathrm{Tr}(A))^2} = \frac{1}{(d-1) + f(\lambda_1/\lambda_2)} \leq \frac{1}{(d-1) + 2\mu_1\mu_2/(\mu_1^2 + \mu_2^2)} \quad \text{a.e. in } \Omega.$$

Thus the Cordes condition (3.3.12) holds for any choice of $\varepsilon \in (0, 2\mu_1\mu_2/(\mu_1^2 + \mu_2^2))$.

We now present Lemma 1 from [111], which generalises estimate (3.3.18), based on the fact that for two collections of scalars $\{x_\alpha\}_{\alpha\in\Lambda}, \{y_\alpha\}_{\alpha\in\Lambda}$, one has that

$$\left| \left(\sup_{\alpha\in\Lambda} x_\alpha \right) - \left(\sup_{\alpha\in\Lambda} y_\alpha \right) \right| \leq \sup_{\alpha\in\Lambda} |x_\alpha - y_\alpha|.$$

Lemma 3.3.7 *Let Ω be a bounded open subset of $\mathbb{R}^d$, and suppose that (3.3.6) and (3.3.12) both hold. Then, for any open set $U \subset \Omega$ and $u, v \in H^2(\Omega)$, the following inequality holds a.e. in U:*

$$|F_\gamma[u] - F_\gamma[v] - \Delta(u - v)| \leq \sqrt{1 - \varepsilon}|D^2(u - v)|. \tag{3.3.19}$$

Theorem 3.3.8 *Assume that H is a subset of $H^2(\Omega)$, that satisfies the MT estimate property (3.3.9)–(3.3.10). Furthermore, assume that the collection of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satisfy (3.3.6) and (3.3.12). Then, there exists a unique $u \in H$ such that*

$$\int_\Omega F_\gamma[u]\, \Delta v = 0 \quad \forall v \in H. \tag{3.3.20}$$

Proof: Firstly, since (3.3.9)–(3.3.10) hold, we see that $(H, \langle \cdot, \cdot \rangle_H)$ forms a Hilbert space, when we define

$$\langle u, v \rangle_H := \int_\Omega \Delta u \, \Delta v, \quad u, v \in H.$$

Then, we define $\mathscr{A} : H \to H'$ (where H' is the dual space of H) by

$$\langle \mathscr{A}(u), v \rangle := \langle F_\gamma[u], v \rangle_H, \quad u, v \in H.$$

Now we will prove that $\mathscr{A}$ is Lipschitz continuous and strongly monotone, so that we can apply the Browder–Minty theorem [107]. This part of the proof is very similar to the proof of Theorem 3 from [111], but it is useful for the reader to see the approach. Furthermore, the current framework is more general, since it allows for more general boundary conditions (which are incorporated in the definition of the space H).

We see that for $u, v \in H$

$$\begin{aligned}
\langle \mathscr{A}(u) - \mathscr{A}(v), u - v \rangle &= \int_\Omega (F_\gamma[u] - F_\gamma[v]) \Delta(u - v) \\
&= \int_\Omega (\Delta(u - v))^2 + \int_\Omega (F_\gamma[u] - F_\gamma[v] - \Delta(u - v)) \, \Delta(u - v) \\
&= \|\Delta(u - v)\|_{2,\Omega}^2 + \int_\Omega (F_\gamma[u] - F_\gamma[v] - \Delta(u - v)) \, \Delta(u - v) \\
&\geq \|\Delta(u - v)\|_{2,\Omega}^2 - \int_\Omega |F_\gamma[u] - F_\gamma[v] - \Delta(u - v)||\Delta(u - v)| \\
&\geq \|\Delta(u - v)\|_{2,\Omega}^2 - \sqrt{1 - \varepsilon} \int_\Omega |D^2(u - v)||\Delta(u - v)|,
\end{aligned}$$

where the final inequality follows from (3.3.19). Applying the Cauchy–Schwarz inequality, followed by (3.3.9), we obtain

$$\begin{aligned}
\langle \mathscr{A}(u) - \mathscr{A}(v), u - v \rangle &\geq \|\Delta(u - v)\|_{2,\Omega}^2 - \sqrt{1 - \varepsilon} |u - v|_{H^2(\Omega)} \|\Delta(u - v)\|_{2,\Omega} \\
&\geq (1 - \sqrt{1 - \varepsilon}) \|\Delta(u - v)\|_{2,\Omega}^2.
\end{aligned} \tag{3.3.21}$$

Finally, applying (3.3.10) yields

$$\langle \mathscr{A}(u) - \mathscr{A}(v), u - v \rangle \geq C \|u - v\|_{H^2(\Omega)}^2, \tag{3.3.22}$$

i.e., $\mathscr{A}$ is strongly monotone.

We then see that for $u, v, z \in H$

$$\begin{aligned}
\langle \mathscr{A}(u) - \mathscr{A}(v), z\rangle &\leq \int_\Omega |F_\gamma[u] - F_\gamma[v]||\Delta z| \\
&= \int_\Omega \left| \sup_{\alpha\in\Lambda}\{A^\alpha : D^2 u - f^\alpha\} - \sup_{\alpha\in\Lambda}\{A^\alpha : D^2 v - f^\alpha\} \right| |\Delta z| \\
&\leq \int_\Omega \sup_{\alpha\in\Lambda} \left\{ \left| A^\alpha : D^2(u-v) \right| \right\} |\Delta z| \\
&\leq \sup_{\alpha\in\Lambda} \|A^\alpha\|_{\infty,\Omega} |u-v|_{H^2(\Omega)} \|\Delta z\|_{2,\Omega} \\
&\leq C \|u-v\|_{H^2(\Omega)} \|z\|_{H^2(\Omega)},
\end{aligned} \tag{3.3.23}$$

where the penultimate inequality follows from the fact that Λ is compact. Estimate (3.3.23) implies that $\mathscr{A}$ is Lipschitz continuous.

Since $\mathscr{A}$ is strongly monotone and Lipschitz continuous, it follows by the Browder–Minty theorem that there exists a unique $u \in H$ such that $\mathscr{A}(u) = 0$, i.e.,

$$\langle \mathscr{A}(u), v\rangle = \int_\Omega F_\gamma[u]\, \Delta v = 0 \quad \forall v \in H,$$

that is, (3.3.20) holds. $\square$

Corollary 3.3.9 (Existence and uniqueness) *Assume that H, and the collection of linear operators, $\{L^\alpha\}_{\alpha\in\Lambda}$, satisfy the assumptions of Theorem 3.3.8. Assume further, that $\Delta : H \to L^2(\Omega)$ is a surjection. Then, there exists a unique $u \in H$ such that*

$$F_\gamma[u] = 0 \quad \textit{a.e. in } \Omega. \tag{3.3.24}$$

Furthermore, u is the unique element of H that satisfies

$$F[u] = 0 \quad \textit{a.e. in } \Omega. \tag{3.3.25}$$

Proof: Since the assumptions of Theorem 3.3.8 are satisfied, we know that there exists a unique $u \in H$ such that

$$\int_\Omega F_\gamma[u]\, \Delta v = 0 \quad \forall v \in H.$$

Since $\Delta : H \to L^2(\Omega)$ is a surjection, we obtain

$$\int_\Omega F_\gamma[u]\, v = 0 \quad \forall v \in L^2(\Omega),$$

and thus $F_\gamma[u] = 0$ a.e. in Ω, which proves (3.3.24). We will now prove that this is equivalent to $F[u] = 0$ a.e. in Ω.

Since $\gamma^\alpha \geq \gamma_0 > 0$ for all $\alpha \in \Lambda$, we obtain

$$\gamma^\alpha (L^\alpha u - f^\alpha) \leq 0 \quad \forall \alpha \in \Lambda,$$

if and only if

$$L^\alpha u - f^\alpha \leq 0 \quad \forall \alpha \in \Lambda,$$

and thus

$$F_\gamma[u] \leq 0 \iff F[u] \leq 0.$$

Since Λ is compact, and the functions $A, f, \gamma \in C(\overline{\Omega} \times \Lambda)$, it follows that the suprema of $F[u]$ and $F_\gamma[u]$ are attained by an element of Λ, this yields

$$F_\gamma[u] \geq 0 \iff F[u] \geq 0.$$

Overall, we have obtained

$$F_\gamma[u] = 0 \iff F[u] = 0. \qquad \square$$

Remark 3.3.10 *Now that we have proven Theorem 3.3.8 and Corollary 3.3.9, it is clear that the unique solvability of the PDE (3.3.1) subject to particular boundary conditions, is equivalent to finding an appropriate space H that satisfies the following:*

1. *$v \in H$ implies that v satisfies the particular boundary condition;*
2. *$\Delta : H \to L^2(\Omega)$ is a surjection;*
3. *The space H satisfies the MT estimate property (3.3.9)–(3.3.10).*

3.3.2 The homogeneous Dirichlet case

As mentioned in Remark 3.3.2, the space $H := H^2(\Omega) \cap H_0^1(\Omega) \subset H^2(\Omega)$ satisfies Definition 3.3.1, if $\Omega \subset \mathbb{R}^d$ is convex. We will now provide a justification of this with the additional assumption that Ω has a C^2 boundary. It is proven in [110] (see Theorem 2), that the C^2 assumption may be relaxed, this is due to the fact that one can approximate a convex domain by a sequence of C^2 convex domains (see Section 3.2.1 of [60]). Moreover, we will see that the assumption that $\partial\Omega$ has nonnegative mean curvature is necessary to obtain a constant of 1 on the right-hand side of (3.3.9), which is an essential ingredient for the proof of Theorem 3.3.8. The following Theorem is from [60] (see Theorem 3.1.1.1).

Theorem 3.3.11 *Let Ω be a bounded open subset of $\mathbb{R}^d$ with a C^2 boundary, and let $v \in [H^1(\Omega)]^d$. Then, we have*

$$\sum_{i,j=1}^{d} \int_\Omega \frac{\partial v_i}{\partial x_j}\frac{\partial v_j}{\partial x_i} = \int_\Omega |\operatorname{div} v|^2 + \int_{\partial\Omega} (v_{\mathbf{T}})^T \nabla_{\mathbf{T}} n_{\partial\Omega}^T v_{\mathbf{T}} - \mathcal{H}_{\partial\Omega}(v\cdot n_{\partial K})^2 + 2\langle v_{\mathbf{T}} | \nabla_{\mathbf{T}}(v\cdot n_{\partial\Omega})\rangle_{(H^{1/2}(\partial\Omega)\times H^{1/2}(\partial\Omega)')}, \tag{3.3.26}$$

where $n_{\partial\Omega}$ is the unit outward normal to $\partial\Omega$, $v_{\mathbf{T}} := v - (v\cdot n_{\partial\Omega})n_{\partial\Omega}$, $\nabla_{\mathbf{T}} := \nabla - n_{\partial\Omega}\frac{\partial}{\partial n_{\partial\Omega}}$, and $\mathcal{H}_{\partial\Omega} := \nabla_{\mathbf{T}}\cdot n_{\partial\Omega}$ is the mean curvature of $\partial\Omega$.

We now deduce the following corollary.

Corollary 3.3.12 *Let Ω be a bounded open convex subset of $\mathbb{R}^d$ with a C^2 boundary, and let $u \in H^2(\Omega)\cap H_0^1(\Omega)$. Then, we have that*

$$|u|_{H^2(\Omega)} \le \|\Delta u\|_{L^2(\Omega)}. \tag{3.3.27}$$

Proof: Let $u \in H^2(\Omega)\cap H_0^1(\Omega)$, and let $v = \nabla u$. We then see that $v_{\mathbf{T}}|_{\partial\Omega} = \nabla_{\mathbf{T}} u|_{\partial\Omega} = 0$, since $u|_{\partial\Omega} = 0$. Substituting $v = \nabla u$ into (3.3.26), yields

$$|u|^2_{H^2(\Omega)} = \|\Delta u\|^2_{L^2(\Omega)} - \int_{\partial\Omega} \mathcal{H}_{\partial\Omega}\left(\frac{\partial u}{\partial n_{\partial\Omega}}\right)^2. \tag{3.3.28}$$

Since Ω is convex, it follows that $\mathcal{H}_{\partial\Omega} \ge 0$ on $\partial\Omega$, and so

$$|u|_{H^2(\Omega)} \le \|\Delta u\|_{L^2(\Omega)},$$

which is (3.3.27). $\square$

Remark 3.3.13 *Let us assume that Ω is a C^2 domain, with a boundary portion $\Gamma \subset \partial\Omega$, of nonzero boundary measure, for which the mean curvature $\mathcal{H}_{\partial\Omega}|_\Gamma < 0$. Taking a function $u \in H^2(\Omega)\cap H_0^1(\Omega)$ with $\frac{\partial u}{\partial n_{\partial\Omega}}|_{\partial\Omega} = g \in H^{1/2}(\partial\Omega)$, that satisfies $g > 0$ on Γ, $g = 0$ on $\partial\Omega\setminus\Gamma$, and substituting this function into (3.3.28) would lead to the following estimate*

$$\begin{aligned} |u|^2_{H^2(\Omega)} &= \|\Delta u\|^2_{L^2(\Omega)} - \int_{\partial\Omega} \mathcal{H}_{\partial\Omega}\left(\frac{\partial u}{\partial n_{\partial\Omega}}\right)^2 \\ &= \|\Delta u\|^2_{L^2(\Omega)} - \int_\Gamma \mathcal{H}_{\partial\Omega}\, g^2 > \|\Delta u\|^2_{L^2(\Omega)}, \end{aligned}$$

contradicting (3.3.28).

We now provide Theorem 2 from [110], which allows us to relax the C^2 regularity assumptions of Corollary 3.3.12.

Theorem 3.3.14 *Let Ω be a bounded open convex subset of $\mathbb{R}^d$, and let $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Then, we have that*

$$|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}. \tag{3.3.29}$$

We now show that the space $H^2(\Omega) \cap H_0^1(\Omega)$ also satisfies estimate (3.3.10), provided that Ω is convex.

Lemma 3.3.15 *Assume that $\Omega \subset \mathbb{R}^d$ is a convex domain. Then, there exists a constant C such that*

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2(\Omega)} \quad \forall u \in H^2(\Omega) \cap H_0^1(\Omega).$$

Proof: Let us take $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Firstly, by the Poincaré inequality [46], we see that there exists a constant C depending on Ω and the dimension d, but independent of u, such that

$$\|u\|_{L^2(\Omega)} \leq C|u|_{H^1(\Omega)}. \tag{3.3.30}$$

Furthermore, an application of integration by parts yields

$$\begin{aligned} |u|^2_{H^1(\Omega)} = \int_\Omega |\nabla u|^2 &= -\int_\Omega u\,\Delta u + \int_{\partial\Omega} \frac{\partial u}{\partial n_{\partial\Omega}} u \\ &\leq \|u\|_{L^2(\Omega)} \|\Delta u\|_{L^2(\Omega)} \\ &\leq C|u|_{H^1(\Omega)} \|\Delta u\|_{L^2(\Omega)}, \end{aligned}$$

where the final inequality follows from (3.3.30). Thus, it follows that

$$|u|_{H^1(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}.$$

From the above estimate, (3.3.30), and (3.3.29), we obtain

$$\begin{aligned} \|u\|^2_{H^2(\Omega)} &= \|u\|^2_{L^2(\Omega)} + |u|^2_{H^1(\Omega)} + |u|^2_{H^2(\Omega)} \\ &\leq (C+1)|u|^2_{H^1(\Omega)} + \|\Delta u\|^2_{L^2(\Omega)} \\ &\leq C\|\Delta u\|^2_{L^2(\Omega)}. \end{aligned}$$

Taking square roots above, we obtain (3.3.30). $\square$

Theorem 3.3.16 (Existence and uniqueness for the Dirichlet case) *Assume that $\Omega \subset \mathbb{R}^d$ is convex, and that the collection of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satisfies the assumptions of Theorem 3.3.8. Then, there exists a unique $u \in H := H^2(\Omega) \cap H_0^1(\Omega)$ such that*

$$\begin{cases} \sup\limits_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \quad \textit{a.e. in} \quad \Omega, \\ \qquad\qquad\qquad u = 0 \quad \textit{on} \quad \partial\Omega. \end{cases} \tag{3.3.31}$$

Proof: Let $H = H^2(\Omega) \cap H_0^1(\Omega)$, and assume that Ω is convex. We then see that $u \in H$ implies that $u|_{\partial\Omega} = 0$ (in the sense of traces), and so u satisfies the boundary condition. Then, Theorem 3.3.14 and Lemma 3.3.15 imply that H satisfies (3.3.9) and (3.3.10), respectively, i.e., the space H satisfies Definition 3.3.1. Finally, as Ω is convex, standard elliptic existence, uniqueness, and regularity theory tells us that $\Delta : H \to L^2(\Omega)$ is a surjection (see [60]).

Since the family of operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satisfy the assumptions of Theorem 3.3.8, i.e., they satisfy (3.3.6) and (3.3.12), we may apply Corollary 3.3.9 to deduce the existence of a unique $u \in H$ that satisfies (3.3.31). $\square$

Remark 3.3.17 (Convexity, and the surjectivity of the Laplacian) *As we have seen in the proof of Theorem 3.3.16, if $\Omega \subset \mathbb{R}^d$ is convex, then $H^2(\Omega) \cap H_0^1(\Omega)$ satisfies Definition 3.3.1. Furthermore, if the collection of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$, satisfy (3.3.9) and (3.3.10), Corollary 3.3.9 gives the existence and uniqueness of $u \in H^2(\Omega) \cap H_0^1(\Omega)$ that is a strong solution of*

$$F[u] = 0 \quad \text{a.e. in } \Omega, \tag{3.3.32}$$

provided that the Laplacian is a surjection from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$.

As discussed in Remark 3.3.13, if Ω is C^2, then, in order to satisfy Definition 3.3.1, $\partial\Omega$ must have nonnegative mean curvature. In the case that Ω is polytopal, Theorem 2.2.1 of [59] states that on any polygonal domain $\Omega \subset \mathbb{R}^2$, estimate (3.3.9) holds with equality, and Theorem 2.2.3 of [59] states that estimate (3.3.10) also holds (under the same domain assumptions), so one may consider that Theorem 3.3.16 holds on arbitrary polytopal domains (rather than only those that are convex).

However, the regularity of weak solutions of the Poisson problem on polygonal domains (see [59]) depends upon the size of the angles at each vertex. In particular, if one assumes that the polygonal domain $\Omega \subset \mathbb{R}^2$ has a largest corner angle of $\theta_{\max}$ (i.e., it has a corner of angle $\theta_{\max}$, and the angles of all of the other corners do not exceed $\theta_{\max}$), then a weak solution $u \in H_0^1(\Omega)$ of

$$\begin{cases} -\Delta u = f, & \text{a.e. in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{3.3.33}$$

with $f \in L^2(\Omega)$, belongs to $H^{\min\{\pi/\theta_{\max}+1-\delta,2\}}(\Omega)$, for arbitrary $\delta > 0$. In particular, if $\theta_{\max} > \pi$, then $\pi/\theta_{\max} < 1$, and u does not belong to $H^2(\Omega)$. This implies that the Laplacian is not a surjection from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$, for such a domain, and we cannot apply Corollary 3.3.9 to prove the existence and uniqueness of a strong

solution to (3.3.32). An example of such a domain is the L-shaped domain, which has $\theta_{\max} = 3\pi/2 > \pi$. *However, a convex polygonal domain must satisfy* $\theta_{\max} < \pi$, *and for such a domain, the Laplacian is a surjection from* $H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$ *(take the unit square, for example).*

3.3.3 The inhomogeneous Dirichlet case

We will now extend Theorem 3.3.16 to the case of inhomogeneous boundary data. In the case of well-posed, linear, uniformly elliptic Dirichlet boundary-value problems, one can view boundary data as playing a similar role to a source term; this is because we can define a new function, by extending the boundary data into Ω, and subtracting it from the supposed solution to the problem satisfying the inhomogeneous boundary data. Due to the linearity of the equation, assuming sufficient regularity of the boundary data, we arrive at a homogeneous boundary-value problem that is uniquely solvable.

In the case of nonlinear PDEs, this approach does not work in general, as when we substitute the new function into the equation we often end up with a different nonlinear equation, that is, the boundary data does not simply reduce to a modified source term. We will see that the particular structure of the HJB equation (3.3.1) allows for an analogue of the technique normally applied to uniformly elliptic, linear equations in the case of inhomogeneous boundary data.

However, this proof requires us to define matrix-valued functions using the property that they maximise the operator defined by (3.3.17). A function $u \in H^2(\Omega)$ defines a matrix-valued function $\mathbf{M} \in L^2(\Omega; \mathbb{R}^{d\times d})$ through $\mathbf{M} = D^2u$. Let $\mathbf{M} \in \mathbb{R}^{d\times d}$, and define

$$F_\gamma(x, \mathbf{M}) := \sup_{\alpha\in\Lambda}\{\gamma^\alpha(A^\alpha : \mathbf{M} - f^\alpha)|_x\}. \tag{3.3.34}$$

For each $(x, \mathbf{M}) \in \Omega \times \mathbb{R}^{d\times d}$, we define

$$\Lambda(x, \mathbf{M}) := \{\alpha \in \Lambda \text{ such that the supremum in (3.3.34) is attained}\}. \tag{3.3.35}$$

This defines a set-valued map $(x, \mathbf{M}) \mapsto \Lambda(x, \mathbf{M})$. For $u \in H^2(\Omega)$, let

$$\begin{aligned}\Lambda[u] := \{\alpha : \Omega \to \Lambda, \text{ Lebesgue measurable} : \alpha(x) \in \Lambda(x, \mathbf{M}(x)) \\ \text{for a.e. } x \in \Omega, \text{ where } \mathbf{M} = D^2u\}.\end{aligned} \tag{3.3.36}$$

Then, due to the following lemma quoted from [111] and theorem, available in [75], we may select a Lebesgue measurable function $\alpha \in \Lambda[u]$, and use this to define the corresponding matrix-valued function A^α.

Lemma 3.3.18 *Let Ω be a bounded open subset of $\mathbb{R}^d$, let Λ be a compact metric space, and let the data A, f be continuous on $\overline{\Omega} \times \Lambda$, and suppose that (3.3.6) holds. Then, for each $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, $\Lambda(x, \mathbf{u})$ is a non-empty subset of Λ. The set-valued map $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$ is upper semicontinuous; that is, for every $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, and any open neighbourhood U of $\Lambda(x, \mathbf{u})$, there exists an open neighbourhood V of $(x, \mathbf{u})$ such that $\Lambda(y, \mathbf{v}) \subset U$ for every $(y, \mathbf{v}) \in V$.*

Theorem 3.3.19 *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set, let Λ be a compact metric space, and let $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$ be an upper semicontinuous set-valued function from $\Omega \times \mathbb{R}^m$ to the subsets of Λ, such that $\Lambda(x, \mathbf{u})$ is nonempty and closed for every $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$. Then, for any Lebesgue measurable function $\mathbf{u} : \Omega \to \mathbb{R}^m$, there exists a Lebesgue measurable selection $\alpha : \Omega \to \Lambda$ such that $\alpha(x) \in \Lambda(x, \mathbf{u}(x))$ for a.e. $x \in \Omega$.*

Theorem 3.3.20 *Assume that $\Omega \subset \mathbb{R}^d$ is a convex domain, and that the collection of linear operators $\{L^\alpha\}_{\alpha \in \Lambda}$ satisfies the assumptions of Theorem 3.3.8. Furthermore assume that g is the trace of a $H^2(\Omega)$ function $\tilde{g}$. Then, there exists a unique strong solution $u \in H^2(\Omega)$ of the following HJB equation:*

$$\begin{aligned} \sup_{\alpha \in \Lambda} \{L^\alpha u - f^\alpha\} = 0 \quad &\text{a.e. in } \Omega, \\ u = g \quad &\text{on } \partial\Omega. \end{aligned} \tag{3.3.37}$$

Proof: Consider the following HJB problem: find $\tilde{u} \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying

$$\begin{aligned} \sup_{\alpha \in \Lambda} \{L^\alpha \tilde{u} - f^\alpha + L^\alpha \tilde{g}\} = 0 \quad &\text{a.e. in } \Omega, \\ u = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Notice that

$$L^\alpha \tilde{g} = \sum_{i,j=1}^{d} A_{ij}^\alpha D_{ij}^2 \tilde{g}.$$

Defining $\tilde{f^\alpha} := f^\alpha - L^\alpha \tilde{g} \in L^2(\Omega)$, we can apply Theorem 3.3.16 to deduce the existence and uniqueness of $\tilde{u} \in H^2(\Omega) \cap H_0^1(\Omega)$, for any given $\tilde{g} \in H^2(\Omega)$. We now define $u := \tilde{u} + \tilde{g}$, and we see that

$$\sup_{\alpha \in \Lambda} \{L^\alpha u - f^\alpha\} = \sup_{\alpha \in \Lambda} \{L^\alpha \tilde{u} - f^\alpha + L^\alpha \tilde{g}\} = \sup_{\alpha \in \Lambda} \{L^\alpha \tilde{u} - \tilde{f^\alpha}\} = 0, \tag{3.3.38}$$

and that

$$u|_{\partial\Omega} = \tilde{u}|_{\partial\Omega} + \tilde{g}|_{\partial\Omega} = g.$$

Thus there exists a $u \in H^2(\Omega)$ that satisfies (3.3.37).

We must now show that such a u is unique. Assume that there exist $u_1,\ u_2 \in H^2(\Omega)$ that satisfy (3.3.37). Now for $i = 1, 2$, by Lemma 3.3.18 and Theorem 3.3.19 there exists a measurable function $\alpha_i : \Omega \to \Lambda$ such that $\alpha_i(x) \in \Lambda$ for a.e. $x \in \Omega$, and

$$\alpha_i(x) = \{\text{argsup}_{\alpha\in\Lambda}\{L^\alpha u_i - f^\alpha\}\}(x).$$

Note that the compactness of Λ and the continuity of A, f, and γ imply the existence of a maximiser. We then define the operators $L^{\alpha_i} : H^2(\Omega) \to L^2(\Omega)$ as follows

$$\{L^{\alpha_i} v\}(x) := L^{\alpha_i(x)} v(x) \quad \text{a.e. } x \in \Omega.$$

Note that we define the functions f^{α_i}, γ^{α_i} similarly. By the definition of the supremum, it is clear that

$$\begin{aligned} &L^{\alpha_i} u_i - f^{\alpha_i} = 0 \quad \text{a.e. in } \Omega, \quad i = 1, 2, \\ &L^{\alpha_i} u_j - f^{\alpha_i} \le 0 \quad \text{a.e. in } \Omega, \quad i, j = 1, 2,\ i \ne j. \end{aligned} \tag{3.3.39}$$

From this we see that

$$L^{\alpha_1}(u_1 - u_2) = f^{\alpha_1} - L^{\alpha_1} u_2 \ge f^{\alpha_1} - f^{\alpha_1} = 0,$$

and

$$L^{\alpha_2}(u_2 - u_1) = f^{\alpha_2} - L^{\alpha_2} u_1 \ge f^{\alpha_2} - f^{\alpha_2} = 0.$$

Let us denote $w = u_1 - u_2 \in H^2(\Omega) \cap H_0^1(\Omega)$; then, the above gives us $L^{\alpha_1} w \ge 0$ and $L^{\alpha_2} w \le 0$ a.e. in Ω, and so

$$\begin{aligned} 0 &\ge \int_{\{\Delta w \le 0\}} \gamma^{\alpha_1} L^{\alpha_1} w \Delta w + \int_{\{\Delta w > 0\}} \gamma^{\alpha_2} L^{\alpha_2} w \Delta w \\ &= \|\Delta w\|^2_{L^2(\Omega)} + \int_{\{\Delta w \le 0\}} (\gamma^{\alpha_1} L^{\alpha_1} w - \Delta w)\Delta w + \int_{\{\Delta w > 0\}} (\gamma^{\alpha_2} L^{\alpha_2} w - \Delta w)\Delta w \\ &\ge \|\Delta w\|^2_{L^2(\Omega)} - \sqrt{1-\varepsilon} \int_\Omega |D^2 w||\Delta w| \\ &\ge \|\Delta w\|^2_{L^2(\Omega)} - \sqrt{1-\varepsilon} |w|_{H^2(\Omega)} \|\Delta w\|_{L^2(\Omega)} \\ &\ge (1 - \sqrt{1-\varepsilon}) \|\Delta w\|^2_{L^2(\Omega)} \\ &\ge C \|w\|^2_{H^2(\Omega)}. \end{aligned} \tag{3.3.40}$$

Note that the second inequality follows from (3.3.19), and the last two inequalities follow from the fact that $H := H^2(\Omega) \cap H_0^1(\Omega)$ satisfies the MT estimate property (3.3.9)–(3.3.10). It is now clear that $w \equiv 0$, i.e., $u_1 \equiv u_2$. $\square$

3.3.4 The oblique case

Our next aim is to prove a similar existence and uniqueness result for the oblique boundary-value problem in the planar case, where $\Omega \subset \mathbb{R}^2$ is a C^2 domain. The oblique boundary-value problem is given as follows:

$$\begin{cases} \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \quad \text{a.e. in } \Omega, \\ \qquad\qquad \beta \cdot \nabla u \quad \text{is constant on } \partial\Omega. \end{cases} \tag{3.3.41}$$

Such types of equations arise in the linearisation of PDEs with nonlinear boundary conditions (for example the MA optimal transport problem (3.2.2)). The unit vector-valued function

$$\beta \in C^1(\partial\Omega; \mathbb{S}^1) \tag{3.3.42}$$

is hereby called the "oblique vector". Notice that since Ω is a C^2 domain, $\beta := n_{\partial\Omega}$, where $n_{\partial\Omega}$ denotes the unit outward normal to $\partial\Omega$, is such an example, since then $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$, and so (3.3.42) holds. Furthermore, the corresponding PDE is

$$\begin{cases} \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \quad \text{a.e. in } \Omega, \\ \qquad\qquad \dfrac{\partial u}{\partial n_{\partial\Omega}} \quad \text{is constant on } \partial\Omega, \end{cases} \tag{3.3.43}$$

i.e., a Neumann boundary-value problem, where the normal derivative of the solution is prescribed on the boundary. Furthermore, taking the particular case that Λ is a singleton set, and $L^\alpha u = \Delta u$, we obtain the Poisson problem with a Neumann boundary condition

$$\begin{cases} \Delta u = f \quad \text{a.e. in } \Omega, \\ \dfrac{\partial u}{\partial n_{\partial\Omega}} \quad \text{is constant on } \partial\Omega. \end{cases} \tag{3.3.44}$$

If we were to impose the more familiar boundary condition

$$\frac{\partial u}{\partial n_{\partial\Omega}} = g \quad \text{on } \partial\Omega,$$

for some given function $g \in H^{1/2}(\partial\Omega)$, then the boundary datum, g, and the right-hand side f would have to satisfy a compatibility condition, emerging from an application of the divergence theorem. The choice of boundary condition in (3.3.44) (where $g \equiv 0$) absorbs this compatibility condition.

We will see later on that the condition

$$\beta \cdot \nabla u \quad \text{is constant on } \partial\Omega,$$

removes the complication of eventual compatibility conditions for the oblique boundary-value problem, and also allows for $\Delta : H \to L^2(\Omega)$ to be a surjection for the corresponding space H (which we will define momentarily).

As we saw in the proof of Theorem 3.3.8, estimates (3.3.9) and (3.3.10) play an essential role, allowing us to obtain estimates (3.3.21) and (3.3.22). In the case that $H = H^2(\Omega) \cap H_0^1(\Omega)$, with $\Omega \subset \mathbb{R}^d$ convex, Theorem 3.3.14 and Lemma 3.3.15 prove that estimates (3.3.9) and (3.3.10) hold for any dimension d. For the oblique boundary-value problem (3.3.41), we will consider the space

$$\begin{aligned} H &= \{v \in H^2(\Omega) : \beta \cdot \nabla v|_{\partial\Omega} \text{ is constant}\} \cap L_0^2(\Omega) \\ &= H_{\beta,0}^2(\Omega), \end{aligned} \tag{3.3.45}$$

where $\Omega \subset \mathbb{R}^2$ is assumed to have a C^2 boundary. When $H = H_{\beta,0}^2(\Omega)$, the proof of the Miranda–Talenti estimates (3.3.9) and (3.3.10) that we provide relies on the assumption that the dimension $d = 2$, leading us to consider the planar case (the proof of these estimates for $d \geq 3$ is currently an open problem). The significance of the requirement that $d = 2$ is demonstrated by the following identities, valid for $u \in H_{\beta,0}^2(\Omega)$:

$$\begin{aligned} |u|_{H^2(\Omega)}^2 + 2\int_\Omega \det D^2 u &= \int_\Omega \sum_{i,j=1}^{d} (D_{ij}^2 u)^2 + 2(D_{11}^2 u D_{22}^2 u - (D_{12}^2 u)^2) \\ &= \int_\Omega (D_{11}^2 u + D_{22}^2 u)^2 = \|\Delta u\|_{L^2(\Omega)}^2, \end{aligned} \tag{3.3.46}$$

and

$$\int_\Omega \det D^2 u = \int_\Omega D_{11}^2 u D_{22}^2 u - (D_{12}^2 u)^2 = \frac{1}{2}\int_{\partial\Omega} |\nabla u|^2 (\beta_1(\partial_{\mathbf{T}_2}\beta_2) - (\partial_{\mathbf{T}_2}\beta_1)\beta_2), \tag{3.3.47}$$

where $\partial_{\mathbf{T}_2} := \mathbf{T}_2 \cdot \nabla$, and $\mathbf{T}_2 := (-[n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1)^T$ is the unit tangent vector to $\partial\Omega$ obtained by rotating the unit outward normal vector, $n_{\partial\Omega}$, anticlockwise by $\pi/2$.

Identity (3.3.46) is an algebraic identity that holds due to the fact that the dimension $d = 2$, and so $D^2 u \in \mathbb{R}^{2\times 2}_{\text{Sym}}$ a.e. in Ω. Furthermore, the identity (3.3.47) is proven in [89] (see Lemma 1.5.5), and also relies on the fact that $d = 2$. Upon showing (under certain assumptions that will be provided) that the right-hand side of (3.3.47) is nonpositive, we obtain (3.3.9) directly from (3.3.46).

We will prove (3.3.47) as a consequence of a more general result, which we will obtain using a similar approach to that of [89], applied to the "determinant-type" bilinear form $\mathscr{B} : H^2(\Omega) \times H^2(\Omega) \to \mathbb{R}$ defined by

$$\mathscr{B}(u, v) := \int_\Omega D_{11}^2 u D_{22}^2 v + D_{22}^2 u D_{11}^2 v - 2 D_{12}^2 u D_{12}^2 u.$$

Indeed, one can see that for $u \in H^2(\Omega)$,

$$\mathscr{B}(u,u) = \int_\Omega D^2_{11}uD^2_{22}u + D^2_{22}uD^2_{11}u - 2D^2_{12}uD^2_{12}u = 2\int_\Omega \det(D^2u).$$

Furthermore, this result will guide the design of the numerical method we propose in Chapter 6, and as such the result will need to also be applicable to an arbitrary element of our triangulation. The elements of our triangulation will always be assumed to be Lipschitz continuous and at least piecewise C^2, which provides us with the domain hypotheses of the following Lemma.

Lemma 3.3.21 *Assume that $E \subset \mathbb{R}^2$ is a bounded, Lipschitz, piecewise C^2 domain, and that $\beta \in C^1(\Gamma_n;\mathbb{S}^1)$ for each C^2 portion Γ_n of ∂E, $n = 1,\dots,N$, $N \in \mathbb{N}$. Then, for any $u, v \in H^s(E)$, $s > 5/2$, we have that*

$$\begin{aligned}\int_E & D^2_{11}uD^2_{22}v + D^2_{22}uD^2_{11}v - 2D^2_{12}uD^2_{12}v \\ &= \int_{\partial E} (\beta_1\partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1)\,(\beta^\perp\cdot\nabla u\,\beta^\perp\cdot\nabla v + \beta\cdot\nabla u\,\beta\cdot\nabla v) \\ &\quad + \int_{\partial E} \left(\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u)\beta\cdot\nabla v - \partial_{\mathbf{T}_2}(\beta\cdot\nabla u)\beta^\perp\cdot\nabla v\right),\end{aligned} \tag{3.3.48}$$

where $\beta^\perp := (-\beta_2,\beta_1) \in C^1(\Gamma_n;\mathbb{S}^1)$ for each C^2 portion Γ_n of ∂E, $n = 1,\dots,N$, $N \in \mathbb{N}$.

Proof: Let us momentarily assume that $u, v \in C^2(\overline{E}) \cap C^3(E)$, and note that an application of integration by parts gives us

$$\begin{aligned}\int_E & D^2_{11}uD^2_{22}v + D^2_{22}uD^2_{11}v - 2D^2_{12}uD^2_{12}v \\ &= \int_E D_1(D_1vD^2_{22}u - D_2vD^2_{12}u) - D_2(D_1vD^2_{21}u - D_2vD^2_{11}u) \\ &= \int_{\partial E} (D_1vD^2_{22}u - D_2vD^2_{12}u)n_1 - (D_1vD^2_{21}u - D_2vD^2_{11}u)n_2.\end{aligned} \tag{3.3.49}$$

Now, denoting $C^1_u := \beta^\perp\cdot\nabla u$, $C^1_v := \beta^\perp\cdot\nabla v$, $C^2_u := \beta\cdot\nabla u$, and $C^2_v := \beta\cdot\nabla u$, we obtain the following linear systems on Γ_n, $n = 1,\dots,N$:

$$\begin{cases} \beta_1D_1u + \beta_2D_2u = C^2_u, \\ -\beta_2D_1u + \beta_1D_2u = C^1_u, \end{cases} \tag{3.3.50}$$

$$\begin{cases} \beta_1D_1v + \beta_2D_2v = C^2_v, \\ -\beta_2D_1v + \beta_1D_2v = C^1_v, \end{cases} \tag{3.3.51}$$

with corresponding unique solutions

$$\begin{cases} D_1u = -C_u^1\beta_2 + C_u^2\beta_1, \quad D_2u = C_u^1\beta_1 + C_u^2\beta_2, \\ D_1v = -C_v^1\beta_2 + C_v^2\beta_1, \quad D_2v = C_v^1\beta_1 + C_v^2\beta_2. \end{cases} \tag{3.3.52}$$

Substituting in the values of D_1v and D_2v present in (3.3.52) into (3.3.49), (denoting $n_i = [n_{\partial\Omega}]^i$, $i = 1, 2$) we obtain the following:

$$\begin{aligned} &\int_E D_{11}^2 u D_{22}^2 v + D_{22}^2 u D_{11}^2 v - 2D_{12}^2 u D_{12}^2 v \\ &\quad = \int_{\partial E} -C_v^1\beta_2 D_{22}^2 u\, n_1 + C_v^2\beta_1 D_{22}^2 u\, n_1 - C_v^1\beta_1 D_{12}^2 u\, n_1 - C_v^2\beta_2 D_{12}^2 u\, n_1 \\ &\qquad + \int_{\partial E} C_v^1\beta_2 D_{21}^2 u\, n_2 - C_v^2\beta_1 D_{21}^2 u\, n_2 + C_v^1\beta_1 D_{11}^2 n_2 + C_v^2\beta_2 D_{11}^2 u\, n_2 \\ &\quad = \int_{\partial E} C_v^1(-\beta_2 D_{22}^2 u\, n_1 - \beta_1 D_{12}^2 u\, n_1 + \beta_2 D_{21}^2 u\, n_2 + \beta_1 D_{11}^2 u\, n_1) \\ &\qquad + \int_{\partial E} C_v^2(\beta_1 D_{22}^2 u\, n_1 - \beta_2 D_{12}^2 u\, n_1 - \beta_1 D_{21}^2 n_2 + \beta_2 D_{11}^2 u\, n_2). \end{aligned} \tag{3.3.53}$$

Taking the directional derivative of the equations in (3.3.50) with respect to $\mathbf{T}$ yields (denoting $\dot{} := \partial_{\mathbf{T}_2}$)

$$\begin{cases} \dot\beta_1 D_1u + \beta_1 D_{11}^2 u(-n_2) + \beta_1 D_{12}^2 u(n_1) + \dot\beta_2 D_2u + \beta_2 D_{12}^2 u(-n_2) + \beta_2 D_{22}^2 u(n_1) = \dot C_u^2, \\ -\dot\beta_2 D_1u - \beta_2 D_{11}^2 u(-n_2) - \beta_2 D_{12}^2 u(n_1) + \dot\beta_1 D_2u + \beta_1 D_{12}^2 u(-n_2) + \beta_1 D_{22}^2 u(n_1) = \dot C_u^1, \end{cases}$$

thus,

$$\begin{cases} -(\beta_1 D_{11}^1 u\, n_2 - \beta_1 D_{12}^2 u\, n_1 + \beta_2 D_{12}^2 u\, n_2 - \beta_2 D_{22}^2 u\, n_1) = \dot C_u^2 - (\dot\beta_1 D_1u + \dot\beta_2 D_2u), \\ (\beta_1 D_{22}^2 u\, n_1 - \beta_2 D_{12}^2 u\, n_1 - \beta_1 D_{12}^2 u\, n_2 + \beta_2 D_{11}^2 u\, n_2) = \dot C_u^1 + \dot\beta_2 D_1u - \dot\beta_1 D_2u. \end{cases} \tag{3.3.54}$$

Substituting the values for D_1u and D_2u present in (3.3.52) gives us

$$\begin{aligned} \dot\beta_1 D_1u + \dot\beta_2 D_2u &= -C_u^1\beta_2\dot\beta_1 + C_u^2\beta_1\cdot\beta_1 + C_u^1\beta_1\dot\beta_2 + C_u^2\beta_2\dot\beta_2 \\ &= C_u^1(\beta_1\dot\beta_2 - \beta_2\dot\beta_1) + C_u^2(\beta_1\dot\beta_1 + \beta_2\dot\beta_2) \\ &= C_u^1(\beta_1\dot\beta_2 - \beta_2\dot\beta_1); \end{aligned} \tag{3.3.55}$$

note that the latter equality holds, since

$$\beta_1\dot\beta_1 + \beta_2\dot\beta_2 = \frac{1}{2}\partial_{\mathbf{T}_2}(|\beta|^2) = \frac{1}{2}\partial_{\mathbf{T}_2}(1) = 0.$$

Similarly, we obtain

$$\dot\beta_2 D_1u - \dot\beta_1 D_2u = C_u^2(\beta_1\dot\beta_2 - \beta_2\dot\beta_1). \tag{3.3.56}$$

Substituting (3.3.55) and (3.3.56) into (3.3.54), and then substituting the resulting equations into (3.3.53), we obtain

$$\begin{aligned}
&\int_E D_{11}^2 u D_{22}^2 v + D_{22}^2 u D_{11}^2 v - 2D_{12}^2 u D_{12}^2 v \\
&\qquad + \int_{\partial E} C_v^1 (C_u^1(\beta_1\dot\beta_2 - \beta_2\dot\beta_1) - \dot C_u^1) + C_v^2 (C_u^2(\beta_1\dot\beta_2 - \beta_2\dot\beta_1) + \dot C_u^2) \\
&= \int_{\partial E} (\beta_1\dot\beta_2 - \beta_2\dot\beta_1)(C_u^1 C_v^1 + C_u^2 C_v^2) + \dot C_u^1 C_v^2 - \dot C_u^2 C_v^1 \\
&= \int_{\partial E} (\beta_1 \partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1)\,(\beta^\perp\cdot\nabla u\,\beta^\perp\cdot\nabla v + \beta\cdot\nabla u\,\beta\cdot\nabla v) \\
&\qquad + \int_{\partial E} \big(\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u)\beta\cdot\nabla v - \partial_{\mathbf{T}_2}(\beta\cdot\nabla u)\beta^\perp\cdot\nabla v\big)\,,
\end{aligned}$$

which is exactly (3.3.48). This identity extends to $u, v \in H^s(E)$, $s > 5/2$, by density, which concludes the proof. $\square$

We now provide (3.3.47) (i.e., Lemma 1.5.5 of [89]) as a consequence of Lemma 3.3.21.

Corollary 3.3.22 *Assume that $\Omega \subset \mathbb{R}^2$ is a C^2 domain, and that $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$. Then, for any $u \in H^2_{\beta,0}(\Omega)$, identity (3.3.47) holds.*

Proof: First, we note that since $\Omega \subset \mathbb{R}^2$ is a C^2 domain, it is both Lipschitz continuous and piecewise C^2. Furthermore, as $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$, it follows that $\beta \in C^1(\Gamma_{\partial\Omega};\mathbb{S}^1)$ for each C^2 portion $\Gamma_{\partial\Omega}$ of $\partial\Omega$; indeed one may take $\Gamma_{\partial\Omega} = \partial\Omega$. Now, let us assume that $u \in C^3(\overline{\Omega}) \cap H^2_{\beta,0}(\Omega)$, so that $u \in H^s(\Omega)$, with $s > 5/2$. Setting $v = u$, it then follows that the hypotheses of Lemma 3.3.21 are satisfied, and so (3.3.48) gives us

$$\begin{aligned}
\mathscr{B}(u,u) &= \int_{\partial\Omega} (\beta_1 \partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1)\,(\beta^\perp\cdot\nabla u\,\beta^\perp\cdot\nabla u + \beta\cdot\nabla u\,\beta\cdot\nabla u) \\
&\qquad + \int_{\partial\Omega} \big(\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u)\beta\cdot\nabla u - \partial_{\mathbf{T}_2}(\beta\cdot\nabla u)\beta^\perp\cdot\nabla u\big) \\
&= \int_{\partial\Omega} |\nabla u|^2\,(\beta_1\partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1) + C\int_{\partial\Omega} \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u),
\end{aligned}$$

where the latter equality holds due to the fact that $\beta\cdot\nabla u|_{\partial\Omega} = C$ for some constant C. Then, since $\partial\Omega$ is a compact hypersurface, integrating yields

$$C\int_{\partial\Omega} \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u) = 0.$$

Thus, we have obtained

$$\frac{1}{2}\mathscr{B}(u,u) = \frac{1}{2}\int_{\partial\Omega} |\nabla u|^2\,(\beta_1\partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1)\,,$$

which is (3.3.47). This extends to $u \in H^2_{\beta,0}(\Omega)$ by density. $\square$

Corollary 3.3.23 *Assume that $E \subset \mathbb{R}^2$ is a bounded, Lipschitz, piecewise C^2 domain, and that $\beta \in C^1(\Gamma_n; \mathbb{S}^1)$ for each C^2 portion Γ_n of ∂E, $n = 1, \dots, N$, $N \in \mathbb{N}$. Then, for any $u, v \in H^s(E)$, $s > 5/2$, we have that*

$$\begin{aligned}\int_E D^2u : D^2v + \int_{\partial E} (\beta_1 \partial_{\mathbf{T}_2} \beta_2 - \beta_2 \partial_{\mathbf{T}_2} \beta_1) \, (\beta^\perp \cdot \nabla u \, \beta^\perp \cdot \nabla v + \beta \cdot \nabla u \, \beta \cdot \nabla v) \\ + \int_{\partial E} \big(\partial_{\mathbf{T}_2} (\beta^\perp \cdot \nabla u) \beta \cdot \nabla v - \partial_{\mathbf{T}_2} (\beta \cdot \nabla u) \beta^\perp \cdot \nabla v \big) \\ = \int_E \Delta u \, \Delta v.\end{aligned}$$

Proof: First note that for $u, v \in H^2(E)$,

$$D^2u : D^2v + D^2_{11}u D^2_{22}v + D^2_{22}u D^2_{11}v - 2D^2_{12}u D^2_{12}v = \Delta u \, \Delta v \quad \text{a.e. in } \Omega,$$

and apply Lemma 3.3.21. $\square$

Definition 3.3.24 (Oblique angle) *We define the "oblique angle", $\Theta : \partial\Omega \to \mathbb{R}$, to be the (anticlockwise) oriented angle between the oblique vector, β, and the unit outward normal, $n_{\partial\Omega}$ (see Figure 3.1). In particular, one has that*

$$\Theta \pmod{2\pi} = \pi + \arctan2(\beta_1 [n_{\partial\Omega}]^2 - \beta_2 [n_{\partial\Omega}]^1, \beta \cdot n_{\partial\Omega})),$$

where

$$\arctan2(y, x) := \begin{cases} \arctan(y/x) & \text{if} \quad x > 0, \\ \arctan(y/x) + \pi & \text{if} \quad x < 0 \text{ and } y \geq 0, \\ \arctan(y/x) - \pi & \text{if} \quad x < 0 \text{ and } y < 0, \\ \dfrac{\pi}{2} & \text{if} \quad x = 0 \text{ and } y > 0, \\ -\dfrac{\pi}{2} & \text{if} \quad x = 0 \text{ and } y < 0. \end{cases}$$

Note that $\arctan2(y, x)$ *is undefined for $x = y = 0$.*

Our next goal is to prove an identity for the expression $\beta_1(\partial_{\mathbf{T}_2}\beta_2) - (\partial_{\mathbf{T}_2}\beta_1)\beta_2$ that appears on the right-hand side of (3.3.46). We will express this in terms of the derivative of the oblique angle, $\partial_{\mathbf{T}_2}\Theta$, and the mean curvature of the boundary, $\mathcal{H}_{\partial\Omega}$. This provides us with a condition between the curvature of the boundary, and the oblique vector that is sufficient to yield the MT estimates (3.3.9) and (3.3.10). This identity is provided in [89] pg 48-49, employing a C^2 parametrisation of the boundary with respect to the natural parameter $\varphi \in [0, L]$ (i.e., the parametrisation $(x_1(\varphi), x_2(\varphi)) \in C^2[0, L]$ is of unit speed with respect to the parameter φ, that is, $(x_1'(\varphi), x_2'(\varphi)) : [0, L] \mapsto \mathbb{S}^1$). We, however, prove this result without reparametrising

the boundary. Furthermore, it is useful for the reader to see how this identity arises.

Lemma 3.3.25 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 domain, and assume that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Then, on $\partial\Omega$, we have that*

$$\beta_1(\partial_{\mathbf{T}_2}\beta_2) - (\partial_{\mathbf{T}_2}\beta_1)\beta_2 = \partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega}, \tag{3.3.57}$$

where $\mathcal{H}_{\partial\Omega}$ is the mean curvature of $\partial\Omega$.

Proof: We first define the functions $\psi, \omega \in C^1(\partial\Omega)$ that satisfy

$$[\mathbf{T}_2]^1 = \cos\psi, \quad [\mathbf{T}_2]^2 = \sin\psi, \tag{3.3.58}$$

$$\beta_1 = \cos\omega, \quad \beta_2 = \sin\omega. \tag{3.3.59}$$

That is, $\psi(x_1, x_2)$ is the (anticlockwise) oriented angle between the x_1-axis and the tangent, $\mathbf{T}_2(x_1, x_2)$, to $\partial\Omega$ at the point (x_1, x_2), and $\omega(x_1, x_2)$ is the (anticlockwise) oriented angle between the x_1-axis and $\beta(x_1, x_2)$ at the point (x_1, x_2) (see Figure 3.1 for a visualisation of these angles). This provides us with the identity

$$\psi = \omega - \Theta + \frac{\pi}{2} \quad \text{on } \partial\Omega. \tag{3.3.60}$$

Furthermore, taking the directional derivative of all of the identities in (3.3.58) and (3.3.59) with respect to $\mathbf{T}_2$, yields

$$\begin{aligned} \partial_{\mathbf{T}_2}[\mathbf{T}_2]^1 &= -\sin\psi\,\partial_{\mathbf{T}_2}\psi, \quad \partial_{\mathbf{T}_2}[\mathbf{T}_2]^2 = \cos\psi\,\partial_{\mathbf{T}_2}\psi, \\ \partial_{\mathbf{T}_2}\beta_1 &= -\sin\omega\,\partial_{\mathbf{T}_2}\omega, \quad \partial_{\mathbf{T}_2}\beta_2 = \cos\omega\,\partial_{\mathbf{T}_2}\omega. \end{aligned}$$

It then follows that

$$\begin{aligned} \partial_{\mathbf{T}_2}\psi &= \cos^2\psi\,\partial_{\mathbf{T}_2}\psi + \sin^2\psi\,\partial_{\mathbf{T}_2}\psi \\ &= \cos\psi\,\partial_{\mathbf{T}_2}[\mathbf{T}_2]^2 - \sin\psi\,\partial_{\mathbf{T}_2}[\mathbf{T}_2]^1 \\ &= [\mathbf{T}_2]^1\,\partial_{\mathbf{T}_2}[n_{\partial\Omega}]^1 + [\mathbf{T}_2]^2\,\partial_{\mathbf{T}_2}[n_{\partial\Omega}]^2 \\ &= \mathbf{T}_2 \cdot \partial_{\mathbf{T}_2} n_{\partial\Omega}, \end{aligned} \tag{3.3.61}$$

and

$$\begin{aligned} \partial_{\mathbf{T}_2}\omega &= \cos^2\omega\,\partial_{\mathbf{T}_2}\omega + \sin^2\omega\,\partial_{\mathbf{T}_2}\omega \\ &= \cos\omega\,\partial_{\mathbf{T}_2}\beta_2 - \sin\omega\,\partial_{\mathbf{T}_2}\beta_1 \\ &= \beta_1\,\partial_{\mathbf{T}_2}\beta_2 - \beta_2\,\partial_{\mathbf{T}_2}\beta_1. \end{aligned} \tag{3.3.62}$$

Now, by definition $\mathcal{H}_{\partial\Omega} = \nabla_{\mathbf{T}} \cdot n_{\partial\Omega} = \mathrm{Tr}(\nabla_{\mathbf{T}} n_{\partial\Omega}^T) = \mathrm{Tr}(\mathcal{W})$, where we recall that $\mathcal{W}$ is the Weingarten map, defined by $\mathcal{W} := \nabla_{\mathbf{T}} n_{\partial\Omega}^T$. Furthermore, since $\mathcal{H}_{\partial\Omega}$ is the

trace of $\mathcal{W}$, it is equal to the sum of the eigenvalues of $\mathcal{W}$, which has one trivial eigenvalue of zero, with a corresponding unit eigenvector, $n_{\partial\Omega}$. Since $d = 2$, we must have that $\mathcal{H}_{\partial\Omega} = \kappa_1$, where κ_1 is the remaining nontrivial eigenvalue of $\mathcal{W}$, and that any corresponding eigenvector must be a scalar multiple of $\mathbf{T}_2$. In particular, a corresponding unit eigenvector is given by $\mathbf{T}_2$. I.e.,

$$\mathcal{W}\mathbf{T}_2 = \mathcal{H}_{\partial\Omega}\mathbf{T}_2.$$

Taking an inner product with $\mathbf{T}_2$ above, we find that

$$\mathcal{H}_{\partial\Omega} = \mathbf{T}_2^T \, \mathcal{W} \, \mathbf{T}_2 = \mathbf{T}_2^T \, \nabla_{\mathbf{T}} n_{\partial\Omega}^T \, \mathbf{T}_2.$$

We then calculate

$$\begin{aligned}
\mathcal{H}_{\partial\Omega} &= \mathbf{T}_2^T \, \nabla_{\mathbf{T}} n_{\partial\Omega}^T \, \mathbf{T}_2 \\
&= \mathbf{T}_2^T \, \nabla n_{\partial\Omega}^T \, \mathbf{T}_2 \\
&= \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}([n_{\partial\Omega}]^j)[\mathbf{T}_2]^i[\mathbf{T}_2]^j \\
&= \sum_{j=1}^{2}[\mathbf{T}_2]^j \sum_{i=1}^{d} \frac{\partial}{\partial x_i}([n_{\partial\Omega}]^j)[\mathbf{T}_2]^i \qquad (3.3.63) \\
&= \sum_{j=1}^{2}[\mathbf{T}_2]^j \partial_{\mathbf{T}_2}([n_{\partial\Omega}]^j) \\
&= \mathbf{T}_2 \cdot \partial_{\mathbf{T}_2} n_{\partial\Omega} \\
&= \partial_{\mathbf{T}_2}\psi,
\end{aligned}$$

where the final equality follows from (3.3.61). Differentiating (3.3.60) with respect to $\mathbf{T}_2$, and applying (3.3.63) and (3.3.62) gives us,

$$\mathcal{H}_{\partial\Omega} = \partial_{\mathbf{T}_2}\psi = \partial_{\mathbf{T}_2}\left(\omega - \Theta + \frac{\pi}{2}\right) = (\beta_1 \, \partial_{\mathbf{T}_2}\beta_2 - \beta_2 \, \partial_{\mathbf{T}_2}\beta_1) - \partial_{\mathbf{T}_2}\Theta,$$

which, upon rearranging, yields (3.3.57). □

Lemma 3.3.26 (Miranda–Talenti estimate) *Let $\Omega \subset \mathbb{R}^2$ be a C^2 domain, and assume that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Furthermore, assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} \geq 0 \text{ on } \partial\Omega.$$

Then, we have that

$$|u|^2_{H^2(\Omega)} := \int_\Omega \sum_{i,j=1}^{2} (D^2_{ij}u)^2 \leq \int_\Omega (\Delta u)^2 = \|\Delta u\|^2_{L^2(\Omega)},$$

for all $u \in H^2_{\beta,0}(\Omega)$.

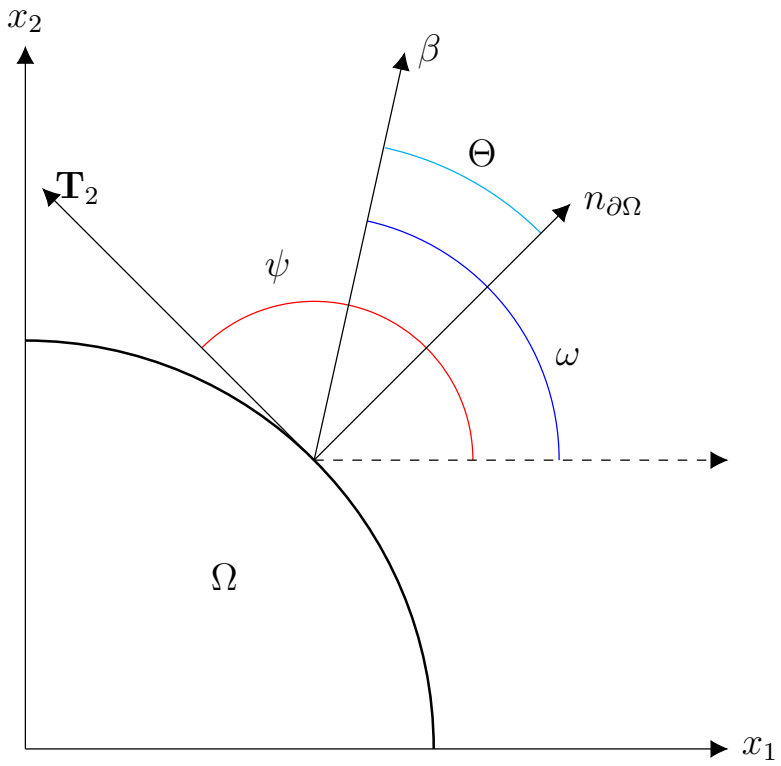


Figure 3.1: Visualisation of the vectors $\beta, n_{\partial\Omega}$, T_2, and the angles Θ, ψ, ω. Θ is the (anticlockwise) oriented angle between β and $n_{\partial\Omega}$, ψ is the (anticlockwise) oriented angle between $\mathbf{T}_2$ and the x_1-axis, and ω is the (anticlockwise) oriented angle between β and the x_1-axis.

Proof: Take $u \in H^2_{\beta,0}(\Omega)$, then, we see that

$$|u|^2_{H^2(\Omega)} + 2\int_\Omega \det D^2 u = \int_\Omega \sum_{i,j=1}^{d} (D^2_{ij}u)^2 + 2(D^2_{11}uD^2_{22}u - (D^2_{12}u)^2)$$
$$= \int_\Omega (D^2_{11}u + D^2_{22}u)^2 = \|\Delta u\|^2_{L^2(\Omega)}.$$

We then apply Corollary 3.3.22, yielding

$$|u|^2_{H^2(\Omega)} + \int_{\partial\Omega} |\nabla u|^2(\beta_1\, \partial_{\mathbf{T}_2}\beta_2 - \beta_2\, \partial_{\mathbf{T}_2}\beta_1) = \|\Delta u\|^2_{L^2(\Omega)}.$$

Now we apply Lemma 3.3.57, which gives us

$$|u|^2_{H^2(\Omega)} + \int_{\partial\Omega} |\nabla u|^2(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega}) = \|\Delta u\|^2_{L^2(\Omega)}.$$

Finally, by the hypotheses of the lemma, we have have that $\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} \geq 0$ on $\partial\Omega$, and thus we obtain the following:

$$|u|^2_{H^2(\Omega)} \leq \|\Delta u\|^2_{L^2(\Omega)},$$

as desired. □

Lemma 3.3.27 (Gradient estimate) *Let $\Omega \subset \mathbb{R}^2$ be a C^2 domain, and assume that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Furthermore, assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \text{ on } \partial\Omega,$$

we have that

$$\int_\Omega |\nabla u|^2 \leq C \int_\Omega (\Delta u)^2,$$

for all $u \in H^2_\beta(\Omega)$, where C is a constant, independent of u.

Proof: See [89], Lemma 1.5.8. □

Lemma 3.3.28 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 domain, and assume that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Furthermore, assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \text{ on } \partial\Omega. \tag{3.3.64}$$

Then, the space $H^2_{\beta,0}(\Omega)$ satisfies the MT estimate property (3.3.9)–(3.3.10).

Proof: Since $\partial\Omega \in C^2$, and (3.3.64) holds, Lemma 3.3.26 implies that

$$|u|^2_{H^2(\Omega)} := \int_\Omega \sum_{i,j=1}^{2} (D^2_{ij} u)^2 \leq \int_\Omega (\Delta u)^2 = \|\Delta u\|^2_{L^2(\Omega)}, \tag{3.3.65}$$

for all $u \in H^2_{\beta,0}(\Omega)$. I.e., (3.3.9) holds. We then apply Lemma 3.3.27, which tells us that

$$\int_\Omega |\nabla u|^2 \leq C \int_\Omega (\Delta u)^2, \tag{3.3.66}$$

for all $u \in H^2_{\beta,0}(\Omega)$, where C is a constant independent of u. Now, since $u \in H^2_{\beta,0}(\Omega)$, it follows that

$$u \in L^2_0(\Omega) = \left\{ v \in L^2(\Omega) : \int_\Omega v = 0 \right\}.$$

Thus, the Poincaré inequality yields

$$\|u\|_{2,\Omega} = \left\| u - \frac{1}{|\Omega|} \int_\Omega u \right\|_{2,\Omega} \leq C|u|_{H^1(\Omega)} \quad \forall u \in H^2_{\beta,0}(\Omega), \tag{3.3.67}$$

where C is a constant independent of u. Combining estimates (3.3.65)–(3.3.67), we obtain

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{2,\Omega} \quad \forall u \in H^2_{\beta,0}(\Omega), \tag{3.3.68}$$

where C is a constant independent of u. Thus, we have proven that $H^2_{\beta,0}(\Omega)$ satisfies (3.3.9) and (3.3.10). $\square$

Theorem 3.3.29 (Existence and uniqueness for the oblique case) *Let $\Omega \subset \mathbb{R}^2$ be a C^2 domain, assume that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$, and that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \text{ on } \partial\Omega.$$

Furthermore assume that that the collection of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satisfies (3.3.6). Then, there exists a unique function $u \in H := H^2_{\beta,0}(\Omega)$ such that

$$\begin{cases} \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \quad \text{a.e. in } \Omega, \\ \qquad\qquad \beta \cdot \nabla u \quad \text{is constant on } \partial\Omega. \end{cases} \tag{3.3.69}$$

Proof: We see that $u \in H$ implies that

$$\beta \cdot \nabla u|_{\partial\Omega} \quad \text{is constant},$$

and thus u satisfies the boundary condition. By, Lemma 3.3.28 it follows that H satisfies the MT estimate property (3.3.9)–(3.3.10). Furthermore, as $d = 2$, and

the family of operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satsifes (3.3.6), we have that the family of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$ also satisfies (3.3.12). Finally, it follows from [89], Page 56, that $\Delta : H \to L^2(\Omega)$ is a surjection. Thus, an application of Corollary 3.3.9 yields the existence of a unique $u \in H$ that satisfies (3.3.69). $\square$

Remark 3.3.30 (Example and counter example of condition (3.3.64)) *Recall that (3.3.64) holds, for a C^2 domain $\Omega \subset \mathbb{R}^2$, and a unit vector-valued function $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$, if*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega, \tag{3.3.70}$$

where Θ is the (anticlockwise) oriented angle between β and $n_{\partial\Omega}$ (depicted in Figure 3.1), and $\mathcal{H}_{\partial\Omega}$ is the mean curvature of $\partial\Omega$.

- Example class: *A class of examples of domain-vector pairs (Ω,β) that satisfy (3.3.70) are C^2 domains $\Omega \subset \mathbb{R}^2$ of* strictly positive *mean curvature (for example ellipse domains), and $\beta = Rn_{\partial\Omega}$ for some $R \in \mathrm{SO}(2)$, where $\mathrm{SO}(2)$ is the set of 2×2 rotation matrices (a particular example is $R = I_2$, the 2×2 identity matrix, i.e., the Neumann boundary condition). In this case $n_{\partial\Omega} \in C^1(\partial\Omega;\mathbb{S}^1)$, and so $\beta = Rn_{\partial\Omega} \in C^1(\partial\Omega;\mathbb{S}^1)$. Furthermore, the angle between β and $n_{\partial\Omega}$ is constant, thus $\partial_{\mathbf{T}_2}\Theta = 0$ on $\partial\Omega$. Overall, we have that*

  $$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

  A particular example of such a pair is $\Omega := \{x = (x_1,x_2) \in \mathbb{R}^2 : |x| < 1\}$, and $\beta := n_{\partial\Omega}$.

- Counterexample class: *A class of examples of domain-vector pairs (Ω,β) that do not satisfy (3.3.70) are C^2 domains $\Omega \subset \mathbb{R}^2$ that have* nonpositive *mean curvature at some point $x \in \partial\Omega$, and $\beta = Rn_{\partial\Omega}$ for some $R \in \mathrm{SO}(2)$. In this case $n_{\partial\Omega} \in C^1(\partial\Omega;\mathbb{S}^1)$, and so $\beta = Rn_{\partial\Omega} \in C^1(\partial\Omega;\mathbb{S}^1)$. Furthermore, the angle between β and $n_{\partial\Omega}$ is constant, thus $\partial_{\mathbf{T}_2}\Theta = 0$ on $\partial\Omega$. However, as the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega}$, is nonpositive at $x \in \partial\Omega$, we have that*

  $$(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega})(x) = \mathcal{H}_{\partial\Omega}(x) \leq 0,$$

  which contradicts (3.3.70). A particular example of such a pair is $\Omega := \{x = (x_1,x_2) \in \mathbb{R}^2 : x_1^4 + x_2^4 < 1\}$, and $\beta := n_{\partial\Omega}$. This domain has zero mean curvature at the points $\pm(0,1)$, $\pm(1,0)$.

One should note that the example class above in non exhaustive, that is, there are other examples of domain-vector pairs that satisfy (3.3.70), in particular one may allow for the case that either $\mathcal{H}_{\partial\Omega}$ or $\partial_{\mathbf{T}_2}\Theta$ *become nonpositive on* $\partial\Omega$*, so long as* $\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0$ *on* $\partial\Omega$*, i,e., it is not necessary for the domain to be convex, or for the oblique angle,* Θ*, to be increasing (with respect to an anticlockwise orientation) on* $\partial\Omega$*.*

This is of importance when considering applications to problems with nonlinear gradient boundary conditions. In particular, the MA optimal transport problem, where the boundary condition is of the form

$$b(\nabla u(x)) = 0 \quad x \in \Omega, \tag{3.3.71}$$

and $b : \mathbb{R}^2 \to \mathbb{R}^2$ *is a given function. Upon linearisation, one obtains a sequence of boundary conditions of the form:*

$$\beta_n \cdot \nabla u_{n+1}(x) = g_n(x), \quad x \in \partial\Omega,$$

for $n \in \mathbb{N}_0$*, where the functions* β_n *and* g_n *are known (see Chapter 9, Section 9.5.3 for further details on the boundary condition (3.3.71), and its linearisation). It then follows that the validity of (3.3.70) is dependent upon the nature of the* β_n *and* g_n*, which depend on the previous iterate* u_n*. However, for the MA optimal transport problem, it is often the case that* Ω *is uniformly convex, and so* $\mathcal{H}_{\partial\Omega} > 0$ *on* $\partial\Omega$*, which means that it is not necessary for the corresponding oblique angles,* Θ_n*,* $n \in \mathbb{N}_0$*, to be increasing (with respect to the anticlockwise orientation) on* $\partial\Omega$*.*

3.4 Krylov's HJB formulation of MA type equations

We shall now state an essential lemma from [74]; recall that the Krylov control set, X, is given by $X := \{W \in \mathbb{R}^{d\times d}_{\mathrm{Sym}} : W \geq 0, \operatorname{Tr} W = 1\}$.

Lemma 3.4.1 *For* $A \in \mathbb{R}^{d\times d}_{\mathrm{Sym}}$*, and* $\psi \in \mathbb{R}^+_0$*, the following holds:*

$$\max_{W\in X}\{-W : A + (\det W)^{1/2}\psi\} = 0$$

if and only if

$$A \geq 0 \quad \textit{and} \quad \psi^d = d^d \det A.$$

We apply this lemma to the MAD problem (3.2.1) with $f(x, u, \nabla u) := f(x)$, where $f \in C^{0,\alpha}(\overline{\Omega}; \mathbb{R}^+)$, $\alpha \in (0,1)$ is a given, uniformly positive function, and the boundary datum $\phi \equiv 0$, that is,

$$\begin{cases} \det D^2 u(x) = f(x), \ x \in \Omega, \\ u(x) = 0, \ x \in \partial\Omega. \end{cases} \tag{3.4.1}$$

We see that a convex function, $u \in C^{2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, satisfies (3.4.1) if and only if

$$\begin{cases} \max\limits_{W \in X} \{-W : D^2 u + (\det W)^{1/d} \psi\}(x) = 0, \ x \in \Omega, \\ u(x) = 0, \ x \in \partial\Omega, \end{cases} \tag{3.4.2}$$

where

$$\psi(x) := d(f(x))^{1/d}. \tag{3.4.3}$$

Due to Krylov's characterisation, we see that the classical solvability the MA equation (3.4.1), with a homogeneous Dirichlet boundary condition and right-hand side $f \in C^{0,\alpha}(\overline{\Omega}; \mathbb{R}^+)$, $\alpha \in (0,1)$, is equivalent to solving the HJB equation (3.4.2).

3.5 Numerical motivation - selection criteria

Krylov's characterisation (3.4.2) of the MA equation (3.4.1), establishes a key link between MA and HJB type equations. That is, the former is a special case of the latter. This provides us with an initial motivation for designing a finite element method for the approximation of classical solutions to MA type problems. Indeed, the discontinuous Galerkin finite element method (DGFEM) of Smears and Süli proposed in [111] is stable and consistent, independent of the discretisation parameter (for further details of these properties, see Chapter 7). The linearisation scheme used to iteratively solve the corresponding discrete nonlinear problem is a semi-smooth Newton's method, corresponding to Howards' method, which can be proven to converge superlinearly.

However, the DGFEM proposed in [111] can be used to approximate strong solutions of HJB Dirichlet boundary-value problems, provided that the following assumptions hold:

1. The family of linear operators $\{L^\alpha\}_{\alpha \in \Lambda}$ must satisfy the uniform ellipticity condition (3.3.6), and the Cordes condition (3.3.12);

2. The domain Ω must be convex and *polytopal*.

One can see that the set X contains matrices that are positive semidefinite, rather than positive definite. This means that the boundary value problem (3.4.2) falls into the class of "degenerate elliptic equations". In fact, a motive for Krylov's formulation of the MA problem as a HJB problem was to provide an approach for understanding the degenerate case of the MA problem, utilising the theory of viscosity solutions of HJB problems.

However, if a family of operators $\{L^\alpha\}_{\alpha\in\Lambda} = \{A^\alpha : D^2\}_{\alpha\in\Lambda}$ has a member L^α, for which the corresponding matrix valued function $A^\alpha \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\text{Sym}})$, takes degenerate values on a set $E \subset \Omega$ of nonzero Lebesgue measure, the family $\{L^\alpha\}_{\alpha\in\Lambda}$ cannot satisfy the Cordes condition (3.3.12). For example, take the case that $d = 2$, then, denoting λ_2 to be the largest eigenvalue of A^α, on E, we have that

$$\frac{|A^\alpha|^2}{(\operatorname{Tr} A^\alpha)^2} = \frac{\lambda_2^2}{\lambda_2^2} = 1 > \frac{1}{1+\varepsilon},$$

for any $\varepsilon \in (0, 1]$, contradicting (3.3.12).

Theorems 3.3.16 and 3.3.20, concern the existence and uniqueness of strong solutions to HJB Dirichlet boundary-value problems, and their respective hypotheses rely upon the validity of the Cordes condition (3.3.12); as such we shall seek to reformulate the HJB problem (3.4.2) where the supremum is taken over a subset of X for which the collection of such matrices satisfies the Cordes condition (3.3.12). We will see that this requires the solution $u \in C^{2,\alpha}(\overline{\Omega})$, $\alpha \in (0, 1)$ of the MA problem (3.4.1) to be uniformly convex. The existence of such a solution is a consequence of Theorem 3.2.1, provided that the domain Ω has a $C^{2,\alpha}$ boundary, $\alpha \in (0, 1)$, with positive minimal principal curvature. Polytopal domains, however, do not satisfy such regularity and curvature assumptions, and thus we must provide an extension of the DGFEM proposed in [111] that allows for domains with curved boundaries (see Chapter 7 for further details on this extension).

One may now ask what it is that motivates us to consider this link between the MA and HJB equation, and why such an approach may lead to the development a method with properties that are advantageous (when compared to other potential methods). Aside from the scheme being stable and consistent (leading to solvability and optimal approximation results), a key motivation for this approach is included in the statement of Lemma 3.4.1. In particular, it is the fact that a uniformly $C^{2,\alpha}$, $\alpha \in (0, 1)$ solution to the MA equation (3.4.2) solves the HJB equation (3.4.1) *provided* that the solution is convex. The dichotomy of convex and concave solutions is key to the uniqueness of solution to MA type equations, and plays an important role in the design of numerical methods.

For example, let us assume that $\Omega \subset \mathbb{R}^2$ is convex, and that a convex function $u \in C^{2,\alpha}(\overline{\Omega})$ satisfies

$$\begin{cases} \det D^2 u(x) = f(x), \ x \in \Omega, \\ \qquad u(x) = 0, \ x \in \partial\Omega, \end{cases} \tag{3.5.1}$$

then one can see that $\tilde{u} := -u$ is *concave*, and satisfies

$$\begin{cases} \det D^2 \tilde{u}(x) = \det(-D^2 u(x)) = \det(D^2 u(x)) = f(x), \ x \in \Omega, \\ \qquad \tilde{u}(x) = -u(x) = 0, \ x \in \partial\Omega, \end{cases}$$

i.e., $\tilde{u}$ also solves (3.5.1). One would hope that numerical methods that approximate solutions of (3.5.1) may have the same uniqueness property, that is, that there exists at most two solutions to the numerical method. However, this is not always the case. In [49], the authors implement a standard nine-point stencil finite difference method for the problem (3.5.1) with a smooth right-hand side, and a smooth solution u (unique up to sign), with the choice of domain $\Omega = (0,1)^2$. Upon implementing this method on a 4×4 grid, solving the resulting nonlinear system by applying Newton's method, they obtain sixteen different numerical solutions by varying the initial guess of the Newton's method.

As mentioned in [49], one may conjecture that this phenomena extrapolates, causing Newton's method to potentially converge to $2^{(N-2)^2}$ different solutions on and $N \times N$ grid, by varying the initial guess. When designing a numerical scheme, it is important that one knows which solution the method is converging to, without needing too much (Newton's method is well known to be conditionally convergent, and a prerequisite to convergence is often sufficient proximity to the true solution) prior knowledge of the true solution. Indeed, the aforementioned finite difference method implemented in [49], was proposed in [14], with an additional *selection* criteria, which in essence singles out a particular numerical solution.

Denoting $u_{i,j}$ to be the approximation of $u(ih, jh)$, and denoting $f_{i,j} := f(ih, jh)$, $i, j = 0, \dots N-1$, $h = 1/(N-1)$; using a nine-point stencil to approximate the second order partial derivatives of u in (3.5.1), leads one to the discretisation

$$\begin{aligned} D^2_{11} u_{i,j} D^2_{22} u_{i,j} - (D^2_{12} u_{i,j})^2 &= f_{i,j}, \quad i, j = 1 \dots, N-2, \\ u_{0,j} = u_{N-1,j} = u_{j,0} = u_{j,N-1} &= 0, \quad j = 0, \dots N-1, \end{aligned} \tag{3.5.2}$$

where $D^2_{11} u_{i,j} := (u_{i+1,j} + u_{i-1,j} - 2u_{i,j})/h^2$, $D^2_{22} u_{i,j} := (u_{i,j+1} + u_{i,j-1} - 2u_{i,j})/h^2$, and $D^2_{12} u_{i,j} := (u_{i+1,j+1} + u_{i-1,j-1} - u_{i-1,j+1} - u_{i+1,j-1})/4h^2$. This provides the following system of quadratic equations for $u_{i,j}$, $i, j = 1, \dots N-2$,

$$4(a_1 - u_{i,j})(a_2 - u_{i,j}) - \frac{1}{4}(a_3 - a_4)^2 = h^4 f_{i,j}, \quad i, j = 1, \dots, N-2, \tag{3.5.3}$$

where

$$a_1 = \frac{u_{i+1,j} - u_{i-1,j}}{2}, \quad a_2 = \frac{u_{i,j+1} + u_{i,j-1}}{2},$$
$$a_3 = \frac{u_{i+1,j+1} - u_{i-1,j-2}}{2}, \quad a_2 = \frac{u_{i-1,j+1} + u_{i+1,j-1}}{2}.$$

The quadratic system (3.5.3) has solutions

$$u^{\pm}_{i,j} = \frac{1}{2}(a_1+a_2) \pm \frac{1}{2}\sqrt{(a_1 - a_3)^2 + \frac{1}{2}(a_3 - a_4)^2 + h^4 f_{i,j}}, \quad i,j = 1,\ldots,N-2. \tag{3.5.4}$$

That is, at each internal grid point, there are two solutions to choose from, either $u^+_{i,j}$ or $u^-_{i,j}$. In [14], the authors make the selection $u_{i,j} := u^-_{i,j}$, which enforces the local convexity condition

$$u(x) \leq \frac{u(x+k) - u(x-k)}{2}, \tag{3.5.5}$$

along all grid directions k. Of course, from the offset one is free to choose either $u^+_{i,j}$ or $u^-_{i,j}$ at any internal grid point, providing $2^{(N-2)^2}$ potential methods (such methods would not preserve (3.5.5) in general).

Returning to the example of [49], for a 4×4 grid, there are four internal grid points, and thus four values $u_{1,1}, u_{1,2}, u_{2,1}, u_{2,2}$ to determine from (3.5.4) (the remaining 12 values are determined by the boundary condition), and thus choosing $u_{i,j} = u^+_{i,j}$ or $u^-_{i,j}$ for $i,j = 1,2$ results in sixteen methods, which is the number of solutions the authors of [49] obtain by varying the initial guess for the corresponding Newton's method.

The example of [49] and the method of [14] demonstrate the importance of a selection criteria in the design of numerical methods for the approximation of solutions to MA type problems. In particular, the method of [14] directly imposes a selection criteria in the design of the method (i.e., selecting $u_{i,j} = u^-_{i,j}$). Whereas, the example of [49] shows that a direct application of Newton's method for this scheme imposes a selection criteria via the choice of the particular initial guess (one would expect that the proximity of the initial guess for the Newton's method to the solution of one of the sixteen methods causes the Newton's method to converge to the corresponding solution), which leads to a lack of uniqueness (of a greater multiplicity than that of the continuous problem (3.5.1)). Although the numerical scheme of [14] is well-posed, the authors of [14] state that the method is not monotone, and thus a proof of convergence of this scheme is an open problem (this lack of monotonicity is also commented upon in [49]).

The numerical methods that we propose in Chapters 8 and 9 involve a selection criteria in the form of an initial guess for Newton's method; however, in contrast to the

example of [49], we empirically observe convergence to the "chosen" solution, that is more robust with respect to the initial guess. In particular, we see examples where the choice of either a convex or concave initial guess results in the convergence of Newton's method to a numerical solution that approximates the true convex or concave solution of (3.2.1) respectively (for further details see Section 8.6). In Chapter 9 we also implement a method that updates the selection criteria, by first applying Newton's method on a coarse mesh (which appears to be more robust than a fine mesh with respect to the initial guess), and then using the coarse mesh solution as an initial guess for Newton's method on the fine mesh, resulting in a scheme that is more robust with respect to the initial guess.

As previously mentioned, a key motivation for considering Krylov's HJB characterisation (3.4.2) of the MA equation (3.4.1) is the fact that a uniformly $C^{2,\alpha}$, $\alpha \in (0,1)$ solution to the MA equation (3.4.2) solves the HJB equation (3.4.1) *provided* that the solution is convex. Thus, by formulating a DGFEM for the solution of the HJB equation, we provide a scheme with a stronger selection property; that is, when the Newton's method converges, it converges to the *globally unique*[1] numerical solution that approximates the *unique* convex solution of (3.4.2). One should note that we are not the first to consider Krylov's characterisation. For example in [50, 66] a semi-Lagrangian method is employed, approximating solutions of the Krylov HJB problem (3.4.2), providing a numerical approximation of viscosity solutions to the MA equation (the method proposed in [50] allows for the degenerate case, as well as inhomogeneous Dirichlet boundary data, and [66] provides numerical evidence that numerical solutions of the method proposed in [50] converge without the assumption of domain and Dirichlet boundary data convexity).

3.5.1 Uniformly elliptic HJB-MA equations

Our goal for the remainder of this Chapter is to prove that we may indeed reformulate problem (3.4.2) as a uniformly elliptic HJB problem.

The link between the HJB equation and the MA equation allows us to deduce more about the nature of the set X. Note that the next lemma holds for any finite dimension d.

[1]This is in contrast to the numerical solution of the method proposed Chapter 8. In particular, Theorem 8.2.3 provides existence of a numerical solution in a ball of radius $h^{2+\alpha}$, for some $\alpha > 0$, of the true solution, and thus the numerical solution is only proved to be *locally unique*.

Lemma 3.5.1 *The set X satisfies*

$$\sup_{W\in X} \det W = \frac{1}{d^d},$$

and a maximiser is given by $\frac{1}{d}I_d$.

Proof: Consider the MA equation

$$\det D^2 u = 1.$$

A solution to this problem is given by $u(x) = \frac{1}{2}|x|^2$; note that $D^2u = I_d$. Applying Lemma 3.4.1 we obtain

$$\begin{aligned} 0 &= \sup_{W\in X}\{-W : D^2u + d(\det W)^{1/d}\} \\ &= \sup_{W\in X}\{-\operatorname{Tr}(W) + d(\det W)^{1/d}\} \\ &= d\left(\sup_{W\in X}(\det W)^{1/d}\right) - 1; \end{aligned}$$

rearranging, and noting that $a^{\frac{1}{d}} \geq b$ implies that $a \geq b^d$ for $a, b \geq 0$, we obtain

$$\sup_{W\in X} \det W = \frac{1}{d^d}.$$

It is also clear that $\frac{1}{d}I_d \in X$ and $\det(\frac{1}{d}I_d) = \frac{1}{d^d}$. $\square$

Theorem 3.5.2 *Let Ω be a bounded convex open subset of $\mathbb{R}^2$. Let $X_\xi := \{W \in X : \det W \geq \xi\}$. Then, for any constant $\frac{1}{4} \geq \xi > 0$, there exists a unique solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of the following HJB equation*

$$\begin{aligned} \sup_{W\in X_\xi}\{-W : D^2u + (\det W)^{1/2}\psi\}(x) = 0,\ x \in \Omega, \\ u(x) = 0,\ x \in \partial\Omega, \end{aligned} \tag{3.5.6}$$

where ψ is given by (3.4.3).

Proof: First note that as $\xi \leq 1/4$, Lemma 3.5.1 guarantees that $X_\xi \neq \emptyset$. The set X_ξ contains positive definite matrices, and in two dimensions uniform ellipticity implies the Cordes condition. Then, setting $\Lambda = X_\xi$, we can see that X_ξ is a compact metric space; using the Euclidean distance as a metric, and noting that $X_\xi = \mathscr{D}^{-1}([\xi, 1/4])$, where $\mathscr{D} : \Lambda \to \mathbb{R}$ given by

$$\mathscr{D}(W) := \det(W), \quad W \in X_\xi,$$

is a continuous function, we deduce that X_ξ is closed. Combining the determinant upper bound given by Lemma 3.5.1, and the unit trace constraint that X_ξ satisfies, we see that X_ξ is also bounded. Thus X_ξ is compact.

We can apply Theorem 3.3.16, yielding existence of a unique $v \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying

$$\begin{cases} \sup\limits_{W \in X_\xi} \{W : D^2 v + (\det W)^{1/2}\psi\} = 0 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{3.5.7}$$

We then (uniquely) define $u := -v$. $\square$

Theorem 3.5.3 *Let $d = 2$, and assume that Ω is uniformly convex, and that $\partial\Omega \in C^{2,\alpha}$, $f \in C^{0,\alpha}(\overline{\Omega}; \mathbb{R}^+)$ is uniformly positive, for some $\alpha \in (0,1)$. Then, there exists a unique uniformly convex $u \in C^{2,\alpha}(\overline{\Omega})$ that satisfies the MAD problem (3.4.1). Furthermore, there exists $1/4 \geq \xi > 0$ dependent upon $\|u\|_{C^2(\overline{\Omega})}$ such that u is also the unique solution to*

$$\begin{cases} \sup\limits_{W \in X_\xi} \{-W : D^2 u + (\det W)^{1/2}\psi\} = 0 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases}$$

where $\psi(x) = 2(f(x))^{1/2}$.

Proof: First, we must show that there exists a $u \in C^{2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, that satisfies the MAD problem (3.4.1). The existence of such a function u follows from Theorem 3.2.1, so long as (3.2.4) holds.

Now, since Ω is uniformly convex, the minimal principal curvature of $\partial\Omega$, $\kappa_{\partial\Omega}$, is uniformly bounded below by some constant $\kappa_0 > 0$. Furthermore, in this case, we have that the boundary datum $\phi \equiv 0$, so that

$$\rho_\phi := (1 + \frac{4d}{\kappa_{\partial\Omega}^4})\|\phi\|_{C^2(\partial\Omega)} = (1 + \frac{8}{\kappa_{\partial\Omega}^4})\|\phi\|_{C^2(\partial\Omega)} = 0.$$

So it remains to find a nondecreasing function $\zeta(t) \geq 1$ such that $f(x,z,q) = f(x) \leq \zeta(|q|)$ for all $(x,z,p) \in \Omega \times \mathbb{R} \times \mathbb{R}^2$, and

$$1 < d\kappa_{\partial\Omega}^d \int_0^\infty \frac{t^{d-1}}{\zeta(t+\rho_\phi)}\, dt = 2\kappa_{\partial\Omega}^2 \int_0^\infty \frac{t}{\zeta(t)}\, dt.$$

Now, we let $\zeta(t) = \lambda(1 + e^{(\kappa_0 t)/\lambda})$, where $\lambda := \max\{1, \|f\|_{C^0(\Omega)}\}$. Then we see that

$$\begin{aligned}
2\kappa_{\partial\Omega}^2 \int_0^\infty \frac{t}{\zeta(t)}\, dt &= 2\kappa_{\partial\Omega}^2 \int_0^\infty \frac{t}{\lambda(1 + e^{(\kappa_0 t)/\lambda})}\, dt \\
&= \frac{2\kappa_{\partial\Omega}^2}{\lambda} \frac{\lambda^2}{\kappa_0^2} \int_0^\infty \frac{s}{1 + e^s}\, ds \\
&= \frac{2\kappa_{\partial\Omega}^2}{\lambda} \frac{\lambda^2}{\kappa_0^2} \frac{\pi^2}{12} \\
&\geq \frac{\kappa_0^2}{\lambda} \frac{\lambda^2}{\kappa_0^2} \frac{\pi^2}{6} \\
&\geq \frac{\pi^2}{6} > 1.
\end{aligned}$$

Furthermore, ζ is a strictly increasing function, satisfying

$$\zeta(|p|) = \lambda(1 + e^{(\kappa_0 |p|)/\lambda}) \geq \lambda \geq \|f\|_{C^0(\Omega)} \geq f(x),\ \forall (x, z, p) \in \Omega \times \mathbb{R} \times \mathbb{R}^2.$$

Thus, by Theorem 3.2.1, there exists a uniformly convex function $u \in C^{2,\alpha}(\overline{\Omega})$ that satisfies the MAD problem (3.4.1).

Let us define the map $A_u : \overline{\Omega} \to \mathbb{R}^{d\times d}$ by:

$$A_u(x) := \frac{\operatorname{Cof}(D^2 u)}{\Delta u}, \tag{3.5.8}$$

note that this map is well defined, since u is uniformly convex, and so, its Laplacian is uniformly positive. Also, since $u \in C^{2,\alpha}(\overline{\Omega})$ for some $\alpha \in (0, 1)$, we have that $A_u \in C^{0,\alpha}(\overline{\Omega})$. Furthermore, $\operatorname{Cof}(D^2 u)$ is symmetric, and

$$\operatorname{Tr}(A_u) = \frac{1}{\Delta u} \operatorname{Tr}(\operatorname{Cof}(D^2 u)) = \frac{\Delta u}{\Delta u} = 1,$$

and so $A_u : \overline{\Omega} \to X$. We see that A_u satisfies

$$\begin{aligned}
&- A_u(x) : D^2 u(x) + \det(A_u(x))^{1/2} \psi(x) \\
&= \frac{1}{\Delta u(x)} (-\operatorname{Cof}(D^2 u(x)) : D^2 u(x) + 2(\det(\operatorname{Cof} D^2 u(x)))^{1/2} f(x)^{1/2}) \\
&= \frac{2}{\Delta u(x)} (-\det D^2 u(x) + \det(D^2 u(x))^{1/2} f(x)^{1/2}) \\
&= \frac{2}{\Delta u(x)} (-\det D^2 u(x) + f(x)) = 0.
\end{aligned} \tag{3.5.9}$$

We also obtain a lower bound on the determinant of A_u:

$$\begin{aligned}
\det(A_u) &= \det\left(\frac{\operatorname{Cof}(D^2 u)}{\Delta u}\right) \\
&= \frac{\det(D^2 u)}{(\Delta u)^2} \\
&= \frac{f}{(\Delta u)^2} \geq \frac{\delta}{2|u|_{C^2(\overline{\Omega})}} =: \xi,
\end{aligned}$$

where $\delta = \inf_{x\in\overline{\Omega}} f(x) > 0$, and so, $\xi > 0$.

Let us consider the following HJB equation: find $v \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$\begin{cases} \sup\limits_{W\in X_\xi} \{-W : D^2 v + (\det W)^{1/2}\psi\} = 0, \ x \in \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad v = 0, \ x \in \partial\Omega. \end{cases} \tag{3.5.10}$$

We also note that since $A_u(x) \in X$ for all $x \in \overline{\Omega}$, by Lemma 3.5.1 we have that

$$\xi \le \det(A(x)) \le \frac{1}{4} \quad \forall x \in \overline{\Omega}.$$

There is an important difference between the set X and the set $X_\xi := \{W \in X : \det W \ge \xi\}$, which is that the latter set consists entirely of positive definite matrices. It then follows from Theorem 3.5.2 that there exists a unique $v \in H^2(\Omega) \cap H_0^1(\Omega)$ that solves (3.5.10).

We then see that the solution u of the MA equation satisfies (noting that $X_\xi \subseteq X$)

$$\begin{aligned} \sup_{W\in X_\xi} \{-W : D^2 u + (\det W)^{1/2}\psi(x)\} &\le \sup_{W\in X} \{-W : D^2 u + (\det W)^{1/2}\psi(x)\} \\ &= 0. \end{aligned}$$

Since $A(x) \in X_\xi$ for each $x \in \Omega$, from (3.5.9), we obtain

$$\sup_{W\in X_\xi} \{-W : D^2 u + (\det W)^{1/2}\psi(x)\} \ge -A(x) : D^2 u(x) + (\det A(x))^{1/2}\psi(x) = 0.$$

By combining these results, we obtain

$$\sup_{W\in X_\xi} \{-W : D^2 u(x) + (\det W)^{1/2}\psi(x)\} = 0.$$

Since $u = 0$ on $\partial\Omega$, and $C^2(\overline{\Omega}) \subset H^2(\Omega)$, we see that $u \in H^2(\Omega) \cap H_0^1(\Omega)$, and thus by uniqueness $u = v$. Furthermore, if $u_1, u_2 \in C^{2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, are two uniformly convex functions satisfying (3.4.1), then $u_1, u_2 \in H^2(\Omega) \cap H_0^1(\Omega)$, and they both satisfy (3.5.10), with $\xi = \xi^*$, where

$$\xi^* := \min\left\{ \frac{\delta}{2|u_1|_{C^2(\overline{\Omega})}}, \frac{\delta}{2|u_2|_{C^2(\overline{\Omega})}} \right\}.$$

Thus, by uniqueness, $u_1 \equiv u_2$, that is, the MAD problem (3.4.1) is also uniquely solvable in the class of uniformly convex $C^{2,\alpha}(\overline{\Omega})$ functions. $\square$

Chapter 4

Finite element theory on curved domains

4.1 Motivation

We begin this Chapter with a motivation for considering domains with curved boundaries. As mentioned in the Introduction, a main goal of this thesis is to design and analyse FEMs for the approximation of solutions to MA type equations. When considering classical solutions, the nature of MA type equations often lead one to pose equations on domains that are either (at least) C^2, or uniformly convex, or both. The set of such domains includes ones with curved boundary; indeed, if one assumes that the $\Omega \subset \mathbb{R}^d$ is bounded, and has a flat boundary, i.e., a polytope, then $\partial\Omega$ cannot be C^2 (due to the boundary portions meeting at a face or vertex), and Ω cannot be uniformly convex, since, for example, when $d = 2$, a flat boundary portion must contain two vertices of the polygon, as well as the open line segment connecting them, contradicting Definition 2.1.7.

The following example demonstrates the problems that may arise when considering even a "simple" MA type problem on a domain without boundary curvature. Consider the MA equation on the unit square $\Omega := (0,1)^2$: find $u : (0,1)^2 \to \mathbb{R}$ such that

$$\begin{cases} \det D^2 u(x) = f(x), \quad x \in (0,1)^2, \\ \qquad u(x) = 0, \quad x \in \{0,1\} \times (0,1) \cup [0,1] \times \{0,1\}, \end{cases} \tag{4.1.1}$$

where $f \in C^\infty([0,1]^2; \mathbb{R}^+)$. Let us assume that there exists a function $u \in C^2([0,1]^2)$ which satisfies (4.1.1). It then follows that

$$D_{11}^2 u = 0 \quad \text{on } \{x \in [0,1]^2 : x_2 = 0\} =: \Gamma \subset \partial\Omega,$$

since on this boundary portion the x_1 derivative coincides with the tangential derivative on Γ, and, furthermore, $u|_\Gamma = 0$. This implies that

$$\det D^2 u(x) = D^2_{11} u(x) D^2_{22} u(x) - (D^2_{12} u(x))^2 = -(D^2_{12} u(x))^2 \leq 0 \quad x \in \Gamma,$$

which contradicts the assumption that $f \in C^\infty([0,1]^2; \mathbb{R}^+)$. In order to stay in an elliptic framework, we must assume that $f \geq 0$ in $\overline{\Omega}$ (if we were to assume that $f < 0$, we would move into the regime of hyperbolic MA equations, see [62, 64, 115, 123]); we have seen from our example that we cannot hope to find a uniformly convex $C^2(\overline{\Omega})$ solution to this problem (as this would require $\det D^2 u > 0$ in $\overline{\Omega}$).

One could pose (4.1.1) with inhomogeneous boundary data, in particular, uniformly convex boundary data (this was in fact a motivation for Theorem 3.3.20). However, in this case, one would in general only expect global $H^{3-\delta}(\Omega)$-regularity for arbitrary $\delta > 0$, due to the angle of the corners of the unit square (see [60] for more details). Furthermore, our approach for proving existence of a numerical solution to the CGFEM in Chapter 8 presupposes that the true solution belongs to $C^{p+2,\alpha}(\overline{\Omega})$, $p \geq 3$, $\alpha \in (0,1)$, and as such, it makes sense to consider smoother domains.

Furthermore, when considering oblique boundary-value problems (which arise from the linearisation of the MA optimal transport problem (3.2.2)), we have seen that a key assumption for well-posedness is that $\partial_{\mathbf{T}_2} \Theta + \mathcal{H}_{\partial\Omega} > 0$ on $\partial\Omega$, and that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$, where Θ is defined in Definition 3.3.24, and $\mathcal{H}_{\partial\Omega}$ is the mean curvature of $\partial\Omega$. However, it is not immediately clear that such a condition can hold on a polygonal domain.

Overall, this example serves to highlight potential pitfalls for the well-posedness of MA equations, and the lack of smoothness of solutions, on domains with flat boundaries. This can be remedied if we consider FEMs that allow for curved boundaries, leading us to the work of C. Bernardi [16] (see also [33, 35, 36]).

4.2 New contributions and existing results

The goal of this chapter is to prove various finite element estimates in curved Lagrange finite element spaces, that will be utilised in Chapters 5 to 8. The results we prove are listed as follows (note that more detail on each of these estimates will be provided later in this Chapter).

1. Inverse estimates in H^s and $W^{s,\infty}$ norms (and seminorms) for integer s.

2. Optimal interpolation estimates in H^s and $W^{s,\infty}$ norms for integer s.

3. Optimal interpolation estimates in H^s norms (both simplicial and trace) for non integer s.

4. A two-dimensional discrete Sobolev inequality, bounding the L^∞ norm in terms of the H^1-norm (with a ln factor of the meshsize, h).

5. Stability estimates for the L^2-projection operator.

6. Tangential operator identities (i.e., the tangential gradient and tangential Laplacian).

7. Curvature bounds for curved simplices.

Many of these results are known in the case that $\Omega \subset \mathbb{R}^d$ is a polytopal domain, where one may triangulate Ω *exactly* by *straight* d-simplices.
Known results in the polytopal case:

- Inverse estimates in $W^{s,q}$ norms (and seminorms) for $s \in \mathbb{N}_0$, $q \in [1, \infty]$ are proven in [23, 33].

- Optimal interpolation estimates $W^{s,q}$ norms for integer s, $q \in [1, \infty]$ are proven in [23, 33].

- Optimal interpolation estimates in H^s norms (both simplicial and trace) for non integer s in [5].

- The discrete Sobolev inequality is proven in [24].

- Stability estimates for the L^2-projection operator are proven in [45].

- The tangential operator identities are proven in [110].

Known results in the curved case:

- Optimal interpolation estimates in $W^{m,p}$ norms for integer m are proven in [16], where the domain approximation is *exact* $1 \leq p \leq \infty$, and in [35] in the case of isoparametric finite elements $1 \leq p < \infty$.

- Important scaling arguments required for several of the results of this Chapter are proven in [16].

- Similar inverse estimates are proven in [33], in H^s norms for $s = 1, 2, 3$ for the case of quadratic isoparametric finite elements.

We also remark that the results on curvature bounds for curved simplices (result 1 from the list above) do not appear present in the current literature on finite element theory, most likely due to the fact that the polytopal counterpart of this is trivial, since polytopal domains, and their simplicial approximations are not curved. Furthermore, the inverse estimates (result 7 from the list above) that we prove exhibit a structure that is different to their polytopal counterparts. This of particular significance when considering inverse estimates in *seminorms* (see the discussion that follows Remark 4.4.16 for further details), and adds extra technical details to the proofs present in Chapters 5 and 7.

One may characterise the curvature bounds for curved simplices and inverse estimates as new results, whereas results 2 to 6 from the list above, may be viewed as reviewing estimate that are well known in the polytopal case, in the context of curved Lagrange finite elements (indeed, in each case, though the results are largely the same, there are technical differences in the corresponding proofs).

4.3 Notation

Definition 4.3.1 (Face and vertex sets) *Given a mesh $\mathscr{T}_h$, we denote by $\mathscr{E}_h^{i,b}$, the set of faces of $\mathscr{T}_h$, by $\mathscr{E}_h^i$ the set of interior faces of $\mathscr{T}_h$, and by $\mathscr{E}_h^b$, the set of boundary faces. In the case that the dimension, $d = 2$, we denote by $\mathscr{V}_h^b$ the set of boundary vertices of $\mathscr{T}_h$.*

Remark 4.3.2 *The boundary vertex set, $\mathscr{V}_h^b$, will only be used in Chapters 6 and 7 in the context of two-dimensional oblique boundary-value problems. Note, however, that it is only used to prove particular statements, and is not necessary for computations.*

Definition 4.3.3 (Jump and average operators) *For each face $F \in \mathscr{E}_h^{i,b}$, we have that $F = \overline{K} \cap \overline{K'}$ for some $K, K' \in \mathscr{T}_h$ (in the case that $F \in \mathscr{E}_h^b$ take $F = \overline{K} \cap \partial\Omega$), with corresponding unit normal vector n_F (which, for convention, is chosen so that it is the outward normal to K, see Figure 4.1), we define the jump operator, $[\![\cdot]\!]$, over F by*

$$[\![\phi]\!] = \begin{cases} (\phi|_K)|_F - (\phi|_{K'})|_F \text{ if } F \in \mathscr{E}_h^i, \\ (\phi|_K)|_F \text{ if } F \in \mathscr{E}_h^b, \end{cases} \tag{4.3.1}$$

and the average operator, $\langle\!\langle \cdot \rangle\!\rangle$, by

$$\langle\!\langle \phi \rangle\!\rangle = \begin{cases} \frac{1}{2}((\phi|_K)|_F + (\phi|_{K'})|_F) \text{ if } F \in \mathscr{E}_h^i, \\ (\phi|_K)|_F \text{ if } F \in \mathscr{E}_h^b. \end{cases} \tag{4.3.2}$$

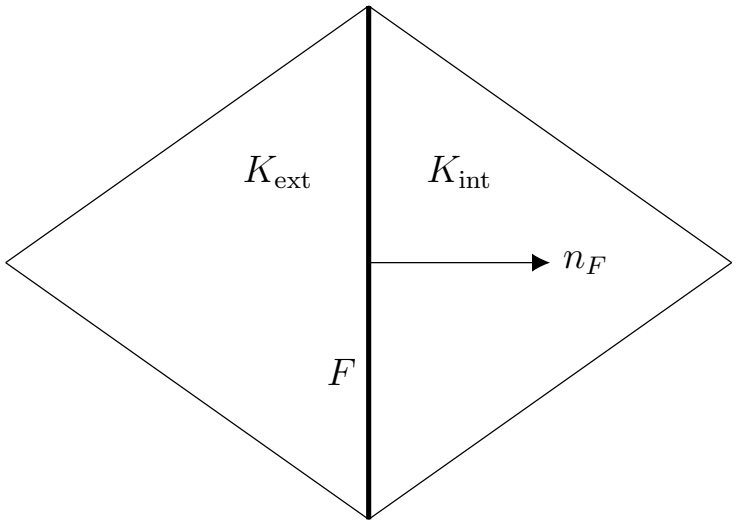


Figure 4.1: Example of two triangles K_{int} and K_{ext} of the mesh, and a face $F = \overline{K_{\text{int}}} \cap \overline{K_{\text{ext}}}$. The unit normal, n_F, is chosen so that n_F is outward pointing for K_{ext}.

For each $e \in \mathscr{V}_h^b$, $e = \overline{F} \cap \overline{F'}$ for some $F, F' \in \mathscr{E}_h^b$, and so we define the jump and average of a function over a vertex analogously to (4.3.1) and (4.3.2).

Definition 4.3.4 (Element L^2-inner product) *For an element K, we define the inner product $\langle \cdot, \cdot \rangle_K$ by*

$$\langle u, v \rangle_K := \begin{cases} \displaystyle\int_K u\, v \ \textit{if } u, v \in L^2(K), \\ \displaystyle\int_K u \cdot v \ \textit{if } u, v \in L^2(K; \mathbb{R}^d), \\ \displaystyle\int_K u : v \ \textit{if } u, v \in L^2(K; \mathbb{R}^{d \times d}). \end{cases} \tag{4.3.3}$$

Any ambiguity in this notation will be resolved by the arguments of the bilinear form. The bilinear forms $\langle \cdot, \cdot \rangle_{\partial K}$ and $\langle \cdot, \cdot \rangle_F$ for $F \in \mathscr{E}_h^{i,b}$, are defined similarly.

4.4 Non-affine meshes

Definition 4.4.1 (Curved d-simplex) *An open set $K \subset \mathbb{R}^d$ is called a curved d-simplex if there exists a C^1 mapping F_K that maps a straight reference d-simplex $\hat{K}$ onto K, and that is of the form*

$$F_K = \tilde{F}_K + \Phi_K, \tag{4.4.1}$$

where

$$\tilde{F}_K : \hat{x} \mapsto \tilde{B}_K \hat{x} + \tilde{b}_K \tag{4.4.2}$$

is an invertible map and $\Phi_K \in C^1(\hat{K};\mathbb{R}^d)$ *satisfies*

$$C_K := \sup_{\hat{x}\in\hat{K}} \|D\Phi_K(\hat{x})\tilde{B}_K^{-1}\| < 1, \tag{4.4.3}$$

where $\|\cdot\|$ *denotes the induced Euclidean norm on* $\mathbb{R}^{d\times d}$.

Definition 4.4.2 (Associated straight d-simplex) *Given a curved d-simplex K, with the associated straight reference d-simplex $\hat{K}$, and map $F_K : \hat{K} \to K$, with $F_K = \tilde{F}_K + \Phi_K$, we define the* associated straight d-simplex*:*

$$\tilde{K} := \tilde{F}_K(\hat{K}).$$

Remark 4.4.3 *The associated d-simplex, $\tilde{K}$, is a straight d-simplex that "approximates" K.*

Lemma 4.4.4 (Affine invariance of C_K) *Given a d-simplex triple $(K,\hat{K},\tilde{K})$, another reference d-simplex $\hat{K}'$, and a map $\tilde{F}_{K'} \in GL(\mathbb{R}^d)$ that maps $\hat{K}'$ onto $\hat{K}$, there is a map $F_{K'} : \hat{K}' \to K$ that also satisfies (4.4.3). Moreover, $C_{K'} = C_K$.*

Proof: This proof is given in [16], but for clarity, we shall provide it. Take a map $\tilde{F}_{K'} \in GL(\mathbb{R}^d)$ as in the statement of the lemma. Then, it must be of the form $\tilde{F}_{K'}(\cdot) = \hat{A}(\cdot) + \hat{a}$, where $\hat{A}$ is an invertible matrix, and $\hat{a} \in \mathbb{R}^d$. We then define $F'_K(\cdot) = F_K(\hat{A}(\cdot) + \hat{a})$, which maps $\hat{K}'$ onto $\hat{K}$, and is of the form

$$F'_K(\cdot) = (\tilde{B}_K\hat{A}(\cdot) + \tilde{B}_K\hat{a} + \tilde{b}_K) + \Phi'_K(\cdot), \quad \text{with} \quad \Phi'_K(\cdot) = \Phi_K(\hat{A}(\cdot) + \hat{a}),$$

which gives us

$$C_{K'} := \sup_{\hat{x}'\in\hat{K}'} \|D\Phi'_K(\hat{x}')\cdot(\tilde{B}_K\hat{A}(\cdot))^{-1}\| = \sup_{\hat{x}\in\hat{K}} \|D\Phi_K(\hat{x})\hat{A}\cdot\hat{A}^{-1}\tilde{B}_K^{-1}\| = C_K. \qquad \square$$

Remark 4.4.5 *Lemma 4.4.4 is rather significant, as it means that given $K \in \mathscr{T}_h$ with an approximating straight simplex $\tilde{K}$, we can always redefine the map F_K (without relabelling it), so that $F_K : \hat{K} \to K$, and $\hat{K} := \tilde{K}/\mathrm{diam}(\tilde{K})$, and the reference simplex enjoys similar shape regularity to $\tilde{K}$. So, in general, we assume a given reference simplex is of this form. This fact will be used in the proof of Corollary 4.4.25.*

Remark 4.4.6 (Affine mesh) *In the case that the domain has a flat boundary, one employs an affine approximation of the domain, in which case, the corresponding functions Φ_K in (4.4.1) are all zero.*

Definition 4.4.7 (Mesh size) *For each $K \in \mathscr{T}_h$, let $h_K := \operatorname{diam}(\tilde{K}) \geq C(d)\|\tilde{B}_K\|$ (where $\tilde{K} = \tilde{B}_K(\hat{K})$). It is assumed that $h = \max_{K\in\mathscr{T}_h} h_K$ for each mesh $\mathscr{T}_h$.*

Definition 4.4.8 (Face-mesh size) *For each face $F \in \mathscr{E}_h^{i,b}$, we define*

$$\tilde{h}_F := \begin{cases} \min(h_K, h_{K'}) & \text{if } F \in \mathscr{E}_h^i, \\ h_K & \text{if } F \in \mathscr{E}_h^b. \end{cases} \tag{4.4.4}$$

where K and K' are such that $F = \partial K \cap \partial K'$ if $F \in \mathscr{E}_h^i$, or $F \subset \partial K \cap \partial\Omega$ if $F \in \mathscr{E}_h^b$.

Mesh conditions. We shall adopt the following assumptions on the meshes.

Assumption 4.4.9 *We assume that any two elements sharing a face have commensurate diameters, i.e., there is a $C_{\mathscr{T}} \geq 1$, independent of h, such that*

$$\max(h_K, h_{K'}) \leq C_{\mathscr{T}} \min(h_K, h_{K'}), \tag{4.4.5}$$

for any K and K' in $\mathscr{T}_h$ that share a face.

Finally, we assume that each $F \in \mathscr{E}_h^b$ satisfies

$$F = F \cap \Gamma_n, \tag{4.4.6}$$

for some $n \in \{1, \ldots, N\}$, with Γ_n given as in (2.5.1). This implies that each boundary face is completely contained in a boundary portion Γ_n, as well as ensuring that our approximation of the domain Ω is exact.

Remark 4.4.10 *The assumptions on the mesh given by Assumption 4.4.9, in particular (4.4.5), show that if F is a face of K, then*

$$h_K \leq C_{\mathscr{T}} \tilde{h}_F. \tag{4.4.7}$$

Definition 4.4.11 (Class m curved d-simplex) *A curved d-simplex K is of class C^m, $m \geq 1$, if the mapping F_K is of class C^m on $\hat{K}$.*

The proofs of the next four lemmas can be found in [16] (i.e., Lemmas 2.1, 2.2, 2.3 and 2.4).

Lemma 4.4.12 *The mapping F_K is a C^1-diffeomorphism from $\hat{K}$ onto K and satisfies*

$$\sup_{\hat{x}\in\hat{K}} \|DF_K(\hat{x})\| \leq (1 + C_K)\|\tilde{B}_K\|, \tag{4.4.8}$$

$$\sup_{x\in K} \|DF_K^{-1}(x)\| \leq (1 - C_K)^{-1}\|\tilde{B}_K^{-1}\|, \tag{4.4.9}$$

$$\forall \hat{x} \in \hat{K}, \quad (1 - C_K)^d |\det \tilde{B}_K| \leq |\det DF_K(\hat{x})| \leq (1 + C_K)^d |\det \tilde{B}_K|. \tag{4.4.10}$$

Lemma 4.4.13 *Let us denote by c_ℓ, $2 \le \ell \le m$, $m \in \mathbb{N}$, the constants*

$$c_\ell(K) := \sup_{\hat{x}\in\hat{K}} \|D^\ell F_K(\hat{x})\| \|\tilde{B}_K\|^{-\ell}. \tag{4.4.11}$$

There exist constants $c_{-\ell}$, $2 \le \ell \le m$, depending continuously on c_K, $c_2(K), \ldots, c_m(K)$, such that

$$\sup_{x\in K} \|D^\ell F_K^{-1}(x)\| \le c_{-\ell} \|\tilde{B}_K\|^{2(\ell-1)} \|\tilde{B}_K^{-1}\|^\ell. \tag{4.4.12}$$

Lemma 4.4.14 *Assume that K is a curved d-simplex of class C^m. Let k be an integer, $0 \le k \le m$, and $q \in \{2, \infty\}$. A function v belongs to $W^{m,q}(K)$ if and only if the function $\hat{v} := v \circ F_K$ belongs to $W^{m,q}(\hat{K})$. We also have for any $v \in W^{m,q}(K)$*

$$|v|_{W^{k,q}(K)} \le C |\det \tilde{B}_K|^{1/q} \|\tilde{B}_K^{-1}\|^k \left(\sum_{r=\min\{k,1\}}^{k} \|\tilde{B}_K\|^{2(k-r)} |\hat{v}|_{W^{r,q}(\hat{K})} \right), \tag{4.4.13}$$

$$|\hat{v}|_{W^{k,q}(\hat{K})} \le C |\det \tilde{B}_K|^{-1/q} \|\tilde{B}_K\|^k \left(\sum_{r=\min\{k,1\}}^{k} |v|_{W^{r,q}(K)} \right), \tag{4.4.14}$$

where the constants C depend continuously on c_K, $c_2(K), \ldots, c_m(K)$.

Lemma 4.4.15 *Assume that K is a curved d-simplex of class C^m, and that F is a face of K; we denote by $\tilde{B}_F$ the restriction of $\tilde{B}_K$ to $\hat{F} := F_K^{-1}(F)$. Let k be an integer, $1 \le k \le m$, $s \in [0, k - 1/2)$. Then, for any $v \in H^k(K)$, the function $\tau_F(v)$ belongs to $H^s(F)$, and we have*

$$\|v\|_{H^s(F)} \le C |\det \tilde{B}_F|^{1/2} |\det \tilde{B}_K|^{-1/2} \|\tilde{B}_K^{-1}\|^s (\|v\|_{L^2(K)} + \|\tilde{B}_K\|^k |v|_{H^k(K)}), \tag{4.4.15}$$

where the constant C depends continuously on c_K, $c_2(K), \ldots c_m(K)$.

Remark 4.4.16 (Interpretation of the Lemmas 4.4.13–4.4.15) *One may view the seemingly rather abstract Lemmas 4.4.13, 4.4.14, and 4.4.15 as a necessary prerequisite for the standard scaling argument used to prove optimal interpolation estimates on affine meshes, inverse inequalities in $W^{m,p}$-norms, and trace inequalities in H^s-norms, respectively. In particular, the following example should demonstrate the importance of having a uniform upper bound on the value c_ℓ.*

A key tool in the derivation of optimal interpolation estimates on affine meshes is the following scaling argument (see Theorem 3.1.2 of [33]): for $k \in \mathbb{N}_0$, $p \in [1, \infty]$, assuming $v \in W^{k,p}(\tilde{K})$, and $\hat{v} := v \circ F_K \in W^{k,p}(\hat{K})$, we have

$$|\hat{v}|_{W^{k,p}(\hat{K})} \le C \|\tilde{B}_K\|^k |\det \tilde{B}_K|^{-1/p} |v|_{W^{k,p}(\tilde{K})}. \tag{4.4.16}$$

Here, we are considering the *affine equivalent* straight d-simplices $\hat{K}$ and $\tilde{K}$, and an invertible affine map F_K. That is, $\tilde{K} = F_K(\hat{K})$, where F_K is of the form (4.4.1) with $\Phi_K \equiv 0$.

One can see that (4.4.16) and (4.4.14) are similar. The main difference is the presence of the lower order seminorms on the right-hand side of (4.4.14).

Let us look at the particular example of the H^2-seminorm when F_K is not affine. The chain rule, and the multivariable change of variables formula yields

$$\begin{aligned}
|\hat{v}|^2_{H^2(\hat{K})} &= \int_{\hat{K}} |D^2\hat{v}] \\
&= \int_{\hat{K}} |\nabla(D(v \circ F_K))|^2 \\
&= \int_{\hat{K}} |\nabla((Dv \circ F_K)DF_K)|^2 \\
&= \int_{\hat{K}} \left|[D((Dv \circ F_K)DF_K)]^T\right|^2 \\
&= \int_{\hat{K}} \left|[D^2v \circ F_K(DF_K)^2 + Dv \circ F_K D^2F_K]^T\right|^2 \\
&= \int_{\hat{K}} \left|((DF_K)^2)^T(D^2v)^T \circ F_K + (D^2F_K)^T(Dv \circ F_K)^T\right|^2 \\
&= \int_{\hat{K}} \left|((DF_K)^2)^T(D^2v \circ F_K) + (D^2F_K)^T(\nabla v \circ F_K)\right|^2 \\
&\leq C(d) \int_{\hat{K}} \|DF_K\|^4 |D^2v \circ F_K|^2 + \|D^2F_K\|^2 |\nabla v \circ F_K|^2 \\
&\leq C(d) \left(\sup_{\hat{x}\in\hat{K}} \|DF_K(\hat{x})\|^4 \int_{\hat{K}} |(D^2v) \circ F_K|^2 \right. \\
&\qquad\qquad \left. + \sup_{\hat{x}\in\hat{K}} \|D^2F_K(\hat{x})\|^2 \int_{\hat{K}} |(\nabla v) \circ F_K|^2 \right) \\
&\leq C(d) \sup_{x\in K} |\det DF_K^{-1}(x)| \left(\sup_{\hat{x}\in\hat{K}} \|D^2F_K(\hat{x})\|^2 |v|^2_{H^1(K)} \right. \\
&\qquad\qquad \left. + \sup_{\hat{x}\in\hat{K}} \|DF_K(\hat{x})\|^4 |v|^2_{H^2(K)} \right).
\end{aligned} \tag{4.4.17}$$

Thus, taking square roots, we obtain

$$\begin{aligned}
|\hat{v}|_{H^2(\hat{K})} \leq C(d) \sup_{x\in K} |\det DF_K^{-1}(x)|^{1/2} &\left(\sup_{\hat{x}\in\hat{K}} \|D^2F_K(\hat{x})\| |v|_{H^1(K)} \right. \\
&\left. + \sup_{\hat{x}\in\hat{K}} \|DF_K(\hat{x})\|^2 |v|_{H^2(K)} \right).
\end{aligned} \tag{4.4.18}$$

Note that if F_K were affine, then $DF_K = \tilde{B}_K$, $DF_K^{-1} = \tilde{B}_K^{-1}$, and $D^2F_K \equiv 0$, thus from the above, we immediately obtain (4.4.16) with $k = p = 2$.

A sufficient assumption that yields an estimate of the same order as (4.4.16) with $k = p = 2$ (in terms of $\|\tilde{B}_K\|$), is to assume that c_ℓ, given by (4.4.11), is uniformly bounded for $\ell = 2$. This, coupled with the fact that $C_K < 1$ gives us

$$\begin{aligned}\sup_{\hat{x}\in\hat{K}} \|DF_K(\hat{x})\| = \sup_{\hat{x}\in\hat{K}} \|(DF_K(\hat{x})\tilde{B}_K^{-1})\tilde{B}_K\| &= \sup_{\hat{x}\in\hat{K}} \|(I_d + D\Phi_K\tilde{B}_K^{-1})\tilde{B}_K\| \\ &\leq (1 + C_K)\|\tilde{B}_K\|,\end{aligned}$$

and

$$\sup_{\hat{x}\in\hat{K}} \|D^2F_K(\hat{x})\| = (\sup_{\hat{x}\in\hat{K}} \|D^2F_K(\hat{x})\|\tilde{B}_K\|^{-2})\|\tilde{B}_K\|^2 = c_2\|\tilde{B}_K\|^2.$$

Applying these two estimates to (4.4.18) yields

$$|\hat{v}|_{H^2(\hat{K})} \leq C \sup_{x\in K} |\det DF_K^{-1}(x)|^{1/2}\|\tilde{B}_K\|^2(|v|_{H^1(K)} + |v|_{H^2(K)}).$$

In order to appropriately bound the determinant term, one must note that $DF_K^{-1} = (DF_K)^{-1}$, and so

$$|\det DF_K^{-1}| = |\det DF_K|^{-1} \leq |\det \tilde{B}_K|^{-1}(1 - C_K)^d.$$

Ultimately, this gives us

$$|\hat{v}|_{H^2(\hat{K})} \leq C|\det \tilde{B}_K|^{-1/2}\|\tilde{B}_K\|^2(|v|_{H^1(K)} + |v|_{H^2(K)}). \tag{4.4.19}$$

This motivates the two definitions that proceed the following remark, generalising the prerequisite assumptions, allowing one to obtain analogous estimates in higher order seminorms.

Remark 4.4.17 (Minor abuse of notation) *One may notice that from the seventh equality of (4.4.17) the "D^2" applied to v is different to the "D^2" applied to F_K. In particular, the first "D^2" is the Hessian, and the latter "D^2" is the second order derivative defined in Definition 2.2.2. This stems from the fact that we define the Hessian of v, D^2v, in the following way:*

$$D^2v = \nabla(Dv) = [D(Dv)]^T = [D^2v]^T,$$

where the final D^2 above is the second derivative defined in Definition 2.2.2.

Definition 4.4.18 *The family $(\mathscr{T}_h)_h$ of meshes is said to be regular if there exist two constants, σ and c, independent of h, such that, for each h, any $K \in \mathscr{T}_h$ satisfies*

$$h_K/\rho_K \leq \sigma, \tag{4.4.20}$$

where ρ_K is the diameter of the sphere inscribed in $\tilde{K}$. Furthermore, we have

$$\sup_h \sup_{K\in\mathscr{T}_h} C_K \leq c < 1. \tag{4.4.21}$$

Definition 4.4.19 *The family of meshes $(\mathscr{T}_h)_h$ is said to be quasi-uniform, if there exists $\rho \in (0,1]$ such that*

$$\min_{K\in\mathscr{T}_h} \rho_K \geq \rho\, h, \tag{4.4.22}$$

for all $h > 0$.

Remark 4.4.20 *Condition (4.4.20) is referred to as* nondegeneracy *(for example in [24]).*

Definition 4.4.21 *The family $(\mathscr{T}_h)_h$ of meshes is said to be regular of order m if it is regular and if, for each h, any $K \in \mathscr{T}_h$ is of class C^{m+1}, with*

$$\sup_h \sup_{K\in\mathscr{T}_h} \sup_{\hat{x}\in\hat{K}} \|D^l F_K(\hat{x})\| \|\tilde{B}_K\|^{-l} < \infty, \quad 2 \leq l \leq m+1. \tag{4.4.23}$$

Definition 4.4.22 (Curved and flat partition of $\mathscr{T}_h$) *Given a mesh $\mathscr{T}_h$, we denote by $\mathscr{T}_h^f$ the set of all $K \in \mathscr{T}_h$ that satisfy $D^2 F_K \equiv 0$, and define $\mathscr{T}_h^c := \mathscr{T}_h \setminus \mathscr{T}_h^f$.*

Definition 4.4.23 (Curved and flat partition of $\mathscr{E}_h^b$) *Given a mesh $\mathscr{T}_h$, we denote by $\mathscr{E}_h^{b,c}$ the set of all faces $F \in \mathscr{E}_h^b$ that satisfy $F \subset \partial K$, where $K \in \mathscr{T}_h^c$, and define $\mathscr{E}_h^{b,f} := \mathscr{E}_h^b \setminus \mathscr{E}_h^{b,c}$.*

A final, necessary step, before providing optimal interpolation estimates and inverse estimates for (continuous and discontinuous) curved Lagrange finite element spaces, is to relate the estimates of this section to the local mesh size, h_K. The general rule of thumb in this context is that $\|\tilde{B}_K\|$ is of order h_K, and $\|\tilde{B}_K^{-1}\|$ is of order h_K^{-1}. This notion is made more concrete by the following theorem from [33].

Theorem 4.4.24 *Let $\hat{K}$ and $\tilde{K} = \tilde{F}_K(\hat{K})$ be two affine-equivalent open subsets of $\mathbb{R}^d$, where $\tilde{F}_K : \hat{x} \to \tilde{B}_K\hat{x} + \tilde{b}_K$ is an invertible affine mapping. Then we have the upper bounds*

$$\|\tilde{B}_K\| \leq \frac{h(\tilde{K})}{\rho(\hat{K})} \quad \textit{and} \quad \|\tilde{B}_K^{-1}\| \leq \frac{h(\hat{K})}{\rho(\tilde{K})}, \tag{4.4.24}$$

where, for a given open subset E of $\mathbb{R}^d$, we define

$$\begin{aligned} &h(E) = \operatorname{diam}(E), \\ &\rho(E) = \sup\{\operatorname{diam}(S) : S \text{ is a ball contained in } E\}. \end{aligned} \tag{4.4.25}$$

Corollary 4.4.25 *Assume that the family $(\mathscr{T}_h)_h$ of meshes satisfies (4.4.20). Then, there exists a positive constant C depending only on σ, such that for any $K \in \mathscr{T}_h$ with an associated straight element $\tilde{K}$, that*

$$\|\tilde{B}_K\| \leq Ch_K \quad \text{and} \quad \|\tilde{B}_K^{-1}\| \leq Ch_K^{-1}. \tag{4.4.26}$$

Proof: Firstly, by Lemma 4.4.4, we may, without loss of generality, assume that the reference simplex $\hat{K} = \frac{1}{\operatorname{diam}(\tilde{K})}\tilde{K}$. Thus, $\tilde{K}$ and $\hat{K}$ are affine equivalent. We apply (4.4.24), which gives us

$$\|\tilde{B}_K\| \leq \frac{h(\tilde{K})}{\rho(\hat{K})} \quad \text{and} \quad \|\tilde{B}_K^{-1}\| \leq \frac{h(\hat{K})}{\rho(\tilde{K})}.$$

Recall that we define $h_K := h(\tilde{K})$, and $\rho_K := \rho(\tilde{K})$ and so we have $\|\tilde{B}_K\| \leq h_K/\rho(\hat{K})$. By (4.4.20), we also have $\|\tilde{B}_K^{-1}\| \leq h(\hat{K})/\rho(\tilde{K}) \leq \sigma h(\hat{K})h_K^{-1} = \sigma h_K^{-1}$. To obtain the first estimate of (4.4.26), we note that

$$\begin{aligned} \rho(\hat{K}) &= \sup\{\operatorname{diam}(S) : S \text{ is a ball contained in } \hat{K}\} \\ &= \sup\left\{\operatorname{diam}(S) : S \text{ is a ball contained in } \frac{1}{\operatorname{diam}\tilde{K}}\tilde{K}\right\} \\ &= \frac{1}{\operatorname{diam}\tilde{K}}\rho(\tilde{K}) \geq \frac{h(\tilde{K})}{\sigma h(\tilde{K})} = \frac{1}{\sigma} \quad \text{(by 4.4.20).} \end{aligned}$$

Thus $\|\tilde{B}_K\| \leq \sigma h_K$. $\square$

Definition 4.4.26 (v, $\hat{v}$, and v^*) *Given a triple $(K^*, \hat{K}, K)$ (fixed reference simplex, reference simplex, and curved simplex), a pair of invertible maps $(G_K : K^* \to \hat{K}, F_K : \hat{K} \to K)$, and a function $v : K \to \mathbb{R}^N$, for some $N \in \mathbb{N}$, we define the functions $\hat{v} : \hat{K} \to \mathbb{R}^N$, $v^* : K^* \to \mathbb{R}^N$, as follows:*

$$\hat{v} := v \circ F_K, \quad v^* := \hat{v} \circ G_K = v \circ F_K \circ G_K. \tag{4.4.27}$$

Furthermore, given $v^* : K^* \to \mathbb{R}^N$, we also define

$$\hat{v} := v^* \circ G_K^{-1}, \quad v := \hat{v} \circ F_K^{-1} = v^* \circ G_K^{-1} \circ F_K^{-1}. \tag{4.4.28}$$

4.5 Finite element spaces and optimal interpolation estimates

The finite element spaces we consider in the thesis either consist of *continuous* piecewise polynomial functions, or *discontinuous* piecewise polynomial functions, which fall into the class of continuous and discontinuous (curved) Lagrange finite element spaces, respectively. In order to define a finite element space, we must define a *finite element*. In general, a finite element is a triple (K, P_K, Σ_K) where K is a subset of $\mathbb{R}^d$, P_K is a finite dimensional space on K, and Σ_K is a set of continuous linear forms on P_K, which we will call the degrees of freedom. In the context of Lagrange finite element spaces, the continuous linear forms are given by (local) point evaluations. In the simplicial case, the placement of these points is naturally described using the barycentric coordinates of the simplex.

Definition 4.5.1 (Barycentric coordinates) *Given a straight d-simplex $\hat{K}$, with vertices $\hat{a}_1, \ldots, \hat{a}_{d+1} \in \mathbb{R}^d$, we define the barycentric coordinates of $\hat{K}$, $\lambda_1, \ldots, \lambda_d, \lambda_{d+1}$ via the following (invertible) system*

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ (\hat{a}_1)_1 & (\hat{a}_2)_1 & \ldots & (\hat{a}_{d+1})_1 \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{a}_1)_d & (\hat{a}_2)_d & \ldots & (\hat{a}_{d+1})_d \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{d+1} \end{bmatrix} = \begin{bmatrix} 1 \\ \hat{x}_1 \\ \vdots \\ \hat{x}_d \end{bmatrix}, \tag{4.5.1}$$

where $\hat{x} = (x_1, \ldots, x_d)^T \in \hat{K}$.

Definition 4.5.2 (Straight Lagrange finite element) *For a straight d-simplex $\hat{K}$ with vertices $\hat{a}_1, \ldots, \hat{a}_{d+1} \in \mathbb{R}^d$, with barycentric coordinates $\lambda_1, \ldots, \lambda_{d+1}$, we set*

$$J(p) = \{\alpha \in \mathbb{N}_0^{d+1} : |\alpha| = p\}, \tag{4.5.2}$$

and for any $\alpha \in J(p)$, we associate the point $\hat{a}_\alpha \in \hat{K}$ with barycentric coordinates $\lambda_i = \alpha_i/p$, $i = 1, \ldots d+1$. Then, we call $(\hat{K}, \hat{P}_K, \hat{\Sigma}_K)$ a straight Lagrange finite element of type p, where

$$\hat{P}_K = \mathbb{P}^p(\hat{K}), \quad \hat{\Sigma}_K = \{\hat{\mu}_\alpha, \alpha \in J(p)\}, \tag{4.5.3}$$

with $\hat{\mu}_\alpha(\hat{f}) := \hat{f}(\hat{a}_\alpha)$, for $f \in \hat{P}_K$, and we recall that $\mathbb{P}^p(K)$ is the space of all polynomials with total degree less than or equal to p.

Definition 4.5.3 (Curved Lagrange finite element) *The triple (K, P_K, Σ_K) is a curved Lagrange finite element of type (m,p) if K is a curved d-simplex of class C^{m+1}, and*

$$P_K = \{\rho = \hat{\rho} \circ F_K^{-1}, \hat{\rho} \in \hat{P}_K = \mathbb{P}^p(\hat{K})\}, \tag{4.5.4}$$

$$\Sigma_K = \{\mu : \forall v \in C^0(\overline{K}), \mu(v) = \hat{\mu}(v \circ F_K),\ \hat{\mu} \in \hat{\Sigma}_K\}, \tag{4.5.5}$$

where $(\hat{K}, \hat{P}_K, \hat{\Sigma}_K)$ is a straight Lagrange finite element of type p.

Definition 4.5.4 (Discontinuous Galerkin finite element space) *The discontinuous Galerkin finite element space $V_{h,p}$ is defined by*

$$V_{h,p} := \{v \in L^2(\Omega) :\ v|_K = \hat{\rho} \circ F_K^{-1},\ \hat{\rho} \in \mathbb{P}^p(\hat{K}),\ \forall K \in \mathscr{T}_h\}, \tag{4.5.6}$$

where $p \in \mathbb{N}_0$. We also define the subspace, $V_{h,p,0} := V_{h,p} \cap L_0^2(\Omega)$.

Definition 4.5.5 (Continuous Galerkin finite element space) *The continuous Galerkin finite element space $\mathbb{V}_{h,p}$ is defined by*

$$\mathbb{V}_{h,p} := \{v \in C^0(\overline{\Omega}) :\ v|_K = \hat{\rho} \circ F_K^{-1},\ \hat{\rho} \in \mathbb{P}^p(\hat{K}),\ \forall K \in \mathscr{T}_h\}, \tag{4.5.7}$$

where $p \in \mathbb{N}$. We also define the corresponding zero trace spaces

$$\mathring{\mathbb{V}}_{h,p} := \mathbb{V}_{h,p} \cap H_0^1(\Omega). \tag{4.5.8}$$

Remark 4.5.6 *One could equivalently define $V_{h,p} := \cup_{K \in \mathscr{T}_h} P_K$, where P_K is a curved Lagrange finite element of type (m,p), and then define $\mathbb{V}_{h,p} := V_{h,p} \cap C(\overline{\Omega})$, and $\mathring{\mathbb{V}}_{h,p} := \mathbb{V}_{h,p} \cap H_0^1(\Omega)$.*

Piecewise polynomial functions naturally satisfy a property of piecewise regularity. This is accurately captured by considering the notion of broken Sobolev spaces.

Definition 4.5.7 (Broken Sobolev spaces) *Let $\mathbf{s} = (s_K : K \in \mathscr{T}_h)$ denote a vector of nonnegative real numbers and let $r \in [1, \infty]$.*
The broken Sobolev space $W^{\mathbf{s},r}(\Omega; \mathscr{T}_h)$ is defined by

$$W^{\mathbf{s},r}(\Omega; \mathscr{T}_h) := \{v \in L^2(\Omega) :\ v|_K \in W^{s_K,r}(K)\ \forall K \in \mathscr{T}_h\}. \tag{4.5.9}$$

We denote $H^{\mathbf{s}}(\Omega; \mathscr{T}_h) := W^{\mathbf{s},2}(\Omega; \mathscr{T}_h)$, and set $W^{s,r}(\Omega; \mathscr{T}_h) := W^{\mathbf{s},r}(\Omega; \mathscr{T}_h)$, in the case that $s_K = s,\ s \geq 0$, for all $K \in \mathscr{T}_h$. For $v \in W^{1,r}(\Omega; \mathscr{T}_h)$, let $\nabla_h v \in L^r(\Omega; \mathbb{R}^d)$ denote the discrete (also known as broken) gradient of v, i.e., $(\nabla_h v)|_K = \nabla(v|_K)$ for

all $K \in \mathscr{T}_h$. Higher order discrete derivatives are defined in a similar way. We define a norm on $W^{s,r}(\Omega; \mathscr{T}_h)$ by

$$\|v\|^r_{W^{s,r}(\Omega;\mathscr{T}_h)} := \sum_{K\in\mathscr{T}_h} \|v\|^r_{W^{s,r}(K)} \tag{4.5.10}$$

with the usual modification when $r = \infty$.

Definition 4.5.8 *We define the following for $K \in \mathscr{T}_h$, $s \in \mathbb{N}_0$, $r \in [1, \infty)$:*

$$|v|^r_{W^{s,r}_*(K)} := \sum_{j=\min\{1,s\}}^{s} |v|^r_{W^{j,r}(K)}, \tag{4.5.11}$$

$$|v|^r_{W^{s,r}_*(\Omega;\mathscr{T}_h)} := \sum_{K\in\mathscr{T}_h} |v|^r_{W^{s,r}_*(K)}, \tag{4.5.12}$$

with the usual modification when $r = \infty$. Note that $|\cdot|_{W^{s,r}_*(\Omega;\mathscr{T}_h)}$ is a norm when $s = 0$, and a semi-norm when $s \in \mathbb{N}$. We also define $|\cdot|_{H^s_*(K)}$ and $|\cdot|_{H^s_*(\Omega;\mathscr{T}_h)}$ in the usual way.

Remark 4.5.9 *We have the following norm relation:*

$$|v|^r_{W^{s,r}_*(\Omega;\mathscr{T}_h)} \le \|v\|^r_{W^{s,r}(\Omega;\mathscr{T}_h)} = |v|^r_{W^{0,r}_*(\Omega;\mathscr{T}_h)} + \min\{s, 1\}|v|^r_{W^{s,r}_*(\Omega;\mathscr{T}_h)}, \tag{4.5.13}$$

for any $r \in [1, \infty)$, $s \in \mathbb{N}_0$ (similarly for $r = \infty$). We can use these semi-norms to equivalently phrase estimates such as (4.4.19), which can now be written as

$$|\hat{v}|_{H^2(\hat{K})} \le C|\det \tilde{B}_K|^{-1/2}\|\tilde{B}_K\|^2|v|_{H^2_*(K)}.$$

The proofs of the following lemmas can both be found in [16], i.e., Theorem 4.1 and Corollary 4.1; one must note that they are both given in a more general context. However, we are considering Lagrange finite element spaces, which satisfy the hypotheses of Theorem 4.1 and Corollary 4.1 (see examples 1 and 2 on page 1221 of [16]).

Lemma 4.5.10 (Optimal local interpolation in $V_{h,p}$) *Assume that the family $(\mathscr{T}_h)_h$ is regular of order m. Let $k, \ell, p \in \mathbb{N}_0$, $p \ge 2$, with $\ell \le k \le \min\{p, m\} + 1$. Then for any $K \in \mathscr{T}_h$, and any $u \in H^k(\Omega; \mathscr{T}_h)$, there exists a $z_h \in V_{h,p}$ such that*

$$\|u - z_h\|_{H^\ell(K)} \le Ch_K^{k-\ell}|u|_{H^k_*(K)}, \tag{4.5.14}$$

where the constant C is independent of h_K, u, and K.

Lemma 4.5.11 (Optimal global interpolation in $\mathbb{V}_{h,p}$) *Under the hypotheses of Lemma 4.5.10, let $k, \ell, p \in \mathbb{N}_0$, $p \geq 1$, with $\ell \leq k \leq \min\{p, m\} + 1$. If the family of triangulations $(\mathscr{T}_h)_h$ is quasi-uniform, then, for any $\omega \in H^k(\Omega; \mathscr{T}_h)$, there exists a $z_h \in \mathbb{V}_{h,p}$ such that*

$$\|\omega - z_h\|_{H^\ell(\Omega;\mathscr{T}_h)} \leq Ch^{k-\ell}|\omega|_{H^k_*(\Omega;\mathscr{T}_h)}. \tag{4.5.15}$$

Furthermore, for any $\omega \in W^{k,\infty}(\Omega; \mathscr{T}_h)$, there exists a $z_h \in \mathbb{V}_{h,p}$ such that

$$\|\omega - z_h\|_{W^{\ell,\infty}(\Omega;\mathscr{T}_h)} \leq Ch^{k-\ell}|\omega|_{W^{k,\infty}_*(\Omega;\mathscr{T}_h)}. \tag{4.5.16}$$

In each case, the constant C is independent of h and ω.

Definition 4.5.12 (Classical Lagrange interpolation operator) *A function u in a finite element space $\mathbb{V}_h$ can be represented as follows:*

$$u = \sum_i u_i \phi_i,$$

where each $u_i \in \mathbb{R}$, and the set $\{\phi_i\}$ forms a global basis of $\mathbb{V}_h$ (note that we are summing over the set of all basis functions of the finite element space).

Motivated by this, for $v \in C(\overline{\Omega})$, we define the classical Lagrange interpolation operator $\pi_h : C(\overline{\Omega}) \to \mathbb{V}_h$, by

$$\pi_h(v) = \sum_i v(x_{\phi_i})\phi_i, \tag{4.5.17}$$

where the points $\{x_{\phi_i}\}$ represent the degrees of freedom of $\mathbb{V}_h$.

Furthermore, we can define the interpolant of a function defined only on the boundary of Ω, similarly, by summing over the degrees of freedom that lie on the boundary, and taking the trace of the resulting function.

We note that the classical Lagrange interpolation operator can only applied functions that have well defined point values. Even in two dimensions, it is not in general the case that functions in H^1 have well defined point values. This leads one to define other interpolation operators that require less regularity, in particular, we define a local interpolation operator that is well defined on L^2 functions (one of the first examples is due to P. Clemént [36], using local averaging; however the one we will define is provided in [16] and is slightly different).

Definition 4.5.13 (Local L^2 projection) *For $v \in L^2(\Omega)$, and $K \in \mathscr{T}_h$, we define $\hat{\rho}_v$ to be the unique element of $\mathbb{P}^p(\hat{K})$ that satisfies*

$$\int_{\hat{K}} (\hat{v} - \hat{\rho}_v)\hat{\rho} \quad \forall \hat{\rho} \in \mathbb{P}^p(\hat{K}). \tag{4.5.18}$$

Definition 4.5.14 (Local Lagrange interpolation operator) *For $K \in \mathscr{T}_h$, we define the Lagrange interpolation operator $\Pi_h : L^2(K) \to P_K$, where (K, P_K, Σ_K) is a curved Lagrange finite element of type (m, p), by*

$$\Pi_h(v) = \sum_{\mu \in \Sigma_K} \mu(\rho_v)\rho_\mu, \tag{4.5.19}$$

where $\rho_v = \hat{\rho}_v \circ F_K^{-1}$, with $\hat{\rho}_v$ satisfying (4.5.18), and $\{\rho_\mu\}_{\mu \in \Sigma_K}$ forms a basis of P_K.

Definition 4.5.15 (Global L^2 projection) *For $v \in [L^2(\Omega)]^{m \times n}$, $m, n \in \mathbb{N}$, given a finite element space $V_h \subset [L^2(\Omega)]^{m \times n}$, we define the global V_h-projection operator $\mathcal{P}_{V_h} : [L^2(\Omega)]^{m \times n} \to V_h$, by*

$$\int_\Omega \mathcal{P}_{V_h}(v) : \Phi = \int_\Omega v : \Phi \quad \forall \Phi \in V_h. \tag{4.5.20}$$

Definition 4.5.16 ($\lesssim$ and $\approx$ symbols) *Herein we write $a \lesssim b$ for $a, b \in \mathbb{R}$, if there exists a constant $C > 0$, such that*

$$a \leq Cb,$$

independent of $\mathbf{h} := \{h_K : K \in \mathscr{T}_h\}$, and u, but otherwise possibly dependent on the polynomial degree, p, the shape-regularity constants of $\mathscr{T}_h$, $C_{\mathcal{P}}$, $C_{\mathcal{T}}$, $\mathbf{s}$, and d. Furthermore, we write $a \approx b$ if both $a \lesssim b$ and $b \lesssim a$.

4.6 Inverse and trace estimates

We now state and prove trace and inverse estimates that we will be utilised frequently.

Lemma 4.6.1 *Assume that $\mathscr{T}_h$ is a regular mesh on $\overline{\Omega}$. Then, for any $K \in \mathscr{T}_h$, we have that*

$$\|v\|_{2,\partial K}^2 \lesssim (h_K^{-1}\|v\|_{2,K}^2 + h_K\|\nabla v\|_{2,K}^2) \quad \forall v \in H^1(K). \tag{4.6.1}$$

In particular, if the triangulation is quasi-uniform, we have

$$\sum_{K \in \mathscr{T}_h} \|v\|_{L^2(\partial K)}^2 \lesssim (h^{-1}\|v\|_{2,\Omega}^2 + h\|v\|_{H^1(\Omega;\mathscr{T}_h)}^2) \quad \forall v \in H^1(\Omega;\mathscr{T}_h). \tag{4.6.2}$$

Proof: Applying (4.4.15) of Lemma 4.4.15 with $k = m = 1$ and $s = 0$, for any $K \in \mathscr{T}_h$ and any face F of K, we obtain

$$\|v\|_{L^2(F)}^2 \leq C|\det \tilde{B}_F||\det \tilde{B}_K|^{-1}(\|v\|_{L^2(K)}^2 + \|\tilde{B}_K\|^2|v|_{H^1(K)}^2),$$

where we recall that $\tilde{B}_F$ is the restriction of $\tilde{B}_K$ to $\hat{F} := F_K^{-1}(F)$ (and thus acts as a map on $\mathbb{R}^{d-1}$). Now, applying (4.4.26) yields

$$\|\tilde{B}_K\| \leq Ch_K \quad \text{and} \quad \|\tilde{B}_K^{-1}\| \leq Ch_K^{-1},$$

where the constant C is independent of K and h_K. Thus, as the determinant is a continuous d-linear ($(d-1)$-linear in the case of $\tilde{B}_F$) map, we obtain

$$\begin{aligned} \|v\|_{L^2(F)}^2 &\leq C\|\tilde{B}_K\|^{d-1}\|\tilde{B}_K^{-1}\|^d(\|v\|_{L^2(K)}^2 + \|\tilde{B}_K\|^2|v|_{H^1(K)}) \\ &\leq Ch_K^{-1}(\|v\|_{L^2(K)}^2 + h_K^2|v|_{H^1(K)}) = C(h_K^{-1}\|v\|_{L^2(K)}^2 + h_K|v|_{H^1(K)}). \end{aligned} \tag{4.6.3}$$

Since the number of faces of an element $K \in \mathscr{T}_h$ is uniformly bounded with respect to the dimension, we obtain (4.6.1) by summing (4.6.3) over all faces $F \subset \partial K$. If we assume that the mesh is quasi-uniform, then $h_K \lesssim h$ and $h_K^{-1} \lesssim h^{-1}$ for all $K \in \mathscr{T}_h$, and we obtain (4.6.2) by summing (4.6.3) over all $K \in \mathscr{T}_h$. $\square$

Lemma 4.6.2 (Non integer order trace estimate) *Assume that $\{\mathscr{T}_h\}_h$ is a regular family of triangulations on $\overline{\Omega}$. Then, for any $K \in \mathscr{T}_h$, and any $(d-1)$ face F of K, we have that*

$$\|v\|_{L^2(F)} \leq Ch_K^{-1/2}(\|v\|_{L^2(K)} + h_K^r|v|_{H^r(K)}), \tag{4.6.4}$$

for all $v \in H^r(K)$, $1/2 < r < 1$. Furthermore, the constant C is independent of h_K and the choice of $K \in \mathscr{T}_h$.

Proof: From the multivariable change of variables formula, we obtain

$$\|v\|_{L^2(F)} \leq C|\det \tilde{B}_F|^{1/2}\|\hat{v}\|_{L^2(\hat{F})},$$

where $\tilde{B}_F$ is the restriction of $\tilde{B}_K$ to $\hat{F} = F_K^{-1}(F)$. Under a second change of variables, we obtain

$$\|\hat{v}\|_{L^2(\hat{F})} = |\det \tilde{A}_{\hat{F}}|^{1/2}\|v^*\|_{L^2(F^*)},$$

where F^* is a $(d-1)$-face of a *fixed* reference d-simplex, K^*, and $G_K : K^* \to \hat{K}$, $G_K(x^*) := \tilde{A}_{\hat{K}}x^* + \tilde{a}_{\hat{K}}$, with $\tilde{A}_{\hat{K}} \in GL(\mathbb{R}^d)$, $\tilde{a}_{\hat{K}} \in \mathbb{R}^d$, and $\tilde{A}_{\hat{F}}$ is the restriction of $\tilde{A}_{\hat{K}}$ to $F^* = G_K^{-1}(\hat{F})$.

We apply Theorem 2.3.5, yielding

$$\begin{aligned} \|v^*\|_{L^2(F^*)} &\leq C(K^*, d)(\|v^*\|_{L^2(K^*)} + |v^*|_{H^r(K^*)}) \\ &\leq C(K^*, d)(\chi_1(A_K)\|\hat{v}\|_{L^2(\hat{K})} + \chi_2(A_K)|\hat{v}|_{H^r(\hat{K})}), \end{aligned} \tag{4.6.5}$$

where χ_1, and χ_2 are positive, continuous functions that we will soon provide.

Recall the definition of the non integer H^r-semi norm:

$$|\hat{v}|^2_{H^r(\hat{K})} := \int_{\hat{K}}\int_{\hat{K}} \frac{|\hat{v}(\hat{x}_1)-\hat{v}(\hat{x}_2)|^2}{|\hat{x}_1-\hat{x}_2|^{d+2r}}. \tag{4.6.6}$$

We note that since $\hat{x}_1, \hat{x}_2 \in \hat{K}$,

$$|F_K(\hat{x}_1)-F_K(\hat{x}_2)| \le C(d)\sup_{\hat{x}\in\hat{K}}\|DF_K(\hat{x})\||\hat{x}_1-\hat{x}_2|,$$

which, when applied to (4.6.6), gives us

$$\begin{aligned}\int_{\hat{K}}\int_{\hat{K}} \frac{|\hat{v}(\hat{x}_1)-\hat{v}(\hat{x}_2)|^2}{|\hat{x}_1-\hat{x}_2|^{d+2r}} &= \int_{\hat{K}}\int_{\hat{K}} \frac{(C(d)\sup_{\hat{x}\in\hat{K}}\|DF_K(\hat{x})\|)^{d+2r}|\hat{v}(\hat{x}_1)-\hat{v}(\hat{x}_2)|^2}{(C(d)\sup_{\hat{x}\in\hat{K}}\|DF_K(\hat{x})\||\hat{x}_1-\hat{x}_2|)^{d+2r}} \\ &\le \int_{\hat{K}}\int_{\hat{K}} \frac{(C(d)\sup_{\hat{x}\in\hat{K}}\|DF_K(\hat{x})\|)^{d+2r}|\hat{v}(\hat{x}_1)-\hat{v}(\hat{x}_2)|^2}{|F_K(\hat{x}_1)-F_K(\hat{x}_2)|^{d+2r}} \\ &\le C\|\tilde{B}_K\|^{d+2r}\int_{\hat{K}}\int_{\hat{K}} \frac{|\hat{v}(\hat{x}_1)-\hat{v}(\hat{x}_2)|^2}{|F_K(\hat{x}_1)-F_K(\hat{x}_2)|^{d+2r}}.\end{aligned} \tag{4.6.7}$$

We apply the multivariable change of variables formula once more, obtaining

$$\begin{aligned}|\hat{v}|^2_{H^r(\hat{K})} &\le C\|\tilde{B}_K\|^{d+2r}\int_K\int_K \frac{|v(x_1)-v(x_2)|^2}{|x_1-x_2|^{d+2r}}|\det(DF_K^{-1}(x_1))||\det(DF_K^{-1}(x_2))| \\ &\le C\|\tilde{B}_K\|^{d+2r}\|\tilde{B}_K^{-1}\|^{2d}|v|^2_{H^r(K)}.\end{aligned} \tag{4.6.8}$$

Of course, we also have

$$\|\hat{v}\|_{L^2(\hat{K})} \le C\|\tilde{B}_K^{-1}\|^d\|v\|_{L^2(K)}.$$

We obtain the functions χ_1 and χ_2 in a similar manner, except since G_K is affine, the scaling argument is simpler, and we have that

$$\chi_1(A) = |\det A^{-1}|, \quad \text{and} \quad \chi_2(A) = |\det A^{-1}|^2\|A\|^{d+2r}.$$

From the nondegeneracy condition (4.4.20), it follows (from the proof of Theorem 4.4.20 in [24]) that the collection of the invertible matrices given by the affine maps from K^* to $\hat{K}$ is contained in a *compact* subset $BL := \{B \in GL(\mathbb{R}^d) : |\det B| \ge \varepsilon, |B_{ij}| \le r\}$ of $GL(\mathbb{R}^d)$ (where $GL(\mathbb{R}^d)$ is the set of $d\times d$ invertible matrices), where $\varepsilon = \varepsilon(\sigma, d, K^*)$, and $r = r(K^*)$. That is, if

$$\tilde{G}_{\hat{K}} : K^* \to \hat{K}, \quad K^* \ni x^* \mapsto \tilde{A}_{\hat{K}}x^* + \tilde{a}_{\hat{K}} \in \hat{K},$$

then $\tilde{A}_{\hat{K}} \in BL$. Thus we have

$$\chi_i(\tilde{A}_{\hat{K}})^2 \le \sup_{A\in BL}\chi_i(A)^2 \le C(K^*,\sigma), \quad i = 1,2.$$

Overall, we have obtained

$$\begin{aligned}\|v\|_{L^2(F)} &\leq C(d,\sigma,K^*)|\det \tilde{B}_F|^{\frac{1}{2}}\|\tilde{B}_K^{-1}\|^{\frac{d}{2}}(\|v\|_{L^2(K)} + \|\tilde{B}_K^{-1}\|^{\frac{d}{2}}\|\tilde{B}_K\|^{\frac{d}{2}}\|\tilde{B}_K\|^r|v|_{H^r(K)}) \\ &\leq Ch_K^{-1/2}(\|v\|_{L^2(K)} + h_K^r|v|_{H^r(K)}),\end{aligned}$$

where the final inequality follows from (4.4.26). Furthermore, the estimate is independent of h_K, and the choice of K. Thus, we have obtained the desired estimate. □

Corollary 4.6.3 (Integer and non integer regularity interpolation estimates) *Let $\{\mathscr{T}_h\}_h$ be a family of triangulations on $\overline{\Omega}$ that is regular of order $m \geq 2$. Let $u \in H^{\mathbf{s}}(\Omega;\mathscr{T}_h)$ with $s_K > 5/2$ for all $K \in \mathscr{T}_h$. Then, there exists a $z_h \in V_{h,p}$, $p \geq 2$, and a constant C, independent of u, h_K and p, but dependent on $\max_K s_K$, such that for each $K \in \mathscr{T}_h$, each nonnegative integer $q \leq 2$, and each multi-index β with $|\beta| = q$, we have*

$$\begin{aligned}\|u - z_h\|_{H^q(K)} &\leq Ch_K^{t_K-q}\|u\|_{H^{s_K}(K)}, \\ \|D^\beta(u - z_h)\|_{L^2(\partial K)} &\leq Ch_K^{t_K-q-1/2}\|u\|_{H^{s_K}(K)},\end{aligned} \tag{4.6.9}$$

where $t_K = \min\{p+1, m+1, s_K\}$.

Proof: We will first discuss how we will obtain the second bound of (4.6.9). Since the family of triangulations is regular of order $m \geq 2$, it follows that for any β such that $|\beta| = q$, for a nonnegative integer $q \leq 2$, and any $v \in V_{h,p}$, that $D^\beta(u-v) \in H^{t_K-q}(K)$. In particular, $t_K - q > 5/2 - q \geq 1/2$. Thus, we may apply the trace estimate (4.6.4) with $r_K = t_K - q > 1/2$, obtaining

$$\|D^\beta(u-v)\|_{L^2(\partial K)} \leq Ch_K^{-1/2}(\|D^\beta(u-v)\|_{L^2(K)} + h_K^{r_K}|D^\beta(u-v)|_{H^{r_K}(K)}).$$

Let us assume that there exists a $z_h \in V_{h,p}$ satisfying the first estimate of (4.6.9). Then, setting $v = z_h$ above we obtain

$$\begin{aligned}\|D^\beta(u - z_h)\|_{L^2(\partial K)} &\leq Ch_K^{-1/2}(\|D^\beta(u-z_h)\|_{L^2(K)} + h_K^r|D^\beta(u-z_h)|_{H^{r_K}(K)}) \\ &\leq Ch_K^{-1/2}(h_K^{t_K-q}\|u\|_{H^{s_K}(K)} + h_K^r|D^\beta(u-z_h)|_{H^{r_K}(K)}).\end{aligned} \tag{4.6.10}$$

Thus, to obtain both estimates of (4.6.9), it suffices to prove that the there exists a $z_h \in V_{h,p}$ such that the first estimate of (4.6.9) holds, as well as the following:

$$|u - z_h|_{H^{q+r_K}(K)} = |u - z_h|_{H^{t_K}(K)} \leq Ch_K^{t_K-q-r_K}\|u\|_{H^{s_K}(K)}. \tag{4.6.11}$$

Since, applying the above estimate to (4.6.10), and noting the factor h_K^r in the second inequality of (4.6.10), we obtain the second estimate of (4.6.9). Note that we already

have such bounds in the case that s_K is an integer, and as such, we shall assume from this point on that $s_K \notin \mathbb{N}$.

We will now prove the first estimate of (4.6.9). Let β satisfy $|\beta| = q$ for a nonnegative integer $q \leq 2$, and let K^* be a fixed reference simplex. Then, from (4.4.13) we obtain

$$\begin{aligned} |D^\beta(u - z_h)|_{L^2(K)} &\leq |u - z_h|_{H^q(K)} \\ &\leq C|\det \tilde{B}_K|^{1/2}\|\tilde{B}_K^{-1}\|^q \sum_{j=\min\{1,q\}}^{q} \|\tilde{B}_K\|^{2(q-j)}|\hat{u} - \hat{z}_h|_{H^j(\hat{K})} \\ &\leq C(K^*,\sigma)|\det \tilde{B}_K|^{1/2}\|\tilde{B}_K^{-1}\|^q \sum_{j=\min\{1,q\}}^{q} \|\tilde{B}_K\|^{2(q-j)}|u^* - z_h^*|_{H^j(K^*)}. \end{aligned} \tag{4.6.12}$$

We take the function $z_h \in V_{h,p}$, defined as follows: $z_h|_K = \Pi_h u$ where Π_h is the local interpolation operator, given by (4.5.19). Due to (4.5.18), this operator reproduces polynomials in $\mathbb{P}^p(\hat{K})$, and so we may apply Theorem 5 of [34] in conjunction with Theorem 1.8 of [69] (applying Theorem 1.8 of [69] allows us to consider non-integer Sobolev spaces when applying the Bramble–Hilbert Lemma), obtaining for $\min\{1, q\} \leq j \leq q$

$$\begin{aligned} |u^* - z_h^*|_{H^j(K^*)} &\leq \|u^* - z_h^*\|_{H^{t_K}(K^*)} \\ &\leq C(K^*)|u^*|_{H^{t_K}(K^*)} \\ &\leq C(K^*,\sigma)|\hat{u}|_{H^{t_K}(\hat{K})}, \end{aligned} \tag{4.6.13}$$

where, by assumption $t_K > 5/2 > 2 \geq q$ (note that the final inequality follows from a scaling argument similar to the one used in estimate (4.6.5), noting that K^* and $\hat{K}$ are affine equivalent, and the mesh is shape regular). We now decompose $t_K = \ell_K + r_K$, where $\ell_K \geq 2$ is an integer, and $r_K \in (0, 1)$. We see that

$$|\hat{u}|^2_{H^{t_K}(\hat{K})} = |\hat{u}|^2_{H^{\ell_K + r_K}(\hat{K})} = \int_{\hat{K}}\int_{\hat{K}} \frac{|D^{\ell_K}\hat{u}(\hat{x}_1) - D^{\ell_K}\hat{u}(\hat{x}_2)|^2}{|\hat{x}_1 - \hat{x}_2|^{d+2r_K}} = |D^{\ell_K}\hat{u}|^2_{H^{r_K}(\hat{K})}.$$

Let us recall the formula

$$D^r(f \circ g) = \sum_{i=1}^{r} (D^i f \circ g)\left(\sum_{\alpha \in E(r,i)} c_\alpha \prod_{l=1}^{r} (D^l g)^{\alpha_l}\right), \tag{4.6.14}$$

where $E(r, i)$ is the set

$$E(r,i) := \left\{\alpha \in \mathbb{N}_0^r : |\alpha| = i \quad \text{and} \quad \sum_{l=1}^{r} l\alpha_l = r\right\}, \tag{4.6.15}$$

and the c_α's, $\alpha \in E(m,r)$ are some given constants, bounded independently of h_K. By the triangle inequality, we obtain

$$|D^{\ell_K}\hat{u}|_{H^{r_K}(\hat{K})} \leq \sum_{i=1}^{\ell_K} \left| (D^i u \circ F_K) \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right|_{H^{r_K}(\hat{K})}.$$

We now apply (2.3.17) to the above estimate, obtaining (noting that $\mathrm{diam}(\hat{K}) = 1$)

$$\begin{aligned} |D^{\ell_K}\hat{u}|_{H^{r_K}(\hat{K})} \leq C(d,r) \sum_{i=1}^{\ell_K} \left\| \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right\|_{L^\infty(\hat{K})} |D^i u \circ F_K|_{H^{r_K}(\hat{K})} \\ + C(d,r) \sum_{i=1}^{\ell_K} \left| \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right|_{C^{0,1}(\overline{\hat{K}})} \|D^i u \circ F_K\|_{L^2(\hat{K})}. \end{aligned} \tag{4.6.16}$$

By (4.4.8), and the fact that the triangulation is regular of order $m \geq 2$ (and that $\mathbb{N} \ni \ell_K < t_K \leq m+1$, so $\ell_K \leq m$), we estimate the first term on the right-hand side of (4.6.16) as follows

$$\begin{aligned} & C(d,r) \sum_{i=1}^{\ell_K} \left\| \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right\|_{L^\infty(\hat{K})} |D^i u \circ F_K|_{H^{r_K}(\hat{K})} \\ & \leq C \sum_{i=1}^{\ell_K} \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} \|\tilde{B}_K\|^{l\alpha_l} |D^i u \circ F_K|_{H^{r_K}(\hat{K})} \\ & \leq C \sum_{i=1}^{\ell_K} \|\tilde{B}_K\|^{\ell_K} |D^i u \circ F_K|_{H^{r_K}(\hat{K})}. \end{aligned} \tag{4.6.17}$$

For the second term, we see that

$$\begin{aligned} & C(d,r) \sum_{i=1}^{\ell_K} \left| \sum_{\alpha \in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right|_{C^{0,1}(\overline{\hat{K}})} \|D^i u \circ F_K\|_{L^2(\hat{K})} \\ & \leq C \sum_{i=1}^{\ell_K} \left\| \sum_{\alpha \in E(\ell_K,i)} c_\alpha D \left(\prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right) \right\|_{L^\infty(\hat{K})} \|D^i u \circ F_K\|_{L^2(\hat{K})} \\ & \leq C \sum_{i=1}^{\ell_K} \sum_{\alpha \in E(\ell_K,i)} c_\alpha \left\| D \left(\prod_{l=1}^{\ell_K} (D^l F_K)^{\alpha_l} \right) \right\|_{L^\infty(\hat{K})} \|D^i u \circ F_K\|_{L^2(\hat{K})} \end{aligned} \tag{4.6.18}$$

Furthermore, for $\alpha \in E(\ell_K, i)$

$$\begin{aligned}\left\| D\left(\prod_{l=1}^{\ell_K}(D^l F_K)^{\alpha_l}\right)\right\|_{L^\infty(\hat{K})} &= \left\|\sum_{l=1}^{\ell_K} D((D^l F_K)^{\alpha_l}) \prod_{j=1, j\neq l}^{\ell_K}(D^j F_K)^{\alpha_j}\right\|_{L^\infty(\hat{K})} \\ &\leq C(d,\ell_K)\sum_{l=1}^{\ell_K}\alpha_l \sup_{\hat{x}\in\hat{K}}\|D^{l+1}F_K(\hat{x})\| \sup_{\hat{x}\in\hat{K}}\|D^{\alpha_l}F_K(\hat{x})\|^{\alpha_l-1}\prod_{j=1,j\neq l}^{\ell_K}\sup_{\hat{x}\in\hat{K}}\|D^j F_K(\hat{x})\|^{\alpha_j},\end{aligned}$$

since the triangulation is regular of order $m \geq 2$ (and $\ell_K + 1 \leq m+1$), applying (4.4.8) to the above yields

$$\begin{aligned}\left\| D\left(\prod_{l=1}^{\ell_K}(D^l F_K)^{\alpha_l}\right)\right\|_{L^\infty(\hat{K})} &\leq C\sum_{l=1}^{\ell_K}\alpha_l\|\tilde{B}_K\|^{l+1+l(\alpha_l-1)}\prod_{j=1,j\neq l}^{\ell_K}\|\tilde{B}_K\|^{j\alpha_j} \\ &\leq C\sum_{l=1}^{\ell_K}\alpha_l\|\tilde{B}_K\|^{l+1+l(\alpha_l-1)}\|\tilde{B}_K\|^{\sum_{j=1}^{\ell_K} j\alpha_j - l\alpha_l} \\ &= C\sum_{l=1}^{\ell_K}\alpha_l\|\tilde{B}_K\|^{l+1+l(\alpha_l-1)-\ell_K-l\alpha_l} \\ &= C\|\tilde{B}_K\|^{1+\ell_K}\sum_{l=1}^{\ell_K}\alpha_l = Ci\|\tilde{B}_K\|^{1+\ell_K}.\end{aligned} \tag{4.6.19}$$

Applying (4.6.19) to (4.6.18) gives us

$$\begin{aligned}&C(d,r)\sum_{i=1}^{\ell_K}\left|\sum_{\alpha\in E(\ell_K,i)} c_\alpha \prod_{l=1}^{\ell_K}(D^l F_K)^{\alpha_l}\right|_{C^{0,1}(\overline{\hat{K}})}\|D^i u\circ F_K\|_{L^2(\hat{K})} \\ &\leq C\sum_{i=1}^{\ell_K} i\sum_{\alpha\in E(\ell_K,i)} c_\alpha\|\tilde{B}_K\|^{1+\ell_K}\|D^i u\circ F_K\|_{L^2(\hat{K})} \\ &\leq C\|\tilde{B}_K\|^{1+\ell_K}\sum_{i=1}^{\ell_K}\|D^i u\circ F_K\|_{L^2(\hat{K})}.\end{aligned} \tag{4.6.20}$$

We now apply (4.6.17) and (4.6.20) to (4.6.16), obtaining

$$|D^{\ell_K}\hat{u}|_{H^{r_K}(\hat{K})} \leq C\|\tilde{B}_K\|^{\ell_K}\sum_{i=1}^{\ell_K}|D^i u\circ F_K|_{H^{r_K}(\hat{K})} + \|\tilde{B}_K\|\|D^i u\circ F_K\|_{L^2(\hat{K})}. \tag{4.6.21}$$

Applying the change of variables formula in the L^2-norms in (4.6.21), and the scaling argument (4.6.7)–(4.6.8) to the $|D^i u\circ F_K|_{H^{r_K}(\hat{K})}$ term for $i = \ell_K$ (noting that this argument is valid for any $r_K \in (0,1)$, as long as the function has H^{r_K}-regularity)

in (4.6.21), in conjunction with (4.4.26), we obtain

$$\begin{aligned}
|D^{\ell_K}\hat{u}|_{H^{r_K}(\hat{K})} &\leq Ch_K^{\ell_K-\frac{d}{2}}(h_K^{r_K}|D^{\ell_K}u|_{H^{r_K}(K)} + \sum_{i=1}^{\ell_K} h_K|u|_{H^i(K)}) \\
&\qquad + Ch_K^{\ell_K}\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^{r_K}(\hat{K})} \\
&\leq Ch_K^{\ell_K+r_K-d/2}\|u\|_{H^{\ell_K+r_K}(K)} + Ch_K^{\ell_K}\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^{r_K}(\hat{K})} \qquad (4.6.22)\\
&\leq Ch_K^{t_K-d/2}\|u\|_{H^{t_K}(K)} + Ch_K^{\ell_K}\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^{r_K}(\hat{K})} \\
&\leq Ch_K^{t_K-d/2}\|u\|_{H^{s_K}(K)} + Ch_K^{\ell_K}\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^{r_K}(\hat{K})},
\end{aligned}$$

where the constant C is independent of h_K and the choice of $K \in \mathscr{T}_h$ (note that we have utilised the continuous embedding $H^{s_K}(K) \subseteq H^{t_K}(K)$, where the constant in the embedding only depends upon d and r_K, due to Proposition 2.1 of [44]). We note, however, that the terms of the sum on the right-hand side of the final inequality of (4.6.22) are not present in the H^{t_K}-norm. Furthermore, for $1 \leq i < \ell_K$, we note the following:

$$\begin{aligned}
|D^iu\circ F_K|_{H^{r_K}(\hat{K})} &\leq C(\sigma,K^*)|(D^iu)^*|_{H^{r_K}(K^*)} \\
&= C(\sigma,K^*)|(D^iu)^* - M|_{H^{r_K}(K^*)},
\end{aligned} \qquad (4.6.23)$$

for any $M \in [\mathbb{P}^0(K^*)]^{\dim(D^iu)}$, where the first inequality follows from a scaling argument, and the fact that the mesh is regular, and the final equality holds due to the fact that constant functions are in the kernel of $|\cdot|_{H^r}$. We now use the fact that the embedding $H^1(K^*) \subseteq H^{r_K}(K^*)$ is continuous, obtaining

$$\begin{aligned}
|D^iu\circ F_K|_{H^{r_K}(\hat{K})} &\leq C(K^*,\sigma,d,r_K)\inf_{M\in[\mathbb{P}^0(K^*)]^{\dim(D^iu)}}\|(D^iu)^* - M\|_{H^1(K^*)} \\
&\leq C(K^*,\sigma,d,r_K)|(D^iu)^*|_{H^1(K^*)} \qquad (4.6.24)\\
&\leq C(K^*,\sigma,d,r_K)|D^iu\circ F_K|_{H^1(\hat{K})},
\end{aligned}$$

where the penultimate inequality follows from an application of Theorem 1.8 of [69] (or alternatively, Poincaré's inequality), and the final inequality follows from the fact that the mesh is regular.

Thus, we obtain

$$\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^{r_K}(\hat{K})} \leq C\sum_{i=1}^{\ell_K-1}|D^iu\circ F_K|_{H^1(\hat{K})} \leq Ch_K^{-d/2+1}\sum_{i=1}^{\ell_K-1}|D^iu|_{H^1(K)}.$$

Applying the above to (4.6.22) gives us

$$|D^{\ell_K}\hat{u}|_{H^{r_K}(\hat{K})} \leq Ch_K^{t_K-\frac{d}{2}}\|u\|_{H^{s_K}(K)} + Ch_K^{\ell_K+1-\frac{d}{2}}\sum_{i=1}^{\ell_K-1}|D^i u|_{H^1(K)} \leq Ch_K^{t_K-\frac{d}{2}}\|u\|_{H^{s_K}(K)}. \tag{4.6.25}$$

Finally, applying (4.6.25), (4.6.13), and (4.4.26) to (4.6.12), we obtain

$$|D^\beta(u-z_h)|_{L^2(K)} \leq Ch_K^{-q}\sum_{j=\min\{1,q\}}^{q} h_K^{2(q-j)}h_K^{t_K}\|u\|_{H^{s_K}(K)} \leq Ch_K^{t_K-q}\|u\|_{H^{s_K}(K)},$$

which is the first estimate of (4.6.10). Estimate (4.6.11) is obtained in a similar manner, utilising (4.4.12). $\square$

Lemma 4.6.4 *Assume that $(\mathscr{T}_h)_h$ is a family of meshes on $\overline{\Omega}$ that is regular of order $m \in \mathbb{N}$. For any $v \in V_{h,p}$, the following inverse estimate holds for any $K \in \mathscr{T}_h$, with $0 \leq s \leq m$, and $q \in \{2,\infty\}$:*

$$|v|_{W_*^{m,q}(K)} \lesssim h_K^{s-m}|v|_{W_*^{s,q}(K)}. \tag{4.6.26}$$

If the family of meshes is quasi-uniform, then we have that

$$|v|_{W_*^{m,q}(\Omega;\mathscr{T}_h)} \lesssim h^{s-m}|v|_{W_*^{s,q}(\Omega;\mathscr{T}_h)}. \tag{4.6.27}$$

Proof: We first note that both (4.6.26) and (4.6.27) are trivial when $m = 0$, since then $s = 0$, and $|\cdot|_{W_*^{m,q}} = |\cdot|_{W_*^{s,q}} = \|\cdot\|_{L^q}$, so we will assume that $m \geq 1$. We will first prove (4.6.26) when $s = 0$. By (4.4.13), for $j \in \mathbb{N}$, $1 \leq j \leq m$, $q \in \{2,\infty\}$, and any $K \in \mathscr{T}_h$, we have

$$|v|_{W^{j,q}(K)} \leq C|\det \tilde{B}_K|^{1/q}\|\tilde{B}_K^{-1}\|^j\left(\sum_{r=\min\{1,j\}}^{j}\|\tilde{B}_K\|^{2(j-r)}|\hat{v}|_{W^{r,q}(\hat{K})}\right). \tag{4.6.28}$$

Now, for $0 \leq r \leq j$,

$$\|\tilde{B}_K\|^{2(j-r)}|\hat{v}|_{W^{r,q}(\hat{K})} \leq C(\sigma)h_K^{2(j-r)}|\hat{v}|_{W^{r,q}(\hat{K})},$$

where the inequality is due to (4.4.26). Now, let K^* be a *fixed* reference element, and take $\tilde{G}_{\hat{K}} : K^* \to \hat{K}$, with $\tilde{G}_{\hat{K}}(x^*) = \tilde{A}_{\hat{K}}x^* + \tilde{a}_{\hat{K}}$, with $\tilde{A}_{\hat{K}} \in GL(\mathbb{R}^d)$ and $\tilde{a}_{\hat{K}} \in \mathbb{R}^d$. As in the proof of Lemma 4.6.2, it follows that $\tilde{A}_{\hat{K}}$ belongs to a compact subset BL of $GL(\mathbb{R}^d)$.

Now, defining $v^*(x^*) = \hat{v}(\tilde{G}_{\hat{K}}(x^*))$, it follows that $v^* \in \mathbb{P}^p(K^*)$, where $\mathbb{P}^p(K^*)$ is of finite dimension, depending only on K^*, d and p, thus by the equivalence of norms on finite dimensional spaces, we see that

$$\begin{aligned}
|\hat{v}|_{W^{r,q}(\hat{K})} &\leq \|\tilde{A}_{\hat{K}}\|^r |\det \tilde{A}_{\hat{K}}|^{1/q} |v^*|_{W^{r,q}(K^*)} \\
&\leq \|\tilde{A}_{\hat{K}}\|^r |\det \tilde{A}_{\hat{K}}|^{1/q} \|v^*\|_{W^{r,q}(K^*)} \\
&\leq C(d,p,K^*)\|\tilde{A}_{\hat{K}}\|^r |\det \tilde{A}_{\hat{K}}|^{1/q} \|v^*\|_{L^q(K^*)} \\
&\leq C(d,p,K^*)\|\tilde{A}_{\hat{K}}\|^r |\det \tilde{A}_{\hat{K}}|^{1/q} |\det \tilde{A}_{\hat{K}}^{-1}|^{1/q} \|\hat{v}\|_{L^q(\hat{K})} \\
&= C(d,p,K^*)\|\tilde{A}_{\hat{K}}\|^r \|\hat{v}\|_{L^q(\hat{K})} \\
&\leq C(d,p,K^*)(\max_{B\in BL} \|\tilde{A}_{\hat{K}}\|^r)\|\hat{v}\|_{L^q(\hat{K})} \\
&\leq C(d,p,\sigma,K^*,r)\|\hat{v}\|_{L^q(\hat{K})}.
\end{aligned} \tag{4.6.29}$$

Thus, applying the above inequality, (4.4.14) with $k = 0$, and (4.4.26), to (4.6.28), we obtain

$$\begin{aligned}
|v|_{W^{j,q}(K)} &\leq C(d,p,\sigma,K^*)|\det \tilde{B}_K|^{1/q} \|\tilde{B}_K^{-1}\|^j \|\hat{v}\|_{L^q(\hat{K})} \\
&\leq C(d,p,\sigma,K^*)|\det \tilde{B}_K|^{1/q} \|\tilde{B}_K^{-1}\|^j |\det \tilde{B}_K|^{-1/q} \|v\|_{L^q(K)} \\
&\leq C(d,p,\sigma,K^*,j)h_K^{-j}\|v\|_{L^q(K)} \\
&\leq C(d,p,\sigma,K^*,m)h_K^{-j}\|v\|_{L^q(K)}.
\end{aligned}$$

Since our choice of $1 \leq j \leq m$ was arbitrary, we may take $1 \leq k \leq m$, and sum the above over $1 \leq j \leq k$, obtaining

$$|v|_{W_*^{k,q}(K)} \leq C(d,p,\sigma,K^*,m)h_K^{-k}\|v\|_{L^q(K)} \quad 1 \leq k \leq m. \tag{4.6.30}$$

We obtain (4.6.26) with $s = 0$, by setting $k = m$ above. We will now prove (4.6.26) for $1 \leq s \leq m$.

In this case we will argue by induction, and as our base case, we shall prove the result for $s = 1$. Take $1 \leq j \leq m$, and let $|\alpha| = j$. Then we may write $D^\alpha v = D^\beta(D^\gamma v)$ for some $|\beta| = j-1$, $|\gamma| = 1$. One must note that by the chain rule, $Dv|_K = D(\hat{v} \circ F_K^{-1})|_K = (D\hat{v} \circ F_K^{-1})DF_K^{-1}$, where the components of $(D\hat{v} \circ F_K^{-1})DF_K^{-1}$ do not necessarily belong to $\mathbb{P}^p(\hat{K})$. It is the case, however, that $D^\delta \hat{v} \in \mathbb{P}^p(\hat{K})$ for any $|\delta| = 1$.

One can see that

$$\begin{aligned}\|D^\alpha v\|_{L^q(K)} &\leq |D^\gamma v|_{W^{j-1,q}_*(K)} \\ &\leq |Dv|_{W^{j-1,q}_*(K)} \\ &= \left|(D\hat{v}\circ F_K^{-1})DF_K^{-1}\right|_{W^{j-1,q}_*(K)} \\ &\lesssim \sum_{r=\min\{1,j-1\}}^{j-1} \sup_{x\in K}\|D^r(DF_K^{-1}(x))\||D\hat{v}\circ F_K^{-1}|_{W^{j-1-r,q}_*(K)} \\ &\lesssim \max_{\min\{2,j\}\leq r\leq j}\sup_{x\in K}\|D^r F_K^{-1}(x)\||D\hat{v}\circ F_K^{-1}|_{W^{j-1,q}_*(K)}.\end{aligned} \tag{4.6.31}$$

By (4.4.9) and (4.4.12), we have that

$$\max_{\min\{2,j\}\leq r\leq j}\sup_{x\in K}\|D^r F_K^{-1}(x)\| \leq \max_{\min\{2,j\}\leq r\leq j} c_{-r}\|\tilde{B}_K\|^{2(r-1)}\|\tilde{B}_K^{-1}\|^r, \tag{4.6.32}$$

where we are denoting $c_{-1} := 1/(1-C_K)$. Furthermore, since $D\hat{v}\in[\mathbb{P}^{p-1}(\hat{K})]^d\subset[\mathbb{P}^p(\hat{K})]^d$, we can apply (4.6.30) with $k=j-1$, obtaining

$$|D\hat{v}\circ F_K^{-1}|_{W^{j-1,q}_*(K)} \lesssim h_K^{1-j}\|D\hat{v}\circ F_K^{-1}\|_{L^q(K)}. \tag{4.6.33}$$

We also have that

$$\begin{aligned}\|D\hat{v}\circ F_K^{-1}\|_{L^q(K)} &= \|(D\hat{v}\circ F_K^{-1}DF_K^{-1})(DF_K^{-1})^{-1}\|_{L^q(K)} \\ &\leq \sup_{\hat{x}\in\hat{K}}\|DF_K\|\|Dv\|_{L^q(K)} \\ &= \sup_{\hat{x}\in\hat{K}}\|DF_K\||v|_{W^{1,q}_*(K)}.\end{aligned} \tag{4.6.34}$$

Applying (4.6.32), (4.6.33), and (4.6.34) to (4.6.31), and summing over all $|\alpha|=j$, we obtain

$$|v|_{W^{j,q}(K)} \lesssim \max_{\min\{2,j\}\leq r\leq j} c_{-r}\|\tilde{B}_K\|^{2(r-1)}\|\tilde{B}_K^{-1}\|^r \sup_{\hat{x}\in\hat{K}}\|DF_K\|h_K^{1-j}|v|_{W^{1,q}_*(K)}.$$

Lastly, applying (4.4.8) and (4.4.26) to the above estimate, we obtain (noting that $\mathscr{T}_h$ is regular of order m)

$$|v|_{W^{j,q}(K)} \lesssim \max_{\min\{2,j\}\leq r\leq j} h_K^{r-1}h_K^{1-j}|v|_{W^{1,q}_*(K)} \leq h_K^{1-j}|v|_{W^{1,q}_*(K)}. \tag{4.6.35}$$

Again, our choice of $1\leq j\leq m$ was arbitrary, and so we can sum (4.6.35) over $1\leq j\leq k$ for any $1\leq k\leq m$, obtaining

$$|v|_{W^{k,q}_*(K)} \lesssim \max_{\min\{2,k\}\leq r\leq k} h_K^{r-1}h_K^{1-k}|v|_{W^{1,q}_*(K)} \leq h_K^{1-k}|v|_{W^{1,q}_*(K)} \quad 1\leq k\leq m.$$

To proceed to argue by induction, we will assume that for $1 \leq s \leq k \leq m-1$ we have

$$|v|_{W_*^{k,q}(K)} \lesssim h_K^{s-k} |v|_{W_*^{s,q}(K)}, \tag{4.6.36}$$

and we will use this to show that

$$|v|_{W_*^{k,q}(K)} \lesssim h_K^{s+1-k} |v|_{W_*^{s+1,q}(K)},$$

for $1 \leq s+1 \leq k \leq m$.

To this end, let us take $s+1 \leq j \leq m$ and let $|\alpha| = j$. Again we write $D^\alpha v = D^\beta(D^\gamma v)$ for some $|\beta| = j-1$, and $|\gamma| = 1$, and so, analogously to our previous argument, we obtain

$$\begin{aligned} \|D^\alpha v\|_{L^q(K)} \leq |D^\gamma v|_{W_*^{j-1,q}(K)} &\leq |Dv|_{W_*^{j-1,q}(K)} \\ &\lesssim h_K^{-1} |D\hat{v} \circ F_K^{-1}|_{W_*^{j-1,q}(K)}. \end{aligned}$$

Applying our inductive hypothesis (4.6.36) with $k = j-1 \geq s$, we obtain

$$\|D^\alpha v\|_{L^q(K)} \lesssim h_K^{-1} h_K^{s-(j-1)} |D\hat{v} \circ F_K^{-1}|_{W_*^{s,q}(K)}. \tag{4.6.37}$$

Now,

$$\begin{aligned} |D\hat{v} \circ F_K^{-1}|_{W_*^{s,q}(K)} &= |(D\hat{v} \circ F_K^{-1} DF_K^{-1}) DF_K \circ F_K^{-1}|_{W_*^{s,q}(K)} \\ &= |(Dv) DF_K \circ F_K^{-1}|_{W_*^{s,q}(K)} \\ &\leq \sum_{r=\min\{1,s\}}^{s} \sup_{x \in K} \|D^r (DF_K \circ F_K^{-1})(x)\| |Dv|_{W_*^{s-r,q}(K)} \\ &\leq \sum_{r=\min\{1,s\}}^{s} \sup_{x \in K} \|D^r (DF_K \circ F_K^{-1})(x)\| |Dv|_{W_*^{s,q}(K)} \\ &\leq \sum_{r=\min\{1,s\}}^{s} \sup_{x \in K} \|D^r (DF_K \circ F_K^{-1})(x)\| |v|_{W_*^{s+1,q}(K)}. \end{aligned}$$

Applying the above to (4.6.37), we obtain

$$\|D^\alpha v\|_{L^q(K)} \lesssim \left(h_K^{-1} \sum_{r=\min\{1,s\}}^{s} \sup_{x \in K} \|D^r (DF_K \circ F_K^{-1})(x)\| \right) h_K^{s+1-j} |v|_{W_*^{s+1,q}(K)}.$$

Let us momentarily assume that, for any $1 \leq s \leq m-1$,

$$h_K^{-1} \sum_{r=\min\{1,s\}}^{s} \sup_{x \in K} \|D^r (DF_K \circ F_K^{-1})(x)\| \lesssim 1. \tag{4.6.38}$$

Then we obtain

$$\|D^\alpha v\|_{L^q(K)} \lesssim h_K^{s+1-j}|v|_{W_*^{s+1,q}(K)},$$

where $|\alpha| = j$, and $s+1 \leq j \leq m$ was arbitrary. Summing over all $|\alpha| = j$, and then all $s+1 \leq j \leq k \leq m$, we obtain

$$\sum_{s+1\leq|\alpha|\leq k} \|D^\alpha v\|_{L^q(K)} \lesssim h_K^{s+1-k}|v|_{W_*^{s+1,q}(K)} \quad s+1 \leq k \leq m.$$

It is also clear that

$$\sum_{\min\{1,s\}\leq|\alpha|\leq s} \|D^\alpha v\|_{L^q(K)} \lesssim |v|_{W_*^{s,q}(K)} \leq |v|_{W_*^{s+1,q}(K)} \lesssim h_K^{s+1-k}|v|_{W_*^{s+1,q}(K)},$$

and so we obtain

$$|v|_{W_*^{k,q}(K)} \lesssim \sum_{\min\{1,s\}\leq|\alpha|\leq s} \|D^\alpha v\|_{L^q(K)} + \sum_{s+1\leq|\alpha|\leq k} \|D^\alpha v\|_{L^q(K)} \lesssim h_K^{s+1-k}|v|_{W_*^{s+1,q}(K)},$$

$s+1 \leq k \leq m$, which concludes our inductive argument, and yields (4.6.26) for $1 \leq s \leq m$, by taking $k = m$. It remains to show that (4.6.38) is in fact true. Let us recall the formula

$$D^r(f\circ g) = \sum_{i=1}^r (D^i f \circ g)\left(\sum_{\alpha\in E(r,i)} c_\alpha \prod_{l=1}^r (D^l g)^{\alpha_l}\right),$$

where $E(r,i)$ is the set given by (4.6.15), and the c_α's, $\alpha \in E(m,r)$ are some given constants, bounded independently of h_K. From this, we obtain

$$\begin{aligned}
&h_K^{-1}\sum_{r=\min\{1,s\}}^s \sup_{x\in K}\|D^r(DF_K\circ F_K^{-1})(x)\| \\
&\quad = h_K^{-1}\sum_{r=\min\{1,s\}}^s \sup_{x\in K}\left\|\sum_{i=1}^r (D^{i+1}F_K)\circ F_K^{-1}(x)\sum_{\alpha\in E(r,i)} c_\alpha\prod_{l=1}^r (D^l F_K^{-1})^{\alpha_l}(x)\right\| \\
&\quad \lesssim h_K^{-1}\sum_{r=\min\{1,s\}}^s\sum_{i=1}^r \sup_{x\in K}\|(D^{i+1}F_K)\circ F_K^{-1}(x)\|\sum_{\alpha\in E(r,i)}\prod_{l=1}^r \sup_{x\in K}\|(D^l F_K^{-1})(x)\|^{\alpha_l} \\
&\quad \lesssim h_K^{-1}\sum_{r=\min\{1,s\}}^s\sum_{i=1}^r c_{i+1}\|\tilde{B}_K\|^{i+1}\sum_{\alpha\in E(r,i)}\prod_{l=1}^r \|\tilde{B}_K\|^{2(l-1)\alpha_l}\|\tilde{B}_K^{-1}\|^{l\alpha_l},
\end{aligned}$$

where the final inequality follows from (4.4.12), and the fact that the mesh is regular of order $m \geq s+1$. Applying (4.4.26), and noting that by definition, if $\alpha \in E(r,i)$, then $|\alpha| = i$ and $\sum_{l=1}^r l\alpha_l = r$, we obtain

$$h_K^{-1}\sum_{r=\min\{1,s\}}^s \sup_{x\in K}\|D^r(DF_K\circ F_K^{-1})(x)\| \lesssim$$

$$
\begin{aligned}
&\lesssim h_K^{-1} \sum_{r=\min\{1,s\}}^{s} \sum_{i=1}^{r} h_K^{i+1} \sum_{\alpha\in E(r,i)} h_K^{\sum_{l=1}^{r}(2l\alpha_l-2\alpha_l)} h_K^{\sum_{l=1}^{r}\alpha_l} \\
&= h_K^{-1} \sum_{r=\min\{1,s\}}^{s} \sum_{i=1}^{r} h_K^{i+1} \sum_{\alpha\in E(r,i)} h_K^{2(r-i)} h_K^{r} \\
&\lesssim h_K^{-1} \sum_{r=\min\{1,s\}}^{s} h_K^{r} \sum_{i=1}^{r} h_K^{1-i} \\
&= \sum_{r=\min\{1,s\}}^{s} h_K^{r} \sum_{i=1}^{r} h_K^{-i} \lesssim \sum_{r=0}^{s} h_K^{r} h_K^{-r} \lesssim 1,
\end{aligned}
$$

as desired.

Note that the estimates we have derived are *independent* of the choice of $K \in \mathscr{T}_h$.

In the case that the mesh is quasi-uniform, we have that $h_K^{-1} \lesssim h^{-1}$ for all $K \in \mathscr{T}_h$, and thus we may sum (4.6.26) over all $K \in \mathscr{T}_h$, which yields (4.6.27) (noting that $\mathbb{V}_{h,p} \subset V_{h,p}$). $\square$

Corollary 4.6.5 *Under the hypotheses of Lemma 4.6.4, (4.6.26) and (4.6.27) hold for any $v \in W^{m,\infty}(\Omega;\mathscr{T}_h)$ satisfying $v|_K = \hat{v} \circ F_K^{-1}$ for $\hat{v} \in \mathbb{P}^{p_K}(\hat{K})$ for all $K \in \mathscr{T}_h$, where $p_K \leq p_{\max} \in \mathbb{N}$.*

Proof: The proof is essentially the same. Indeed given an arbitrary function v satisfying the statement of the corollary, we see that for each $K \in \mathscr{T}_h$, each corresponding function $v^* \in \mathbb{P}^{p_K}(K^*)$, where $\mathbb{P}^{p_K}(K^*)$ is a finite dimensional space, and so we obtain estimates such as (4.6.29); the difference being that the constant now depends upon p_K, which in turn is bounded above by $p_{\max}$. $\square$

4.7 Discrete Poincaré–Friedrichs' and Sobolev inequalities

The following Lemmas will be utilised in Chapters 5 and 7, and are proven directly using integration by parts identities.

Lemma 4.7.1 (Discrete Poincaré–Friedrichs' inequality) *Assume that $\{\mathscr{T}_h\}_h$ is regular of order 2 family of triangulations, and let $v \in V_{h,p}$. Then, the following inequality holds*

$$
\|v\|^2_{L^2(\Omega)} \leq C \left(|v|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^b} \|u_h\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![u_h]\!]\|^2_{L^2(F)} \right), \tag{4.7.1}
$$

where the positive constant, C, depends only on C_{Tr}, d, and Ω.

Proof: Let $K \in \mathscr{T}_h$, and take $v \in V_{h,p}$. We see that

$$\begin{aligned}\int_K |v|^2 &= \frac{1}{d}\int_K \nabla\cdot(xv^2) - 2v\sum_{i=1}^d x_i D_i v \\ &\leq \frac{1}{d}\left(\int_{\partial K}(xv^2)\cdot n_{\partial K} + \int_K \frac{d}{2}|v|^2 + 2\sum_{i=1}^d x_i^2 |D_i v|^2\right),\end{aligned}$$

subtracting $(1/2)\int_K v^2$ from each side and multiplying by 2 yields

$$\int_K |v|^2 \leq \frac{2}{d}\left(\int_{\partial K}(xv^2)\cdot n_{\partial K} + 2\sum_{i=1}^d x_i^2 |D_i v|^2\right).$$

Summing the above over all $K \in \mathscr{T}_h$, and denoting n_F to be a *fixed* choice of unit normal to $F \in \mathscr{E}_h^{i,b}$, we obtain

$$\begin{aligned}\sum_{K\in\mathscr{T}_h}\|v\|^2_{L^2(K)} &\leq \frac{2}{d}\left(\sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![xv^2]\!]\cdot n_F + \sum_{K\in\mathscr{T}_h} 2\sum_{i=1}^d \int_K x_i^2|D_i v|^2\right) \\ &\leq \frac{2}{d}\left(\sum_{F\in\mathscr{E}_h^{i}}\int_F [\![xv^2]\!]\cdot n_F + C(\Omega)\sum_{F\in\mathscr{E}_h^{b}}\|v\|^2_{L^2(F)} + 2C(\Omega)^2|v|^2_{H^1(\Omega;\mathscr{T}_h)}\right),\end{aligned} \tag{4.7.2}$$

where $C(\Omega) := \max_{x\in\overline{\Omega}}\max_{i=1,\dots,d}|x_i|$. Furthermore, we have that

$$\begin{aligned}\sum_{F\in\mathscr{E}_h^i}\int_F [\![xv^2]\!]\cdot n_F &= \sum_{F\in\mathscr{E}_h^i}\int_F ([\![x]\!]\langle\!\langle v^2\rangle\!\rangle + \langle\!\langle x\rangle\!\rangle [\![v^2]\!])\cdot n_F \\ &= \sum_{F\in\mathscr{E}_h^i}\int_F (2x\langle\!\langle v\rangle\!\rangle [\![v]\!])\cdot n_F \\ &\leq 2C(\Omega)\sum_{F\in\mathscr{E}_h^i}\|\langle\!\langle v\rangle\!\rangle\|_{L^2(F)}\|[\![v]\!]\|_{L^2(F)} \\ &\leq \sum_{F\in\mathscr{E}_h^i} C(\Omega)^2(\delta\tilde{h}_F)^{-1}\|[\![v]\!]\|^2_{L^2(F)} + \delta\tilde{h}_F\|\langle\!\langle v\rangle\!\rangle\|^2_{L^2(F)},\end{aligned}$$

for any $\delta > 0$. We then apply the trace inequality (4.6.1), obtaining

$$\begin{aligned}\sum_{F\in\mathscr{E}_h^i}\int_F [\![xv^2]\!]\cdot n_F &\leq C(\Omega)^2\sum_{F\in\mathscr{E}_h^i}(\delta\tilde{h}_F)^{-1}\|[\![v]\!]\|^2_{L^2(F)} \\ &\quad + \delta C(d)\sum_{K\in\mathscr{T}_h : F\subset\partial K}\tilde{h}_F(h_K^{-1}\|v\|^2_{L^2(K)} + h_K\|\nabla v\|^2_{L^2(K)}) \\ &\leq C(\Omega)^2\sum_{F\in\mathscr{E}_h^i}(\delta\tilde{h}_F)^{-1}\|[\![v]\!]\|^2_{L^2(F)} \\ &\quad + \delta C(d)\sum_{K\in\mathscr{T}_h}\|v\|^2_{L^2(K)} + h_K\tilde{h}_F\|\nabla v\|^2_{L^2(K)}).\end{aligned}$$

Applying the above estimate to (4.7.2), we obtain, for any $\delta > 0$,

$$\sum_{K\in\mathscr{T}_h} \|v\|^2_{L^2(K)} \leq \frac{2}{d}\left(2C(\Omega)^2 |v|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i} C(\Omega)^2(\delta\tilde{h}_F)^{-1}\|[\![v]\!]\|^2_{L^2(F)} \right.$$

$$\left. + \delta C(d) \sum_{K\in\mathscr{T}_h} \|v\|^2_{L^2(K)} + \|\nabla v\|^2_{L^2(K)} + C(\Omega) \sum_{F\in\mathscr{E}_h^b} \|v\|^2_{L^2(F)} \right). \tag{4.7.3}$$

Choosing δ sufficiently small, so that $2\delta C(d)/d \leq 1/2$, subtracting $(1/2)\|v\|^2_{L^2(\Omega)}$ from each side of (4.7.3) and multiplying by 2 we obtain the desired estimate. □

Lemma 4.7.2 (Gradient Poincaré–Friedrichs' inequality) *Assume that $\{\mathscr{T}_h\}_h$ is regular of order 2 family of triangulations, and let $v \in V_{h,p}$. Then, the following inequality holds*

$$|v|^2_{H^1(\Omega;\mathscr{T}_h)} \leq C\left(|v|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1}\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^{i,b}} \tilde{h}_F^{-1}\|[\![v]\!]\|^2_{L^2(F)} \right), \tag{4.7.4}$$

where the positive constant, C, depends only on C_{Tr}, d, and Ω.

Proof: Let $v \in V_{h,p}$, and take any $K \in \mathscr{T}_h$. An application of the divergence theorem gives us

$$\int_K |\nabla v|^2 = -\int_K (\Delta v)v + \int_{\partial K} (\nabla v\cdot n_{\partial K})v.$$

Summing this equality over all $K \in \mathscr{T}_h$, gives us

$$\sum_{K\in\mathscr{T}_h} |v|^2_{H^1(K)} = -\sum_{K\in\mathscr{T}_h} \langle \Delta v, v\rangle_K + \sum_{F\in\mathscr{E}_h^{i,b}} \langle [\![\nabla v\cdot n_F]\!], \langle\!\langle v\rangle\!\rangle\rangle_F + \langle \langle\!\langle \nabla v\cdot n_F\rangle\!\rangle, [\![v]\!]\rangle_F$$

$$\leq \sum_{K\in\mathscr{T}_h} \frac{\delta}{2}\|v\|^2_{L^2(K)} + \frac{1}{2\delta}C(d)|v|^2_{H^2(K)} + \sum_{F\in\mathscr{E}_h^b} \frac{\delta\tilde{h}_F}{2}\|\nabla v\|^2_{L^2(F)} + \frac{1}{2\delta\tilde{h}_F}\|v\|^2_{L^2(F)}$$

$$+ \sum_{F\in\mathscr{E}_h^i} \frac{1}{2\delta\tilde{h}_F}\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)} + \frac{\delta\tilde{h}_F}{2}\|\langle\!\langle v\rangle\!\rangle\|^2_{L^2(F)}$$

$$+ \sum_{F\in\mathscr{E}_h^i} \frac{1}{2\delta\tilde{h}_F}\|[\![v]\!]\|^2_{L^2(F)} + \frac{\delta\tilde{h}_F}{2}\|\langle\!\langle \nabla v\cdot n_F\rangle\!\rangle\|^2_{L^2(F)}, \tag{4.7.5}$$

for any $\delta > 0$. Applying the trace estimate (4.6.1), we obtain (noting that $\tilde{h}_F \leq h_K$)

$$\begin{aligned}
&\sum_{F\in\mathscr{E}_h^i} \frac{\delta\tilde{h}_F}{2}(\|\langle\!\langle v\rangle\!\rangle\|^2_{L^2(F)} + \|\langle\!\langle \nabla v\cdot n_F\rangle\!\rangle\|^2_{L^2(F)}) + \sum_{F\in\mathscr{E}_h^b} \frac{\delta\tilde{h}_F}{2}\|\nabla v\|^2_{L^2(F)} \\
&\leq C\frac{\delta}{2}\sum_{F\in\mathscr{E}_h^{i,b}}\sum_{K\in\mathscr{T}_h:F\subset\partial K} \tilde{h}_F(\|v\|^2_{L^2(\partial K)} + \|\nabla v\|^2_{L^2(\partial K)}) \\
&\leq C\frac{\delta}{2}\sum_{F\in\mathscr{E}_h^{i,b}}\sum_{K\in\mathscr{T}_h:F\subset\partial K} \tilde{h}_F(h_K^{-1}\|v\|^2_{L^2(K)} + (h_K + h_K^{-1})\|\nabla v\|^2_{L^2(K)} + h_K^{-1}|\nabla v|^2_{H^1(K)}) \\
&\leq C\frac{\delta}{2}\|v\|^2_{H^2(\Omega;\mathscr{T}_h)};
\end{aligned}$$

applying this to (4.7.5) gives us

$$\begin{aligned}
|v|^2_{H^1(\Omega;\mathscr{T}_h)} \leq{}& C\delta\|v\|^2_{L^2(\Omega)} + C\delta|v|^2_{H^1(\Omega;\mathscr{T}_h)} + C(\delta^{-1}+\delta)|v|^2_{H^2(\Omega;\mathscr{T}_h)} \\
&+ \delta^{-1}\sum_{F\in\mathscr{E}_h^b}\tilde{h}_F^{-1}\|v\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}(\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)} + \|[\![v]\!]\|^2_{L^2(F)}).
\end{aligned}$$

We now apply (4.7.1) to the estimate above, which yields (noting that $1 \leq \tilde{h}_F^{-1}$)

$$\begin{aligned}
|v|^2_{H^1(\Omega;\mathscr{T}_h)} \leq{}& 2C\delta|v|^2_{H^1(\Omega;\mathscr{T}_h)} + C(\delta^{-1}+\delta)|v|^2_{H^2(\Omega;\mathscr{T}_h)} \\
&+ \delta^{-1}\sum_{F\in\mathscr{E}_h^b}\tilde{h}_F^{-1}\|v\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}(\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)} + \|[\![v]\!]\|^2_{L^2(F)}).
\end{aligned}$$

We now choose δ sufficiently small, so that $2C\delta \leq 1/2$, which gives us

$$\begin{aligned}
|v|^2_{H^1(\Omega;\mathscr{T}_h)} \leq{}& \frac{1}{2}|v|^2_{H^1(\Omega;\mathscr{T}_h)} \\
&+ C\left[|v|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^{i,b}}\tilde{h}_F^{-1}\|[\![v]\!]\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)}\right].
\end{aligned}$$

Subtracting $(1/2)|v|^2_{H^1(\Omega;\mathscr{T}_h)}$ from both sides of the inequality and multiplying by 2 yields the desired estimate. □

Corollary 4.7.3 *Under the hypotheses of Lemma 4.7.2, we have that*

$$\|v\|^2_{H^1(\Omega;\mathscr{T}_h)} \leq C\left(|v|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^{i,b}}\tilde{h}_F^{-1}\|[\![v]\!]\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}\|[\![\nabla v\cdot n_F]\!]\|^2_{L^2(F)})\right). \tag{4.7.6}$$

Proof: Let $v \in V_{h,p}$. An application of (4.7.1) gives us

$$\begin{aligned}\|v\|^2_{H^1(\Omega;\mathscr{T}_h)} &= \|v\|^2_{L^2(\Omega)} + |v|^2_{H^1(\Omega;\mathscr{T}_h)} \\ &\leq C\left(|v|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^b}\|u_h\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}\|[\![u_h]\!]\|^2_{L^2(F)}\right) \\ &\leq C\left(|v|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^{i,b}}\tilde{h}_F^{-1}\|[\![u_h]\!]\|^2_{L^2(F)}\right).\end{aligned}$$

We then apply (4.7.4) to the above estimate, which yields

$$\|v\|^2_{H^1(\Omega;\mathscr{T}_h)} \leq C\left(|v|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^{i,b}}\tilde{h}_F^{-1}\|[\![u_h]\!]\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^{i}}\tilde{h}_F^{-1}\|[\![\nabla v_h]\!]\|^2_{L^2(F)}\right),$$

as desired. □

Poincaré–Friedrichs' estimate for the oblique problem: In the proof of existence and uniqueness of a numerical solution to the DGFEM we will propose in Chapter 6, we require a slightly different Poincaré–Friedrichs' inequality, namely one that does not include boundary penalty terms of form $\tilde{h}_F^{-1}\|u_h\|^2_{L^2(F)}$, as appears in (4.7.4). This is due to the fact that we are treating the oblique boundary-value problem, and so we no not expect to impose a Dirichlet boundary condition via penalty terms as we do in the DGFEM of Chapter 5. Such a result has been proven in [21] in the L^2-norm in dimension $d = 2, 3$ (and generalised in [80] to more general L^p-norms, and $d \geq 3$, however, in the context of Chapter 6, estimates in the L^2-norm, with $d = 2$ are sufficient) in the case of simplicial (and also more general) partitions of an arbitrary polyhedral domain (polygonal for $d = 2$); we will prove Theorem 5.1 [21] in the context of exactly approximated, possibly curved domains (this requires some different constructions to the affine case).

Theorem 4.7.4 *Let $d = 2$, and assume that $\{\mathscr{T}_h\}_h$ is regular of order 2 family of triangulations, and let $v \in H^1(\Omega;\mathscr{T}_h)$ for a given h. Then, there exists a positive constant C, dependent upon the shape regularity of the mesh, and upon Ω, such that*

$$|v|^2_{L^2(\Omega)} \leq C\left(|v|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}\|[\![v]\!]\|^2_{L^2(F)} + \left|\frac{1}{|\partial\Omega|}\int_{\partial\Omega} v\right|^2\right). \tag{4.7.7}$$

Proof: We define $\mathcal{V}^c_{\mathcal{T}}$ to be the *curved* nonconforming $\mathbb{P}^1$ Crouzeix–Raviart finite element space as follows:

$$\begin{aligned}\mathcal{V}^c_{\mathcal{T}} := \{v \in L^2(\Omega) : \forall K \in \mathscr{T}_h, v_K = v|_K = \hat{v}\circ F_K^{-1},\ \hat{v} \in \mathbb{P}^1(\hat{K}) :\ & v \text{ is continuous at} \\ & \text{the centre of the side of two common edges}\}.\end{aligned}$$

We define the local projection operator $\Pi_K : H^1(K) \to \mathbb{P}^1(\hat{K}) \circ F_K^{-1}$ by

$$(\Pi_K \zeta)(b_F) = \begin{cases} \dfrac{1}{|\hat{F}|} \displaystyle\int_{\hat{F}} \hat{\zeta}, & \text{if } F \subset \partial K \cap \partial\Omega, \\ \dfrac{1}{|F|} \displaystyle\int_F \zeta, & \text{if } F \subset \partial K \setminus \partial\Omega, \end{cases}$$

where $b_F := F_K(\hat{b}_{\hat{F}})$, and $b_{\hat{F}}$ denotes the centre point of the edge $\hat{F}$. We define the interpolation operator $\mathcal{I} : H^1(\Omega; \mathscr{T}_h) \to \mathcal{V}_{\mathcal{T}}^c$ by

$$(\mathcal{I}\zeta)(b_F) = \begin{cases} (\Pi_K \zeta)(b_F), & \text{if } F \in \mathscr{E}_h^b, \\ \dfrac{1}{|F|} \displaystyle\int_F \langle\!\langle \zeta \rangle\!\rangle, & \text{if } F \in \mathscr{E}_h^i. \end{cases}$$

Given these definitions, we see that the difference of the two interpolants on a given $K \in \mathscr{T}_h$ can be expressed as

$$(\mathcal{I}\zeta - \Pi_K \zeta)(b_F) = \begin{cases} \dfrac{1}{2|F|} \displaystyle\int_F [\![\zeta]\!] & \text{if } F \in \partial K \setminus \partial\Omega, \\ 0 & \text{if } F \in \partial K \cap \partial\Omega. \end{cases} \tag{4.7.8}$$

For a given $K \in \mathscr{T}_h$, one has that

$$\Pi_K \zeta = \sum_{F \subset \partial K \setminus \partial\Omega} \left(\frac{1}{|F|} \int_F \zeta \right) \hat{\rho}_F \circ F_K^{-1} + \sum_{F \subset \partial K \cap \partial\Omega} \left(\frac{1}{|\hat{F}|} \int_{\hat{F}} \hat{\zeta} \right) \hat{\rho}_F \circ F_K^{-1}.$$

Let $\hat{K}$ be the reference simplex associated with K. Upon noting that the basis functions $\hat{\rho}_F \in \mathbb{P}^1(\hat{K})$, determined by the barycentric coordinates of $\hat{K}$ (see Example 4 of Section 5 of [38]) satisfy

$$\frac{1}{|\hat{F}|} \int_{\hat{F}} \hat{\rho}_{F'} = \delta_{F,F'}, \quad F, F' \subset \partial K,$$

we obtain

$$\begin{aligned} \Pi_K \rho_{F'} &= \sum_{F \subset \partial K \setminus \partial\Omega} \left(\frac{1}{|F|} \int_F \rho_{F'} \right) \hat{\rho}_F \circ F_K^{-1} + \sum_{F \subset \partial K \cap \partial\Omega} \left(\frac{1}{|\hat{F}|} \int_{\hat{F}} \hat{\rho}_{F'} \right) \hat{\rho}_F \circ F_K^{-1} \\ &= \sum_{F \subset \partial K \setminus \partial\Omega} \left(\frac{1}{|F|} \int_F \rho_{F'} \right) \hat{\rho}_F \circ F_K^{-1} + \sum_{F \subset \partial K \cap \partial\Omega} \delta_{F,F'} \hat{\rho}_F \circ F_K^{-1}. \end{aligned}$$

Furthermore, if $F \in \partial K \setminus \partial\Omega$, then (since $d = 2$) F is flat, and $\hat{F} = F_K^{-1}|_F(F)$, where $F_K^{-1}|_F$ is affine. Thus, applying the change of variables formula, we have that

$$\sum_{F \subset \partial K \setminus \partial\Omega} \left(\frac{1}{|F|} \int_F \rho_{F'} \right) \hat{\rho}_F \circ F_K^{-1} = \sum_{F \subset \partial K \setminus \partial\Omega} \left(\frac{|\det D(F_K^{-1}|_F)|}{|F|} \int_{\hat{F}} \hat{\rho}_{F'} \right) \hat{\rho}_F \circ F_K^{-1}$$

$$
\begin{aligned}
&= \sum_{F\subset\partial K\setminus\partial\Omega} \left(\frac{1}{|F|}\frac{|F|}{|\hat{F}|}\int_{\hat{F}}\hat{\rho}_{F'}\right)\hat{\rho}_F\circ F_K^{-1}\\
&= \sum_{F\subset\partial K\setminus\partial\Omega} \left(\frac{1}{|\hat{F}|}\int_{\hat{F}}\hat{\rho}_{F'}\right)\hat{\rho}_F\circ F_K^{-1} \qquad (4.7.9)\\
&= \sum_{F\subset\partial K\setminus\partial\Omega} \delta_{F,F'}\hat{\rho}_F\circ F_K^{-1}.
\end{aligned}
$$

Overall, we obtain

$$\Pi_K\rho_{F'} = \sum_{F\subset\partial K}\delta_{F,F'}\hat{\rho}_F\circ F_K^{-1} = \hat{\rho}_{F'}\circ F_K^{-1} = \rho_{F'} \quad \forall F'\subset\partial K.$$

That is, Π_K determines the basis elements, and we may write $\widehat{\Pi_K\rho_{F'}} = \hat{\rho}_{F'}$, which means that the linear map $\mathscr{L} : L^2(\hat{K}) \to L^2(\hat{K})$, given by $\mathscr{L}(\hat{\rho}) = \hat{\rho} - \widehat{\Pi_K\rho}$ for all $\hat{\rho}\in L^2(\hat{K})\supset H^1(\hat{K})$, vanishes on $\mathbb{P}^1(\hat{K})$. This allows us to apply Theorem 5 of [38] on the reference simplex, which yields

$$\|\hat{\zeta} - \widehat{\Pi_K\zeta}\|_{H^1(\hat{K})} \le \frac{Ch(\hat{K})}{\rho(\hat{K})}|\hat{\zeta}|_{H^1(\hat{K})} \le C|\hat{\zeta}|_{H^1(\hat{K})},$$

where C depends upon the shape-regularity constant σ, but is independent of K.

Via scaling arguments, we obtain:

$$
\begin{aligned}
\|\zeta - \Pi_K\zeta\|^2_{L^2(K)} &\le C|\det\tilde{B}_K|\|\hat{\zeta} - \widehat{\Pi_K\zeta}\|^2_{L^2(\hat{K})}\\
&\le C|\det\tilde{B}_K||\hat{\zeta}|^2_{H^1(\hat{K})}\\
&\le C|\det\tilde{B}_K||\det\tilde{B}_K|^{-1}\|\tilde{B}_K\|^2|\zeta|^2_{H^1(K)}\\
&\le Ch_K^2|\zeta|^2_{H^1(K)},
\end{aligned}
$$

and

$$
\begin{aligned}
|\zeta - \Pi_K\zeta|^2_{H^1(K)} &\le C|\det\tilde{B}_K|\|\tilde{B}_K^{-1}\|^2|\hat{\zeta} - \widehat{\Pi_K\zeta}|^2_{H^1(\hat{K})}\\
&\le C|\det\tilde{B}_K|\|\tilde{B}_K^{-1}\|^2\|\hat{\zeta} - \widehat{\Pi_K\zeta}\|^2_{H^1(\hat{K})}\\
&\le C|\det\tilde{B}_K|\|\tilde{B}_K^{-1}\|^2|\hat{\zeta}|^2_{H^1(\hat{K})}\\
&\le C|\det\tilde{B}_K|\|\tilde{B}_K^{-1}\|^2|\det\tilde{B}_K^{-1}|\|\tilde{B}_K\|^2|\zeta|^2_{H^1(K)}\\
&\le C|\zeta|_{H^1(K)},
\end{aligned}
$$

the latter of which also implies that

$$|\Pi_K\zeta|_{H^1(K)} \le |\zeta - \Pi_K\zeta|_{H^1(K)} + |\zeta|_{H^1(K)} \le C|\zeta|_{H^1(K)}.$$

Overall, we obtain

$$\|\zeta - \Pi_K\zeta\|^2_{L^2(K)} + h_K^2|\Pi_K\zeta|^2_{H^1(K)} \lesssim h_K^2|\zeta|^2_{H^1(K)}.$$

By the inverse estimate (4.6.26), and the fact that the functions $\hat{p}_F$ are uniformly bounded on the reference simplex, for a given $K \in \mathscr{T}_h$, we obtain

$$\begin{aligned}
|\mathcal{I}\zeta - \Pi_K\zeta|^2_{H^1(K)} &\lesssim h_K^{-2}\|\mathcal{I}\zeta - \Pi_K\zeta\|^2_{L^2(K)} \\
&\leq h_K^{-2}\left(\sum_{F\subset\partial K\setminus\partial\Omega}[(\mathcal{I}\zeta - \Pi_K\zeta)(b_F)]^2\right)\left(\sum_{F\subset\partial K\setminus\partial\Omega}\|p_F\|^2_{L^2(K)}\right) \\
&\lesssim h_K^{-2}|\det\tilde{B}_K|\left(\sum_{F\subset\partial K\setminus\partial\Omega}[(\mathcal{I}\zeta - \Pi_K\zeta)(b_F)]^2\right)\left(\sum_{F\subset\partial K\setminus\partial\Omega}\|\hat{p}_F\|^2_{L^2(\hat{K})}\right) \\
&\lesssim h_K^{d-2}\left(\sum_{F\subset\partial K\setminus\partial\Omega}[(\mathcal{I}\zeta - \Pi_K\zeta)(b_F)]^2\right) \\
&= h_K^{d-2}\sum_{F\subset\partial K\setminus\partial\Omega}\left(\frac{1}{2|F|}\int_F[\![\zeta]\!]\right)^2 \lesssim h_K^{-d}\sum_{F\subset\partial K\setminus\partial\Omega}\left(\int_F[\![\zeta]\!]\right)^2,
\end{aligned}$$

where we have used the fact that $h_K \approx \tilde{h}_F \approx |F|^{1/(d-1)}$ when $F \subset \partial K \setminus \partial\Omega$. Similarly, we obtain

$$\begin{aligned}
\|\mathcal{I}\zeta - \Pi_K\zeta\|^2_{L^2(K)} &= h_K^2(h_K^{-2}\|\mathcal{I}\zeta - \Pi_K\zeta\|^2_{L^2(K)}) \\
&\lesssim h_K^{2-d}\sum_{F\subset\partial K\setminus\partial\Omega}\left(\int_F[\![\zeta]\!]\right)^2.
\end{aligned}$$

Combining all of our estimates yields the following for all $K \in \mathscr{T}_h$:

$$\begin{aligned}
|\mathcal{I}\zeta|^2_{H^1(K)} &\leq |\mathcal{I}\zeta - \zeta|^2_{H^1(K)} + |\zeta|^2_{H^1(K)} \\
&\leq |\mathcal{I}\zeta - \Pi_K\zeta|^2_{H^1(K)} + |\zeta|^2_{H^1(K)} + |\Pi_K\zeta|^2_{H^1(K)} \\
&\lesssim |\mathcal{I}\zeta - \Pi_K\zeta|^2_{H^1(K)} + |\zeta|^2_{H^1(K)} \\
&\lesssim h_K^{-d}\sum_{F\subset\partial K\setminus\partial\Omega}\left(\int_F[\![\zeta]\!]\right)^2 + |\zeta|^2_{H^1(K)},
\end{aligned} \tag{4.7.10}$$

and similarly,

$$\begin{aligned}
\|\zeta - \mathcal{I}\zeta\|^2_{L^2(K)} &\leq \|\zeta - \Pi_K\zeta\|^2_{L^2(K)} + \|\Pi_K\zeta - \mathcal{I}\zeta\|^2_{L^2(K)} \\
&\lesssim h_K^2|\zeta|^2_{H^1(K)} + h_K^{2-d}\sum_{F\subset\partial K\setminus\partial\Omega}\left(\int_F[\![\zeta]\!]\right)^2.
\end{aligned} \tag{4.7.11}$$

Summing estimate (4.7.10) and estimate (4.7.11) over all $K \in \mathscr{T}_h$ yields

$$\begin{cases}
|\mathcal{I}\zeta|^2_{H^1(\Omega;\mathscr{T}_h)} \lesssim |\zeta|^2_{H^1(\Omega;\mathscr{T}_h)} + \displaystyle\sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-d}\left(\int_F[\![\zeta]\!]\right)^2, \\
\|\zeta - \mathcal{I}\zeta\|^2_{L^2(\Omega)} \lesssim \displaystyle\sum_{K\in\mathscr{T}_h}h_K^2|\zeta|^2_{H^1(K)} + \sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{2-d}\left(\int_F[\![\zeta]\!]\right)^2.
\end{cases}$$

Let us denote the curved continuous Lagrange finite element space of degree 2 by $\mathcal{W}_{\mathcal{T}}^c$ (in fact one has $\mathcal{W}_{\mathcal{T}}^c = \mathbb{V}_{h,p}$ for $p = 2$). It is important to note that the nodal variables of $\mathcal{V}_{\mathcal{T}}^c$ are also nodal variables of $\mathcal{W}_{\mathcal{T}}^c$. For $K \in \mathscr{T}_h$, we denote by $C(K)$, the set of the centres of the sides of K (where we define the centre as the point $p := F_K(\hat{p})$ where $\hat{p}$ is the centre of the side of $\hat{K}$), and denote the set of other nodes by $N(K)$. We also define $C(\mathscr{T}_h) := \cup_{K\in\mathscr{T}_h} C(K)$, and $N(\mathscr{T}_h) := \cup_{K\in\mathscr{T}_h} N(T)$.

We now define two operators $E : \mathcal{V}_{\mathcal{T}}^c \to \mathcal{W}_{\mathcal{T}}^c$, and $F : \mathcal{W}_{\mathcal{T}}^c \to \mathcal{V}_{\mathcal{T}}^c$ by:

$$\begin{cases} (Ev)(p) = \dfrac{1}{|\Xi_p|} \displaystyle\sum_{K\in\Xi_p} v_K(p) \quad \forall p \in N(\mathscr{T}_h) \cup C(\mathscr{T}_h), \\ (Fw)(p) = w(p) \quad \forall p \in C(\mathscr{T}_h), \end{cases}$$

where $\Xi_p := \{K \in \mathscr{T}_h : p \in \partial K\}$ is the set of the simplexes sharing p as a vertex, $|\Xi|_p$ is the cardinality of the set Ξ_p. We now show that,

$$\begin{cases} \|Ev - v\|^2_{L^2(\Omega)} \lesssim \displaystyle\sum_{K\in\mathscr{T}_h} h_K^2 |v|^2_{H^1(K)}, \\ \|Fw - w\|^2_{L^2(\Omega)} \lesssim \displaystyle\sum_{K\in\mathscr{T}_h} h_K^2 |w|^2_{H^1(K)}. \end{cases} \tag{4.7.12}$$

The argument to obtain these estimates is effectively the same as those in the proof of Lemma 3.2 in [21]. However there are some small modifications, which we will provide. Let $p \in N(\mathscr{T}_h)$, and $K_\sharp, K_\flat \in \Xi_p$. We can find a sequence $c_1, \ldots, c_m$ in $C(\mathscr{T}_h)$ so that $c_1 \in \partial K_\sharp$, $c_m \in \partial K_\flat$ and c_j, c_{j+1} belong to the boundary of $K_j \in \Xi_p$ for $j = 1, \ldots, m-1$. Note that $|\Xi_p|$ and hence m are bounded by a constant depending continuously on the shape-regularity constants of the mesh. Hence, it follows from the Cauchy–Schwarz inequality and the mean-value theorem that

$$\begin{aligned} (v_{K_\sharp}(p) - v_{K_\flat}(p))^2 &\lesssim [v_{K_\sharp}(p) - v_{K_\sharp}(c_1)]^2 + \sum_{j=1}^{m-1} [v_{K_j}(c_j) - v_{K_j}(c_{j+1})]^2 \\ &\quad + [v_{K_\flat}(c_m) - v_{K_\flat}(p)]^2 \\ &\lesssim |\nabla v_{K_\sharp}(\xi_{K_\sharp})|^2 |p - c_1|^2 + \sum_{j=1}^{m-1} |\nabla v_{K_j}(\xi_{K_j})|^2 |c_j - c_{j+1}|^2 \\ &\quad + |\nabla v_{K_\flat}(\xi_{K_\flat})|^2 |c_m - p|^2, \end{aligned}$$

where $\xi_{K_\sharp} \in (c_1, p) \subset \partial K_\sharp$, $\xi_{K_\flat} \in (c_m, p) \subset \partial K_\flat$, and $\xi_{K_j} \in (c_j, c_{j+1}) \subset K_j$ for $j = 1, \ldots, m-1$ (where (a, b) denotes the open line segment between $a, b \in \mathbb{R}^d$).

We also see that $|c_1 - p|^2 \leq h_{K_\sharp}^{2-d} |K_\sharp|$, $|p - c_m|^2 \leq h_{K_\flat}^{2-d} |K_\flat|$, and $|c_j - c_{j+1}|^2 \leq h_{K_j}^{2-d} |K_j|$. By the chain rule, for $K \in \Xi_p$, $\nabla v_K = (DF_K^{-1})^T (\nabla \hat{v} \circ F_K^{-1})$, but since

$\hat{v} \in \mathbb{P}^1(\hat{K})$, it follows that $(\nabla \hat{v} \circ F_K^{-1})$ is constant on K. Thus, for $j = 1, \ldots, m-1$, we have

$$\begin{aligned}
|\nabla v_{K_j}(\xi_{K_j})|^2 \, |c_j - c_{j+1}|^2 &\le h_{K_j}^{2-d} \sup_{x \in K_j} \|DF_K^{-1}\|^2 \int_{K_j} |\nabla \hat{v} \circ F_K|^2 \\
&\lesssim h_{K_j}^{2-d} \sup_{x \in K_j} |DF_K^{-1}(x)|^2 \sup_{\hat{x} \in \hat{K}_j} \|DF_K(\hat{x})\|^2 \int_{K_j} |(DF_K^{-1})^T \nabla \hat{v} \circ F_K|^2 \\
&= h_{K_j}^{2-d} \sup_{x \in K_j} \|DF_K^{-1}(x)\|^2 \sup_{\hat{x} \in \hat{K}_j} \|DF_K(\hat{x})\|^2 |v|^2_{H^1(K_j)} \\
&\lesssim h_{K_j}^{2-d} |v|^2_{H^1(K_j)}.
\end{aligned}$$

Analogous estimates hold for $K_\sharp$ and $K_\flat$. Thus, we obtain

$$(v_{K_\sharp}(p) - v_{K_\flat}(p))^2 \lesssim \sum_{K \in \Xi_p} h_K^{2-d} |v|^2_{H^1(K)}.$$

We then see that

$$\begin{aligned}
((Ev - v_K)(p))^2 &= \frac{1}{|\Xi_p|} \left(\sum_{K' \in \Xi_p} (v_{K'} - v_K)(p) \right)^2 \\
&\lesssim \sum_{K' \in \Xi_p} h_{K'}^{2-d} |v|^2_{H^1(K')}.
\end{aligned}$$

Let $K \in \mathscr{T}_h$. By the above, we have

$$\begin{aligned}
\|Ev - v\|^2_{L^2(K)} &= \int_K (Ev - v_K)^2 \\
&\lesssim |K| \sum_{p \in N(K) \cup C(K)} [(Ev - v_K)(p)]^2 \\
&= |K| \sum_{p \in N(K)} [(Ev - v_K)(p)]^2 \quad \text{(by the continuity of } v \text{ across } p \in C(K)) \\
&\lesssim \sum_{p \in N(K)} \sum_{K' \in \Xi_p} h_K^{d+2-d} |v|^2_{H^1(K')} \\
&= \sum_{p \in N(K)} \sum_{K' \in \Xi_p} h_K^2 |v|^2_{H^1(K')}.
\end{aligned}$$

Summing the above over all $K \in \mathscr{T}_h$ yields the first estimate of (4.7.12).

Now, on each $K \in \mathscr{T}_h$, Fw coincides with the linear classical Lagrange interpolation operator defined by (4.5.17), which satisfies the estimate

$$\|Fw - w\|^2_{L^2(K)} \lesssim h_K^4 |w|^2_{H^2_*(K)},$$

which, followed by an inverse estimate, gives us

$$\|Fw - w\|^2_{L^2(K)} \lesssim h_K^2 |w|^2_{H^1(K)}.$$

Summing the above over all $K \in \mathscr{T}_h$, we obtain

$$\|Fw - w\|^2_{L^2(K)} \lesssim \sum_{K \in \mathscr{T}_h} h_K^2 |w|^2_{H^1(K)},$$

i.e., the second estimate of (4.7.12). The main result of the theorem now follows in the same manner as in [21] (we have provided all of the necessary estimates). $\square$

Remark 4.7.5 (Constants in Theorem 5.1 of [16]) *We note that the estimates of Theorem 5.1 of [16] are of the form $A \leq \kappa(\theta_{\mathscr{T}_h})B$, where $\theta_{\mathscr{T}_h}$ is the minimum angle of the simplices of $\mathscr{T}_h$, and $\kappa : \mathbb{R}^+ \to \mathbb{R}^+$ is a continuous function, independent of $\mathscr{T}_h$. In the context of Theorem 4.7.4, we assume that the family of meshes $\{\mathscr{T}_h\}_h$ is regular, so, in particular, (4.4.20) holds. This gives us a uniform lower bound on $\theta_{\mathscr{T}_h}$ independent of h, and thus we bound the corresponding functions κ above by a constant depending on this uniform lower bound (i.e., dependent upon σ).*

In the context of curved domain approximation, the discrete gradient $\nabla_h : V_{h,p} \to H^1(\Omega; \mathscr{T}_h)$, and so we may apply Theorem 4.7.4 obtaining estimates on $|v|_{H^1(\Omega;\mathscr{T}_h)}$, i.e., the following Corollary.

Corollary 4.7.6 *Under the hypotheses of Theorem 4.7.4, we have that*

$$|v|^2_{H^1(\Omega;\mathscr{T}_h)} \lesssim |v|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F \in \mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![\nabla v]\!]\|^2_{L^2(F)} + \frac{1}{|\partial\Omega|} \sum_{F \in \mathscr{E}_h^b} \|\nabla v\|^2_{L^2(F)}, \tag{4.7.13}$$

for all $v \in V_{h,p}$.

Proof: This is a direct consequence of Theorem 4.7.4. $\square$

Corollary 4.7.7 *Under the hypotheses of Theorem 4.7.4, let $c_h \in V_{h,0}$. Then, we have that*

$$\sum_{F \in \mathscr{E}_h^b} \tilde{h}_F \|c_h\|^2_{L^2(F)} \lesssim \sum_{F \in \mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![c_h]\!]\|^2_{L^2(F)} + \left| \frac{1}{|\partial\Omega|} \int_{\partial\Omega} c_h \right|^2. \tag{4.7.14}$$

Proof: Let $c_h \in V_{h,p,0}$. Then, applying the trace estimate (4.6.1), for $F \in \mathscr{E}_h^b$ and $K \in \mathscr{T}_h$ such that $F \subset \partial K$, we obtain

$$\begin{aligned} \tilde{h}_F \|c_h\|^2_{L^2(F)} \leq \tilde{h}_F \|c_h\|^2_{L^2(\partial K)} &\lesssim \tilde{h}_F h_K^{-1} (\|c_h\|^2_{L^2(K)} + h_K |c_h|^2_{H^1(K)}) \\ &= \tilde{h}_F h_K^{-1} \|c_h\|^2_{L^2(K)} \leq \|c_h\|^2_{L^2(K)}, \end{aligned}$$

where the above estimates hold, due to the fact that c_h is piecewise constant, and by the definition of $\tilde{h}_F$. We then sum the above inequality over all $F \in \mathscr{E}_h^b$, and apply (4.7.7), yielding

$$\begin{aligned}\sum_{F\in\mathscr{E}_h^b} \tilde{h}_F \|c_h\|^2_{L^2(F)} &\lesssim \sum_{F\in\mathscr{E}_h^b} \sum_{K\in\mathscr{T}_h : F\subset\partial K} \|c_h\|^2_{L^2(K)} \\ &\lesssim \|c_h\|^2_{L^2(\Omega)} \lesssim |c_h|^2_{H^1(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![c_h]\!]\|^2_{L^2(F)} + \left| \frac{1}{|\partial\Omega|} \int_{\partial\Omega} c_h \right|^2 \\ &= \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![c_h]\!]\|^2_{L^2(F)} + \left| \frac{1}{|\partial\Omega|} \int_{\partial\Omega} c_h \right|^2,\end{aligned}$$

as desired. □

The following lemma was proven in [23] (see Lemma 4.9.1) in the case that $\Omega \subset \mathbb{R}^2$ is a polygonal domain. As noted by the authors of [23], the proof they provide is very similar to the proof of Lemma 3.3 in [20]. Indeed, in extending the result to the context of curved finite elements, the approach is the same, and simply relies on proving an inverse estimate. However, we include the additional assumption of domain convexity (note that we require this estimate in Chapter 8, where already assume convexity of the domain), as it is then simpler to prove that $\operatorname{diam}(K) \approx \operatorname{diam}(\tilde{K})$, for any $K \in \mathscr{T}_h$ with approximating straight simplex $\tilde{K}$ (since this assumption implies that $\tilde{K} \subset K$).

Lemma 4.7.8 (Discrete Sobolev inequality) *Let $\Omega \subset \mathbb{R}^2$ be convex and piecewise C^2, and let $\{\mathscr{T}_h\}_{h>0}$ be a family of regular quasi-uniform meshes on $\overline{\Omega}$. Then, for h sufficiently small, we have that*

$$\|v\|_{\infty,\Omega} \lesssim (1 + |\ln h|)^{1/2} \|v\|_{H^1(\Omega)}, \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{4.7.15}$$

Proof: Since Ω is Lipschitz continuous, Ω satisfies the cone property. That is, each $x \in \Omega$ is the vertex of a cone $\mathcal{C}_x$ that is congruent to the cone $\mathcal{C}$ defined in polar coordinates by

$$\mathcal{C} := \{(r,\theta) : 0 < r < \delta < \infty,\ 0 < \theta < \omega < 2\pi\},$$

where the values δ and ω depend upon Ω. Let us assume that $h < \delta/2$. We take $K \in \mathscr{T}_h$, and let $c \in \mathbb{R}^2$ be the centroid of $\tilde{K}$, where $\tilde{K}$ is the approximating straight simplex to K. For simplicity, we may take c to be the origin, and the cone $\mathcal{C}_c$ to be $\mathcal{C}$.

The quasi-uniformity of $\mathscr{T}_h$ implies that there exists $\eta \in (0,1)$ that is independent of K and h, such that the cone

$$\mathcal{C}_\eta := \{(r,\theta) : 0 < r < \eta h,\ 0 < \theta < \omega\} \subset K.$$

Now let $v \in \mathbb{V}_{h,p}$ be arbitrary, and let $\alpha = v(c)$. It follows from the fundamental theorem of calculus that

$$\alpha = v(r,\theta) - \int_0^r \frac{\partial v}{\partial r}(\rho,\theta) \quad \text{for} \quad \frac{\delta}{2} < r < \delta,$$

and hence,

$$\alpha^2 \leq 2(v(r,\theta))^2 + 2\left(\int_0^r \frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \quad \text{for} \quad \frac{\delta}{2} < r < \delta. \tag{4.7.16}$$

We may estimate the integral in (4.7.16) as follows (noting that $\eta h < h < \delta/2 < r$)

$$\begin{aligned}
\int_0^r \frac{\partial v}{\partial r}(\rho,\theta) &= \int_0^{\eta h} \frac{\partial v}{\partial r}(\rho,\theta) + \int_{\eta h}^r \frac{\partial v}{\partial r}(\rho,\theta) \\
&\leq \eta h \left\| \frac{\partial v}{\partial r} \right\|_{L^\infty(\mathcal{C}_\eta)} + \left(\int_{\eta h}^r \left(\frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \rho\right)^{\frac{1}{2}} \left(\int_{\eta h}^r \rho^{-1}\right)^{\frac{1}{2}} \\
&\leq \eta h |v|_{W^{1,\infty}(K)} + \sqrt{\ln(\delta/\eta h)} \left(\int_{\eta h}^r \left(\frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \rho\right)^{\frac{1}{2}}.
\end{aligned}$$

Thus, we see that

$$\left(\int_0^r \frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \leq 2(\eta h)^2 |v|^2_{W^{1,\infty}(K)} + 2\ln(\delta/\eta h) \int_{\eta h}^r \left(\frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \rho. \tag{4.7.17}$$

We then multiply (4.7.16) by r, integrate over $(0,\omega) \times (\delta/2, \delta)$, and apply (4.7.17), obtaining

$$\begin{aligned}
\alpha^2 \int_0^\omega \int_{\delta/2}^\delta r \leq 2 \int_0^\omega \int_{\delta/2}^\delta (v(r,\theta))^2 r + 4(\eta h)^2 |v|^2_{W^{1,\infty}(K)} \int_0^\omega \int_{\delta/2}^\delta r \\
+ 4\ln(\delta/\eta h) \int_{\delta/2}^\delta \left(\int_0^\omega \int_{\eta h}^r \left(\frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \rho\right) r.
\end{aligned} \tag{4.7.18}$$

Furthermore, we see that

$$\int_0^\omega \int_{\delta/2}^\delta (v(r,\theta))^2 r \leq \|v\|^2_{L^2(\Omega)}, \quad \int_0^\omega \int_{\eta h}^r \left(\frac{\partial v}{\partial r}(\rho,\theta)\right)^2 \rho \leq |v|^2_{H^1(\Omega)}, \tag{4.7.19}$$

and

$$\int_0^\omega \int_{\delta/2}^\delta r \geq \int_0^\omega \int_{\delta/2}^\delta \frac{\delta}{2} = \frac{1}{4}\omega\delta^2. \tag{4.7.20}$$

Dividing (4.7.18) by $\int_0^\omega \int_{\delta/2}^\delta r$, and then applying (4.7.19) and (4.7.20), it then follows that

$$\begin{aligned}\alpha^2 &\leq \frac{8\|v\|^2_{L^2(\Omega)}}{\omega\delta^2} + 4(\eta h)^2|v|^2_{W^{1,\infty}(K)} + \frac{16\ln(\delta/\eta h)}{\omega\delta^2}\frac{\delta}{2}|v|^2_{H^1(\Omega)} \\ &\leq C_1(1+|\ln h|)\|v\|^2_{H^1(\Omega)} + 4(\eta h)^2|v|^2_{W^{1,\infty}(K)},\end{aligned}$$

where $C_1 := \frac{8}{\omega\delta}\max\{1/\delta, |\ln(\delta/\eta)|, 1\}$.

Let us momentarily assume that there exists a constant C_2 independent of h, K and v such that

$$|v|_{W^{1,\infty}(K)} \leq C_2 h^{-1}|v|_{H^1(K)}. \tag{4.7.21}$$

It then follows that

$$\alpha^2 \leq C_1(1+|\ln h|)\|v\|^2_{H^1(\Omega)} + 4C_2\eta^2|v|^2_{H^1(K)},$$

and so

$$|v(c)| = |\alpha| \leq C(1+|\ln h|)^{\frac{1}{2}}\|v\|_{H^1(K)}.$$

Then, take $x \in K$, and since $c \in K$, we have that

$$\begin{aligned}|v(x) - v(c)| &\leq |v|_{W^{1,\infty}(K)}|x-c| \\ &\leq \operatorname{diam}(K)|v|_{W^{1,\infty}(K)} \\ &\leq C\operatorname{diam}(\tilde{K})|v|_{W^{1,\infty}(K)} \\ &\leq Ch|v|_{W^{1,\infty}(K)} \leq CC_2|v|_{H^1(K)},\end{aligned}$$

where the last inequality follows from (4.7.21), and the third inequality is proven in the proof Lemma 4.8.3 (see estimate (4.8.10)); furthermore, the constants C and C_2 are independent of K, v and h.

Thus, for $x \in K$,

$$|v(x)| \leq |v(x) - v(c)| + |v(c)| \leq C(1+|\ln h|)^{\frac{1}{2}}\|v\|_{H^1(\Omega)}.$$

Since the choice of $K \in \mathscr{T}_h$ and $x \in K$ was arbitrary, we obtain

$$\|v\|_{L^\infty(\Omega)} \leq C(1+|\ln h|)^{\frac{1}{2}}\|v\|_{H^1(\Omega)},$$

as desired. It now remains to prove (4.7.21).

Let us take $K \in \mathscr{T}_h$ and $v \in \mathbb{V}_{h,p}$. Then, taking K^* to be a fixed reference simplex, by (4.4.20), (4.4.14), (4.4.9), (4.4.21), (4.4.26), (4.4.22), and the equivalence of norms on finite dimensional spaces, we see that

$$\begin{aligned}|v|_{W^{1,\infty}(K)} = \|\nabla v\|_{L^\infty(K)} &= \|\nabla(\hat{v}\circ F_K)\|_{L^\infty(K)} \\ &= \|(DF_K^{-1})^T(\nabla\hat{v}\circ F_K)\|_{L^\infty(K)} \\ &\leq \sup_{x\in K}\|DF_K^{-1}\|\|\nabla\hat{v}\circ F_K\|_{L^\infty(K)} \\ &\leq C(K^*,\sigma)\sup_{x\in K}\|DF_K^{-1}\|\|\nabla v^*\|_{L^\infty(K^*)}\end{aligned}$$

$$
\begin{aligned}
&\leq C(K^*,\sigma,p)\sup_{x\in K}\|DF_K^{-1}\|\|\nabla v^*\|_{L^2(K^*)}\\
&\leq C(K^*,\sigma,p)\sup_{x\in K}\|DF_K^{-1}\|\|\nabla \hat{v}\|_{L^2(\hat{K})}\\
&= C(K^*,\sigma,p)\sup_{x\in K}\|DF_K^{-1}\|\hat{v}|_{H^1(\hat{K})}\\
&\leq C\sup_{x\in K}\|DF_K^{-1}\|\,|\det\tilde{B}_K\|^{-1/2}\|\tilde{B}_K\||v|_{H^1(K)}\\
&\leq C\|\tilde{B}_K^{-1}\|^{-2}\|\tilde{B}_K\||v|_{H^1(K)}\\
&\leq Ch_K^{-1}|v|_{H^1(K)}\\
&\leq Ch^{-1}|v|_{H^1(K)},
\end{aligned}
$$

which concludes the proof. □

4.8 L^q-stability of the L^2 projection operator

In Chapter 8, we will utilise the L^q-stability of the L^2 operator, $\mathcal{P}_{V_h}$, defined by (4.5.20). The stability results that we need have been proven in [45], and require the validation of the following hypotheses (stated equivalently using our notational convention).

Let $\Omega\subset\mathbb{R}^2$ be a bounded domain, and let $\{K_j\}_{j=1}^J$ be a collection of subsets of Ω such that each $K_j\cap K_k=\emptyset$ if $j\neq k$, and

$$\overline{\Omega}=\bigcup_{j=1}^J\overline{K_j}.$$

Let

$$h^*=\max_j \operatorname{diam}(\overline{K_j}), \tag{4.8.1}$$

and assume that for each j there is an open ball $B_j\subset K_j$ such that

$$h^*\leq N\operatorname{diam}(B_j), \tag{4.8.2}$$

for some constant N. The finite element spaces under consideration can be described in terms of local bases. Let $\{\phi_\ell\}_{\ell=1}^L$ be a linearly independent set in $L^\infty(\Omega)$ and set

$$\mathscr{M}:=\operatorname{span}\{\phi_\ell:\ell=1,\dots,\ell\}.$$

Assume that for the same constant N, the collection $\{\phi_\ell\}_{\ell=1}^L$ satisfies

$$\|\phi_\ell\|_{L^\infty(\Omega)}\leq 1,\ \ell=1,\dots,L, \tag{4.8.3}$$

$$\operatorname{diam}(\operatorname{supp}(\phi_\ell))\leq Nh^*,\ \ell=1,\dots,L, \tag{4.8.4}$$

$$\text{the number of } \phi_\ell\text{'s which are nonzero in } \overline{K_j} \text{ is bounded by } N,\ j=1,\dots,J, \tag{4.8.5}$$

$$\text{if } U=\sum_{\ell=1}^L U_\ell\phi_\ell \text{ and } K_j\cap\operatorname{supp}(\phi_k)\neq\emptyset, \text{ then } |U_k|\leq Nh^{-1}\|U\|_{L^2(K_j)}. \tag{4.8.6}$$

Define $P : L^1(\Omega) \to \mathscr{M}$ by

$$\int_\Omega (Pu - u)v = 0, \quad \forall v \in \mathscr{M}. \tag{4.8.7}$$

Then we have the following Theorem and Corollary.

Theorem 4.8.1 *There is a constant C depending only on K such that for $u \in L^q(\Omega)$, $1 \leq q \leq \infty$,*

$$\|Pu\|_{L^q(\Omega)} \leq C^\theta \|u\|_{L^q(\Omega)}, \quad \theta = |1 - 2/q|. \tag{4.8.8}$$

Corollary 4.8.2 *With C and q as in the hypotheses of Theorem 4.8.1 and $u \in L^q(\Omega)$,*

$$\|u - Pu\|_{L^q(\Omega)} \leq (1 + C)^\theta \inf_{v \in \mathscr{M}} \|u - v\|_{L^q(\Omega)}, \quad \theta = |1 - 2/q|. \tag{4.8.9}$$

We will now justify that (4.8.3)–(4.8.6) still hold in the context of curved continuous Lagrange finite element spaces (indeed, in the conclusion of [45], the authors state that the results of the paper hold in the case of isoparametric Lagrange finite elements, which bear resemblance to the finite element spaces we consider), and so Theorem 4.8.1 and Lemma 4.8.2 hold.

Lemma 4.8.3 *Assume that $\Omega \subset \mathbb{R}^2$ is convex and piecewise C^2, and that the family $\{\mathscr{T}_h\}_{h>0}$ is quasi-uniform and regular. Then, for any $h > 0$, (4.8.3)–(4.8.6) hold for space $\mathbb{V}_{h,p}$.*

Proof: Let $\{\mathscr{T}_h\}_{h>0}$ be as in the statement of the lemma, and take $h > 0$. Then, taking J to be the cardinality of the set $\{K : K \in \mathscr{T}_h\}$, indexing the collection $\{K : K \in \mathscr{T}_h\} = \{K_1, \ldots, K_J\}$, we have that $K_i \cap K_j = \emptyset$ if $i \neq j$, and

$$\overline{\Omega} = \bigcup_{K \in \mathscr{T}_h} \overline{K} = \bigcup_{j=1}^{J} \overline{K_j}.$$

Furthermore, since Ω is convex, it follows that for each K_j, the associated straight d-simplex, $\tilde{K}_j$ is contained in K_j, and thus $h_{K_j} = \operatorname{diam}(\tilde{K}_j) \leq \operatorname{diam}(K_j) = \operatorname{diam}(\overline{K_j})$.

Furthermore, since the triangulation is regular, we have that

$$\begin{aligned}
\operatorname{diam}(\overline{K_j}) &= \operatorname{diam}(K_j) \\
&= \sup_{x,y\in K_j} \|x-y\| \\
&= \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|F_{K_j}(\hat{x}) - F_{K_j}(\hat{y})\| \qquad \text{(since } F_{K_j} \text{ is one-to-one)} \\
&= \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|(\tilde{F}_{K_j}(\hat{x}) - \tilde{F}_{K_j}(\hat{y})) + (\Phi_{K_j}(\hat{x}) - \Phi_{K_j}(\hat{y}))\| \\
&\le \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|\tilde{F}_{K_j}(\hat{x}) - \tilde{F}_{K_j}(\hat{y})\| + \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|\Phi_{K_j}(\hat{x}) - \Phi_{K_j}(\hat{y})\| \\
&\le \sup_{\tilde{x},\tilde{y}\in \tilde{K}_j} \|\tilde{x}-\tilde{y}\| + (\sup_{\hat{z}\in \hat{K}_j} \|D\Phi_K(\hat{z})\|) \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|\hat{x}-\hat{y}\| \\
&= \operatorname{diam}(\tilde{K}_j) + (\sup_{\hat{z}\in \hat{K}_j} \|(D\Phi_K(\hat{z})\tilde{B}_K^{-1})\tilde{B}_K\|) \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|\tilde{B}_K^{-1}(\tilde{F}_K(\hat{x}) - \tilde{F}_K(\hat{y}))\| \\
&\le \operatorname{diam}(\tilde{K}_j) + (\sup_{\hat{z}\in \hat{K}_j} \|D\Phi_K(\hat{z})\tilde{B}_K^{-1}\|)\|\tilde{B}_K\|\|\tilde{B}_K^{-1}\| \sup_{\hat{x},\hat{y}\in \hat{K}_j} \|\tilde{F}_K(\hat{x}) - \tilde{F}_K(\hat{y})\| \\
&\le C \operatorname{diam}(\tilde{K}_j) = C \operatorname{diam}(\overline{\tilde{K}_j}),
\end{aligned} \tag{4.8.10}$$

where the final inequality follows from (4.4.3) and (4.4.26), and the constant C is independent of j. This implies that $h^* \approx h$, where we recall that $h := \max_{K\in\mathscr{T}_h} h_K = \max_{1\le j\le J} h_{K_j}$ and so, we may equivalently prove (4.8.3)–(4.8.6) with h^* replaced by h. We also note that (4.8.2) is a consequence of the nondegeneracy condition (4.4.20), with $N = \sigma$.

We now take the collection $\{\phi_\ell\}_{\ell=1}^L$ to be the collection of basis functions that satisfy $\phi_j(x_i) = \delta_{ij}$, $i,j = 1,\dots,L$, where the collection of points $\{x_\ell\}_{\ell=1}^L$ represent the degrees of freedom of $\mathbb{V}_{h,p}$. Such basis functions attain their supremum value of 1 at one of the degrees of freedom. This implies (4.8.3) with equality.

Furthermore, we see that $\operatorname{supp}\phi_\ell = \{K \in \{K_j\}_{j=1}^L : \overline{K} \cap \{x_\ell\} \neq \emptyset\}$. Due to the degeneracy condition (4.4.20) there exists a constant M_1, dependent only on σ, that provides an upper bound for the cardinality of $\{K \in \{K_j\}_{j=1}^J : \overline{K} \cap \{x_\ell\} \neq \emptyset\}$, and thus

$$\operatorname{diam}(\operatorname{supp}\phi_\ell) \le M_1 \max_{K\in\{K\in\{K_j\}_{j=1}^J : \overline{K}\cap\{x_\ell\}\neq\emptyset\}} \operatorname{diam}(K) \le M_1 h,$$

and so (4.8.4) holds with $N = M_1$.

Furthermore, for a given $j \in \{1,\dots,J\}$ the number of ϕ_ℓ's that are nonzero in $\overline{K_j}$ is bounded by the number of degrees of freedom of K_j, which is bounded by a constant M_2 which depends on only on p, and so (4.8.5) holds with $N = M_2$.

Finally, if $U = \sum_{\ell=1}^L U_\ell\phi_\ell$ and $j \in \{1,\dots,J\}$ such that $K_j \cap \operatorname{supp}(\phi_k) \neq \emptyset$, it follows that the point $x_k \in \{x_\ell\}_{\ell=1}^L$ representing the degree of freedom for which

$\phi_k(x_k) = 1$, we have that $x_k \in \overline{K_j}$, and that $U(x_k) = \sum_{\ell=1}^{L} U_\ell \phi_\ell(x_k) = U_k$, and thus $|U_k| = |U(x_k)| \leq \|U\|_{L^\infty(K_j)}$. Furthermore, for a fixed reference simplex K^*, we have that

$$\begin{aligned}
\|U\|_{L^\infty(K_j)} &= \|U^*\|_{L^\infty(K^*)} \\
&\leq C(K^*)\|U^*\|_{L^2(K^*)} \\
&\leq C(K^*, \sigma, d, p)\|\hat{U}\|_{L^2(\hat{K})} \qquad \text{(by (4.4.20))} \\
&\leq C(K^*, \sigma) C h_K^{-d/2} \|U\|_{L^2(K)} \qquad \text{(by scaling)} \\
&\leq M_3 h^{-d/2} \|U\|_{L^2(K)},
\end{aligned}$$

where the first and final inequality above follow due to the equivalence of norms on finite dimensional spaces, and the fact that the mesh is quasi-uniform, respectively. Since $|U_k| \leq \|U\|_{L^\infty(K_j)}$, it follows that (4.8.6) holds with $N = M_3$, where M_3 depends on K^*, σ, d and p, but is independent of j. We now define $N := \max\{\sigma, M_1, M_2, M_3\}$, which concludes the proof. $\square$

Corollary 4.8.4 *Assume that $\Omega \subset \mathbb{R}^d$ is convex and piecewise C^2, and that the family $\{\mathscr{T}_h\}_{h>0}$ is quasi-uniform and regular, and for any $h > 0$, let $V_h = [\mathbb{V}_{h,p}]^{m\times n}$, for some $m, n \in \mathbb{N}$. Then, for $1 \leq q \leq \infty$ and $u \in L^q(\Omega)$, the projection operator, $\mathcal{P}_{V_h} : L^1(\Omega) \to V_h$, satisfies the following inequalities*

$$\|\mathcal{P}_{V_h} u\|_{L^q(\Omega)} \leq C^\theta \|u\|_{L^q(\Omega)}, \tag{4.8.11}$$

and

$$\|u - \mathcal{P}_{V_h} u\|_{L^q(\Omega)} \leq (1 + C)^\theta \inf_{v \in \mathscr{M}} \|u - v\|_{L^q(\Omega)}, \tag{4.8.12}$$

where $\theta = |1 - 2/q|$, and the constant C is independent of h.

Proof: Let the finite element space V_h satisfy the hypotheses of the Corollary. Note that the operator P defined by (4.8.7) coincides with the operator $\mathcal{P}_{V_h}$, with the choice $\mathscr{M} = V_h$. Estimates (4.8.11) and (4.8.12) then follow from Theorem 4.8.1 and Corollary 4.8.2, respectively, which hold for $\mathscr{M} = \mathbb{V}_{h,p}$, due to Lemma 4.8.3, and thus estimates (4.8.11) and (4.8.12) analogously hold for $\mathscr{M} = V_h = [\mathbb{V}_{h,p}]^{m\times n}$. $\square$

4.9 Constructing curved triangulations

We will now provide details on how one may construct curved triangulations (note that this description is given in Section 6 of [16]); both those that fit the boundary *exactly*, and those that use polynomial approximation, restoring optimal convergence rates when the polynomial degree is chosen in an appropriate manner. The approach

we provide is an extension of Lenoir's method [82], and is implemented iteratively in terms of the dimension, d. Now, let us assume that $\Omega \subset \mathbb{R}^d$ has a piecewise C^{m+1} boundary, i.e., there exists a finite number of charts $\psi : \omega \to \mathbb{R}^d$, where ω is a bounded domain in $\mathbb{R}^{d-1}$, such that each ψ belongs to $C^{m+1}(\overline{\omega})$, and

$$\partial\Omega = \bigcup_{\psi} \{\psi(\xi) : \xi \in \overline{\omega}\}.$$

We assume that for each pair (ψ, ω), there exists family $\{\mathscr{T}_{h,\omega}\}_h$ of exact triangulations of the domain ω that is also regular of order m (note that in the case $d = 2$, the existence of such a family is trivial, as the boundary has no curvature). The following property is assumed: for two chart-domain pairs (ψ, ω) and (ψ', ω'), with corresponding families of triangulations $\{\mathscr{T}_{h,\omega}\}_h$, and $\{\mathscr{T}_{h,\omega'}\}_h$ it holds that $\{\psi(\alpha) : \alpha \text{ is a vertex of } \kappa \in \mathscr{T}_{h,\omega}\} \cap \psi(\overline{\omega}) \cap \psi'(\overline{\omega'})$ and $\{\psi'(\alpha') : \alpha' \text{ is a vertex of } \kappa' \in \mathscr{T}_{h,\omega'}\} \cap \psi(\overline{\omega}) \cap \psi'(\overline{\omega'})$ coincide for each h.

Let $\hat{\kappa}$ be a reference $(d-1)$-simplex with vertices $\hat{\alpha}_1, \ldots, \hat{\alpha}_d$. For any (possibly curved) $(d-1)$-simplex $\kappa \in \mathscr{T}_{h,\omega}$, there exists a C^{m+1}-mapping $f_\kappa : \hat{\kappa} \to \kappa$ such that (4.4.21)and (4.4.23) are satisfied.

Then, for any integer l, $0 \leq l \leq m$, we associate with each κ in $\mathscr{T}_{h,\omega}$ a curved Lagrange finite element of order l; we denote by π_κ^l the classical interpolation operator from $C^0(\kappa)$ into the corresponding finite element space that is given by Definition 4.5.12. Moreover, we introduce a regular family $(\tilde{\mathscr{T}}_h)_h$ of triangulations of Ω by *straight* d-simplices, such that

1. The set of all of the vertices of $\tilde{K}$ in $\tilde{\mathscr{T}}_h$ which belong to $\partial\Omega$ is given by $\cup_\psi \{\psi(\alpha) : \alpha \text{ is a vertex of } \kappa \in \mathscr{T}_{h,\omega}\}$, for each h.

2. If an element $\tilde{K}$ of $\tilde{\mathscr{T}}_h$ has two or more vertices on $\partial\Omega$, these vertices belong to the same $\psi(\overline{\omega})$, for one chart ψ.

Let $\hat{K}$ be a reference d-simplex; we denote by $\lambda_1, \ldots, \lambda_{d+1}$ the barycentric coordinates of $\hat{K}$ with respect to its vertices $\hat{a}_1, \ldots, \hat{a}_{d+1}$, where we set $\lambda_{d+1} = 1 - \lambda_1 - \ldots - \lambda_d$.

Since the family of triangulations $(\tilde{\mathscr{T}}_h)_h$ are regular, for any given *straight* d-simplex $\tilde{K}$ in $\tilde{\mathscr{T}}_h$, there exists an *affine* mapping $\tilde{F}_K : \hat{K} \to \tilde{K}$. We then define a C^{m+1}-mapping $\Phi_K : \hat{K} \to \mathbb{R}^d$ such that, defining $F_K := \tilde{F}_K + \Phi_K$, and $K := F_K(\hat{K})$, we have that $\mathscr{T}_h := \{K : \tilde{K} \in \tilde{\mathscr{T}}_h\}$ is an exact triangulation of Ω, and that the family $(\mathscr{T}_h)_h$ is regular of order m.

4.9.1 Lenoir's procedure

Let $\tilde{K}$ be a d-simplex of $\tilde{\mathscr{T}}_h$, with vertices $a_i = \tilde{F}_K(\hat{a}_i)$, $1 \le i \le d+1$. Then:

1. If at most one vertex of $\tilde{K}$ belongs to $\partial\Omega$, we set $\Phi_K \equiv 0$.

2. If j vertices of $\tilde{K}$, $2 \le j \le d$, belong to $\partial\Omega$, for instance $a_i = \psi(\alpha_i)$, $1 \le i \le j$, we consider a $(d-1)$-simplex $\kappa = f_\kappa(\hat{\kappa})$ that has $\alpha_i = f_\kappa(\hat{\alpha}_i)$, $1 \le i \le j$, among its vertices, and we set

$$\Phi_K(\lambda_1, \ldots, \lambda_d) = \left(\left(\sum_{k=1}^{j} \lambda_k \right)^{m+2} (\psi - \pi_\kappa^m \psi) + \sum_{l=2}^{m} \left(\sum_{k=1}^{j} \lambda_k \right)^{l} (\pi_\kappa^l \psi - \pi_\kappa^{l-1} \psi) \right) \circ f_\kappa((\lambda_1 \hat{\alpha}_1 + \ldots + \lambda_j \hat{\alpha}_j)/(\lambda_1 + \ldots + \lambda_j)). \tag{4.9.1}$$

Remark 4.9.1 *To obtain an isoparametric triangulation of Ω, it suffices to replace (4.9.1) by the simpler expression*

$$\Phi_K(\lambda_1, \ldots, \lambda_d) = \left(\sum_{l=2}^{m} (\lambda_1 + \ldots + \lambda_j)^l (\pi_\kappa^l \psi - \pi_\kappa^{l-1} \psi) \right) \circ f_\kappa((\lambda_1 \hat{\alpha}_1 + \ldots + \lambda_j \hat{\alpha}_j)/(\lambda_1 + \ldots + \lambda_j)), \tag{4.9.2}$$

which is the construction given by Lenoir in [82].

Remark 4.9.2 (Two-dimensional case) *In the case where $d = j = 2$, (4.9.1) becomes:*

$$\Phi_K(\lambda_1, \lambda_2) = \left[(\lambda_1 + \lambda_2)^{m+2} (\psi - \pi_\kappa^m \psi) + \sum_{l=2}^{m} (\lambda_1 + \lambda_2)^l (\pi_\kappa^l \psi - \pi_\kappa^{l-1} \psi) \right] \frac{\lambda_1 \alpha_1 + \lambda_2 \alpha_2}{\lambda_1 + \lambda_2}. \tag{4.9.3}$$

This tell us that on the side $[\hat{a}_1, \hat{a}_2]$ (where $\lambda_1 + \lambda_2 = 1$),

$$\Phi_K(\lambda_1, \lambda_2) = \psi(\lambda_1 \alpha_1 + \lambda_2 \alpha_2) - \lambda_1 a_1 - \lambda_2 a_2,$$

and on the sides $[\hat{a}_1, \hat{a}_3]$, and $[\hat{a}_2, \hat{a}_3]$ (where $\lambda_2 = 0$, and $\lambda_1 = 0$, respectively)

$$\Phi_K(\lambda_1, \lambda_2) = 0.$$

So that K has one curved edge and two straight edges. Moreover, the straight edges are internal, and the restriction of F_K to such edges is indeed affine, justifying our use of this assertion in (4.7.9) in the proof of Theorem 4.7.4.

Remark 4.9.3 *Section 6 of [16] also provides a proof of the fact that the procedure we have defined produces exact triangulations that are regular of order m (see Theorem 6.2 and Corollary 6.2 of [16]).*

4.10 Tangential operators and curved simplex curvature bounds

Tangential differential operators. For $F \in \mathscr{E}^{i,b}$, denote for $s > 1/2$ the space of H^s-regular tangential vector fields on F by $H^s_{\mathbf{T}}(F) := \{v \in H^s(F)^d : v \cdot n_F = 0 \text{ on } F\}$. Below we define the tangential gradient $\nabla_{\mathbf{T}} : H^s(F) \to H^{s-1}_T(F)$ and the tangential divergence $\mathrm{div}_{\mathbf{T}} : H^s_{\mathbf{T}}(F) \to H^{s-1}(F)$, where $1 \leq s \leq 2$ (note that in the case that $\partial\Omega$ is piecewise C^m, with $m \geq 2$, we are able to consider $1 \leq s \leq m$). We see that $F \subset \partial K$, for some $K \in \mathscr{T}_h$. Since K is piecewise C^2 (see the proof of Lemma 4.10.3), for a.e. $x \in \partial K$, there exists a neighbourhood W_x of x in ∂K, sufficiently small to allow the existence of a family of C^2 curves that satisfy the following: a curve of each family passes through every point of W_x, and the unit tangent vectors to these curves form an orthonormal system (assumed to be oriented with respect to $\overline{n}$, where $\overline{n}$ is the unit outward normal to ∂K) at every point of W_x. We take the lengths $s_1, \ldots, s_{d-1}$ along each of these curves, respectively, to be the local coordinate system, and denote $t_1, \ldots, t_{d-1}$ to be the unit tangent vectors along each curve, respectively. In this notation, we have the following for $v : \partial K \to \mathbb{R}^d$:

$$v = v_{\mathbf{T}} + (v \cdot \overline{n})\overline{n}, \quad v_{\mathbf{T}} := \sum_{j=1}^{d-1} (v \cdot t_j) t_j.$$

For $\phi \in C^1(\overline{K})$, and $\psi \in C^1(\overline{K})^d$, with $\psi|_{\partial K} = \sum_{j=1}^{d-1} \psi_j t_j$, we obtain

$$\nabla \phi|_{\partial K} = \nabla_{\mathbf{T}} \phi + \frac{\partial \phi}{\partial \overline{n}} \overline{n}, \quad \nabla_{\mathbf{T}} \phi = \sum_{j=1}^{d-1} \frac{\partial \phi}{\partial s_j} t_j, \tag{4.10.1}$$

and

$$\mathrm{div}_{\mathbf{T}}\, \psi = \nabla_{\mathbf{T}} \cdot \psi = \sum_{j=1}^{d-1} \frac{\partial \psi_j}{\partial s_j}, \tag{4.10.2}$$

which extend to $\phi \in H^s(K)$, $s > 3/2$, by density and the construction of the trace operator. Furthermore, one can see that by rearranging the first identity of (4.10.1), that $\nabla_{\mathbf{T}} = \nabla - \overline{n} \frac{\partial}{\partial \overline{n}}$ (and thus $\mathrm{div}_{\mathbf{T}}$) is well defined *a.e.* on ∂K, and is independent of the choice of normal $\overline{n}$. We approach (4.10.1) and (4.10.2) in the context of traces and Sobolev spaces, in the following lemma. In particular we are able to decompose the Laplacian, Δ, in terms of the tangential Laplacian $\Delta_{\mathbf{T}} := \mathrm{div}_{\mathbf{T}} \nabla_{\mathbf{T}}$, as well as other terms.

Lemma 4.10.1 *Let Ω be a piecewise C^2 domain, and let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes on $\overline{\Omega}$ that is regular of order 1 and satisfies Assumption 4.4.9. Then, for any $h > 0$, for each $K \in \mathscr{T}_h$ and each face $F \subset \partial K$, the following identities hold:*

$$\tau_F(\nabla v) = \nabla_{\mathbf{T}}(\tau_F v) + \left(\tau_F \frac{\partial v}{\partial n_F}\right) n_F \quad \forall v \in H^s(K),\, s > 3/2, \tag{4.10.3}$$

$$\tau_F(\Delta v) = \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}}(\tau_F v) + \mathcal{H}_F \left(\tau_F \frac{\partial v}{\partial n_F}\right) + \tau_F \frac{\partial}{\partial n_F}(\nabla v \cdot n_F), \tag{4.10.4}$$

for all $v \in H^s(K)$, $s > 5/2$, where $\mathcal{H}_F$ is the mean curvature of the face F, and τ_F is the trace operator from K to F.

Proof: Let us take $U \in C^3(\overline{K})$, and for $F \in \mathscr{E}_h^{i,b}$, let $u = U|_F$. Then, as the family of meshes $\{\mathcal{T}_h\}_{h>0}$ is regular of order 1, it follows that $F \subset \partial K$ for some $K \in \mathscr{T}_h$, where K is piecewise C^2 (see the proof of Theorem 4.10.3). Thus, we may extend (without relabelling) the unit normal to F, n_F (note that this choice of unit normal is fixed, and that (4.10.3) is independent of this choice), by $n_F \in C^1(\overline{K})$ (note that the extension may not be normal to the other faces of ∂K, when restricted there), and so also define an extension of the tangential gradient, $\nabla_{\mathbf{T}} : C^3(\overline{K}) \to C^1(\overline{K})^d$, by

$$\nabla_{\mathbf{T}} U = \nabla U - \frac{\partial U}{\partial n_F} n_F.$$

This can be rearranged to yield

$$\nabla U = \nabla_{\mathbf{T}} U + \frac{\partial U}{\partial n_F} n_F.$$

Upon restricting to F, we obtain

$$\begin{aligned}
\nabla U|_F &= \nabla_{\mathbf{T}} U|_F + \left(\left(\frac{\partial U}{\partial n_F}\right) n_F\right)\Big|_F \\
&= \nabla_{\mathbf{T}}|_F(U|_F) + \left(\frac{\partial U}{\partial n_F}\right)\Big|_F n_F|_F \\
&= \nabla_{\mathbf{T}}(U|_F) + \left(\frac{\partial U}{\partial n_F}\right)\Big|_F n_F
\end{aligned}$$

Thus, by density and the construction of the trace operator, this extends to $u \in H^s(K)$, $s > 3/2$, giving us

$$\tau_F(\nabla u) = \nabla_{\mathbf{T}}(\tau_F u) + \left(\tau_F \frac{\partial u}{\partial n_F}\right) n_F, \tag{4.10.5}$$

which is (4.10.3).

For the identity (4.10.4), we follow a similar approach to [106], in which the statement is essentially proven for $d = 2, 3$. Now, for $x \in F$ let us take a local coordinate system $s_1, \dots, s_{d-1}$, on a neighbourhood W_x of x in F. Expressing F locally as the graph of a C^2 function ϕ, we see that

$$u(s_1, \dots, s_{d-1}) = U(s_1, \dots, s_{d-1}, \phi(s_1, \dots, s_{d-1})).$$

Furthermore, let us assume that the coordinates have been chosen so that $\nabla_{s'}\phi(0) = 0$ (denoting $s' = (s_1, \dots, s_{d-1})$), so that the local coordinates $\{s', s_d\} = \{s', \phi(s')\}$ are tangent to the hyperplane $\{s_d = 0\}$ at $x = (0, \phi(0))$. Then, in W_x, we have that

$$\operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u = \sum_{j=1}^{d-1} \frac{\partial^2 u}{\partial s_j^2},$$

where, for $j = 1, \dots, d-1$,

$$\begin{aligned}
\frac{\partial^2 u}{\partial s_j^2} &= \frac{\partial}{\partial s_j}\left(\frac{\partial}{\partial s_j}(U(s', \phi(s')))\right) \\
&= \frac{\partial}{\partial s_j}\left(U_j(s', \phi(s')) + \frac{\partial \phi}{\partial s_j} U_d(s', \phi(s'))\right) \\
&= U_{jj}(s', \phi(s')) + 2\frac{\partial \phi}{\partial s_j} U_{dj}(s', \phi(s')) + \frac{\partial^2 \phi}{\partial s_j^2} U_d(s', \phi(s')) + \left(\frac{\partial \phi}{\partial s_j}\right)^2 U_{dd}(s', \phi(s')),
\end{aligned}$$

where U_j, U_{jk} denote the first and second order partial derivatives in the j and j, k components of U, respectively. Thus, at x, i.e., at $s' = 0$, we have

$$\operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u = \sum_{j-1}^{d-1} U_{jj}(0, \phi(0)) + U_d(0, \phi(0)) \sum_{j=1}^{d-1} \frac{\partial^2 \phi(0)}{\partial s_j^2}.$$

Moreover, at x, $U_{dd} = \frac{\partial^2 U}{\partial n_F^2}$, $U_d = \frac{\partial U}{\partial n_F}$, and $\sum_{j=1}^{d-1} \frac{\partial^2 \phi}{\partial s_j^2} = -\mathcal{H}_F$. Thus, at x,

$$\Delta U = \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u + \mathcal{H}_F \frac{\partial U}{\partial n_F} + \frac{\partial^2 U}{\partial n_F^2}.$$

This decomposition is valid at any $x \in F$, and so we obtain

$$\begin{aligned}
\Delta U|_F &= \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u + \mathcal{H}_F \frac{\partial U}{\partial n_F}\bigg|_F + \frac{\partial^2 U}{\partial n_F^2}\bigg|_F \\
&= \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} (U|_F) + \mathcal{H}_F \frac{\partial U}{\partial n_F}\bigg|_F + \frac{\partial^2 U}{\partial n_F^2}\bigg|_F.
\end{aligned}$$

Thus, by density, applying (4.10.5), for $u \in H^s(K)$, $s > 5/2$, we obtain

$$\tau_F(\Delta u) = \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} (\tau_F u) + \mathcal{H}_F \left(\tau_F \frac{\partial u}{\partial n_F}\right) + \tau_F \frac{\partial}{\partial n_F}(\nabla u \cdot n_F),$$

which is (4.10.4). $\square$

The proof of the following lemma follows the proof of Theorem 3.1.1.12 of [60], generalising the result to the associated bilinear form. That is, Theorem 3.1.1.12 of [60] essentially provides the following lemma in the case that $u = v$.

Lemma 4.10.2 (Integration by parts identity) *For any $K \in \mathscr{T}_h$, and any $u, v \in H^s(K)$, $s > 5/2$, we have that*

$$\begin{aligned}\int_K \Delta u\, \Delta v = \int_K D^2u : D^2v + \int_{\partial K} \mathrm{div}_{\mathbf{T}}\, \nabla_{\mathbf{T}} u \frac{\partial v}{\partial \overline{n}} - \nabla_{\mathbf{T}} \left(\frac{\partial u}{\partial \overline{n}}\right) \cdot \nabla_{\mathbf{T}} v \\ + \mathcal{H}_{\partial K} \frac{\partial u}{\partial \overline{n}} \frac{\partial v}{\partial \overline{n}} + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v,\end{aligned} \tag{4.10.6}$$

where $\overline{n}$ is the unit outward normal to ∂K, and $\mathcal{H}_{\partial K} := \nabla_{\mathbf{T}} \cdot \overline{n}$.

Proof: First let us assume that $u, v \in C^3(\overline{K})$. Then, repeated applications of integration by parts gives us

$$\begin{aligned}\int_K \Delta u\, \Delta v &= \sum_{i,j=1}^{d} \int_K \partial^2_{ii} u\, \partial^2_{jj} v \\ &= -\sum_{i,j=1}^{d} \int_K \partial^3_{iji} u\, \partial_i v + \sum_{i,j=1}^{d} \int_{\partial K} \partial^2_{jj} u\, \partial_j v\, \overline{n}_j \\ &= \sum_{i,j=1}^{d} \int_K \partial^2_{ij} u\, \partial^2_{ij} v + \sum_{i,j=1}^{d} \int_{\partial K} \partial^2_{jj} u\, \partial_j v\, \overline{n}_j - \partial^2_{ij} u\, \partial_j v\, \overline{n}_i \\ &= \int_K D^2u : D^2v + \int_{\partial K} \Delta u \frac{\partial v}{\partial \overline{n}} - (\nabla v \cdot \nabla(Du)) \cdot \overline{n}.\end{aligned} \tag{4.10.7}$$

We see that for a given $x \in \partial K$, and a sufficiently small neighbourhood W_x of x in ∂K, that in W_x

$$\begin{aligned}\Delta u &= \mathrm{div}(\nabla u) \\ &= \sum_{j=1}^{d-1} \frac{\partial}{\partial s_j}(\nabla u) \cdot t_j + \frac{\partial}{\partial \overline{n}}(\nabla u) \cdot \overline{n} \\ &= \sum_{j=1}^{d-1} \frac{\partial}{\partial s_j} \left(\sum_{k=1}^{d-1} \frac{\partial u}{\partial s_k} t_k + \frac{\partial u}{\partial \overline{n}} \overline{n}\right) \cdot t_j + \frac{\partial}{\partial \overline{n}} \left(\sum_{k=1}^{d-1} \frac{\partial u}{\partial s_k} t_k + \frac{\partial u}{\partial \overline{n}} \overline{n}\right) \cdot \overline{n} \\ &= \sum_{j,k=1}^{d-1} \left(\frac{\partial^2 u}{\partial s_j \partial s_k} t_k + \frac{\partial u}{\partial t_k} \frac{\partial t_k}{\partial s_j}\right) \cdot t_j + \sum_{j=1}^{d-1} \left(\frac{\partial^2 u}{\partial \overline{n} \partial s_j} + \frac{\partial u}{\partial \overline{n}} \frac{\partial \overline{n}}{\partial t_j}\right) \cdot t_j \\ &\quad + \sum_{k=1}^{d-1} \left(\frac{\partial^2 u}{\partial \overline{n} \partial s_k} t_k + \frac{\partial u}{\partial s_k} \frac{\partial t_k}{\partial \overline{n}}\right) \cdot \overline{n} + \left(\frac{\partial^2 u}{\partial \overline{n}^2} \overline{n} + \frac{\partial u}{\partial \overline{n}} \frac{\partial \overline{n}}{\partial \overline{n}}\right) \cdot \overline{n}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{d-1} \frac{\partial^2 u}{\partial s_j^2} + \sum_{j,k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\cdot t_j + \sum_{j=1}^{d-1} \frac{\partial u}{\partial \overline{n}}\frac{\partial \overline{n}}{\partial s_j}\cdot t_j \\
&\quad + \sum_{k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial \overline{n}}\cdot \overline{n} + \frac{\partial^2 u}{\partial \overline{n}^2},
\end{aligned} \tag{4.10.8}
$$

where the final equality is due to the fact that $t_i \cdot t_j = \delta_{ij}$, $t_i \cdot \overline{n} = 0$, and $\frac{\partial \overline{n}}{\partial \overline{n}} \cdot \overline{n} = 0$. In a similar fashion, we obtain

$$
\begin{aligned}
\nabla v \cdot \nabla(Du) &= \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial}{\partial s_j}(Du) + \frac{\partial v}{\partial \overline{n}}\frac{\partial}{\partial \overline{n}}(Du) \\
&= \sum_{j,k=1}^{d-1} \frac{\partial v}{\partial s_j}\left(\frac{\partial^2 u}{\partial s_k \partial s_j}t_k + \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\right) + \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\left(\frac{\partial^2 u}{\partial s_j \partial \overline{n}}\overline{n} + \frac{\partial u}{\partial \overline{n}}\frac{\partial \overline{n}}{\partial s_j}\right) \\
&\quad + \frac{\partial v}{\partial \overline{n}}\left(\sum_{k=1}^{d-1} \frac{\partial^2 u}{\partial s_k \partial \overline{n}}t_k + \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial \overline{n}}\right) + \frac{\partial v}{\partial \overline{n}}\left(\frac{\partial^2 u}{\partial \overline{n}^2}\overline{n} + \frac{\partial u}{\partial \overline{n}}\frac{\partial \overline{n}}{\partial \overline{n}}\right).
\end{aligned}
$$

Then, since $t_i \cdot \overline{n} = 0$, $\frac{\partial \overline{n}}{\partial s_j} \cdot \overline{n} = 0$, and $\frac{\partial \overline{n}}{\partial \overline{n}} \cdot \overline{n} = 0$, from the above, we obtain

$$
(\nabla v \cdot \nabla(Du)) \cdot \overline{n} = \sum_{j,k=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\cdot \overline{n} + \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial^2 u}{\partial s_j \partial \overline{n}} + \frac{\partial v}{\partial \overline{n}}\left(\sum_{k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial \overline{n}}\cdot \overline{n} + \frac{\partial^2 u}{\partial \overline{n}^2}\right). \tag{4.10.9}
$$

Noting the cancellation of particular terms, (4.10.8) and (4.10.9) gives us

$$
\begin{aligned}
\Delta u \frac{\partial v}{\partial \overline{n}} - (\nabla v \cdot \nabla(Du)) \cdot \overline{n} &= \sum_{j=1}^{d-1} \frac{\partial^2 u}{\partial s_j^2}\frac{\partial v}{\partial \overline{n}} + \sum_{j,k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\cdot t_j \frac{\partial v}{\partial \overline{n}} + \sum_{j=1}^{d-1} \frac{\partial u}{\partial \overline{n}}\frac{\partial \overline{n}}{\partial s_j}\cdot t_j \frac{\partial v}{\partial \overline{n}} \\
&\quad - \sum_{j,k=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\cdot \overline{n} - \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial^2 u}{\partial s_j \partial \overline{n}}.
\end{aligned} \tag{4.10.10}
$$

One can see that

$$
\sum_{j=1}^{d-1} \frac{\partial u}{\partial \overline{n}}\frac{\partial \overline{n}}{\partial s_j}\cdot t_j \frac{\partial v}{\partial \overline{n}} = \left(\sum_{j=1}^{d-1} \frac{\partial \overline{n}}{\partial s_j}\cdot t_j\right)\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} = (\nabla_{\mathbf{T}} \cdot \overline{n})\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} = \mathcal{H}_{\partial K}\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}}. \tag{4.10.11}
$$

Furthermore, for $j, k = 1 \ldots, d-1$, we see that

$$
0 = \frac{\partial}{\partial s_j}(\overline{n} \cdot t_k) = \frac{\partial \overline{n}}{\partial s_j} \cdot t_k + \overline{n} \cdot \frac{\partial t_k}{\partial s_j} \Rightarrow \frac{\partial t_k}{\partial s_j} \cdot \overline{n} = -\frac{\partial \overline{n}}{\partial s_j} \cdot t_k,
$$

and so

$$
-\sum_{j,k=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\cdot \overline{n} = \sum_{j,k=1}^{d-1} \frac{\partial \overline{n}}{\partial s_j}\cdot t_k \frac{\partial v}{\partial s_j}\frac{\partial u}{\partial s_k} = (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v. \tag{4.10.12}
$$

Substituting (4.10.11) and (4.10.12) into (4.10.10), we obtain

$$\Delta u\frac{\partial v}{\partial \overline{n}} - (\nabla v \cdot \nabla(Du)) \cdot \overline{n} = \sum_{j=1}^{d-1} \frac{\partial^2 u}{\partial s_j^2}\frac{\partial v}{\partial \overline{n}} + \sum_{j,k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j} \cdot t_j \frac{\partial v}{\partial \overline{n}} + \mathcal{H}_{\partial K}\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v - \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial^2 u}{\partial s_j \partial \overline{n}}. \tag{4.10.13}$$

Finally, we calculate

$$\begin{aligned} \operatorname{div}_{\mathbf{T}}\left(\frac{\partial v}{\partial \overline{n}}\nabla_{\mathbf{T}} u\right) &= \sum_{j=1}^{d-1} \frac{\partial}{\partial s_j}\left(\frac{\partial v}{\partial \overline{n}}\nabla_{\mathbf{T}} u\right) \cdot t_j \\ &= \sum_{j,k=1}^{d-1} \frac{\partial}{\partial s_j}(\frac{\partial v}{\partial \overline{n}}\frac{\partial u}{\partial s_k} t_k) \cdot t_j \\ &= \sum_{j,k=1}^{d-1} \left(\frac{\partial^2 v}{\partial s_j \partial \overline{n}}\frac{\partial u}{\partial s_k} t_k + \frac{\partial v}{\partial \overline{n}}\frac{\partial^2 u}{\partial s_j \partial s_k} t_k + \frac{\partial v}{\partial \overline{n}}\frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j}\right) \cdot t_j \\ &= \sum_{j=1}^{d-1} \frac{\partial^2 u}{\partial s_j^2}\frac{\partial v}{\partial \overline{n}} + \sum_{j,k=1}^{d-1} \frac{\partial u}{\partial s_k}\frac{\partial t_k}{\partial s_j} \cdot t_j \frac{\partial v}{\partial \overline{n}} + \sum_{j=1}^{d-1} \frac{\partial^2 v}{\partial s_j \partial \overline{n}}\frac{\partial u}{\partial s_j}. \end{aligned} \tag{4.10.14}$$

Applying (4.10.14) to (4.10.13), we obtain

$$\begin{aligned} \Delta u\frac{\partial v}{\partial \overline{n}} - (\nabla v \cdot \nabla(Du)) \cdot \overline{n} &= \operatorname{div}_{\mathbf{T}}\left(\frac{\partial v}{\partial \overline{n}}\nabla_{\mathbf{T}} u\right) - \sum_{j=1}^{d-1} \frac{\partial^2 v}{\partial s_j \partial \overline{n}}\frac{\partial u}{\partial s_j} \\ &\quad + \mathcal{H}_{\partial K}\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} - (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v - \sum_{j=1}^{d-1} \frac{\partial v}{\partial s_j}\frac{\partial^2 u}{\partial s_j \partial \overline{n}} \\ &= \operatorname{div}_{\mathbf{T}}\left(\frac{\partial v}{\partial \overline{n}}\nabla_{\mathbf{T}} u\right) - \nabla_{\mathbf{T}}\left(\frac{\partial u}{\partial \overline{n}}\right) \cdot \nabla_{\mathbf{T}} v - \nabla_{\mathbf{T}}\left(\frac{\partial v}{\partial \overline{n}}\right) \cdot \nabla_{\mathbf{T}} u \\ &\quad + \mathcal{H}_{\partial K}\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v \\ &= \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u\frac{\partial v}{\partial \overline{n}} + \mathcal{H}_{\partial K}\frac{\partial u}{\partial \overline{n}}\frac{\partial v}{\partial \overline{n}} \\ &\quad + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v - \nabla_{\mathbf{T}}\left(\frac{\partial u}{\partial \overline{n}}\right) \cdot \nabla_{\mathbf{T}} v. \end{aligned} \tag{4.10.15}$$

Note that the identity above does not depend on the choice of local coordinates in W_x, and so, varying W_x, we deduce that the identity holds a.e. on ∂K, and thus

applying this to (4.10.7), we obtain

$$\int_K \Delta u\, \Delta v = \int_K D^2 u : D^2 v + \int_{\partial K} \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u \frac{\partial v}{\partial \overline{n}} - \nabla_{\mathbf{T}} \left(\frac{\partial u}{\partial \overline{n}} \right) \cdot \nabla_{\mathbf{T}} v + \mathcal{H}_{\partial K} \frac{\partial u}{\partial \overline{n}} \frac{\partial v}{\partial \overline{n}} + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v,$$

which is exactly (4.10.6). This extends to $u, v \in H^s(K)$, $s > 5/2$, by density. □

Lemma 4.10.3 *Let Ω be a piecewise C^2 domain, and let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes on $\overline{\Omega}$ that is regular of order 1, and satisfies Assumption 4.4.9. Then, there exists a constant C, depending on Ω, d and the family of triangulations $\{\mathcal{T}_h\}_{h>0}$, such that for $\mathscr{E}_h^{i,b} \ni F \subset \partial K$, on F we have that*

$$(\nabla_{\mathbf{T}} v)^T \nabla_{\mathbf{T}} n_F^T \nabla_{\mathbf{T}} w \leq C |\nabla_{\mathbf{T}} v| |\nabla_{\mathbf{T}} w| \quad \forall v, w \in H^s(K),\ s > 3/2. \tag{4.10.16}$$

If, in addition, Ω is a piecewise C^2 piecewise convex domain, then, for $\mathscr{E}_h^b \ni F \subset \partial K$, on F we have that

$$\mathcal{H}_F \geq 0. \tag{4.10.17}$$

Proof: First, let us assume that $F \in \mathscr{E}_h^b$. Then, since Ω is piecewise C^2, $F \subset \Gamma_n \subset \partial\Omega$, where Γ_n is a C^2 portion of $\partial\Omega$. It then follows that for a given $F \in \mathscr{E}_h^b$, n_F is of class $C^1(F)$. For any two vectors $\xi^1, \xi^2 : F \to \mathbb{R}^d$ tangent to F, it then follows that

$$(\xi^1)^T \nabla_{\mathbf{T}} n_F^T \xi^2 \leq \sup_{x \in F} |\nabla_{\mathbf{T}} n_F^T(x)| |\xi_1| |\xi_2| \leq \sup_{x \in \Gamma_n} |\nabla_{\mathbf{T}} n_{\Gamma_n}^T(x)| |\xi_1| |\xi_2|.$$

Thus, for an arbitrary $F \in \mathscr{E}_h^b$,

$$(\xi^1)^T \nabla_{\mathbf{T}} n_F^T \xi^2 \leq \max_{i=1,\ldots,N} \sup_{x \in \Gamma_n} |\nabla_{\mathbf{T}} n_{\Gamma_n}^T(x)| |\xi_1| |\xi_2| = C(\Omega) |\xi_1| |\xi_2|,$$

where the constant above depends on Ω, as the portions Γ_n are determined by Ω. If $F \in \mathscr{E}_h^i$, then we may express F locally as the graph of a function determined by one of the maps F_K that make up the mesh $\mathscr{T}_h$; we also have that $F_K \in C^2$, as the family of meshes is regular of order 1. That is, since $F \subset \partial K$ for some $K \in \mathscr{T}_h$, there exists a (straight) reference face $\hat{F}$, such that $F = F_K(\hat{F})$. Furthermore, there exists a straight approximating face $\tilde{F} = \tilde{F}_K(\hat{F})$ ($\tilde{F}_K$ is the affine part of F_K), which provides us with a local coordinate system. As $\tilde{F}$ is flat, after a suitable change of coordinates, one has that $\tilde{F} \subset \{(x', 0) : x' \in \mathbb{R}^{d-1}\}$. Furthermore, without loss of generality, we may assume that F does not intersect $\tilde{F}$, since in such a case, we may define another flat face $\tilde{F}_a$, and an invertible affine map $\tilde{A} : \tilde{F}_a \to \tilde{F}$ that consists only of rotation and translation, which does not affect the bounds that we are about

to obtain (i.e., the Euclidean norm of the matrix $D\tilde{A}$ is equal to 1). Let us denote $\tilde{F}' = \{x' \in \mathbb{R}^{d-1} : (x', 0) \in \tilde{F}\}$. Now, defining $\varphi_{F_K} : \tilde{F}' \to \mathbb{R}$ by

$$\varphi_{F_K}(x') = [F_K]^d \circ \tilde{F}_K^{-1}(x', 0), \quad x' \in \tilde{F}',$$

we see that $F = \{(x', \varphi_{F_K}(x')) : x' \in \tilde{F}'\}$. Note that we have now shown that all $F \in \mathscr{E}_h^{i,b}$ are of class C^2, and furthermore, for any $K \in \mathscr{T}_h$, ∂K may be expressed as the finite union of the closures of $F \in \mathscr{E}_h^{i,b}$, and thus for all $K \in \mathscr{T}_h$, ∂K is piecewise C^2.

Furthermore, expressing F as the zero level set of the function $h_{F_K}(x', x_d) = x_d - \varphi_{F_K}(x')$, we see that

$$n_F = \frac{\nabla h_{F_K}}{|\nabla h_{F_K}|} = -\frac{(\nabla_{x'}\varphi_{F_K}, -1)^T}{|(\nabla_{x'}\varphi_{F_K}, -1)|} = -\frac{(\nabla_{x'}\varphi_{F_K}, -1)^T}{\sqrt{1 + |\nabla_{x'}\varphi_{F_K}|^2}}.$$

Then, since $\nabla_{\mathbf{T}} n_F^T = \nabla n_F^T - n_F \frac{\partial n_F^T}{\partial n_F}$, for any two vectors $\xi^1, \xi^2 : F \to \mathbb{R}^d$ tangent to F (with components $\xi_1^k, \ldots, \xi_d^k$, $k = 0, 1$), and hence orthogonal to n_F, it follows that $(\xi^1)^T \nabla_{\mathbf{T}} n_F^T \xi^2 = (\xi^1)^T \nabla n_F^T \xi^2$. Furthermore, denoting $\delta^{ij} := 1 - \delta_{ij}$, where δ_{ij} is the Kronecker-delta symbol, we see that

$$\begin{aligned}
\frac{\partial [n_F]^j}{\partial x_i} &= -\delta^{id} \frac{\partial}{\partial x_i} \left(\frac{(\delta^{jd} \frac{\partial \varphi_{F_K}}{\partial x_j} - \delta_{jd})}{\sqrt{|\nabla_{x'}\varphi_{F_K}|^2 + 1}} \right) \\
&= -\delta^{id} \frac{\delta^{jd}(|\nabla_{x'}\varphi_{F_K}|^2 + 1) \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_j} - (\delta^{jd} \frac{\partial \varphi_{F_K}}{\partial x_j} - \delta_{jd}) \sum_{k=1}^{d-1} \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_k} \frac{\partial \varphi_{F_K}}{\partial x_k}}{(|\nabla_{x'}\varphi_{F_K}|^2 + 1)^{3/2}},
\end{aligned}$$

and so

$$\begin{aligned}
(\xi^1)^T \nabla n_F^T \xi^2 &= -\frac{\sum_{i,j=1}^{d-1} \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_j} \xi_i^1 \xi_j^2}{\sqrt{|\nabla_{x'}\varphi_{F_K}|^2 + 1}} + \frac{\sum_{i,k=1}^{d-1} \xi_i^1 \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_k} \frac{\partial \varphi_{F_K}}{\partial x_k} \sum_{j=1}^{d} (\delta^{jd} \frac{\partial \varphi_{F_K}}{\partial x_j} - \delta_{jd}) \xi_j^2}{(|\nabla_{x'}\varphi_{F_K}|^2 + 1)^{3/2}} \\
&= -\frac{\sum_{i,j=1}^{d-1} \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_j} \xi_i^1 \xi_j^2}{\sqrt{|\nabla_{x'}\varphi_{F_K}|^2 + 1}} - \frac{\sum_{i,k=1}^{d-1} \xi_i^1 \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_k} \frac{\partial \varphi_{F_K}}{\partial x_k} (\xi^2 \cdot n_F)}{(|\nabla_{x'}\varphi_{F_K}|^2 + 1)} \\
&= -\frac{\sum_{i,j=1}^{d-1} \frac{\partial^2 \varphi_{F_K}}{\partial x_i \partial x_j} \xi_i^1 \xi_j^2}{\sqrt{|\nabla_{x'}\varphi_{F_K}|^2 + 1}} \leq \|D_{x'}^2 \varphi_{F_K}\| |\xi^1| |\xi^2|.
\end{aligned}$$

One also has that

$$\begin{aligned}\sup_{x'\in\tilde{F}'}\|D^2_{x'}\varphi_{F_K}(x')\| &\leq \sup_{x'\in\tilde{F}'}\|D^2_{x'}(F_K\circ\tilde{F}_K^{-1})(x',0)\|\\ &\leq \sup_{x\in\tilde{F}}\|D^2_{x}(F_K\circ\tilde{F}_K^{-1})(x)\|\\ &= \sup_{x\in\tilde{F}}\|D^2F_K\circ\tilde{F}_K^{-1}(x)(\tilde{B}_K^{-1})^2\|\\ &= \sup_{x\in\hat{F}}\|D^2F_K(\hat{x})(\tilde{B}_K^{-1})^2\|\\ &\leq \sup_{x\in\hat{F}}\|D^2F_K(\hat{x})\|\|\tilde{B}_K^{-1}\|^2 = c_2 \leq C_{\text{int}},\end{aligned}$$

where the final inequality follows from (4.4.23), and C_{int} is independent of both h, and the choice of F, since the family of meshes is regular of order 1. Thus, defining $C := \max\{C(\Omega), C_{\text{int}}\}$. For all $F \in \mathscr{E}_h^{i,b}$, we have

$$(\xi^1)^T\nabla_{\mathbf{T}}n_F^T\,\xi^2 \leq C|\xi^1||\xi^2|, \tag{4.10.18}$$

on F, for any tangent vectors to F. Upon noting that $\nabla_{\mathbf{T}}u$, and $\nabla_{\mathbf{T}}v$ are tangent vectors to F, we obtain (4.10.16).

Now, let us assume further that Ω is also piecewise convex. Then, for any $F \in \mathscr{E}_h^{\partial b}$, $F \subset \Gamma_n \subset \partial\Omega$, where Γ_n is a C^2 portion of $\partial\Omega$, and so F can be expressed locally as the graph of a uniformly C^2 concave function φ_n. Take $x \in F$, then, after a suitable change of coordinates, we may assume that the tangent plane to $\partial\Omega$ at x (and consequently Γ_n and F) is given by $\{(x',0), x' \in \mathbb{R}^{d-1}\}$. Then, by By Definition 2.5.3, at a given $x \in F$, there exists a local neighbourhood V_n of x, with a local coordinate system $\{y_1,\ldots,y_d\}$, such that $F \subset \Gamma_n = \{y^n = (y^{n\prime}, y_d^n) \in V_n : y_d^n = \varphi_n(y^{n\prime})\}$. Let us assume further that $\nabla_{y^{n\prime}}\varphi_n(0) = 0$. This means that the new coordinates have been chosen in a manner so that the hyperplane $\{y_d^n = 0\}$ is tangent to the tangent plane of Γ_n at x. Furthermore, since the tangent plane to Γ_n at x is given by $\{(x',0) : x' \in \mathbb{R}^{d-1}\}$, it follows that $\{y_d^n = 0\} = \{(y^{n\prime},0) : y^{n\prime} \in \mathbb{R}^{d-1}\}$. We may also express Γ_n locally via the level set

$$\Gamma_n \cap V_n = \{(y^{n\prime}, y_d^n) \in V : g_n(y^{n\prime}, y_d^n) = y_d^n - \varphi_n(y^{n\prime}) = 0\},$$

and so we may express the unit normal n_F (corresponding to the unit outward normal of $\partial\Omega$), as

$$n_F = \frac{\nabla_{y^n}g_n}{|\nabla_{y^n}g_n|} = -\frac{(\nabla_{y^{n\prime}}\varphi_n,-1)^T}{|(\nabla_{y^{n\prime}}\varphi_n,-1)^T|} = -\frac{(\nabla_{y^{n\prime}}\varphi_n,-1)^T}{\sqrt{|\nabla_{y^{n\prime}}\varphi_n|^2+1}}.$$

We see that for $i, j = 1 \ldots, d$,

$$\begin{aligned}\left[\nabla_{\mathbf{T}} n_F^T\right]_j^i &= -\delta^{id}\frac{\partial}{\partial y_i^n}\left(\frac{(\delta^{jd}\frac{\partial \varphi_n}{\partial y_j^n} - \delta_{jd})}{\sqrt{|\nabla_{y^{n\prime}}\varphi_n|^2 + 1}}\right) \\ &= -\delta^{di}\frac{\delta^{jd}\frac{\partial^2 \varphi_n}{\partial y_i \partial y_j} - (\delta^{jd}\frac{\partial \varphi_n}{\partial y_j} - \delta_{jd})\sum_{k=1}^{d-1}\frac{\partial^2 \varphi_n}{\partial y_i \partial y_k}\frac{\partial \varphi_n}{\partial y_k}}{(|\nabla_{y^{n\prime}}\varphi_n|^2 + 1)^{3/2}}\end{aligned}$$

Then, since $\nabla_{y^{n\prime}}\varphi(0) = 0$, at x we have for $i, j = 1, \ldots, d$

$$\left[\nabla_{\mathbf{T}} n_F^T\right]_j^i = -\delta^{id}\delta^{jd}\frac{\partial^2 \varphi_n(0)}{\partial y_i \partial y_j}.$$

Thus, taking ξ, to be a tangent vector to $\partial\Omega$ at x, with $\{\xi_1, \ldots, \xi_{d-1}\}$ denoting the components of ξ in the directions $y_1^n, \ldots, y_{d-1}^n$, we see that at x

$$\xi^T \nabla_{\mathbf{T}} n_F^T \xi = -\sum_{i,j=1}^{d-1}\frac{\partial^2 \varphi_n(0)}{\partial y_i \partial y_j}\xi_i \xi_j \geq 0, \tag{4.10.19}$$

since φ_n is concave, and so $-\varphi_n$ is convex. The inequality (4.10.19) is independent of x, and we thus deduce that it holds everywhere on F. One can see that (4.10.19) implies (4.10.17). Indeed, let us take $\psi \in \mathbb{R}^d$, and decompose ψin terms of its tangential and normal components, i.e., $\psi = \psi_{\mathbf{T}} + \psi_{n_F} n_F$. Then, we see that on F

$$\begin{aligned}\psi^T \nabla_{\mathbf{T}} n_F^T \psi &= (\psi_{\mathbf{T}} + \psi_{n_F} n_F)^T \nabla_{\mathbf{T}} n_F^T (\psi_{\mathbf{T}} + \psi_{n_F} n_F) \\ &= (\psi_{\mathbf{T}})^T \nabla_{\mathbf{T}} n_F^T \psi_{\mathbf{T}} + (\psi_{n_F} n_F) n_F^T \nabla_{\mathbf{T}} n_F^T (\psi_{\mathbf{T}} + \psi_{n_F} n_F) \\ &\quad + (\psi_{n_F}\psi_{n_F}) n_F^T \nabla_{\mathbf{T}} n_F^T n_F \\ &= (\psi_{\mathbf{T}})^T \nabla_{\mathbf{T}} n_F^T \psi_{\mathbf{T}} \\ &\geq 0,\end{aligned} \tag{4.10.20}$$

where the final inequality is due to (4.10.19). Thus, $\nabla_{\mathbf{T}} n_F^T$ is positive semidefinite on F, and it follows that on F,

$$\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F = \mathrm{Tr}(\nabla_{\mathbf{T}} n_F^T) \geq 0,$$

which is (4.10.17). $\square$

Lemma 4.10.4 *Assume that Ω is piecewise C^2, and let $\{\mathscr{T}_h\}_{h>0}$ be a family of meshes on $\overline{\Omega}$ that satisfies Assumption 4.4.9. Then, there exists a constant C depending on the family $\{\mathscr{T}_h\}_{h>0}$, d, and Ω such that for any $F \in \mathscr{E}_h^b$, the following*

estimates hold on F:

$$\sup_{x\in F}|\mathcal{H}_F(x)| \le C(d)\sup_{x\in F}|\nabla_{\mathbf{T}} n_F^T(x)| \le C, \tag{4.10.21}$$

$$\left|\nabla_{\mathbf{T}}\left(\tau_F\frac{\partial v}{\partial n_F}\right)\right| \le C(|\tau_F(D^2v)| + |\tau_F(\nabla v)|), \tag{4.10.22}$$

$$|\operatorname{div}_{\mathbf{T}}\nabla_{\mathbf{T}}\tau_F(v)| \le C(|\tau_F(D^2v)| + |\tau_F(\nabla v)|), \tag{4.10.23}$$

for all $v \in H^s(K)$, $s > 5/2$, where $F \subset \partial K$, and τ_F is the trace operator from K to F.

Proof: Let $F \in \mathscr{E}_h^{i,b}$. Then, by definition, we see that

$$\sup_{x\in F}|\mathcal{H}_F(x)| = \sup_{x\in F}|\nabla_{\mathbf{T}}\cdot n_F(x)| \le C(d)|\sup_{x\in F}\nabla_{\mathbf{T}} n_F^T(x)|. \tag{4.10.24}$$

Furthermore, let us take $\xi^1, \xi^2 \in \mathbb{R}^d$, and decompose them in terms of their tangential and normal components, i.e., $\xi^k = (\xi^k)_{\mathbf{T}} + (\xi^k_{n_F})n_F$, $k = 1, 2$. Then, we see that on F

$$\begin{aligned}(\xi^1)^T\nabla_{\mathbf{T}} n_F^T\,\xi^2 &= (\xi^1_{\mathbf{T}})^T\nabla_{\mathbf{T}} n_F^T\,\xi^2_{\mathbf{T}}\\ &\le C|\xi^1_{\mathbf{T}}||\xi^2_{\mathbf{T}}| \le C|\xi^1||\xi^2|,\end{aligned} \tag{4.10.25}$$

where the last inequality is due to (4.10.18), as $(\xi^k)_{\mathbf{T}}$ are tangent vectors. Since this holds for all $\xi^1, \xi^2 \in \mathbb{R}^d$, we deduce that

$$\sup_{x\in F}|\nabla_{\mathbf{T}} n_F^T(x)| \le C, \tag{4.10.26}$$

which, combined with (4.10.24) yields

$$\sup_{x\in F}|\mathcal{H}_F(x)| \le C(d)\sup_{x\in F}|\nabla_{\mathbf{T}} n_F^T(x)| \le C, \tag{4.10.27}$$

which is (4.10.21).

Then, by (4.10.4) and (4.10.27) we see that on F

$$\begin{aligned}|\operatorname{div}_{\mathbf{T}}\nabla_{\mathbf{T}}\tau_F(v)| &= \left|\tau_F(\Delta v) + \mathcal{H}_F\tau_F\frac{\partial v}{\partial n_F} + \tau_F\frac{\partial^2 v}{\partial n_F^2}\right|\\ &\le C(d)|\tau_F(D^2v)| + \sup_{x\in F}|\mathcal{H}_F(x)||\tau_F(\nabla u)|\\ &\le C(|\tau_F(D^2v)| + |\tau_F(\nabla v)|),\end{aligned}$$

which is (4.10.23). Finally, from (4.10.27) we obtain the following

$$\begin{aligned}\left|\nabla_{\mathbf{T}}\tau_F\left(\frac{\partial v}{\partial n_F}\right)\right| &= \left|\nabla_{\mathbf{T}}(\tau_F(Dv))\cdot n_F + (\nabla_{\mathbf{T}} n_F^T)\cdot\nabla v\right|\\ &\le |\nabla_{\mathbf{T}}\tau_F(Dv)| + |\nabla_{\mathbf{T}} n_F^T||\nabla v|\\ &\le C(|\tau_F(D^2v)| + |\tau_F(\nabla v)|),\end{aligned}$$

which is (4.10.22). □

4.11 Finite element Hessian

Definition 4.11.1 *For $v \in H^1(\Omega)$ with $[\nabla v\, n^T]^i_j \in (H^{1/2}(\partial\Omega))'$, $i,j = 1,\dots,d$, we define the ($\mathbb{V}_{h,p}$-)finite element Hessian, $\mathbf{H}_h v$, to be the unique (via the Riesz representation theorem) element of $\mathbb{W}_{h,p}$ that satisfies*

$$\langle [\mathbf{H}_h v]^i_j, \Phi\rangle_\Omega = \langle [D^2 v]^i_j | \Phi\rangle \quad \forall \Phi \in \mathbb{V}_{h,p}, 1 \le i \le d, i \le j \le d, \tag{4.11.1}$$

where $D^2 v$ is the generalised Hessian of v given by (2.4.1).

Remark 4.11.2 *If $v \in \mathbb{V}_{h,p}$, then $\nabla v|_{\partial\Omega} \in L^2(\partial\Omega)^d$. Furthermore, if $\partial\Omega$ is Lipschitz continuous, then $n \in L^\infty(\partial\Omega)$, and so $[\nabla v\, n^T]^i_j \in L^2(\partial\Omega) \subset (H^{1/2}(\partial\Omega))'$. In (4.11.1), we take $\Phi \in \mathbb{V}_{h,p} \subset H^1(\Omega)$, and so $\tau_{\partial\Omega}(\Phi) \in H^{1/2}(\partial\Omega)$. It then follows that*

$$\begin{aligned}
\langle [\mathbf{H}_h v]^i_j, \Phi\rangle_\Omega &= \langle [D^2 v]^i_j | \Phi\rangle \\
&= -\left\langle \frac{\partial v}{\partial x_i}, \frac{\partial \Phi}{\partial x_j}\right\rangle_\Omega + \langle [\nabla v\, n^T]^i_j | \Phi\rangle_{(H^{1/2}(\partial\Omega))'\times H^{1/2}(\partial\Omega)} \\
&= -\left\langle \frac{\partial v}{\partial x_i}, \frac{\partial \Phi}{\partial x_j}\right\rangle_\Omega + \langle [\nabla v\, n^T]^i_j, \Phi\rangle_{\partial\Omega}
\end{aligned}$$

for all $i,j = 1,\dots,d$. That is, each duality pairing is given by an L^2-inner product over the boundary.

Definition 4.11.3 (Finite element convexity) *A function $v \in H^1(\Omega)$, with $[\nabla v\, n^T]^i_j \in (H^{1/2}(\partial\Omega))'$, $i,j = 1,\dots,d$, is said to be uniformly finite element convex with respect to $\mathbb{V}_{h,p}$, or $\mathbb{V}_{h,p}$-uniformly convex, if and only if*

$$\langle \mathbf{H}_h v, \Phi\rangle_\Omega \quad \textit{is symmetric positive definite} \quad \forall \Phi \in \mathbb{V}_{h,p} \setminus \{0\} : \Phi \ge 0. \tag{4.11.2}$$

The following lemma is from [92].

Lemma 4.11.4 *Let $v \in \mathbb{V}_{h,p}$. Then, the finite element Hessian of v, $\mathbf{H}_h v$, for $v \in \mathbb{V}_{h,p}$, satisfies*

$$\mathbf{H}_h v = \mathcal{P}_{\mathbb{W}_{h,p}}(D^2_h v) + \mathcal{L}(\nabla v), \tag{4.11.3}$$

where $\mathcal{L} : [L^2(\mathscr{E}^i_h)]^d \to \mathbb{W}_{h,p}$ is defined by

$$\int_\Omega \mathcal{L}(V) : \Phi = -\sum_{F\in\mathscr{E}^i_h} \int_F [\![[\langle\!\langle \Phi \rangle\!\rangle V] \cdot n_F]\!] \quad \forall \Phi \in \mathbb{W}_{h,p},\ \forall\, V \in [L^2(\mathscr{E}^i_h)]^d, \tag{4.11.4}$$

where n_F is a fixed *choice of unit normal to $F \in \mathscr{E}^i_h$.*

Corollary 4.11.5 (Finite element Hessian of H^2-regular functions) *If $v \in H^2(\Omega)$, then $\mathbf{H}_h v = \mathcal{P}_{\mathbb{W}_{h,p}} D^2 v$.*

Proof: Defining $\mathbf{H}_h v$ by (4.11.3), we see that if $v \in H^2(\Omega)$, then $\nabla v \in H^1(\Omega)$ and thus has zero jump across internal faces. It then follows that $\mathbf{H}_h v = \mathcal{P}_{\mathbb{W}_{h,p}}(D_h^2 v) = \mathcal{P}_{\mathbb{W}_{h,p}}(D^2 v)$. $\square$

The following is an inverse estimate for the finite element Hessian, found in [92] (Lemma 3.1).

Lemma 4.11.6 *Let the finite element Hessian be defined by (4.11.3). Then we have that*

$$\|\mathbf{H}_h v\|_{2,\Omega} \leq C h^{-1} \|v\|_{H^1(\Omega)} \quad \forall v \in \mathbb{V}_{h,p}. \tag{4.11.5}$$

Chapter 5

A DGFEM for linear elliptic equations with Dirichlet boundary conditions

In this chapter, we present and analyse a discontinuous Galerkin finite element method for the approximation of solutions to linear elliptic equations with Dirichlet boundary conditions, that satisfy the Cordes condition (3.3.12). Such problems arise in the linearisation of the MAD problem (3.4.1), and the HJB problem (3.3.31).

5.1 The PDE

Consider the following second-order elliptic boundary–value problem: find $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} Lu = f & \text{in} \quad \Omega, \\ u = 0 & \text{on} \quad \partial\Omega, \end{cases} \tag{5.1.1}$$

$\Omega \subset \mathbb{R}^d$ is a convex domain,

$$Lu := \sum_{i,j=1}^{d} A_{ij} D_{ij}^2 u, \tag{5.1.2}$$

where $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\text{Sym}})$, is uniformly elliptic, that is, there exist $0 < \mu_1 \leq \mu_2 < \infty$ such that

$$\mu_1 |\xi|^2 \leq \xi^T A(x) \xi \leq \mu_2 |\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega, \tag{5.1.3}$$

and that $f \in L^2(\Omega)$ is a given function. Furthermore, we assume that A satisfies the *Cordes condition*: there is an $\varepsilon \in (0, 1]$ such that

$$\frac{|A|^2}{(\text{Tr}(A))^2} = \frac{\sum_{i,j=1}^d A_{ij}^2}{(\sum_{i=1}^d A_{ii})^2} \leq \frac{1}{d - 1 + \varepsilon} \quad \text{a.e. in } \Omega. \tag{5.1.4}$$

Theorem 5.1.1 *Assume that $\Omega \subset \mathbb{R}^d$ is convex, and that $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\mathrm{Sym}})$ satisfies (5.1.3) and (5.1.4). Then, there exists a unique $u \in H^2(\Omega) \cap H^1_0(\Omega)$ that is a strong solution of (5.1.1).*

Proof: We simply apply Theorem 3.4.1, with Λ as a singleton set, corresponding to the single operator L given by (5.1.2), and the single right-hand side $f \in L^2(\Omega)$. We then see that

$$F[u] = \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = Lu - f.$$

Thus Theorem 3.4.1 yields the existence of a unique $u \in H := H^2(\Omega) \cap H^1_0(\Omega)$ that satisfies (5.1.1). $\square$

5.2 Existing framework and original contributions

Existing framework: In [110], the authors proposed a dicontinuous Galerkin finite element method (DGFEM) for the approximation of strong solutions of (5.1.1), with the assumption that the computational domain, $\Omega \subset \mathbb{R}^d$ is both convex and polytopal (notice that this is a stronger assumption than is necessary for the existence and uniqueness of a strong solution, which only requires convexity). The authors were successful in proving that the proposed DGFEM is stable and consistent, implying the existence and uniqueness of a numerical solution that satisfies optimal error estimates in a H^2-type norm.

Original contributions: Our goal in this Chapter is to extend this scheme, allowing for domains that have curved boundaries. In doing so, we are also able to prove that the new scheme (and in fact the scheme proposed in [110]) is stable under the weaker assumption that the computational domain is piecewise convex and piecewise C^3. The analysis of this extended scheme utilises the trace, inverse, Poincaré–Friedrichs' and interpolation estimates developed in Chapter 4.

The method proposed in this chapter will provide a starting point for the development of the the numerical schemes proposed in Chapters 6 and 7, where we will tackle the oblique boundary-value problems, and both HJB problems and MA-HJB problems, respectively. In particular, we will see that that this scheme generalises nicely to the nonlinear setting of HJB problems.

5.2.1 Computational domain assumptions

As mentioned in Section 5.2, the DGFEM we propose in this chapter allows for domains with curved boundaries. In particular, we will assume that the domain $\Omega \subset \mathbb{R}^d$ is piecewise C^3 and piecewise convex. Recall however, that a piecewise convex domain is not necessarily convex; two such examples being the "key-hole" shaped domain (see Figure 2.1), and the L-shaped domain, which are both piecewise C^∞ and piecewise convex (since the boundaries of both domains have piecewise nonnegative curvature).

This means that there are cases where a numerical solution exists, but the hypotheses of Theorem 5.1.1 are not satisfied, and so we cannot deduce the existence and uniqueness of the corresponding strong solution of (5.1.1).

It is however, important to note that any polytopal domain in $\mathbb{R}^d$ is piecewise C^∞ and piecewise convex, and so the DGFEM we propose *extends* the class of domains considered in [110].

5.3 The design of the numerical method

In order to understand the terms that arise in the bilinear form for the numerical method of this chapter, it is first useful to recap the proof of Theorem 3.3.8, particularly in the context of the linear nondivergence form problem (5.1.1). Indeed, the existence and uniqueness in the linear case follows from the fact that (5.1.1) is a special case of the well posed nonlinear HJB problem (as mentioned in the proof of Theorem 5.1.1). To this end, for $\Omega \subset \mathbb{R}^d$, convex, let $H = H^2(\Omega) \cap H_0^1(\Omega)$, and define $A_\gamma : H \times H \to \mathbb{R}$ by

$$A_\gamma(u, v) := \int_\Omega \gamma A : D^2 u \, \Delta v \quad \forall u, v \in H,$$

where we recall that the renormalisation factor $\gamma : \Omega \to \mathbb{R}^+$, is defined by

$$\gamma := \frac{\operatorname{Tr} A}{|A|^2}.$$

Taking into account that γ is uniformly positive (when restricted to matrix–valued functions with uniformly positive trace), we see that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ solves

$$\begin{cases} \gamma L u = \gamma f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{5.3.1}$$

uniquely, if and only if it is the unique solution of (5.1.1). Thus, we consider the following strong formulation of (5.3.1), that is: find $u \in H := H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$A_\gamma(u, v) = \int_\Omega \gamma f \Delta v =: L_f(v) \quad \forall v \in H. \tag{5.3.2}$$

Upon showing that A_γ is coercive and bounded on $H \times H$, and that L_f is bounded on H, the Lax–Milgram Theorem [46] yields the existence and uniqueness of a $u \in H$ that satisfies (5.3.2). However, in order to prove that A_γ is indeed coercive, we utilise the Miranda–Talenti (MT) estimates (3.3.9) and (3.3.10), which hold for the space H (by virtue of Theorems 3.3.14 and 3.3.15), but they do not hold for general functions that belong to the DG finite element space, which we recall is defined by

$$V_{h,p} := \{v \in L^2(\Omega) : \ v|_K = \hat{\rho} \circ F_K^{-1}, \, \hat{\rho} \in \mathbb{P}^p(\hat{K}), \ \forall K \in \mathscr{T}_h\}.$$

For clarity, we provide the proof that A_γ is coercive on $H \times H$. In particular, we highlight the step where we add and subtract the term $\|\Delta u\|^2_{L^2(\Omega)}$. One has that

$$\begin{aligned}
A_\gamma(u, u) &= \int_\Omega \gamma A : D^2 u \, \Delta u \\
&= \int_\Omega \gamma A : D^2 u \, \Delta u + (\Delta u)^2 - (\Delta u)^2 && \text{(add and subtract } \|\Delta u\|^2_{L^2(\Omega)}) \\
&= \|\Delta u\|^2_{L^2(\Omega)} + \int_\Omega (\gamma A - I_d) : D^2 u \, \Delta u \\
&\geq \|\Delta u\|^2_{L^2(\Omega)} + \sqrt{1-\varepsilon} |u|_{H^2(\Omega)} \|\Delta u\|_{L^2(\Omega)} && \text{(by the Cordes condition)} \\
&\geq (1 - \sqrt{1-\varepsilon}) \|\Delta u\|^2_{L^2(\Omega)} && \text{(by (3.3.9))} \\
&\geq C \|u\|^2_{H^2(\Omega)}, && \text{(by (3.3.10))}
\end{aligned}$$

i.e., A_γ is coercive (note that the boundedness of A_γ and L_f follow in a similar manner). Our first step into designing the scheme is to provide an analogue of A_γ on $V_{h,p} \times V_{h,p}$, i.e., the following

$$A^h_\gamma(u_h, v_h) := \sum_{K \in \mathscr{T}_h} \langle \gamma A : D^2 u_h, \Delta v_h \rangle_K \quad \forall u_h, v_h \in V_{h,p}.$$

There is no reason, however, that such a bilinear form should be coercive on $V_{h,p} \times V_{h,p}$. We may follow the proof of coercivity for A_γ, obtaining

$$A^h_\gamma(u_h, u_h) \geq \sum_{K \in \mathscr{T}_h} \|\Delta u_h\|^2_{L^2(K)} - \sqrt{1-\varepsilon} |u|_{H^2(\Omega; \mathscr{T}_h)} \left(\sum_{K \in \mathscr{T}_h} \|\Delta u_h\|^2_{L^2(K)} \right)^{\frac{1}{2}},$$

but there does not seem to be any way to proceed to the next step without the MT estimate (3.3.9). We must instead numerically enforce the MT estimates. This is done

by adding an additional bilinear form to A^h_γ, that "a priori" includes the addition and subtraction of the $\|\Delta u\|^2_{L^2(\Omega)}$ term; that is, we define

$$\begin{aligned} A^{\mathcal{D}}_h(u_h, v_h) &:= A^h_\gamma(u_h, v_h) + B^{\mathcal{D}}_{h,1/2}(u_h, v_h) - \sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K \\ &= \sum_{K\in\mathscr{T}_h} \langle \gamma A : D^2 u_h, \Delta v_h\rangle_K + B^{\mathcal{D}}_{h,1/2}(u_h, v_h) - \sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K, \end{aligned} \tag{5.3.3}$$

where we *claim* that the bilinear form $B^{\mathcal{D}}_{h,1/2}$ is coercive (this claim is justified by Lemma 5.6.2), and that

$$B^{\mathcal{D}}_{h,1/2}(w, v_h) = \sum_{K\in\mathscr{T}_h} \langle \Delta w, \Delta v_h\rangle_K \tag{5.3.4}$$

for $w \in H^2(\Omega)\cap H^1_0(\Omega)\cap H^s(\Omega;\mathscr{T}_h)$, $s > 5/2$ (we will discuss the assertion of this broken regularity later on), and for all $v_h \in V_{h,p}$. It is then clear that (5.3.4) implies

$$A^{\mathcal{D}}_h(w, v_h) = A^h_\gamma(w, v_h) \quad \forall v_h \in V_{h,p}, \tag{5.3.5}$$

for such w. Note that the above identity is referred to as the *consistency* of the scheme, and plays an important role in the derivation of error estimates for the numerical solution. Furthermore, we remark that the bilinear form, $B^{\mathcal{D}}_{h,1/2}$, will take the following form

$$B^{\mathcal{D}}_{h,1/2}(u_h, v_h) := \frac{1}{2} B^{\mathcal{D}}_{h,*}(u_h, v_h) + \frac{1}{2}\sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K + J^{\mathcal{D}}_h(u_h, v_h), \tag{5.3.6}$$

and that the bilinear forms $B^{\mathcal{D}}_{h,*}, J^{\mathcal{D}}_h : V_{h,p}\times V_{h,p} \to \mathbb{R}$ satisfy

$$B^{\mathcal{D}}_{h,*}(w, v_h) = \sum_{K\in\mathscr{T}_h} \langle \Delta w, \Delta v_h\rangle_K, \quad \text{and} \quad J^{\mathcal{D}}_h(w, v_h) = 0,$$

for the same choice of w, and all $v_h \in V_{h,p}$. Moreover, one can see that the above implies (5.3.4) which in turn implies (5.3.5). We also remark that the bilinear form $J^{\mathcal{D}}_h$ plays no role in the consistency identity (5.3.5) (other than by its absence), and is in fact a jump penalty term that enforces regularity that is consistent with that of the true solution. In particular, if $w \in H^2(\Omega)\cap H^1_0(\Omega)$ (which is the space that the strong solution of (5.1.1) belongs to) then we see that

$$[\![w]\!] = [\![\nabla w \cdot n_F]\!] = [\![\nabla_{\mathbf{T}} w]\!] = 0 \quad \forall F \in \mathscr{E}^i_h, \tag{5.3.7}$$

and furthermore, since $\tau_F(w) = 0$ for all $F \in \mathscr{E}^b_h$, it follows that

$$[\![w]\!] = [\![\nabla_{\mathbf{T}} w]\!] = 0 \quad \forall F \in \mathscr{E}^b_h. \tag{5.3.8}$$

$J_h^{\mathcal{D}}$ also enforces the Dirichlet boundary condition, and leads to the bilinear form $B_{h,1/2}^{\mathcal{D}}$ being provably coercive (in a particular broken H^2-type norm on $V_{h,p}$). In particular we define $J_h^{\mathcal{D}}$ as follows:

$$\begin{aligned} J_h^{\mathcal{D}}(u_h, v_h) := &\sum_{F \in \mathscr{E}_h^{i,b}} [\mu_F \langle [\![\nabla_{\mathbf{T}} u_h]\!], [\![\nabla_{\mathbf{T}} v_h]\!] \rangle_F + \eta_F \langle [\![u_h]\!], [\![v_h]\!] \rangle_F] \\ &+ \sum_{F \in \mathscr{E}_h^{i}} \mu_F \langle [\![\nabla u_h \cdot n_F]\!], [\![\nabla v_h \cdot n_F]\!] \rangle_F, \end{aligned} \tag{5.3.9}$$

where the positive edge-dependent quantities μ_F and η_F will be specified later, and they will be chosen in a manner that allows us to prove that $B_{h,1/2}^{\mathcal{D}}$ is coercive (see Lemma 5.6.2). Furthermore, (5.3.7) and (5.3.8) imply that

$$J_h^{\mathcal{D}}(w, v_h) = 0 \tag{5.3.10}$$

for $w \in H^2(\Omega) \cap H_0^1(\Omega) \cap H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$, and all $v_h \in V_{h,p}$.

The bilinear form $B_{h,*}^{\mathcal{D}}$ plays a key role (when paired with the remaining term in the definition of $B_{h,1/2}^{\mathcal{D}}$) in identity (5.3.4), and its structure is motivated by Lemma 4.10.2, the statement of which we recall.

Statement of Lemma 4.10.2: For any $K \in \mathscr{T}_h$, and any $u, v \in H^s(K)$, $s > 5/2$, we have that

$$\begin{aligned} \int_K \Delta u \, \Delta v = &\int_K D^2 u : D^2 v + \int_{\partial K} \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} u \frac{\partial v}{\partial \overline{n}} - \nabla_{\mathbf{T}} \left(\frac{\partial u}{\partial \overline{n}} \right) \cdot \nabla_{\mathbf{T}} v \\ &+ \int_{\partial K} \mathcal{H}_{\partial K} \frac{\partial u}{\partial \overline{n}} \frac{\partial v}{\partial \overline{n}} + (\nabla_{\mathbf{T}} u)^T \nabla_{\mathbf{T}} \overline{n}^T \nabla_{\mathbf{T}} v, \end{aligned} \tag{5.3.11}$$

where $\overline{n}$ is the unit outward normal to ∂K, and $\mathcal{H}_{\partial K} := \nabla_{\mathbf{T}} \cdot \overline{n}$.

Designing the bilinear form: Now, let us take $w \in H^2(\Omega) \cap H_0^1(\Omega) \cap H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$, and $v_h \in V_{h,p}$, then, in particular $w, v_h \in H^s(K)$, $s > 5/2$, for all $K \in \mathscr{T}_h$. Summing (5.3.11) with $u = w$, and $v = v_h$, over all $K \in \mathscr{T}_h$, we obtain

$$\begin{aligned} &\sum_{K \in \mathscr{T}_h} \langle D^2 w, D^2 v_h \rangle_K + \sum_{F \in \mathscr{E}_h^{i,b}} \int_F [\![(\operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} w)(\nabla v_h \cdot n_F) - \nabla_{\mathbf{T}} (\nabla w \cdot n_F) \cdot \nabla_{\mathbf{T}} v_h]\!] \\ &+ \sum_{F \in \mathscr{E}_h^{i,b}} \int_F [\![\mathcal{H}_F \frac{\partial w}{\partial n_F} \frac{\partial v_h}{\partial n_F} + (\nabla_{\mathbf{T}} w)^T \nabla_{\mathbf{T}} n_F^T \nabla_{\mathbf{T}} v_h]\!] = \sum_{K \in \mathscr{T}_h} \langle \Delta w, \Delta v_h \rangle_K, \end{aligned} \tag{5.3.12}$$

where n_F is now a *fixed* choice of unit normal to F, and $\mathcal{H}_F := \nabla_{\mathbf{T}} \cdot n_F$ is the mean

curvature of the face, determined by the unit normal, n_F. Let us define

$$I(w,v_h) := \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![(\mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}} w)(\nabla v_h\cdot n_F) - \nabla_{\mathbf{T}}(\nabla w\cdot n_F)\cdot\nabla_{\mathbf{T}} v_h]\!] + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![\mathcal{H}_F\frac{\partial w}{\partial n_F}\frac{\partial v_h}{\partial n_F} + (\nabla_{\mathbf{T}} w)^T\nabla_{\mathbf{T}} n_F^T\nabla_{\mathbf{T}} v_h]\!],$$

so, by (5.3.12), it is clear that

$$\sum_{K\in\mathscr{T}_h}\langle D^2 w, D^2 v_h\rangle_K + I(w,v_h) = \sum_{K\in\mathscr{T}_h}\langle\Delta w,\Delta v_h\rangle_K. \tag{5.3.13}$$

Applying the following identity (valid for any $f,g\in H^s(\Omega;\mathscr{T}_h)$, $s>1/2$)

$$\sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![fg]\!] = \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![f]\!]\langle\!\langle g\rangle\!\rangle + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F \langle\!\langle f\rangle\!\rangle [\![g]\!],$$

to $I(w,v_h)$, we obtain

$$\begin{aligned} I(w,v_h) &= \sum_{F\in\mathscr{E}_h^{i}}\int_F \mathrm{div}_{\mathbf{T}}[\![\nabla_{\mathbf{T}} w]\!]\langle\!\langle\nabla v_h\cdot n_F\rangle\!\rangle - \nabla_{\mathbf{T}}[\![\nabla w\cdot n_F]\!]\cdot\langle\!\langle\nabla_{\mathbf{T}} v_h\rangle\!\rangle \\ &\quad + \sum_{F\in\mathscr{E}_h^{i}}\int_F \mathcal{H}_F[\![\nabla w\cdot n_F]\!]\langle\!\langle\nabla v_h\cdot n_F\rangle\!\rangle + [\![\nabla_{\mathbf{T}} w]\!]^T\nabla_{\mathbf{T}} n_F^T\langle\!\langle\nabla_{\mathbf{T}} v_h\rangle\!\rangle \\ &\quad + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F \mathrm{div}_{\mathbf{T}}\langle\!\langle\nabla_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!] - \nabla_{\mathbf{T}}\langle\!\langle\nabla w\cdot n_F\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!] \\ &\quad + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F \mathcal{H}_F\langle\!\langle\nabla w\cdot n_F\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!] + \langle\!\langle\nabla_{\mathbf{T}} w\rangle\!\rangle^T\nabla_{\mathbf{T}} n_F^T[\![\nabla_{\mathbf{T}} v_h]\!] \\ &= \sum_{F\in\mathscr{E}_h^{i}}\int_F \mathrm{div}_{\mathbf{T}}\langle\!\langle\nabla_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!] + \langle\!\langle\nabla_{\mathbf{T}} w\rangle\!\rangle^T\nabla_{\mathbf{T}} n_F^T[\![\nabla_{\mathbf{T}} v_h]\!] \\ &\quad + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F \mathcal{H}_F\langle\!\langle\nabla w\cdot n_F\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!] - \nabla_{\mathbf{T}}\langle\!\langle\nabla w\cdot n_F\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!], \end{aligned} \tag{5.3.14}$$

where the final equality is a direct consequence of (5.3.7) and (5.3.8). A further application of (5.3.7) and (5.3.8) to (5.3.14) consistently symmetrises of the terms present in (5.3.14), which will lead to the symmetrisation of the associated bilinear

form, $B^{\mathcal{D}}_{h,*}$. That is, we obtain

$$\begin{aligned} I(w,v_h) = &\sum_{F\in\mathscr{E}_h^i}\int_F \mathrm{div}_{\mathbf{T}}\langle\!\langle \nabla_{\mathbf{T}} w\rangle\!\rangle [\![\nabla v_h\cdot n_F]\!] + \mathrm{div}_{\mathbf{T}}\langle\!\langle \nabla_{\mathbf{T}} v_h\rangle\!\rangle [\![\nabla w\cdot n_F]\!] \\ &+\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle \nabla_{\mathbf{T}} w\rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} v_h]\!] + \langle\!\langle \nabla_{\mathbf{T}} v_h\rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} w]\!] \\ &+\sum_{F\in\mathscr{E}_h^i}\int_F \mathcal{H}_F\langle\!\langle \nabla v_h\cdot n_F\rangle\!\rangle [\![\nabla w_h\cdot n_F]\!] + \sum_{F\in\mathscr{E}_h^{i,b}}\int_F \mathcal{H}_F\langle\!\langle \nabla w\cdot n_F\rangle\!\rangle [\![\nabla v_h\cdot n_F]\!] \\ &-\sum_{F\in\mathscr{E}_h^{i,b}}\int_F \nabla_{\mathbf{T}}\langle\!\langle \nabla w\cdot n_F\rangle\!\rangle\cdot [\![\nabla_{\mathbf{T}} v_h]\!] - \nabla_{\mathbf{T}}\langle\!\langle \nabla v_h\cdot n_F\rangle\!\rangle\cdot [\![\nabla_{\mathbf{T}} w]\!] =: \tilde{I}(w,v_h). \end{aligned} \tag{5.3.15}$$

Overall, we have obtained

$$\begin{aligned} \sum_{K\in\mathscr{T}_h}\langle\Delta w,\Delta v_h\rangle_K &= \sum_{K\in\mathscr{T}_h}\langle D^2 w, D^2 v_h\rangle_K + I(w,v_h) \\ &= \sum_{K\in\mathscr{T}_h}\langle D^2 w, D^2 v_h\rangle_K + \tilde{I}(w,v_h) \\ &=: B^{\mathcal{D}}_{h,*}(w,v_h). \end{aligned} \tag{5.3.16}$$

This gives us our definition of $B^{\mathcal{D}}_{h,*}(u_h,v_h)$, that is

$$B^{\mathcal{D}}_{h,*}(u_h,v_h) := \sum_{K\in\mathscr{T}_h}\langle D^2 w, D^2 v_h\rangle_K + \tilde{I}(w,v_h),$$

and, by virtue of (5.3.16), we have that

$$B^{\mathcal{D}}_{h,*}(w,v_h) = \sum_{K\in\mathscr{T}_h}\langle\Delta w,\Delta v_h\rangle_K, \tag{5.3.17}$$

for $w\in H^2(\Omega)\cap H^1_0(\Omega)\cap H^s(\Omega;\mathscr{T}_h)$, $s>5/2$, and all $v_h\in V_{h,p}$. Furthermore, since the right-hand side of (5.3.15) is symmetric, it follows that $B^{\mathcal{D}}_{h,*}(u_h,v_h)$ is also symmetric, i.e., $B^{\mathcal{D}}_{h,*}(u_h,v_h) = B^{\mathcal{D}}_{h,*}(v_h,u_h)$ for all $u_h,v_h\in V_{h,p}$. We are now ready to define the numerical method of this chapter.

5.4 The numerical method

The definition of the numerical scheme requires the following bilinear forms, derived in Section 5.3, and concisely defined as follows. Firstly, the bilinear form $B^{\mathcal{D}}_{h,*}$:

$V_{h,p} \times V_{h,p} \to \mathbb{R}$ is defined by

$$
\begin{aligned}
B^{\mathcal{D}}_{h,*}(u_h, v_h) &:= \sum_{K \in \mathscr{T}_h} \langle D^2 u_h, D^2 v_h \rangle_K \\
&+ \sum_{F \in \mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}} \langle\!\langle \nabla_{\mathbf{T}} u_h \rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!] \rangle_F + \langle \mathrm{div}_{\mathbf{T}} \langle\!\langle \nabla_{\mathbf{T}} v_h \rangle\!\rangle, [\![\nabla u_h \cdot n_F]\!] \rangle_F \\
&+ \sum_{F \in \mathscr{E}_h^i} \langle \langle\!\langle \nabla_{\mathbf{T}} u_h \rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} v_h]\!] \rangle_F + \langle \langle\!\langle \nabla_{\mathbf{T}} v_h \rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} u_h]\!] \rangle_F \\
&+ \sum_{F \in \mathscr{E}_h^i} \langle \mathcal{H}_F \langle\!\langle \nabla v_h \cdot n_F \rangle\!\rangle, [\![\nabla u_h \cdot n_F]\!] \rangle_F + \sum_{F \in \mathscr{E}_h^{i,b}} \langle \mathcal{H}_F \langle\!\langle \nabla u_h \cdot n_F \rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!] \rangle_F \\
&- \sum_{F \in \mathscr{E}_h^{i,b}} \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla u_h \cdot n_F \rangle\!\rangle, [\![\nabla_{\mathbf{T}} v_h]\!] \rangle_F - \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla v_h \cdot n_F \rangle\!\rangle, [\![\nabla_{\mathbf{T}} u_h]\!] \rangle_F
\end{aligned}
\tag{5.4.1}
$$

where $\mathcal{H}_F$ is the mean curvature of the face F, and $u_h,\ v_h \in V_{h,p}$ throughout this chapter. Then, for positive edge-dependent quantities μ_F and η_F to be specified later, the jump stabilization bilinear form $J_h^{\mathcal{D}} : V_{h,p} \times V_{h,p} \to \mathbb{R}$ is defined by

$$
\begin{aligned}
J_h^{\mathcal{D}}(u_h, v_h) := &\sum_{F \in \mathscr{E}_h^{i,b}} [\mu_F \langle [\![\nabla_{\mathbf{T}} u_h]\!], [\![\nabla_{\mathbf{T}} v_h]\!] \rangle_F + \eta_F \langle [\![u_h]\!], [\![v_h]\!] \rangle_F] \\
&+ \sum_{F \in \mathscr{E}_h^i} \mu_F \langle [\![\nabla u_h \cdot n_F]\!], [\![\nabla v_h \cdot n_F]\!] \rangle_F.
\end{aligned}
\tag{5.4.2}
$$

For each $\theta \in [0, 1]$, we define the bilinear form $B^{\mathcal{D}}_{h,\theta} : V_{h,p} \times V_{h,p} \to \mathbb{R}$ by

$$
B^{\mathcal{D}}_{h,\theta}(u_h, v_h) := \theta B^{\mathcal{D}}_{h,*}(u_h, v_h) + (1 - \theta) \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K + J_h^{\mathcal{D}}(u_h, v_h). \tag{5.4.3}
$$

Finally, the bilinear form $A_h^{\mathcal{D}} : V_{h,p} \times V_{h,p} \to \mathbb{R}$ is defined by

$$
A_h^{\mathcal{D}}(u_h, v_h) := \sum_{K \in \mathscr{T}_h} \langle \gamma A : D^2 u_h, \Delta v_h \rangle_K + B^{\mathcal{D}}_{h,1/2}(u_h, v_h) - \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K. \tag{5.4.4}
$$

The scheme for approximating the solution of (5.1.1) is to find $u_h \in V_{h,p}$ such that

$$
A_h^{\mathcal{D}}(u_h, v_h) = \sum_{K \in \mathscr{T}_h} \langle \gamma f, \Delta v_h \rangle_K \quad \forall w_h \in V_{h,p}. \tag{5.4.5}
$$

Remark 5.4.1 *In (5.4.1)–(5.4.4) we have defined the bilinear forms $B^{\mathcal{D}}_{h,*}, J_h^{\mathcal{D}}, B^{\mathcal{D}}_{h,\theta}$, $A_h^{\mathcal{D}} : V_{h,p} \times V_{h,p} \to \mathbb{R}$. The main difference between these bilinear forms and the bilinear forms given presented in Section 3 of [110] is in the bilinear form $B^{\mathcal{D}}_{h,*}$ (and*

thus, by definition, in $B^{\mathcal{D}}_{h,\theta}$ and $A^{\mathcal{D}}_h$). In particular, the bilinear form $B^{\mathcal{D}}_{h,*}$ (given by (5.4.1)) contains the following additional terms:

$$\sum_{F\in\mathscr{E}_h^i} \langle\langle\nabla_{\mathbf{T}} w\rangle\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} v_h]\!]\rangle_F + \langle\langle\nabla_{\mathbf{T}} v_h\rangle\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} w]\!]\rangle_F$$
$$+ \sum_{F\in\mathscr{E}_h^i} \langle \mathcal{H}_F \langle\langle\nabla v_h \cdot n_F\rangle\rangle, [\![\nabla w_h \cdot n_F]\!]\rangle_F + \sum_{F\in\mathscr{E}_h^{i,b}} \langle \mathcal{H}_F \langle\langle\nabla w \cdot n_F\rangle\rangle, [\![\nabla v_h \cdot n_F]\!]\rangle_F$$

which arise due to the piecewise curvature of $\partial\Omega$. Indeed, if Ω is polytopal (which is a necessary assumption of Lemmas 5, 7, and 8, as well as Theorems 8 and 9 in [110]), then all of the faces $F \in \mathscr{E}_h^b$ are flat, and so

$$\mathcal{H}_F = 0 \quad \textit{and} \quad \nabla_{\mathbf{T}} n_F = 0, \quad \textit{for all } F \in \mathscr{E}_h^b,$$

which means that the additional terms vanish, and we retain the scheme introduced in [110]. In experiment 5.9.4, the results imply the necessity of these extra terms when $\partial\Omega$ is curved. Furthermore, the presence of these additional terms requires the application of new techniques, in order to prove that the numerical method is consistent (see Lemma 5.5.1), admits a unique solution (see Theorem 5.6.3), and that the resulting solution satisfies optimal error bounds (see Theorem 5.7.1).

5.5 Consistency of the method

We will now provide a consistency result for our method. This method is central to the error analysis discussed in Section 5.7, as it allows for a "Galerkin orthogonality" type argument.

Lemma 5.5.1 *Let Ω be a piecewise C^3 and piecewise convex domain, and let $\{\mathscr{T}_h\}_h$ be a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Let $w \in H^s(\Omega;\mathscr{T}_h) \cap H^2(\Omega) \cap H_0^1(\Omega)$, $s > 5/2$. Then, for every $v_h \in V_{h,p}$, we have the identities*

$$B^{\mathcal{D}}_{h,*}(w, v_h) = \sum_{K\in\mathscr{T}_h} \langle \Delta w, \Delta v_h\rangle_K \quad \textit{and} \quad J_h^{\mathcal{D}}(w, v_h) = 0. \tag{5.5.1}$$

Proof: Assume that $w \in H^s(\Omega;\mathscr{T}_h) \cap H^2(\Omega) \cap H_0^1(\Omega)$, $s > 5/2$, and $v_h \in V_{h,p}$. The identities in (5.5.1) are given by (5.3.10) and (5.3.17). □

The following corollary shows that the method is consistent, that is, if the true solution, u, of (5.1.1) is sufficiently piecewise smooth then u also satisfies (5.4.5).

Corollary 5.5.2 *Let Ω be a piecewise C^3 and piecewise convex domain, and let $\{\mathscr{T}_h\}_h$ be a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Assume that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfies (5.1.1). If $u \in H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$, then u satisfies*

$$A_h^{\mathcal{D}}(u_h, v_h) = \sum_{K \in \mathscr{T}_h} \langle \gamma f, \Delta v_h \rangle_K \quad \forall w_h \in V_{h,p}. \tag{5.5.2}$$

Proof: This follows simply by noting that u satisfies

$$\begin{cases} \gamma L u = \gamma f, & \text{a.e in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases}$$

as well as the regularity assumptions necessary for Lemma 5.5.1 to hold. □

Remark 5.5.3 (Conforming regularity) *The hypotheses of Corollary 5.5.2 assume that the strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of (5.1.1) also belongs to $H^s(\Omega; \mathscr{T}_h)$, for some $s > 5/2$. However, the coefficient matrix, A, describing the PDE (5.1.1), belongs to $L^\infty(\Omega)$, and so in general, would not expect such broken regularity. There are of course cases (for instance the Poisson problem), where such regularity would be expected (particularly since our formulation allows for C^∞ domains, for example). This assumption of higher broken regularity is utilised in Section 5.7 (see Theorem 5.7.1), to prove optimal error estimates, where one is required to substitute the true solution into the first argument of the bilinear form $A_h^{\mathcal{D}}$. Where such regularity is not available, we prove an error estimate (see Theorem 5.7.2) that only requires the expected H^2-regularity of the strong solution, by asserting that the finite element interpolant also belongs to $H_0^1(\Omega) \cap H^2(\Omega)$, which implies that the method of this chapter is as least as accurate as a fully conforming method (however we do not have to enforce H^2-conformity directly into the finite element space, which can be computationally expensive).*

5.6 Stability of the method

Let c_* be a positive constant independent of h and to be determined later. For each $\theta \in (0, 1]$ define the functional $\|\cdot\|_{h,\theta} : V_{h,p} \to \mathbb{R}^+$ by

$$\|v_h\|_{h,\theta}^2 := \sum_{K \in \mathscr{T}_h} [\theta |v_h|_{H^2(K)}^2 + (1-\theta)\|\Delta v_h\|_{L^2(K)}^2] + \frac{\theta}{2} \sum_{F \in \mathscr{E}_h^b} \left\| \mathcal{H}_F^{1/2} \frac{\partial v_h}{\partial n_F} \right\|_{L^2(F)}^2 + c_* J_h^{\mathcal{D}}(v_h, v_h). \tag{5.6.1}$$

Lemma 5.6.1 *Assume that $\{\mathscr{T}_h\}_h$ is a regular of order 2 family of triangulations on $\overline{\Omega}$. Then, for any $\theta \in (0,1]$, $\|\cdot\|_{h,\theta}$ is a norm on $V_{h,p}$.*

Proof: Homogeneity and the triangle inequality are clear. It remains to show that if $\|v_h\|_{h,\theta} = 0$, then $v_h = 0$ for $v_h \in V_{h,p}$. Let $v_h \in V_{h,p}$ satisfy $\|v_h\|_{h,\theta} = 0$ for some $\theta \in (0,1]$. Since $\theta \in (0,1]$, it follows that $|v_h|_{H^2(\Omega;\mathscr{T}_h)} = 0$, and thus v_h must be piecewise affine. Furthermore, $J_h^{\mathcal{D}}(v_h, v_h) = 0$ implies that $[\![\nabla v_h]\!] = 0$ for all $F \in \mathscr{E}_h^i$, and $[\![v_h]\!] = 0$ for all $F \in \mathscr{E}_h^{i,b}$. It follows that v_h is an affine function that satisfies $v_h|_{\partial\Omega} = 0$, and so $v_h \equiv 0$. $\square$

Lemma 5.6.2 *Let Ω be a piecewise C^3 and piecewise convex domain, and let $\{\mathscr{T}_h\}_h$ be a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Then, for each constant $\kappa > 1$, there exists a positive constant c_{stab}, independent of h, p, and θ, such that for any $v_h \in V_{h,p}$ and any $\theta \in (0,1]$, we have*

$$\begin{aligned}\kappa B_{h,\theta}^{\mathcal{D}}(v_h, v_h) \geq \theta |v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + (1-\theta)\sum_{K\in\mathscr{T}_h} \|\Delta v_h\|^2_{L^2(K)} \\ + \frac{\theta}{2}\sum_{F\in\mathscr{E}_h^b} \left\| \mathcal{H}_F^{1/2} \frac{\partial v_h}{\partial n_F}\right\|^2_{L^2(F)} + c_* J_h^{\mathcal{D}}(v_h, v_h),\end{aligned} \tag{5.6.2}$$

where, for some fixed constant $\sigma \geq 1$, the jump penalty parameters μ_F and η_F satisfy

$$\mu_F = \frac{\sigma c_{\text{stab}}}{\tilde{h}_F} \quad \textit{and} \quad \eta_F = \frac{\sigma c_{\text{stab}}}{\tilde{h}_F^3}. \tag{5.6.3}$$

Proof: The proof is similar to that of [111], Section 6, Lemma 6; in this case we must now deal with the extra terms arising in the bilinear form $B_{h,*}^{\mathcal{D}}$ due to the curvature of the boundary, $\partial\Omega$, and the resulting curvature of internal faces $F \in \mathscr{E}_h^i$ that are contained in the boundary of an element $K \in \mathscr{T}_h$ for which F_K is a nonaffine map.

Firstly, for $v_h \in V_{h,p}$, we have

$$B_{h,\theta}^{\mathcal{D}}(v_h, v_h) = \theta |v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + (1-\theta)\sum_{K\in\mathscr{T}_h} \|\Delta v_h\|^2_{L^2(K)} + J_h^{\mathcal{D}}(v_h, v_h) + \theta \sum_{i=1}^{5} I_i,$$

where

$$I_1 := 2\sum_{F\in\mathscr{E}_h^i} \langle \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \langle\!\langle v_h \rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!]\rangle_F,\ I_2 := -2\sum_{F\in\mathscr{E}_h^{i,b}} \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla v_h \cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}} v_h]\!]\rangle_F,$$

$$I_3 := 2\sum_{F\in\mathscr{E}_h^i} \left\langle \langle\!\langle \nabla_{\mathbf{T}} v_h \rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} v_h]\!] \right\rangle_F,\ I_4 := \sum_{F\in\mathscr{E}_h^i} \langle \mathcal{H}_F \langle\!\langle \nabla v_h \cdot n_F\rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!]\rangle_F,$$

$$I_5 := \sum_{F\in\mathscr{E}_h^b} \langle \mathcal{H}_F(\nabla v_h \cdot n_F), (\nabla v_h \cdot n_F)\rangle_F.$$

We first see that for any $\delta > 0$,

$$\begin{aligned}
|I_1| &= 2\left|\sum_{F\in\mathscr{E}_h^i} \langle \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \langle\!\langle v_h \rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!] \rangle_F\right| \\
&\leq \sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \| \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \langle\!\langle v_h \rangle\!\rangle\|^2_{L^2(F)} + (\tilde{h}_F\delta)^{-1} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)} \\
&\leq \sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h : F\subset\partial K} \| \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} v_h\|^2_{L^2(\partial K)} \right) + \sum_{F\in\mathscr{E}_h^i} (\tilde{h}_F\delta)^{-1} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}.
\end{aligned}$$

Furthermore, applying (4.10.23), followed by the trace estimate (4.6.1), we obtain

$$\begin{aligned}
&\sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h : F\subset\partial K} \| \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \tau_{\partial K}(v_h|_K)\|^2_{L^2(\partial K)} \right) \\
&\leq C \sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h^c : F\subset\partial K} \|D^2 v_h\|^2_{L^2(\partial K)} + \|\nabla v_h\|^2_{L^2(\partial K)} + \sum_{K\in\mathscr{T}_h^f : F\subset\partial K} \|D^2 v_h\|^2_{L^2(\partial K)} \right) \\
&\leq C \sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h^c : F\subset\partial K} h_K^{-1}\|D^2 v_h\|^2_{L^2(K)} + h_K\|D^2 v_h\|^2_{H^1(K)} + h_K^{-1}\|\nabla v_h\|^2_{L^2(K)} + \right. \\
&\qquad \left. + \sum_{K\in\mathscr{T}_h^f : F\subset\partial K} h_K^{-1}\|D^2 v_h\|^2_{L^2(K)} + h_K|D^2 v_h|^2_{H^1(K)} \right) \\
&\leq C \sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h^c : F\subset\partial K} h_K^{-1}|v_h|^2_{H^2_*(K)} + h_K|v_h|^2_{H^3_*(K)} \right. \\
&\qquad \left. + \sum_{K\in\mathscr{T}_h^f : F\subset\partial K} h_K^{-1}|v_h|^2_{H^2(K)} + h_K|D^2 v_h|^2_{H^1_*(K)} \right).
\end{aligned}$$

We now note that for $K \in \mathscr{T}_h^f$, the map F_K is affine, and so $D^2 v_h|_K$ is a piecewise polynomial, which means that we may apply the inverse estimate (4.6.26) with $m = 1, q = 2$, obtaining $|D^2 v_h|^2_{H^1_*(K)} \leq Ch_K^{-2}\|D^2 v_h\|^2_{W^{0,2}_*(K)} = Ch_K^{-2}|v|^2_{H^2(K)}$. However, for $K \in \mathscr{T}_h^c$, this is no longer the case, and so we must apply the same estimate, with $m = 3$, which gives us $|v_h|^2_{H^3_*(K)} \leq Ch_K^{-2}|v_h|^2_{H^2_*(K)}$. Applying this to the above

estimate, we obtain (noting that by definition $\tilde{h}_F \leq h_K$ if $F \subset \partial K$)

$$\begin{aligned}
&\sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \left(\sum_{K\in\mathscr{T}_h^c : F\subset\partial K} \|\operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} v_h\|^2_{L^2(\partial K)} + \sum_{K\in\mathscr{T}_h^f : F\subset\partial K} \|\operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} v_h\|^2_{L^2(\partial K)} \right) \\
&\leq C \sum_{F\in\mathscr{E}_h^i} \delta \left(\sum_{K\in\mathscr{T}_h^c : F\subset\partial K} \|v_h\|^2_{H^2(K)} + \sum_{K\in\mathscr{T}_h^f : F\subset\partial K} |v_h|^2_{H^2(K)} \right) \\
&\leq \delta C C(d) \left(\sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |v_h|^2_{H^2(K)} \right).
\end{aligned}$$

This gives us, for any $\delta > 0$,

$$|I_1| \leq \delta C C(d) \left(\sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |v_h|^2_{H^2(K)} \right) + \delta^{-1} \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}. \tag{5.6.4}$$

Analogously (utilising (4.10.22)) for I_2, we obtain, for any $\delta > 0$,

$$|I_2| \leq \delta C C(d) \left(\sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |v_h|^2_{H^2(K)} \right) + \delta^{-1} \sum_{F\in\mathscr{E}_h^{i,b}} \tilde{h}_F^{-1} \|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)}. \tag{5.6.5}$$

For I_3, due to (4.10.16), and the trace estimate (4.6.1) we see that

$$\begin{aligned}
|I_3| &= 2 \left| \sum_{F\in\mathscr{E}_h^i} \langle \langle\!\langle \nabla_{\mathbf{T}} v_h \rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T [\![\nabla_{\mathbf{T}} v_h]\!] \rangle_F \right| \\
&\leq C \left(\sum_{F\in\mathscr{E}_h^i} \delta\tilde{h}_F \|\langle\!\langle \nabla_{\mathbf{T}} v_h \rangle\!\rangle\|^2_{L^2(F)} + (\delta\tilde{h}_F)^{-1} \|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)} \right) \\
&\leq C \left(\sum_{F\in\mathscr{E}_h^i} \sum_{K\in\mathscr{T}_h^c : F\subset\partial K} \delta\tilde{h}_F \|\nabla_{\mathbf{T}} v_h\|^2_{L^2(\partial K)} + (\delta\tilde{h}_F)^{-1} \|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)} \right) \\
&\leq C \left(\sum_{F\in\mathscr{E}_h^i} \sum_{K\in\mathscr{T}_h^c : F\subset\partial K} \delta\tilde{h}_F h_K^{-1} \|\nabla v_h\|^2_{L^2(K)} + h_K |\nabla v_h|^2_{H^1(K)}) \right) \\
&\qquad + C \sum_{F\in\mathscr{E}_h^i} (\delta\tilde{h}_F)^{-1} \|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)} \\
&\leq \delta C C(d) \sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} + C \sum_{F\in\mathscr{E}_h^i} (\delta\tilde{h}_F)^{-1} \|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)}.
\end{aligned} \tag{5.6.6}$$

Similarly, for I_4, we obtain

$$|I_4| \leq \delta CC(d) \sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} + C\sum_{F\in\mathscr{E}_h^i} (\delta\tilde{h}_F)^{-1} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}. \tag{5.6.7}$$

An application of (4.7.4) yields

$$\begin{aligned}\sum_{K\in\mathscr{T}_h^c} |v_h|^2_{H^2_*(K)} &\leq |v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + |v|^2_{H^1(\Omega;\mathscr{T}_h)} \\ &\leq C\left(|v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^{i,b}} \tilde{h}_F^{-1}\|[\![v_h]\!]\|^2_{L^2(F)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1}\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)})\right).\end{aligned} \tag{5.6.8}$$

Applying (5.6.8) to (5.6.4)–(5.6.7), and summing the resulting estimates, we obtain (noting that $\nabla v_h|_F = \nabla_{\mathbf{T}} v_h + (\nabla v_h \cdot n_F)n_F$), for any $\delta > 0$,

$$\begin{aligned}\sum_{i=1}^4 |I_i| &\leq \delta CC(d)|v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + (\delta^{-1} + \delta CC(d)) \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1}\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)} \\ &+ (\delta^{-1} + \delta CC(d))\left(\sum_{F\in\mathscr{E}_h^{i,b}} \tilde{h}_F^{-1}(\|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)} + \|[\![v_h]\!]\|^2_{L^2(F)})\right).\end{aligned} \tag{5.6.9}$$

One can also see that

$$I_5 = \sum_{F\in\mathscr{E}_h^b} \left\| \mathcal{H}_F^{1/2} \frac{\partial v_h}{\partial n}\right\|^2_{L^2(F)},$$

where, due to (4.10.17), it follows that $\mathcal{H}_F \geq 0$ on each F, and so one can always make sense of $\mathcal{H}_F^{1/2}$. Now that we have bounds on $I_1, \ldots, I_5$, we obtain the following:

$$B^{\mathcal{D}}_{h,\theta}(v_h, v_h) \geq \sum_{i=1}^6 A_i,$$

where

$$\begin{aligned}A_1 &= \theta(1-\delta CC(d))|v_h|^2_{H^2(\Omega;\mathscr{T}_h)}, \quad A_2 = (1-\theta)\sum_{K\in\mathscr{T}_h} \|\Delta v_h\|^2_{L^2(K)}, \\ A_3 &= \sum_{F\in\mathscr{E}_h^i} \left(\mu_F - \theta(\delta^{-1} + \delta CC(d))\tilde{h}_F^{-1}\right)\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)}, \\ A_4 &= \sum_{F\in\mathscr{E}_h^{i,b}} \left(\mu_F - \theta\left(\left(\delta^{-1} + \delta CC(d)\right)\tilde{h}_F^{-1}\right)\right)\|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)}, \\ A_5 &= \frac{\theta}{2}\sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F^{1/2}\frac{\partial v_h}{\partial n_F}\right\|^2_{L^2(F)}, \\ A_6 &= \sum_{F\in\mathscr{E}_h^{i,b}} (\eta_F - \theta(\delta^{-1} + \delta CC(d))\tilde{h}_F^{-1}))\|v_h\|^2_{L^2(F)}.\end{aligned}$$

For any given $\kappa > 1$, there is a $\delta > 0$ such that

$$1 - \delta CC(d) > \kappa^{-1} \quad \text{and} \quad \delta CC(d) < \delta^{-1}.$$

Set $c_{\text{stab}} = 4/\delta$, $c_* = \kappa/2$ so that the following inequalities hold for any $\theta \in (0, 1]$:

$$A_1 \geq \theta\kappa^{-1}|v_h|^2_{H^2(\Omega;\mathscr{T}_h)},\ A_2 \geq (1-\theta)\kappa^{-1}\sum_{K\in\mathscr{T}_h}\|\Delta v_h\|^2_{L^2(K)},$$

$$A_3 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^i}\mu_F\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)} = \kappa^{-1}c_*\sum_{F\in\mathscr{E}_h^i}\mu_F\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)},$$

$$A_4 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^{i,b}}\mu_F\|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)} = \kappa^{-1}c_*\sum_{F\in\mathscr{E}_h^{i,b}}\mu_F\|[\![\nabla_{\mathbf{T}} v_h]\!]\|^2_{L^2(F)},$$

$$A_5 \geq \frac{\theta}{2\kappa}\sum_{F\in\mathscr{E}_h^b}\left\|\mathcal{H}_F^{1/2}\frac{\partial v_h}{\partial n_F}\right\|^2_{L^2(F)},\ A_6 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^b}\eta_F\|v_h\|^2_{L^2(F)} = \kappa^{-1}c_*\sum_{F\in\mathscr{E}_h^b}\eta_F\|v_h\|^2_{L^2(F)},$$

whenever μ_F and η_F satisfy (5.6.3). Hence,

$$\kappa B^{\mathcal{D}}_{h,\theta}(v_h, v_h) \geq \theta|v_h|^2_{H^2(\Omega;\mathscr{T}_h)} + (1-\theta)\sum_{K\in\mathscr{T}_h}\|\Delta v_h\|^2_{L^2(K)} + c_* J^{\mathcal{D}}_h(v_h, v_h) + \frac{\theta}{2}\sum_{F\in\mathscr{E}_h^b}\|\mathcal{H}_F^{1/2}\frac{\partial v_h}{\partial n}\|^2_{L^2(F)},$$

as has been asserted. □

Theorem 5.6.3 *Assume that $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\text{Sym}})$ satisfies (5.1.3) and (5.1.4). Under the hypotheses of Lemma 5.6.2, let c_{stab}, $c_{\mathcal{H}}$, η_F and μ_F be chosen so that Lemma 5.6.2 holds with $\kappa < (1-\varepsilon)^{-1/2}$. Then, for every $v_h \in V_{h,p}$, we have*

$$\|v_h\|^2_{h,1} \leq \frac{2\kappa}{1-\kappa(1-\varepsilon)}A^{\mathcal{D}}_h(v_h, v_h). \tag{5.6.10}$$

Therefore, there exists a unique solution $u_h \in V_{h,p}$ of the numerical scheme (5.4.5). Furthermore, we have the bound

$$\|u_h\|_{h,1} \leq \frac{2\kappa\sqrt{d}\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa^2(1-\varepsilon)}\|f\|_{L^2(\Omega)}. \tag{5.6.11}$$

Proof: The proof is the same as the proof of Theorem 8, in [110], Section 4, which relies upon the stability estimate (5.6.2), in order to prove that $A^{\mathcal{D}}_h$ is coercive on $V_{h,p}\times V_{h,p}$, yielding the existence and uniqueness of a numerical solution (note that this result is proven in a more general context in the proof of Theorem 7.4.1). □

5.7 Error estimates

Theorem 5.7.1 *Assume that $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\mathrm{Sym}})$ satisfies (5.1.3) and (5.1.4). Let Ω be a piecewise C^{m+1} and piecewise convex domain, $m \in \mathbb{N}$, $m \geq 2$, and let $\{\mathscr{T}_h\}_h$ be a regular of order m family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Furthermore, let $u \in H^2(\Omega) \cap H^1_0(\Omega)$ be the unique solution of (5.1.1), and assume that $u \in H^{\mathbf{s}}(\Omega; \mathscr{T}_h)$ with $s_K > 5/2$ for each $K \in \mathscr{T}_h$. Let c_{stab}, μ_F, and η_F be chosen as in Theorem 5.6.3 for all $F \in \mathscr{E}_h^{i,b}$. Then, there exists a positive constant C independent of h and u, but depending on $\max_K s_K$, such that for the unique solution u_h of (5.4.5), we have*

$$\|u - u_h\|^2_{h,1} \leq C \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)}, \tag{5.7.1}$$

where $t_K = \min(p+1, s_K, m+1)$ for each $K \in \mathscr{T}_h$.

Note that for the special case of quasi-uniform meshes, the a priori estimate (5.7.1) simplifies to

$$\|u - u_h\|_{h,1} \leq C h^{\min(p+1,s,m+1)-2} \|u\|_{H^s(\Omega;\mathscr{T}_h)}.$$

Therefore, the convergence rates are optimal with respect to the mesh size.

Proof: The proof is analogous to the proof of Theorem 9 in [110], Section 5. It is noteworthy that the proof relies on the existence of a $z_h \in V_{h,p}$ satisfying (4.6.9). Let us denote $\xi_h := z_h - u$ and $\psi_h := z_h - u_h$. Then, from the triangle inequality, we obtain

$$\|u - u_h\|_{h,1} \leq \|u - z_h\|_{h,1} + \|z_h - u_h\|_{h,1} = \|\xi_h\|_{h,1} + \|\psi_h\|_{h,1}. \tag{5.7.2}$$

Then, applying the coercivity estimate (5.6.10) yields (noting that $\psi_h \in V_{h,p}$)

$$\begin{aligned}
\|\psi_h\|^2_{h,1} &\lesssim A_h^{\mathcal{D}}(\psi_h, \psi_h) \\
&= A_h^{\mathcal{D}}(z_h - u_h, \psi_h) \\
&= A_h^{\mathcal{D}}(z_h, \psi_h) - A_h^{\mathcal{D}}(u_h, \psi_h) \\
&= A_h^{\mathcal{D}}(z_h, \psi_h) - \sum_{K\in\mathscr{T}_h} \langle \gamma f, \Delta\psi_h\rangle_K \\
&= A_h^{\mathcal{D}}(z_h, \psi_h) - \sum_{K\in\mathscr{T}_h} \langle \gamma A : D^2 u, \Delta\psi_h\rangle_K \\
&= A_h^{\mathcal{D}}(z_h, \psi_h) - A_h^{\mathcal{D}}(u, \psi_h) = A_h^{\mathcal{D}}(\xi_h, \psi_h).
\end{aligned}$$

Note that we have used the fact that u_h is the numerical solution, and the consistency result (5.5.1), which allows for a "Galerkin orthogonality" type argument.

We now proceed to show that

$$\|\psi_h\|_{h,1}^2 \lesssim A_h^{\mathcal{D}}(\xi_h,\psi_h) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2}\|\psi_h\|_{h,1},$$

which gives us

$$\|\psi_h\|_{h,1} \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2}.$$

We see that

$$A_h^{\mathcal{D}}(\xi_h,\psi_h) = \sum_{K\in\mathscr{T}_h}\left[\langle\gamma A : D^2\xi_h, \Delta\psi_h\rangle_K + \frac{1}{2}\langle\Delta\xi_h,\Delta\psi_h\rangle_K\right] + \frac{1}{2}B_{h,*}^{\mathcal{D}}(\xi_h,\psi_h) + J_h^{\mathcal{D}}(\xi_h,\psi_h).$$

Utilising the first estimate of (5.7.2), it is clear that

$$\begin{aligned}\sum_{K\in\mathscr{T}_h}\left[\langle\gamma A : D^2\xi_h, \Delta\psi_h\rangle_K + \frac{1}{2}\langle\Delta\xi_h,\Delta\psi_h\rangle_K\right] &\lesssim |\xi_h|_{H^2(\Omega;\mathscr{T}_h)}|\psi_h|_{H^2(\Omega;\mathscr{T}_h)}\\ &\lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2}\|\psi_h\|_{h,1}.\end{aligned} \tag{5.7.3}$$

Furthermore,

$$J_h^{\mathcal{D}}(\xi_h,\psi_h) \lesssim J_h^{\mathcal{D}}(\xi_h,\xi_h)^{1/2}J_h^{\mathcal{D}}(\psi_h,\psi_h)^{1/2} \lesssim J_h^{\mathcal{D}}(\xi_h,\xi_h)^{1/2}\|\psi_h\|_{h,1},$$

where, by our assumptions on μ_F and η_F, and applying the second estimate of (4.6.9)

$$\begin{aligned}J_h^{\mathcal{D}}(\xi_h,\xi_h) &\lesssim \sum_{F\in\mathscr{E}_h^{i,b}}\sum_{K\in\mathscr{T}_h:F\subset\partial K}\tilde{h}_F^{-1}\|\nabla\xi_h\|_{L^2(\partial K)}^2 + \tilde{h}_F^{-3}\|\xi_h\|_{L^2(\partial K)}^2\\ &\lesssim \sum_{F\in\mathscr{E}_h^{i,b}}\sum_{K\in\mathscr{T}_h:F\subset\partial K}\tilde{h}_F^{-1}h_K^{2t_K-3}\|u\|_{H^{s_K}(K)}^2 + \tilde{h}_F^{-3}h_K^{2t_K-1}\|u\|_{H^{s_K}(K)}^2\\ &\lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2,\end{aligned} \tag{5.7.4}$$

and so,

$$J_h^{\mathcal{D}}(\xi_h,\psi_h) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2}\|\psi_h\|_{h,1}. \tag{5.7.5}$$

We obtain an estimate for $B^{\mathcal{D}}_{h,*}(\xi_h,\psi_h)$ using techniques similar to those utilised in the proof of Theorem 5.6.2. We see that

$$\begin{aligned} B^{\mathcal{D}}_{h,*}(\xi_h,\psi_h) &\leq |\xi_h|_{H^2(\Omega;\mathscr{T}_h)}|\psi_h|_{H^2(\Omega;\mathscr{T}_h)} + \sum_{i=1}^{8} A_i \\ &\lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{1/2} \|\psi_h\|_{h,1} + \sum_{i=1}^{8} A_i, \end{aligned} \tag{5.7.6}$$

where we define the terms $A_1,\ldots,A_8$ as follows:

$$A_1 := \sum_{F\in\mathscr{E}_h^i} \langle \mathcal{H}_F\langle\!\langle \nabla\xi_h\cdot n_F\rangle\!\rangle, [\![\nabla\psi_h\cdot n_F]\!]\rangle_F\,,\ A_2 := \sum_{F\in\mathscr{E}_h^{i,b}} \langle \mathcal{H}_F[\![\nabla\psi_h\cdot n_F]\!], \langle\!\langle\nabla\xi_h\cdot n_F\rangle\!\rangle\rangle_F$$

$$A_3 := \sum_{F\in\mathscr{E}_h^i} \langle \langle\!\langle \nabla_{\mathbf{T}}\xi_h\rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T\,[\![\nabla_{\mathbf{T}}\psi_h]\!]\rangle_F,\ A_4 := \sum_{F\in\mathscr{E}_h^i} \langle \langle\!\langle \nabla_{\mathbf{T}}\psi_h\rangle\!\rangle^T \nabla_{\mathbf{T}} n_F^T\,[\![\nabla_{\mathbf{T}}\xi_h]\!]\rangle_F$$

$$A_5 := \sum_{F\in\mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle\xi_h\rangle\!\rangle, [\![\nabla\psi_h\cdot n_F]\!]\rangle_F,\ A_6 := \sum_{F\in\mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle\psi_h\rangle\!\rangle, [\![\nabla\xi_h\cdot n_F]\!]\rangle_F]$$

$$A_7 := -\sum_{F\in\mathscr{E}_h^{i,b}} \langle \nabla_{\mathbf{T}}\langle\!\langle \nabla\xi_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}\psi_h]\!]\rangle_F,\ A_8 = -\sum_{F\in\mathscr{E}_h^{i,b}} \langle \nabla_{\mathbf{T}}\langle\!\langle \nabla\psi_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}\xi_h]\!]\rangle_F].$$

First we estimate A_1+A_3, due to the Cauchy–Schwarz inequality, and (4.10.21), we see that

$$\begin{aligned} A_1+A_3 &\lesssim \left(\sum_{F\in\mathscr{E}_h^i} \|\langle\!\langle\nabla\xi_h\rangle\!\rangle\|^2_{L^2(F)}\right)^{\frac12} J_h(\psi_h,\psi_h)^{\frac12} \\ &\lesssim \left(\sum_{K\in\mathscr{T}_h} \|\nabla\xi_h\|^2_{L^2(\partial K)}\right)^{\frac12} J_h(\psi_h,\psi_h)^{\frac12} \\ &\lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-3}\|u\|^2_{H^{s_K}(K)}\right)^{\frac12} \|\psi_h\|_{h,1} \leq \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{\frac12} \|\psi_h\|_{h,1}. \end{aligned}$$

Furthermore, due to the Cauchy–Schwarz inequality, (4.10.21), the trace estimate (4.6.1), the inverse estimate (4.6.26), and estimate (4.7.4), for A_2, we obtain

$$\begin{aligned} A_2 &\lesssim \left(\sum_{K\in\mathscr{T}_h} \tilde{h}_F^{-1}\|\nabla\xi_h\|^2_{L^2(\partial K)}\right)^{\frac12} \left(\sum_{K\in\mathscr{T}_h} \tilde{h}_F\|\nabla\psi_h\|^2_{L^2(\partial K)}\right)^{\frac12} \\ &\quad + \left(\sum_{F\in\mathscr{E}_h^b} \|\nabla\xi_h\|^2_{L^2(F)}\right)^{\frac12} \left(\sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F\frac{\partial\psi_h}{\partial n_F}\right\|^2_{L^2(F)}\right)^{\frac12} \\ &\lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{\frac12} \left(\left(\sum_{K\in\mathscr{T}_h} |\psi_h|^2_{H^2_*(K)}\right)^{\frac12} + \left(\sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F\frac{\partial\psi_h}{\partial n_F}\right\|^2_{L^2(F)}\right)^{\frac12}\right) \end{aligned}$$

$$\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1}.$$

Similarly, due to the Cauchy–Schwarz inequality, (4.10.21), the trace estimate (4.6.1), the inverse estimate (4.6.26), and estimate (4.7.4), for A_4, we obtain

$$\begin{aligned} A_4 &\lesssim \left(\sum_{K \in \mathscr{T}_h} \tilde{h}_F^{-1} \|\nabla \xi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathscr{T}_h} \tilde{h}_F \|\nabla \psi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \left(\sum_{K \in \mathscr{T}_h} |\psi_h|^2_{H^2_*(K)} \right)^{\frac{1}{2}} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1}. \end{aligned}$$

For A_5, we have

$$\begin{aligned} A_2 &\lesssim \left(\sum_{F \in \mathscr{E}_h^i} \sum_{K \in \mathscr{T}_h : F \subset \partial K} \tilde{h}_F \|D^2 \xi_h\|^2_{L^2(\partial K)} + \|\nabla \xi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1} \end{aligned}$$

Indeed, for A_6, by (5.7.4), (4.6.1), (4.6.26) and (4.7.4) we see that

$$\begin{aligned} A_6 &\lesssim J_h^{\mathcal{D}}(\xi_h, \xi_h)^{\frac{1}{2}} \left(\sum_{F \in \mathscr{E}_h^i} \sum_{K \in \mathscr{T}_h : F \subset \partial K} \tilde{h}_F \|D^2 \psi_h\|^2_{L^2(\partial K)} + \|\nabla \psi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \\ &\lesssim \left[\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right]^{\frac{1}{2}} \left[\sum_{K \in \mathscr{T}_h} |\psi_h|^2_{H^2_*(K)} \right]^{\frac{1}{2}} \lesssim \left[\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right]^{\frac{1}{2}} \|\psi_h\|_{h,1}. \end{aligned}$$

Similar to the estimate for A_5, we obtain

$$\begin{aligned} A_7 &\lesssim \left(\sum_{F \in \mathscr{E}_h^{i,b}} \sum_{K \in \mathscr{T}_h : F \subset \partial K} \tilde{h}_F \|D^2 \xi_h\|^2_{L^2(\partial K)} + \|\nabla \xi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1} \end{aligned}$$

Again, applying (5.7.4), (4.6.1), (4.6.26) and (4.7.4) we obtain

$$A_8 \lesssim J_h^{\mathcal{D}}(\xi_h,\xi_h)^{\frac{1}{2}} \left(\sum_{F\in\mathscr{E}_h^{i,b}} \sum_{K\in\mathscr{T}_h : F\subset\partial K} \tilde{h}_F \|D^2\psi_h\|^2_{L^2(\partial K)} + \|\nabla\psi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}}$$

$$\lesssim \left[\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right]^{\frac{1}{2}} \left[\sum_{K\in\mathscr{T}_h} |\psi_h|^2_{H^2_*(K)} \right]^{\frac{1}{2}} \lesssim \left[\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right]^{\frac{1}{2}} \|\psi_h\|_{h,1}.$$

Thus, applying our estimates for $A_1,\ldots,A_8$ to (5.7.6) we obtain the following:

$$B^{\mathcal{D}}_{h,*}(\xi_h,\psi_h) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2} \|\psi_h\|_{h,1}. \tag{5.7.7}$$

From (5.7.3), (5.7.5), and (5.7.7), it follows that

$$A^{\mathcal{D}}_h(\xi_h,\psi_h) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|\psi_h\|_{h,1},$$

and so,

$$\|\psi_h\|_{h,1} \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}}.$$

Furthermore, applying both estimates of (4.6.9), in conjunction with (5.7.4), we obtain

$$\|\xi_h\|^2_{h,1} \lesssim |\xi_h|^2_{H^2(\Omega;\mathscr{T}_h)} + \frac{1}{2} \sum_{F\in\mathscr{E}_h^b} \left\| \mathcal{H}_F^{1/2} \frac{\partial \xi_h}{\partial n_F} \right\|^2_{L^2(F)} + J_h^{\mathcal{D}}(\xi_h,\xi_h) \lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)},$$

and so

$$\|u-u_h\|_{h,1} \leq \|\xi_h\|_{h,1} + \|\psi_h\|_{h,1} \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2},$$

as desired. □

We now provide an error estimate that only assumes that the strong solution of (5.1.1) belongs to $H^2(\Omega)\cap H^1_0(\Omega)$, i.e., no higher broken Sobolev regularity is assumed. The proof is similar to the proof of Proposition 10 in [110].

Theorem 5.7.2 *Assume that $A \in L^\infty(\Omega;\mathbb{R}^{d\times d}_{\mathrm{Sym}})$ satisfies (5.1.3) and (5.1.4). Let Ω be a piecewise C^3 and piecewise convex domain, and let $\{\mathscr{T}_h\}_h$ be a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Furthermore, let $u \in H^2(\Omega)\cap H^1_0(\Omega)$ be the unique solution of (5.1.1). Let c_{stab}, μ_F, and η_F be*

chosen as in Theorem 5.6.3 for all $F \in \mathscr{E}_h^{i,b}$. *Then, there exists a positive constant* C *independent of* h *and* u, *such that for the unique solution* u_h *of (5.4.5), we have*

$$\|u - u_h\|_{h,1}^2 \leq C \inf \left\{ \|u - z_h\|_{H^2(\Omega)} : z_h \in V_{h,p} \cap H^2(\Omega) \cap H_0^1(\Omega) \right\}. \tag{5.7.8}$$

Proof: Let $z_h \in V_{h,p} \cap H^2(\Omega) \cap H_0^1(\Omega)$ be arbitrary, and set $\xi_h := z_h - u$, $\psi_h = z_h - u_h$. Then, by the triangle inequality, we see that

$$\|u - u_h\|_{h,1} \leq \|\xi_h\|_{h,1} + \|\psi_h\|_{h,1}. \tag{5.7.9}$$

Furthermore, by (5.6.10), along with the fact that u_h satisfies (5.4.5), and that $z_h \in V_{h,p} \cap H^2(\Omega) \cap H_0^1(\Omega)$ satisfies (5.5.1) we see that

$$\begin{aligned} \|\psi_h\|_{h,1}^2 &\lesssim A_h^{\mathcal{D}}(\psi_h, \psi_h) \\ &= A_h^{\mathcal{D}}(z_h, \psi_h) - A_h^{\mathcal{D}}(u_h, \psi_h) \\ &= \sum_{K \in \mathscr{T}_h} \langle \gamma (A : D^2 z_h - f), \Delta \psi_h \rangle_K. \end{aligned}$$

We also have that $A : D^2 u = f$ a.e. in Ω, and so

$$\begin{aligned} \|\psi_h\|_{h,1}^2 &\lesssim \sum_{K \in \mathscr{T}_h} \langle \gamma A : D^2 (z_h - u), \Delta \psi_h \rangle_K \\ &\lesssim |u - z_h|_{H^2(\Omega)} \left(\sum_{K \in \mathscr{T}_h} \|\Delta \psi_h\|_{L^2(K)}^2 \right)^{\frac{1}{2}} \\ &\lesssim |u - z_h|_{H^2(\Omega)} \|\psi_h\|_{h,1}, \end{aligned}$$

which yields

$$\|\psi_h\|_{h,1} \lesssim |u - z_h|_{H^2(\Omega)}. \tag{5.7.10}$$

Then, by (4.10.21), we also see that

$$\begin{aligned} \|\xi_h\|_{h,1}^2 &= \|u - z_h\|_{h,1}^2 \\ &= |u - z_h|_{H^2(\Omega)}^2 + \frac{1}{2} \sum_{F \in \mathscr{E}_h^b} \left\| \mathcal{H}_F^{1/2} \frac{\partial}{\partial n_F} (u - z_h) \right\|_{L^2(F)}^2 \\ &\lesssim |u - z_h|_{H^2(\Omega)}^2 + \|\nabla (u - z_h)\|_{L^2(\partial\Omega)}^2 \\ &\lesssim \|u - z_h\|_{H^2(\Omega)}^2, \end{aligned}$$

where the last inequality follows due to the fact that the trace operator is continuous from $H^1(\Omega) \to L^2(\partial\Omega)$. Thus, we have that

$$\|\xi_h\|_{h,1} \lesssim \|u - z_h\|_{H^2(\Omega)}. \tag{5.7.11}$$

Applying (5.7.10) and (5.7.11) to (5.7.9), we obtain

$$\|u - u\|_{h,1} \lesssim \|u - z_h\|_{H^2(\Omega)}. \tag{5.7.12}$$

Notice that our choice of $z_h \in V_{h,p} \cap H^2(\Omega) \cap H_0^1(\Omega)$ was arbitrary, and so we may take an infimum over all such z_h in (5.7.12), which yields (5.7.8). $\square$

5.8 Implementation

Software and code: The experiments in this Chapter have been implemented in the most recent version of the Firedrake software [105, 87] (as of 3rd July 2018), which interfaces directly with PETSc [6, 7] running through a Python interface [39, 63]. A working Firedrake script, Curved-Dirichlet-DGFEM.py, used to generate the experiments of this Chapter is available in the Github repository:
https://github.com/ekawecki/FiredrakeNDV.

Linear systems and condition numbers: The bilinear form $A_h^{\mathcal{D}}$ defined by (5.4.4) can be considered to be similar to those present in finite element methods for fourth-order elliptic boundary-value problems (see [114, 25] for example), in the sense that the evaluation of $A_h^{\mathcal{D}}(u_h, v_h)$ for $u_h, v_h \in V_{h,p}$ involves the integration of products of second order partial derivatives. This typically leads to the matrix $A^{\mathcal{D}}$, describing the linear system given by (5.4.5), to have a Euclidean norm condition number of order h^{-4}. This can pose difficulties when applying iterative methods to solve the linear system, and thus to ensure that we solve the linear system with sufficiently high accuracy as the mesh size h decreases, we apply the Iterative refinement algorithm, i.e., Algorithm 1.1 of [32]. We implement the Iterative refinement algorithm by using the following choices in the Firedrake "solve" function.

```
t = time()
solve(A_gamma == L, U,
      solver_parameters = {
      "snes_type": "newtonls",
      "ksp_type": "preonly",
      "pc_type": "lu",
      "snes_monitor": False,
      "snes_rtol": 1e-16,
      "snes_atol": 1e-25})
tt.append(time()-t)
```

One can also see that when executing the script in Firedrake, we record the runtimes by way of the first and last line above, so that we only record the time that it takes to solve the linear system.

Two-dimensional curved boundary approximation: When implementing curved finite elements, we use a piecewise quadratic polynomial mapping to obtain a higher order approximation of the domain boundary. This is implemented in Firedrake by first using Gmsh [56] (version 3.0.1) to generate an affine triangulation Ω_h that approximates Ω (our assumption of piecewise convexity ensures that there is such a triangulation that is also a subset of Ω). We then define the *continuous* Lagrange finite element space $\mathbb{V} := \{v \in C(\overline{\Omega_h}; \mathbb{R}^2) : v \in \mathbb{P}^2(K; \mathbb{R}^2) \, \forall K \in \Omega_h\}$. Then, we take $\psi_i : \omega_i \to \mathbb{R}^2$, $\omega_i \subset \mathbb{R}$, $i = 1, \ldots, n$, to be the collection of charts that locally describe $\partial\Omega$, and denote $\{x_j\}_{j=1}^N$ to be the degrees of freedom of $\mathbb{V}$. We partition the collection of degrees of freedom by defining $J_{\text{ext}} = \{j \in \{1, \ldots, N\} : x_j \in \partial\Omega_h\}$, and $J_{\text{int}} = \{1, \ldots, N\} \setminus J_{\text{ext}}$, and so $\{x_j\}_{j=1}^N = \{x_j\}_{j \in J_{\text{int}}} \cup \{x_j\}_{j \in J_{\text{ext}}}$. We then define the the function $T \in \mathbb{V}$ by

$$\begin{cases} T(x_j) = x_j, & j \in J_{\text{int}}, \\ T(x_j) = \psi_i(x_j), & j \in J_{\text{ext}}, \quad i \in \{1, \ldots, n\} \text{ such that } x_j \in \omega_i. \end{cases} \tag{5.8.1}$$

Finally, we define our computational finite element space $V_{h,p}^{\text{comp}} := \{v \in L^2(\Omega) : v \circ T^{-1} \in \mathbb{P}^p(\hat{K})\}$. This procedure is implemented in Firedrake, in the code snippet below, utilising the Firedrake "Mesh" function. In this case Ω is the unit disk, and so there is only one chart, $\psi := x/|x|$. Furthermore, when we refine the mesh in our experiments, the meshes at each refinement level are not related to one another (the one exception being Experiment 6.11.2 of Chapter 6). That is, there is no hierarchical mesh structure, i.e., at each refinement level, we "remesh". A collection of the meshes used for the computations of this thesis can be found in the folder "Meshes" in the Github repository: https://github.com/ekawecki/FiredrakeNDV.

```
#Affine mesh of the unit disk, generated in Gmsh
mesh = Mesh("quasiunifrefdisk.msh")
# Implementing quadratic domain approximation
V = FunctionSpace(mesh, "CG", 2)
# Defining a function that identifies the curved portion of the boundary
bdry_indicator = Function(V)
bc = DirichletBC(V, Constant(1.0), 1)
bc.apply(bdry_indicator)
# Defining the continuous, piecewise quadratic vector-valued finite element↩
    space
VV = VectorFunctionSpace(mesh, "CG", 2)
T = Function(VV)
T.interpolate(SpatialCoordinate(mesh))
# Defining the function T given by (5.8.1)
T.interpolate(conditional(abs(1-bdry_indicator) < 1e-5, T/sqrt(inner(T,T)),↩
    T))
# Defining the curved mesh
mesh = Mesh(T)
# Defining the space V_{h,p}^{comp}
FES = FunctionSpace(mesh,"DG",deg)
```

Remark 5.8.1 (Computational parameters) *In the following experiments, we employ the following parameter choices:* $c_{\text{stab}} = 2$, $\mu_F = c_{\text{stab}}(p-1)^2/2\tilde{h}_F$, $\eta_F = 3c_{\text{stab}}(p-1)^4/8\tilde{h}_F^3$. *The order of the computational parameters with respect to* $\tilde{h}_F$ *was guided by (5.6.3) in the statement of Lemma 5.6.2, and the order of the computational parameter choices with respect to* p *was guided by the parameter choices employed in the experiments in Section 6 of [110]. Finally, the choice of* c_{stab} *was obtained experimentally.*

5.9 Experiments

In this chapter, we test the robustness of the scheme (5.4.5), with the computational domain Ω taken to be the unit disk, and the "key-hole" shaped domain (2.5.2), and consider various elliptic operators, L, that satisfy the Cordes condition (5.1.4). In each case, we see that the convergence rates are of the expected order in the $\|\cdot\|_{h,1}$–norm, for which we have proven the error bound (5.7.1).

5.9.1 Experiment 1

In this experiment, we consider the following problem

$$\begin{cases} \Delta u = f, & \text{in} \quad \Omega, \\ u = 0, & \text{on} \quad \partial\Omega, \end{cases} \tag{5.9.1}$$

where $\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$. For this problem, f is chosen so that the solution of (5.9.1) is given by

$$u(x_1, x_2) = \frac{1}{4}\sin(\pi(x_1^2 + x_2^2)).$$

We can also directly calculate the renormalisation parameter, γ, and provide the largest value of ε for which the Cordes condition (5.1.4) holds. In particular, we have that

$$\gamma := \frac{\text{Tr}(A)}{|A|^2} = \frac{\text{Tr}(I_d)}{|I_d|^2} = \frac{I_d : I_d}{I_d : I_d} = 1, \quad \text{and} \quad \varepsilon = 1.$$

Furthermore, since Ω is the unit disk, $\partial\Omega = \mathbb{S}^1$, and it follows that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore, $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. For the internal faces, the mean curvature is calculated directly as $\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F$, where n_F is a fixed choice of unit normal to F.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\text{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly,

we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u-u_h\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 5.1, and report the exact values in Table 5.1, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 5.2.

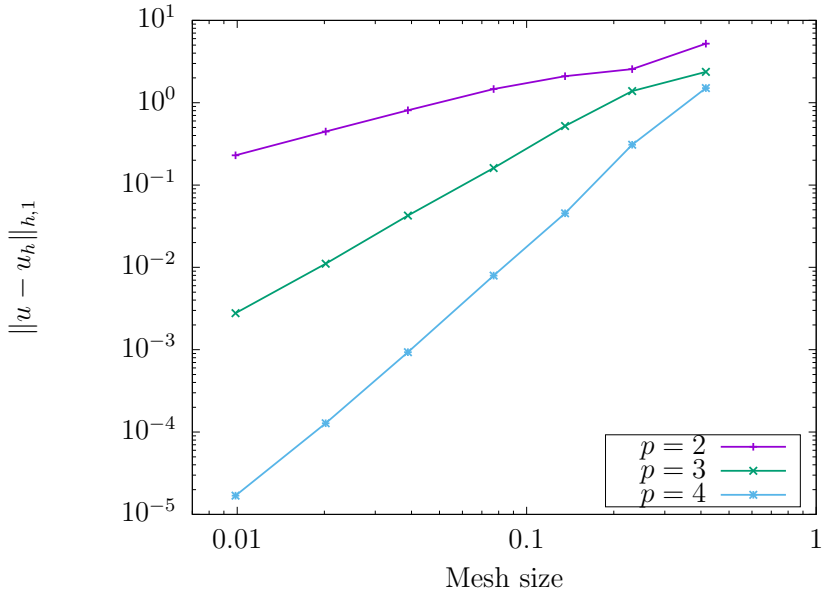


Figure 5.1: Convergence rates for the numerical scheme applied to problem (5.9.1). The error $\|u-u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p = 2$ to $p = 4$. We observe the optimal rate of convergence $\|u-u_h\|_{h,1} = \mathcal{O}(h^{p-1})$ for all values of p.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4158	5.21		2.37		1.51	
0.2314	2.56	(1.21)	1.38	(0.92)	3.08×10^{-1}	(2.71)
0.1357	2.10	(0.38)	5.20×10^{-1}	(1.83)	4.54×10^{-2}	(3.59)
0.0769	1.46	(0.63)	1.61×10^{-1}	(2.07)	7.94×10^{-3}	(3.07)
0.0389	8.09×10^{-1}	(0.87)	4.25×10^{-2}	(1.95)	9.31×10^{-4}	(3.14)
0.0202	4.47×10^{-1}	(0.91)	1.11×10^{-2}	(2.06)	1.28×10^{-4}	(3.03)
0.0099	2.30×10^{-1}	(0.93)	2.78×10^{-3}	(1.93)	1.69×10^{-5}	(2.82)

Table 5.1: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 5.9.1.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.4158	0.47	7.20	7.90	96	160	240
0.2314	0.17	0.18	0.19	384	640	960
0.1357	0.19	0.20	0.28	1044	1740	2610
0.0769	0.23	0.34	0.61	3420	5700	8550
0.0389	0.51	1.25	3.06	13920	23200	34800
0.0202	2.28	7.52	20.85	52476	87460	131190
0.0099	14.77	55.51	176.64	205848	343080	514620

Table 5.2: Runtimes and number of DoFs for Experiment 5.9.1, for each mesh size h, and each polynomial degree, p.

5.9.2 Experiment 2

In this experiment, we consider the following problem

$$\begin{cases} \displaystyle\sum_{i,j=1}^{2}(1+\delta_{ij})\frac{x_i}{|x_i|}\frac{x_j}{|x_j|}D^2_{ij}u = f, \quad \text{in} \quad \Omega, \\ \qquad\qquad\qquad\qquad u = 0, \quad \text{on} \quad \partial\Omega, \end{cases} \tag{5.9.2}$$

where $\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$. In this case, f is chosen so that the solution of (5.9.2) is given by

$$u(x_1, x_2) = \frac{1}{4}\sin(\pi(x_1^2 + x_2^2)).$$

We can also directly calculate the renormalisation parameter, γ, and provide the largest value of ε for which the Cordes condition (5.1.4) holds. In particular, we have that

$$\gamma = \frac{\mathrm{Tr}(A)}{|A|^2} = \frac{2 + x_1^2/|x_1|^2 + x_2^2/|x_2|^2}{8 + 2x_1^2x_2^2/(|x_1|^2|x_2|^2)} = 2/5, \quad \text{and} \quad \varepsilon = 3/5.$$

Furthermore, since Ω is the unit disk, $\partial\Omega = \mathbb{S}^1$, and it follows that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore, $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. For the internal faces, the mean curvature is calculated directly as $\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F$, where n_F is a fixed choice of unit normal to F.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u - u_h\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 5.2, and report the exact values in Table 5.3, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of

freedom (DoFs) and run times for each computation in Table 5.4. One can see that the error values in the $\|\cdot\|_{h,1}$ norm, and runtimes vary only slightly from those of experiment 5.9.1 (see Tables 5.2 and 5.1), highlighting the robustness of this method with respect to the choice of coefficient matrix $A \in L^\infty(\Omega;\mathbb{R}^{d\times d}_{\mathrm{Sym}}) \setminus C(\Omega;\mathbb{R}^{d\times d}_{\mathrm{Sym}})$. In particular, for this example, the off diagonal entries of A are discontinuous across the set $\{x = 0 \text{ or } y = 0\}$.

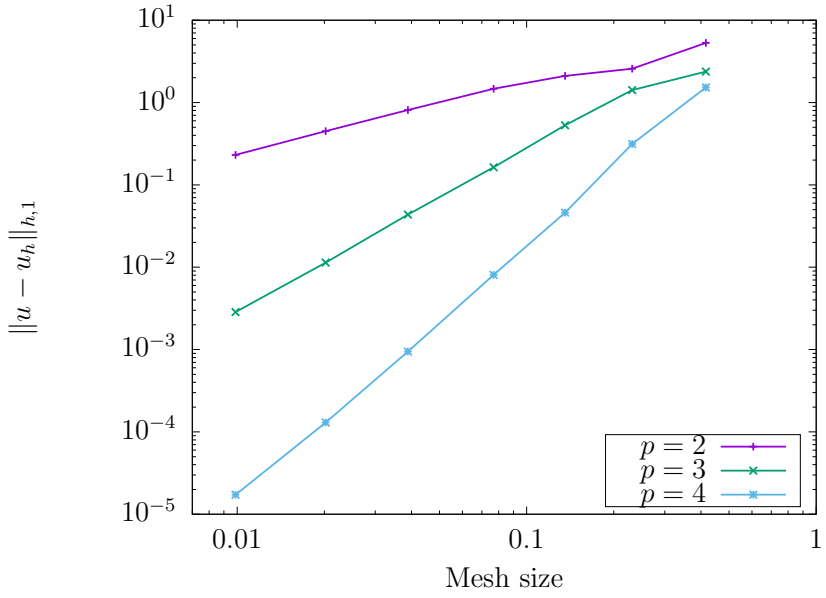


Figure 5.2: Convergence rates for the numerical scheme applied to problem (5.9.2). The error $\|u-u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p = 2$ to $p = 4$. We observe the optimal rate of convergence $\|u-u_h\|_{h,1} = \mathcal{O}(h^{p-1})$ for all values of p.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4158	5.30		2.38		1.53	
0.2314	2.57	(1.23)	1.42	(0.88)	3.14×10^{-1}	(2.70)
0.1357	2.10	(0.38)	5.29×10^{-1}	(1.85)	4.59×10^{-2}	(3.60)
0.0769	1.47	(0.63)	1.64×10^{-1}	(2.07)	8.07×10^{-3}	(3.06)
0.0389	8.13×10^{-1}	(0.87)	4.36×10^{-2}	(1.94)	9.46×10^{-4}	(3.14)
0.0202	4.49×10^{-1}	(0.91)	1.14×10^{-2}	(2.05)	1.30×10^{-4}	(3.03)
0.0099	2.31×10^{-1}	(0.93)	2.85×10^{-3}	(1.93)	1.72×10^{-5}	(2.82)

Table 5.3: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 5.9.2.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.4158	5.56	6.46	6.66	96	160	240
0.2314	0.16	0.17	0.19	384	640	960
0.1357	0.18	0.27	0.29	1044	1740	2610
0.0769	0.23	0.39	0.65	3420	5700	8550
0.0389	0.53	1.33	3.37	13920	23200	34800
0.0202	2.32	7.95	22.52	52476	87460	131190
0.0099	16.30	56.82	174.80	205848	343080	514620

Table 5.4: Runtimes and number of DoFs for Experiment 5.9.2, for each mesh size h, and each polynomial degree, p.

5.9.3 Experiment 3

In this experiment, we consider the PDE given by (5.9.2), and so the renormalisation parameter, γ, and the largest value of ε for which the Cordes condition (5.1.4) holds, are given by

$$\gamma = \frac{\mathrm{Tr}(A)}{|A|^2} = \frac{2 + x_1^2/|x_2|^2 + x_1^2/|x_2|^2}{8 + 2x_1^2x_2^2/(|x_1|^2|x_2|^2)} = 2/5, \quad \text{and} \quad \varepsilon = 3/5.$$

In this case f is chosen so that the solution of (5.9.2) is also given by

$$u(x_1, x_2) = \frac{1}{4}\sin(\pi(x_1^2 + x_2^2)).$$

We have also taken Ω to be the "key-hole" shaped domain (see Figure 2.1) given by

$$\Omega = \{x_1^2 + x_2^2 < 1 : x_2 \geq 1/\sqrt{2}\} \cup [-1/\sqrt{2}, 1/\sqrt{2}] \times [-3, 1/\sqrt{2}]. \tag{5.9.3}$$

This domain is piecewise convex, as it has piecewise nonnegative curvature, but it is not a convex domain, demonstrating the robustness of this method with respect to the choice of domain.

Here, $\partial\Omega = \cup_{i=1}^4 \overline{\Gamma_i}$, where

$$\Gamma_1 = \{(x_1, x_2) \in \mathbb{S}^1 : x_2 > 1/\sqrt{2}\}, \tag{5.9.4}$$

$$\Gamma_2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = -1/\sqrt{2}, -3 < x_2 < 1/\sqrt{2}\}, \tag{5.9.5}$$

$$\Gamma_3 = \{(x_1, x_2) \in \mathbb{R}^2 : -1/\sqrt{2} < x_1 < 1/\sqrt{2}, x_2 = -3\}, \tag{5.9.6}$$

$$\Gamma_4 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 1/\sqrt{2}, -3 < x_2 < 1/\sqrt{2}\}. \tag{5.9.7}$$

Since $\Gamma_1 \subset \mathbb{S}^1$, it follows that $\mathcal{H}_{\Gamma_1} = 1$, and so $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$ that are contained in Γ_1. We can also see that Γ_2, Γ_3, and Γ_4 are flat, and so $\mathcal{H}_F = 0$ for the remaining $F \in \mathscr{E}_h^b$ that are not contained in Γ_1. For the internal faces, the mean curvature is calculated directly as $\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F$, where n_F is a fixed choice of unit normal to F.

Furthermore, the boundary value problem considered is inhomogeneous. In order to extend our numerical method (5.4.5) to this case, we simply modify the right hand side as follows (denoting g to be the restriction of u the boundary, $\partial\Omega$)

$$A_h^{\mathcal{D}}(u_h, v_h) = \sum_{K \in \mathscr{T}_h} \langle \gamma f, \Delta v_h \rangle_K + \sum_{F \in \mathscr{E}_h^b} [\mu_F \langle \nabla_{\mathbf{T}} g, \nabla_{\mathbf{T}} v_h \rangle_F + \eta_F \langle g, v_h \rangle_F] \\ - \frac{1}{2} \sum_{F \in \mathscr{E}_h^b} [\langle \mathrm{div}_{\mathbf{T}} \nabla_{\mathbf{T}} g, \nabla v_h \cdot n_F \rangle_F + \langle \nabla_{\mathbf{T}} (\nabla v_h \cdot n_F), \nabla_{\mathbf{T}} g \rangle_F - \langle \nabla_{\mathbf{T}} g \, \nabla_{\mathbf{T}} n_F^T \, \nabla_{\mathbf{T}} v_h \rangle_F]. \tag{5.9.8}$$

In this particular example, since $\partial\Omega = \cup_{i=1}^4 \overline{\Gamma_i}$, where $\Gamma_1, \ldots, \Gamma_4$ are given by (5.9.4)–(5.9.7), one can calculate that for any $F \in \mathscr{E}_h^b$ that is contained in Γ_1, and any $x \in F$, $\tau_1 \nabla_{\mathbf{T}} n_F^T \tau_2 = \tau_1 \cdot \tau_2$ for any two tangent vectors τ_1, τ_2 at x. Furthermore, $\nabla_{\mathbf{T}} n_F^T = 0$ for any $F \in \mathscr{E}_h^b$ that are not contained in Γ_1. Thus, the final term of the final sum in (5.9.8), is calculated directly as $\langle \nabla_{\mathbf{T}} g \, \nabla_{\mathbf{T}} n_F^T \, \nabla_{\mathbf{T}} v_h \rangle_F = \langle \nabla_{\mathbf{T}} g, \nabla_{\mathbf{T}} v_h \rangle_F$ if $F \in \mathscr{E}_h^b$ is contained in Γ_1, and $\langle \nabla_{\mathbf{T}} g \, \nabla_{\mathbf{T}} n_F^T \, \nabla_{\mathbf{T}} v_h \rangle_F = 0$ if $F \in \mathscr{E}_h^b$ is not contained in Γ_1.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u - u_h\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 5.3, and report the exact values in Table 5.5, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 5.6.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2586	6.99×10^1		4.59×10^1		3.66×10^1	
0.1457	5.84×10^1	(0.31)	2.19×10^1	(1.29)	8.54	(2.54)
0.0757	3.62×10^1	(0.73)	6.31	(1.90)	9.84×10^{-1}	(3.30)
0.0397	2.08×10^1	(0.86)	1.77	(1.97)	1.61×10^{-1}	(2.80)
0.0197	1.02×10^1	(1.01)	4.35×10^{-1}	(1.99)	1.87×10^{-2}	(3.07)
0.0101	5.13	(1.04)	1.11×10^{-1}	(2.05)	2.38×10^{-3}	(3.10)

Table 5.5: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 5.9.3.

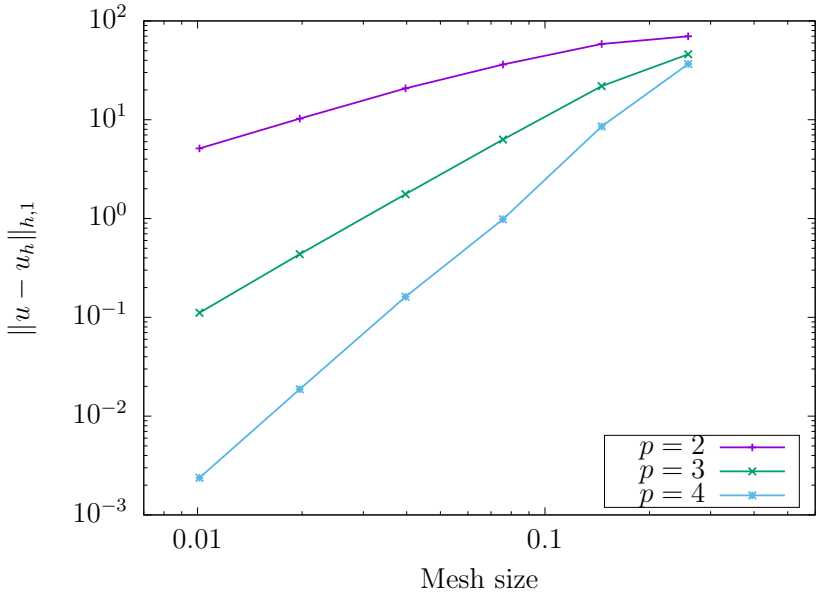


Figure 5.3: Convergence rates for the numerical scheme applied to problem (5.9.2), with Ω given by (5.9.3). The error $\|u - u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p = 2$ to $p = 4$. We observe the optimal rate of convergence $\|u - u_h\|_{h,1} = \mathcal{O}(h^{p-1})$ for all values of p.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.2586	0.52	3.57	4.42	582	970	1455
0.1457	0.20	0.26	0.42	1878	3130	4695
0.0757	0.31	0.54	1.13	6606	11010	16515
0.0397	0.96	2.63	6.92	25818	43030	64545
0.0197	5.39	18.66	53.54	101682	169470	254205
0.0101	34.53	140.70	595.29	397416	662360	993540

Table 5.6: Runtimes and number of DoFs for Experiment 5.9.3, for each mesh size h, and each polynomial degree, p.

5.9.4 Experiment 4 - Consistency

As mentioned in the introduction, the bilinear form $B_{h,*}^{\mathcal{D}}$ defined by (5.4.1) includes terms that are necessary for the consistency of the method, arising from the curvature of the boundary. These terms are not present in the method presented in [110], and the following experiment shows the necessity of including these new terms; in particular, we see both a lack of consistency, and error results inferior to those produced by the new method (5.4.5).

In the results that follow, we provide the consistency residual

$$\mathrm{Res}(w_h) := B^{\mathcal{D}}_{h,*}(w_h, w_h) - \sum_{K\in\mathscr{T}_h} \langle \Delta w_h, \Delta w_h\rangle_K,$$

where the function $w_h \in V^{\mathrm{comp}}_{h,2}$ is given by the Lagrange interpolant of the function $w(x_1, x_2) := x_1^2 + x_2^2 - 1 \in H^s(\Omega; \mathscr{T}_h) \cap H^2(\Omega) \cap H^1_0(\Omega)$, $s > 5/2$, which numerically validates Lemma 5.5.1, as well as the error results arising from one mesh refinement. In the first set of results, we implement the method presented in [110], which we shall call "Method A", for problem (5.9.1), and in the second set, we implement the method presented in this chapter, which we shall call "Method B", for the same problem. In both cases, the solution space is $V^{\mathrm{comp}}_{h,2}$, and $\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$. The results are presented in Table 5.7.

Method	Ref. no.	Mesh size	Res(w)	$\|u - u_h\|_{h,1}$	
A	1	0.02	-25.129	1.60	
A	2	0.01	-25.132	1.56	(0.03)
B	1	0.02	-6.980×10^{-8}	4.46×10^{-1}	
B	2	0.01	-4.473×10^{-9}	2.30×10^{-1}	(0.93)

Table 5.7: Residual values and $\|\cdot\|_{h,1}$-norm error values with EOCs for Experiment 5.9.4, under one mesh refinement.

5.10 Concluding remarks for this method

We have extended the framework introduced in [110], allowing for domains with curved boundaries. We have tested the robustness of this new method (given by (5.4.5)) with numerical experiments involving elliptic operators with discontinuous coefficients, and a uniformly convex domain that has a curved boundary. Furthermore, experiment 5.9.4 validated the necessity of the modifications to the method found in [110], that are present in our new method (5.4.5).

In Experiments 5.9.1, 5.9.2 and 5.9.4, the computational domain we considered was the unit disc; in order to verify the error estimate of Theorem 5.7.1, we used a mesh consisting of curved triangles whose edges were defined by polynomial mappings.

For Experiment 5.9.3, we considered the "key-hole" shaped domain (2.5.2). One should note that this domain is not convex and thus Theorem 5.1.1 does not guarantee the existence and uniqueness of a strong solution to the corresponding boundary-value problem.

The type of problems under consideration (problems in nondivergence form on curved domains) pose many analytical and computational difficulties, whilst housing a large variety of applications; in this chapter we have developed a method that produces optimal error results. This inference has been validated by the analysis in Section 5.7, and the numerical experiments found in Section 5.9.

Chapter 6

A DGFEM for planar oblique boundary-value problems

6.1 New contributions and existing methods

The goal of this chapter is to design and analyse a discontinuous Galerkin finite element method (DGFEM) for the approximation of strong solutions to linear nondivergence form elliptic equations with oblique boundary conditions on curved planar domains. To our knowledge, this is the first DGFEM for such problems.

Existing methods: As will be mentioned in the following section, the available literature on finite element methods for oblique boundary value problems is rather sparse.

- In [110], the authors propose a *hp*-DGFEM for the approximation of strong solutions to linear elliptic equations in nondivergence form with *Dirichlet* boundary conditions on *polytopal* domains. As such, the method of this chapter extends the h-version of this method to the oblique boundary condition on curved planar domains.

- In [55], the author proposes a mixed finite element method for the approximation of solutions of strong solutions to linear nondivergence form elliptic equations with oblique boundary conditions on curved planar domains. The author provides proofs of existence and uniqueness of a numerical solution, as well as a priori and a posteriori error bounds in the case of a continuous piecewise linear approximation space for the approximation of the gradient of the solution, and a range of finite dimensional subspaces of H^1 for the approximation of the solution.

Original contributions of this chapter:

- We provide a DGFEM for the approximation of strong solutions to linear elliptic equations with oblique boundary conditions on domains with *curved* boundaries.
- We prove that the linear operator $A_h^{\mathcal{O}}$ defined in Section 6.6 is coercive, yielding existence and uniqueness of a numerical solution;
- We prove error estimates in broken H^2-type norms for H^2 solutions of sufficient broken Sobolev regularity, and for those that are only assumed to be H^2-regular. In the case of sufficient broken regularity, the error estimate is optimal with respect to the mesh size. For solutions possessing only H^2 regularity, we show that the method is at least as accurate as a method with an approximation space V that varies between being fully conforming and nonconforming (see Remark 6.9.4 for further details).

6.2 A brief introduction to oblique boundary-value problems

The model problem that we consider in this chapter is the following oblique boundary-value problem: find $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} \displaystyle\sum_{i.j=1}^{2} A_{ij} D_{ij}^2 u = f \text{ in } \Omega, \\ \qquad \beta \cdot \nabla u \text{ is constant on } \partial\Omega, \end{cases} \tag{6.2.1}$$

where Ω is a given, C^2 domain in $\mathbb{R}^2$, $f \in L^2(\Omega)$, and $A = (A_{ij})_{i,j=1}^2 \in L^\infty(\Omega)$ satisfies

$$x^T A(\xi) x \geq \lambda |x|^2 \quad \forall x \in \mathbb{R}^2, \text{ for a.e. } \xi \in \Omega,$$

for some constant $\lambda > 0$. The constant present in the boundary condition of (6.2.1) is there to absorb potential compatibility conditions (consider solving the Poisson problem with a homogeneous Neumann boundary condition imposed). The vector-valued function $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$, is called the "oblique vector".

The oblique boundary-value problem appears in several interesting applications, often dependent upon which dimension, d, is considered, and whether or not the oblique boundary-value problem is *strict*. For $d \geq 2$, and $\beta \in C^1(\partial\Omega; \mathbb{S}^d)$, the boundary-value problem is referred to as *strictly* oblique, if there exits a constant $\delta > 0$ such that

$$\beta \cdot n_{\partial\Omega} \geq \delta \quad \text{on } \partial\Omega, \tag{6.2.2}$$

where $n_{\partial\Omega}$ is the unit outward normal to $\partial\Omega$. If (6.2.2) may only hold with $\delta = 0$, then the boundary-value problem is called *degenerate* oblique. As we have seen in Chapter 3 (see Section 3.3.4), in the case that $d = 2$, we do not require the condition (6.2.2), that is, the oblique vector may become tangential to the boundary, and may even fully rotate around the normal vector $n_{\partial\Omega}$. For $d \geq 3$, (6.2.2) is necessary for the well-posedness of the boundary-value problem, with [102] (pg 13–14) providing counter examples to uniqueness for the Poisson problem, in the case that the oblique vector becomes tangential to the boundary, even on a set of zero boundary measure.

That said, the degenerate (or tangential) oblique problem (falling into the class of degenerate elliptic problems), arises naturally in the (geodetic) problem of determining the gravitational fields of celestial bodies [97]. This problem was discovered by Poincaré [100] during his work on the theory of tides. In the case that $d = 2$, the oblique boundary-value problem arises in systems of conservation laws in [121, 30], where the latter focuses on a mixed elliptic-hyperbolic problem that requires the boundary condition to be strictly oblique. For an overview of the problem for $d = 2$ one should refer to [83], and for the case $d \geq 3$, one should seek [84]. A particular, and broad subclass of the oblique boundary-value problem is the case when $\beta \equiv n_{\partial\Omega}$, which is in fact the Neumann boundary-value problem.

Our interest in this type of boundary-value problem stems from applications to fully nonlinear second-order elliptic partial differential equations (PDEs). In particular equations of Monge–Ampère (MA) and Hamilton–Jacobi–Bellman (HJB) type, with oblique boundary conditions. Upon linearising such equations (for instance by the application of Newton's method), one arrives at an infinite sequence of problems of the form (6.2.1), and as such, the linear theory contained in this chapter will be applicable when considering these nonlinear problems in the later chapters.

The problem (6.2.1) poses several difficulties, both analytically and numerically. For instance, it is in non-divergence form (and due to the generality of our assumptions, it *cannot* be written in divergence form, so the standard weak formulation cannot be used here), and as such, well-posedness is not guaranteed (see [57, 89, 108]). In [89], well posedness of (6.2.1) is proven using the method of continuity; recall that in Section 3.3.4 we used a different method, to prove Theorem 3.3.29, which is similar to the proof of Theorem 3 in [110]. The technique we employed relied upon a variant of the Miranda–Talenti estimate (see Lemma 3.3.26). Our motivation for using this different technique is the fact that it is applicable to both linear elliptic equations

and fully nonlinear elliptic (HJB) equation, for which the method present in [89] does not apply.

The Miranda–Talenti estimate is known to hold for $d \geq 2$ when considering functions in a suitable Sobolev space, whose trace or normal derivative vanishes on $\partial\Omega$. However, for $d \geq 3$, this estimate remains an open problem if one assumes that the oblique derivative $(\beta \cdot \nabla u)$ is constant on the boundary [43], restricting us to a two-dimensional framework.

Our goal in this chapter is not only to recap on the analytical framework for two-dimensional oblique boundary-value problems, but also to present and analyse a discontinuous Galerkin finite element method (DGFEM) that approximates solutions to such problems. Interestingly, our method also gives an approximation of the constant that arises in the compatibility condition for problems of the form (6.2.1). In the case of conormal boundary-value problems, i.e.,

$$\begin{cases} -\nabla \cdot (A\nabla u) = f, \text{ in } \Omega, \\ \quad (A\nabla u) \cdot n \text{ is constant, on } \partial\Omega, \end{cases}$$

it is quite straightforward (by an application of the Divergence Theorem) to determine the constant in terms of the function f. In the problems we consider, however, there does not appear to be any explicit relationship between the constant and the right hand side, in general.

Our approach extends the framework of [110] and [70] (the first of which applies to the Dirichlet boundary condition on polytopal domains, and the second applies to the Dirichlet boundary condition on curved, uniformly convex domains) to the oblique case. To our knowledge, there is a sparse amount of work on finite element methods for oblique boundary-value problems present in the existing literature, for instance [119, 47, 12], where [47] and [12] apply to a particular geodetic and free boundary problem respectively. As such, a motivation of this chapter is to widen the scope of the current numerical framework for oblique boundary-value problems. This chapter is organised as follows: Firstly, we discuss the design of the numerical method, motivated by the need to enforce a numerical analogue of the Miranda–Talenti (MT) estimates (3.3.9) and (3.3.10). We then prove an important consistency result, and proceed to prove a stability result for our numerical scheme; this stability result is then used as a main tool in the proof of existence and uniqueness of a numerical solution. Finally we prove an error estimate that is optimal in terms of the mesh size that assume sufficient broken regularity of the solution, and an error estimate that only assumes conforming regularity. We test our method by running several

numerical experiments where the true solution is known. This allows us to verify the error estimates, and to test the robustness of the scheme by considering operators with discontinuous coefficients, and various oblique vectors. We finalise by giving concluding remarks on this method.

6.3 Existence and uniqueness

We now provide a theorem for the existence and uniqueness of a strong solution to (6.2.1), as a consequence of Theorem 3.3.29. We also recall some of the relevant notation from Section 3.3.4.

We denote by $\partial_{\mathbf{T}_2} := \mathbf{T}_2 \cdot \nabla$, and $\mathbf{T}_2 := (-[n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1)^T$ is the unit tangent vector to $\partial\Omega$ obtained by rotating the unit outward normal vector, $n_{\partial\Omega}$, anticlockwise by $\pi/2$. The "oblique angle", $\Theta : \partial\Omega \to \mathbb{R}$, is the (anticlockwise) oriented angle between the oblique vector, β, and the unit outward normal, $n_{\partial\Omega}$. Furthermore, $\mathcal{H}_{\partial\Omega} := \nabla_{\mathbf{T}_2} \cdot n_{\partial\Omega}$ is the mean curvature of $\partial\Omega$.

Theorem 6.3.1 *Assume that $\Omega \subset \mathbb{R}^2$, has a C^2 boundary, $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$, and that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Then, there exists a unique $u \in H^2_{\beta,0}(\Omega)$ that is a strong solution of (6.2.1).

Proof: We simply apply Theorem 3.3.29, with Λ as a singleton set, corresponding to the single operator L given by (5.1.2), and the single right-hand side $f \in L^2(\Omega)$. We then see that

$$F[u] = \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = Lu - f.$$

Thus, Theorem 3.4.1 yields the existence of a unique $u \in H := H^2_{\beta,0}(\Omega)$ that satisfies (6.2.1). $\square$

6.4 Computational domain assumptions

For the DGFEM of this chapter, we impose an additional assumption upon the domain $\Omega \subset \mathbb{R}^2$ that we consider for computations. In particular, as in the statement of Theorem 6.3.1, we shall assume that Ω is a C^2 domain, that $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$ and that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Furthermore, we will assume that Ω is also piecewise C^3.

This additional assumption stems from the fact that we require that an arbitrary finite element function $v_h \in V_{h,p}$ belongs to $H^3(\Omega; \mathscr{T}_h)$. Let us recall the definition of the space $V_{h,p}$:

$$V_{h,p} := \{v \in L^2(\Omega) : \ v|_K = \hat{\rho} \circ F_K^{-1}, \, \hat{\rho} \in \mathbb{P}^p(\hat{K}), \ \forall K \in \mathscr{T}_h\},$$

in particular, one can see that $v_h|_K = \hat{\rho} \circ F_K^{-1}$, where the function $F_K^{-1} \in C^m(K)$, $m \in \mathbb{N}$ if Ω is piecewise C^m. Since the functions $\hat{\rho}$ are smooth (as they are polynomials), we require that $m \geq 3$, in order to deduce that $v_h|_K \in H^3(K)$ for all $K \in \mathscr{T}_h$, which requires Ω to be piecewise C^3.

Two particular examples of when such broken regularity is utilised are (6.5.12) (here we require $v_h \in H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$), in the discussion in Section 6.5, and the application of the trace estimate that leads to (6.8.4) in the proof of Lemma 6.8.2 (stability of the method) requires that $v_h \in H^3(\Omega; \mathscr{T}_h)$.

6.5 The design of the numerical method

As in Chapter 5, we shall discuss how the terms in the definition bilinear form that defines the method of this chapter, arise. We are motivated by the desire to numerically enforce the Miranda–Talenti (MT) estimates (3.3.9) and (3.3.10), whilst producing a scheme that is both consistent and symmetric (the latter occurs when the operator $A : D^2$ is isotropic).

In this numerical method, we solve for both $u_h \in V_{h,p,0} := V_{h,p} \cap L^2_0(\Omega)$, which approximates the strong solution $u \in H^2_{\beta,0}(\Omega)$ of (6.2.1), and $c_h \in V_{h,0}$, which approximates the compatibility constant of (6.2.1), that is, c_h approximates the value of $C = \beta \cdot \nabla u|_{\partial\Omega}$ (the value of C is a priori *unknown*). As such, our finite element space will be

$$M_h := V_{h,p,0} \times V_{h,0}.$$

We first note that the bilinear form will take the following structure:

$$\begin{aligned} A_h^{\mathcal{O}}((u_h, c_h); (v_h, \mu_h)) := \sum_{K \in \mathscr{T}_h} \langle \gamma A : D^2 u_h, \Delta v_h \rangle_K + B^{\mathcal{O}}_{h,1/2}((u_h, c_h); (v_h, \mu_h)) \\ - \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K \quad \forall (u_h, c_h), (v_h, \mu_h) \in M_h. \end{aligned} \tag{6.5.1}$$

We *claim* that the bilinear form $B^{\mathcal{O}}_{h,1/2}$ is coercive on $M_h \times M_h$, and that

$$B^{\mathcal{O}}_{h,1/2}((w, c), (v_h, \mu)) = \sum_{K \in \mathscr{T}_h} \langle \Delta w, \Delta v_h \rangle_K, \quad \forall (v_h, \mu) \in V_{h,p,0} \times \mathbb{R}, \tag{6.5.2}$$

when $w \in H^2_{\beta,0}(\Omega) \cap H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$, and $c = \beta \cdot \nabla w|_{\partial\Omega}$.

It is then clear that (6.5.2) implies that

$$A_h^{\mathcal{O}}((w,c);(v_h,\mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma A : D^2 u_h, \Delta v_h\rangle_K = A_\gamma^h(w, v_h), \quad \forall (v_h,\mu) \in V_{h,p,0} \times \mathbb{R}, \tag{6.5.3}$$

for the aforementioned choice of w and c.

Recall that A_γ^h is the numerical discretisation of A_γ, defined by

$$A_\gamma(u,v) = \int_\Omega \gamma A : D^2 u \, \Delta v \quad \forall u, v \in H^2_{\beta,0}(\Omega),$$

which is also present in the definition (5.4.4) of the bilinear form $A_h^{\mathcal{D}}$. Furthermore, the term

$$-\sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K$$

is also present in both (6.5.1) and (5.4.4), and thus, the key differences between the the method of Chapter 5 and the method we present in this chapter are present in the bilinear forms $B^{\mathcal{O}}_{h,1/2}$ and $B^{\mathcal{D}}_{h,1/2}$.

The bilinear form $B^{\mathcal{O}}_{h,1/2} : M_h \times M_h \to \mathbb{R}$ takes the following form

$$\begin{aligned} B^{\mathcal{O}}_{h,1/2}((u_h,c_h);(v_h,\mu_h)) := \frac{1}{2} B^{\mathcal{O}}_{h,*}((u_h,c_h);(v_h,\mu_h)) + \frac{1}{2}\sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K \\ + J_h^{\mathcal{O}}((u_h,c_h);(v_h,\mu_h)), \end{aligned} \tag{6.5.4}$$

where the bilinear forms $B^{\mathcal{O}}_{h,*}, J_h^{\mathcal{O}} : M_h \times M_h \to \mathbb{R}$ satisfy

$$B^{\mathcal{O}}_{h,*}((w,c),(v_h,\mu)) = \sum_{K\in\mathscr{T}_h} \langle \Delta w, \Delta v_h\rangle_K, \quad \text{and} \quad J_h^{\mathcal{O}}((w,c),(v_h,\mu)) = 0, \tag{6.5.5}$$

for all $(v_h,\mu) \in V_{h,p,0} \times \mathbb{R}$, when $w \in H^2_{\beta,0}(\Omega) \cap H^s(\Omega;\mathscr{T}_h)$, $s > 5/2$, and $c = \beta \cdot \nabla w|_{\partial\Omega}$.

Moreover, one can see that the above implies (6.5.2) which in turn implies (6.5.3). We also remark that the bilinear form $J_h^{\mathcal{O}}$ plays no role in the consistency identity (6.5.3) (other than by its absence), and is in fact a jump penalty term that enforces regularity that is consistent with that of the true solution. In particular, if $w \in H^2_{\beta,0}(\Omega) \cap H^1_0(\Omega)$ (which is the space that the strong solution of (6.2.1) belongs to), and $c = \beta \cdot \nabla w|_{\partial\Omega}$, then we see that

$$[\![c]\!] = [\![w]\!] = [\![\nabla w \cdot n_F]\!] = [\![\nabla_{\mathbf{T}} w]\!] = 0 \quad \forall F \in \mathscr{E}_h^i, \tag{6.5.6}$$

and furthermore, since $\tau_F(\beta \cdot \nabla w) = c$ for all $F \in \mathscr{E}_h^b$, it follows that

$$[\![\beta \cdot \nabla w - c]\!] = [\![\partial_{\mathbf{T}_2}(\beta \cdot \nabla w)]\!] = 0 \quad \forall F \in \mathscr{E}_h^b. \tag{6.5.7}$$

$J_h^{\mathcal{O}}$ also enforces the oblique boundary condition, and leads to the bilinear form $B_{h,1/2}^{\mathcal{O}}$ being provably coercive (in a particular H^2-type norm on M_h). In particular we define $J_h^{\mathcal{O}}$ as follows:

$$\begin{aligned} J_h^{\mathcal{O}}((u_h,\lambda),(v_h,\mu)) &:= \sum_{F\in\mathscr{E}_h^i} \mu_F \langle [\![\nabla_{\mathbf{T}} u_h]\!], [\![\nabla_{\mathbf{T}} v_h]\!]\rangle_F] \\ &+ \sum_{F\in\mathscr{E}_h^i} [\mu_F \langle [\![\nabla u_h \cdot n_F]\!], [\![\nabla v_h \cdot n_F]\!]\rangle_F + \eta_F \langle [\![u_h]\!], [\![v_h]\!]\rangle_F + \ell_F \langle [\![\lambda]\!], [\![\mu]\!]\rangle_F] \\ &+ \sum_{F\in\mathscr{E}_h^b} \sigma_F \langle \beta\cdot\nabla u_h - \lambda, \beta\cdot\nabla v_h - \mu\rangle_F, \end{aligned} \tag{6.5.8}$$

where the positive edge-dependent quantities μ_F, η_F, ℓ_F, and σ_F will be specified later, and their particular choice will become apparent when we prove that $B_{h,1/2}^{\mathcal{O}}$ is coercive (see Lemma 6.8.2). Furthermore, (6.5.6) and (6.5.7) imply that

$$J_h^{\mathcal{O}}((w,c),(v_h,\mu)) = 0, \quad \forall (v_h,\mu)\in V_{h,p,0}\times\mathbb{R}, \tag{6.5.9}$$

when $w \in H^2_{\beta,0}(\Omega)\cap H^s(\Omega;\mathscr{T}_h)$, $s>5/2$, and $c=\beta\cdot\nabla w|_{\partial\Omega}$.

The bilinear form $B_{h,*}^{\mathcal{O}}$ plays a key role (when paired with the remaining term in the definition of $B_{h,1/2}^{\mathcal{O}}$) in identity (6.5.2), and its structure is motivated by Corollary 3.3.23 and Lemma 3.3.25, the statements of which we recall.

Statement of Corollary 3.3.23: Assume that $E\subset\mathbb{R}^2$ is a bounded, Lipschitz, piecewise C^2 domain, and that $\beta\in C^1(\Gamma_n;\mathbb{S}^1)$ for each C^2 portion Γ_n of ∂E, $n=1,\dots,N$, $N\in\mathbb{N}$. Then, for any $u,v\in H^s(E)$, $s>5/2$, we have that

$$\begin{aligned} \int_E D^2u : D^2v &+ \int_{\partial E} (\beta_1\partial_{\mathbf{T}_2}\beta_2 - \beta_2\partial_{\mathbf{T}_2}\beta_1)\,(\beta^\perp\cdot\nabla u\,\beta^\perp\cdot\nabla v + \beta\cdot\nabla u\,\beta\cdot\nabla v) \\ &+ \int_{\partial E} \big(\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u)\beta\cdot\nabla v - \partial_{\mathbf{T}_2}(\beta\cdot\nabla u)\beta^\perp\cdot\nabla v\big) \\ &= \int_E \Delta u\,\Delta v. \end{aligned} \tag{6.5.10}$$

Statement of Lemma 3.3.25: Let $\Omega\subset\mathbb{R}^2$ be a C^2 domain, and assume that $\beta\in C^1(\partial\Omega;\mathbb{S}^1)$. Then, on $\partial\Omega$, we have that

$$\beta_1(\partial_{\mathbf{T}_2}\beta_2) - (\partial_{\mathbf{T}_2}\beta_1)\beta_2 = \partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega}. \tag{6.5.11}$$

Designing the bilinear form: Let us consider $K\in\mathscr{T}_h$ that satisfies $|\partial K\cap\partial\Omega|\neq 0$ (this allows for elements with one curved edge that lies on $\partial\Omega$, but excludes elements that only intersect $\partial\Omega$ at a vertex). Note that $K\subset\mathbb{R}^2$ is bounded with a Lipschitz continuous, piecewise C^3 boundary. K also has three edges F_K^1, F_K^2, F_K^3, (each of

which are C^3 (and hence C^2) portions of ∂K) and three vertices e_K^1, e_K^2, e_K^3. Let e_K^1 and e_K^2 be the two vertices that lie on $\partial\Omega$, and let F_K^1 be the *curved* side that lies on $\partial\Omega$ and also connects e_K^1 and e_K^2. Finally, let F_K^2 be the *straight* edge of K that connects e_K^2, and e_K^3. It then follows that F_K^3 is the remaining *straight* edge that connects e_K^3, and e_K^1.

Now define $\tilde{\beta} : \partial K \to \mathbb{S}^1$ by

$$\begin{cases} \tilde{\beta}|_{F_K^1} = \beta, \\ \tilde{\beta}|_{F_K^2} = \beta(e_K^2), \\ \tilde{\beta}|_{F_K^3} = \beta(e_K^1), \end{cases}$$

and so $\tilde{\beta} \in C^1(F_K^j; \mathbb{S}^1)$, $j = 1, 2, 3$, where $\partial K = \cup_{j=1}^3 \overline{F_K^j}$.

Then, noting that $w, v_h \in H^s(K)$, $s > 5/2$, applying (6.5.10) with $E := K$, and $\beta := \tilde{\beta}$, we obtain

$$\begin{aligned} &\int_K D_{11}^2 w D_{22}^2 v_h + D_{22}^2 w D_{11}^2 v_h - 2 D_{12}^2 w D_{12}^2 v_h = \\ &\qquad \int_{\partial K} \left(\tilde{\beta}_1 \partial_{\mathbf{T}_2} \tilde{\beta}_2 - \tilde{\beta}_2 \partial_{\mathbf{T}_2} \tilde{\beta}_1 \right) (\tilde{\beta}^\perp \cdot \nabla w \, \tilde{\beta}^\perp \cdot \nabla v_h + \tilde{\beta} \cdot \nabla w \, \tilde{\beta} \cdot \nabla v_h) \\ &\qquad\quad + \int_{\partial K} \left(\partial_{\mathbf{T}_2} (\tilde{\beta}^\perp \cdot \nabla w) \tilde{\beta} \cdot \nabla v_h - \partial_{\mathbf{T}_2} (\tilde{\beta} \cdot \nabla w) \tilde{\beta}^\perp \cdot \nabla v_h \right) \\ &= \int_{F_K^1} (\beta_1 \partial_{\mathbf{T}_2} \beta_2 - \beta_2 \partial_{\mathbf{T}_2} \beta_1) (\beta^\perp \cdot \nabla w \, \beta^\perp \cdot \nabla v_h + \beta \cdot \nabla w \, \beta \cdot \nabla v_h) \\ &\qquad + \int_{F_K^1} \left(\partial_{\mathbf{T}_2} (\beta^\perp \cdot \nabla w) \beta \cdot \nabla v_h - \partial_{\mathbf{T}_2} (\beta \cdot \nabla w) \beta^\perp \cdot \nabla v_h \right) \\ &\qquad + \int_{F_K^2 \cup F_K^3} \left(\partial_{\mathbf{T}_2} (\tilde{\beta}^\perp \cdot \nabla w) \tilde{\beta} \cdot \nabla v_h - \partial_{\mathbf{T}_2} (\tilde{\beta} \cdot \nabla w) \tilde{\beta}^\perp \cdot \nabla v_h \right) \end{aligned} \tag{6.5.12}$$

Furthermore, upon noting that $\tilde{\beta}$, $\tilde{\beta}^\perp$, and the unit normal to ∂K, are all constant on F_K^2 and F_K^3, one can calculate the following:

$$\begin{aligned} &\int_{F_K^2 \cup F_K^3} \left(\partial_{\mathbf{T}_2} (\tilde{\beta}^\perp \cdot \nabla w) \tilde{\beta} \cdot \nabla v_h - \partial_{\mathbf{T}_2} (\tilde{\beta} \cdot \nabla w) \tilde{\beta}^\perp \cdot \nabla v_h \right) \\ &\qquad = \int_{F_K^2 \cup F_K^3} \Delta w \, (\nabla v_h \cdot n_{\partial K}) - \nabla (\nabla w \cdot n_{\partial K}) \cdot \nabla v_h. \end{aligned} \tag{6.5.13}$$

Since $F_K^1 \subset \partial\Omega$, we may apply identity (6.5.11), obtaining

$$\begin{aligned} &\int_{F_K^1} (\beta_1 \partial_{\mathbf{T}_2} \beta_2 - \beta_2 \partial_{\mathbf{T}_2} \beta_1) (\beta^\perp \cdot \nabla w \, \beta^\perp \cdot \nabla v_h + \beta \cdot \nabla w \, \beta \cdot \nabla v_h) \\ &\qquad = \int_{F_K^1} \left(\partial_{\mathbf{T}_2} \Theta + \mathcal{H}_{F_K^1} \right) (\beta^\perp \cdot \nabla w \, \beta^\perp \cdot \nabla v_h + \beta \cdot \nabla w \, \beta \cdot \nabla v_h), \end{aligned} \tag{6.5.14}$$

where $\mathcal{H}_{F_K^1} = \mathcal{H}_{\partial\Omega}|_{F_K^1}$.

We now consider an element $K \in \mathscr{T}_h$ that satisfies $|\partial K \cap \partial\Omega| = 0$. An application of integration by parts (noting that the unit outward normal to ∂K is constant on each edge of K) yields

$$\int_K D^2 w_h : D^2 v_h + \int_{\partial K} \Delta w \, (\nabla v_h \cdot n_{\partial K}) - \nabla(\nabla w \cdot n_{\partial K}) \cdot \nabla v_h = \int_K \Delta w \, \Delta v_h. \quad (6.5.15)$$

One can also see that for any $K \in \mathscr{T}_h$, and thus in particular, for those $K \in \mathscr{T}_h$ that satisfy $|\partial K \cap \partial\Omega| \neq 0$,

$$\int_K D^2 w : D^2 v_h + (D_{11}^2 w D_{22}^2 v_h + D_{22}^2 w D_{11}^2 v_h - 2 D_{12}^2 w D_{12}^2 v_h) = \int_K \Delta w \, \Delta v_h. \quad (6.5.16)$$

Now, applying identities (6.5.13) and (6.5.14) to (6.5.12), and summing (6.5.15) over all $K \in \mathscr{T}_h$ such that $|\partial K \cap \partial\Omega| = 0$, and (6.5.16) over all $K \in \mathscr{T}_h$ such that $|\partial K \cap \partial\Omega| \neq 0$, we obtain

$$\begin{aligned}
&\sum_{K \in \mathscr{T}_h} \int_K D^2 w : D^2 v_h + \sum_{F \in \mathscr{E}_h^i} \int_F [\![\Delta w \, \nabla v_h \cdot n_F - \nabla(\nabla w \cdot n_F) \cdot \nabla v_h]\!] \\
&\quad + \sum_{F \in \mathscr{E}_h^b} \int_F \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla w) \beta \cdot \nabla v_h - \partial_{\mathbf{T}_2}(\beta \cdot \nabla w) \beta^\perp \cdot \nabla v_h \\
&\qquad + \sum_{F \in \mathscr{E}_h^b} \int_F (\partial_{\mathbf{T}_2} \Theta + \mathcal{H}_F) \, \beta^\perp \cdot \nabla w \, \beta^\perp \cdot \nabla v_h + (\partial_{\mathbf{T}_2} \Theta + \mathcal{H}_F) \, \beta \cdot \nabla w \, \beta \cdot \nabla v_h \\
&\qquad\quad = \sum_{K \in \mathscr{T}_h} \int_K \Delta w \, \Delta v_h,
\end{aligned} \quad (6.5.17)$$

where n_F is now a *fixed* choice of unit normal to F, and $\mathcal{H}_F := \mathcal{H}_{\partial\Omega}|_F$.

Utilising the tangential operator identities (4.10.3) and (4.10.4), we obtain

$$\begin{aligned}
&\sum_{F \in \mathscr{E}_h^i} \int_F [\![\Delta w \, \nabla v_h \cdot n_F - \nabla(\nabla w \cdot n_F) \cdot \nabla v_h]\!] \\
&\qquad = \sum_{F \in \mathscr{E}_h^i} \int_F [\![\Delta_{\mathbf{T}} w \, \nabla v_h \cdot n_F - \nabla_{\mathbf{T}}(\nabla w \cdot n_F) \cdot \nabla_{\mathbf{T}} v_h]\!]
\end{aligned} \quad (6.5.18)$$

We then apply the identity (valid for any $f, g \in H^s(\Omega; \mathscr{T}_h)$, $s > 1/2$)

$$\sum_{F \in \mathscr{E}_h^i} \int_F [\![fg]\!] = \sum_{F \in \mathscr{E}_h^i} \int_F [\![f]\!] \langle\!\langle g \rangle\!\rangle + \sum_{F \in \mathscr{E}_h^i} \int_F \langle\!\langle f \rangle\!\rangle [\![g]\!],$$

along with (6.5.6), to (6.5.18), which gives us

$$\begin{aligned}
&\sum_{F\in\mathscr{E}_h^i}\int_F [\![\Delta w\,\nabla v_h\cdot n_F-\nabla(\nabla w\cdot n_F)\cdot\nabla v_h]\!]\\
&\qquad=\sum_{F\in\mathscr{E}_h^i}\int_F [\![\Delta_{\mathbf{T}} w]\!]\langle\!\langle\nabla v_h\cdot n_F\rangle\!\rangle+\langle\!\langle\Delta_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!]\\
&\qquad\quad-\sum_{F\in\mathscr{E}_h^i}\int_F [\![\nabla_{\mathbf{T}}(\nabla w\cdot n_F)]\!]\cdot\langle\!\langle\nabla_{\mathbf{T}} v_h\rangle\!\rangle+\langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!]\\
&\qquad=\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\Delta_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!]-\langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!].
\end{aligned}\tag{6.5.19}$$

Then, again by (6.5.6) we may consistently symmetrise the final right-hand side of (6.5.19), yielding

$$\begin{aligned}
&\sum_{F\in\mathscr{E}_h^i}\int_F [\![\Delta w\,\nabla v_h\cdot n_F-\nabla(\nabla w\cdot n_F)\cdot\nabla v_h]\!]\\
&\qquad=\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\Delta_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!]+\langle\!\langle\Delta_{\mathbf{T}} v_h\rangle\!\rangle[\![\nabla w\cdot n_F]\!]\\
&\qquad\quad-\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!]+\langle\!\langle\nabla_{\mathbf{T}}(\nabla v_h\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} w]\!].
\end{aligned}\tag{6.5.20}$$

From (6.5.7), we obtain

$$\sum_{F\in\mathscr{E}_h^b}\int_F \partial_{\mathbf{T}_2}(\beta\cdot\nabla w)\beta^\perp\cdot\nabla v_h=0,\tag{6.5.21}$$

for all $F\in\mathscr{E}_h^b$. Furthermore, on $F\in\mathscr{E}_h^b$, one has that

$$\beta^\perp\cdot\nabla w\,\beta^\perp\cdot\nabla v_h+\beta\cdot\nabla w\,\beta\cdot\nabla v_h=\nabla w\cdot\nabla v_h.\tag{6.5.22}$$

Applying (6.5.20), (6.5.21), and (6.5.22) to (6.5.17) we obtain

$$\begin{aligned}
&\sum_{K\in\mathscr{T}_h}\int_K D^2w\colon D^2v_h+\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\Delta_{\mathbf{T}} w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!]+\langle\!\langle\Delta_{\mathbf{T}} v_h\rangle\!\rangle[\![\nabla w\cdot n_F]\!]\\
&\qquad-\sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!]+\langle\!\langle\nabla_{\mathbf{T}}(\nabla v_h\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} w]\!]\\
&\qquad\quad+\sum_{F\in\mathscr{E}_h^b}\int_F \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla w)\beta\cdot\nabla v_h+(\partial_{\mathbf{T}_2}\Theta+\mathcal{H}_F)\,\nabla w\cdot\nabla v_h\\
&\qquad\quad=\sum_{K\in\mathscr{T}_h}\int_K \Delta w\,\Delta v_h.
\end{aligned}\tag{6.5.23}$$

So far, all of the applications of (6.5.6) and (6.5.7) have been made with consistency and symmetry in mind. We make one further alteration, which is necessary for the coercivity of $B^{\mathcal{O}}_{h,1/2}$. In particular, notice that each term of each integrand on the left-hand side of (6.5.23) either has a sign if we take $w = v_h$ (in particular, $D^2 w : D^2 v_h$ and $(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)\nabla w \cdot \nabla v_h$), or consists of the product of two terms, one of which is present in the definition (6.5.8) of the jump stabilisation bilinear form, $J^{\mathcal{O}}_h$, except for the integrand $\partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla w)\beta \cdot \nabla v_h$.

To this end, let us denote by e^+_F and e^-_F the two vertices of an edge $F \in \mathscr{E}^b_h$, and notice that for any $\mu \in \mathbb{R}$,

$$\begin{aligned}\sum_{F\in\mathscr{E}^b_h}\int_F \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla w)\mu &= \sum_{F\in\mathscr{E}^b_h}(\beta^\perp\cdot\nabla w)\mu\big|_{e^-_F}^{e^+_F}\\ &= \mu\sum_{e\in\mathscr{V}^b_h}[\![\beta^\perp\cdot\nabla w]\!]\\ &= 0,\end{aligned} \tag{6.5.24}$$

where the jumps in (6.5.24) are considered across boundary vertices $e \in \mathscr{V}^b_h$. Note that the final equality holds, due to the fact that $\beta^\perp \in C^1(\partial\Omega)$, and $\nabla w \in H^{1/2}(\partial\Omega)$, and thus neither function may jump across boundary vertices.

Applying (6.5.24) to (6.5.23), we obtain

$$\begin{aligned}&\sum_{K\in\mathscr{T}_h}\int_K D^2 w : D^2 v_h + \sum_{F\in\mathscr{E}^i_h}\int_F \langle\!\langle \Delta_{\mathbf{T}} w\rangle\!\rangle [\![\nabla v_h\cdot n_F]\!] + \langle\!\langle\Delta_{\mathbf{T}} v_h\rangle\!\rangle[\![\nabla w\cdot n_F]\!]\\ &\qquad - \sum_{F\in\mathscr{E}^i_h}\int_F \langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} v_h]\!] + \langle\!\langle\nabla_{\mathbf{T}}(\nabla v_h\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}} w]\!]\\ &\qquad + \sum_{F\in\mathscr{E}^b_h}\int_F \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla w)(\beta\cdot\nabla v_h - \mu) + (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)\nabla w\cdot\nabla v_h\\ &\qquad = \sum_{K\in\mathscr{T}_h}\int_K \Delta w\,\Delta v_h.\end{aligned} \tag{6.5.25}$$

Alas, the left-hand side of (6.5.25) is not symmetric. However, by (6.5.7), it follows that

$$\sum_{F\in\mathscr{E}^b_h}\big\langle\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla v_h), \beta\cdot\nabla w - c\big\rangle_F = 0, \tag{6.5.26}$$

which, when applied to (6.5.25), gives us

$$\begin{aligned}\sum_{K\in\mathscr{T}_h}\int_K D^2w\!:\!D^2v_h &+ \sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\Delta_{\mathbf{T}}w\rangle\!\rangle[\![\nabla v_h\cdot n_F]\!] + \langle\!\langle\Delta_{\mathbf{T}}v_h\rangle\!\rangle[\![\nabla w\cdot n_F]\!]\\ &- \sum_{F\in\mathscr{E}_h^i}\int_F \langle\!\langle\nabla_{\mathbf{T}}(\nabla w\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}}v_h]\!] + \langle\!\langle\nabla_{\mathbf{T}}(\nabla v_h\cdot n_F)\rangle\!\rangle\cdot[\![\nabla_{\mathbf{T}}w]\!]\\ &+ \sum_{F\in\mathscr{E}_h^b}\int_F \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla v_h)(\beta\cdot\nabla w - c) + \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla w)(\beta\cdot\nabla v_h-\mu)\\ &+ \sum_{F\in\mathscr{E}_h^b}\int_F (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)\,\nabla w\cdot\nabla v_h\\ &= \sum_{K\in\mathscr{T}_h}\int_K \Delta w\,\Delta v_h,\end{aligned} \tag{6.5.27}$$

consistently restoring symmetry.

We then define $B^{\mathcal{O}}_{h,*}$ by the left-hand side of (6.5.27). That is,

$$\begin{aligned}B^{\mathcal{O}}_{h,*}((u_h,\lambda),(v_h,\mu)) &:= \sum_{K\in\mathscr{T}_h}\langle D^2u_h, D^2v_h\rangle_K\\ &+ \sum_{F\in\mathscr{E}_h^i}[\langle\mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle u_h\rangle\!\rangle, [\![\nabla v_h\cdot n_F]\!]\rangle_F + \langle\mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle v_h\rangle\!\rangle, [\![\nabla u_h\cdot n_F]\!]\rangle_F]\\ &- \sum_{F\in\mathscr{E}_h^i}[\langle\nabla_{\mathbf{T}}\langle\!\langle\nabla u_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}v_h]\!]\rangle_F + \langle\nabla_{\mathbf{T}}\langle\!\langle\nabla v_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}u_h]\!]\rangle_F]\\ &+ \sum_{F\in\mathscr{E}_h^b}[\langle(\partial_{\mathbf{T}_2}\Theta+\mathcal{H}_F)\,\nabla u_h, \nabla v_h\rangle_F]\\ &+ \sum_{F\in\mathscr{E}_h^b}\left[\left\langle\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u_h), \beta\cdot\nabla v_h-\mu\right\rangle_F + \left\langle\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla v_h), \beta\cdot\nabla u_h-\lambda\right\rangle_F\right],\end{aligned}$$

for all $(u_h,\lambda),(v_h,\mu)\in M_h$. It follows from (6.5.27) that the bilinear form $B^{\mathcal{O}}_{h,*}$ satisfies the first identity of (6.5.5). We are now ready to define the numerical method of this chapter.

6.6 The numerical method

The definition of the numerical scheme requires the following bilinear forms, derived in Section 6.5, and concisely defined as follows. $B^{\mathcal{O}}_{h,*}, J^{\mathcal{O}}_h, B^{\mathcal{O}}_{h,\theta} : M_h\times M_h\to\mathbb{R}$, given

by

$$
\begin{aligned}
B^{\mathcal{O}}_{h,*}((u_h,\lambda),(v_h,\mu)) &:= \sum_{K\in\mathscr{T}_h} \langle D^2 u_h, D^2 v_h\rangle_K \\
&+ \sum_{F\in\mathscr{E}_h^i} [\langle \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \langle\!\langle u_h \rangle\!\rangle, [\![\nabla v_h \cdot n_F]\!]\rangle_F + \langle \operatorname{div}_{\mathbf{T}} \nabla_{\mathbf{T}} \langle\!\langle v_h \rangle\!\rangle, [\![\nabla u_h \cdot n_F]\!]\rangle_F] \\
&- \sum_{F\in\mathscr{E}_h^i} [\langle \nabla_{\mathbf{T}} \langle\!\langle \nabla u_h \cdot n_F \rangle\!\rangle, [\![\nabla_{\mathbf{T}} v_h]\!]\rangle_F + \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla v_h \cdot n_F \rangle\!\rangle, [\![\nabla_{\mathbf{T}} u_h]\!]\rangle_F] \\
&+ \sum_{F\in\mathscr{E}_h^b} [\langle (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)\nabla u_h, \nabla v_h\rangle_F] \\
&+ \sum_{F\in\mathscr{E}_h^b} \left[\left\langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla u_h), \beta\cdot\nabla v_h - \mu\right\rangle_F + \left\langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla v_h), \beta\cdot\nabla u_h - \lambda\right\rangle_F\right],
\end{aligned}
\tag{6.6.1}
$$

$$
\begin{aligned}
J^{\mathcal{O}}_h((u_h,\lambda),(v_h,\mu)) &:= \sum_{F\in\mathscr{E}_h^i} \mu_F \langle [\![\nabla_{\mathbf{T}} u_h]\!], [\![\nabla_{\mathbf{T}} v_h]\!]\rangle_F] \\
&+ \sum_{F\in\mathscr{E}_h^i} [\mu_F \langle [\![\nabla u_h\cdot n_F]\!], [\![\nabla v_h\cdot n_F]\!]\rangle_F + \eta_F \langle [\![u_h]\!], [\![v_h]\!]\rangle_F + \ell_F \langle [\![\lambda]\!], [\![\mu]\!]\rangle_F] \\
&+ \sum_{F\in\mathscr{E}_h^b} \sigma_F \langle \beta\cdot\nabla u_h - \lambda, \beta\cdot\nabla v_h - \mu\rangle_F,
\end{aligned}
\tag{6.6.2}
$$

and

$$
\begin{aligned}
B^{\mathcal{O}}_{h,\theta}((u_h,\lambda),(v_h,\mu)) &:= \theta B^{\mathcal{O}}_{h,*}((u_h,\lambda),(v_h,\mu)) + (1-\theta)\sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K \\
&+ J^{\mathcal{O}}_h((u_h,\lambda),(v_h,\mu)),
\end{aligned}
\tag{6.6.3}
$$

for $\theta \in [0,1]$. Note that n_F denotes a *fixed* choice of unit normal to F, $\mathcal{H}_F := \mathcal{H}_{\partial\Omega}|_F$, and that the penalty parameters μ_F, η_F, ℓ_F, and σ_F will be given later on in the chapter.

We define

$$
A^{\mathcal{O}}_h((u_h,\lambda),(v_h,\mu)) := \sum_{K\in\mathscr{T}_h} \langle \gamma L u_h, \Delta v_h\rangle_K + B^{\mathcal{O}}_{h,1/2}((u_h,\lambda),(v_h,\mu)) - \sum_{K\in\mathscr{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K.
\tag{6.6.4}
$$

Now we are ready to state the finite element method: find $(u_h, c_h) \in M_h$ such that

$$
A^{\mathcal{O}}_h((u_h,c_h),(v_h,\mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma f, \Delta v_h\rangle_K \quad \forall (v_h,\mu)\in M_h.
\tag{6.6.5}
$$

Remark 6.6.1 (Extension of the bilinear forms) *The bilinear forms $B^{\mathcal{O}}_{h,*}$ and $J^{\mathcal{O}}_h$ are both defined on $(V_{h,p,0} \times V_{h,0}) \times (V_{h,p,0} \times V_{h,0})$, but one must note that they are both well defined on $(H^s(\Omega;\mathscr{T}_h) \cap H^2_\beta(\Omega) \times V_{h,0}) \times (V_{h,p,0} \times V_{h,0})$ for $s > 5/2$, of which $(H^s(\Omega;\mathscr{T}_h) \cap H^2_\beta(\Omega) \times \mathbb{R}) \times (V_{h,p,0} \times \mathbb{R})$ is a proper subset, that the functions in the following lemma belong to.*

6.7 Consistency of the method

We now provide a lemma on the consistency of the DGFEM of this chapter.

Lemma 6.7.1 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 and piecewise C^3 domain, and let $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$. Furthermore, assume that $\{\mathscr{T}_h\}_h$ is a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9. Let $(w,c) \in H^s(\Omega;\mathscr{T}_h) \cap H^2_\beta(\Omega) \times \mathbb{R}$, $s > 5/2$, where $\beta \cdot \nabla w|_{\partial\Omega} = c$. Then, for every $(v_h,\mu) \in V_{h,p} \times \mathbb{R}$, we have the identities*

$$B^{\mathcal{O}}_{h,*}((w,c),(v_h,\mu)) = \sum_{K\in\mathscr{T}_h} \langle \Delta w, \Delta v_h\rangle_K \quad \textit{and} \quad J^{\mathcal{O}}_h((w,c),(v_h,\mu)) = 0. \tag{6.7.1}$$

Proof: Assume that the pair (w,c) satisfies the hypotheses of the lemma. Then, the identities of (6.7.1) follow from (6.5.27) and (6.5.9). $\square$

6.8 Stability of the method

We now aim to show that $B^{\mathcal{O}}_{h,\theta}$ is coercive in a particular norm on M_h. Before we prove that $B^{\mathcal{O}}_{h,\theta}$ is coercive, we must define the norm in which the bilinear form is coercive. To this end, let us define the following family of functionals, $\|(\cdot,\cdot)\|_{h,\theta} : M_h \to [0,\infty)$ for $\theta \in (0,1]$:

$$\begin{aligned}\|(u_h,\lambda)\|^2_{h,\theta} := \sum_{K\in\mathscr{T}_h} [\theta |u_h|^2_{H^2(K)} + (1-\theta)\|\Delta u_h\|^2_{L^2(K)}] \\ + c_* J^{\mathcal{O}}_h((u_h,\lambda),(u_h,\lambda)) + \frac{\theta}{2}\sum_{F\in\mathscr{E}^b_h} \left\| (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla u_h \right\|^2_{L^2(F)},\end{aligned} \tag{6.8.1}$$

where c_* is a positive constant to be determined.

Lemma 6.8.1 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 and piecewise C^3 domain, and let $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$. Assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \textit{on } \partial\Omega.$$

Furthermore, assume that $\{\mathscr{T}_h\}_h$ *is a regular of order* 2 *family of triangulations on* $\overline{\Omega}$ *satisfying Assumption 4.4.9. Then, for each* $\theta \in (0,1]$, $\|\cdot\|_{h,\theta} : M_h \to [0,\infty)$ *defines a norm on* M_h.

Proof: First we note that homogeneity and the triangle inequality are clear. Now let us assume that the pair $(v_h, \mu) \in M_h$ satisfies

$$\|(v_h,\mu)\|_{h,\theta} = 0,$$

for some $\theta \in (0,1]$. It then follows that $|v_h|_{H^2(\Omega;\mathscr{T}_h)} = 0$, and so v_h is piecewise affine. Moreover

$$[\![\mu]\!]_F = [\![v_h]\!]_F = [\![\nabla v_h]\!]_F = 0 \text{ for } F \in \mathscr{E}_h^i,$$

and, as $\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0$ on $\partial\Omega$, it follows that

$$[\![\nabla v_h]\!]_F = 0 \text{ for } F \in \mathscr{E}_h^b.$$

It then follows that v_h is affine, i.e, $v_h = a^T x + b$, with $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and that μ is constant. But then we see that

$$0 = \nabla v_h|_F = a,$$

for $F \in \mathscr{E}_h^b$, and thus $a = 0$, i.e., $v_h = b$. Then, since $v_h \in V_{h,p,0}$, $0 = \int_\Omega v_h = |\Omega| b$, and so $b = 0$, i.e., $v_h \equiv 0$.

Finally, we see that $J_h^{\mathcal{O}}((v_h,\mu),(v_h,\mu)) = 0$, and it follows that

$$0 = \beta \cdot \nabla v_h = \mu \text{ on } \partial\Omega,$$

and so $\mu \equiv 0$. Overall, we have obtained $(v_h,\mu) \equiv (0,0)$. $\square$

Lemma 6.8.2 *Let* $\Omega \subset \mathbb{R}^2$ *be a* C^2 *and piecewise* C^3 *domain, and let* $\beta \in C^1(\partial\Omega;\mathbb{S}^1)$. *Assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Furthermore, assume that $\{\mathscr{T}_h\}_h$ *is a regular of order* 2 *family of triangulations on* $\overline{\Omega}$ *satisfying Assumption 4.4.9. Then, for each constant* $\kappa > 1$, *there exist positive constants* c_{stab} *and* c_*, *independent of* h, *and* θ, *such that*

$$B_{h,\theta}^{\mathcal{O}}((u_h,\lambda),(u_h,\lambda)) \geq \kappa^{-1}\|(u_h,\lambda)\|_{h,\theta}^2 \ \ \forall (u_h,\lambda) \in M_h, \forall \theta \in [0,1], \tag{6.8.2}$$

whenever

$$\mu_F \geq \frac{c_{\text{stab}}}{\tilde{h}_F}, \sigma_F \geq \frac{c_{\text{stab}}}{\tilde{h}_F} \text{ and } \eta_F, \ell_F > 0. \tag{6.8.3}$$

Proof: We see that for $(u_h, \lambda) \in M_h$,

$$
\begin{aligned}
B^{\mathcal{O}}_{h,\theta}((u_h,\lambda),(u_h,\lambda)) &= \sum_{K\in\mathscr{T}_h} [\theta\langle D^2 u_h, D^2 u_h\rangle_K + (1-\theta)\langle \Delta u_h, \Delta u_h\rangle_K] \\
&+ 2\theta \sum_{F\in\mathscr{E}_h^i} [\langle \mathrm{div}_{\mathbf{T}}\, \nabla_{\mathbf{T}} \langle\!\langle u_h \rangle\!\rangle, [\![\nabla u_h \cdot n_F]\!]\rangle_F - \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla u_h \cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}} u_h]\!]\rangle_F] \\
&+ \theta \sum_{F\in\mathscr{E}_h^b} \left[\left\| (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2} \nabla u_h \right\|^2_{L^2(F)} \right] \\
&+ 2\theta \sum_{F\in\mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u_h), \beta\cdot\nabla u_h - \lambda\rangle_F + \sum_{F\in\mathscr{E}_h^i} \mu_F [\|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} \|[\![\nabla u_h \cdot n_F]\!]\|^2_{L^2(F)}] \\
&+ \sum_{F\in\mathscr{E}_h^i} [\eta_F \|[\![u_h]\!]\|^2_{L^2(F)} + \ell_F \|[\![\lambda]\!]\|^2_{L^2(F)}] + \sum_{F\in\mathscr{E}_h^b} \sigma_F \|\beta\cdot\nabla u_h - \lambda\|^2_{L^2(F)}.
\end{aligned}
$$

By (5.6.4), we have, for any $\delta > 0$,

$$
\begin{aligned}
I_1 := \left| 2 \sum_{F\in\mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}}\, \nabla_{\mathbf{T}} \langle\!\langle u_h \rangle\!\rangle, [\![\nabla u_h \cdot n_F]\!]\rangle_F \right| &\le \delta^{-1} \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![\nabla u_h \cdot n_F]\!]\|^2_{L^2(F)} \\
&+ \delta C C(d) \left(\sum_{K\in\mathscr{T}_h^c} |u_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |u_h|^2_{H^2(K)} \right).
\end{aligned} \tag{6.8.4}
$$

Similarly, we obtain (noting that the sum in the sequel is over internal edges)

$$
\begin{aligned}
I_2 := \left| 2 \sum_{F\in\mathscr{E}_h^i} \langle \nabla_{\mathbf{T}} \langle\!\langle \nabla u_h \cdot n_F\rangle\!\rangle_F, [\![\nabla_{\mathbf{T}} u_h]\!]_F\rangle_F \right| &\le \delta^{-1} \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} \\
&+ \delta C C(d) \left(\sum_{K\in\mathscr{T}_h^c} |u_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |u_h|^2_{H^2(K)} \right).
\end{aligned} \tag{6.8.5}
$$

We now apply (4.7.13), and noting that on F, $\nabla u_h = \nabla_{\mathbf{T}} u_h + (\nabla u_h \cdot n_F) n_F$, obtaining

$$
\begin{aligned}
\sum_{K\in\mathscr{T}_h^c} |u_h|^2_{H_1(K)} &\le C \left(|u_h|^2_{H_2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} \|[\![\nabla u_h]\!]\|^2_{L^2(F)} + \frac{1}{|\partial\Omega|} \sum_{F\in\mathscr{E}_h^b} \|\nabla u_h\|^2_{L^2(F)} \right) \\
&\le C \left(|u_h|^2_{H_2(\Omega;\mathscr{T}_h)} + \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} (\|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} + \|[\![\nabla u_h \cdot n_F]\!]\|^2_{L^2(F)}) \right. \\
&\qquad \left. + \frac{1}{|\partial\Omega|} \sum_{F\in\mathscr{E}_h^b} \|\nabla u_h\|^2_{L^2(F)} \right).
\end{aligned} \tag{6.8.6}
$$

Applying the above estimate to (6.8.4) and (6.8.5), and summing the result, we obtain

$$I_1 + I_2 \leq \delta CC(d) \left(|u_h|^2_{H^2(\Omega;\mathscr{T}_h)} + \frac{1}{|\partial\Omega|} \sum_{F\in\mathscr{E}_h^b} \|\nabla u_h\|^2_{L^2(F)} \right) \\ + (\delta CC(d) + \delta^{-1}) \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} (\|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} + \|[\![\nabla u_h \cdot n_F]\!]\|^2_{L^2(F)})$$

Since $\beta^\perp \in C^1(\partial\Omega;\mathbb{S}^1)$, we also see that for any $\delta > 0$,

$$\begin{aligned} I_3 &:= \sum_{F\in\mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u_h), \beta \cdot \nabla u_h - \lambda \rangle_F \\ &\leq C \sum_{F\in\mathscr{E}_h^b} (\|\nabla u_h\|_{L^2(F)} + \|D^2 u_h\|_{L^2(F)}) \|\beta \cdot \nabla u_h - \lambda\|_{L^2(F)} \\ &\leq \sum_{F\in\mathscr{E}_h^b} \left[\frac{1}{2\delta}(1 + 1/\tilde{h}_F)\|\beta \cdot \nabla u_h - \lambda\|^2_{L^2(F)} \right. \\ &\qquad \left. + \frac{\delta}{2} \left(\|\nabla u_h\|^2_{L^2(F)} + \sum_{K\in\mathscr{T}_h : F\subset\partial K} \tilde{h}_F \|D^2 u_h\|^2_{L^2(\partial K)} \right) \right] \\ &\leq \sum_{F\in\mathscr{E}_h^b} \frac{1}{\delta\tilde{h}_F} \|\beta \cdot \nabla u_h - \lambda\|^2_{L^2(F)} + \frac{\delta}{2}\|\nabla u_h\|^2_{L^2(F)} \\ &\qquad + \frac{CC(d)\delta}{2} \left(\sum_{K\in\mathscr{T}_h^c} |u_h|^2_{H^2_*(K)} + \sum_{K\in\mathscr{T}_h^f} |u_h|^2_{H^2(K)} \right) \\ &\leq \sum_{F\in\mathscr{E}_h^b} \frac{1}{\delta\tilde{h}_F} \|\beta \cdot \nabla u_h - \lambda\|^2_{L^2(F)} + \frac{\delta(1 + CC(d))}{2} \|\nabla u_h\|^2_{L^2(F)} + \delta CC(d) |u_h|^2_{H^2(\Omega;\mathscr{T}_h)} \\ &\qquad + \delta CC(d) \sum_{F\in\mathscr{E}_h^i} \tilde{h}_F^{-1} (\|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} + \|[\![\nabla u_h \cdot n_F]\!]\|^2_{L^2(F)}) \end{aligned} \tag{6.8.7}$$

Note that the penultimate inequality above follows from an application of the trace estimate (4.6.1) and inverse estimate (4.6.26), and the final inequality follows from (6.8.6).

Now, notice that for any $F \in \mathscr{E}_h^b$,

$$\begin{aligned} \|\nabla u\|^2_{L^2(F)} &= \int_F \frac{1}{\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F} ((\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2} |\nabla u|)^2 \\ &\leq (\min_{F\in\mathscr{E}_h^b} \inf_F (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F))^{-1} \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2} \nabla u\|^2_{L^2(F)}, \end{aligned}$$

and let us denote

$$\Theta_* := (\min_{F\in\mathscr{E}_h^b} \inf_F (\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F))^{-1}.$$

Applying the above estimate to (6.8.7), we obtain

$$I_3 \leq \sum_{F\in\mathscr{E}_h^b} \frac{1}{\delta\tilde{h}_F}\|\beta\cdot\nabla u_h - \lambda\|^2_{L^2(F)} + \frac{\delta\Theta_*(1+CC(d))}{2}\|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}$$
$$+ \delta CC(d)\sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}(\|[\![\nabla_{\mathbf{T}}u_h]\!]\|^2_{L^2(F)} + \|[\![\nabla u_h\cdot n_F]\!]\|^2_{L^2(F)}) + \delta CC(d)|u_h|^2_{H^2(\Omega;\mathscr{T}_h)}.$$

Our bounds for I_1, I_2 and I_3 now give us

$$B^{\mathcal{O}}_{h,\theta}((u_h,\lambda),(u_h,\lambda)) \geq \sum_{i=1}^{7} A_i,$$

where

$$A_1 := \theta(1-2\delta CC(d))\sum_{K\in\mathscr{T}_h}\|D^2u_h\|^2_{L^2(K)},\ A_2 := (1-\theta)\sum_{K\in\mathscr{T}_h}\|\Delta u_h\|^2_{L^2(K)},$$
$$A_3 := \sum_{F\in\mathscr{E}_h^i}\left(\mu_F - \frac{2\theta(\delta CC(d)+\delta^{-1})}{\tilde{h}_F}\right)\|[\![\nabla u_h\cdot n_F]\!]\|^2_{L^2(F)},$$
$$A_4 := \sum_{F\in\mathscr{E}_h^i}\eta_F\|[\![u_h]\!]\|^2_{L^2(F)} + \ell_F\|[\![\lambda]\!]\|^2_{L^2(F)},$$
$$A_5 := \sum_{F\in\mathscr{E}_h^i}\left(\mu_F - \frac{2\theta(\delta CC(d)+\delta^{-1})}{\tilde{h}_F}\right)\|[\![\nabla_{\mathbf{T}}u_h]\!]\|^2_{L^2(F)},$$
$$A_6 := \sum_{F\in\mathscr{E}_h^b}\left(\sigma_F - \frac{\theta}{\delta\tilde{h}_F}\right)\|\beta\cdot\nabla u_h - \lambda\|^2_{L^2(F)},$$
$$A_7 := \theta\left(1 - \frac{\delta\Theta_*(1+CC(d))}{2}\right)\sum_{F\in\mathscr{E}_h^b}\|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}.$$

Now let $\kappa > 1$ be given. Then, since $\kappa^{-1} < 1$, there exists an $\delta > 0$ sufficiently small such that

$$1 - 2\delta CC(d) > \kappa^{-1},\ \delta\Theta_*(1+CC(d)) < 1,\ \text{and}\ \delta CC(d) < \delta^{-1},$$

we then choose $c_{\text{stab}} = 4/\delta$, $c_* = \kappa/2$ and note that by assumption, $\mu_F \geq c_{\text{stab}}/\tilde{h}_F$ and $\sigma_F \geq c_{\text{stab}}/\tilde{h}_F$. Therefore, for any $\theta\in[0,1]$,

$$A_1 \geq \theta\kappa^{-1}|u_h|^2_{H^2(\Omega;\mathscr{T}_h)},\ A_2 \geq (1-\theta)\kappa^{-1}\sum_{K\in\mathscr{T}_h}\|\Delta u_h\|^2_{L^2(K)},$$
$$A_3 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^i}\mu_F\|[\![\nabla u_h\cdot n_F]\!]\|^2_{L^2(F)} = \kappa^{-1}c_*\sum_{F\in\mathscr{E}_h^i}\|[\![\nabla u_h\cdot n_F]\!]\|^2_{L^2(F)},$$

$$A_4 \geq \frac{1}{2}A_4 = \kappa^{-1}c_* \sum_{F\in\mathscr{E}_h^i} \eta_F \|[\![u_h]\!]\|^2_{L^2(F)} + \ell_F \|[\![\lambda]\!]\|^2_{L^2(F)},$$

$$A_5 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^i} \mu_F \|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)} = \kappa^{-1}c_* \sum_{F\in\mathscr{E}_h^i} \mu_F \|[\![\nabla_{\mathbf{T}} u_h]\!]\|^2_{L^2(F)}$$

$$A_6 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^b} \sigma_F \|\beta\cdot\nabla u_h - \lambda\|^2_{L^2(F)} = \kappa^{-1}c_* \sum_{F\in\mathscr{E}_h^b} \sigma_F \|\beta\cdot\nabla u_h - \lambda\|^2_{L^2(F)},$$

$$A_7 \geq \frac{1}{2}\sum_{F\in\mathscr{E}_h^b} \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}.$$

Thus, we obtain

$$\kappa B^{\mathcal{O}}_{h,\theta}((u_h,\lambda),(u_h,\lambda)) \geq \sum_{K\in\mathscr{T}_h} [\theta\|D^2 u_h\|^2_{L^2(K)} + (1-\theta)\|\Delta u_h\|^2_{L^2(K)}]$$
$$+ c_* J^{\mathcal{O}}_h((u_h,\lambda),(u_h,\lambda)) + \frac{\theta}{2}\sum_{F\in\mathscr{E}_h^b} \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}. \qquad \square$$

We will now prove that $A^{\mathcal{O}}_h$ is coercive in $\|\cdot\|_{h,1}$.

Theorem 6.8.3 *Under the assumptions of Lemma 6.8.2, let c_{stab} and c_*, μ_F, η_F, and σ_F be chosen so that (6.8.2) and (6.8.3) hold with $\kappa < (1-\varepsilon)^{-1/2}$. Let the operator L be uniformly elliptic (and thus satisfy the Cordes condition (3.3.12)). Then, the operator $A^{\mathcal{O}}_h$ is coercive in $\|\cdot\|_{h,1}$. In particular, for any $(v_h,\mu)\in M_h$, we have*

$$\|(v_h,\mu)\|^2_{h,1} \leq \frac{2\kappa}{1-\kappa^2(1-\varepsilon)} A^{\mathcal{O}}_h((v_h,\mu),(v_h,\mu)). \tag{6.8.8}$$

Therefore, there exists a unique solution pair $(u_h,c_h)\in M_h$ of (6.6.5). Moreover, the pair (u_h,c_h) satisfies

$$\|(u_h,c_h)\|_{h,1} \leq \frac{2\sqrt{2}\kappa\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa^2(1-\varepsilon)}\|f\|_{L^2(\Omega)}. \tag{6.8.9}$$

Proof: Let $(v_h,\mu)\in M_h$, then we have that, for any $K\in\mathscr{T}_h$:

$$\langle\gamma L v_h - \Delta v_h, \Delta v_h\rangle_K \leq \|(\gamma L-\Delta)v_h\|_{L^2(K)}\|\Delta v_h\|_{L^2(K)}$$
$$\leq \sqrt{1-\varepsilon}\|D^2 v_h\|_{L^2(K)}\|\Delta v_h\|_{L^2(K)}.$$

The Cauchy–Schwarz inequality with a parameter and (6.8.2) then give us

$$A^{\mathcal{O}}_h((v_h,\mu)) = B^{\mathcal{O}}_{h,1/2}((v_h,\mu)) + \sum_{K\in\mathscr{T}_h}\langle\gamma L v_h - \Delta v_h,\Delta v_h\rangle_K + \|\Delta v_h\|^2_{L^2(K)}$$
$$\geq \kappa^{-1}\|(v_h,\mu)\|^2_{h,1/2} + \sum_{K\in\mathscr{T}_h}[\|\Delta v_h\|^2_{L^2(K)} - \sqrt{1-\varepsilon}\|D^2 v_h\|_{L^2(K)}\|\Delta v_h\|_{L^2(K)}]$$

$$
\begin{aligned}
&\geq \kappa^{-1}\left(\sum_{K\in\mathscr{T}_h}\frac{1}{2}\|D^2v_h\|^2_{L^2(K)}+\frac{1}{2}\|\Delta v_h\|^2_{L^2(K)}\right.\\
&\qquad\left.+\,c_*J_h^{\mathcal{O}}((v_h,\mu),(v_h,\mu))+\frac{1}{2}\sum_{F\in\mathscr{E}_h^b}\|(\partial_{\mathbf{T}_2}\Theta+\mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}\right)\\
&\qquad\qquad-\sum_{K\in\mathscr{T}_h}\left[\frac{\kappa(1-\varepsilon)}{2}\|D^2v_h\|^2_{L^2(K)}+\frac{\kappa^{-1}}{2}\|\Delta v_h\|^2_{L^2(K)}\right]\\
&=\sum_{K\in\mathscr{T}_h}\frac{\kappa^{-1}}{2}\|D^2v_h\|^2_{L^2(K)}-\frac{\kappa(1-\varepsilon)}{2}\|D^2v_h\|^2_{L^2(K)}\\
&\qquad+\kappa^{-1}\left(c_*J_h^{\mathcal{O}}((v_h,\mu),(v_h,\mu))+\frac{1}{2}\sum_{F\in\mathscr{E}_h^b}\|(\partial_{\mathbf{T}_2}\Theta+\mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}\right)\\
&\geq\frac{1-\kappa^2(1-\varepsilon)}{2\kappa}\left(\sum_{K\in\mathscr{T}_h}\|D^2v_h\|^2_{L^2(K)}+c_*J_h^{\mathcal{O}}((v_h,\mu),(v_h,\mu))\right.\\
&\qquad\left.+\sum_{F\in\mathscr{E}_h^b}\|(\partial_{\mathbf{T}_2}\Theta+\mathcal{H}_F)^{1/2}\nabla u_h\|^2_{L^2(F)}\right)\\
&=\frac{1-\kappa^2(1-\varepsilon)}{2\kappa}\|(v_h,\mu)\|^2_{h,1}.
\end{aligned}
$$

Thus, we obtain

$$
\|(v_h,\mu)\|^2_{h,1}\leq\frac{2\kappa}{1-\kappa^2(1-\varepsilon)}A_h^{\mathcal{O}}((v_h,\mu),(v_h,\mu)). \tag{6.8.10}
$$

By Lemma 6.8.1 $\|\cdot\|_{h,1}$ is a norm on M_h. It then follows that there exists a unique pair $(u_h,c_h)\in V_{h,p,0}\times V_{h,0}$ such that

$$
A_h^{\mathcal{O}}((u_h,c_h),(v_h,\mu))=\sum_{K\in\mathscr{T}_h}\langle\gamma f,\Delta v_h\rangle_K\qquad\forall(v_h,\mu)\in M_h.
$$

Finally, taking $(v_h,\mu)=(u_h,c_h)$ in (6.8.10) gives us:

$$
\begin{aligned}
\|(u_h,c_h)\|^2_{h,1}&\leq\frac{2\kappa}{1-\kappa^2(1-\varepsilon)}A_h^{\mathcal{O}}((u_h,c_h),(u_h,c_h))\\
&=\frac{2\kappa}{1-\kappa^2(1-\varepsilon)}\sum_{K\in\mathscr{T}_h}\langle\gamma f,\Delta u_h\rangle_K\\
&\leq\frac{2\kappa\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa^2(1-\varepsilon)}\sum_{K\in\mathscr{T}_h}\|f\|_{L^2(K)}\|\Delta u_h\|_{L^2(K)}\\
&\leq\frac{2\sqrt{2}\kappa\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa^2(1-\varepsilon)}\|f\|_{L^2(\Omega)}\|(u_h,c_h)\|_{h,1};
\end{aligned}
$$

note that the factor of $\sqrt{2}$ comes from the fact that $\|\Delta u_h\|_{L^2(K)}\leq\sqrt{2}\|D^2u_h\|_{L^2(K)}$ for $K\in\mathscr{T}_h$. Dividing through by $\|(u_h,c_h)\|_{h,1}$, we obtain (6.8.9). $\square$

6.9 Error analysis

Theorem 6.9.1 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 and piecewise C^{m+1} domain, $m \in \mathbb{N}$, $m \geq 2$, and let $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Furthermore, assume that $\{\mathscr{T}_h\}_h$ is a regular of order m family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9.

Let $(u, c) \in H^2_{\beta,0}(\Omega) \times \mathbb{R}$ be the unique strong solution of (5.1.1). Assume that $u \in H^{\mathbf{s}}(\Omega; \mathscr{T}_h)$ with $s_K > 5/2$ for all $K \in \mathscr{T}_h$. Let c_{stab}, c_*, μ_F and σ_F be chosen as in Theorem 6.8.3, and choose $\eta_F \lesssim 1/\tilde{h}_F^3$, $\sigma_F \lesssim 1/\tilde{h}_F$, $F \in \mathscr{E}_h^{i,b}$ and $\tilde{h}_F^{1-\alpha} \lesssim \ell_F$ for all $F \in \mathscr{E}_h^i$, for some $\alpha > 2$. Then, there exists a constant $C > 0$, independent of h, and u, but depending on $\max_K s_K$, such that

$$\|(u - u_h, c - c_h)\|_{h,1} \leq C \left(\left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2} + \frac{\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F \in \mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega; \mathscr{T}_h)} \right), \tag{6.9.1}$$

where $t_K := \min\{p + 1, s_K, m + 1\}$. In the case that the family of triangulations is quasiuniform, this becomes

$$\|(u - u_h, c - c_h)\|_{h,1} \leq C(h^{2t_K - 4} + h^{(\alpha-2)/2}) \|u\|_{H^{\mathbf{s}}(\Omega; \mathscr{T}_h)}.$$

Proof: Let us take $z_h \in V_{h,p}$, and define $\psi_h := z_h - u_h$, $\xi_h := u - z_h$, and $\mu_h := c - c_h$. Then, we see that

$$\begin{aligned} \|(u - u_h, c - c_h)\|_{h,1} &= \|(\xi_h + \psi_h, \mu_h)\|_{h,1} \\ &= \|(\xi_h, 0) + (\psi_h, \mu_h)\|_{h,1} \leq \|(\xi_h, 0)\|_{h,1} + \|(\psi_h, \mu_h)\|_{h,1}. \end{aligned} \tag{6.9.2}$$

As in the proof of Theorem 5.7.1, we require the existence of a $z_h \in V_{h,p}$ that satisfies (4.6.9). Due to our assumptions upon the parameters μ_F, η_F, and σ_F, by applying the estimates in (4.6.9), we obtain

$$\|(\xi_h, 0)\|_{h,1} \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2},$$

thus, by (6.9.2), it is sufficient to obtain the following estimate:

$$\|(\psi_h, \mu_h)\|_{h,1} \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2} + \frac{\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F \in \mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega; \mathscr{T}_h)}. \tag{6.9.3}$$

Now, applying the coercivity result from Theorem 6.8.3, and noting that c is a constant, we can apply the consistency result (6.7.1), obtaining

$$\begin{aligned}
\|(\psi_h,\mu_h)\|_{h,1}^2 &= \|(z_h-u_h,c-c_h)\|_{h,1}^2\\
&\lesssim A_h^{\mathcal{O}}((z_h-u_h,c-c_h),(z_h-u_h,c-c_h))\\
&= A_h^{\mathcal{O}}((z_h,c),(z_h-u_h,c-c_h)) - A_h^{\mathcal{O}}((u_h,c_h),(z_h-u_h,c-c_h))\\
&= A_h^{\mathcal{O}}((z_h,c),(z_h-u_h,c-c_h)) - \sum_{K\in\mathscr{T}_h}\langle \gamma f,\Delta(z_h-u_h)\rangle_K\\
&= A_h^{\mathcal{O}}((z_h,c),(z_h-u_h,c-c_h)) - A_h^{\mathcal{O}}((u,c),(z_h-u_h,c))\\
&= A_h^{\mathcal{O}}((z_h,c),(z_h-u_h,c-c_h)) - A_h^{\mathcal{O}}((u,c),(z_h-u_h,c-c_h)+(0,c_h))\\
&= A_h^{\mathcal{O}}((z_h-u,0),(z_h-u_h,c-c_h)) - A_h^{\mathcal{O}}((u,c),(0,c_h))\\
&= A_h^{\mathcal{O}}((\xi_h,0),(\psi_h,\mu_h)) - A_h^{\mathcal{O}}((u,c),(0,c_h)).
\end{aligned}$$

From this, we obtain

$$\|(\psi_h,\mu_h)\|_{h,1}^2 \lesssim \sum_{i=1}^{6} A_i, \tag{6.9.4}$$

where

$$\begin{aligned}
A_1 &:= \sum_{K\in\mathscr{T}_h}\langle D^2\xi_h, D^2\psi_h\rangle_K,\ A_2 := \sum_{K\in\mathscr{T}_h}\langle(\gamma L-\Delta)\xi_h,\Delta\psi_h\rangle_K,\\
A_3 &:= \sum_{K\in\mathscr{T}_h}\frac{1}{2}\langle\Delta\xi_h,\Delta\psi_h\rangle_K,\ A_4 := \frac{1}{2}B_{h,*}^{\mathcal{O}}((\xi_h,0),(\psi_h,\mu_h)),\\
A_5 &:= J_h^{\mathcal{O}}((\xi_h,0),(\psi_h,\mu_h)),\ A_6 := -B_{h,1/2}^{\mathcal{O}}((u,c),(0,c_h)).
\end{aligned}$$

We see that

$$|A_1|,|A_2|,|A_3| \lesssim |\xi_h|_{H^2(\Omega;\mathscr{T}_h)}|\psi_h|_{H^2(\Omega;\mathscr{T}_h)} \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{\frac{1}{2}}\|(\psi_h,\mu_h)\|_{h,1}, \tag{6.9.5}$$

and

$$\begin{aligned}
|A_5| &\le J_h^{\mathcal{O}}((\xi_h,0),(\xi_h,0))^{1/2}J_h^{\mathcal{O}}((\psi_h,\mu_h),(\psi_h,\mu_h))^{1/2}\\
&\le J_h^{\mathcal{O}}((\xi_h,0),(\xi_h,0))^{1/2}\|(\psi_h,\mu_h)\|_{h,1}.
\end{aligned} \tag{6.9.6}$$

Applying the first estimate in (4.6.9) to the estimates in (6.9.5), we obtain

$$|A_1|,|A_2|,|A_3| \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2}\|(\psi_h,\mu_h)\|_{h,1}.$$

We also see that

$$J_h^{\mathcal{O}}((\xi_h,0),(\xi_h,0))^{1/2} \lesssim (e_1+e_2+e_3)^{1/2},$$

where, based on the assumption that $\mu_F, \sigma_F \lesssim 1/\tilde{h}_F, \eta_F \lesssim 1/\tilde{h}_F^3$,

$$e_1 := \sum_{F\in\mathscr{E}_h^i} \mu_F[\|[\![\nabla\xi_h\cdot n_F]\!]\|^2_{L^2(F)} + \|[\![\nabla_{\mathbf{T}}\xi_h]\!]\|^2_{L^2(F)}] \lesssim \sum_{K\in\mathscr{T}_h} \frac{1}{\tilde{h}_F}\|\nabla\xi_h\|^2_{L^2(\partial K)},$$
$$\lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}$$

$$e_2 \lesssim \sum_{F\in\mathscr{E}_h^i} \frac{1}{\tilde{h}_F^3}\|[\![\xi_h]\!]\|^2_{L^2(F)} \lesssim \sum_{K\in\mathscr{T}_h} \frac{1}{\tilde{h}_F^3}\|\xi_h\|^2_{L^2(\partial K)} \lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)},$$

and

$$e_3 \lesssim \sum_{F\in\mathscr{E}_h^b} \frac{1}{\tilde{h}_F}\|\nabla\xi_h\|^2_{L^2(F)} \lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}.$$

Thus

$$J_h^{\mathcal{O}}((\xi_h,0),(\xi_h,0)) \lesssim \sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}, \tag{6.9.7}$$

and so

$$|A_5| \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{\frac{1}{2}} \|(\psi_h,\mu_h)\|_{h,1}.$$

Now we must obtain a bound for $A_4 = \frac{1}{2}B^{\mathcal{O}}_{h,*}((\xi_h,0),(\psi_h,\mu_h))$. One can see that

$$B^{\mathcal{O}}_{h,*}((\xi_h,0),(\psi_h,\mu_h)) = \sum_{i=1}^{8} I_i,$$

where

$$I_1 := \sum_{K\in\mathscr{T}_h} \langle D^2\xi_h, D^2\psi_h\rangle_K,\ I_2 := \sum_{F\in\mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle\xi_h\rangle\!\rangle, [\![\nabla\psi_h\cdot n_F]\!]\rangle_F$$
$$I_3 := \sum_{F\in\mathscr{E}_h^i} \langle \mathrm{div}_{\mathbf{T}}\,\nabla_{\mathbf{T}}\langle\!\langle\psi_h\rangle\!\rangle, [\![\nabla\xi_h\cdot n_F]\!]\rangle_F,\ I_4 := -\sum_{F\in\mathscr{E}_h^i} \langle \nabla_{\mathbf{T}}\langle\!\langle\nabla\xi_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}\psi_h]\!]\rangle_F$$
$$I_5 := -\sum_{F\in\mathscr{E}_h^i} \langle \nabla_{\mathbf{T}}\langle\!\langle\nabla\psi_h\cdot n_F\rangle\!\rangle, [\![\nabla_{\mathbf{T}}\xi_h]\!]\rangle_F,\ I_6 := \sum_{F\in\mathscr{E}_h^b} \langle(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)\nabla\xi_h, \nabla\psi_h\rangle_F,$$
$$I_7 := \sum_{F\in\mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\xi_h), \beta\cdot\nabla\psi_h - \mu_h\rangle_F,\ I_8 := \sum_{F\in\mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\psi_h), \beta\cdot\nabla\xi_h\rangle_F.$$

We have that

$$I_1 \lesssim |\xi_h|_{H^2(\Omega;\mathscr{T}_h)}|\psi_h|_{H^2(\Omega;\mathscr{T}_h)} \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{\frac{1}{2}} \|(\psi_h,\mu_h)\|_{h,1}.$$

Our approach for bounding $I_2, \ldots, I_4$, is similar to that of obtaining the estimates for $A_2, \ldots, A_5$ in the proof of Theorem 5.7.1. We find that

$$\begin{aligned} I_2, I_4 &\lesssim \left(\sum_{K \in \mathscr{T}_h} \tilde{h}_F \|D^2 \xi_h\|^2_{L^2(\partial K)} + \|\nabla \xi_h\|^2_{L^2(\partial K)} \right)^{\frac{1}{2}} J_h^{\mathcal{O}}((\psi_h, \mu_h), (\psi_h, \mu_h))^{\frac{1}{2}} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1}. \end{aligned} \tag{6.9.8}$$

From (6.9.7), and Corollary 4.7.6 we obtain

$$\begin{aligned} I_3, I_5 &\lesssim J_h^{\mathcal{O}}((\xi_h, 0), (\xi_h, 0))^{\frac{1}{2}} \left(\sum_{K \in \mathscr{T}_h} |\psi_h|^2_{H^2_*(K)} \right)^{\frac{1}{2}} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1}. \end{aligned}$$

Now, for $I_6, \ldots, I_8$,

$$\begin{aligned} I_6 &\lesssim \left(\sum_{F \in \mathscr{E}_h^b} \|\nabla \xi_h\|^2_{L^2(F)} \right)^{1/2} \left(\sum_{F \in \mathscr{E}_h^b} \|(\partial_{\mathbf{T}_2} \Theta + \mathcal{H}_F)^{1/2} \nabla \psi_h\|^2_{L^2(F)} \right)^{1/2} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1}, \end{aligned}$$

Furthermore,

$$\begin{aligned} I_7 &\lesssim \left(\sum_{F \in \mathscr{E}_h^b} \tilde{h}_F \|D^2 \xi_h\|^2_{L^2(F)} + \|\nabla \xi_h\|^2_{L^2(F)} \right)^{\frac{1}{2}} \left(\sum_{F \in \mathscr{E}_h^b} (\tilde{h}_F^{-1} + 1) \|\beta \cdot \nabla \psi_h - \mu_h\|^2_{L^2(F)} \right)^{\frac{1}{2}} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1} \end{aligned}$$

and

$$\begin{aligned} I_8 &:= \sum_{F \in \mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2} (\beta^\perp \cdot \nabla \psi_h), \beta \cdot \nabla \xi_h \rangle_F \\ &\lesssim \left(\sum_{F \in \mathscr{E}_h^b} \frac{1}{\tilde{h}_F} \|\nabla \xi_h\|^2_{L^2(F)} \right)^{1/2} \left(\sum_{F \in \mathscr{E}_h^b} \tilde{h}_F \|\partial_{\mathbf{T}_2} (\beta^\perp \cdot \nabla \psi_h)\|^2_{L^2(F)} \right)^{1/2} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1}, \end{aligned}$$

note that obtaining the final inequality in the estimate for I_8 is analogous to (6.8.7). We now see that

$$|A_4| \leq \sum_{i=1}^{8} I_i \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|_{H^{s_K}(K)}^2 \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1}.$$

By (6.9.4), it then follows that

$$\|(\psi_h, \mu_h)\|_{h,1}^2 \leq \sum_{i=1}^{6} |A_i| \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|_{H^{s_K}(K)}^2 \right)^{\frac{1}{2}} \|(\psi_h, \mu_h)\|_{h,1} + A_6. \tag{6.9.9}$$

Furthermore, for A_6, we see that

$$\begin{aligned}
A_6 &= -B_{h,1/2}^{\mathcal{O}}((u,c),(0,c_h)) \\
&= \sum_{F \in \mathscr{E}_h^b} \left[\frac{1}{2} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u), c_h \rangle_F + \sigma_F \langle \beta \cdot \nabla u - c, c_h \rangle_F \right] + \sum_{F \in \mathscr{E}_h^i} \ell_F \langle [\![c]\!], [\![c_h]\!] \rangle_F \\
&= \frac{1}{2} \sum_{F \in \mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u), c_h \rangle_F \\
&= \frac{1}{2} \sum_{F \in \mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u), c_h - M \rangle_F,
\end{aligned}$$

for any constant function M, where the penultimate equality holds due to the fact that $\beta \cdot \nabla u - c|_F = 0$ for all $F \in \mathscr{E}_h^b$, and as c is constant, it cannot jump across internal edges, and the final equality holds due to the following argument, valid for an arbitrary constant function M. Integrating by parts gives us

$$\begin{aligned}
\sum_{F \in \mathscr{E}_h^b} \langle \partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u), M \rangle_F &= \sum_{e \in \mathscr{V}_h^b} [\![M \beta^\perp \cdot \nabla u]\!]_e \\
&= \sum_{e \in \mathscr{V}_h^b} M([\![\beta^\perp]\!]_e \cdot \langle\!\langle \nabla u \rangle\!\rangle_e + \langle\!\langle \beta^\perp \rangle\!\rangle_e \cdot [\![\nabla u]\!]_e) = 0,
\end{aligned}$$

where the final equality holds, due to the fact that $\beta^\perp \in C^1(\partial\Omega; \mathbb{S}^1)$, and so cannot jump across vertices, furthermore, $\nabla u \in H^{1/2}(\partial\Omega)$, and thus, since $\partial\Omega$ is a one-dimensional hypersurface, neither can ∇u.

By the Cauchy–Schwarz inequality with a parameter, we see that for any $\delta > 0$

$$\begin{aligned}
A_6 &\leq \frac{1}{2} \left(\sum_{F \in \mathscr{E}_h^b} (\delta \tilde{h}_F)^{-1} \|\partial_{\mathbf{T}_2}(\beta^\perp \cdot \nabla u)\|_{L^2(F)}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathscr{E}_h^b} \delta \tilde{h}_F \|c_h - M\|_{L^2(F)}^2 \right)^{\frac{1}{2}} \\
&\lesssim \left[\sum_{F \in \mathscr{E}_h^b} (\delta \tilde{h}_F)^{-1} (\|\nabla u\|_{L^2(F)}^2 + \|D^2 u\|_{L^2(F)}^2) \right]^{\frac{1}{2}} \left[\sum_{F \in \mathscr{E}_h^b} \delta \tilde{h}_F \|(c_h - M)\|_{L^2(F)}^2 \right]^{\frac{1}{2}}
\end{aligned}$$

$$\lesssim \left[\sum_{F\in\mathscr{E}_h^b}\sum_{K\in\mathscr{T}_h:F\subset\partial K}(\delta\tilde{h}_F h_K)^{-1}\|u\|^2_{H^{s_K}(K)}\right]^{\frac{1}{2}} \times \left[\sum_{F\in\mathscr{E}_h^i}\tilde{h}_F^{-1}\delta\|[\![c_h-M]\!]\|^2_{L^2(F)}+\left|\frac{\delta}{|\partial\Omega|}\int_{\partial\Omega}(c_h-M)\right|^2\right]^{\frac{1}{2}}. \tag{6.9.10}$$

where the final inequality follows from (4.6.1) and Corollary 4.7.7. We note that M and $c=\beta\cdot\nabla u|_{\partial\Omega}$ are both constant, and so

$$[\![c_h-M]\!]=[\![c_h-c]\!],\quad F\in\mathscr{E}_h^i. \tag{6.9.11}$$

The choice of M was arbitrary, so we take

$$M:=\frac{1}{|\partial\Omega|}\int_{\partial\Omega}c_h \Rightarrow \int_{\partial\Omega}c_h-M=0. \tag{6.9.12}$$

We now choose $\delta:=\min_{F\in\mathscr{E}_h^i}\ell_F\tilde{h}_F$, and so $\delta\tilde{h}_F^{-1}\leq\ell_F$ for all $F\in\mathscr{E}_h^i$, and, furthermore, $\delta^{-1}=(\min_{F\in\mathscr{E}_h^i}\ell_F\tilde{h}_F)^{-1}=\max_{F\in\mathscr{E}_h^i}\ell_F^{-1}\tilde{h}_F^{-1}$. Applying these estimates, along with (6.9.12) to (6.9.10), we obtain

$$\begin{aligned}A_6&\lesssim\left[\sum_{F\in\mathscr{E}_h^b}\sum_{K\in\mathscr{T}_h:F\subset\partial K}(\tilde{h}_F h_K)^{-1}\max_{F\in\mathscr{E}_h^i}(\ell_F^{-1}\tilde{h}_F^{-1})\|u\|^2_{H^{s_K}(K)}\right]^{\frac{1}{2}}\|(\psi_h,\mu_h)\|_{h,1}\\&\lesssim(\max_{F\in\mathscr{E}_h^b}\tilde{h}_F^{-1})(\max_{F\in\mathscr{E}_h^i}\ell_F^{-1}\tilde{h}_F^{-1})^{1/2}\|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)}\|(\psi_h,\mu_h)\|_{1,h}\end{aligned} \tag{6.9.13}$$

Furthermore, we have assumed that $\tilde{h}_F^{1-\alpha}\lesssim\ell_F$, for some $\alpha>2$, which gives us $\max_{F\in\mathscr{E}_h^i}\ell_F^{-1}\tilde{h}_F^{-1}\lesssim\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha}$. It then follows that

$$(\max_{F\in\mathscr{E}_h^b}\tilde{h}_F^{-1})(\max_{F\in\mathscr{E}_h^i}\ell_F^{-1}\tilde{h}_F^{-1})^{1/2}\lesssim\frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha/2}}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F}.$$

Applying the above estimate to (6.9.13) yields

$$A_6\lesssim\frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha/2}}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F}\|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)}\|(\psi_h,\mu_h)\|_{1,h}, \tag{6.9.14}$$

which we apply to (6.9.9), obtaining

$$\|(\psi_h,\mu_h)\|^2_{h,1}\lesssim\left[\left(\sum_{K\in\mathscr{T}_h}h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{\frac{1}{2}}+\frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F}\|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)}\right]\|(\psi_h,\mu_h)\|_{h,1}.$$

We then divide through by $\|(\psi_h,\mu_h)\|_{h,1}$ in the above estimate, which gives us (6.9.3). This completes the proof. $\square$

6.9.1 An error estimate in the case of conforming regularity

The hypotheses of Theorem 5.7.1 includes the sufficient condition that the strong solution, u, is piecewise-sufficiently regular, so that one may substitute (u, c) into the left-hand argument of the operator, $A_h^{\mathcal{O}}$. However, the assumption, that the true solution $u \in H^s(\Omega; \mathscr{T}_h)$, $s > 5/2$ may not hold in general, particularly when we consider that the coefficient matrix $A \in L^\infty(\Omega)$.

In the following lemma, we provide an error estimate for strong solutions $u \in H^2_{\beta,0}$, i.e., the expected regularity of strong solutions implied by Theorem 6.3.1. As in estimate (6.9.1), one can see the error contribution arising from the inconsistency of c_h belonging to $V_{h,0}$ as opposed to $\mathbb{R}$. Similarly, this contribution is zero if c_h does not jump across boundary vertices. This shows that our method provides an approximation that is at least as accurate in the $\|(\cdot,\cdot)\|_{h,1}$-norm, as a H^2-conforming finite element method. Note that the design of such a conforming space would require knowledge of the compatibility constant, c, otherwise further terms (quantifying the lack of conformity in the oblique derivative) would arise in the following error estimate.

Lemma 6.9.2 *Let $\Omega \subset \mathbb{R}^2$ be a C^2 and piecewise C^3 domain, and let $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Furthermore, assume that $\{\mathscr{T}_h\}_h$ is a regular of order 2 family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9.

Let $(u, c) \in H^2_{\beta,0}(\Omega) \times \mathbb{R}$ be the unique strong solution of (5.1.1). Let c_{stab}, c_*, and μ_F be chosen as in Theorem 6.8.3, and choose $\eta_F \lesssim 1/\tilde{h}_F^3$, $\sigma_F \lesssim 1/\tilde{h}_F$, and $\tilde{h}_F^{1+\alpha} \lesssim \ell_F$, where $\alpha > 2 + p^*/2$, and $p^* := 2\,\mathrm{sgn}(p-2)$. Then, we have the following error estimate

$$\begin{aligned}\|(u-u_h, c-c_h)\|_{h,1} \lesssim &+ \left(\frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^\alpha}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F^{2+p^*}}\right)^{\frac{1}{2}} \|u\|_{H^2(\Omega)} \\ &+ \inf_{z_h\in V}\left\{\|z_h-u\|_{H^2(\Omega)} + \left[\sum_{F\in\mathscr{E}_h^b}\frac{1}{\tilde{h}_F}(\|\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h)\|^2_{L^2(F)} + \|\beta\cdot\nabla z_h - c\|^2_{L^2(F)})\right]^{\frac{1}{2}}\right\}.\end{aligned} \tag{6.9.15}$$

where $V := V_{h,p,0} \cap H^2(\Omega)$.

Proof: First we assume that $z_h \in H^2(\Omega) \cap V_{h,p,0}$. Then, we see that

$$\|(u-u_h, c-c_h)\|_{h,1} \le \|(\xi_h, 0)\|_{h,1} + \|(\psi_h, \mu_h)\|,$$

where ξ_h, ψ_h and μ_h are given as in the proof of Theorem 5.7.1. Since we only assume that z_h is in $H^2(\Omega) \cap V_{h,p,0}$, only the consistency properties of the bilinear form $A_h^{\mathcal{O}}$ that depend on the piecewise regularity and H^2-regularity of z_h hold, i.e., we may utilise (6.5.6). In particular (6.5.21) does not hold, yielding the following H^2-consistency results,

$$\begin{aligned} B_{h,*}^{\mathcal{O}}((z_h,c),(\psi_h,\mu)) &= \sum_{K\in\mathscr{T}_h} \langle \Delta z_h, \Delta\psi_h\rangle_K \\ &\quad + \sum_{F\in\mathscr{E}_h^b} [\langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\psi_h), \beta\cdot\nabla z_h - c\rangle_F - \langle \partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h), \beta^\perp\cdot\nabla\psi_h\rangle_F], \end{aligned} \tag{6.9.16}$$

for any $\mu \in \mathbb{R}$, and

$$J_h^{\mathcal{O}}((z_h,c),(\psi_h,\mu_h)) = \sum_{F\in\mathscr{E}_h^b} \sigma_F \langle \beta\cdot\nabla z_h - c, \beta\cdot\nabla\psi_h - \mu_h\rangle_F. \tag{6.9.17}$$

From (6.9.16) and (6.9.17), we obtain

$$\begin{aligned} A_h^{\mathcal{O}}((z_h,c),(\psi_h,\mu_h)) &= \sum_{K\in\mathscr{T}_h} \langle \gamma L z_h, \Delta\psi_h\rangle_K + J_h^{\mathcal{O}}((z_h,c),(\psi_h,\mu_h)) \\ &\quad + \frac{1}{2}\left(B_{h,*}^{\mathcal{O}}((z_h,c),(\psi_h,\mu_h)) - \sum_{K\in\mathscr{T}_h} \langle \Delta z_h, \Delta\psi_h\rangle_K \right) \\ &= \sum_{K\in\mathscr{T}_h} \langle \gamma L z_h, \Delta\psi_h\rangle_K + \sum_{F\in\mathscr{E}_h^b} \sigma_F \langle \beta\cdot\nabla z_h - c, \beta\cdot\nabla\psi_h - \mu_h\rangle_F \\ &\quad + \frac{1}{2}\left(B_{h,*}^{\mathcal{O}}((z_h,c),(\psi_h,c)) - \sum_{K\in\mathscr{T}_h} \langle \Delta z_h, \Delta\psi_h\rangle_K - B_{h,*}^{\mathcal{O}}((z_h,c),(0,c_h)) \right) \\ &= \sum_{K\in\mathscr{T}_h} \langle \gamma L z_h, \Delta\psi_h\rangle_K + \sum_{F\in\mathscr{E}_h^b} \sigma_F \langle \beta\cdot\nabla z_h - c, \beta\cdot\nabla\psi_h - \mu_h\rangle_F - \frac{1}{2} B_{h,*}^{\mathcal{O}}((z_h,c),(0,c_h)) \\ &\quad + \frac{1}{2}\left(\sum_{F\in\mathscr{E}_h^b} [\langle \partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\psi_h), \beta\cdot\nabla z_h - c\rangle_F - \langle \partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h), \beta^\perp\cdot\nabla\psi_h\rangle_F] \right). \end{aligned} \tag{6.9.18}$$

Then, by the coercivity estimate (6.8.8), (6.9.18), and the fact that (u_h, c_h) satis-

fies (6.6.5) it follows that

$$\begin{aligned}\|(\psi_h,\mu_h)\|_{h,1}^2 &\lesssim A_h^{\mathcal{O}}((z_h,c),(\psi_h,\mu_h)) - A_h^{\mathcal{O}}((u_h,c_h),(\psi_h,\mu_h))\\ &= \sum_{K\in\mathscr{T}_h}\langle\gamma L z_h,\Delta\psi_h\rangle_K - \sum_{K\in\mathscr{T}_h}\langle\gamma f,\Delta\psi_h\rangle_K - \frac{1}{2}B_{h,1/2}^{\mathcal{O}}((z_h,c),(0,c_h))\\ &\quad+\frac{1}{2}\left(\sum_{F\in\mathscr{E}_h^b}[\langle\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\psi_h),\beta\cdot\nabla z_h - c\rangle_F - \langle\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h),\beta^\perp\cdot\nabla\psi_h\rangle_F]\right)\\ &\quad+\sum_{F\in\mathscr{E}_h^b}\sigma_F\langle\beta\cdot\nabla z_h - c,\beta\cdot\nabla\psi_h-\mu_h\rangle_F.\end{aligned} \tag{6.9.19}$$

Firstly, we see that

$$\begin{aligned}\sum_{K\in\mathscr{T}_h}\langle\gamma L z_h,\Delta\psi_h\rangle_K - \sum_{K\in\mathscr{T}_h}\langle\gamma f,\Delta\psi_h\rangle_K &= \sum_{K\in\mathscr{T}_h}\langle\gamma A:D^2(z_h-u),\Delta\psi_h\rangle_K\\ &\lesssim |z_h-u|_{H^2(\Omega;\mathscr{T}_h)}\|(\psi_h,\mu_h)\|_{h,1}.\end{aligned} \tag{6.9.20}$$

We obtain the following estimate

$$-\frac{1}{2}B_{h,1/2}^{\mathcal{O}}((z_h,c),(0,c_h)) \lesssim \max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha/2}\left(\sum_{F\in\mathscr{E}_h^b}\sum_{\substack{K\in\mathscr{T}_h:\\ F\subset\partial K}}\tilde{h}_F^{-2}\|z_h\|_{H^3(K)}^2\right)^{\frac{1}{2}}\|(\psi_h,\mu_h)\|_{h,1}, \tag{6.9.21}$$

using the same techniques employed in the derivation of (6.9.14). Indeed the argument only utilised the piecewise H^{s_K}-regularity of u, with $s_K > 5/2$, and we have that $z_h \in H^3(\Omega;\mathscr{T}_h)$. Furthermore, we have the following bounds:

$$\sum_{F\in\mathscr{E}_h^b}\sigma_F\langle\beta\cdot\nabla z_h - c,\beta\cdot\nabla\psi_h-\mu_h\rangle_F \lesssim \left(\sum_{F\in\mathscr{E}_h^b}\sigma_F\|\beta\cdot\nabla z_h - c\|_{L^2(F)}^2\right)^{1/2}\|(\psi_h,\mu_h)\|_{h,1}, \tag{6.9.22}$$

and

$$\begin{aligned}&\sum_{F\in\mathscr{E}_h^b}\langle\partial_{\mathbf{T}_2}(\beta^\perp\cdot\nabla\psi_h),\beta\cdot\nabla z_h - c\rangle_F \lesssim\\ &\qquad\left(\sum_{F\in\mathscr{E}_h^b}\sigma_F\|\beta\cdot\nabla z_h - c\|_{L^2(F)}^2\right)^{1/2}\left(\sum_{F\in\mathscr{E}_h^b}\|\nabla\psi_h\|_{L^2(F)}^2+\tilde{h}_F\|D^2\psi_h\|_{L^2(F)}^2\right)^{1/2}\\ &\qquad\lesssim\left(\sum_{F\in\mathscr{E}_h^b}\sigma_F\|\beta\cdot\nabla z_h - c\|_{L^2(F)}^2\right)^{1/2}\|(\psi_h,\mu_h)\|_{h,1}.\end{aligned} \tag{6.9.23}$$

Note that the first bound follows directly from an application of the Cauchy–Schwarz inequality, and the definition of the $\|(\cdot,\cdot)\|_{h,1}$-norm, and the second inequality is obtained analogously to (6.8.7) in the proof of Lemma 6.8.2, utilising the trace estimate (4.6.1), and the inverse inequality (4.6.26), followed by the Poincaré–Friedrichs' inequality (4.7.13).

Again, by applying the trace estimate (4.6.1), we obtain

$$\sum_{F\in\mathscr{E}_h^b}\langle\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h),\beta^\perp\cdot\nabla\psi_h\rangle_F \lesssim \left(\sum_{F\in\mathscr{E}_h^b}\frac{1}{\tilde{h}_F}\|\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h)\|^2_{L^2(F)}\right)^{1/2}\|(\psi_h,\mu_h)\|_{h,1}. \tag{6.9.24}$$

Applying estimates (6.9.20)–(6.9.24) to (6.9.19), and noting that $\sigma_F\lesssim 1/\tilde{h}_F$, yields

$$\begin{aligned}\|(\psi_h,\mu_h)\|^2_{h,1} \lesssim \Bigg(&|z_h-u|_{H^2(\Omega)}+\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha/2}\left(\sum_{F\in\mathscr{E}_h^b}\sum_{\substack{K\in\mathscr{T}_h:\\ F\subset\partial K}}\tilde{h}_F^{-2}\|z_h\|^2_{H^3(K)}\right)^{\frac{1}{2}}\\ &+\left(\sum_{F\in\mathscr{E}_h^b}\frac{1}{\tilde{h}_F}(\|\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h)\|^2_{L^2(F)}+\|\beta\cdot\nabla z_h-c\|^2_{L^2(F)})\right)^{\frac{1}{2}}\Bigg)\|(\psi_h,\mu_h)\|_{h,1}.\end{aligned} \tag{6.9.25}$$

Now, if $p=2$, we have that $\|z_h\|_{H^3(K)}\lesssim\|z_h\|_{H^2(K)}$ for all $K\in\mathscr{T}_h$, and if $p\geq 3$, by the inverse inequality (4.6.26), we have that $\|z_h\|_{H^3(K)}\lesssim h_K^{-1}\|z_h\|_{H^2(K)}$. I.e., for any $p\geq 2$, $\|z_h\|^2_K\lesssim h_K^{-p^*}\|z_h\|^2_{H^2(K)}$, where $p^*:=2\,\mathrm{sgn}(p-2)$. Furthermore, we see that

$$\begin{aligned}\sum_{F\in\mathscr{E}_h^b}\sum_{\substack{K\in\mathscr{T}_h:\\ F\subset\partial K}}\tilde{h}_F^{-2}\|z_h\|^2_{H^3(K)} &\lesssim \sum_{F\in\mathscr{E}_h^b}\sum_{\substack{K\in\mathscr{T}_h:\\ F\subset\partial K}}\tilde{h}_F^{-2}h_K^{-p^*}\|z_h\|^2_{H^2(K)}\\ &\lesssim \sum_{F\in\mathscr{E}_h^b}\sum_{\substack{K\in\mathscr{T}_h:\\ F\subset\partial K}}\tilde{h}_F^{-(2+p^*)}(\|u-z_h\|^2_{H^2(K)}+\|u\|^2_{H^2(K)})\\ &\lesssim (\min_{F\in\mathscr{E}_h^b}\tilde{h}_F^{2+p^*})^{-1}\|u\|^2_{H^2(\Omega)},\end{aligned} \tag{6.9.26}$$

where the penultimate inequality follows from (4.5.14). Applying (6.9.26) to (6.9.25), and dividing through by $\|(\psi_h,\mu_h)\|_{h,1}$, we obtain

$$\begin{aligned}\|(\psi_h,\mu_h)\|_{h,1}\lesssim \|z_h-u\|_{H^2(\Omega)}&+\left(\frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\alpha}}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F^{2+p^*}}\right)^{\frac{1}{2}}\|u\|_{H^2(\Omega)}\\ &+\left[\sum_{F\in\mathscr{E}_h^b}\frac{1}{\tilde{h}_F}(\|\partial_{\mathbf{T}_2}(\beta\cdot\nabla z_h)\|^2_{L^2(F)}+\|\beta\cdot\nabla z_h-c\|^2_{L^2(F)})\right]^{\frac{1}{2}}.\end{aligned}$$

Recalling that $\beta \cdot \nabla u|_{\partial\Omega} = c$, we see that

$$\|(\xi_h, 0)\|_{h,1}^2 \lesssim |u - z_h|_{H^2(\Omega)}^2 + \sum_{F \in \mathscr{E}_h^b} \left[\frac{1}{\tilde{h}_F} \|\beta \cdot \nabla z_h - c\|_{L^2(F)}^2 + \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2} \nabla (u - z_h)\|_{L^2(F)}^2 \right].$$

Furthermore, the trace operator is continuous from $H^1(\Omega) \to L^2(\partial\Omega)$, and so

$$\sum_{F \in \mathscr{E}_h^b} \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2} \nabla (u - z_h)\|_{L^2(F)}^2 \lesssim \|\nabla (u - z_h)\|_{L^2(\partial\Omega)}^2 \lesssim |u - z_h|_{H^2_*(\Omega)}^2.$$

Thus, we obtain

$$\|(u - u_h, c - c_h)\|_{h,1} \lesssim \|z_h - u\|_{H^2(\Omega)} + \left(\frac{\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^\alpha}{\min_{F \in \mathscr{E}_h^b} \tilde{h}_F^{2+p^*}} \right)^{\frac{1}{2}} \|u\|_{H^2(\Omega)} + \left[\sum_{F \in \mathscr{E}_h^b} \frac{1}{\tilde{h}_F} (\|\partial_{\mathbf{T}_2}(\beta \cdot \nabla z_h)\|_{L^2(F)}^2 + \|\beta \cdot \nabla z_h - c\|_{L^2(F)}^2) \right]^{\frac{1}{2}}.$$

Note that our choice of $z_h \in H^2(\Omega) \cap V_{h,p,0}$ was arbitrary, thus we may take an infimum over V above, yielding estimate (6.9.15). $\square$

Remark 6.9.3 (Stabilisation parameter choice) *Note that our assumption $\tilde{h}_F^{1+\alpha} \lesssim \ell_F$, with $\alpha > 2 + p^*/2$ implies that for α large enough, we may control the contribution of the first term on the right-hand side of (6.9.15). Indeed, in the case of quasiuniform meshes, we would have $((\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^\alpha / \min_{F \in \mathscr{E}_h^b} \tilde{h}_F^{2+p^*})^{\frac{1}{2}} \lesssim h^{(\alpha - 2 - p^*)/2}$. Furthermore, we see that refining the mesh size internally (i.e., leaving $\tilde{h}_F$ unchanged for $F \in \mathscr{E}_h^b$) reduces the error, since the minimum in the denominator is over boundary edges, and the maximum in the numerator is over internal edges.*

Remark 6.9.4 (Conforming finite element methods) *Notice that the error bound (6.9.15) incorporates the error arising from the approximation of the oblique derivative (i.e., the last term in the infimum). Furthermore, one can see that if the space V is fully conforming, that is, $V = V_{h,p,0} \cap \{v \in H^2(\Omega) : \beta \cdot \nabla v|_{\partial\Omega} = c\}$, then this term vanishes entirely. However, it is not immediately clear that the space V is non-empty for arbitrary $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$ satisfying the hypotheses of Lemma 3.3.28.*

These considerations also imply that the scheme introduced in this chapter is at least as accurate as any conforming method seeking a numerical solution $u_h \in V$, where $V_{h,p} \cap H^2(\Omega) \subset V \subset V_{h,p,0} \cap \{v \in H^2(\Omega) : \beta \cdot \nabla v|_{\partial\Omega} = c\}$.

6.10 Implementation

Software and code: The experiments in this Chapter have been implemented in the most recent version of the Firedrake software [105, 87] (as of 3rd July 2018), which interfaces directly with PETSc [6, 7] running through a Python interface [39, 63]. A working Firedrake script, Curved-oblique-DGFEM.py, used to generate the experiments of this Chapter is available in the Github repository:
https://github.com/ekawecki/FiredrakeNDV.

Linear systems and condition numbers: Akin to the the bilinear form $A_h^{\mathcal{D}}$ defined by (5.4.4), the bilinear form $A_h^{\mathcal{O}}$ defined by (6.6.4) can also be considered to be similar to those present in finite element methods for fourth-order elliptic boundary-value problems (see [114, 25] for example), in the sense that the evaluation of $A_h^{\mathcal{O}}((u_h, \lambda_h); (v_h, \mu_h))$ for $(u_h, \lambda_h), (v_h, \mu_h) \in M_h$ involves the integration of products of second order partial derivatives. This typically leads to the matrix $A^{\mathcal{O}}$, describing the linear system given by (6.6.5), to have a Euclidean norm condition number of order h^{-4}. This can pose difficulties when applying iterative methods to solve the linear system, and thus to ensure that we solve the linear system with sufficiently high accuracy as the mesh size h decreases, we apply the Iterative refinement algorithm, i.e., Algorithm 1.1 of [32]. We implement the Iterative refinement algorithm by using the following choices in the Firedrake "solve" function.

```
        # implementing nullspace, as solution should have zero sum
        V_basis = VectorSpaceBasis(constant=True)
        nullspace = MixedVectorSpaceBasis(S, [V_basis, S[1]])

        # begin timing of linear system solve
        t = time()

        # solving linear system
        solve(A_gamma == L,Uh,nullspace = nullspace,
              solver_parameters = {"mat_type": "aij",
              "snes_type": "newtonls",
              "ksp_type": "preonly",
              "pc_type": "lu",
              "snes_monitor": False,
              "snes_rtol": 1e-16,
              "snes_atol": 1e-25})
        # end timing of linear system solve
        tt.append(time()-t)
```

One can also see that when executing the script in Firedrake, we record the runtimes by way of the sixth and last line above, so that we only record the time that it takes to solve the linear system.

Furthermore, the solver choices differ slightly from those present in Section 5.8, i.e., we also include "nullspace = nullspace", where "nullspace" is defined on line 3,

and "mat_type = aij". The first choice imposes that the numerical solution u_h (from the pair $(u_h, c_h) \in M_h$ that satisifies (6.6.5)) has a zero-sum, and the latter essentially informs the solver that the solution consists of two parts, i.e., u_h and c_h, and that the system may be treated in block formation.

Two-dimensional curved boundary approximation: When implementing curved finite elements, we use a piecewise quadratic polynomial mapping to obtain a higher order approximation of the domain boundary. This is implemented in exactly the same manner as discussed in Section 5.8. As in Section 5.8, we define the space $V_{h,p}^{\text{comp}} := \{v \in L^2(\Omega) : v \circ T^{-1} \in \mathbb{P}^p(\hat{K})\}$, where the piecewise quadratic function T is defined by (5.8.1). In this case, we then define $V_{h,p,0}^{\text{comp}} := V_{h,p}^{\text{comp}} \cap L_0^2(\Omega)$ and $M_h^{\text{comp}} := V_{h,p,0}^{\text{comp}} \times V_{h,0}^{\text{comp}}$.

Remark 6.10.1 (Computational parameters) *In the following experiments, we employ the following parameter choices:* $c_{\text{stab}} = 2.5$, $\mu_F = 2c_{\text{stab}}(p-1)^2/\tilde{h}_F$, $\eta_F = 15(p-1)^4/16\tilde{h}_F^3$, $\sigma_F = 2c_{\text{stab}}p^2/\tilde{h}_F^2$, *and* $\ell_F = c_{\text{stab}}\tilde{h}_F^{-3}$. *The order of the computational parameters with respect to* $\tilde{h}_F$ *were guided by the hypotheses of Theorem 6.9.1. The orders with respect to* p *for* η_F *and* μ_F *were guided by the experiments in Section 5.9. Finally, the value of* c_{stab}, *and the orders with respect to* p *of* σ_F *and* ℓ_F *(in the case of* ℓ_F, *the parameter is in fact independent of* p*) were obtained experimentally.*

6.11 Experiments

In this section, we test the robustness of the scheme (6.6.5), with the computational domain Ω taken to be the unit disk, and consider various elliptic operators, L, that satisfy the Cordes condition (5.1.4). In each case, we see that the convergence rates are of the expected order in the various broken Sobolev norms considered, and in particular in the $\|\cdot\|_{h,1}$–norm, for which we have proven the error bound (6.9.1).

6.11.1 Experiment 1

In this experiment, we consider the following problem

$$\begin{cases} \Delta u = f, \quad \text{in} \quad \Omega, \\ \beta \cdot \nabla u \text{ is constant on } \partial\Omega, \end{cases} \tag{6.11.1}$$

where $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$, and $\beta \equiv n_{\partial\Omega}$. In this case f is chosen so that the solution of (6.11.1) is given by

$$u(x) = \frac{1}{6}|x|^6 - \frac{1}{2}|x|^2 + \frac{5}{24}.$$

Since β coincides with the unit outward normal to $\partial\Omega$, this problem corresponds to a Neumann boundary-value problem. Moreover, one can see that the oriented angle, Θ, between β and $n_{\partial\Omega}$, satisfies $\Theta = 0$ on $\partial\Omega$, and thus $\partial_{\mathbf{T}_2}\Theta = 0$ on $\partial\Omega$. Furthermore, since Ω is the unit disk, it follows that $\partial\Omega = \mathbb{S}^1$, and that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. It is also then clear that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = 1 > 0 \quad \text{on } \partial\Omega.$$

Since the solution in known, one can directly calculate that $\nabla u|_{\partial\Omega} = (|x|^5 - 1)x$, and so $\beta \cdot \nabla u|_{\partial\Omega} = n_{\partial\Omega} \cdot ((|x|^5 - 1)x) = |x|^6 - |x|^2 = 0$ (since $|x| = 1$ on $\partial\Omega$), and thus, the compatibility constant $c = 0$. We can also directly calculate the renormalisation parameter, γ, and provide the largest value of ε for which the Cordes condition (5.1.4) holds. In particular, we have that

$$\gamma := \frac{\mathrm{Tr}(A)}{|A|^2} = \frac{\mathrm{Tr}(I_d)}{|I_d|^2} = \frac{I_d : I_d}{I_d : I_d} = 1, \quad \text{and} \quad \varepsilon = 1.$$

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p,0}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|(u - u_h, c - c_h)\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm, and plot the error arising in the approximation of the compatibility constant in Figure 6.1, and report the exact values in Tables 6.1 and 6.2, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and runtimes for each computation in Table 6.3.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	2.75		1.16		3.09×10^{-1}	
0.2828	1.84	(0.70)	3.61×10^{-1}	(2.06)	4.33×10^{-2}	(3.47)
0.1627	1.10	(0.94)	1.17×10^{-1}	(2.03)	7.35×10^{-3}	(3.21)
0.0973	6.00×10^{-1}	(1.17)	3.45×10^{-2}	(2.39)	1.06×10^{-3}	(3.76)
0.0508	2.93×10^{-1}	(1.10)	8.46×10^{-3}	(2.16)	1.27×10^{-4}	(3.27)
0.0269	1.47×10^{-1}	(1.08)	2.11×10^{-3}	(2.18)	1.47×10^{-5}	(3.38)
0.0138	7.24×10^{-2}	(1.06)	5.12×10^{-4}	(2.11)	1.70×10^{-6}	(3.22)

Table 6.1: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 6.11.1.

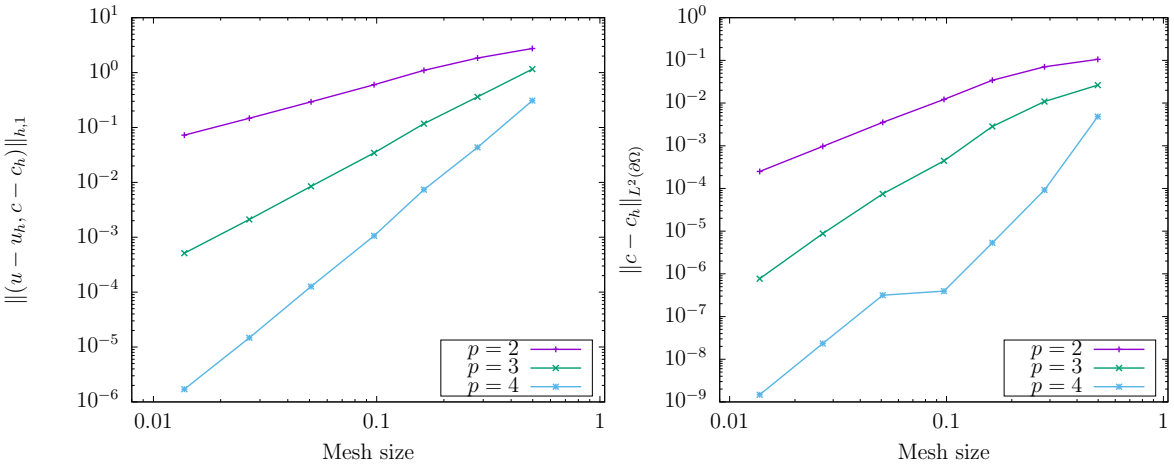


Figure 6.1: Convergence rates for the numerical scheme applied to problem (6.11.1). We provide the error values $\|(u-u_h, c-c_h)\|_{h,1}$ (left), and $\|c-c_h\|_{L^2(\partial\Omega)}$ (right). We observe that the convergence rates in the $\|\cdot\|_{h,1}$ norm are optimal with respect to the choice of polynomial degree, p. That is, $\|(u-u_h, c-c_h)\|_{h,1} = \mathcal{O}(h^{p-1})$. Furthermore, we observe that $\|c-c_h\|_{L^2(\partial\Omega)} = \mathcal{O}(h^p)$.

Mesh size	$p=2$		$p=3$		$p=4$	
0.4981	1.06×10^{-1}		2.63×10^{-2}		4.82×10^{-3}	
0.2828	7.06×10^{-2}	(0.72)	1.09×10^{-2}	(1.56)	9.19×10^{-5}	(6.99)
0.1627	3.42×10^{-2}	(1.31)	2.83×10^{-3}	(2.44)	5.32×10^{-6}	(5.16)
0.0973	1.22×10^{-2}	(2.01)	4.44×10^{-4}	(3.60)	3.94×10^{-7}	(5.06)
0.0508	3.53×10^{-3}	(1.91)	7.48×10^{-5}	(2.74)	3.16×10^{-7}	(0.34)
0.0269	9.70×10^{-4}	(2.03)	8.79×10^{-6}	(3.36)	2.34×10^{-8}	(4.08)
0.0138	2.50×10^{-4}	(2.02)	7.72×10^{-7}	(3.63)	1.48×10^{-9}	(4.13)

Table 6.2: $\|c-c_h\|_{L^2(\partial\Omega)}$ error values and EOCs for Experiment 6.11.1.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.4981	0.72	0.88	0.89	112	176	256
0.2828	0.37	0.38	0.41	448	704	1024
0.1627	0.38	0.42	0.93	1218	1914	2784
0.0973	0.46	0.58	0.89	3990	6270	9120
0.0508	0.84	1.75	3.95	16240	25520	37120
0.0269	3.31	9.38	25.09	61222	96206	139936
0.0138	20.96	66.31	199.07	240156	377388	548928

Table 6.3: Runtimes and number of DoFs for Experiment 6.11.1, for each mesh size h, and each polynomial degree, p.

6.11.2 Experiment 2

In this experiment, we consider the following problem

$$\begin{cases} \sum_{i,j=1}^{2}(1+\delta_{ij})\frac{x_i}{|x_i|}\frac{x_j}{|x_j|}D^2_{ij}u = f, \quad \text{in} \quad \Omega, \\ \qquad\qquad\qquad\qquad \beta\cdot\nabla u \text{ is constant on } \partial\Omega, \end{cases} \tag{6.11.2}$$

where $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$, and β is a $\pi/4$ anticlockwise rotation of the normal, $n_{\partial\Omega}$. That is

$$\beta = \frac{1}{\sqrt{2}}([n_{\partial\Omega}]^1 - [n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1 + [n_{\partial\Omega}]^2)^T.$$

In this case, f is chosen so that the solution of (6.11.2) is given by

$$u(x) = |x|^{1.5} - 0.75|x|^2 - \frac{1}{\pi}\int_\Omega (|x|^{1.5} - 0.75|x|^2).$$

One can see that the oriented angle, Θ, between β and $n_{\partial\Omega}$, satisfies $\Theta = \pi/4$ on $\partial\Omega$, and thus $\partial_{\mathbf{T}_2}\Theta = 0$ on $\partial\Omega$. Furthermore, since Ω is the unit disk, it follows that $\partial\Omega = \mathbb{S}^1$, and that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. It is also then clear that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = 1 > 0 \quad \text{on } \partial\Omega.$$

Since the solution in known, one can directly calculate that $\nabla u|_{\partial\Omega} = 1.5(|x|^{-1/2} - 1)x$, and so $\beta \cdot \nabla u|_{\partial\Omega} = \frac{1}{\sqrt{2}}([n_{\partial\Omega}]^1 - [n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1 + [n_{\partial\Omega}]^2) \cdot (1.5(|x|^{-1/2} - 1)x) = 0$ (since $|x| = 1$ on $\partial\Omega$), and thus, the compatibility constant $c = 0$. We can also directly

calculate the renormalisation parameter, γ, and provide the largest value of ε for which the Cordes condition (5.1.4) holds. In particular, we have that

$$\gamma = \frac{\mathrm{Tr}(A)}{|A|^2} = \frac{2 + x_1^2/|x_1|^2 + x_2^2/|x_2|^2}{8 + 2x_1^2 x_2^2/(|x_1|^2|x_2|^2)} = 2/5, \quad \text{and} \quad \varepsilon = 3/5.$$

In this experiment, the true solution $u \in H^2(\Omega)$, and, in particular, $u \in H^{5/2-\delta}(\Omega)$ for arbitrary $\delta > 0$. However, the H^s-broken Sobolev regularity of u fails for $s \geq 5/2$, and we must appeal to the minimal regularity estimate of Lemma 6.9.2. In this experiment we successively increase the degree, p, of the finite element space $V_{h,p,0}^{\mathrm{comp}}$ from 2 to 4. Furthermore, we compute the numerical solution both on sequence of meshes refined towards the origin (where the solution lacks regularity, an example of such a mesh is given in Figure 6.11.2), and on a sequence of quasi-uniformly refined meshes (that in particular does not prioritise refinement towards the origin). We plot the error arising in both cases (adapted mesh refinement and non adapted mesh refinement) in the broken H^2-seminorm, against the number of DoFs in Figure 6.11.2, and report the error values, with experimental orders of convergence in brackets (calculated with respect to NDoFs), along with the runtimes for each computation in Tables 6.4 and 6.5. For $p = 2, 3, 4$, we see a reduction in error from the adapted mesh sequence. In particular, for $p = 3$ and $p = 4$ we see a reduction in the order of error from $\mathcal{O}(\mathrm{ndofs}^{-1/4})$ to $\mathcal{O}(\mathrm{ndofs}^{-1/2})$.

Quasi-uniformly refined mesh							
$p = 2$				$p = 3$			
NDoFs	Error		Runtime	NDoFs	Error		Runtime
112	1.38		0.79	176	8.99×10^{-1}		0.90
448	1.18	(-0.12)	0.50	704	6.95×10^{-1}	(-0.19)	0.43
1218	9.17×10^{-1}	(-0.25)	0.52	1914	5.11×10^{-1}	(-0.31)	0.53
3990	6.47×10^{-1}	(-0.29)	0.67	6270	3.61×10^{-1}	(-0.29)	0.91
16240	4.95×10^{-1}	(-0.19)	1.72	25520	2.73×10^{-1}	(-0.20)	3.03
61222	3.43×10^{-1}	(-0.28)	6.56	96206	1.87×10^{-1}	(-0.28)	14.31

$p = 4$			
NDoFs	Error		Runtime
256	6.63×10^{-1}		0.88
1024	5.05×10^{-1}	(-0.20)	0.45
2784	3.60×10^{-1}	(-0.34)	0.67
9120	3.01×10^{-1}	(-0.15)	1.37
37120	1.93×10^{-1}	(-0.32)	5.90
139936	1.40×10^{-1}	(-0.24)	36.48

Table 6.4: NDoFs, error values in the $|\cdot|_{H^2(\Omega;\mathscr{T}_h)}$-seminorm with EOCs in brackets (calculated with respect to NDoFs), and corresponding runtimes for Experiment 6.11.2, with a quasi-uniformly refined mesh.

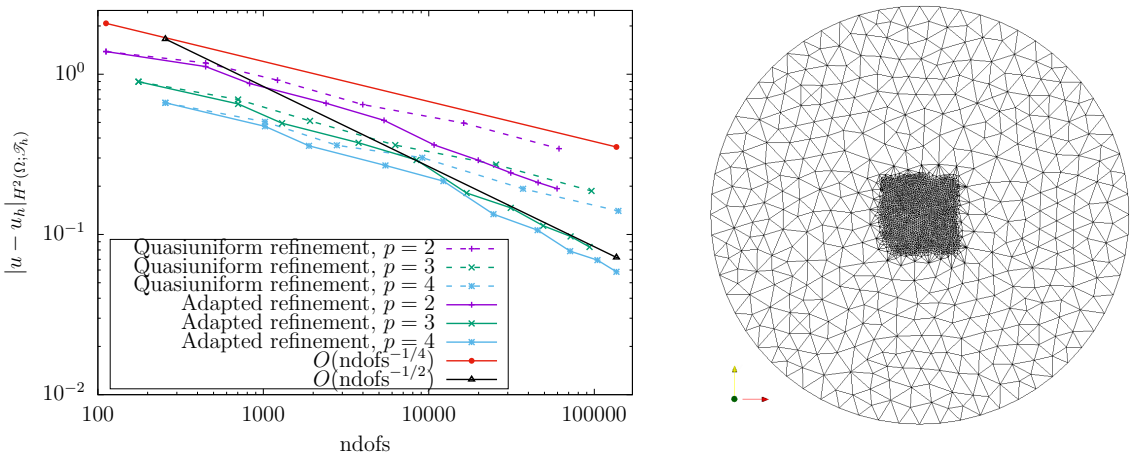


Figure 6.2: Convergence rates for the numerical scheme applied to problem (6.11.2), with a true solution of minimal regularity. On the left, we provide the error values in the $|\cdot|_{H^2(\Omega)}$ seminorm, where the numerical scheme is implemented on a quasiuniformly refined mesh, and an adapted mesh, with refinement towards the origin. On the right we provide an example of this adapted mesh, at refinement level 7, consisting of 4532 elements.

Adapted mesh						
$p = 2$				$p = 3$		
NDoFs	Error		Runtime	NDoFs	Error		Runtime
112	1.38		0.81	176	8.99×10^{-1}		0.98
448	1.12	(-0.16)	0.48	704	6.53×10^{-1}	(-0.23)	0.43
826	8.75×10^{-1}	(-0.40)	0.44	1298	4.95×10^{-1}	(-0.45)	0.47
2394	6.59×10^{-1}	(-0.27)	0.56	3762	3.73×10^{-1}	(-0.26)	0.76
5362	5.15×10^{-1}	(-0.30)	0.82	8426	2.91×10^{-1}	(-0.31)	1.30
10766	3.62×10^{-1}	(-0.51)	1.33	16918	1.81×10^{-1}	(-0.68)	2.99
19894	2.90×10^{-1}	(-0.36)	2.42	31262	1.47×10^{-1}	(-0.35)	6.03
31276	2.42×10^{-1}	(-0.40)	4.27	49148	1.13×10^{-1}	(-0.57)	11.35
45752	2.11×10^{-1}	(-0.37)	6.61	71896	9.72×10^{-2}	(-0.40)	18.11
59570	1.93×10^{-1}	(-0.33)	8.38	93610	8.35×10^{-2}	(-0.58)	22.21

$p = 4$			
NDoFs	Error		Runtime
256	6.63×10^{-1}		1.00
1024	4.73×10^{-1}	(-0.24)	0.51
1888	3.57×10^{-1}	(-0.46)	0.63
5472	2.70×10^{-1}	(-0.26)	1.04
12256	2.15×10^{-1}	(-0.28)	2.17
24608	1.34×10^{-1}	(-0.68)	5.05
45472	1.06×10^{-1}	(-0.38)	11.07
71488	7.86×10^{-2}	(-0.67)	25.11
104576	6.91×10^{-2}	(-0.34)	42.45
136160	5.84×10^{-2}	(-0.64)	56.49

Table 6.5: NDoFs, error values in the $|\cdot|_{H^2(\Omega;\mathscr{T}_h)}$-seminorm with EOCs in brackets (calculated with respect to NDoFs), and corresponding runtimes for Experiment 6.11.2, with an adapted mesh.

6.11.3 Experiment 3

In this experiment, we consider problem (6.11.2), where $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$. We take β to be the anti-clockwise rotation of the normal by the angle $\varphi(x_1, x_2) := \pi/4 + \arctan(\frac{x_2}{x_1})$, for $(x_1, x_2) \in \partial\Omega$, that is,

$$\beta(x_1, x_2) = \begin{bmatrix} \cos\varphi(x_1, x_2) & -\sin\varphi(x_1, x_2) \\ \sin\varphi(x_1, x_2) & \cos\varphi(x_1, x_2) \end{bmatrix} \begin{bmatrix} [n_{\partial\Omega}]^1 \\ [n_{\partial\Omega}]^2 \end{bmatrix} \quad (x_1, x_2) \in \partial\Omega.$$

Furthermore, the function f on the right-hand side of (6.11.2) is chosen so that the solution u is given by

$$u(x_1, x_2) = \frac{1}{4}\cos(\pi(x_1^2 + x_2^2)) - \frac{1}{\pi}\int_\Omega \frac{1}{4}\cos(\pi(x_1^2 + x_2^2)).$$

As in Experiment 6.11.2, we can directly calculate the renormalisation parameter, γ, and provide the largest value of ε for which the Cordes condition (5.1.4) holds. In particular, we have that $\gamma = \frac{\mathrm{Tr}(A)}{|A|^2} = 2/5$ and $\varepsilon = 3/5$.

We can also see that the oriented angle, Θ, between β and $n_{\partial\Omega}$ is given by $\Theta(x_1, x_2) = \pi/4 + \varphi(x_1, x_2)$ for $(x_1, x_2) \in \partial\Omega$. It then follows that on $\partial\Omega$,

$$\begin{aligned} \partial_{\mathbf{T}_2}\Theta &= \nabla(\varphi(x_1, x_2)) \cdot (-[n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1) \\ &= \nabla(\arctan(x_2/x_1)) \cdot (-x_2, x_1) \\ &= \frac{1}{x_1^2 + x_2^2}(-x_2, x_1) \cdot (-x_2, x_1) = 1. \end{aligned} \tag{6.11.3}$$

Furthermore, since Ω is the unit disk, and $\partial\Omega = \mathbb{S}^1$, the oblique vector β rotates by exactly 2π, as we traverse $\partial\Omega$ in a fixed direction, and so $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Since the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, by (6.11.3) we also see that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = 2 > 0.$$

This experiment serves to demonstrate the robustness of this method with respect to the choice of oblique vector, β. In particular, β performs a full rotation around the normal vector.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p,0}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly. In Figure 6.11.3 we plot the convergence rates in the $\|\cdot\|_{h,1}$-norm and the $|\cdot|_{H^1(\Omega;\mathscr{T}_h)}$-seminorm, and report the exact values in Tables 6.6 and 6.7, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and runtimes for each computation in Table 6.8. We observe the optimal convergence rates $\|(u - u_h, c - c_h)\|_{h,1} = \mathcal{O}(h^{p-1})$, and $|u - u_h|_{H^1(\Omega;\mathscr{T}_h)} = \mathcal{O}(h^p)$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	8.64		5.05		1.13	
0.2828	6.86	(0.41)	1.03	(2.80)	2.50×10^{-1}	(2.66)
0.1627	3.66	(1.14)	3.10×10^{-1}	(2.18)	6.72×10^{-2}	(2.38)
0.0973	1.86	(1.32)	1.16×10^{-1}	(1.91)	1.41×10^{-2}	(3.04)
0.0508	8.80×10^{-1}	(1.15)	2.96×10^{-2}	(2.10)	1.81×10^{-3}	(3.15)
0.0269	4.42×10^{-1}	(1.08)	8.31×10^{-3}	(1.99)	2.41×10^{-4}	(3.17)
0.0138	2.21×10^{-1}	(1.04)	2.15×10^{-3}	(2.02)	3.02×10^{-5}	(3.10)

Table 6.6: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 6.11.3.

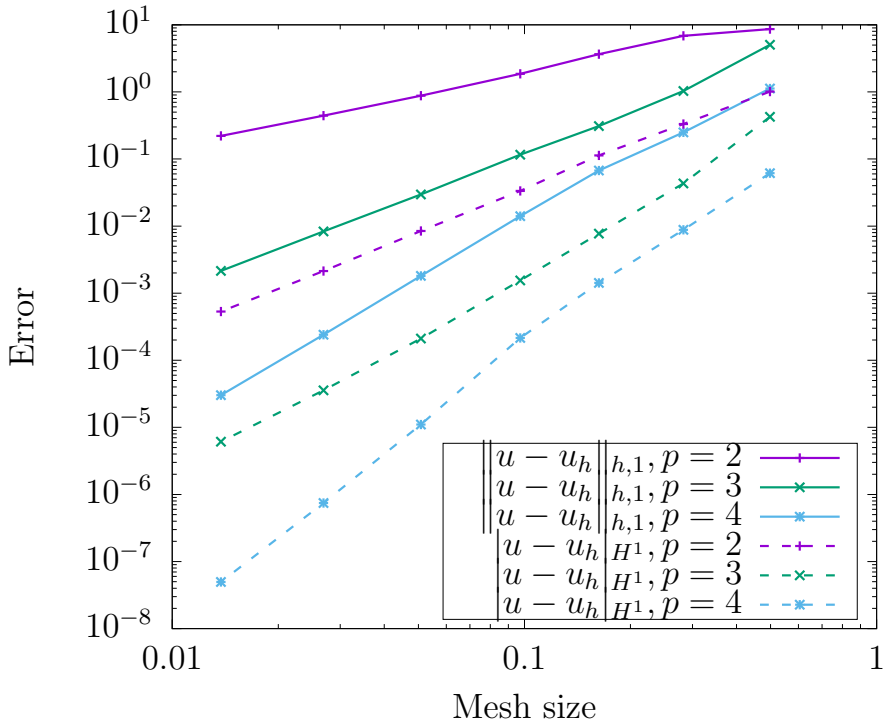


Figure 6.3: Convergence rates for Experiment 6.11.3. We provide the error values $\|(u - u_h, c - c_h)\|_{h,1}$, and $|u - u_h|_{H^1(\Omega;\mathscr{T}_h)}$, along with the final (i.e., from the final mesh refinement) and mean experimental order of convergence. We observe that the convergence rates in the $\|\cdot\|_{h,1}$ norm are optimal with respect to the choice of polynomial degree, p. That is, $\|(u - u_h, c - c_h)\|_{h,1} = \mathcal{O}(h^{p-1})$. Furthermore, we observe that $|u - u_h|_{H^1(\Omega;\mathscr{T}_h)} = \mathcal{O}(h^p)$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	1.01		4.26×10^{-1}		6.14×10^{-2}	
0.2828	3.32×10^{-1}	(1.96)	4.33×10^{-2}	(4.04)	8.78×10^{-3}	(3.43)
0.1627	1.13×10^{-1}	(1.95)	7.71×10^{-3}	(3.13)	1.42×10^{-3}	(3.30)
0.0973	3.34×10^{-2}	(2.37)	1.55×10^{-3}	(3.12)	2.15×10^{-4}	(3.67)
0.0508	8.45×10^{-3}	(2.11)	2.11×10^{-4}	(3.07)	1.10×10^{-5}	(4.57)
0.0269	2.14×10^{-3}	(2.16)	3.57×10^{-5}	(2.79)	7.45×10^{-7}	(4.23)
0.0138	5.32×10^{-4}	(2.08)	6.10×10^{-6}	(2.64)	4.96×10^{-8}	(4.04)

Table 6.7: Error values in the $|\cdot|_{H^1(\Omega;\mathscr{T}_h)}$-seminorm and EOCs for Experiment 6.11.3.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.4981	0.90	23.88	26.14	112	176	256
0.2828	0.49	0.69	0.54	448	704	1024
0.1627	0.53	0.62	0.78	1218	1914	2784
0.0973	0.78	1.00	1.51	3990	6270	9120
0.0508	1.95	3.23	6.49	16240	25520	37120
0.0269	6.80	15.12	36.20	61222	96206	139936
0.0138	36.57	97.87	261.50	240156	377388	548928

Table 6.8: Runtimes and number of DoFs for Experiment 6.11.3, for each mesh size h, and each polynomial degree, p.

6.12 Concluding remarks for this method

We have extended the framework introduced in [110], and [70] allowing for domains with curved boundaries, as well as oblique boundary conditions. In doing so, we have introduced a new DGFEM for elliptic equations in nondivergence form, that satisfy the Cordes condition.

The computational domain we considered was the unit disc; in order to verify the error estimates present in Section 3 we used a mesh consisting of curved triangles with edges were defined by polynomial mappings. It would be an interesting avenue for future research to consider oblique boundary-value problems in dimensions three and higher; this would require one to prove the Miranda–Talenti estimates (3.3.28) in $H^2_{\beta,0}(\Omega)$, in higher dimensions, which is currently an open problem.

The finite element approximation of solutions to elliptic problems in nondivergence form with oblique boundary conditions is a challenging problem, and as such appears to be underrepresented in the available literature. This chapter provides and analyses a new method, which appears to be the first discontinuous Galerkin finite element method for oblique boundary-value problems; we were successful in proving both a stability estimate (6.8.9), guaranteeing existence and uniqueness of the numerical solution, and an optimal apriori error estimate (6.9.1).

Chapter 7

A DGFEM for HJB equations with Dirichlet and oblique boundary conditions, with applications to MA type problems

7.1 New contributions and existing results

The goal of this Chapter is to design and analyse discontinuous Galerkin finite element methods (DGFEMs) for HJB and MA type equations with Dirichlet and oblique boundary conditions on domains with *curved* boundaries.

Existing results: Some of the contributions of this chapter build upon existing results.

- In [111], the authors provide a *hp*-DGFEM for the approximation of strong solutions to HJB equations with Dirichlet boundary conditions on *polytopal* domains, proving existence and uniqueness of a numerical solution, as well as optimal a prior error bounds in a broken H^2-type norm. We extend the h-version of this method, and its analysis to the case of domains with curved boundaries.

Our original contributions are listed as follows:

1. We provide a DGFEM for the approximation of strong solutions to HJB equations with Dirichlet and oblique boundary conditions on domains with *curved* boundaries. In particular:

    - We prove that the nonlinear operators $\mathscr{A}_h^{\mathcal{D}}$ and $\mathscr{A}_h^{\mathcal{O}}$ defined in Section 7.3 are strongly monotone and Lipschitz continuous, yielding existence and

uniqueness of a numerical solution in the Dirichlet and oblique case, respectively;

- We prove optimal error estimates with respect to the mesh size in broken H^2-type norms.

2. Utilising Theorem 3.5.3, we provide a new DGFEM for the approximation of solutions to MA type problems in the case $d = 2$. Furthermore, our experiments (in particular Experiment 7.9.5) show that the semismooth Newton's method employed to approximate the numerical solution is robust with respect to the choice of initial guess and right-hand side function. In particular, upon convergence, the semismooth Newton's method converges to the unique uniformly convex solution (as mentioned in Section 3.5, such a property is an advantage of this method, and is not always present in numerical methods for MA type problems, sometimes resulting in large scale nonuniqueness).

7.2 Model problems

In this chapter, we propose and analyse a numerical method for HJB equations on domains with curved boundaries. Furthermore, we use the results presented in Section 3.4 to reformulate the MA problem as a HJB problem, and apply the proposed method, resulting in a new DGFEM for the approximation of solutions to MA type problems.

Our model problems will be the following two HJB equations:

$$\begin{cases} \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \text{ a.e. in } \quad \Omega, \\ \qquad\qquad\qquad u = 0 \text{ on } \quad \partial\Omega, \end{cases} \tag{7.2.1}$$

and

$$\begin{cases} \sup_{\alpha\in\Lambda}\{L^\alpha u - f^\alpha\} = 0 \text{ a.e. in } \quad \Omega, \\ \qquad\qquad \beta\cdot\nabla u \text{ is constant on } \partial\Omega. \end{cases} \tag{7.2.2}$$

Remark 7.2.1 *As before, we assume that $\Omega \subset \mathbb{R}^d$ is convex, and that the collection of linear operators $\{L^\alpha\}_{\alpha\in\Lambda}$ satisfies the assumptions of Theorem 3.3.8. Then, recall that there exists a unique $u \in H := H^2(\Omega)\cap H_0^2(\Omega)$ that is a strong solution of (7.2.1), thanks to Theorem 3.3.16.*

In the oblique case, Theorem 3.3.29 guarantees existence of a unique $u \in H^2_{\beta,0}(\Omega)$ that is a strong solution of (7.2.2). This, however, only holds when $d = 2$, and as such we also provide a numerical scheme for the two-dimensional case.

7.3 Numerical schemes

The definitions of the numerical schemes in this chapter rely upon the bilinear forms $B^{\mathcal{D}}_{h,*}, J^{\mathcal{D}}_h, B^{\mathcal{D}}_{h,1/2} : V_{h,p} \times V_{h,p} \to \mathbb{R}$ defined in Chapter 5, and $B^{\mathcal{O}}_{h,*}, J^{\mathcal{O}}_h, B^{\mathcal{O}}_{h,1/2} : M_h \times M_h \to \mathbb{R}$ defined in Chapter 6. Recall that the finite element spaces $V_{h,p}$ and M_h are define as follows:

$$V_{h,p} := \{v \in L^2(\Omega) : \ v|_K = \hat{\rho} \circ F_K^{-1},\ \hat{\rho} \in \mathbb{P}^p(\hat{K}),\ \forall K \in \mathscr{T}_h\},$$

and

$$M_h := V_{h,p,0} \times V_{h,0} = (V_{h,p} \cap L^2_0(\Omega)) \times V_{h,0}.$$

The main difference between the numerical methods considered earlier for linear PDEs and the one for nonlinear HJB problems is in the *semilinear* forms $\mathscr{A}^{\mathcal{D}}_h : V_{h,p} \times V_{h,p} \to \mathbb{R}$, and $\mathscr{A}^{\mathcal{O}}_h : M_h \times M_h \to \mathbb{R}$ defined by

$$\begin{aligned}
\mathscr{A}^{\mathcal{D}}_h(u_h; v_h) &:= \sum_{K \in \mathscr{T}_h} \langle F_\gamma[u_h], \Delta v_h \rangle_K + B^{\mathcal{D}}_{h,1/2}(u_h, v_h) - \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K, \\
\mathscr{A}^{\mathcal{O}}_h((u_h, \lambda); (v_h, \mu)) &:= \sum_{K \in \mathscr{T}_h} \langle F_\gamma[u_h], \Delta v_h \rangle_K + B^{\mathcal{O}}_{h,1/2}((u_h, \lambda); (v_h, \mu)) \\
&\qquad - \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K,
\end{aligned} \tag{7.3.1}$$

where we recall that

$$F_\gamma[u_h] = \sup_{\alpha \in \Lambda} \{\gamma^\alpha (A^\alpha : D^2 u - f^\alpha)\}.$$

The forms $\mathscr{A}^{\mathcal{D}}_h$ and $\mathscr{A}^{\mathcal{O}}_h$ are linear in the second argument and nonlinear in the first argument. The scheme for approximating the solution of (7.2.1) is to find $u_h \in V_{h,p}$ such that

$$\mathscr{A}^{\mathcal{D}}_h(u_h; v_h) = 0 \quad \forall v_h \in V_{h,p}. \tag{7.3.2}$$

Similarly, the scheme for approximating the solution of (7.2.2) is to find $(u_h, c_h) \in M_h$ such that

$$\mathscr{A}^{\mathcal{O}}_h((u_h, c_h); (v_h, \mu)) = 0 \quad \forall (v_h, \mu) \in M_h. \tag{7.3.3}$$

7.4 Monotonicity analysis

Theorem 7.4.1 *Under the hypotheses of Lemma 5.6.2, let c_{stab}, $c_{\mathcal{H}}$, η_F and μ_F be chosen so that (5.6.2) holds with $\kappa < (1-\varepsilon)^{-1/2}$. Then, for every $u_h, v_h \in V_{h,p}$, we have*

$$\|u_h - v_h\|^2_{h,1} \le C(\mathscr{A}^{\mathcal{D}}_h(u_h; u_h - v_h) - \mathscr{A}^{\mathcal{D}}_h(v_h; u_h - v_h)), \tag{7.4.1}$$

where the constant $C := 2\kappa/(1-\kappa(1-\varepsilon))$. *Moreover, there exists a constant* C, *independent of* h, *such that for any* u_h, v_h, *and* z_h *in* $V_{h,p}$,

$$|\mathscr{A}_h^{\mathcal{D}}(u_h; z_h) - \mathscr{A}_h^{\mathcal{D}}(v_h; z_h)| \leq C\|u_h - v_h\|_{h,1}\|z_h\|_{h,1}. \tag{7.4.2}$$

Therefore, there exists a unique solution $u_h \in V_{h,p}$ *to the numerical scheme (7.3.2). Furthermore, we have the bound*

$$\|u_h\|_{h,1} \leq \frac{2\kappa\sqrt{d+1}\sup_{\alpha\in\Lambda}\|\gamma^\alpha\|_{L^\infty(\Omega)}}{1-\kappa(1-\varepsilon)}\|\sup_{\alpha\in\Lambda}|f^\alpha|\|_{L^2(\Omega)}. \tag{7.4.3}$$

Proof: For any $u_h, v_h, z_h, \in V_{h,p}$, (3.3.19) yields

$$\sum_{K\in\mathscr{T}_h}\langle F_\gamma[u_h] - F_\gamma[v_h] - \Delta(u_h - v_h), \Delta z_h\rangle_K \leq \sqrt{1-\varepsilon}\sum_{K\in\mathscr{T}_h}|u_h - v_h|_{H^2(K)}\|\Delta z_h\|_{L^2(K)}. \tag{7.4.4}$$

We first show the Lipschitz continuity of $\mathscr{A}_h^{\mathcal{D}}$. Applying (7.4.4), for any $u_h, v_h, z_h, \in V_{h,p}$, we obtain

$$\begin{aligned}
&\mathscr{A}_h^{\mathcal{D}}(u_h; z_h) - \mathscr{A}_h^{\mathcal{D}}(v_h; z_h)\\
&\quad= \sum_{K\in\mathscr{T}_h}\langle F_\gamma[u_h] - F_\gamma[v_h] - \Delta(u_h - v_h), \Delta z_h\rangle_K + B_{h,1/2}^{\mathcal{D}}(u_h - v_h, z_h)\\
&\quad\leq \sqrt{1-\varepsilon}\sum_{K\in\mathscr{T}_h}|u_h - v_h|_{H^2(K)}\|\Delta z_h\|_{L^2(K)} + B_{h,1/2}^{\mathcal{D}}(u_h - v_h, z_h)\\
&\quad\leq \sqrt{d(1-\varepsilon)}\|u_h - v_h\|_{h,1}\|z_h\|_{h,1} + B_{h,1/2}^{\mathcal{D}}(u_h - v_h, z_h);
\end{aligned}$$

thus, it suffices to show that the bilinear form $B_{h,1/2}^{\mathcal{D}}$ is a bounded on $V_{h,p}\times V_{h,p}$. Applying techniques analogous to those employed in the proof of Lemma 5.6.2, we obtain

$$B_{h,1/2}^{\mathcal{D}}(u_h - v_h, z_h) \leq C\|u_h - v_h\|_{h,1}\|z_h\|_{h,1},$$

where the constant C is independent of u_h, v_h, z_h, and h. This gives us (7.4.2).

We will now prove (7.4.1). For $u_h, v_h \in V_{h,p}$, we denote $w_h := u_h - v_h$. Applying (7.4.4), we have that

$$\begin{aligned}
&\mathscr{A}_h^{\mathcal{D}}(u_h; w_h) - \mathscr{A}_h^{\mathcal{D}}(v_h; w_h)\\
&\quad= \sum_{K\in\mathscr{T}_h}\langle F_\gamma[u_h] - F_\gamma[v_h] - \Delta w_h, \Delta w_h\rangle_K + B_{h,1/2}^{\mathcal{D}}(w_h, w_h)\\
&\quad\geq B_{h,1/2}^{\mathcal{D}}(w_h, w_h) - \sqrt{1-\varepsilon}\sum_{K\in\mathscr{T}_h}\|D^2 w_h\|_{L^2(K)}\|\Delta w_h\|_{L^2(K)}.
\end{aligned}$$

Applying the stability estimate (5.6.2), with $\theta = \frac{1}{2}$, for $\kappa > 1$ satisfying $\kappa < (1-\varepsilon)^{-1/2}$, we obtain

$$
\begin{aligned}
&\mathscr{A}_h^{\mathcal{D}}(u_h; w_h) - \mathscr{A}_h^{\mathcal{D}}(v_h; w_h) \\
&\quad \geq \kappa^{-1}\|w_h\|_{h,1/2}^2 - \sqrt{1-\varepsilon} \sum_{K\in\mathscr{T}_h} \|D^2 w_h\|_{L^2(K)}\|\Delta w_h\|_{L^2(K)} \\
&\quad = \frac{1}{2}\kappa^{-1}\left(\sum_{K\in\mathscr{T}_h} \|D^2 w_h\|_{L^2(K)}^2 + \|\Delta w_h\|_{L^2(K)}^2 + \sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F^{1/2}\frac{\partial w_h}{\partial n_F}\right\|_{L^2(F)}^2\right) \\
&\quad + \kappa^{-1} c_* J_h(w_h, w_h) - \sqrt{1-\varepsilon} \sum_{K\in\mathscr{T}_h} \|D^2 w_h\|_{L^2(K)}\|\Delta w_h\|_{L^2(K)} \\
&\geq \frac{1}{2}\kappa^{-1}\left(\sum_{K\in\mathscr{T}_h} \|D^2 w_h\|_{L^2(K)}^2 + \|\Delta w_h\|_{L^2(K)}^2 + \sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F^{1/2}\frac{\partial w_h}{\partial n_F}\right\|_{L^2(F)}^2\right) \\
&\quad + \kappa^{-1} c_* J_h(w_h, w_h) - \frac{1}{2}\sum_{K\in\mathscr{T}_h} [\kappa(1-\varepsilon)\|D^2 w_h\|_{L^2(K)}^2 + \kappa^{-1}\|\Delta w_h\|_{L^2(K)}^2] \\
&= \frac{1}{2}\kappa^{-1}\left(\sum_{K\in\mathscr{T}_h} (1-\kappa^2(1-\varepsilon))\|D^2 w_h\|_{L^2(K)}^2 + \sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F^{1/2}\frac{\partial w_h}{\partial n_F}\right\|_{L^2(F)}^2\right) \\
&\qquad\qquad + \kappa^{-1} c_* J_h(w_h, w_h) \\
&\geq \frac{1-\kappa^2(1-\varepsilon)}{2\kappa}\left(\sum_{K\in\mathscr{T}_h} \|D^2 w_h\|_{L^2(K)}^2 + \frac{1}{2}\sum_{F\in\mathscr{E}_h^b} \left\|\mathcal{H}_F^{1/2}\frac{\partial w_h}{\partial n_F}\right\|_{L^2(F)}^2 + c_* J_h(w_h, w_h)\right) \\
&= \frac{1-\kappa^2(1-\varepsilon)}{2\kappa}\|w_h\|_{h,1}^2.
\end{aligned}
$$

Thus we have proven (7.4.1), with $C := \frac{2\kappa}{1-\kappa^2(1-\varepsilon)}$. Thus, the Browder–Minty Theorem yields existence and uniqueness of a $u_h \in V_{h,p}$ such that $\mathscr{A}_h^{\mathcal{D}}(u_h; v_h) = 0$ for all $v_h \in V_{h,p}$. Applying (7.4.1) with $v_h \equiv 0$, we obtain, for the numerical solution u_h,

$$\|u_h\|_{h,1}^2 \leq \tfrac{2\kappa}{1-\kappa^2(1-\varepsilon)}\mathscr{A}_h^{\mathcal{D}}(u_h; u_h) \leq \tfrac{2\kappa\sqrt{d+1}\sup_{\alpha\in\Lambda}\|\gamma^\alpha\|_{L^\infty(\Omega)}}{1-\kappa(1-\varepsilon)}\|\sup_{\alpha\in\Lambda}|f^\alpha|\|_{L^2(\Omega)}\|u_h\|_{h,1}.$$

Dividing by $\|u_h\|_{h,1}$, we obtain (7.4.3). □

Theorem 7.4.2 *Under the hypotheses of Lemma 6.8.2, let $c_{\rm stab}$, η_F, μ_F, and σ_F be chosen so that (6.8.2) holds with $\kappa < (1-\varepsilon)^{-1/2}$. Then, for every $(u_h, \lambda), (v_h, \mu) \in M_h$, we have*

$$\|(u_h - v_h, \lambda - \mu)\|_{h,1}^2 \leq C(\mathscr{A}_h^{\mathcal{O}}((u_h, \lambda); (u_h - v_h, \lambda - \mu)) - \mathscr{A}_h^{\mathcal{O}}((v_h, \mu); (u_h - v_h, \lambda - \mu))), \tag{7.4.5}$$

where the constant $C := 2\kappa/(1-\kappa(1-\varepsilon))$. *Moreover, there exists a constant* C, *independent of* h, *such that for any* $(u_h,\lambda),(v_h,\mu)$, *and* (z_h,ν) *in* M_h,

$$|\mathscr{A}_h^{\mathcal{O}}((u_h,\lambda);(z_h,\nu)) - \mathscr{A}_h^{\mathcal{O}}((v_h,\mu);(z_h,\nu))| \leq C\|(u_h-v_h,\lambda-\mu)\|_{h,1}\|(z_h,\nu)\|_{h,1}. \tag{7.4.6}$$

Therefore, there exists a unique solution $(u_h,c_h)\in M_h$ *to the numerical scheme (7.3.3). Furthermore, we have the bound*

$$\|(u_h,c_h)\|_{h,1} \leq \frac{2\sqrt{3}\kappa \sup_{\alpha\in\Lambda}\|\gamma^\alpha\|_{L^\infty(\Omega)}}{1-\kappa(1-\varepsilon)}\|\sup_{\alpha\in\Lambda}|f^\alpha|\|_{L^2(\Omega)}. \tag{7.4.7}$$

Proof: The proof is analogous to the proof of Theorem 7.4.1, utilising the stability estimate from Lemma 6.8.2, and the boundedness of $B^{\mathcal{O}}_{h,1/2}$, in conjunction with (3.3.19), to prove the monotonicity estimate (7.4.5) and Lipschitz continuity estimate (7.4.6), respectively. As in the proof of Theorem 7.4.1, the Browder–Minty Theorem yields the existence and uniqueness of a pair $(u_h,c_h)\in M_h$ such that $\mathscr{A}_h^{\mathcal{O}}((u_h,c_h);(v_h,\mu_h))=0$ for all $(v_h,\mu_h)\in M_h$. Applying (7.4.5) with $(v_h,c_h)\equiv(0,0)$, yields (7.4.7) for the pair (u_h,c_h). $\square$

7.5 Error estimates

Theorem 7.5.1 *Let* Ω *be a piecewise* C^{m+1} *and piecewise convex domain,* $m\in\mathbb{N}$, $m\geq 2$, *and let* $\{\mathscr{T}_h\}_h$ *be a regular of order* m *family of triangulations on* $\overline{\Omega}$ *satisfying Assumption 4.4.9. Furthermore, let* $u\in H^2(\Omega)\cap H_0^1(\Omega)$ *be the unique solution of (7.2.1), and assume that* $u\in H^{\mathbf{s}}(\Omega;\mathscr{T}_h)$ *with* $s_K>5/2$ *for each* $K\in\mathscr{T}_h$. *Let* c_{stab}, μ_F, *and* η_F *be chosen as in Theorem 5.6.3 for all* $F\in\mathscr{E}_h^{i,b}$. *Then, there exists a positive constant* C *independent of* h *and* u, *but depending on* $\max_K s_K$, *such that for the unique solution* u_h *of (7.3.2), we have*

$$\|u-u_h\|_{h,1}^2 \leq C\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}, \tag{7.5.1}$$

where $t_K=\min(p+1,s_K)$ *for each* $K\in\mathscr{T}_h$.

Proof: This proof is similar to the proof of Theorem 5.7.1, except we must deal with the fact that the operator $\mathscr{A}_h^{\mathcal{D}}$, given by (7.3.1), is nonlinear. Now, let us denote $\xi_h := z_h-u$, and $\psi_h := z_h-u_h$. From Theorem 7.4.1, we have that

$$\|\psi_h\|_{h,1}^2 \leq C(\mathscr{A}_h^{\mathcal{D}}(u_h;\psi_h)-\mathscr{A}_h^{\mathcal{D}}(z_h;\psi_h)) = C(\mathscr{A}_h^{\mathcal{D}}(u;\psi_h)-\mathscr{A}_h^{\mathcal{D}}(z_h;\psi_h)),$$

where the second equality follows from the fact that

$$\mathscr{A}_h^{\mathcal{D}}(u_h; v_h) = 0 \quad \forall v_h \in V_{h,p}, \quad \text{and} \quad \mathscr{A}_h^{\mathcal{D}}(u; v_h) = 0 \quad \forall v_h \in V_{h,p},$$

the latter being a consequence of the consistency result of Lemma 5.5.1.

We see that

$$\begin{aligned}\mathscr{A}_h^{\mathcal{D}}(u; \psi_h) - \mathscr{A}_h^{\mathcal{D}}(z_h; \psi_h) = \sum_{K \in \mathscr{T}_h} \left[\langle F_\gamma[u] - F_\gamma[z_h], \Delta\psi_h \rangle_K - \frac{1}{2} \langle \Delta(u - z_h), \Delta\psi_h \rangle_K \right] \\ + \frac{1}{2} B_{h,*}^{\mathcal{D}}(u - z_h, \psi_h) + J_h(u - z_h, \psi_h).\end{aligned}$$

Firstly, we see that

$$\begin{aligned}&\sum_{K \in \mathscr{T}_h} \langle F_\gamma[u] - F_\gamma[z_h], \Delta\psi_h \rangle_K \\ &\quad \leq \left(\sum_{K \in \mathscr{T}_h} \| F_\gamma[u] - F_\gamma[z_h] \|_{L^2(K)}^2 \right)^{1/2} \left(\sum_{K \in \mathscr{T}_h} \| \Delta\psi_h \|_{L^2(K)}^2 \right)^{1/2} \\ &\quad \leq \sqrt{1 - \varepsilon} \left(\sum_{K \in \mathscr{T}_h} \| D^2 \xi_h \|_{L^2(K)}^2 \right)^{1/2} \| \psi_h \|_{h,1},\end{aligned}$$

and

$$\begin{aligned}-\frac{1}{2} \sum_{K \in \mathscr{T}_h} \langle \Delta(u - z_h), \Delta\psi_h \rangle_K &\leq \left(\sum_{K \in \mathscr{T}_h} \| \Delta\xi_h \|_{L^2(K)}^2 \right)^{1/2} \left(\sum_{K \in \mathscr{T}_h} \| \Delta\psi_h \|_{L^2(K)}^2 \right)^{1/2} \\ &\leq C \left(\sum_{K \in \mathscr{T}_h} \| D^2 \xi_h \|_{L^2(K)}^2 \right)^{1/2} \| \psi_h \|_{h,1}.\end{aligned}$$

By applying the Cauchy-Schwarz inequality, and the second estimate in (4.6.9), and noting our assumptions on η_F and μ_F, we obtain

$$J_h(u - z_h, \psi_h) \leq J_h(\xi_h, \xi_h)^{1/2} J_h(\psi_h, \psi_h)^{1/2} \leq C \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \| u \|_{H^{s_K}(K)} \right)^{1/2} \| \psi_h \|_{h,1}.$$

Furthermore, applying the first estimate in (4.6.9), we obtain

$$\left(\sum_{K \in \mathscr{T}_h} \| D^2 \xi_h \|_{L^2(K)}^2 \right)^{1/2} \leq C \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \| u \|_{H^{s_K}(K)} \right)^{1/2}.$$

Thus, we have obtained

$$\mathscr{A}_h^{\mathcal{D}}(u; \psi_h) - \mathscr{A}_h^{\mathcal{D}}(z_h; \psi_h) \leq C \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \| u \|_{H^{s_K}(K)} \right)^{1/2} \| \psi_h \|_{h,1} + B_{h,1/2}^{\mathcal{D}}(\xi_h, \psi_h). \tag{7.5.2}$$

Estimate (5.7.7) gives us

$$B^{\mathcal{D}}_{h,*}(\xi_h,\psi_h) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{1/2} \|\psi_h\|_{h,1}.$$

Applying the above estimate to (7.5.2), yields

$$\begin{aligned}\|\psi_h\|^2_{h,1} &\leq C(\mathscr{A}^{\mathcal{D}}_h(u;\psi_h) - \mathscr{A}^{\mathcal{D}}_h(z_h;\psi_h))\\ &\leq C\left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{1/2} \|\psi_h\|_{h,1}.\end{aligned}$$

Dividing through by $\|\psi_h\|_{h,1}$ on both sides, we obtain the desired estimate. □

Theorem 7.5.2 *Let $\Omega\subset\mathbb{R}^2$ be a C^2 and piecewise C^{m+1} domain, $m\in\mathbb{N}$, $m\geq 2$, and let $\beta\in C^1(\partial\Omega;\mathbb{S}^1)$. Assume that*

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} > 0 \quad \text{on } \partial\Omega.$$

Furthermore, assume that $\{\mathscr{T}_h\}_h$ is a regular of order m family of triangulations on $\overline{\Omega}$ satisfying Assumption 4.4.9.

Let $(u,c)\in H^2_{\beta,0}(\Omega)\times\mathbb{R}$ be the unique strong solution of (7.2.2). Assume that $u\in H^{\mathbf{s}}(\Omega;\mathscr{T}_h)$ with $s_K>5/2$ for all $K\in\mathscr{T}_h$. Let c_{stab}, c_*, μ_F and σ_F be chosen as in Theorem 6.8.3, and choose $\eta_F\lesssim 1/\tilde{h}_F^3$, $\sigma_F\lesssim 1/\tilde{h}_F$, $F\in\mathscr{E}_h^{i,b}$ and $h_F^{1-\alpha}\lesssim \ell_F$ for all $F\in\mathscr{E}_h^i$, for some $\alpha>2$. Then, there exists a constant $C>0$, independent of h, and u, but depending on $\max_K s_K$, such that Then, there exists a constant $C>0$, independent of h, and u, but depending on $\max_K s_K$, such that for the unique solution pair (u_h,c_h) of (7.3.3), we have

$$\|(u-u_h,c-c_h)\|_{h,1} \leq C\left(\left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|^2_{H^{s_K}(K)}\right)^{1/2} + \frac{\max_{F\in\mathscr{E}_h^i}\tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F\in\mathscr{E}_h^b}\tilde{h}_F}\|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)}\right). \tag{7.5.3}$$

where $t_K := \min\{p+1, s_K, m+1\}$.

Proof: This proof combines the techniques present in the proofs of Theorems 5.7.1 and 6.9.1. Let us denote $\psi_h := u_h - z_h$, $\xi_h := z_h - u$, and $\mu_h = c - c_h$, where (u,c) denotes the true strong solution of (7.2.2), (u_h,c_h) denotes the numerical solution of (7.3.3), and $z_h\in V_{h,p,0}$ is arbitrary (and is later taken to be the element of $V_{h,p,0}$ that satisfies the estimates present in (4.6.9)).

Since $(\psi_h, \mu_h) \in V_{h,p,0} \times V_{h,0}$, we may utilise Lemma 7.4.2, yielding

$$\begin{aligned}\|(\psi_h,\mu_h)\|_{h,1} &\lesssim \mathscr{A}_h^{\mathcal{O}}((u_h,c_h);(\psi_h,\mu_h)) - \mathscr{A}_h^{\mathcal{O}}((z_h,c);(\psi_h,\mu_h)) \\ &= \mathscr{A}_h^{\mathcal{O}}((u,c);(\psi_h,c)) - \mathscr{A}_h^{\mathcal{O}}((z_h,c);(\psi_h,\mu_h)) \\ &= \mathscr{A}_h^{\mathcal{O}}((u,c);(\psi_h,\mu_h)) - \mathscr{A}_h^{\mathcal{O}}((z_h,c);(\psi_h,\mu_h)) + \mathscr{A}_h^{\mathcal{O}}((u,c);(0,c_h)) \\ &= \sum_{K\in\mathscr{T}_h} [\langle F_\gamma[u] - F_\gamma[z_h], \Delta\psi_h\rangle_K - \langle \Delta(u-z_h), \Delta\psi_h\rangle_K] \\ &\quad + B_{h,1/2}^{\mathcal{O}}((u-z_h,0),(\psi_h,\mu_h)) + \mathscr{A}_h^{\mathcal{O}}((u,c);(0,c_h)).\end{aligned} \tag{7.5.4}$$

Note that the first equality follows from the fact that

$$\mathscr{A}_h^{\mathcal{O}}((u,c),(\psi_h,c)) = \mathscr{A}_h^{\mathcal{O}}((u_h,c_h),(\psi_h,\mu_h)) = 0,$$

due to the consistency result of Lemma 6.7.1, and that (u_h, c_h) is the numerical solution of (7.3.3).

As in the proof of Theorem 7.5.1, we have that

$$\begin{aligned}&\sum_{K\in\mathscr{T}_h} [\langle F_\gamma[u] - F_\gamma[z_h], \Delta\psi_h\rangle_K - \langle \Delta(u-z_h), \Delta\psi_h\rangle_K] \\ &\qquad \lesssim \left(\sum_{K\in\mathscr{T}_h} \|D^2\xi_h\|_{L^2(K)}^2\right)^{1/2} \left(\sum_{K\in\mathscr{T}_h} \|\Delta\psi_h\|_{L^2(K)}^2\right)^{1/2} \\ &\qquad \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2} \|(\psi_h,\mu_h)\|_{h,1}.\end{aligned}$$

Furthermore, from the proof of Theorem 6.9.1, we have that

$$B_{h,1/2}^{\mathcal{O}}((u-z_h,0),(\psi_h,\mu_h)) \lesssim \left(\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right)^{1/2} \|(\psi_h,\mu_h)\|_{h,1},$$

as well as

$$\begin{aligned}\mathscr{A}_h^{\mathcal{O}}((u,c);(0,c_h)) &= B_{h,1/2}^{\mathcal{O}}((u,c),(0,c_h)) \\ &\lesssim \frac{\max_{F\in\mathscr{E}_h^i} \tilde{h}_F^{\alpha/2}}{\min_{F\in\mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)} \|(\psi_h,\mu_h)\|_{h,1}.\end{aligned}$$

Combining these estimates, we obtain

$$\|(\psi_h,\mu_h)\|_{h,1}^2 \lesssim \left[\left[\sum_{K\in\mathscr{T}_h} h_K^{2t_K-4}\|u\|_{H^{s_K}(K)}^2\right]^{1/2} + \frac{\max_{F\in\mathscr{E}_h^i} \tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F\in\mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega;\mathscr{T}_h)}\right] \|(\psi_h,\mu_h)\|_{h,1},$$

which, upon dividing through by $\|(\psi_h, \mu_h)\|_{h,1}$, yields

$$\|(\psi_h, \mu_h)\|_{h,1} \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2}.$$

Furthermore, due to our assumptions on μ_F, η_F and σ_F, by applying the estimates in (4.6.9), we obtain

$$\|(\xi_h, 0)\|_{h,1} \lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2} + \frac{\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F \in \mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega; \mathscr{T}_h)},$$

thus

$$\begin{aligned} \|(u - u_h, c - c_h)\|_{h,1} &\leq \|(\xi_h, 0)\|_{h,1} + \|(\psi_h, \mu_h)\|_{h,1} \\ &\lesssim \left(\sum_{K \in \mathscr{T}_h} h_K^{2t_K - 4} \|u\|^2_{H^{s_K}(K)} \right)^{1/2} + \frac{\max_{F \in \mathscr{E}_h^i} \tilde{h}_F^{\frac{\alpha}{2}}}{\min_{F \in \mathscr{E}_h^b} \tilde{h}_F} \|u\|_{H^{\mathbf{s}}(\Omega; \mathscr{T}_h)}, \end{aligned}$$

as desired. □

7.6 Semismooth Newton's method - a practical algorithm

We have proven the existence and uniqueness of a solution to the finite element methods given by (7.3.2) and (7.3.3), but of course, it is necessary to compute such solutions. It is not clear, immediately how one may do this, since the forms on the left-hand side of (7.3.2), and (7.3.3), are both nonlinear in their first argument. Normally, one would apply Newton's method to the operator, arriving at a sequence of problems, but this requires the operator to be differentiable, which it is not (due to the presence of the supremum).

In such a case, one must instead use a "semismooth" Newton's method, which does not rely upon the operator being (classically) differentiable.

For $1 \leq r \leq \infty$, a function $u \in W^{2,r}(\Omega; \mathscr{T}_h)$ defines a vector-valued function $\mathbf{u} \in L^r(\Omega; \mathbb{R}^m)$ through $\mathbf{u} = (u, D_h^2 u)$, where $D_h^2 u$ is the broken Hessian of u. Let $\mathbf{u} = (z, \mathbf{M}) \in \mathbb{R}^m$, and define

$$F_\gamma(x, \mathbf{u}) := \sup_{\alpha \in \Lambda} \{\gamma^\alpha (A^\alpha : \mathbf{M} - f^\alpha)|_x\}. \tag{7.6.1}$$

For each $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, we define

$$\Lambda(x, \mathbf{u}) := \{\alpha \in \Lambda \text{ such that the supremum in (7.6.1) is attained}\}. \tag{7.6.2}$$

This defines a set-valued map $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$.

We now give some final definitions, following [111], that are necessary to describe the algorithm for the semismooth Newton's method. For $u \in W^{2,r}(\Omega; \mathscr{T}_h)$, let

$$\begin{aligned}\Lambda[u] := \{\alpha : \Omega \to \Lambda, \text{ Lebesgue measurable} : \alpha(x) \in \Lambda(x, \mathbf{u}(x)) \\ \text{for a.e. } x \in \Omega, \text{ where } \mathbf{u} = (u, D_h^2 u)\}.\end{aligned} \tag{7.6.3}$$

Lemma 3.3.18 and Theorem 3.3.19 show that $\Lambda[u]$ is nonempty for each $u \in W^{2,r}(\Omega; \mathscr{T}_h)$. For measurable $\alpha : \Omega \to \Lambda$, we define $\gamma^\alpha : \Omega \to \mathbb{R}^+$ through $\gamma^\alpha(x) = \gamma(x, \alpha(x))$, where $\gamma : \Omega \times \mathbb{R} \to \mathbb{R}^+$ was defined by (3.3.15). It follows from the uniform continuity of γ over $\Omega \times \Lambda$ that $\gamma^\alpha \in L^\infty(\Omega)$ with $\|\gamma^\alpha\|_{L^\infty(\Omega)} \leq \|\gamma\|_{C(\overline{\Omega} \times \mathbb{R})}$. The functions A^α, f^α and the operators L^α are defined in a similar manner and are likewise bounded. It is clear that if $\alpha \in \Lambda[u]$, then

$$F_\gamma[u] = \gamma^\alpha (L^\alpha u - f^\alpha).$$

7.6.1 The algorithm

We are now ready to give the definition of the semismooth Newton's method used to approximate solutions of (7.3.2) and (7.3.3), and to provide results regarding the superlinear convergence rates of these methods.

Let us define the linear operators required at each step of the semismooth Newton's method. For a given $u_h^k \in V_{h,p}$, let $\Lambda[u_h^k]$ be given by (7.6.3), and select an *arbitrary* $\alpha_k \in \Lambda[u_h^k]$. We then define

$$\begin{aligned}A_h^{\mathcal{D},k}(u_h, v_h) = &\sum_{K \in \mathscr{T}_h} \langle \gamma^{\alpha_k} A^{\alpha_k} : D^2 u_h, \Delta v_h \rangle_K + B_{h,1/2}^{\mathcal{D}}(u_h, v_h) \\ &- \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K \quad \forall u_h, v_h \in V_{h,p},\end{aligned} \tag{7.6.4}$$

and

$$\begin{aligned}A_h^{\mathcal{O},k}((u_h, \lambda); (v_h, \mu)) = &\sum_{K \in \mathscr{T}_h} \langle \gamma^{\alpha_k} A^{\alpha_k} : D^2 u_h, \Delta v_h \rangle_K + B_{h,1/2}^{\mathcal{O}}((u_h, \lambda); (v_h, \mu)) \\ &- \sum_{K \in \mathscr{T}_h} \langle \Delta u_h, \Delta v_h \rangle_K \quad \forall (u_h, \lambda), (v_h, \mu) \in M_h.\end{aligned} \tag{7.6.5}$$

Since each $\alpha_k : \Omega \to \Lambda$ is measurable, each respective operator given by (7.6.4) and (7.6.5) is well defined. In fact, as in the proof of Theorems 7.4.1 and 7.4.2, the Dirichlet and oblique bilinear forms are coercive on $V_{h,p}$ and $V_{h,p,0} \times V_{h,0}$ respectively,

and for each $k \in \mathbb{N}$, we have

$$\begin{aligned} \|v_h\|_{h,1}^2 &\leq \frac{2\kappa}{1-\kappa(1-\varepsilon)} A_h^{\mathcal{D},k}(v_h, v_h) \quad \forall v_h \in V_{h,p}, \\ \|(v_h,\mu)\|_{h,1}^2 &\leq \frac{2\kappa}{1-\kappa(1-\varepsilon)} A_h^{\mathcal{O},k}((v_h,\mu);(v_h,\mu)) \quad \forall (v_h,\mu) \in V_{h,p,0} \times V_{h,0}. \end{aligned} \tag{7.6.6}$$

Therefore, the sequence of iterations $\{u_h^k\}_{k=1}^\infty$ and $\{(u_h^k, c_h^k)\}_{k=1}^\infty$ generated by the semismooth Newton's method in the Dirichlet and oblique case, respectively, are well defined and remain bounded in $(V_{h,p}, \|\cdot\|_{h,1})$, and $(V_{h,p,0} \times V_{h,0}, \|\cdot\|_{h,1})$.

Algorithm 1 HJB semismooth Newton's method for Dirichlet BVP

Require: $\Omega \subset \mathbb{R}^d$, tol $\in \mathbb{R}^+$, itermax $\in \mathbb{N}$, $\mathscr{T}_h$ a mesh on $\overline{\Omega}$, $V_{h,p}$, Λ, $\{A^\alpha, \gamma^\alpha, f^\alpha\}_{\alpha\in\Lambda}$, $u_h^0 \in V_{h,p}$

$k \leftarrow 0$
$r \leftarrow 1$
$u_h^0 \leftarrow u_h^0$
while $k <$ itermax & $r >$ tol **do**
    $\Lambda_k \leftarrow \Lambda[u_h^k]$ defined by (7.6.3)
    Select an arbitrary $\alpha_k \in \Lambda_k$
    Define $A_h^{\mathcal{D},k}$ by (7.6.4)
    $u_h^{k+1} \leftarrow$ the solution of

$$A_h^{\mathcal{D},k}(u_h^{k+1}, v_h) = \sum_{K\in\mathscr{T}_h} \langle \gamma^{\alpha_k} f^{\alpha_k}, \Delta v_h \rangle_K \quad \forall v_h \in V_{h,p} \tag{7.6.7}$$

    $r \leftarrow \|u_h^{k+1} - u_h^k\|_{L^\infty(\Omega)}$
    $u_h^k \leftarrow u_h^{k+1}$
    $k \leftarrow k+1$
end while

We now state the theorem on the convergence of this method for the Dirichlet problem that is present in [111].

Theorem 7.6.1 *Under the hypotheses of Theorem 7.4.1, there exists a constant $R > 0$, possibly depending on h, such that if $\|u_h - u_h^0\|_{h,1} < R$, where u_h solves (7.3.2), then the sequence $\{u_h^k\}_{k=1}^\infty$ converges to u_h with a superlinear convergence rate.*

Remark 7.6.2 *The proof of Theorem 7.6.1 does not change, since it only relies on the nature of the nonlinear terms present in the definition of $\mathscr{A}_h^{\mathcal{D}}$. However, the proof of the following Theorem (Theorem 7.6.3) requires some changes, due to the fact that we also solve for $c_h \in V_{h,0}$, the approximation of the compatibility constant, c.*

Algorithm 2 HJB semismooth Newton's method for oblique BVP

Require: $\Omega \subset \mathbb{R}^2$, $\beta : \partial\Omega \to \mathbb{S}^1$, tol $\in \mathbb{R}^+$, itermax $\in \mathbb{N}$, $\mathscr{T}_h$ a mesh on $\overline{\Omega}$, $V_{h,p}$, Λ, $\{A^\alpha, \gamma^\alpha, f^\alpha\}_{\alpha\in\Lambda}$, $u_h^0 \in V_{h,p,0}$

$k \leftarrow 0$
$r \leftarrow 1$
$u_h^0 \leftarrow u_h^0$
while $k <$ itermax & $r >$ tol **do**
    $\Lambda_k \leftarrow \Lambda[u_h^k]$ defined by (7.6.3)
    Select an arbitrary $\alpha_k \in \Lambda_k$
    Define $A_h^{\mathcal{O},k}$ by (7.6.5)
    $(u_h^{k+1}, c_h^{k+1}) \leftarrow$ the solution of

$$A_h^{\mathcal{O},k}((u_h^{k+1}, c_h^{k+1}); (v_h, \mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma^{\alpha_k} f^{\alpha_k}, \Delta v_h\rangle_K \quad \forall (v_h, \mu) \in M_h \tag{7.6.8}$$

    $r \leftarrow \|u_h^{k+1} - u_h^k\|_{L^\infty(\Omega)}$
    $u_h^k \leftarrow u_h^{k+1}$
    $k \leftarrow k + 1$
end while

Theorem 7.6.3 *Under the hypotheses of Theorem 7.4.2, there exists a constant $R > 0$, possibly depending on h, such that if $\|(u_h - u_h^0, c_h - c_h^0)\|_{h,1} < R$, where u_h solves (7.3.3), then the sequence $\{(u_h^k, c_h^k)\}_{k=1}^\infty$ converges to (u_h, c_h) with a superlinear convergence rate.*

Proof: Since $\alpha_k \in \Lambda[u_h^k]$, we have $F_\gamma[u_h^k] = \gamma^{\alpha_k} L^{\alpha_k} u_h^k - \gamma^{\alpha_k} f^{\alpha_k}$, therefore

$$A_h^{\mathcal{O},k}((u_h^{k+1}, c_h^{k+1}); (v_h, \mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma^{\alpha_k} f^{\alpha_k}, \Delta v_h\rangle_K \quad \forall (v_h, \mu) \in M_h$$

is equivalent to

$$A_h^{\mathcal{O},k}((u_h^{k+1}, c_h^{k+1}), (v_h, \mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h^k - F_\gamma[u_h^k], \Delta v_h\rangle_K \quad \forall (v_h, \mu) \in M_h. \tag{7.6.9}$$

The definition of the numerical scheme for $\mathscr{A}_h^{\mathcal{O}}$ in (7.3.1) implies that the pair (u_h, c_h) satisfies

$$A_h^{\mathcal{O},k}((u_h, c_h), (v_h, \mu)) = \sum_{K\in\mathscr{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h - F_\gamma[u_h], \Delta v_h\rangle_K \quad \forall (v_h, \mu) \in M_h. \tag{7.6.10}$$

After subtracting (7.6.10) from (7.6.9), and applying the second bound in (7.6.6), we obtain

$$\|(u_h^{k+1} - u_h, c_h^{k+1} - c_h)\|_{h,1} \le C_1 \|F_\gamma[u_h^k] - F_\gamma[u_h] - \gamma^{\alpha_k} L^{\alpha_k}(u_h^k - u_h)\|_{L^2(\Omega)}, \tag{7.6.11}$$

where C_1 depends on κ, ε and γ as in (7.4.7), but not on k.

Fix $r > 2$; since $V_{h,p,0}$ is finite-dimensional, there exists a constant C_2, dependent upon h, such that

$$\|v_h\|_{W^{2,r}(\Omega;\mathscr{T}_h)} \leq C_2\|v_h\|_{h,1,*} \leq C_2\|(v_h,\mu)\|_{h,1} \quad \forall (v_h,\mu) \in M_h,$$

where

$$\|v_h\|_{h,1,*}^2 := |v_h|_{H^2(\Omega;\mathscr{T}_h)}^2 + \frac{1}{2}\sum_{F\in\mathscr{E}_h^b} \|(\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_F)^{1/2}\nabla v_h\|_{L^2(F)}^2$$

$$+ c_* \left(\sum_{F\in\mathscr{E}_h^i} [\eta_F \|[\![v_h]\!]\|_{L^2(F)}^2 + \mu_F\|[\![\nabla_{\mathbf{T}} v_h]\!]\|_{L^2(F)}^2] + \mu_F\|[\![\nabla v_h \cdot n_F]\!]\|_{L^2(F)}^2] \right)$$

is a norm on $V_{h,p,0}$ (the proof of this is analogous to the proof of Lemma 6.8.1).

Theorem 13 from [111] shows that for each $\rho \in (0,1)$, there is an $R_\rho > 0$ such that if $\|w_h - u_h\|_{W^{2,r}(\Omega;\mathscr{T}_h)} < R_\rho$, then, for any $\alpha \in \Lambda[w_h]$,

$$\|F_\gamma[w_h] - F_\gamma[u_h] - \gamma^{\alpha_k} L^{\alpha_k}(w_h - u_h)\|_{L^2(\Omega)} \leq \frac{\rho}{C_1 C_2}\|w_h - u_h\|_{W^{2,r}(\Omega;\mathscr{T}_h)}. \tag{7.6.12}$$

If $\|(u_h^0 - u_h, c_h^0 - c_h)\|_{h,1} < R_\rho$ for some $\rho < 1$, then we use (7.6.11) and (7.6.12) to obtain

$$\|(u_h^{k+1} - u_h, c_h^{k+1} - c_h)\|_{h,1} \leq \rho\|(u_h^k - u_h, c_h^k - c_h)\|_{h,1} \quad \forall k \in \mathbb{N}_0,$$

which yields convergence of (u_h^k, c_h^k) to (u_h, c_h). For each $\rho < 1$, $\|u_h^k - u_h\|_{h,1} < R_\rho$ is then eventually satisfied, thus implying a superlinear convergence rate. $\square$

7.7 Applications to the two-dimensional MA equation

In this Section we discuss the application of the DGFEM given by (7.3.2) to MA Dirichlet boundary-value problems of the following type:

$$\begin{cases} \det D^2 u(x) = f(x), \ x \in \Omega, \\ u(x) = \phi, \ x \in \partial\Omega, \end{cases} \tag{7.7.1}$$

where $f \in C^{0,\alpha}(\overline{\Omega};\mathbb{R}^+)$, is uniformly positive, $\phi \in C^{2,\alpha}(\partial\Omega)$ is the restriction of a function $\tilde{\phi} \in C^{2,\alpha}(\overline{\Omega})$ to $\partial\Omega$, and $\Omega \subset \mathbb{R}^2$ is uniformly convex, with $\partial\Omega \in C^{2,\alpha}$, $\alpha \in (0,1)$. In the case that $\phi \equiv 0$, Theorem 3.5.3 implies the existence and uniqueness

of a uniformly convex $u \in C^{2,\alpha}(\overline{\Omega})$ that is both a solution of (7.7.1), and the following HJB problem:

$$\begin{cases} \sup_{W \in X_\xi} \{-W : D^2 u + (\det W)^{1/2}\psi\} = 0, \text{ in } \Omega, \\ u = 0, \text{ on } \partial\Omega. \end{cases} \tag{7.7.2}$$

where $\psi(x) := 2(f(x))^{1/2}$, and $0 < \xi \leq 1/4$ is a constant depending on the uniform C^2-seminorm of u, and we recall that

$$X_\xi := \{W \in X : \det(W) \geq \xi\},$$

where the Krylov control set, X, is defined by

$$X := \{W \in \mathbb{R}^{2\times 2}_{\text{Sym}} : W \geq 0, \text{Tr}(W) = 1\}.$$

Remark 7.7.1 (Inhomogeneous boundary data) *One may of course be interested in the case that the boundary datum, ϕ, is not identically zero. Indeed, if we assume further that f satisfies the growth condition (3.2.3) and that f, ϕ, and the minimal principal curvature, $\kappa_{\partial\Omega}$, of $\partial\Omega$ satisfy (3.2.4), then Theorem 3.3.20 guarantees the existence of a uniformly convex solution $u \in C^{2,\alpha}(\overline{\Omega})$ of (7.7.1). Furthermore, under the same hypotheses, Theorem 3.2.1 guarantees existence and uniqueness of a function $v \in H^2(\Omega)$ that is a strong solution of the HJB problem:*

$$\begin{cases} \sup_{W \in X_\xi} \{W : D^2 v + (\det W)^{1/2}\psi\} = 0, \text{ in } \Omega, \\ v = -\phi, \text{ on } \partial\Omega, \end{cases}$$

for any $0 < \xi \leq 1/4$ (note that in particular, Theorem 3.3.20 only requires that Ω is convex, $\psi \in L^2(\Omega)$, and that ϕ is the trace of a $H^{3/2}(\Omega)$ function). As such, under these assumptions, inhomogeneous Dirichlet boundary conditions should not pose any additional difficulties. Furthermore, in general, we may implement inhomogeneous Dirichlet boundary conditions numerically by modifying the form $\mathscr{A}_h^{\mathcal{D}}$ follows:

$$\mathscr{A}^{\mathcal{D},\text{inhom}}(u_h; v_h) := \mathscr{A}_h^{\mathcal{D}}(u_h; v_h) - \sum_{F \in \mathscr{E}_h^b} [\mu_F \langle \nabla_{\mathbf{T}} g, \nabla_{\mathbf{T}} v_h \rangle_F + \eta_F \langle g, v_h \rangle_F]$$
$$+ \frac{1}{2} \sum_{F \in \mathscr{E}_h^b} [\langle \text{div}_{\mathbf{T}} \nabla_{\mathbf{T}} g, \nabla v_h \cdot n_F \rangle_F + \langle \nabla_{\mathbf{T}} (\nabla v_h \cdot n_F), \nabla_{\mathbf{T}} g \rangle_F - \langle \nabla_{\mathbf{T}} g \, \nabla_{\mathbf{T}} n_F^T \nabla_{\mathbf{T}} v_h \rangle_F],$$

where g is the restriction of the boundary data to $\partial\Omega$. We then then seek $u_h \in V_{h,p}$ such that

$$\mathscr{A}_h^{\mathcal{D},\text{inhom}}(u_h; v_h) = 0 \quad \forall v_h \in V_{h,p}.$$

One retains a HJB problem with Cordes coefficients from (7.7.2), simply by letting $v := -u$, which gives us

$$\begin{cases} \sup\limits_{W \in X_\xi} \{W : D^2 v + (\det W)^{1/2} \psi\} = 0, \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad v = 0, \text{ on } \partial\Omega. \end{cases} \tag{7.7.3}$$

One can see that we have rewritten the MA equation (7.7.1) (with $\phi \equiv 0$) as a HJB equation, and thus we can use the semismooth Newton's method defined by (7.6.1). It is important to note that for $\xi > 0$, X_ξ consists of positive definite matrices, and since $d = 2$, they satisfy the Cordes condition. Moreover, the following calculation shows that since the matrices have unit trace, there is in fact no need to renormalise the problem by multiplying through by

$$\gamma^\alpha := \frac{\operatorname{Tr}(A^\alpha)}{|A^\alpha|^2}.$$

One can see that for any $W \in X_\xi$,

$$\begin{aligned} |W - I_d|^2 &= |W|^2 - 2I_d : W + |I_d|^2 \\ &= |W|^2 - 2\operatorname{Tr}(W) + 2 \\ &= |W|^2 \\ &= (\operatorname{Tr}(W))^2 - 2\det W \\ &= 1 - 2\det W \\ &\leq 1 - 2\xi < 1. \end{aligned} \tag{7.7.4}$$

Thus we have the following lemma.

Lemma 7.7.2 *Let Ω be a bounded open subset of $\mathbb{R}^2$. Then, for any open set $U \subset \Omega$, any $u, v \in H^2(\Omega)$, and any $\xi \in (0, 1/4]$ the following inequality holds a.e. in U:*

$$\left| \sup_{W \in X_\xi} \{W : D^2 u - f^\alpha\} - \sup_{W \in X_\xi} \{W : D^2 v - f^\alpha\} - \Delta(u - v) \right| \leq \sqrt{1 - \varepsilon} |D^2(u - v)|, \tag{7.7.5}$$

with $\varepsilon = 2\xi$.

Proof: Denoting $w := u - v$, one can see that

$$\begin{aligned} \left| \sup_{W \in X_\xi} \{W : D^2 u - f^\alpha\} - \sup_{W \in X_\xi} \{W : D^2 v - f^\alpha\} - \Delta w \right| &\leq \sup_{W \in X_\xi} \left\{ \left| W : D^2 w - \Delta w \right| \right\} \\ &= \sup_{W \in X_\xi} \left\{ \left| (W - I_d) : D^2 w \right| \right\} \\ &\leq \sup_{W \in X_\xi} \{|W - I_d|\} \, |D^2 w|. \end{aligned}$$

Noting that the estimate (7.7.4) is uniform in W over X_ξ, we obtain the desired result. $\square$

Remark 7.7.3 *Removing the renormalisation factor, γ^α, from the HJB equation does not affect any of the existence and uniqueness results, nor does it affect the well-posedness of the algorithm for the semismooth Newton's method. This holds, due to the fact that the necessity of including the renormalisation factor, γ^α, was solely to obtain an estimate of the same form as (7.7.5), so that one can utilise the "addition-subtraction" of the Laplacian technique, followed by an application of the Miranda–Talenti estimate.*

Remark 7.7.4 (The maximisation problem) *At an arbitrary step $k \in \mathbb{N}$ of the semismooth Newton's method, we are required to calculate $\Lambda[u_h^k]$. In the case of the MA–HJB equation (7.7.3), this will be denoted by $X_\xi[u_h^k]$. Implicitly, this involves solving a maximisation problem for each $k \in \mathbb{N}$. As we will see momentarily, in the case of the MA–HJB equation (7.7.3), this can by done analytically, in the case that we* do not *multiply through by the renormalisation factor, γ^α.*

Notice that for a given $v_h \in V_{h,\mathbf{p}}$, we have that $D_h^2 v_h(x) \in \mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ for any $x \in K$, for all $K \in \mathscr{T}_h$. It follows that the maximisation problem of calculating $X_\xi[v_h(x)]$ at a given $x \in \Omega$, is equivalent to finding $W^* \in Y$, satisfying

$$W^* : M + (\det W^*)^{1/2}\psi = \sup_{W\in X_\xi}\{W : M + (\det W)^{1/2}\psi\}, \tag{7.7.6}$$

for some fixed $M \in \mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ and $\psi \in \mathbb{R}^+$.

Lemma 7.7.5 *For each $M \in \mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ and $\psi \in \mathbb{R}^+$, there exists a $W^* = W^*(M,\psi) \in X_\xi$ satisfying the maximisation problem given by (7.7.6). A maximiser is given by*

$$W^* = \begin{bmatrix} \frac{1}{2}(1 + \frac{M_{11}-M_{22}}{\sqrt{\psi^2+(M_{11}-M_{22})^2+4M_{12}^2}}) & \frac{M_{12}}{\sqrt{\psi^2+4M_{12}^2}}\sqrt{1-\frac{(M_{11}-M_{22})^2}{(\psi^2+(M_{11}-M_{22})^2+4M_{12}^2)}} \\ \frac{M_{12}}{\sqrt{\psi^2+4M_{12}^2}}\sqrt{1-\frac{(M_{11}-M_{22})^2}{(\psi^2+(M_{11}-M_{22})^2+4M_{12}^2)}} & 1-\frac{1}{2}(1 + \frac{M_{11}-M_{22}}{\sqrt{\psi^2+(M_{11}-M_{22})^2+4M_{12}^2}}) \end{bmatrix}, \tag{7.7.7}$$

unless W^* does not belong to X_ξ, in which case a maximiser is given by

$$\begin{bmatrix} \frac{1}{2}\left(1 + (M_{11}-M_{22})\sqrt{\frac{1-4\xi}{4M_{12}^2+(M_{11}-M_{22})^2}}\right) & \mathrm{sgn}(M_{12})\sqrt{\frac{1}{4}(1-\frac{(M_{11}-M_{22})^2(1-4\xi)}{4M_{12}^2+(M_{11}-M_{22})^2}) - \xi} \\ \mathrm{sgn}(M_{12})\sqrt{\frac{1}{4}(1-\frac{(M_{11}-M_{22})^2(1-4\xi)}{4M_{12}^2+(M_{11}-M_{22})^2}) - \xi} & 1-\frac{1}{2}\left(1 + (M_{11}-M_{22})\sqrt{\frac{1-4\xi}{4M_{12}^2+(M_{11}-M_{22})^2}}\right) \end{bmatrix}. \tag{7.7.8}$$

Proof: Notice that $X_\xi = \{W \in \mathbb{R}^{2\times 2}_{\text{Sym}} : \operatorname{Tr} W = 1,\ \det W \geq \xi\}$ is in fact isomorphic to the set

$$\left\{(\lambda,\alpha) \in \left[\frac{1}{2}(1-\sqrt{1-4\xi}), \frac{1}{2}(1+\sqrt{1-4\xi})\right] \times \left[-\frac{1}{2},\frac{1}{2}\right] : \lambda(1-\lambda) \geq \alpha^2 + \xi\right\} =: X'_\xi. \tag{7.7.9}$$

This follows, since $W \in X_\xi$ can be represented by

$$\begin{bmatrix} \lambda & \alpha \\ \alpha & 1-\lambda \end{bmatrix},$$

where the values $(\lambda,\alpha) \in [0,1] \times [-1/2, 1/2]$ are constrained by $\det W = \lambda(1-\lambda) - \alpha^2 \geq \xi$.

Now given $M = [M_{ij}]_{i,j=1,2} \in \mathbb{R}^{2\times 2}_{\text{Sym}}$ and $\psi \in \mathbb{R}^+$, let us define $\eta : X'_\xi \to \mathbb{R}$ by

$$\eta(\lambda,\alpha) = \lambda M_{11} + (1-\lambda)M_{22} + 2\alpha M_{12} + (\lambda(1-\lambda)-\alpha^2)^{1/2}\psi. \tag{7.7.10}$$

We see that

$$\begin{aligned}
\partial_\lambda \eta(\lambda,\alpha) &= M_{11} - M_{22} + \frac{1}{2}(1-2\lambda)(\lambda(1-\lambda)-\alpha^2)^{-1/2}\psi,\\
\partial_\alpha \eta(\lambda,\alpha) &= 2M_{12} - \alpha(\lambda(1-\lambda)-\alpha^2)^{-1/2}\psi,\\
\partial^2_{\lambda,\alpha}\eta(\lambda,\alpha) &= \frac{1}{2}\alpha(1-2\lambda)(\lambda(1-\lambda)-\alpha^2)^{-3/2}\psi,\\
\partial^2_{\lambda,\lambda}\eta(\lambda,\alpha) &= -((\lambda(1-\lambda)-\alpha^2)^{-1/2} + \frac{1}{4}(1-2\lambda)^2(\lambda(1-\lambda)-\alpha^2)^{-3/2})\psi,\\
\partial^2_{\alpha,\alpha}\eta(\lambda,\alpha) &= -((\lambda(1-\lambda)-\alpha^2)^{-1/2} + \alpha^2(\lambda(1-\lambda)-\alpha^2)^{-3/2})\psi.
\end{aligned}$$

We also see that $\partial^2_{\alpha,\alpha}\eta,\ \partial^2_{\lambda,\lambda}\eta < 0$, and that

$$\begin{aligned}
\det(D^2\eta) &= \partial^2_{\alpha,\alpha}\eta\partial^2_{\lambda,\lambda} - (\partial^2_{\lambda,\alpha}\eta)^2\\
&= ((\lambda(1-\lambda)-\alpha^2)^{-1} + \frac{1}{4}(1-2\lambda)^2(\lambda(1-\lambda)-\alpha^2)^{-2} + \alpha^2(\lambda(1-\lambda)-\alpha^2)^{-2})\psi^2\\
&\quad + (\frac{1}{4}\alpha^2(1-2\lambda)^2(\lambda(1-\lambda)-\alpha^2)^{-3} - \frac{1}{4}\alpha^2(1-2\lambda)^2(\lambda(1-\lambda)-\alpha^2)^{-3})\psi^2\\
&= ((\lambda(1-\lambda)-\alpha^2)^{-1} + \frac{1}{4}(1-2\lambda)^2(\lambda(1-\lambda)-\alpha^2)^{-2} + \alpha^2(\lambda(1-\lambda)-\alpha^2)^{-2})\psi^2 > 0.
\end{aligned}$$

Since $\eta \in C^\infty(X'_\xi)$, it follows that any critical point of η is in fact a local maximum. Critical points occur where

$$\partial_\lambda \eta(\lambda,\alpha) = \partial_\alpha \eta(\lambda,\alpha) = 0. \tag{7.7.11}$$

The first equality in (7.7.11) gives us

$$M_{11} - M_{22} + \frac{1}{2}(1-2\lambda)(\lambda(1-\lambda)-\alpha^2)^{-1/2}\psi = 0,$$

which is satisfied by

$$\lambda = \frac{1}{2}\left(1 \pm |M_{11} - M_{22}|\sqrt{(1-4\alpha^2)/(\psi^2 + (M_{11}-M_{22})^2)}\right). \tag{7.7.12}$$

The second equality in (7.7.11) gives us

$$2M_{12} - \alpha(\lambda(1-\lambda) - \alpha^2)^{-1/2}\psi = 0,$$

which is satisfied by

$$\alpha = \operatorname{sgn}(M_{12})\sqrt{\frac{4M_{12}^2\lambda(1-\lambda)}{\psi^2 + 4M_{12}^2}}. \tag{7.7.13}$$

Substituting (7.7.13) into (7.7.12) and solving the resulting equation for λ, we obtain

$$\lambda^{\pm} = \frac{1}{2}(1 \pm |M_{11} - M_{22}|/\sqrt{\psi^2 + (M_{11}-M_{22})^2 + 4M_{12}^2});$$

then, substituting this back into (7.7.13), we find that

$$\alpha = \frac{\operatorname{sgn}(M_{12})|M_{12}|}{\sqrt{\psi^2 + 4M_{12}^2}}\sqrt{1 - (M_{11}-M_{22})^2/(\psi^2 + (M_{11}-M_{22})^2 + 4M_{12}^2)}.$$

Notice that the value of $\lambda^{\pm}$ in (7.7.12) is determined up to a sign. Since the determinant of the matrix represented by $(\lambda^{\pm}, \alpha)$ is invariant under this change of sign, it is clear that the maximising value is determined by the maximiser of

$$\lambda M_{11} + (1-\lambda)M_{22}.$$

We see that if $M_{11} - M_{22} > 0$, then the maximiser satisfies $\lambda \geq 1 - \lambda$, which occurs when we take the value $\lambda = \lambda^+$, similarly when $M_{11} - M_{22} < 0$, we take $\lambda = \lambda^-$. This tells us that in either case

$$\begin{aligned}\lambda &= \frac{1}{2}\left(1 + \operatorname{sign}(M_{11}-M_{22})|M_{11}-M_{22}|/\sqrt{\psi^2 + (M_{11}-M_{22})^2 + 4M_{12}^2}\right) \\ &= \frac{1}{2}\left(1 + (M_{11}-M_{22})/\sqrt{\psi^2 + (M_{11}-M_{22})^2 + 4M_{12}^2}\right).\end{aligned}$$

Since $\mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ is unbounded, it may be the case that given $M \in \mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ the critical point of η may lie outside of X'_ξ. For instance, we can always construct a sequence $\{M^k\}_{k=1}^{\infty} \subset \mathbb{R}^{2\times 2}_{\mathrm{Sym}}$ such that $\lambda^k = \lambda(M^k)$ given by (7.7.12) tends to 1 as $k \to \infty$, and $\alpha^k = \alpha(M^k)$, given by (7.7.13) is zero for all members of the sequence. This limit point is not in X'_ξ.

Let us consider the case when a critical point $(\lambda_*, \alpha_*) \in [0,1] \times [-1/2, 1/2] \setminus X'_\xi$. Since η is smooth, X'_ξ is compact and η does not have a critical point in X'_ξ, it

must be that any maximising point of X'_ξ must in fact lie on the boundary of X'_ξ, $\partial X'_\xi = \{[\lambda, \alpha] \in X'_\xi : \lambda(1-\lambda) - \alpha^2 = \xi\}$. We now aim to find a critical point of $\eta|_{\partial X'_\xi}$, that is, where the tangential derivative,

$$\partial_\tau \eta = \nabla \eta \cdot \tau, \tag{7.7.14}$$

is zero. We can represent $\partial X'_\xi$ as the level set

$$\{[\lambda, \alpha] \in X'_\xi : \phi(\lambda, \alpha) = \xi\},$$

where $\phi(\lambda, \alpha) = \lambda(1-\lambda) - \alpha^2$. The level set has the unit normal vector $n = \nabla\phi/|\nabla\phi|$, and corresponding tangential vector

$$\tau = \left[\frac{-\partial_\alpha \phi}{|\nabla\phi|}, \frac{\partial_\lambda \phi}{|\nabla\phi|}\right]^T = \left[\frac{2\alpha}{|\nabla\phi|}, \frac{1-2\lambda}{|\nabla\phi|}\right]^T.$$

Moreover, on $\partial X'_\xi$

$$\begin{aligned} |\nabla\phi| = \sqrt{(1-2\lambda)^2 + (2\alpha)^2} &= \sqrt{1 - 4(\lambda(1-\lambda) - \alpha^2)} \\ &= \sqrt{1 - 4\phi(\lambda, \alpha)} = \sqrt{1 - 4\xi} > 0. \end{aligned}$$

We then see that

$$\partial_\tau(\eta|_{\partial X'_\xi}) = [2\alpha(M_{11} - M_{22}) + 2M_{12}(1-2\lambda)]/\sqrt{1-4\xi},$$

and,

$$\begin{aligned} \partial^2_{\tau,\tau}(\eta|_{\partial X'_\xi}) &= \nabla(\partial_\tau(\eta|_{\partial X'_\xi})) \cdot \tau \\ &= \nabla((2\alpha(M_{11} - M_{22}) + 2M_{12}(1-2\lambda))/\sqrt{1-4\xi}) \cdot \tau \\ &= [2(1-2\lambda)(M_{11} - M_{22}) - 8M_{12}\alpha]/(1-4\xi). \end{aligned} \tag{7.7.15}$$

Note that on $\partial X'_\xi$, we have that $\alpha = \pm\sqrt{\lambda(1-\lambda) - \xi}$. Substituting these values into (7.7.14), we obtain

$$\pm\sqrt{\lambda(1-\lambda) - \xi}(M_{11} - M_{22}) + M_{12}(1-2\lambda) = 0. \tag{7.7.16}$$

Let us consider separately the case when $M_{12} = 0$, i.e., when M is a diagonal matrix. Let us assume further that $M_{11} \neq M_{22}$, since otherwise η is maximised by $(1/2)I_d \in X'_\xi$, and the point $(1/2, 1/2)$, that it represents, does not lie outside of X'_ξ. We see that in this case, (7.7.16) is satisfied by

$$\lambda = \frac{1}{2}(1 \pm \sqrt{1-4\xi}),$$

which implies that $\alpha = 0$.

For the critical point to be a maximum, we require that $\partial^2_{\tau,\tau}\eta$, given by (7.7.15), is negative. This holds for

$$(\lambda, \alpha) = \left(\frac{1}{2}(1 + \operatorname{sgn}(M_{11} - M_{22})\sqrt{1 - 4\xi}), 0\right). \tag{7.7.17}$$

Now we consider the case when $M_{11} - M_{22} = 0$, and $M_{12} \neq 0$. In this case, $\eta|_{\partial X'_\xi}$ is constant in its first variable, and since $\alpha = \pm\sqrt{\lambda(1-\lambda) - \xi}$, the maximiser is given by $(1/2, \operatorname{sgn}(M_{12})\sqrt{1/4 - \xi})$.

Finally we turn to the case when $M_{11} - M_{22} \neq 0$, and $M_{12} \neq 0$. In this case, (7.7.16) is satisfied by

$$\lambda = \frac{1}{2}\left(1 \pm \sqrt{\frac{(M_{11} - M_{22})^2(1 - 4\xi)}{4M_{12}^2 + (M_{11} - M_{22})^2}}\right) \in \left[\frac{1}{2}(1 - \sqrt{1 + 4\xi}), \frac{1}{2}(1 - \sqrt{1 - 4\xi})\right]. \tag{7.7.18}$$

Note that (7.7.18) holds, since we have assumed that $M_{12}, M_{11} - M_{22} \neq 0$, and thus

$$\frac{(M_{11} - M_{22})^2}{4M_{12}^2 + (M_{11} - M_{22})^2} < 1.$$

To ensure that the second tangential derivative given by (7.7.15) is negative (so that the critical points are maxima), we obtain that the corresponding (λ, α) are given by

$$\lambda = \frac{1}{2}\left(1 + \operatorname{sgn}(M_{11} - M_{22})\sqrt{\frac{(M_{11} - M_{22})^2(1 - 4\xi)}{4M_{12}^2 + (M_{11} - M_{22})^2}}\right),$$
$$\alpha = \operatorname{sgn}(M_{12})\sqrt{\frac{1}{4}(1 - \frac{(M_{11} - M_{22})^2(1 - 4\xi)}{4M_{12}^2 + (M_{11} - M_{22})^2}) - \xi}. \qquad \square$$

7.8 Implementation

Software and code: The experiments in this Chapter have been implemented in the most recent version of the Firedrake software [105, 87] (as of 3rd July 2018), which interfaces directly with PETSc [6, 7] running through a Python interface [39, 63]. Two working Firedrake scripts, MA-HJB-Dirichlet.py and MA-HJB-oblique.py used to generate the experiments of this Chapter are available in the Github repository: https://github.com/ekawecki/FiredrakeNDV.

Linear systems and condition numbers: Each step of the semismooth Newton's methods given by Algorithms 1 and 2 involves solving a linear problem of the form (7.6.7)

and (7.6.8), respectively. As discussed in Chapters 5 and 6, the resulting coefficient matrix of each linear system typically has a Euclidean norm condition number of order h^{-4}. This can pose difficulties when applying iterative methods to solve the linear system, and thus to ensure that we solve the linear system with sufficiently high accuracy as the mesh size h decreases, we apply the Iterative refinement algorithm, i.e., Algorithm 1.1 of [32]. We implement the Iterative refinement algorithm by using the following choices in the Firedrake "solve" function. Note that the first code snippet is for the Dirichlet boundary-value problem, and the second code snippet is for the oblique boundary-value problem.

```
t = time()
solve(A_gamma == L, U,
      solver_parameters = {
      "snes_type": "newtonls",
      "ksp_type": "preonly",
      "pc_type": "lu",
      "snes_monitor": False,
      "snes_rtol": 1e-16,
      "snes_atol": 1e-25})
tt.append(time()-t)
```

```
# implementing nullspace, as solution should have zero sum
V_basis = VectorSpaceBasis(constant=True)
nullspace = MixedVectorSpaceBasis(S, [V_basis, S[1]])

# begin timing of linear system solve
t = time()

# solving linear system
solve(A_gamma == L,Uh,nullspace = nullspace,
      solver_parameters = {"mat_type": "aij",
      "snes_type": "newtonls",
      "ksp_type": "preonly",
      "pc_type": "lu",
      "snes_monitor": False,
      "snes_rtol": 1e-16,
      "snes_atol": 1e-25})
# end timing of linear system solve
tt.append(time()-t)
```

Furthermore, the solver choices in the second code snippet above differ slightly from those present in the first code snippet, i.e., we also include "nullspace = nullspace", where "nullspace" is defined on line 3, and "mat_type = aij". The first choice imposes that the numerical solution u_h^{k+1} (from the pair $(u_h^{k+1}, c_h^{k+1}) \in M_h$ that satisifies (7.6.8)) has a zero-sum, and the latter essentially informs the solver that the solution consists of two parts, i.e., u_h^{k+1} and c_h^{k+1}, and that the system may be treated in block formation.

Two-dimensional curved boundary approximation: When implementing curved

finite elements, we use a piecewise quadratic polynomial mapping to obtain a higher order approximation of the domain boundary. This is implemented in exactly the same manner as discussed in Section 5.8. As in Section 5.8, we define the space $V_{h,p}^{\text{comp}} := \{v \in L^2(\Omega) : v \circ T^{-1} \in \mathbb{P}^p(\hat{K})\}$, where the piecewise quadratic function T is defined by (5.8.1). In this case, we then define $V_{h,p,0}^{\text{comp}} := V_{h,p}^{\text{comp}} \cap L_0^2(\Omega)$ and $M_h^{\text{comp}} := V_{h,p,0}^{\text{comp}} \times V_{h,0}^{\text{comp}}$.

Furthermore, when we refine the mesh in our experiments, the meshes at each refinement level are not related to one another. That is, there is no hierarchical mesh structure, i.e., at each refinement level, we "remesh". A collection of the meshes used for the computations of this thesis can be found in the folder "Meshes" in the Github repository: https://github.com/ekawecki/FiredrakeNDV.

Selection process: One can see that lines 5-6 of Algorithms 1 and 2 require one to define the control set $\Lambda_k := \Lambda[u_h^k]$, and then to select an arbitrary $\alpha_k \in \Lambda_k$. This is a core step in the semismooth Newton's method. For the examples that we consider, given $u_h^k \in V_{h,p}^{\text{comp}}$, we can solve the corresponding maximisation problem (7.6.3) that defines $\Lambda[u_h^k]$. In particular, when considering the MA problem, the corresponding pointwise maximisation problem is given by (7.7.6), for which (7.7.7)–(7.7.8) provides a solution. Once a maximiser is known, we are then required to provide an *arbitrary* $\alpha_k \in \Lambda_k$, and so we simply choose the maximiser that we have calculated. This process is implemented for the MA problem in the following code snippet.

```
        # calculating right-hand side function f for the MA problem
        d2udxx = convup*(Constant(1.0)-alp*0.5*pi*cos(pi*rho)+alp*pi**2*x**2*sin(pi↩
            *rho))
        d2udxy = convup*(Constant(0.0)+alp*pi**2*x*y*sin(pi*rho))
        d2udyy = convup*(Constant(1.0)-alp*0.5*pi*cos(pi*rho)+alp*pi**2*y**2*sin(pi↩
            *rho))
        f = 2.0*pow(d2udxx*d2udyy-pow(d2udxy,2.0),0.5)
        # Defining optimal controls for the MA-HJB problem
        xi = 1.0/100.0
        def controls(u00,u01,u11):
            hh = f
            diff = u00-u11
            c00w = 0.5*(1+(diff)/sqrt(hh**2.0+pow(diff,2.0)+4.0*pow(u01,2.0)))
            c11w = 0.5*(1-(diff)/sqrt(hh**2.0+pow(diff,2.0)+4.0*pow(u01,2.0)))
            c01w = (u01/sqrt(hh**2.0+4.0*pow(u01,2.0)))*sqrt(1-(pow(diff,2.0))/(hh↩
                **2+pow(diff,2.0)+4.0*pow(u01,2.0)))
            c00xi = 0.5*(1+diff*sqrt((1-4.0*xi)/(1e-15+4.0*pow(u01,2)+pow(diff,2)))↩
                )
            c11xi = 0.5*(1-diff*sqrt((1-4.0*xi)/(1e-15+4.0*pow(u01,2)+pow(diff,2)))↩
                )
            c01xi = signfct(u01)*sqrt(0.25*(1-(pow(diff,2)*(1-4.0*xi))/(1e-15+4*pow↩
                (u01,2)+pow(diff,2)))-xi)
            det = 0.25*pow(hh,2)/(pow(hh,2)+pow(diff,2)+4.0*pow(u01,2))
            zeta = signfct(makemax(det-xi,0))
            cont = [[zeta*c00w+(1-zeta)*c00xi,zeta*c01w+(1-zeta)*c01xi],[zeta*c01w↩
                +(1-zeta)*c01xi,zeta*c11w+(1-zeta)*c11xi]]
            fcont = -hh*pow(cont[0][0]*cont[1][1]-cont[1][0]*cont[0][1],0.5)
            return cont, fcont
```

Remark 7.8.1 (Computational parameters) *In the following experiments, we consider both Dirichlet and oblique boundary-value problems (BVPs). In the Dirichlet BVP case, we employ the following parameter choices:* $c_{\text{stab}} = 2$, $\mu_F = c_{\text{stab}}(p-1)^2/2\tilde{h}_F$, $\eta_F = 3c_{\text{stab}}(p-1)^4/8\tilde{h}_F^3$. *In the oblique BVP case, we employ the following parameter choices:* $c_{\text{stab}} = 2.5$, $\mu_F = 2c_{\text{stab}}(p-1)^2/\tilde{h}_F$, $\eta_F = 15(p-1)^4/16\tilde{h}_F^3$, $\sigma_F = 2c_{\text{stab}}p^2/\tilde{h}_F^2$, *and* $\ell_F = c_{\text{stab}}\tilde{h}_F^{-3}$.

In both cases, the choice of parameters was guided by results of the experiments of Chapters 5 and 6 (since, in the Dirichlet BVP and oblique BVP case, each step of the semismooth Newton's method requires the solution of a problem of the form similar to the linear problems considered in Sections 5.9 and 6.11, respectively). Note that in Remarks 5.8.1 and 6.10.1 we discuss the choice of computational parameters for the corresponding linear problems.

7.9 Experiments

In the following experiments, we successively increase the degree, p, of the finite element space $V_{h,p}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly. We apply the semismooth Newton's method with an initial guess given by

$$u_h^0 := \cos(\pi x)\cos(\pi y),$$

until the step increment $\|\cdot\|_{L^2(\Omega)}$-norm reaches a tolerance of 10^{-12} (unless stated otherwise). Note that we denote by u_h the final Newton iterate, i.e., for the index N, we have that $u_h = u_h^N$, and $\|u_h^N - u_h^{N-1}\|_{L^2(\Omega)} <$ tolerance.

7.9.1 Experiment 1

In this experiment, we consider the following MA problem:

$$\begin{cases} \det D^2 u(x) = f(x) & x \in \Omega, \\ u(x) = 0 & x \in \partial\Omega, \end{cases} \tag{7.9.1}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and f is chosen so that the true solution of (7.9.1) is given by

$$u(x,y) = 5(x^2+y^2-1) - \frac{1}{8}\sin(\pi(x^2+y^2)).$$

In order to approximate the solution of (7.9.1), we consider the equivalent HJB problem:

$$\begin{cases} \sup_{W \in X_\xi} \{-W : D^2 u + 2(\det W)^{1/2} f^{1/2}\} = 0, \text{ in } \Omega, \\ u = 0, \text{ on } \partial\Omega, \end{cases} \tag{7.9.2}$$

where $\xi = 1/100$, and apply the semismooth Newton's method, given by Algorithm 1, to (7.9.2). Since each $W \in X_\xi$ has unit trace, we use the renormalisation parameter $\gamma^\alpha = 1$ (as opposed to $\gamma^\alpha := \mathrm{Tr}(A^\alpha)/|A^\alpha|^2$, see Remark 7.7.3).

Furthermore, since Ω is the unit disk, $\partial\Omega = \mathbb{S}^1$, and it follows that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore, $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. For the internal faces, the mean curvature is calculated directly as $\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F$, where n_F is a fixed choice of unit normal to F.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u - u_h^N\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 7.1, and report the exact values in Table 7.1, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 7.2.

We also plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 7.2, 7.3, and 7.4, respectively.

Across all polynomial degrees and mesh refinements we see a small variation in the number of Newton iterations required to reach the desired tolerance. In particular, in each case we observe that reaching the desired tolerance of 10^{-12} requires either 5 or 6 iterations.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	5.87		1.36		7.76×10^{-1}	
0.2828	1.88	(2.01)	6.68×10^{-1}	(1.25)	1.57×10^{-1}	(2.82)
0.1627	1.04	(1.08)	2.49×10^{-1}	(1.78)	2.51×10^{-2}	(3.32)
0.0973	5.79×10^{-1}	(1.13)	7.52×10^{-2}	(2.33)	4.63×10^{-3}	(3.28)
0.0508	2.98×10^{-1}	(1.02)	2.00×10^{-2}	(2.04)	6.42×10^{-4}	(3.04)
0.0269	1.55×10^{-1}	(1.03)	5.21×10^{-3}	(2.11)	1.05×10^{-4}	(2.84)
0.0138	7.77×10^{-2}	(1.03)	1.31×10^{-3}	(2.06)	1.78×10^{-5}	(2.65)

Table 7.1: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 7.9.1.

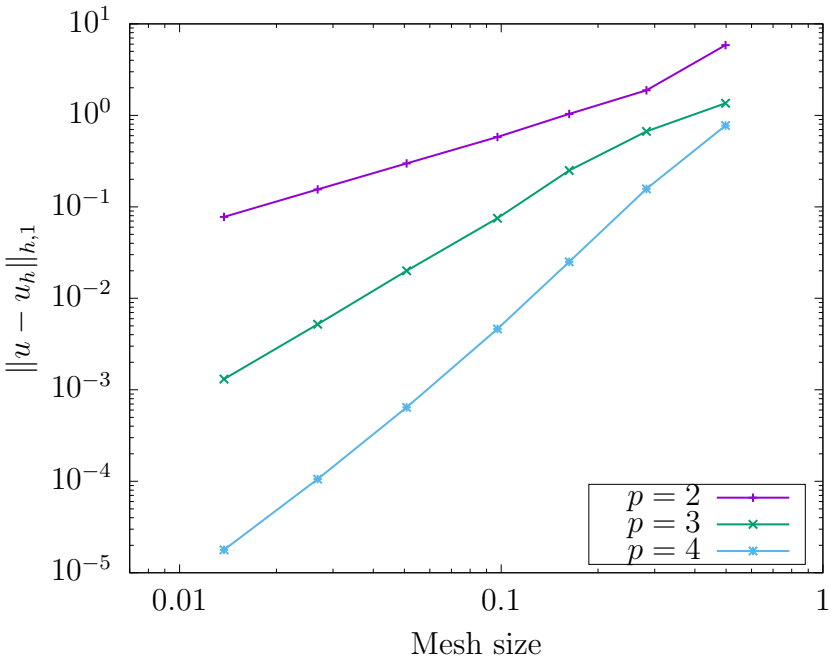


Figure 7.1: Convergence rates for the numerical scheme applied to problem (7.9.2). The error $\|u-u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p = 2$ to $p = 4$. The optimal convergence rates $\|u - u_h\|_{h,1} = \mathrm{O}(h^{p-1})$ are observed for all values of p.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.4981	1.83	2.22	2.38	96	160	240
0.2828	1.61	1.82	1.87	384	640	960
0.1627	2.06	2.01	2.28	1044	1740	2610
0.0973	2.08	2.50	3.72	3420	5700	8550
0.0508	3.45	7.27	17.95	13920	23200	34800
0.0269	13.15	41.45	120.92	52476	87460	131190
0.0138	81.36	306.06	937.53	205848	343080	514620

Table 7.2: Runtimes and number of DoFs for Experiment 7.9.1, for each mesh size h, and each polynomial degree, p.

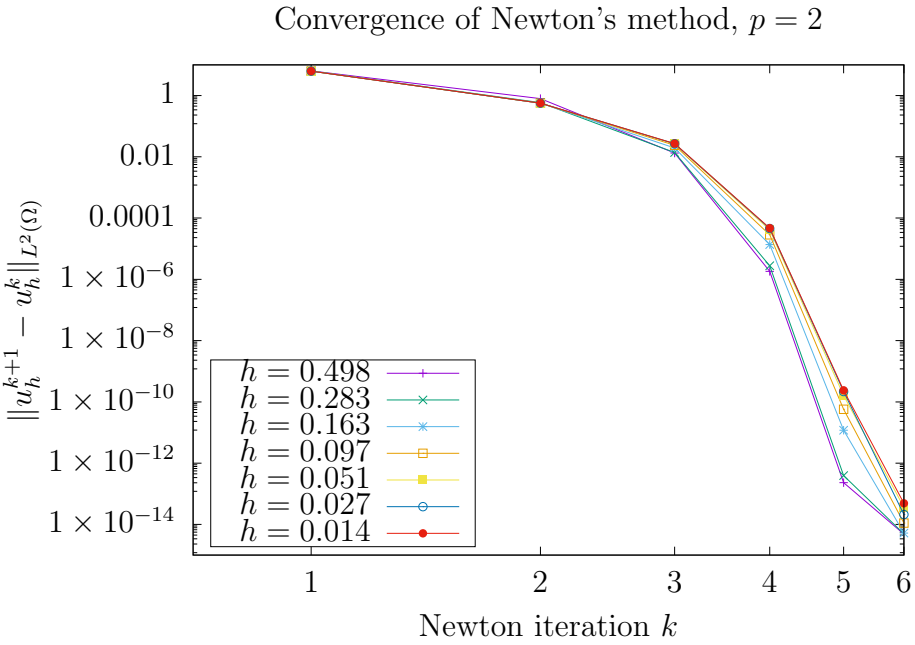


Figure 7.2: Convergence of Newton's method for the numerical scheme applied to problem (7.9.2) with $p = 2$.

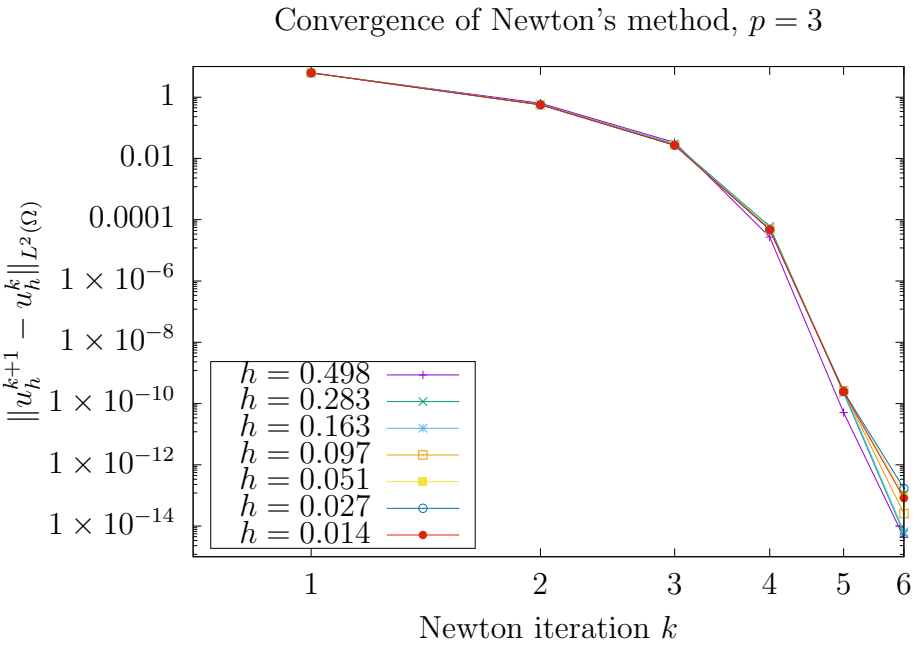


Figure 7.3: Convergence of Newton's method for the numerical scheme applied to problem (7.9.2) with $p = 3$.

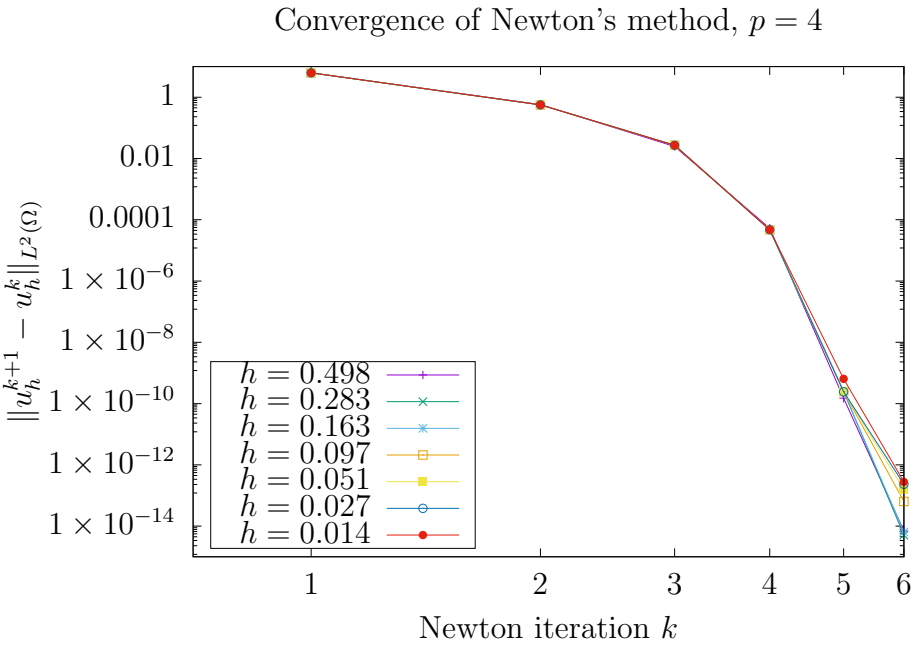


Figure 7.4: Convergence of Newton's method for the numerical scheme applied to problem (7.9.2) with $p = 4$.

7.9.2 Experiment 2

In this experiment, we consider the HJB Dirichlet boundary-value problem,

$$\begin{cases} \sup_{\alpha \in \Lambda} \{A^\alpha : D^2 u - f\} = 0 & \text{a.e. in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{7.9.3}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and the function f in (7.9.3) is chosen so that the true solution is given by

$$u(x, y) = \frac{1}{4} \sin(\pi(x^2 + y^2)).$$

For this problem, the set of controls, $\Lambda := [0, \pi/3] \times \mathrm{SO}(2)$ (note that $\mathrm{SO}(2)$ denotes the set of 2×2 rotation matrices), and the coefficient matrices $\{A^\alpha\}_{\alpha \in \Lambda}$ are defined by

$$A^\alpha := \sigma^\alpha (\sigma^\alpha)^T / 2, \quad \sigma^\alpha := (\sigma_1^\alpha \sigma_2^\alpha) := R^T \begin{bmatrix} 1 & \sin\theta \\ 0 & \cos\theta \end{bmatrix}, \quad \alpha = (\theta, R) \in \Lambda. \tag{7.9.4}$$

Furthermore, since Ω is the unit disk, $\partial\Omega = \mathbb{S}^1$, and it follows that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore, $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. For the internal faces,

the mean curvature is calculated directly as $\mathcal{H}_F = \nabla_{\mathbf{T}} \cdot n_F$, where n_F is a fixed choice of unit normal to F. One can calculate that $\mathrm{Tr}(A^\alpha) = 1$ for all $\alpha \in \Lambda$, and $\det A^\alpha = \frac{1}{4}\cos^2(\theta) \geq 1/16$ for all $\theta \in [0, \pi/3]$, and so the control set Λ satisfies the Cordes condition. Furthermore, as $\mathrm{Tr}\, A^\alpha = 1$ for all $\alpha \in \Lambda$, as in the case of the MA problem, so we may take $\gamma^\alpha := 1$.

As mentioned in Chapter 1, the HJB problem (7.9.3) corresponds to a Markov optimal control problem, where the state equation is given by (1.1.11), with $b \equiv 0$.

In order to approximate the solution of (7.9.3), we apply the semismooth Newton's method, given by Algorithm 2, to (7.9.3). Since each A^α, $\alpha \in \Lambda$ has unit trace, we use the renormalisation parameter $\gamma^\alpha = 1$ (as opposed to $\gamma^\alpha := \mathrm{Tr}(A^\alpha)/|A^\alpha|^2$, see Remark 7.7.3). In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u - u_h\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 7.5, and report the exact values in Table 7.3, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 7.4.

We also plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 7.6, 7.7, and 7.8, respectively. For all polynomial degrees, at each mesh refinements we see an increase in the number of Newton iterations required to reach the desired tolerance. In particular, in each case we observe that reaching the desired tolerance of 10^{-12} requires roughly one extra Newton iteration per mesh refinement.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	4.72		2.71		1.92	
0.2828	2.52	(1.11)	1.59	(0.94)	3.37×10^{-1}	(3.07)
0.1627	2.51	(0.01)	5.48×10^{-1}	(1.93)	4.77×10^{-2}	(3.54)
0.0973	1.48	(1.03)	1.71×10^{-1}	(2.27)	8.34×10^{-3}	(3.39)
0.0508	7.20×10^{-1}	(1.11)	4.59×10^{-2}	(2.02)	9.79×10^{-4}	(3.30)
0.0269	3.66×10^{-1}	(1.06)	1.19×10^{-2}	(2.13)	1.35×10^{-4}	(3.11)
0.0138	1.77×10^{-1}	(1.09)	2.94×10^{-3}	(2.08)	1.81×10^{-5}	(3.00)

Table 7.3: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 7.9.2.

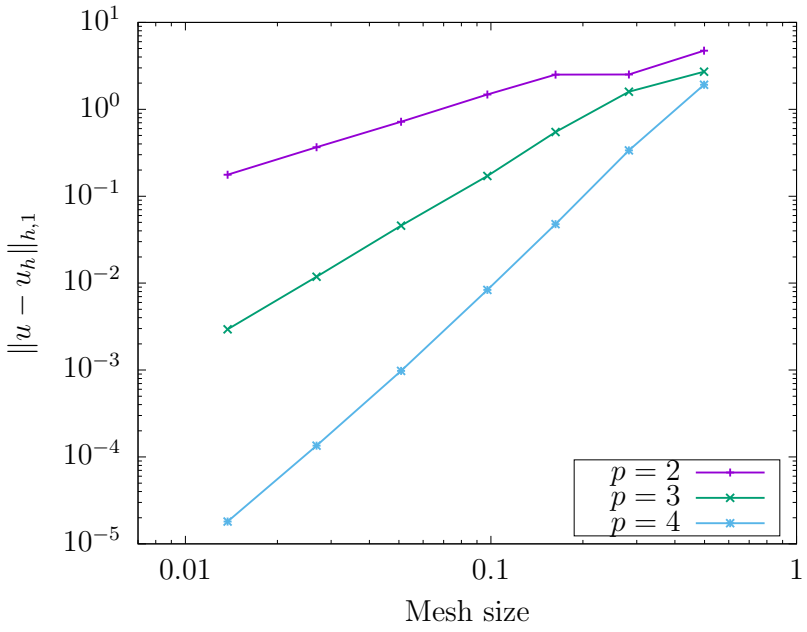


Figure 7.5: Convergence rates for the numerical scheme applied to problem (7.9.3). The error $\|u-u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p = 2$ to $p = 4$. The optimal convergence rates $\|u - u_h\|_{h,1} = \mathrm{O}(h^{p-1})$ are observed for all values of p.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.4981	7.91	8.97	9.79	96	160	240
0.2828	1.83	2.20	2.56	384	640	960
0.1627	2.15	2.57	3.32	1044	1740	2610
0.0973	2.77	3.80	6.55	3420	5700	8550
0.0508	5.77	13.72	37.33	13920	23200	34800
0.0269	25.60	94.24	259.84	52476	87460	131190
0.0138	197.05	669.21	2075.13	205848	343080	514620

Table 7.4: Runtimes and number of DoFs for Experiment 7.9.2, for each mesh size h, and each polynomial degree, p.

7.9.3 Experiment 3

In this experiment, we consider the following MA Neumann boundary-value problem

$$\begin{cases} \det D^2 u(x) = f(x) \quad x \in \Omega, \\ \dfrac{\partial u}{\partial n_{\partial\Omega}} \text{ is constant on } \partial\Omega, \end{cases} \tag{7.9.5}$$

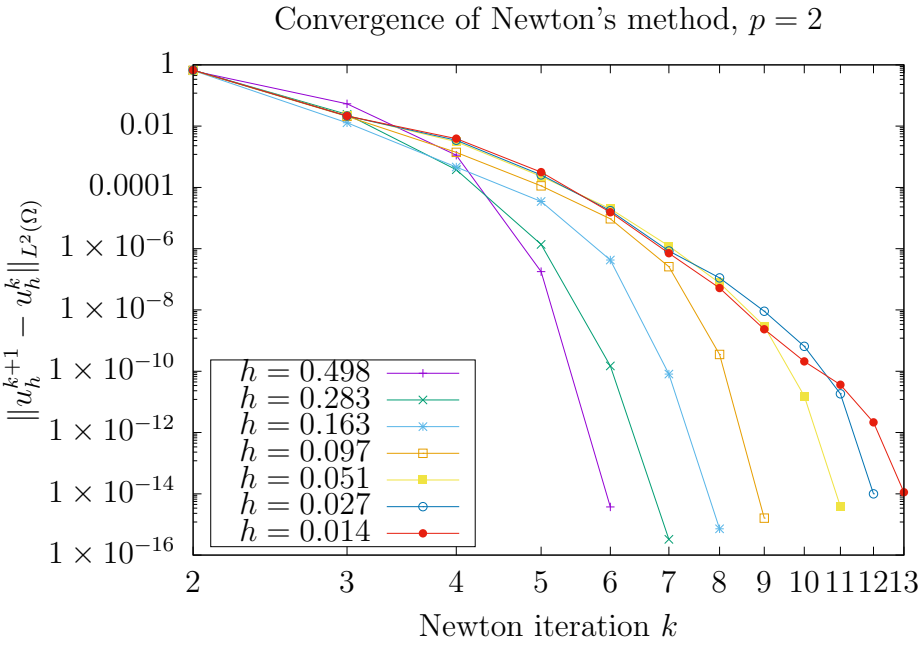


Figure 7.6: Convergence of Newton's method for the numerical scheme applied to problem (7.9.3) with $p = 2$.

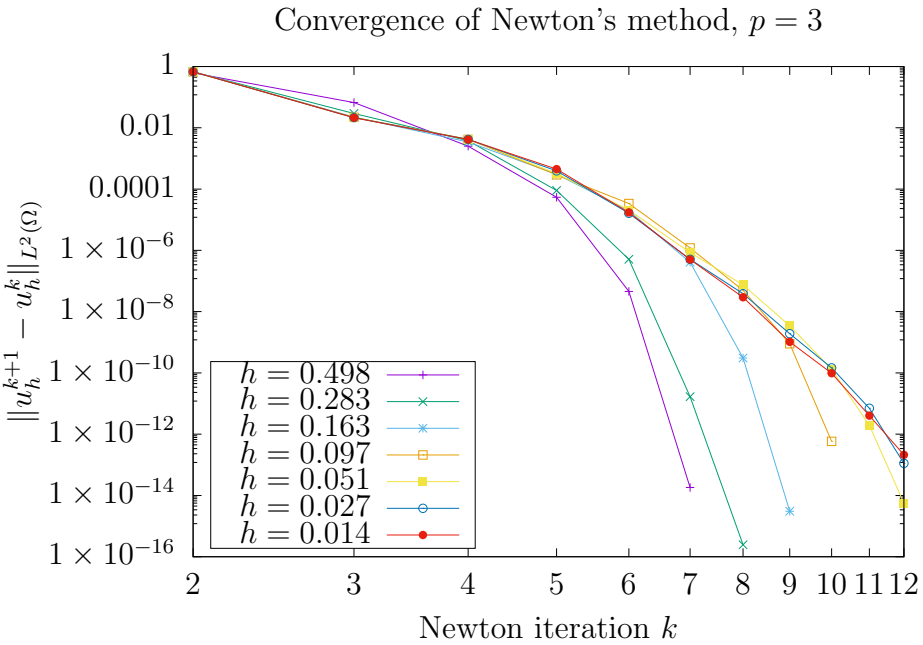


Figure 7.7: Convergence of Newton's method for the numerical scheme applied to problem (7.9.3) with $p = 3$.

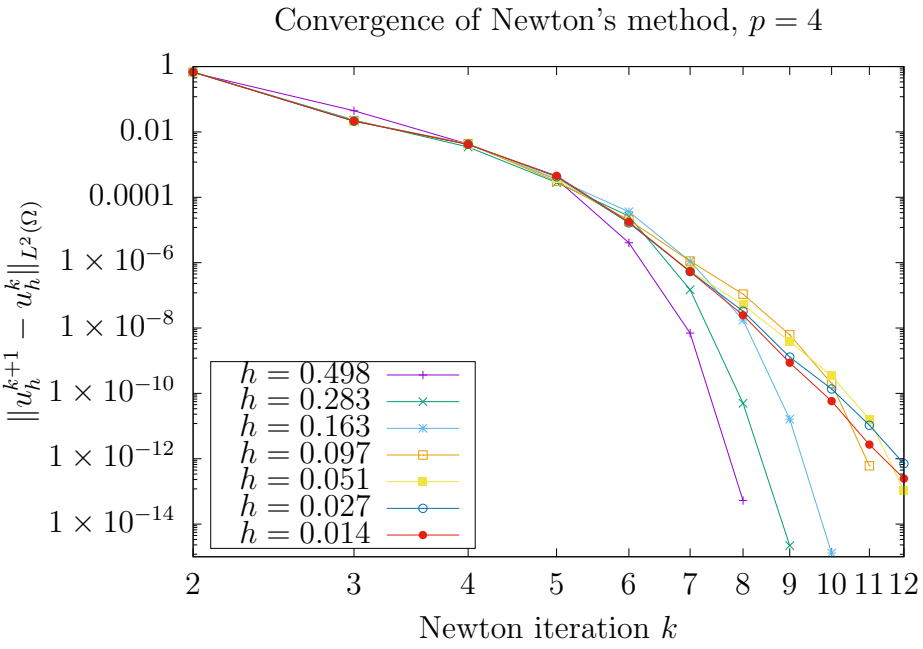


Figure 7.8: Convergence of Newton's method for the numerical scheme applied to problem (7.9.3) with $p = 4$.

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and f is chosen so that the true solution of (7.9.1) is given by

$$u(x, y) = 5(x_1^2 + x_2^2 - 1) + \frac{1}{8}\sin(\pi(x_1^2 + x_2^2)).$$

In order to approximate the solution of (7.9.5), we consider the equivalent HJB problem:

$$\begin{cases} \sup\limits_{W \in X_\xi} \{-W : D^2 u + 2(\det W)^{1/2} f^{1/2}\} = 0, \text{ in } \Omega, \\ \qquad\qquad \dfrac{\partial u}{\partial n_{\partial\Omega}} \text{ is constant on } \partial\Omega, \end{cases} \tag{7.9.6}$$

where $\xi = 1/100$, and apply the semismooth Newton's method, given by Algorithm 2, to (7.9.6), in this case with a step increment tolerance of 1×10^{-11}. Also, in this experiment, we set $c_{\text{stab}} = 25$. Since each $W \in X_\xi$ has unit trace, we use the renormalisation parameter $\gamma^\alpha = 1$ (as opposed to $\gamma^\alpha := \text{Tr}(A^\alpha)/|A^\alpha|^2$, see Remark 7.7.3).

Problem (7.9.6) corresponds to a HJB oblique boundary-value problem, where the oblique vector $\beta \equiv n_{\partial\Omega}$ (i.e., a Neumann BVP). Moreover, one can see that the oriented angle, Θ, between β and $n_{\partial\Omega}$, satisfies $\Theta = 0$ on $\partial\Omega$, and thus $\partial_{\mathbf{T}_2}\Theta = 0$ on $\partial\Omega$. Furthermore, since Ω is the unit disk, it follows that $\partial\Omega = \mathbb{S}^1$, and that the

mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$. It is also then clear that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = 1 > 0 \quad \text{on } \partial\Omega.$$

Since the solution in known, one can directly calculate that

$$\nabla u|_{\partial\Omega} = (10 + \frac{\pi}{4}\cos(\pi(x_1^2 + x_2^2)))x,$$

and so

$$\beta \cdot \nabla u|_{\partial\Omega} = n_{\partial\Omega} \cdot ((10 + \frac{\pi}{4}\cos(\pi(x_1^2 + x_2^2)))x) = 10 - \frac{\pi}{4},$$

(since $n_{\partial\Omega} = x/|x|$, and $|x| = 1$ on $\partial\Omega$), and thus, the compatibility constant $c = 10 - \frac{\pi}{4}$.

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p,0}^{\text{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|(u - u_h, c - c_h)\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 7.1, and report the exact values in Table 7.5, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 7.6.

We also plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 7.10, 7.11, and 7.12, respectively. Across all polynomial degrees and mesh refinements we see a small variation in the number of Newton iterations required to reach the desired tolerance. Except in the case that $p = 4$, at the finest mesh level reaching the desired tolerance of 10^{-11} required 12 iterations, and in the case that $p = 3$, at the penultimate mesh refinement, 30 iterations were required.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	6.34		1.51		8.08×10^{-1}	
0.2828	1.96	(2.08)	7.21×10^{-1}	(1.30)	1.64×10^{-1}	(2.82)
0.1627	1.04	(1.15)	2.71×10^{-1}	(1.77)	2.42×10^{-2}	(3.46)
0.0973	6.21×10^{-1}	(1.00)	8.93×10^{-2}	(2.16)	4.61×10^{-3}	(3.23)
0.0508	3.33×10^{-1}	(0.96)	2.33×10^{-2}	(2.07)	5.67×10^{-4}	(3.23)
0.0269	1.77×10^{-1}	(0.99)	6.08×10^{-3}	(2.11)	8.52×10^{-5}	(2.98)
0.0138	8.96×10^{-2}	(1.02)	1.53×10^{-3}	(2.06)	1.28×10^{-5}	(2.83)

Table 7.5: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 7.9.3.

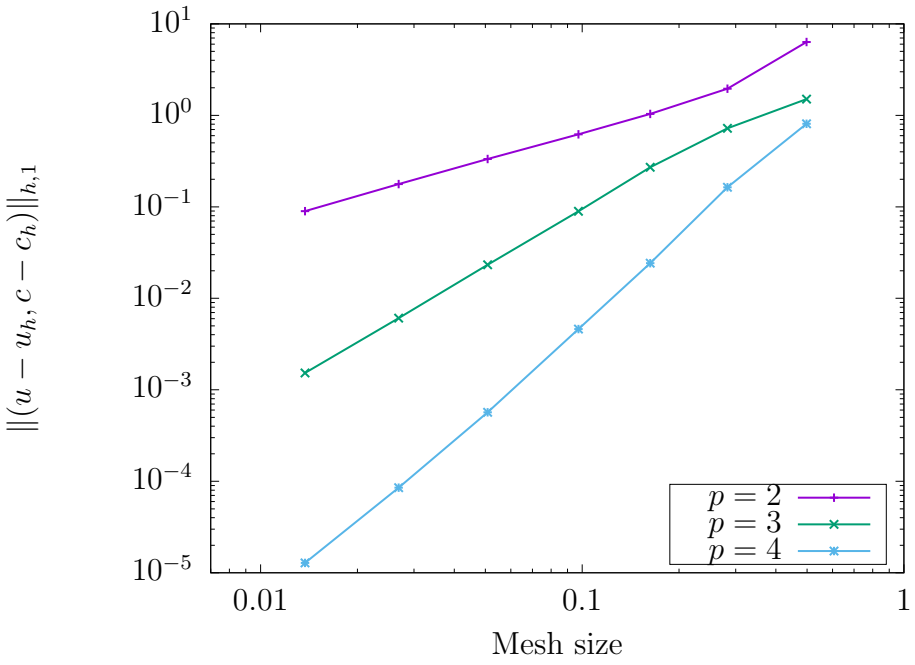


Figure 7.9: Convergence rates for the numerical scheme applied to problem (7.9.6). The error $\|u-u_h\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p=2$ to $p=4$. The optimal convergence rates $\|u-u_h\|_{h,1} = \mathrm{O}(h^{p-1})$ are observed for all values of p.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.4981	29.60	43.02	52.38	112	176	256
0.2828	2.60	3.82	4.23	448	704	1024
0.1627	2.83	3.89	4.84	1218	1914	2784
0.0973	3.93	5.99	10.77	3990	6270	9120
0.0508	10.11	17.55	33.93	16240	25520	37120
0.0269	43.84	534.98	195.06	61222	96206	139936
0.0138	200.66	734.90	3247.93	240156	377388	548928

Table 7.6: Runtimes and number of DoFs for Experiment 7.9.3, for each mesh size h, and each polynomial degree, p.

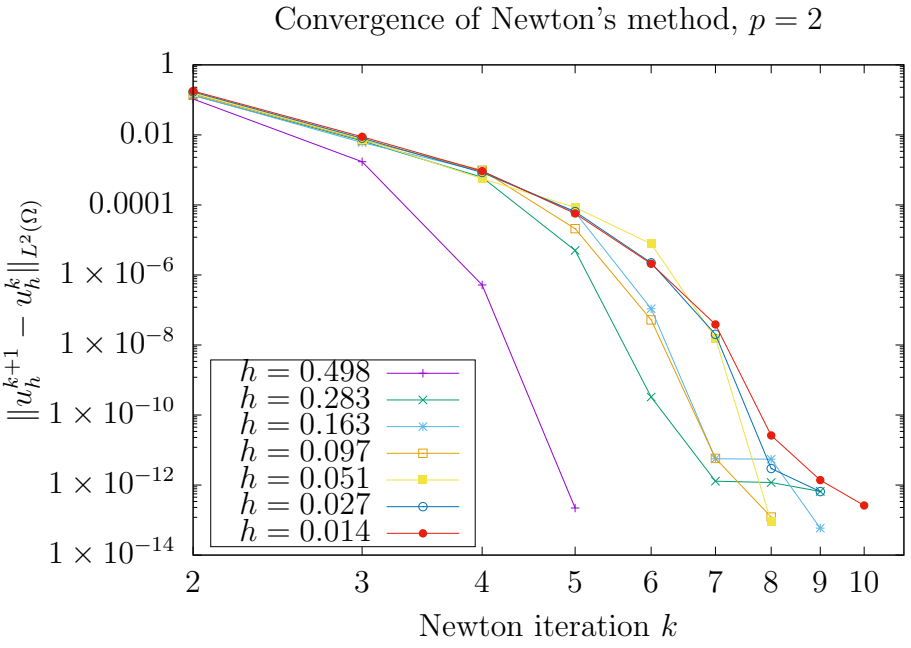


Figure 7.10: Convergence of Newton's method for the numerical scheme applied to problem (7.9.6) with $p = 2$.

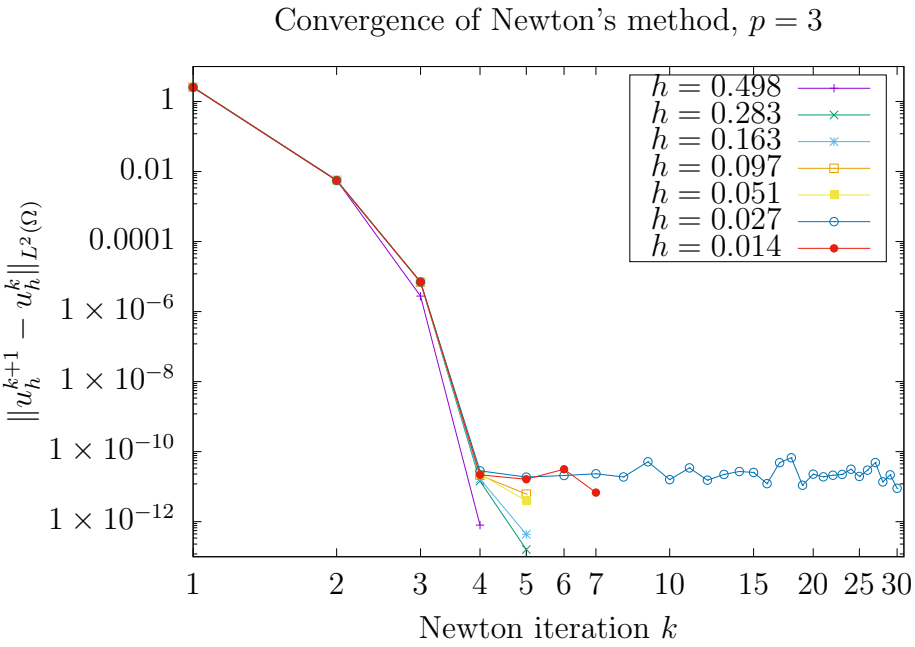


Figure 7.11: Convergence of Newton's method for the numerical scheme applied to problem (7.9.6) with $p = 3$.

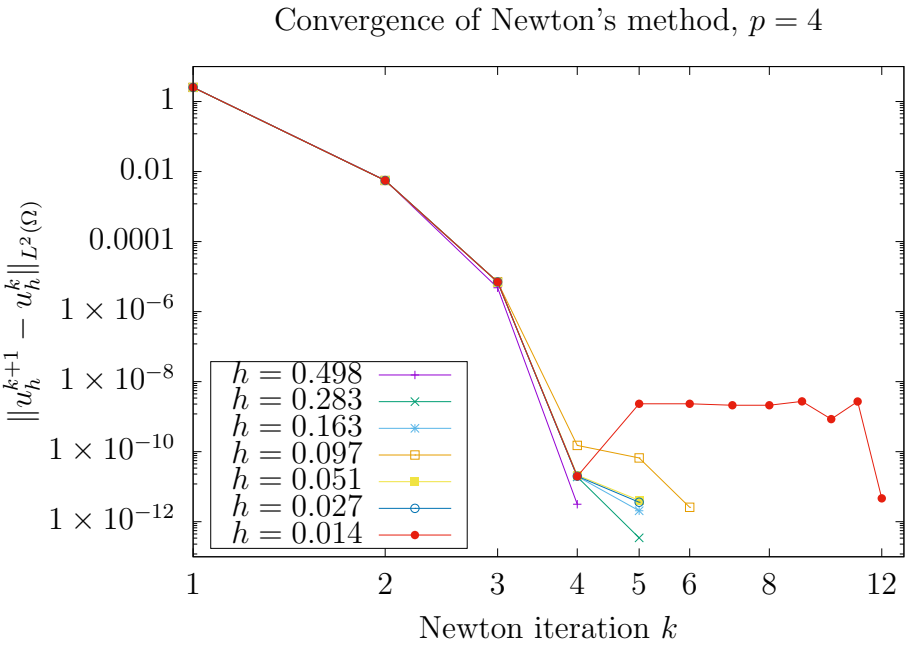


Figure 7.12: Convergence of Newton's method for the numerical scheme applied to problem (7.9.6) with $p = 4$.

7.9.4 Experiment 4

In this experiment, we consider the HJB oblique boundary-value problem,

$$\begin{cases} \sup\limits_{\alpha\in\Lambda}\{A^\alpha : D^2u - f\} = 0 \quad \text{a.e. in } \Omega, \\ \qquad\qquad \beta \cdot \nabla u \text{ is constant on } \partial\Omega, \end{cases} \tag{7.9.7}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, the anti-clockwise rotation of the normal by the angle $\varphi(x_1, x_2) := \pi/4 + \arctan(\frac{x_2}{x_1})$, for $(x_1, x_2) \in \partial\Omega$, that is,

$$\beta(x_1, x_2) = \begin{bmatrix} \cos\varphi(x_1, x_2) & -\sin\varphi(x_1, x_2) \\ \sin\varphi(x_1, x_2) & \cos\varphi(x_1, x_2) \end{bmatrix} \begin{bmatrix} [n_{\partial\Omega}]^1 \\ [n_{\partial\Omega}]^2 \end{bmatrix} \quad (x_1, x_2) \in \partial\Omega.$$

The function f in (7.9.7) is chosen so that the true solution is given by

$$u(x_1, x_2) = \frac{1}{4}\cos(\pi(x_1^2 + x_2^2)) - \frac{1}{|\Omega|}\int_\Omega \frac{1}{4}\cos(\pi(x_1^2 + x_2^2)).$$

We can also see that the oriented angle, Θ, between β and $n_{\partial\Omega}$ is given by $\Theta(x_1, x_2) = \pi/4 + \varphi(x_1, x_2)$ for $(x_1, x_2) \in \partial\Omega$. It then follows that on $\partial\Omega$,

$$\begin{aligned} \partial_{\mathbf{T}_2}\Theta &= \nabla(\varphi(x_1, x_2)) \cdot (-[n_{\partial\Omega}]^2, [n_{\partial\Omega}]^1) \\ &= \nabla(\arctan(x_2/x_1)) \cdot (-x_2, x_1) \\ &= \frac{1}{x_1^2 + x_2^2}(-x_2, x_1) \cdot (-x_2, x_1) = 1. \end{aligned} \tag{7.9.8}$$

Furthermore, since Ω is the unit disk, and $\partial\Omega = \mathbb{S}^1$, the oblique vector β rotates by exactly 2π, as we traverse $\partial\Omega$ in a fixed direction, and so $\beta \in C^1(\partial\Omega; \mathbb{S}^1)$. Since the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, by (7.9.8) we also see that

$$\partial_{\mathbf{T}_2}\Theta + \mathcal{H}_{\partial\Omega} = 2 > 0.$$

Since the solution in known, one can directly calculate that

$$\nabla u|_{\partial\Omega} = -\frac{\pi}{2}\sin(\pi(x_1^2 + x_2^2))x = 0,$$

since $|x| = 1$ on $\partial\Omega$ and so $\beta \cdot \nabla u|_{\partial\Omega} = 0$, and thus, the compatibility constant $c = 0$.

Furthermore, since Ω is the unit disk, $\partial\Omega = \mathbb{S}^1$, and it follows that the mean curvature of $\partial\Omega$, $\mathcal{H}_{\partial\Omega} = 1$, and therefore, $\mathcal{H}_F = 1$ for all $F \in \mathscr{E}_h^b$.

For this problem, the set of controls, $\Lambda := [0, \pi/3] \times \mathrm{SO}(2)$, and the coefficient matrices $\{A^\alpha\}_{\alpha\in\Lambda}$ are defined by

$$A^\alpha := \sigma^\alpha(\sigma^\alpha)^T/2, \quad \sigma^\alpha := (\sigma_1^\alpha \sigma_2^\alpha) := R^T \begin{bmatrix} 1 & \sin\theta \\ 0 & \cos\theta \end{bmatrix}, \quad \alpha = (\theta, R) \in \Lambda. \tag{7.9.9}$$

One can calculate that $\mathrm{Tr}(A^\alpha) = 1$ for all $\alpha \in \Lambda$, and $\det A^\alpha = \frac{1}{4}\cos^2(\theta) \geq 1/16$ for all $\theta \in [0, \pi/3]$, and so the control set Λ satisfies the Cordes condition. Furthermore, as $\mathrm{Tr}\, A^\alpha = 1$ for all $\alpha \in \Lambda$, as in the case of the MA problem, so we may take $\gamma^\alpha := 1$.

As mentioned in Chapter 1, the HJB problem (7.9.7) corresponds to a Markov optimal control problem, where the state equation is given by (1.1.11), with $b \equiv 0$.

In order to approximate the solution of (7.9.7), we apply the semismooth Newton's method, given by Algorithm 2, to (7.9.7). Since each A^α, $\alpha \in \Lambda$ has unit trace, we use the renormalisation parameter $\gamma^\alpha = 1$ (as opposed to $\gamma^\alpha := \mathrm{Tr}(A^\alpha)/|A^\alpha|^2$, see Remark 7.7.3). In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p,0}^{\mathrm{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence in the $\|\cdot\|_{h,1}$-norm are optimal, that is $\|u - u_h\|_{h,1} = \mathcal{O}(h^{p-1})$. We plot the error values in the $\|\cdot\|_{h,1}$-norm in Figure 7.13, and report the exact values in Table 7.7, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 7.8.

We also plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 7.14, 7.15, and 7.16, respectively. For all polynomial degrees, at each mesh refinements we see an increase in the number of Newton iterations required to reach

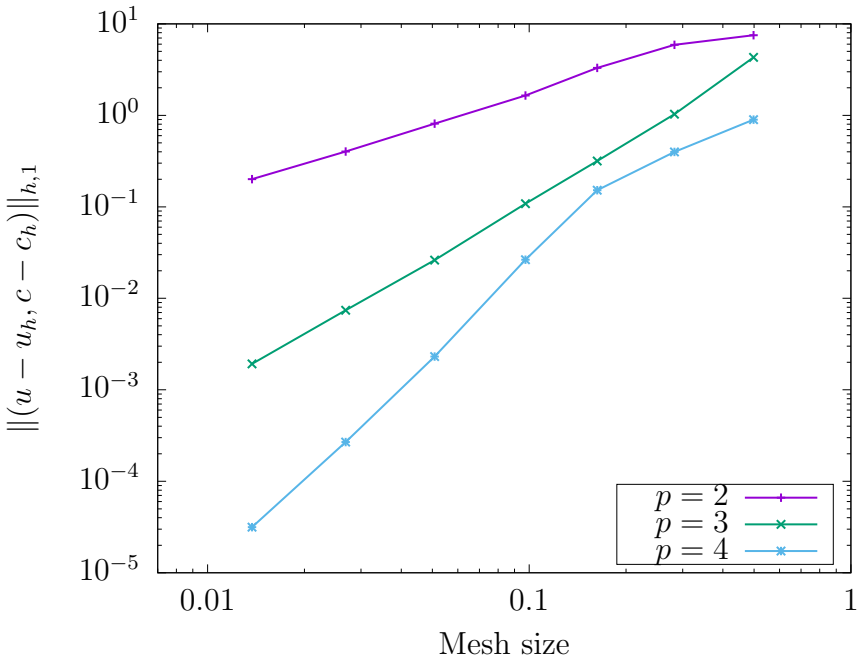


Figure 7.13: Convergence rates for the numerical scheme applied to problem (7.9.7). The error $\|(u-u_h, c-c_h)\|_{h,1}$ is plotted against the mesh size h for polynomial degrees ranging from $p=2$ to $p=4$. The optimal convergence rates $\|(u-u_h, c-c_h)\|_{h,1} = \mathrm{O}(h^{p-1})$ are observed for all values of p.

the desired tolerance. In particular, in each case we observe that reaching the desired tolerance of 10^{-12} requires roughly one extra Newton iteration per mesh refinement.

Mesh size	$p=2$		$p=3$		$p=4$	
0.4981	7.51		4.30		8.96×10^{-1}	
0.2828	5.90	(0.43)	1.03	(2.53)	3.99×10^{-1}	(1.43)
0.1627	3.30	(1.05)	3.17×10^{-1}	(2.13)	1.52×10^{-1}	(1.75)
0.0973	1.65	(1.35)	1.08×10^{-1}	(2.09)	2.65×10^{-2}	(3.40)
0.0508	8.11×10^{-1}	(1.09)	2.61×10^{-2}	(2.19)	2.31×10^{-3}	(3.76)
0.0269	4.03×10^{-1}	(1.10)	7.41×10^{-3}	(1.98)	2.69×10^{-4}	(3.38)
0.0138	2.01×10^{-1}	(1.04)	1.92×10^{-3}	(2.02)	3.15×10^{-5}	(3.20)

Table 7.7: Error values in the $\|\cdot\|_{h,1}$-norm and EOCs for Experiment 7.9.4.

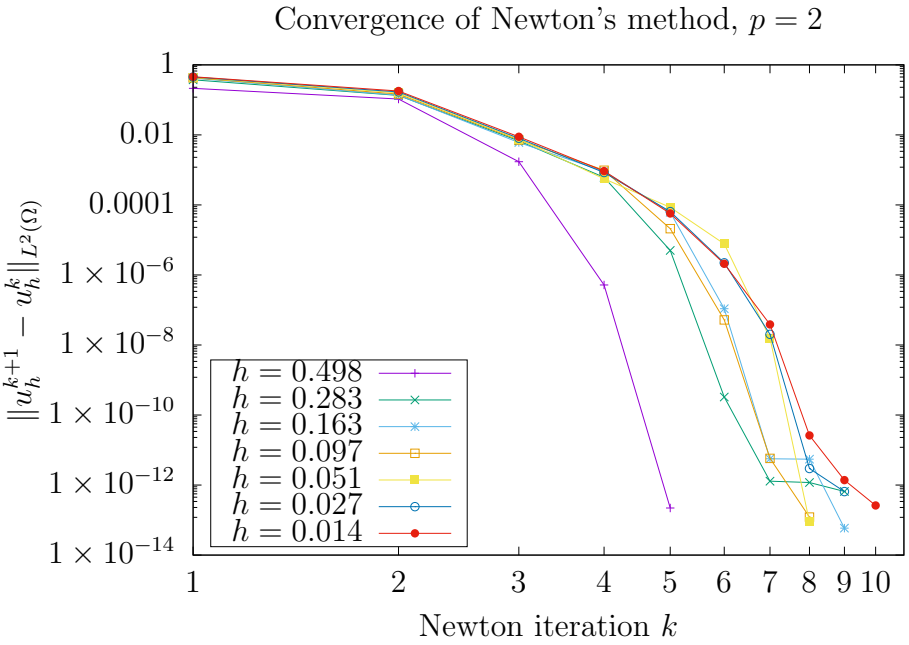


Figure 7.14: Convergence of Newton's method for the numerical scheme applied to problem (7.9.7) with $p = 2$.

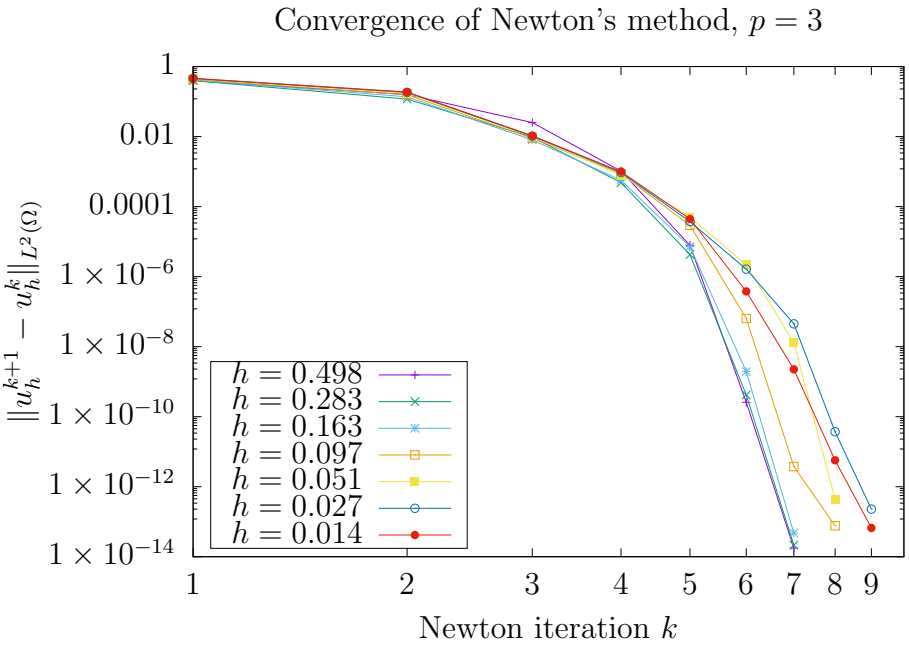


Figure 7.15: Convergence of Newton's method for the numerical scheme applied to problem (7.9.7) with $p = 3$.

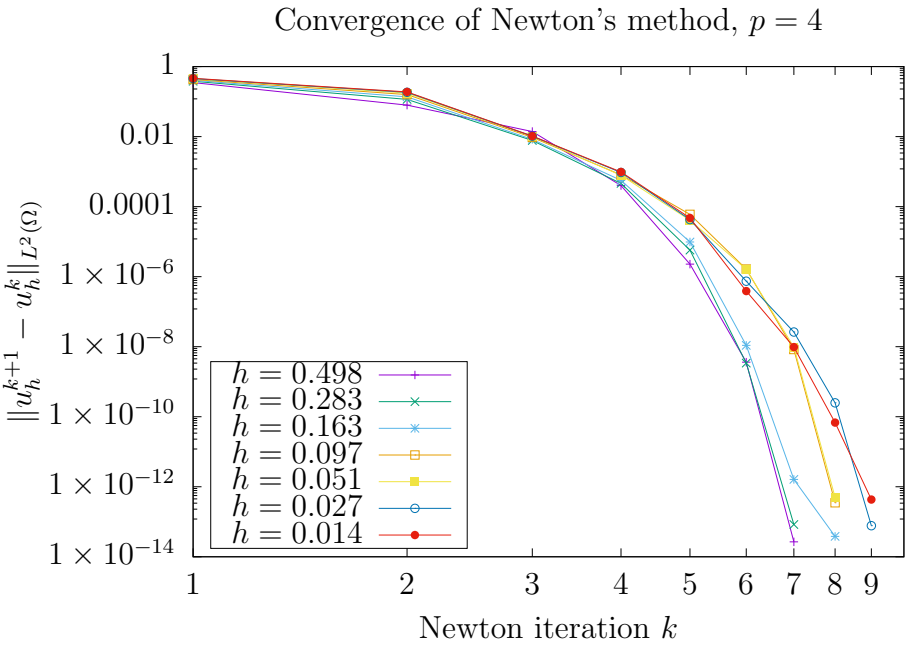


Figure 7.16: Convergence of Newton's method for the numerical scheme applied to problem (7.9.7) with $p = 4$.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.4981	31.25	50.29	53.89	112	176	256
0.2828	4.77	4.68	4.43	448	704	1024
0.1627	5.23	5.16	6.78	1218	1914	2784
0.0973	6.15	9.27	12.46	3990	6270	9120
0.0508	14.55	27.94	52.74	16240	25520	37120
0.0269	59.82	154.72	342.06	61222	96206	139936
0.0138	352.95	911.25	2325.04	240156	377388	548928

Table 7.8: Runtimes and number of DoFs for Experiment 7.9.4, for each mesh size h, and each polynomial degree, p.

7.9.5 Experiment 5 - Robustness of Newton's method

In this experiment, we consider the following 210 MA Dirichlet boundary-value problems

$$\begin{cases} \det D^2 u^i(x) = f^i(x) & x \in \Omega, \\ u(x) = 0 & x \in \partial\Omega, \end{cases} \tag{7.9.10}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and the index $i \in \{1, \ldots, 210\}$. The purpose of this experiment is to demonstrate the robustness of the numerical method with respect to the mesh size, h, the choice of initial guess, $u_h^{i,0}$, the proximity of $u_h^{i,0}$ to the numerical solution u_h^i in the $\|\cdot\|_{h,1}$-norm, and the choice of right-hand side function f^i. In each case both the true solution u^i, and the numerical solution u_h^i of (7.3.2) are unknown, thus instead of considering the value $\|u_h^{i,0} - u_h\|_{h,1}$ (which is unknown), we instead use the (computable) value $\|u_h^{i,0} - u_h^{i,N}\|_{h,1}$, where $u_h^{i,N}$ is the final Newton iterate generated by achieving the desired tolerance. Heuristically, if the Newton's method has converged to a sufficiently small tolerance, then one would expect that $\|u_h^{i,0} - u_h^i\|_{h,1} \approx \|u_h^{i,0} - u_h^{i,N}\|_{h,1}$.

The 210 individual experiments are implemented via the following procedure:

1. We refine the mesh 7 times, and at each refinement, we provide a *randomly chosen* right-hand side function f^i, a *randomly chosen* initial guess $u_h^{i,0}$, and apply the semismooth Newton's method, with the quadratic approximation space $V_{h,2}^{\text{comp}}$, until the step increment L^2-norm is below 10^{-12}.

2. We repeat the previous step 30 times.

The right-hand side functions, f^i, and initial guesses, $u_h^{i,0}$, are *randomly chosen* in the following sense: the function f^i is a polynomial of the form

$$f^i(x_1, x_2) := \sum_{0 \leq \ell + m \leq 3} f_{\ell,m}^i (x_1^2)^\ell (x_2^2)^m, \tag{7.9.11}$$

where each coefficient $f_{\ell,m}^i = a_{\ell,m}^i b_{\ell,m}^i$, where each $a_{\ell,m}^i$ is an integer chosen randomly (with respect to a uniform distribution) from $\{100, \ldots, 999\}$, and $b_{\ell,m}^i$ is a real number chosen randomly (with respect to a uniform distribution) from the interval $[0, 1)$ (note that this particular form of polynomial ensures that the right-hand side function is nonnegative). The initial guess is a polynomial of the form

$$u_h^{i,0}(x_1, x_2) := \sum_{0 \leq \ell + m \leq 3} u_{\ell,m}^i x_1^\ell x_2^m, \tag{7.9.12}$$

where each coefficient $u_{\ell,m}^i = c_{\ell,m}^i d_{\ell,m}^i$, where $c_{\ell,m}^i$ is an integer chosen randomly (with respect to a uniform distribution) from $\{10^4, \ldots, 10^6 - 1\}$, and $d_{\ell,m}^i$ is a real number chosen randomly (with respect to a uniform distribution) from the interval $[-1, 1)$. We generate such coefficients using the "randint" and "uniform" functions of the "NumPy" module [96].

In Figure 7.17 we plot the value $\|u_h^{i,0} - u_h^{i,N}\|_{h,1}$ (where $u_h^{i,N} \in V_{h,2}^{\text{comp}}$ is the Newton iterate that satisfies the requested increment L^2-norm tolerance), against the total number of Newton steps required to reach the desired tolerance (where applicable). We have provided results for all 210 experiments in Tables B.1 to B.6 in Appendix B. In particular, for each $i \in \{1, \ldots, 230\}$, these tables provide the value of $\|u_h^{i,N} - u_h^{i,0}\|_{h,1}$, the mesh size, and the number of Newton steps. Furthermore, Tables B.7 to B.12 provide the values of the initial guess coefficients $\{u_{\ell,m}^i\}_{0 \leq \ell+m \leq 3}$, and Tables B.13 to B.16 provide the values of the right-hand side function coefficients $\{f_{\ell,m}^i\}_{0 \leq \ell+m \leq 3}$.

For 207 of the 210 experiments we observe a small variation in the number of Newton steps required for the step increment L^2-norm error to fall below the tolerance, in particular only 6-10 iterations are required in each case.

The remaining 3 experiments were the cases $i = 56, 118, 154$. For $i = 56$ the total number of steps required was 14, for $i = 118, 154$, after 20 Newton steps the Newton increment L^2-norm error $10^{-12} < \|u_h^{20,i} - u_h^{19,i}\|_{L^2(\Omega)} < 10^{-11}$. That is, the desired tolerance of 10^{-12} was not achieved, but the error did fall below 10^{-11}.

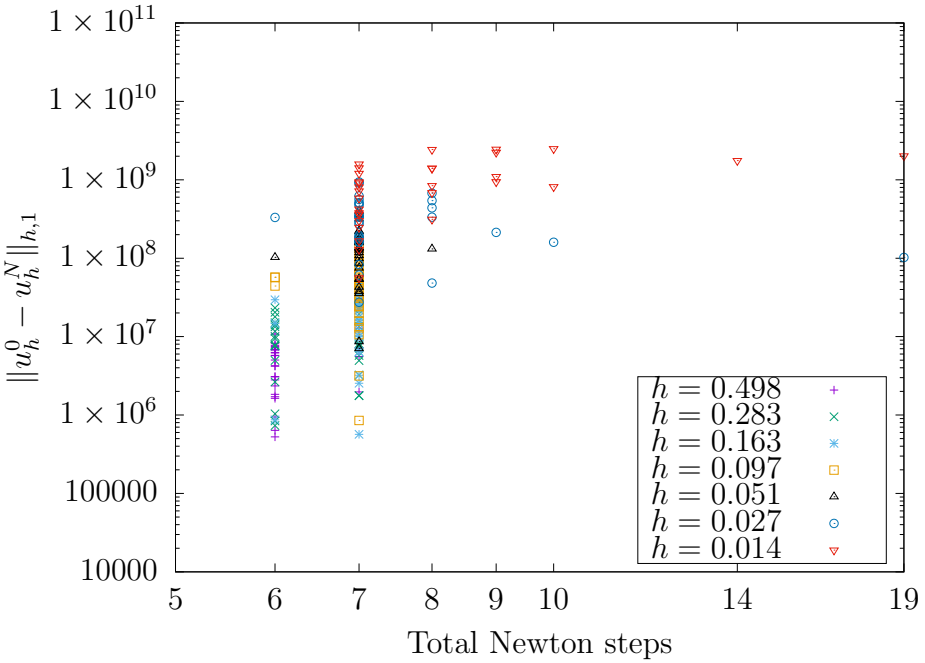


Figure 7.17: Total Newton steps plotted against the value $\|u_h^{i,0} - u_h^{i,N}\|_{h,1}$, where $u_h^{i,N}$ is the approximate solution of (7.9.10) generated by applying the semismooth Newton's method until the increment L^2-norm fell below 10^{-12}.

7.10 Concluding remarks for this method

In this chapter we have introduced a discontinuous Galerkin finite element method for the approximation of strong solutions to nonlinear HJB equations, with Dirichlet and oblique boundary conditions. By reformulating the MA equation as a uniformly elliptic HJB equation (due to Theorem 3.5.3), this has provided us with a DGFEM for the approximation of uniformly convex solutions to the two-dimensional MA equation. Indeed, due to the uniform convexity assumption upon the true solution, we see that the method is always well posed, and we are not required to numerically enforce convexity.

Chapter 8

A CGFEM for the MA Dirichlet problem

8.1 New contributions and existing results

In this chapter, we present a continuous Galerkin FEM for the MA Dirichlet problem (3.2.1) in the case that

$$f(x, u(x), \nabla u(x)) := \frac{f_1(x, u(x))}{f_2(\nabla u(x))}. \tag{8.1.1}$$

That is, the problem of approximating the solution, u, of the following MA Dirichlet problem

$$\begin{cases} \det D^2 u(x) = \dfrac{f_1(x, u(x))}{f_2(\nabla u(x))}, & x \in \Omega, \\ u(x) = \phi(x), & x \in \partial\Omega. \end{cases} \tag{8.1.2}$$

Existing results: Some of the contributions of this chapter build upon existing results.

- In [79] the authors proposed the nonvariational finite element method (NVFEM) for the numerical approximation of solutions to the MA Dirichlet problem (8.1.2) in the case that $f := f(x)$ (i.e., f does not depend on u or ∇u).
- In [92], the author proves the existence of a solution to the method proposed in [79] that satisfies an a priori error estimate in a H^1-type norm, that is optimal with respect to the mesh size. However, this proof appears to be in the context of finite elements on polygonal domains.

The original contributions of this chapter are listed as follows:

1. We extend the method proposed in [79] to the case that f takes the form (8.1.1) (i.e., f *is* allowed to depend on u and ∇u).

2. We extend the proof present in [92] to the case that f takes the form (8.1.1), as well as the context of *curved* finite elements (i.e., we also extend the proof from the polygonal case). In doing so we also prove optimal a priori error estimates in both H^1 and (broken) H^2-type norms, that are optimal with respect to the mesh size.

3. We prove convergence of Newton's method for this problem under the assumption that the initial guess lies within a radius of size $h^{2+\alpha}$, $\alpha > 0$, of the numerical solution (measured in a H^1-type norm).

4. We extend this method to more general nonlinear elliptic equations with a structure similar to those of MA type; a particular example being the Weingarten equation, which is a nonlinear geometric PDE (see Section 8.5 for further details).

The reader must note that both the proof present in [92] and the proof we provide yields existence of a numerical solution that is unique in a ball of radius $h^{2+\alpha}$, $\alpha > 0$, centred at a particular projection of the true solution (see Section 8.3.4 for further details on this projection). As discussed in Section 3.5, numerical methods for MA type problems may exhibit non-uniqueness on a larger scale than that of the underlying PDE (recall the finite difference method given by (3.5.2), for which it is conjectured in [93] that an application of Newton's method to the nonlinear algebraic system can lead to $2^{(N-2)^2}$ solutions on an $N \times N$ grid, by varying the initial guess). However, the experiments of Section 8.7 indicate that the method we propose does not exhibit such large scale nonuniquess, provided that the initial guess for the Newton's method is carefully chosen.

8.2 The numerical method

Let us recall the notational convention that we use for the arguments of the functions f_1, f_2, in order to be precise when referring to the derivatives of f_1 and f_2. In particular we follow the notational convention: $f_1 = f_1(x, z)$, and $f_2 = f_2(q)$.

We require the following assumptions, in order to prove existence of a numerical solution.

Assumption 8.2.1 *Let $d = 2$, and assume that $p \geq 3$. Furthermore, assume that $f_1 \in C^{p,\alpha}(\Omega \times \mathbb{R}; \mathbb{R}^+)$ satisfies $\frac{\partial f_1}{\partial z} \geq 0$, and $f_2 \in C^{p,\alpha}(\mathbb{R}^d; \mathbb{R}^+)$ is uniformly positive.*

We also assume that the boundary datum ϕ, and $\partial\Omega$ satisfy the assumptions of Theorem 3.2.1, with $k = p$. Furthermore, we assume that $\{\mathscr{T}_h\}_h$ is family of triangulations on $\overline{\Omega}$ that is regular of order $p+1$ and quasi-uniform.

Note that under the above assumptions, Theorem 3.2.1 yields existence of a uniformly convex function $u \in C^{p+2,\alpha}(\overline{\Omega})$ that satisfies (8.1.2). Note that the value of p corresponds to the polynomial degree of the finite element space $\mathbb{V}_{h,p}$, which we recall is defined by

$$\mathbb{V}_{h,p} := \{v \in C^0(\overline{\Omega}) : \ v|_K = \hat{\rho} \circ F_K^{-1},\ \hat{\rho} \in \mathbb{P}^p(\hat{K}),\ \forall K \in \mathscr{T}_h\},$$

as well as the zero trace space

$$\mathring{\mathbb{V}}_{h,p} := \mathbb{V}_{h,p} \cap H_0^1(\Omega).$$

The numerical method for the approximation of solutions to the MAD problem (8.1.2) is given as follows: find $u_h \in \mathbb{V}_{h,p}$ with $u_h|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$ satisfying,

$$\langle F_h^{\mathrm{MA}}[u_h], v\rangle = \int_\Omega (f_1(x,u_h) - f_2(\nabla u_h)\det \mathbf{H}_h u_h)v = 0 \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{8.2.1}$$

Remark 8.2.2 *The method (8.2.1) is proposed in d-dimensions, however, the reader must note that our proof of existence of a numerical solution holds* only *in the two-dimensional case.*

Let us define the following norms:

$$\begin{aligned} \|v\|_h^2 &:= |v|_{H^1(\Omega)}^2 + \sum_{F\in\mathscr{E}_h} h_F \|\langle\!\langle \nabla v \rangle\!\rangle\|_{2,F}^2, \\ \|r\|_{-1,h} &:= \sup_{v\in\mathring{\mathbb{V}}_{h,p}\setminus\{0\}} \frac{\langle r, v\rangle}{\|v\|_h}. \end{aligned} \tag{8.2.2}$$

The aim of this chapter will be to prove the following theorem:

Theorem 8.2.3 *Under the hypotheses of Assumption 8.2.1, there exists a uniformly convex function $u \in C^{p+2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, that satisfies (8.1.2), and a constant $h_0 > 0$ such that for $h \le h_0$ there exists a $u_h \in \mathbb{V}_{h,p}$ that satisfies (8.2.1). Moreover, the functions u, u_h satisfy the following error estimate:*

$$\|u - u_h\|_h \le Ch^p, \tag{8.2.3}$$

where C is a constant independent of h and u_h. Furthermore, the function u_h is unique in the ball $\mathbb{B}_{h^{2+\alpha}}(u_*)$ defined by (8.3.16), for some $\alpha > 0$.

Remark 8.2.4 *Theorem 8.2.3 was proven in [92] in the case when $f(x,z,q) := f(x)$. As such, Theorem 8.2.3 extends this result, allowing for more general nonlinearities.*

8.3 Analysis of the numerical method

In order to analyse the method given by (8.2.1), we must define the following spaces:

$$\begin{aligned} V &= W^{2,\infty}(\Omega;\mathscr{T}_h) \cap W^{1,\infty}(\Omega), \\ \mathring{V} &= V \cap W_0^{1,\infty}(\Omega). \end{aligned} \tag{8.3.1}$$

Since the finite element spaces $\mathbb{V}_{h,p}$ and $\mathring{\mathbb{V}}_{h,p}$ consist of *continuous* piecewise polynomial functions, we have that $\mathbb{V}_{h,p} \subset V$, and $\mathring{\mathbb{V}}_{h,p} \subset \mathring{V}$. One can see that $F_h^{\mathrm{MA}} : \mathbb{V}_{h,p} \to (\mathring{\mathbb{V}}_{h,p})'$ is the restriction of $F^{\mathrm{MA}} : V \to \mathring{V}'$ defined by

$$\langle F^{\mathrm{MA}}[w], v\rangle := \int_\Omega (f_1(x,w) - f_2(\nabla w)\det \mathbf{H}_h w)v \quad \forall w \in V,\ \forall v \in \mathring{V}, \tag{8.3.2}$$

to the finite element space.

8.3.1 Taylor expansion of the finite element operator

We have defined $F^{\mathrm{MA}} : V \to \mathring{V}'$ in (8.3.2). First, we note that

$$\det(A+B) = \det(A) + \mathrm{Cof}(A) : B + \det(B), \quad \forall A, B \in \mathbb{R}^{2\times 2}.$$

We can now apply Theorem 2.2.9, in conjunction with Definition 2.2.10, Taylor expanding the integrand in (8.3.2), about the function u, where $u \in C^{p+2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, satisfies (8.1.2), yielding the following for $w \in V$, $v \in \mathring{V}$:

$$\begin{aligned} \langle F^{\mathrm{MA}}[u+w], v\rangle &= \int_\Omega (f_1(x,u) - \det(\mathbf{H}_h u) f_2(\nabla u))v \\ &+ \int_\Omega (D_z f_1(x,u)w - \mathrm{Cof}(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)) : \mathbf{H}_h w) f_2(\nabla u) - \det(\mathbf{H}_h u) D_q f_2(\nabla u)\cdot\nabla w)v \\ &+ \int_\Omega (R^{f_1,u}(u+w) - \det(\mathbf{H}_h w)(f_2(\nabla u) + D_p f_2(\nabla u)\cdot\nabla w))v \\ &- \int_\Omega (R^{f_2,\nabla u}(\nabla u + \nabla w)(\det \mathbf{H}_h w + \mathrm{Cof}(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)) : \mathbf{H}_h w))v \\ &= \langle F^{\mathrm{MA}}[u], v\rangle + \langle L_u[w], v\rangle + \langle R[w], v\rangle, \end{aligned} \tag{8.3.3}$$

where $L_u,\ R : V \to \mathring{V}'$ are defined by

$$\begin{aligned} \langle L_u[w], v\rangle := \int_\Omega &(D_z f_1(x,u)w - \mathrm{Cof}(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)) : \mathbf{H}_h w) f_2(\nabla u) v \\ &- \int_\Omega \det(\mathbf{H}_h u) D_q f_2(\nabla u)\cdot\nabla w\, v, \end{aligned} \tag{8.3.4}$$

and

$$\langle R[w], v\rangle := \int_\Omega (R^{f_1,u}(u+w) - \det(\mathbf{H}_h w)(f_2(\nabla u) + D_p f_2(\nabla u)\cdot\nabla w))v \\ - \int_\Omega (R^{f_2,\nabla u}(\nabla u + \nabla w)(\det \mathbf{H}_h w + \mathrm{Cof}(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)) : \mathbf{H}_h w))v, \tag{8.3.5}$$

respectively.

We remark that the expression $\mathrm{Cof}\,\mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)$ appears in (8.3.3), (8.3.4), and (8.3.5), due to the fact that the definition (8.3.2) of F^{MA} involves the finite element Hessian, $\mathbf{H}_h$. Furthermore, as $u \in C^{p+2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, it follows from Corollary 4.11.5, that $\mathbf{H}_h(D^2u) = \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)$.

Note that $R^{f_1,u}$ and $R^{f_2,\nabla u}$ are the quadratic remainder terms of f_1 and f_2 associated with (x,u) and ∇u respectively, given by Definition 2.2.10.

We then define $L_{u,h},\ R_h : \mathbb{V}_{h,p} \to \mathring{\mathbb{V}}'_{h,p}$ to be the restrictions of L_u and R to $\mathbb{V}_{h,p}$, respectively.

8.3.2 Main theorem proof outline

Now we have the operators $F^{\mathrm{MA}},\ L_u,\ R$, and $F_h^{\mathrm{MA}},\ L_{u,h},\ R_h$, we can give a description of how we will prove Theorem 8.2.3.

1. We define a ball, $\mathbb{B}_\rho(u_*)$ of radius ρ, centred at a particular projection, u_* (for which we will provide more details later on) of the analytical solution, u, onto the finite element space $\mathbb{V}_{h,p}$.

2. We define the operator $M : V \to \mathbb{V}_{h,p}$ by

$$M = L_{u,h}^{-1}(L_u - F^{\mathrm{MA}}), \tag{8.3.6}$$

   and we define $M_h : \mathbb{V}_{h,p} \to \mathbb{V}_{h,p}$ to be its restriction to $\mathbb{V}_{h,p}$, i.e.,

$$M = L_{u,h}^{-1}(L_{u,h} - F_h^{\mathrm{MA}}). \tag{8.3.7}$$

3. We prove that M_h satisfies the hypotheses of Banach's fixed point theorem on the ball of radius $\rho = h^{2+\alpha}$ for some $\alpha > 0$.

4. This proves the existence and uniqueness of a $u_h \in \mathbb{B}_\rho(u_*)$, with $\rho = h^{2+\alpha}$ that is a solution to the finite element method (8.2.1).

5. It also holds that M_h satisfies the hypotheses of Banach's fixed point theorem on the ball, $\mathbb{B}_\rho(u_*)$ of radius $\rho = h^p$. This proves the existence and uniqueness of a $u_h \in \mathbb{B}_{h^p}(u_*) \subset \mathbb{B}_{h^{2+\alpha}}(u_*)$ that is a solution to the finite element method (8.2.1). Since $\mathbb{B}_{h^p}(u_*) \subset \mathbb{B}_{h^{2+\alpha}}(u_*)$, we deduce that the two solutions coincide, and, furthermore, we use the fact that u_h belongs to the ball of radius $\rho = h^p$ in order to prove the desired optimal error estimate.

Remark 8.3.1 *Steps 1–5 require estimates on the operators L_u, $L_{u,h}$ and R, as well as a consistency result for the operator F^{MA}. Some of the proofs of these estimates are rather long, and thus have been left to the end of the Chapter. This makes the proof more accessible to the reader.*

Subsections 8.3.3 to 8.3.8 are now laid out as follows:

1. In subsection 8.3.3 we prove the estimates for L_u and $L_{u,h}$;
2. In subsection 8.3.4 we define the ball of radius ρ, and some estimates for functions in this ball;
3. In subsection 8.3.5 we state the estimate for R (proven at the end of the chapter);
4. In subsection 8.3.6 we prove the consistency result for F^{MA};
5. In subsection 8.3.7 we use the previous estimates to prove estimates for M and M_h that are required for the fixed point argument.
6. In subsection 8.3.8 we use the estimates for M and M_h to conclude the proof of Theorem 8.2.3.

8.3.3 Estimates for L_u and $L_{u,h}$

The proof of the main theorem of this subsection relies on several lemmas and steps. In what follows, we state the lemmas that will be used; the proofs of these results can be found in Appendix A.

Let us denote

$$A := \mathrm{Cof}(D^2 u), \quad \text{and} \quad A_h := \mathcal{P}_{\mathbb{W}_{h,p}}(\mathrm{Cof}(D^2 u)) = \mathcal{P}_{\mathbb{W}_{h,p}}(A).$$

Lemma 8.3.2 *The norm $\|\cdot\|_h$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$ when restricted to the finite element space $\mathring{\mathbb{V}}_{h,p}$, i.e., there exist constants C_1, C_2, independent of h, such that*

$$C_1\|v\|_{H^1(\Omega)} \leq \|v\|_h \leq C_2\|v\|_{H^1(\Omega)} \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{8.3.8}$$

Proof: See Lemma A.1. □

Lemma 8.3.3 *We have that*

$$h^{s-j}\|A_h\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)} + \|A - A_h\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} \lesssim h^{s-j}\|u\|_{W^{s+2,\infty}(\Omega)}, \quad 0 \leq j \leq s \leq p+1. \tag{8.3.9}$$

Proof: See Lemma A.2. □

Lemma 8.3.4 *For any $v \in \mathring{\mathbb{V}}_{h,p}$, if $h \leq h_0$, with $h_0 \in (0,1)$ sufficiently small, we have that*

$$\|Af_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)} \lesssim h^{2-m}\|\nabla v\|_{2,\Omega}, \quad m = 0, 1, \tag{8.3.10}$$

where $C > 0$ depends on the shape-regularity of the mesh, $\|u\|_{W^{p+3,\infty}(\Omega)}$ and for $K := \max_{x\in\overline{\Omega}}|\nabla u(x)|$, the value $\|f_2\|_{C^{p,\alpha}(\{q\in\mathbb{R}^2:|q|\leq K\})}$.

Proof: See Lemma A.5. □

Lemma 8.3.5 *For any $v \in \mathring{\mathbb{V}}_{h,p}$, we have that*

$$\sum_{F\in\mathscr{E}_h^{i,b}} h_F^{-1}\|[\![Af_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v)]\!]\|_{2,F}^2 \lesssim h^2\|\nabla v\|_{2,\Omega}^2. \tag{8.3.11}$$

Proof: This follows directly from the trace inequality (4.6.2), and (8.3.10). □

We will now prove the main stability theorem of this chapter.

Theorem 8.3.6 *There exists a positive constant C, independent of h, such that*

$$\|L_u[w]\|_{-1,h} \leq C\|w\|_h \quad \forall w \in V. \tag{8.3.12}$$

Moreover for h sufficiently small the operator $L_{u,h}$ is invertible and there exists a positive constant C, independent of h, with

$$\|L_{u,h}^{-1}[r]\|_h \leq C\|r\|_{-1,h} \quad \forall r \in \mathring{\mathbb{V}}_p'. \tag{8.3.13}$$

Proof: First we show that estimate (8.3.12) holds. By Lemma A.6, we have the following identity for the operator $L_u : V \to \mathring{V}'$: for all $w \in V$ and $v \in \mathring{V}$,

$$\begin{aligned}\langle L_u[w], v\rangle &= \int_\Omega f_2(\nabla u)A\nabla w\cdot\nabla v - \nabla w\cdot\nabla_h\cdot(A\, f_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)v))\\ &+ \frac{1}{2}\int_\Omega (D^2u + \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)) : (D^2u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))(D_q f_2(\nabla u)\cdot\nabla w)v\\ &+ \int_\Omega D_z f_1(x,u)wv - \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![[(\mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)v) - A\, f_2(\nabla u)v)\langle\!\langle\nabla w\rangle\!\rangle]\cdot n_F]\!].\end{aligned}$$

Thus, for any $w \in V$ and $v \in \mathring{\mathbb{V}}_{h,p}$ we have that

$$\begin{aligned}\langle L_u[w], v\rangle &\le \|A\|_{\infty,\Omega}\|f_2\|_{\infty,\{|q|\le K\}}|w|_{H^1(\Omega)}|v|_{H^1(\Omega)}\\ &\quad + |w|_{H^1(\Omega)}\|\nabla\cdot(A\, f_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v))\|_{2,\Omega}\\ &\quad + \|D_z f_1\|_{\infty,\Omega\times[-K,K]}\|w\|_{2,\Omega}\|v\|_{2,\Omega}\\ &\quad + \|A\|_{\infty,\Omega}\|D_q f_2\|_{\infty,\{|q|\le K\}}|w|_{H^1(\Omega)}\|v\|_{2,\Omega}\\ &+ \left(\sum_{F\in\mathscr{E}_h^{i,b}} h_F^{-1}\|[\![\mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v]\!]\|_{2,F}^2\right)^{1/2}\left(\sum_{F\in\mathscr{E}_h^{i,b}} h_F\|\langle\!\langle\nabla w\rangle\!\rangle\|_{2,F}^2\right)^{1/2}\\ &\lesssim \|w\|_h\|v\|_h.\end{aligned}$$

Note that the final inequality follows from (4.6.2), (4.6.27), (A.1), (8.3.10), and Lemma 8.3.11; we also use the fact that $u \in C^{p+2,\alpha}(\overline{\Omega})$ implies that there exists a constant $K > 0$ such that $|u|, |\nabla u| \le K$ in $\overline{\Omega}$. By the definition of the norm $\|\cdot\|_{-1,h}$ we obtain (8.3.12).

To prove the estimate (8.3.13) it is sufficient to show that $L_{u,h}$ is coercive on $\mathring{\mathbb{V}}_p$ with respect to the norm $\|\cdot\|_h$. It is important to note that since u is uniformly convex, it follows that $\operatorname{Cof} D^2u$ is positive definite. This, and the fact that f_2 is positive and uniformly bounded below, means that there exists a constant $\lambda_u > 0$ such that

$$\int_\Omega f_2(\nabla u)A\nabla w\cdot\nabla w \ge \lambda_u|w|_{H^1(\Omega)}^2 \quad \forall w \in H^1(\Omega).$$

By Lemma A.6, we also obtain the following for $w \in \mathring{\mathbb{V}}_{h,p}$:

$$\begin{aligned}\langle L_u[w], w\rangle &= \int_\Omega f_2(\nabla u)A\nabla w\cdot\nabla w - \nabla w\cdot(\nabla_h\cdot(Af_2(\nabla u)w - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)w)))\\ &+ \frac{1}{2}\int_\Omega (D^2u + \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)) : (D^2u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))(D_q f_2(\nabla u)\cdot\nabla w)w\\ &+ \int_\Omega D_z f_1(x,u)w^2 - \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![[(\mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)w) - Af_2(\nabla u)w)\langle\!\langle\nabla w\rangle\!\rangle]\cdot n_F]\!]\end{aligned}$$

$$
\begin{aligned}
&\geq \lambda_u |w|^2_{H^1(\Omega)} - \sum_{F\in\mathscr{E}_h^{i,b}} \int_F [\![(\mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)w) - A\, f_2(\nabla u)w)\langle\!\langle \nabla w \rangle\!\rangle] \cdot n_F]\!] \\
&\quad - C\|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)\|_{\infty,\Omega} |w|_{H^1(\Omega)} \|w\|_{L^2(\Omega)} \\
&\quad - \int_\Omega \nabla w \cdot \nabla \cdot (A\, f_2(\nabla u)w - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)w)).
\end{aligned}
$$

From (8.3.10) and (A.1) we see that

$$
\begin{aligned}
\int_\Omega \nabla w \cdot \nabla_h \cdot (A\, f_2(\nabla u)w - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)w)) \\
&\leq |w|_{H^1(\Omega)} \|\nabla_h \cdot (Aw - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h w))\|_{2,\Omega} \\
&\leq Ch|w|^2_{H^1(\Omega)}.
\end{aligned}
$$

Then, from (8.3.10), (8.3.11) and (4.6.2), we have

$$
\begin{aligned}
&\sum_{F\in\mathscr{E}_h^{i,b}} \int_F [\![(\mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)w) - A\, f_2(\nabla u)w)\langle\!\langle \nabla w \rangle\!\rangle] \cdot n_F]\!] \leq \\
&\leq \left(\sum_{F\in\mathscr{E}_h^{i,b}} h_F^{-1} \| [\![\mathcal{P}_{\mathbb{W}_{h,p}}(A_h\, f_2(\nabla u)w) - A\, f_2(\nabla u)w]\!] \|^2_{2,F} \right)^{1/2} \left(\sum_{F\in\mathscr{E}_h^{i,b}} h_F \| \langle\!\langle \nabla w \rangle\!\rangle \|^2_{2,F} \right)^{1/2} \\
&\leq Ch|w|^2_{H^1(\Omega)}.
\end{aligned}
$$

We also see that

$$
\begin{aligned}
C\|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)\|_{\infty,\Omega} |w|_{H^1(\Omega)} \|w\|_{L^2(\Omega)} &\leq Ch\|D^2 u\|_{W^{1,\infty}(\Omega)} |w|^2_{H^1(\Omega)} \\
&\leq Ch|w|^2_{H^1(\Omega)}.
\end{aligned}
$$

Combining these results we obtain

$$
\langle L_{u,h}[w], w\rangle \geq (\lambda_u - Ch)|w|^2_{H^1(\Omega)}.
$$

It now follows from the Poincaré inequality and (A.1), that for h sufficiently small we have

$$
\langle L_{u,h}[w], w\rangle \geq C\|w\|^2_{H^1(\Omega)} \quad \forall w \in \mathring{\mathbb{V}}_{h,p}.
$$

This concludes the proof of Theorem 8.3.6. □

8.3.4 The ball of radius ρ

Lemma 8.3.7 *Define $u_* \in \mathbb{V}_{h,p}$ with $u_*|_{\partial\Omega} = \phi_h|_{\partial\Omega} = (\pi_h\phi)|_{\partial\Omega}$ as*

$$
u_* = L^{-1}_{u,h}(L_u[u]), \tag{8.3.14}
$$

i.e.,

$$\langle L_{u,h}[u_*], v\rangle = \langle L_u[u], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_p.$$

Then we have that

$$h\|\mathbf{H}_h[u-u_*]\|_{2,\Omega} + \|u-u_*\|_h \lesssim h^p. \tag{8.3.15}$$

Proof: See [92], proof of Corollary 4.1; the same technique applies here, since it only relies on the boundedness of $L_{u,h}$ and its inverse, $L_{u,h}^{-1}$, which has been established by Theorem 8.3.6. $\square$

Remark 8.3.8 *The existence of a unique $u_* \in \mathbb{V}_{h,p}$ that satisfies the statement of Lemma 8.3.7 in fact follows from standard finite element theory. Since $L_{u,h}[u] \in H^1(\Omega)' \subset \mathbb{V}_{h,p}{}'$, and the map $a : \mathring{\mathbb{V}}_{h,p} \times \mathring{\mathbb{V}}_{h,p} \to \mathbb{R}$ defined by*

$$a(u,v) = \langle L_{u,h}[u], v\rangle \quad \forall u, v \in \mathring{\mathbb{V}}_{h,p},$$

is a bounded, coercive, bilinear functional; the existence and uniqueness can be established using the Lax–Milgram Theorem [46].

We now define the ball of radius ρ, on which we will apply our fixed point argument.

Definition 8.3.9 *Let us define the following closed ball*

$$\mathbb{B}_\rho(u_*) := \{v \in \mathbb{V}_{h,p} : v|_{\partial\Omega} = \phi_h,\ \|u_* - v\|_h \le \rho\}. \tag{8.3.16}$$

8.3.5 Estimate for the quadratic remainder term R

In this chapter we prove an important estimate for the remainder term R defined by (8.3.5). This estimate is in fact a contraction estimate when R is considered on the ball $\mathbb{B}_{h^{2+\alpha}}(u_*)$ for sufficiently small h, and some $\alpha > 0$.

Lemma 8.3.10 *Let $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u := \{v - u : v \in \mathbb{B}_{h^{2+\alpha}}(u_*)\}$, for some $\alpha > 0$. Then, if $h \le h_0$, for some $h_0 \in (0,1)$, we have the following estimate:*

$$\|R[w_1] - R[w_2]\|_{-1,h} \lesssim \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}\|w_1 - w_2\|_h. \tag{8.3.17}$$

Furthermore, for $w \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, we have

$$\|R[w]\|_{-1,h} \lesssim h^{2+\alpha}\sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}. \tag{8.3.18}$$

Proof: See Lemma A.8. $\square$

8.3.6 Consistency result for F^{MA}

We now consider the consistency of our method; that is we see how much "error" arises when we apply F^{MA} to the true solution u.

Lemma 8.3.11 *We have that*

$$\|F^{\mathrm{MA}}[u]\|_{-1,h} \lesssim h^{p+1} \tag{8.3.19}$$

Proof: For $v \in \mathring{\mathbb{V}}_{h,p}$ we see that

$$\begin{aligned}
\langle F^{\mathrm{MA}}[u], v\rangle &= \int_\Omega (f_1(x,u) - f_2(\nabla u)\det \mathbf{H}_h[u])v \\
&= \int_\Omega f_2(\nabla u)(\frac{f_1(x,u)}{f_2(\nabla u)} - \det(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))v \\
&= \int_\Omega f_2(\nabla u)(\det D^2u - \det(\mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))v \\
&= \frac{1}{2}\int_\Omega f_2(\nabla u)(A + A_h) : (D^2u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))v \\
&\lesssim \|A + A_h\|_{2,\Omega}\|D^2u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)\|_{\infty,\Omega}\|v\|_{2,\Omega} \\
&\lesssim h^{p+1}\|u\|_{W^{p+3,\infty}(\Omega)}\|v\|_{2,\Omega} \\
&\lesssim h^{p+1}\|v\|_h.
\end{aligned}$$

By the definition of $\|\cdot\|_{-1,h}$ we obtain the desired estimate. □

8.3.7 Estimates for M and M_h

Lemma 8.3.12 *Let $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*)$, for some $\alpha > 0$. If $h \le h_0$ for some sufficiently small $h_0 \in (0,1)$, then, we have that*

$$\|M_h[w_1] - M_h[v_2]\|_h \lesssim \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}\|w_1 - w_2\|_h. \tag{8.3.20}$$

Proof: Let $w = v - u \in V$, then we see that

$$\begin{aligned}
Mv &= L_{u,h}^{-1}(L_u[v] - F^{\mathrm{MA}}[v]) \\
&= L_{u,h}^{-1}((L_u[v] - L_u[v-u]) + (L_u[v-u] - F^{\mathrm{MA}}[v])) \\
&= L_{u,h}^{-1}((L_u[v] - L_u[v-u]) + (L_u[w] - F^{\mathrm{MA}}[u+w])) \\
&= L_{u,h}^{-1}((L_u[v] - L_u[v-u]) + (L_u[w] - L_u[w] - F^{\mathrm{MA}}[u] - R[w])) \\
&= L_{u,h}^{-1}(L_u[u] - F^{\mathrm{MA}}[u] - R[w]) \\
&= L_{u,h}^{-1}(u_* - F^{\mathrm{MA}}[u] - R[w])
\end{aligned}$$

and thus

$$M[w_1] - M[w_2] = L_{u,h}^{-1}\left(R[w_2 - u] - R[w_1 - u]\right) \quad \forall w_1, w_2 \in V.$$

Now, if $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*)$, for some $\alpha > 0$, then $w_i - u \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, $i = 1, 2$, and so by (8.3.13), and (8.3.17), it follows that

$$\begin{aligned} \|M[w_1] - M[w_2]\|_h &\leq \|L_{u,h}^{-1}(R[w_2 - u] - R[w_1 - u])\|_h \\ &\lesssim \|R[w_2 - u] - R[w_1 - u]\|_{-1,h} \\ &\lesssim \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}\|w_1 - w_2\|_h. \end{aligned} \tag{8.3.21}$$

for all $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*)$, as desired. $\square$

Lemma 8.3.13 *Let $w \in \mathbb{B}_{h^{2+\alpha}}(u_*)$, for some $\alpha > 0$. If $h \leq h_0$ for some sufficiently small $h_0 \in (0, 1)$, then, we have that*

$$\|u_* - M_h[w]\|_h \lesssim h^{p+1} + h^{2+\alpha}\sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}. \tag{8.3.22}$$

Proof: By (8.3.12), (8.3.18), and (8.3.19), we have that

$$\begin{aligned} \|u_* - M_h[w]\|_h &\leq \|L_{u,h}^{-1}(F^{\mathrm{MA}}[u])\|_h + \|L_{u,h}^{-1}(R[w - u])\|_h \\ &\lesssim \|F^{\mathrm{MA}}[u]\|_{-1,h} + \|R[w - u]\|_{-1,h} \\ &\lesssim h^{p+1} + h^{2+\alpha}\sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}, \end{aligned}$$

as desired. $\square$

8.3.8 Concluding the proof

Proof of Theorem 8.2.3: The bounds we have proven for $L_u, L_{u,h}$, and R as well as the consistency error estimate for F_h^{MA} are sufficient to prove that the restriction of the map $M : V \to \mathbb{V}_{h,p}$ defined by

$$M = L_{u,h}^{-1}(L_u - F^{\mathrm{MA}}),$$

to the finite element space is in fact a contraction on the ball $\mathbb{B}_{h^{2+\alpha}}(u_*)$ given by (8.3.16), for some $\alpha > 0$. Note that the restriction of M to the finite element space $\mathbb{V}_{h,p}$ is the map $M_h : \mathbb{V}_{h,p} \to \mathbb{V}_{h,p}$ given by

$$M_h = L_{u,h}^{-1}(L_{u,h} - F_h^{\mathrm{MA}}).$$

If $M_h : \mathbb{B}_{h^{2+\alpha}}(u_*) \to \mathbb{B}_{h^{2+\alpha}}(u_*)$ is a contraction map, then by Banach's fixed point theorem there exists a unique $u_h \in \mathbb{B}_{h^{2+\alpha}}(u_*)$ such that

$$M_h[u_h] = u_h,$$

i.e.,

$$L_{u,h}^{-1}(L_{u,h}[u_h] - F_h^{\mathrm{MA}}[u_h]) = u_h,$$

and thus

$$F_h^{\mathrm{MA}}[u_h] = 0, \quad \text{and} \quad u_h|_{\partial\Omega} = \phi_h.$$

It follows that u_h is the numerical solution. We shall now prove that M_h is a contraction map.

Choose $h_0 \in (0,1)$ sufficiently small so that for some $\alpha > 0$, Lemmas 8.3.12 and 8.3.13 hold, and that if $0 < h \leq h_0$, then

$$C_1 \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} \leq \eta \in (0,1), \tag{8.3.23}$$

and

$$C_2 \left(h^{p-\alpha-1} + \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} \right) \leq 1, \tag{8.3.24}$$

where C_1, C_2 are the fixed positive constants on the right-hand sides of (8.3.20) and (8.3.22), respectively.

Then if $h \leq h_0$, and $w, w_1, w_2 \in \mathbb{B}_{h^p}(u_*)$, by (8.3.20) we have that

$$\begin{aligned} \|M_h[w_1] - M_h[w_2]\|_h &\leq C_1 \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} \|w_1 - w_2\|_h \\ &\leq \eta \|w_1 - w_2\|_h, \end{aligned} \tag{8.3.25}$$

and by (8.3.22), we see that

$$\begin{aligned} \|u_* - M_h[w]\|_h &\leq C_2 \left(h^{p+1} + h^{2+\alpha} \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} \right) \\ &= C_2 \left(h^{p-\alpha-1} + \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} \right) h^{2+\alpha} \\ &\leq h^{2+\alpha}. \end{aligned} \tag{8.3.26}$$

Thus $M_h : \mathbb{B}_{h^{2+\alpha}}(u_*) \to \mathbb{B}_{h^{2+\alpha}}(u_*)$, and is a contraction. By Banach's fixed point theorem, there exists a unique $u_h \in \mathbb{B}_{h^{2+\alpha}}$ that is fixed point of M_h. Furthermore, u_h is a solution of (8.2.1). Thus we have deduced the existence and uniqueness of a

numerical solution u_h of (8.2.1) that is unique in the ball $\mathbb{B}_{h^{2+\alpha}}$. It now remains to prove the error estimate (8.2.3).

Since $p > 2$, we may apply the same argument on the ball $\mathbb{B}_\rho(u_*)$ with $\rho = h^p$, which yields the existence and uniqueness of a $u_h' \in \mathbb{B}_{h^p}(u_*)$ that is a fixed point of M_h and thus a solution of (8.2.1). Moreover (since $0 < h \leq h_0 < 1$), we have that $\mathbb{B}_{h^p}(u_*) \subset \mathbb{B}_{h^{2+\alpha}}(u_*)$, and thus by uniqueness, we deduce that $u_h \equiv u_h'$. In particular, we have that $u_h \in \mathbb{B}_{h^p}(u_*)$, and thus we obtain the following error estimate

$$\|u - u_h\|_h \leq \|u - u_*\|_h + \|u_* - u_h\|_h \lesssim h^p, \tag{8.3.27}$$

as desired. $\square$

Remark 8.3.14 (The ball of radius $\rho = h^{2+\alpha}$ is essential) *One should note that in order for such a $h_0 \in (0,1)$ to exist, such that (8.3.23) and (8.3.24) hold, it is essential that we consider the ball of radius $\rho = h^{2+\alpha}$ for $\alpha > 0$. Otherwise, if $\alpha \leq 0$, one may bound the left-hand side of (8.3.23) as follows:*

$$\begin{aligned} C_1 \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}} &\geq C_1 h^{(1+\alpha)-1}(1 + |\ln h|)^{\frac{1}{2}} \\ &= C_1 h^{\alpha}(1 + |\ln h|)^{\frac{1}{2}} \geq 1 \end{aligned}$$

for small values of h, contradicting the validity of the estimate (8.3.23) for all $h \in (0, h_0]$. If $\alpha \leq 0$, then we may similarly obtain a contradiction of estimate (8.3.24). Note that estimates (8.3.23) and (8.3.24) lead directly to the estimates (8.3.25) and (8.3.26), respectively, for M_h, that allow us to apply Banach's fixed point theorem.

Corollary 8.3.15 *Under the hypotheses of Theorem 8.2.3, the functions u, u_h satisfy the following error estimates*

$$\begin{aligned} \|\mathbf{H}_h u - \mathbf{H}_h u_h\|_{L^2(\Omega)} &\leq C h^{p-1}, \\ \|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)} &\leq C h^{p-1}, \end{aligned} \tag{8.3.28}$$

where the positive constant, C, is independent of the mesh size.

Proof: Note that for u_* defined by (8.3.14), $u_* - u_h \in \mathring{\mathbb{V}}_{h,p}$. We then see that

$$\begin{aligned} \|\mathbf{H}_h u - \mathbf{H}_h u_h\|_{L^2(\Omega)} &\leq \|\mathbf{H}_h u - \mathbf{H}_h u_*\|_{L^2(\Omega)} + \|\mathbf{H}_h u_* - \mathbf{H}_h u_h\|_{L^2(\Omega)} \\ &\leq C h^{p-1} + \|\mathbf{H}_h u_* - \mathbf{H}_h u_h\|_{L^2(\Omega)}, \end{aligned}$$

where the final inequality follows from (8.3.15). Now, by (4.11.5), followed by Poincaré's inequality, we obtain

$$\begin{aligned}\|\mathbf{H}_h u_* - \mathbf{H}_h u_h\|_{L^2(\Omega)} &\leq Ch^{-1}\|u_* - u_h\|_{H^1(\Omega)} \\ &\leq Ch^{-1}|u_* - u_h|_{H^1(\Omega)} \\ &\leq Ch^{-1}\|u_* - u_h\|_h \leq Ch^{p-1}.\end{aligned}$$

Thus, we obtain

$$\|\mathbf{H}_h u - \mathbf{H}_h u_h\|_{L^2(\Omega)} \leq Ch^{p-1}, \tag{8.3.29}$$

which is the first estimate of (8.3.28). Since $p \geq 3$, and $u \in C^{p+2,\beta}(\overline{\Omega}) \subset H^2(\Omega)$, it follows from Corollary 4.11.5 that $\mathbf{H}_h u = \mathcal{P}_{\mathbb{W}_{h,p}} D^2 u$. Thus

$$\begin{aligned}\|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)} &\leq \|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}} D^2 u\|_{L^2(\Omega)} + \|\mathcal{P}_{\mathbb{W}_{h,p}} D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)} \\ &= \|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}} D^2 u\|_{L^2(\Omega)} + \|\mathbf{H}_h u - \mathbf{H}_h u_h\|_{L^2(\Omega)} \\ &\leq \|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}} D^2 u\|_{L^2(\Omega)} + Ch^{p-1},\end{aligned} \tag{8.3.30}$$

where the final inequality follows from (8.3.29). We also see that

$$\|D^2 u - \mathcal{P}_{\mathbb{W}_{h,p}} D^2 u\|_{L^2(\Omega)} \leq \|D^2 u - \pi_h(D^2 u)\|_{L^2(\Omega)} \leq Ch^p.$$

Note that the final inequality follows from the fact that $D^2 u \in C^{p,\beta}(\overline{\Omega}) \subset H^p(\Omega) \subset H^p(\Omega; \mathscr{T}_h)$, which allows the application of (4.5.15). Combining the above estimate with (8.3.30), we obtain the second estimate of (8.3.28). □

8.4 Newton's Method for the MAD problem

One can see that (8.2.1) constitutes a *nonlinear* continuous Galerkin finite element method. Since it is not straightforward how one may find a $u_h \in \mathbb{V}_{h,p}$ with $u_h|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$ that satisfies (8.2.1), we apply instead Newton's method to the nonlinear operator defined by (8.2.1), resulting in the following iterative scheme.

8.4.1 Iterative scheme

Let us recall the finite element method given by (8.2.1): find $u_h \in \mathbb{V}_{h,p}$, with $u_h|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$ such that

$$\langle F_h^{\mathrm{MA}}[u_h], v\rangle = \int_\Omega (f_1(x, u_h) - f_2(\nabla u_h)\det(\mathbf{H}_h u_h))v = 0 \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{8.4.1}$$

Applying Newton's method, we obtain the following sequence of problems: given $u_h^0 \in \mathbb{V}_{h,p}$ satisfying $u_h^0|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$, for $k \in \mathbb{N}_0$ find $u_h^{k+1} \in \mathbb{V}_{h,p}$ satisfying $u_h^{k+1}|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$ such that

$$\langle L_{u_h^k,h}[u_h^{k+1} - u_h^k], v\rangle = -\langle F_h^{\mathrm{MA}}[u_h^k], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{8.4.2}$$

Note that (8.4.2) arises from applying Newton's method, and using the fact that $L_{u,h}[w_h]$ is the derivative of F_h^{MA} at u in the direction $w_h \in \mathbb{V}_{h,p}$. It thus follows that $L_{u_h^k,h}[w_h]$ is the derivative of F_h^{MA} at u_h^k in the direction $w_h \in \mathbb{V}_{h,p}$.

Denoting $\theta^k := u_h^{k+1} - u_h^k$, we see that (8.4.2) is given by

$$\begin{aligned}\int_\Omega (D_z f_1(x, u_h^k)\theta^k - (\mathrm{Cof}(\mathbf{H}_h u_h^k) : \mathbf{H}_h\theta^k) f_2(\nabla u_h^k) - \det(\mathbf{H}_h u_k) D_q f_2(\nabla u_h^k)\cdot\nabla\theta^k)v \\ = \int_\Omega (f_2(\nabla u_h^k)\det(\mathbf{H}_h u_h^k) - f_1(x, u_h^k))v \quad \forall v \in \mathring{\mathbb{V}}_{h,p}.\end{aligned} \tag{8.4.3}$$

Theorem 8.4.1 *Assume that $h \le h_0$ for some $h_0 \in (0,1)$, sufficiently small, and that $u_h \in \mathbb{B}_{h^p}(u_*)$ satisfies (8.2.1), with $f := f(x)$. Then if $u_h^0 \in \mathbb{V}_{h,p}$, such that $u_h^0|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$ satisfies*

$$\|u_h^0 - u_h\|_h \le C_* h^{2+\alpha},$$

for some constants $\alpha, C_* > 0$, then the sequence $\{u_h^k\}_{k=0}^\infty$ generated by (8.4.2) converges to u_h superlinearly. Moreover, there exists a $\xi \in (0,1)$, independent of h, such that

$$\|u_h^{k+1} - u_h\|_h \le \xi\|u_h^k - u_h\|_h \quad \forall k \in \mathbb{N}_0.$$

Proof: First let us define $L_k : \mathbb{V}_{h,p} \to \mathring{\mathbb{V}}'_{h,p}$ by

$$\langle L_k[w], v\rangle = -\int_\Omega \mathrm{Cof}(\mathbf{H}_h[u_k]) : \mathbf{H}_h[w]v \quad \forall w \in \mathbb{V}_{h,p},\ \forall v \in \mathring{\mathbb{V}}_{h,p}.$$

We will proceed to argue by induction; thus we shall assume that u_h^k satisfies $u_h^k|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$, and

$$\|u_h^k - u_h\|_h \le C_* h^{2+\alpha},$$

for some $\alpha > 0$. We define $u_h^{k+1} \in \mathbb{V}_{h,p}$ to be the solution of the Newton iteration given by (8.4.2), which is equivalent to

$$\langle L_k[u_h^{k+1} - u_h^k], v\rangle = -\langle F^{\mathrm{MA}}[u_h^k], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{8.4.4}$$

To prove that there exists a unique $u_h^{k+1} \in \mathbb{V}_{h,p}$ that satisfies $u_h^{k+1}|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$, and (8.4.4), we first recall from Theorem 8.3.6, that the operator $L_u : \mathbb{V}_{h,p} \to \mathring{\mathbb{V}}'_{h,p}$, given by

$$\langle L_u[v], w\rangle = -\int_\Omega \mathcal{P}_{\mathbb{W}_{h,p}}(\mathrm{Cof}(D^2u)) : \mathbf{H}_h[v]w \quad \forall w \in \mathbb{V}_{h,p},\ \forall v \in \mathring{\mathbb{V}}_{h,p},$$

is bounded and coercive on $\mathring{\mathbb{V}}_{h,p} \times \mathring{\mathbb{V}}_{h,p}$. We see that

$$\begin{aligned}\langle L_k[w], v\rangle &= \langle L_u[v], w\rangle + \langle (L_k - L_u)[w], v\rangle \\ &= -\int_\Omega \left(\mathcal{P}_{\mathbb{W}_{h,p}}(\mathrm{Cof}(D^2u)) : \mathbf{H}_h v\right) w \\ &\quad -\int_\Omega \left((\mathcal{P}_{\mathbb{W}_{h,p}}(\mathrm{cof}(D^2u)) - \mathbf{H}_h u_h^k) : \mathbf{H}_h w\right) v.\end{aligned}$$

Let us denote by $C_{u,1},\ C_{u,2}$, the coercivity constant and boundedness constant of L_u, respectively. Let $w, v \in \mathring{\mathbb{V}}_{h,p}$; recalling that $u_h \in \mathbb{B}_{h^p}(u_*)$ satisfies (8.2.1), we have that

$$\begin{aligned}\langle L_k[w], v\rangle &\le C_{u,1}\|w\|_h\|v\|_h + \|\mathcal{P}_{\mathbb{W}_{h,p}}(D^2u) - \mathbf{H}_h[u_h]\|_{2,\Omega}\|w\|_{\infty,\Omega}\|v\|_{2,\Omega} \\ &\quad + \|\mathbf{H}_h[u_h - u_k]\|_{2,\Omega}\|w\|_{\infty,\Omega}\|v\|_{2,\Omega} \\ &\le (C_{u,1} + 2C_3C_4C_p(1 + |\ln h|)h^{-1}(h^p + h^{2+\alpha})\|v\|_h\|w\|_h.\end{aligned}$$

Similarly we see that

$$\langle L_k[w], w\rangle \ge (C_{u,2} - 2C_3C_4C_p(1 + |\ln h|)h^{-1}(h^p + h^{2+\alpha})\|w\|_h^2.$$

Note that the constants $C_{u,1}, C_{u,2}, C_3, C_4$ and C_p do not depend upon h or k, and for $h \le h_0$, with $h_0 \in (0,1)$, sufficiently small, there exists a constant $C_5 > 0$ such that

$$\langle L_k[w], w\rangle \ge C_5\|w\|_h^2 \quad \forall w \in \mathring{\mathbb{V}}_{h,p},$$

where C_5 depends upon $C_{u,1}, C_{u,2}, C_3, C_4$ and C_p, but not upon h or k. Similarly there exists a constant $C_6 > 0$, independent of h and k such that

$$\langle L_k[w], v\rangle \le C_6\|w\|_h\|v\|_h \quad \forall v, w \in \mathring{\mathbb{V}}_{h,p}.$$

Thus L_k is bounded and coercive on $\mathring{\mathbb{V}}_{h,p} \times \mathring{\mathbb{V}}_{h,p}$, and the existence of a unique $\theta^k \in \mathring{\mathbb{V}}_{h,p}$ that satisfies

$$\langle L_k[\theta^k], v\rangle = -\langle F^{\mathrm{MA}}[u_h^k], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_{h,p}$$

follows from the Lax–Milgram theorem. We then define $u_h^{k+1} := \theta^k + u_h^k$, and note that since $\theta^k \in \mathring{\mathbb{V}}_{h,p}$, and $u_h^k|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$, it follows that $u_h^{k+1}|_{\partial\Omega} = \pi_h\phi|_{\partial\Omega}$.

Now since $u_h^{k+1} - u_h \in \mathring{\mathbb{V}}_{h,p}$, the coercivity of L_k yields

$$\|u_h^{k+1} - u_h\|_h^2 \leq \frac{1}{C_6}\langle L_k[u_h^{k+1} - u_h], u_h^{k+1} - u_h\rangle. \tag{8.4.5}$$

Since u_h^{k+1} satisfies (8.4.4), it follows that

$$\langle L_k[u_h^{k+1}], v\rangle = \langle L_k[u_h^k] - F^{\mathrm{MA}}[u_h^k], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_{h,p}.$$

Recall that u_h satisfies

$$\langle F^{\mathrm{MA}}[u_h], v\rangle = 0 \quad \forall v \in \mathring{\mathbb{V}}_{h,p},$$

and from this, we obtain that

$$\langle L_k[u_h], v\rangle = \langle L_k[u_h] - F^{\mathrm{MA}}[u_h], v\rangle \quad \forall v \in \mathring{\mathbb{V}}_{h,p}.$$

Thus for $v \in \mathring{\mathbb{V}}_{h,p}$ we have that

$$\begin{aligned}
\langle L_k[u_h^{k+1} - u_h], v\rangle &= -\langle F^{\mathrm{MA}}[u_h^k] - F^{\mathrm{MA}}[u_h] - L_k[u_h^k - u_h], v\rangle \\
&= -\int_\Omega (\det(\mathbf{H}_h u_h) - \det(\mathbf{H}_h u_h^k) + \mathrm{Cof}(\mathbf{H}_h u_h^k) : \mathbf{H}_h[u_h^k - u_h])v \\
&= -\int_\Omega (\det(\mathbf{H}_h[u_h^k + (u_h - u_h^k)]) - \det(\mathbf{H}_h u_h^k))v \\
&\quad + \int_\Omega (\mathrm{Cof}(\mathbf{H}_h u_h^k) : \mathbf{H}_h[u_h - u_h^k])v \\
&= -\int_\Omega \det(\mathbf{H}_h[u_h - u_h^k])v \\
&\leq \|\mathbf{H}_h[u_h^k - u_h]\|_{2,\Omega}^2 \|v\|_{\infty,\Omega} \\
&\leq C_9 h^{-2}(1 + |\ln h|)^{1/2}\|u_h^k - u_h\|_h^2\|v\|_h \\
&\leq C_9 C_* h^\alpha (1 + |\ln h|)^{1/2}\|u_h^k - u_h\|_h\|v\|_h.
\end{aligned} \tag{8.4.6}$$

Note that the final inequality follows from applying the inductive hypothesis. Taking $v = u_h^{k+1} - u_h$, from (8.4.5) and (8.4.6), we obtain

$$\|u_{k+1} - u_h\|_h \leq \frac{C_9 C_*}{C_6} h^\alpha (1 + |\ln h|)^{\frac{1}{2}}\|u_k - u_h\|_h.$$

Note that C_6, C_9, and C_* are independent of k and h. Thus for $h_0 \in (0,1)$ chosen sufficiently small, so that for $h \leq h_0$, we have

$$\frac{C_9 C_*}{C_6} h^\alpha (1 + |\ln h|)^{\frac{1}{2}} \leq \frac{C_9 C_*}{C_6} h_0^\alpha (1 + |\ln h_0|)^{\frac{1}{2}} \leq \xi,$$

for some $\xi \in (0,1)$, which implies that

$$\|u_h^{k+1} - u_h\|_h \leq \xi\|u_h^k - u_h\|_h,$$

where the value ξ is also independent of k and h. This concludes our inductive arguement.

Thus, by induction we deduce that for $h \leq h_0$, $h_0 \in (0,1)$, sufficiently small, there exists a $\xi \in (0,1)$, such that if

$$\|u_h - u_0\|_h \leq C_* h^{2+\alpha},$$

for some constants $C_*, \alpha > 0$, then

$$\|u_h^{k+1} - u_h\|_h \leq \xi \|u_k - u_h\|_h \quad \forall k \in \mathbb{N}_0,$$

where ξ is independent of h and k. It then follows that

$$\|u_h^{k+1} - u_h\|_h \leq \xi^k \|u_0 - u_h\|_h \leq \xi^k h^{2+\alpha} \to 0 \quad \text{as} \quad k \to \infty.$$

Thus the sequence $\{u_h^k\}_{k=0}^{\infty}$ converges to u_h superlinearly. $\square$

8.5 A modified method

Consider the following MA type equation

$$\begin{cases} \det(D^2 u) + \dfrac{\operatorname{div}(\mathcal{A}(\nabla u)) - f_1(x,u)}{f_2(\nabla u)} = 0, & \text{in } \Omega, \\ u = \phi, & \text{on } \partial\Omega, \end{cases} \tag{8.5.1}$$

where $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth function. The finite element method for the approximation of solutions of (8.5.1) is given as follows: find $u_h \in \mathbb{V}_{h,p}$ satisfying $u_h|_{\partial\Omega} = \pi_h \phi|_{\partial\Omega}$, such that

$$\int_\Omega (f_1(x,u_h) - f_2(\nabla u_h) \det \mathbf{H}_h u_h - D\mathcal{A}(\nabla u_h) : \mathbf{H}_h u_h)\varphi_h \quad \forall \varphi_h \in \mathring{\mathbb{V}}_{h,p}. \tag{8.5.2}$$

This method is a modification of the FEM given by (8.2.1), allowing for the approximation of the nonlinear term $\operatorname{div}(\mathcal{A}(\nabla u))$.

Using the chain rule, one can calculate that $\operatorname{div}(\mathcal{A}(\nabla u)) = D\mathcal{A}(\nabla u) : D^2 u$, and so, it is clear that this method is motivated by replacing $D^2 u$ with the finite element Hessian of $u_h \in \mathring{\mathbb{V}}_{h,p}$. We will see in the following experiments that problems of the form (8.5.1) also arise in differential geometry, in particular, the equations of prescribed Gaussian curvature ($\mathcal{A} \equiv 0$), and prescribed Weingarten curvature.

8.6 Implementation

Software and code: The experiments in this Chapter have been implemented in the most recent version of the Firedrake software [105, 87] (as of 3rd July 2018), which interfaces directly with PETSc [6, 7] running through a Python interface [39, 63]. A working Firedrake script, MA-Dirichlet-NVFEM.py used to generate the experiments of this Chapter is available in the Github repository:
https://github.com/ekawecki/FiredrakeNDV.

Two-dimensional curved boundary approximation: When implementing curved finite elements, we use a piecewise quadratic polynomial mapping to obtain a higher order approximation of the domain boundary. This is implemented in exactly the same manner as discussed in Section 5.8. As in Section 5.8, we define the space $\mathbb{V}_{h,p}^{\text{comp}} := \{v \in C(\overline{\Omega}) : v \circ T^{-1} \in \mathbb{P}^p(\hat{K})\}$, where the piecewise quadratic function T is defined by (5.8.1). We then define $\mathring{\mathbb{V}}_{h,p}^{\text{comp}} := \mathbb{V}_{h,p}^{\text{comp}} \cap H_0^1(\Omega)$, and $\mathbb{W}_{h,p}^{\text{comp}} := [\mathbb{V}_{h,p}^{\text{comp}}]^{2\times 2}$.

Furthermore, when we refine the mesh in our experiments, the meshes at each refinement level are not related to one another. That is, there is no hierarchical mesh structure, i.e., at each refinement level, we "remesh". A collection of the meshes used for the computations of this thesis can be found in the folder "Meshes" in the Github repository: https://github.com/ekawecki/FiredrakeNDV.

Initial guess selection criteria: In the following experiments, we are required to provide an initial guess for the Newton's method. This has an impact upon the convergence of the method, which we shall briefly discuss. Our initial guess $(u_h^0, H_h^0) \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}} \times \mathbb{W}_{h,p}^{\text{comp}}$ takes the following form

$$\begin{aligned} u_{h,r}^0 &:= r(x^2 + y^2 - 1) \\ H_{h,r}^0 &:= 2rI_d, \end{aligned} \tag{8.6.1}$$

for some $r \in \mathbb{R} \setminus \{0\}$, where I_d is the 2×2 identity matrix. The choice of $r \in \mathbb{R} \setminus \{0\}$ corresponds to starting with an initial guess that is either *uniformly finite element convex* ($r > 0$) or *uniformly finite element concave* ($r < 0$). Interestingly, Newton's method appears to converge to the uniformly convex solution of the MA equation if $r > 0$, or the uniformly concave solution if $r < 0$.

Furthermore, as derived in Section 8.4.1, Newton's method applied to (8.4.1) with $f_1(x) := f(x)$ and $f_2(q) \equiv 1$ gives us the following iterative scheme: given

$(u_{h,r}^k, H_{h,r}^k) \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}} \times \mathbb{W}_{h,p}^{\text{comp}}$ find $(u_{h,r}^{k+1}, H_{h,r}^{k+1}) \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}} \times \mathbb{W}_{h,p}^{\text{comp}}$ satisfying

$$\begin{aligned} -\int_\Omega (\text{Cof}(H_{h,r}^k) : \mathbf{H}_h u_{h,r}^{k+1}) v_h &= -\int_\Omega (\text{Cof}(H_{h,r}^k) : \mathbf{H}_h u_h^n + f - \det(H_{h,r}^k)) v_h \\ &= -\int_\Omega (f + \det(H_{h,r}^k) v_h) \quad \forall v_h \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}}. \end{aligned}$$

It is then clear that if the initial guess $(u_{h,r}^0, H_{h,r}^0) \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}} \times \mathbb{W}_{h,p}^{\text{comp}}$ generates the sequence of Newton iterates $\{(u_{h,r}^k, H_{h,r}^k)\}_{k=1}^\infty$, then the initial guess

$$(u_{h,-r}^0, H_{h,-r}^0) = (-u_{h,r}^0, -H_{h,r}^0) \in \mathring{\mathbb{V}}_{h,p}^{\text{comp}} \times \mathbb{W}_{h,p}^{\text{comp}}$$

generates the sequence of Newton iterates

$$\{(u_{h,-r}^k, H_{h,-r}^k)\}_{k=1}^\infty = \{(-u_{h,r}^k, -H_{h,r}^k)\}_{k=1}^\infty.$$

In Experiments 8.7.1 and 8.7.2 we observe that if we start with an initial guess of the form (8.6.1), with a fixed $r = r^* > 0$, then the sequence of Newton iterates converges to the uniformly convex solution, u, of the corresponding MA equations (8.7.1) and (8.7.3). Thus, we also see that starting with the initial guess of the form (8.6.1), with $r = -r^* < 0$, the sequence of of Newton iterates converges to the uniformly concave solution, $-u$, of the corresponding MA equation. From our previous observation (and from running the experiments), we see that the convergence rates are identical for $r = \pm r^*$, and thus, for these experiments, we provide the convergence rates for $r = r^* > 0$, i.e., we choose to approximate the uniformly convex solution.

8.7 Experiments

In the following experiments, we successively increase the degree, p, of the finite element space $\mathbb{V}_{h,p}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly. We implement Newton's method until the increment error, $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$, falls below desired tolerance of 10^{-12}. Note that we denote by u_h the final Newton iterate, i.e., for the index N, we have that $u_h = u_h^N$, and $\|u_h^N - u_h^{N-1}\|_{L^2(\Omega)} < 10^{-12}$.

8.7.1 Experiment 1

In this experiment, we consider the following MA problem:

$$\begin{cases} \det D^2 u(x) = f(x) & x \in \Omega, \\ u(x) = 0 & x \in \partial\Omega, \end{cases} \tag{8.7.1}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and f is chosen so that the true solution of (8.7.1) is given by

$$u(x, y) = 5(x_1^2 + x_2^2 - 1) - \frac{1}{8}\sin(\pi(x^2 + y^2)).$$

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\text{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence for the errors $|u - u_h|_{H^1(\Omega)}$ and $\|u - \mathbf{H}_h u_h\|_{L^2(\Omega)}$ are optimal, that is $\|D^2 u - \mathbf{H}_h u_h\|_{h,1} = \mathcal{O}(h^{p-1})$ and $|u - u_h|_{H^1(\Omega)} = \mathcal{O}(h^p)$. We plot the error values $|u - u_h|_{H^1(\Omega)}$ (left) and $\|u - \mathbf{H}_h u_h\|_{L^2(\Omega)}$ (right) in Figure 8.1, and report the exact values in Table 8.2, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 8.3. In this example, Newton's method requires an initial function u_h^0, and an initial Hessian H_h^0 our initial guesses are given by

$$\begin{aligned} u_h^0 &:= 5(x_1^2 + x_2^2 - 1) \\ H_h^0 &:= 10 I_d, \end{aligned} \tag{8.7.2}$$

where I_d is the 2×2 identity matrix.

We plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 8.2, 8.3, and 8.4, respectively. Across all polynomial degrees and mesh refinements we see that the number of Newton iterations required to reach the desired tolerance is exactly 5.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	5.25×10^{-1}		6.35×10^{-2}		4.85×10^{-2}	
0.2828	9.37×10^{-2}	(3.04)	2.03×10^{-2}	(2.01)	4.97×10^{-3}	(4.02)
0.1627	2.75×10^{-2}	(2.22)	4.29×10^{-3}	(2.81)	5.65×10^{-4}	(3.94)
0.0973	8.28×10^{-3}	(2.34)	7.16×10^{-4}	(3.48)	6.27×10^{-5}	(4.28)
0.0508	2.15×10^{-3}	(2.08)	9.73×10^{-5}	(3.07)	5.33×10^{-6}	(3.79)
0.0269	5.82×10^{-4}	(2.05)	1.28×10^{-5}	(3.18)	4.95×10^{-7}	(3.73)
0.0138	1.48×10^{-4}	(2.05)	1.61×10^{-6}	(3.09)	4.42×10^{-8}	(3.61)

Table 8.1: Error values in the $|\cdot|_{H^1(\Omega)}$-seminorm and EOCs for Experiment 8.7.1.

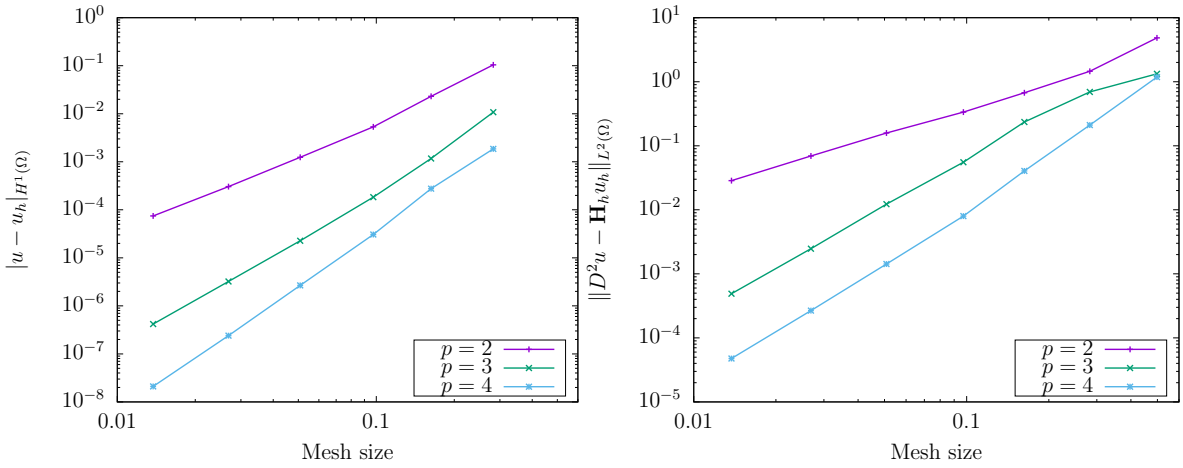


Figure 8.1: We provide the error values $|u - u_h|_{H^1(\Omega)}$ (left), and $\|D^2u - \mathbf{H}_h u_h\|_{L^2(\Omega)}$ (right), along with the experimental orders of convergence. We observe that the convergence rates are optimal with respect to the choice of polynomial degree, p. That is, $|u - u_h|_{H^1(\Omega)} = O(h^p)$, and $\|D^2u - \mathbf{H}_h u_h\|_{L^2(\Omega)} = O(h^{p-1})$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.4981	4.85		1.33		1.18	
0.2828	1.45	(2.13)	6.92×10^{-1}	(1.15)	2.10×10^{-1}	(3.05)
0.1627	6.73×10^{-1}	(1.39)	2.35×10^{-1}	(1.95)	4.05×10^{-2}	(2.98)
0.0973	3.36×10^{-1}	(1.35)	5.52×10^{-2}	(2.82)	7.97×10^{-3}	(3.16)
0.0508	1.58×10^{-1}	(1.16)	1.22×10^{-2}	(2.32)	1.42×10^{-3}	(2.65)
0.0269	6.92×10^{-2}	(1.30)	2.47×10^{-3}	(2.52)	2.67×10^{-4}	(2.63)
0.0138	2.86×10^{-2}	(1.32)	4.90×10^{-4}	(2.41)	4.76×10^{-5}	(2.57)

Table 8.2: The error values $\|u - \mathbf{H}_h u_h\|_{h,1}$ and EOCs for Experiment 8.7.1.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.4981	0.23	0.29	0.26	164	340	580
0.2828	0.19	0.19	0.26	580	1252	2180
0.1627	0.21	0.32	0.54	1508	3304	5796
0.0973	0.46	0.95	2.00	4772	10576	18660
0.0508	2.12	5.93	13.87	18980	42388	75076
0.0269	14.24	46.70	113.62	70788	158656	281508
0.0138	112.60	355.61	886.74	276084	619972	1101092

Table 8.3: Runtimes and number of DoFs for Experiment 8.7.1, for each mesh size h, and each polynomial degree, p.

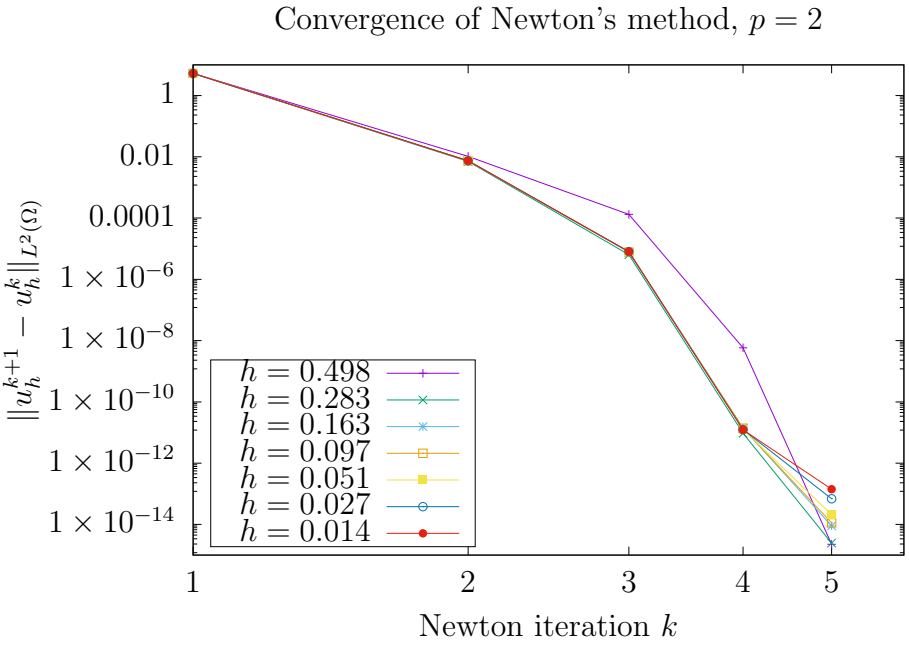


Figure 8.2: Convergence of Newton's method for the numerical scheme applied to problem (8.7.1) with $p = 2$.

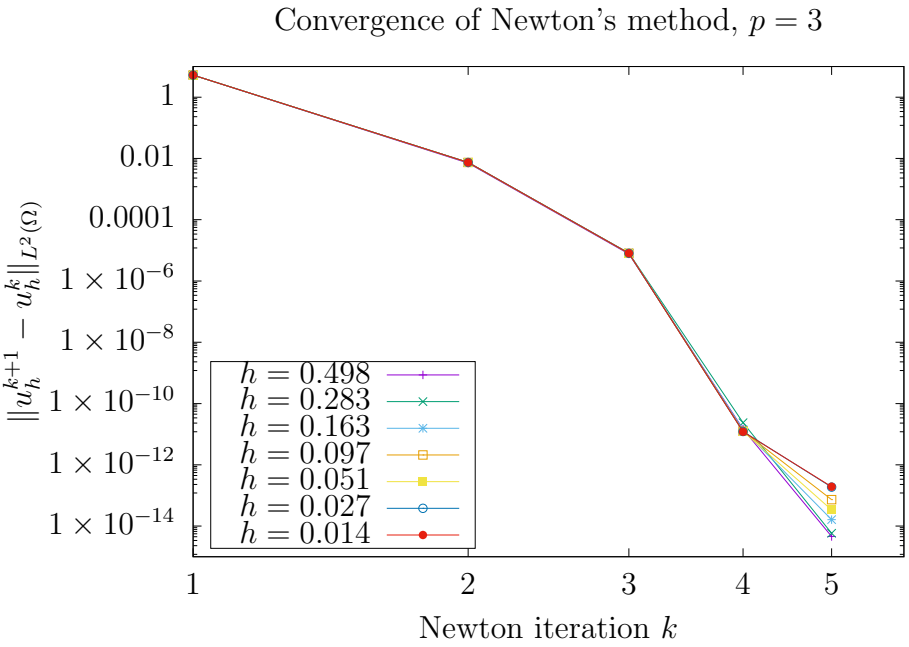


Figure 8.3: Convergence of Newton's method for the numerical scheme applied to problem (8.7.1) with $p = 3$.

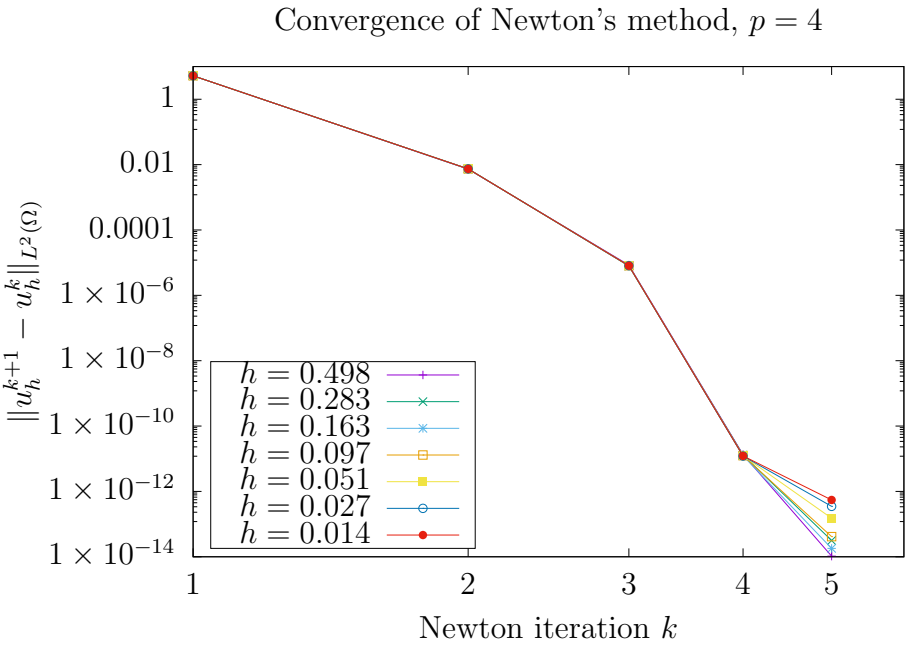


Figure 8.4: Convergence of Newton's method for the numerical scheme applied to problem (8.7.1) with $p = 4$.

8.7.2 Experiment 2

In this experiment, we consider the following problem of prescribed Gaussian curvature:

$$\begin{cases} \det(D^2 u) = K(|\nabla u|^2 + 1)^2, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{8.7.3}$$

where $\Omega = \{(x, y) \in \mathbb{R}^2 : |x| < 1\}$. In this case the function K is chosen so that the solution of (8.7.3) is given by

$$u(x, y) = \frac{5}{2}(x^2 + y^2 - 1) + \frac{1}{20}(\cos \pi(x^2 + y^2) + 1). \tag{8.7.4}$$

In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\text{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence for the errors $|u - u_h^N|_{H^1(\Omega)}$ and $\|u - \mathbf{H}_h u_h^N\|_{L^2(\Omega)}$ are optimal, that is $\|D^2 u - \mathbf{H}_h u_h^N\|_{h,1} = \mathcal{O}(h^{p-1})$ and $|u - u_h^N|_{H^1(\Omega)} = \mathcal{O}(h^p)$. We plot the error values $|u - u_h^N|_{H^1(\Omega)}$ (left) and $\|u - \mathbf{H}_h u_h^N\|_{L^2(\Omega)}$ (right) in Figure 8.5, and report the exact values in Table 8.5, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 8.6.

In this example, Newton's method requires an initial function u_h^0, and an initial Hessian H_h^0 our initial guesses are given by

$$u_h^0 := 0$$
$$H_h^0 := 10 I_d,$$

where I_d is the 2×2 identity matrix.

We plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 8.6, 8.7, and 8.8, respectively. Across all polynomial degrees and mesh refinements we see that the number of Newton iterations required to reach the desired tolerance ranges between 5 and 6.

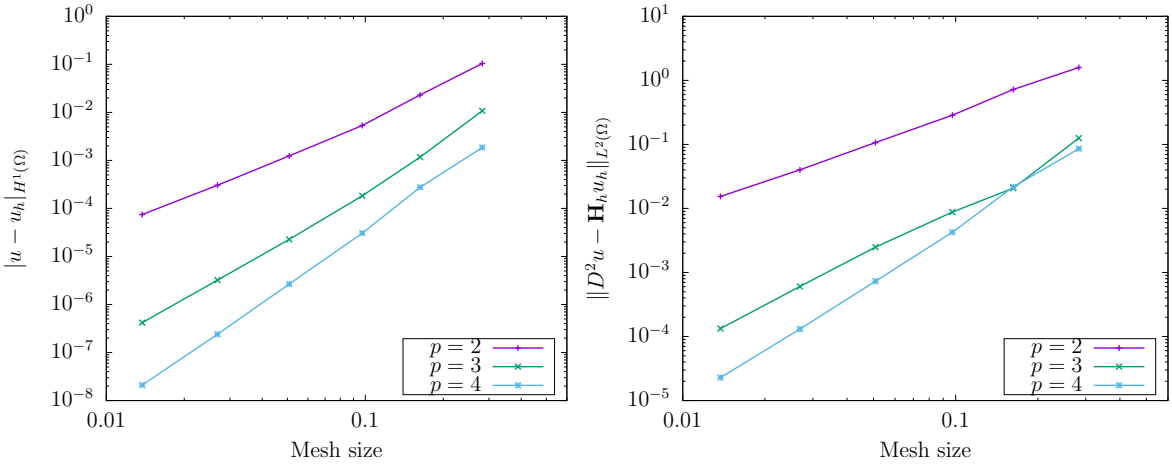


Figure 8.5: We provide the error values $|u - u_h|_{H^1(\Omega)}$ (left), and $\|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)}$ (right), along with the experimental orders of convergence. We observe that the convergence rates are optimal with respect to the choice of polynomial degree, p. That is, $|u - u_h|_{H^1(\Omega)} = O(h^p)$, and $\|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)} = O(h^{p-1})$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	1.04×10^{-1}		1.08×10^{-2}		1.86×10^{-3}	
0.1627	2.31×10^{-2}	(2.73)	1.17×10^{-3}	(4.02)	2.76×10^{-4}	(3.45)
0.0973	5.33×10^{-3}	(2.85)	1.83×10^{-4}	(3.61)	3.07×10^{-5}	(4.28)
0.0508	1.24×10^{-3}	(2.24)	2.26×10^{-5}	(3.22)	2.67×10^{-6}	(3.76)
0.0269	3.05×10^{-4}	(2.20)	3.23×10^{-6}	(3.06)	2.40×10^{-7}	(3.78)
0.0138	7.45×10^{-5}	(2.10)	4.18×10^{-7}	(3.05)	2.10×10^{-8}	(3.64)

Table 8.4: Error values in the $|\cdot|_{H^1(\Omega)}$-seminorm and EOCs for Experiment 8.7.2.

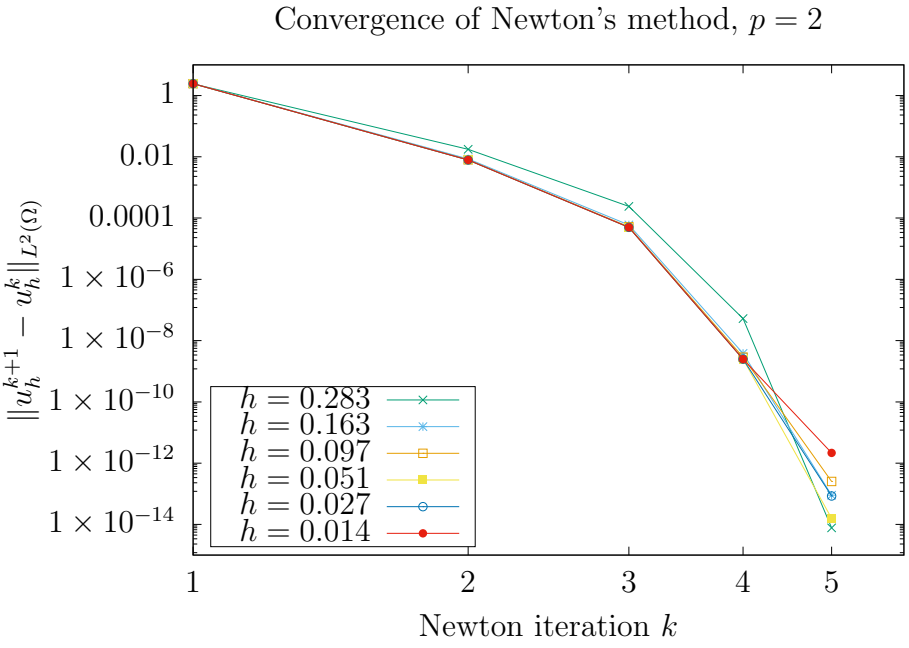


Figure 8.6: Convergence of Newton's method for the numerical scheme applied to problem (8.7.3) with $p = 2$.

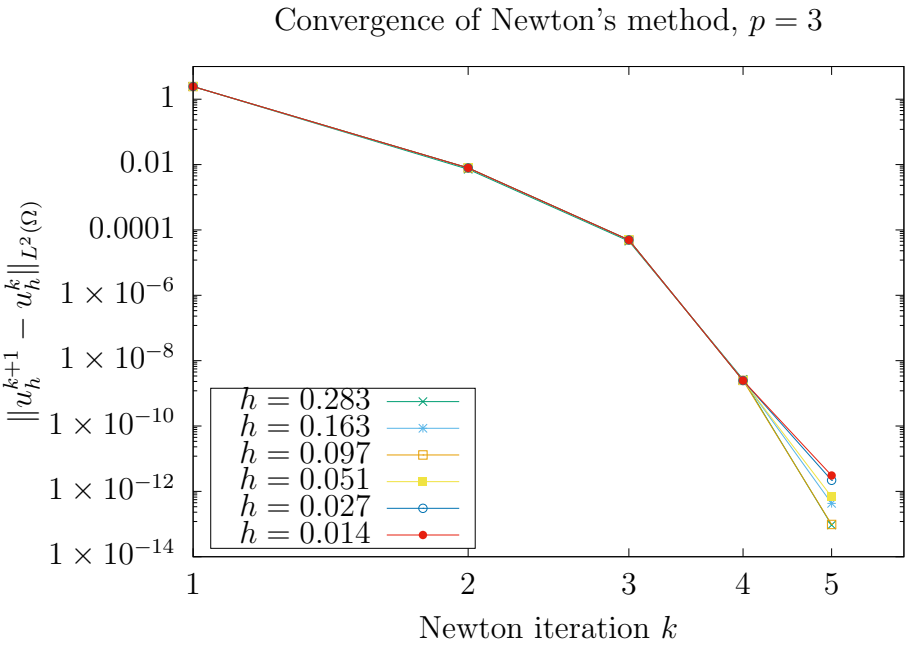


Figure 8.7: Convergence of Newton's method for the numerical scheme applied to problem (8.7.3) with $p = 3$.

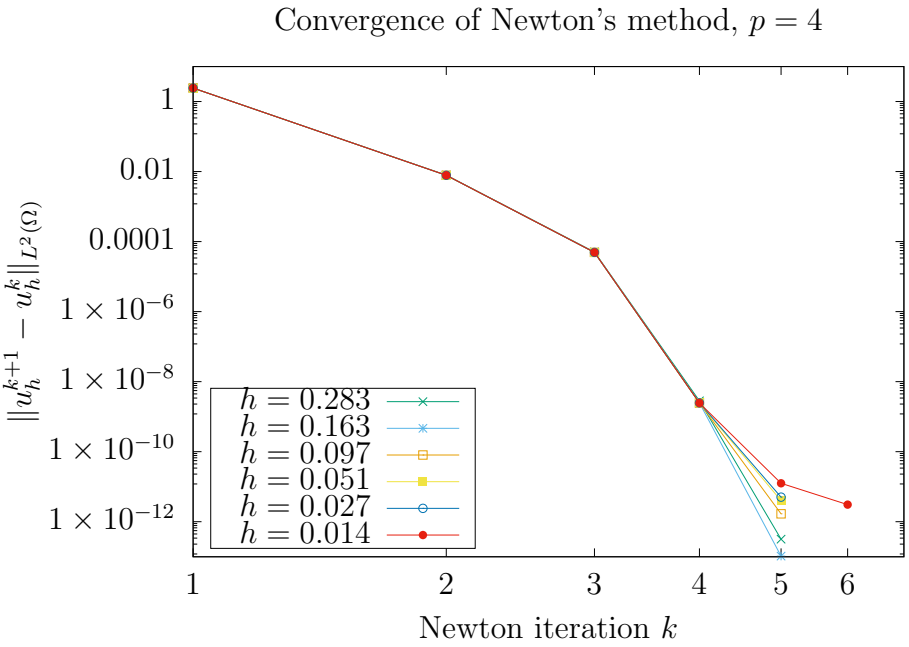


Figure 8.8: Convergence of Newton's method for the numerical scheme applied to problem (8.7.3) with $p = 4$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	1.59		1.25×10^{-1}		8.55×10^{-2}	
0.1627	7.21×10^{-1}	(1.43)	2.07×10^{-2}	(3.26)	2.18×10^{-2}	(2.48)
0.0973	2.86×10^{-1}	(1.80)	8.80×10^{-3}	(1.67)	4.25×10^{-3}	(3.18)
0.0508	1.07×10^{-1}	(1.52)	2.49×10^{-3}	(1.95)	7.31×10^{-4}	(2.71)
0.0269	4.01×10^{-2}	(1.53)	6.06×10^{-4}	(2.22)	1.31×10^{-4}	(2.70)
0.0138	1.54×10^{-2}	(1.43)	1.33×10^{-4}	(2.26)	2.28×10^{-5}	(2.61)

Table 8.5: The error values $\|u - \mathbf{H}_h u_h\|_{h,1}$ and EOCs for Experiment 8.7.2.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.2828	0.28	0.42	0.39	580	1252	2180
0.1627	0.25	0.39	0.59	1508	3304	5796
0.0973	0.48	1.07	2.11	4772	10576	18660
0.0508	2.14	6.88	13.92	18980	42388	75076
0.0269	14.62	50.14	106.34	70788	158656	281508
0.0138	104.76	344.84	970.57	276084	619972	1101092

Table 8.6: Runtimes and number of DoFs for Experiment 8.7.2, for each mesh size h, and each polynomial degree, p.

8.7.3 Experiment 3

In this experiment, we consider the following problem of prescribed Weingarten curvature:

$$\begin{cases} \dfrac{\det(D^2u)}{(|\nabla u|^2+1)^2} + \operatorname{div}\left(\dfrac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right) = W, \quad \text{in } \Omega, \\ u = 0, \quad \text{on} \quad \partial\Omega, \end{cases} \tag{8.7.5}$$

where $\Omega = \{(x,y) \in \mathbb{R}^2 : |x| < 1\}$. In this case the function W is chosen so that the solution of (8.7.5) is given by (8.7.4). Note that the first and second term on the left-hand side of (8.7.5) are the Gaussian curvature and mean curvature of the surface

$$Z := \{(x,y,z) \in \Omega \times \mathbb{R} : z = u(x,y)\},$$

respectively. The inclusion of the mean curvature term in (8.7.5) demonstrates the applicability of the finite element method given by (8.5.2). In the case of (8.7.5), $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ is given by $\mathcal{A}(q) = q/\sqrt{1+|q|^2}$. Furthermore, f_2 is given by $1/(1+|q|^2)^2$, $f_1(x,u) := W(x)$, and $\phi \equiv 0$. In this experiment, we successively increase the degree, p, of the finite element space $V_{h,p}^{\text{comp}}$ from 2 to 4, and for each fixed degree we refine the mesh quasi–uniformly, we observe that the experimental orders of convergence for the errors $|u - u_h^N|_{H^1(\Omega)}$ and $\|u - \mathbf{H}_h u_h^N\|_{L^2(\Omega)}$ are optimal, that is $\|D^2u - \mathbf{H}_h u_h^N\|_{h,1} = \mathcal{O}(h^{p-1})$ and $|u - u_h^N|_{H^1(\Omega)} = \mathcal{O}(h^p)$. We plot the error values $|u - u_h^N|_{H^1(\Omega)}$ (left) and $\|u - \mathbf{H}_h u_h^N\|_{L^2(\Omega)}$ (right) in Figure 8.9, and report the exact values in Table 8.8, with the corresponding experimental orders of convergence given in brackets. Furthermore, we provide the number of degrees of freedom (DoFs) and run times for each computation in Table 8.9. In this example, Newton's method requires an initial function u_h^0, and an initial Hessian H_h^0, our initial guesses are given

by

$$u_h^0 := 5(x_1^2 + x_2^2 - 1)$$
$$H_h^0 := 10 I_d,$$

where I_d is the 2×2 identity matrix.

We plot the incremental L^2-Newton error $\|u_h^{k+1} - u_h^k\|_{L^2(\Omega)}$ against the number of Newton iterations, k, for all levels of mesh refinements, for each degree $p = 2, 3, 4$ in Figures 8.10, 8.11, and 8.12, respectively. Across all polynomial degrees and mesh refinements we see that the number of Newton iterations required to reach the desired tolerance is exactly 5.

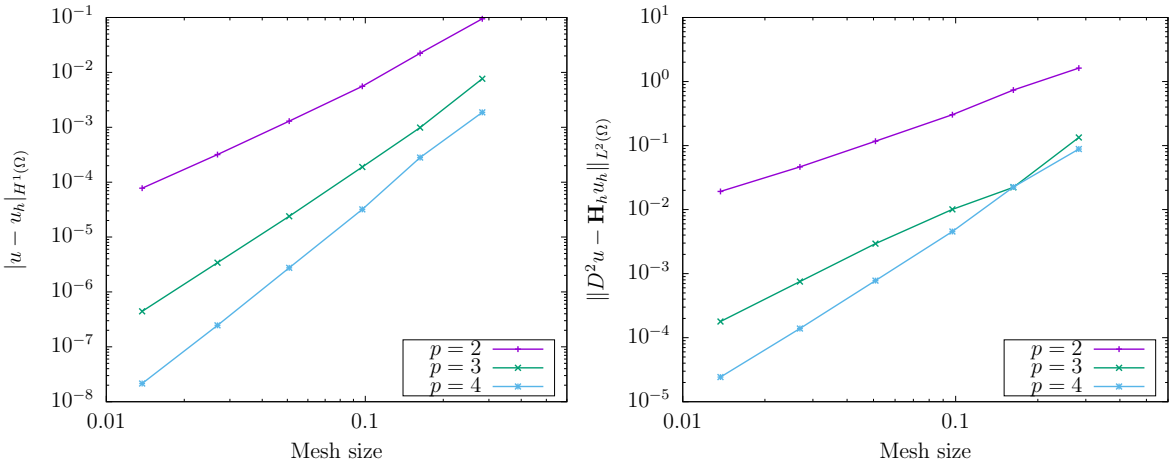


Figure 8.9: We provide the error values $|u - u_h|_{H^1(\Omega)}$ (left), and $\|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)}$ (right), along with the experimental orders of convergence. We observe that the convergence rates are optimal with respect to the choice of polynomial degree, p. That is, $|u - u_h|_{H^1(\Omega)} = O(h^p)$, and $\|D^2 u - \mathbf{H}_h u_h\|_{L^2(\Omega)} = O(h^{p-1})$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	9.48×10^{-2}		7.64×10^{-3}		1.87×10^{-3}	
0.1627	2.23×10^{-2}	(2.62)	9.87×10^{-4}	(3.71)	2.81×10^{-4}	(3.43)
0.0973	5.57×10^{-3}	(2.70)	1.89×10^{-4}	(3.21)	3.20×10^{-5}	(4.23)
0.0508	1.29×10^{-3}	(2.24)	2.39×10^{-5}	(3.19)	2.75×10^{-6}	(3.78)
0.0269	3.18×10^{-4}	(2.20)	3.42×10^{-6}	(3.05)	2.47×10^{-7}	(3.78)
0.0138	7.78×10^{-5}	(2.10)	4.43×10^{-7}	(3.05)	2.15×10^{-8}	(3.64)

Table 8.7: Error values in the $|\cdot|_{H^1(\Omega)}$-seminorm and EOCs for Experiment 8.7.3.

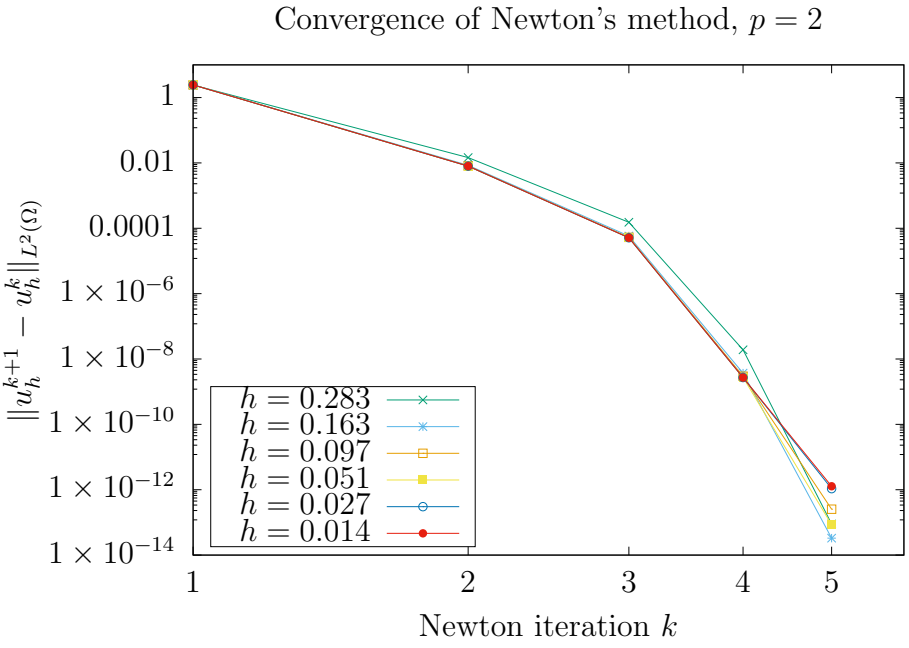


Figure 8.10: Convergence of Newton's method for the numerical scheme applied to problem (8.7.5) with $p = 2$.

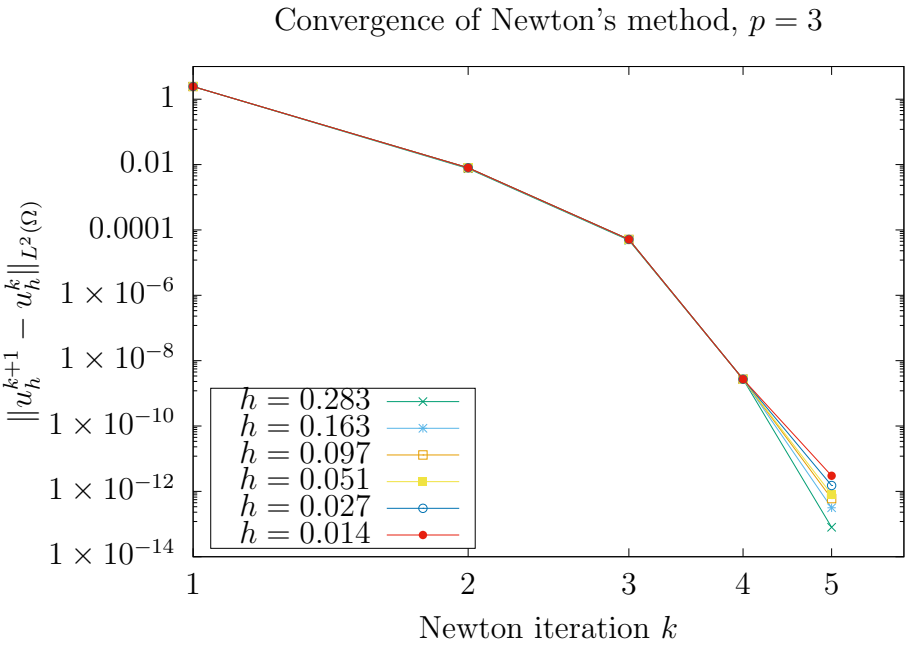


Figure 8.11: Convergence of Newton's method for the numerical scheme applied to problem (8.7.5) with $p = 3$.

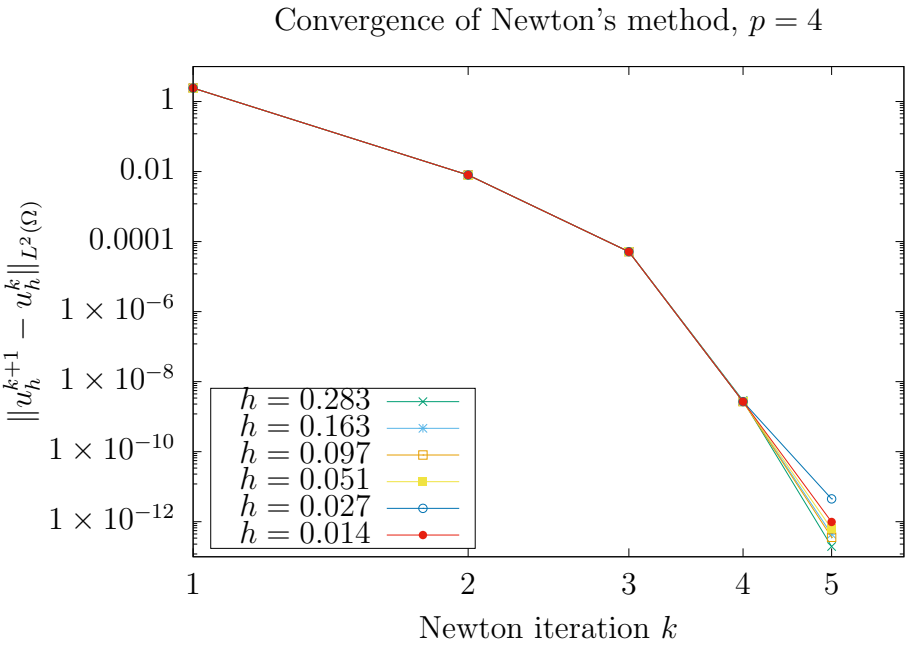


Figure 8.12: Convergence of Newton's method for the numerical scheme applied to problem (8.7.5) with $p = 4$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	1.62		1.34×10^{-1}		8.83×10^{-2}	
0.1627	7.37×10^{-1}	(1.43)	2.23×10^{-2}	(3.25)	2.25×10^{-2}	(2.47)
0.0973	3.03×10^{-1}	(1.73)	1.01×10^{-2}	(1.54)	4.55×10^{-3}	(3.11)
0.0508	1.17×10^{-1}	(1.47)	2.95×10^{-3}	(1.89)	7.77×10^{-4}	(2.72)
0.0269	4.65×10^{-2}	(1.45)	7.52×10^{-4}	(2.14)	1.39×10^{-4}	(2.70)
0.0138	1.92×10^{-2}	(1.32)	1.79×10^{-4}	(2.15)	2.42×10^{-5}	(2.61)

Table 8.8: The error values $\|u - \mathbf{H}_h u_h\|_{h,1}$ and EOCs for Experiment 8.7.3.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
0.2828	0.50	0.49	0.70	580	1252	2180
0.1627	0.45	0.52	0.74	1508	3304	5796
0.0973	0.86	1.24	2.33	4772	10576	18660
0.0508	3.01	6.60	15.86	18980	42388	75076
0.0269	15.63	50.61	117.42	70788	158656	281508
0.0138	125.57	384.11	873.55	276084	619972	1101092

Table 8.9: Runtimes and number of DoFs for Experiment 8.7.3, for each mesh size h, and each polynomial degree, p.

8.7.4 Experiment 4 - A comparison of the methods

In this experiment, we consider the following MA problem:

$$\begin{cases} \det D^2 u(x) = f(x) \quad x \in \Omega, \\ u(x) = 0 \quad x \in \partial\Omega, \end{cases} \tag{8.7.6}$$

where $\Omega := \{x = (x_1, x_2) \in \mathbb{R}^2 : |x| < 1\}$, and f is chosen so that the true solution of (8.7.1) is given by

$$u(x, y) = 1 - \sqrt{2 - x_1^2 - x_2^2}.$$

We apply the semismooth Newton's method given by Algorithm 1 from Chapter 7 (we will call this method A), and the Newton's method given by (8.4.3) (we will call this method B), comparing various properties of the approximations. In both cases, the L^2 increment tolerance was set to 10^{-11}. Furthermore, the computational parameters (initial condition, jump stabilisation parameters, ξ, and $\mathcal{H}_F$) for method A are the same as in Experiment 7.9.1, and the initial guess for method B is given by (8.7.2).

For this discussion, let us denote by $u_{h,A}$ the numerical solution of method A, and $u_{h,B}$ the numerical solution of method B. To assess the error, we use H^2-style quantities that arise naturally in the two different methods. In particular, for method A we calculate $\|u - u_{h,A}\|_{h,1}$ and for method B we calculate $\|D^2 u - \mathbf{H}_h u_{h,B}\|_{L^2(\Omega)}$, we plot these values against the mesh size for $p = 2, 3, 4$ in Figure 8.13, and provide the actual values in Tables 8.10 and 8.11 along with the EOCs in brackets. We observe that the convergence rates for both method are comparable, with method A outperforming method B for $p = 4$, and method B outperforming method A for $p = 2, 3$.

However, we observe a noticeable difference in the number of Newton steps, runtimes and number of degrees of freedom (NDoFs). The convergence of Newtons

method is plotted in Figures 8.14, 8.15, and 8.16, we see that the number of Newton iterations for method A varies between 5 and 6 in all cases, whereas it takes 9 iterations for method B to convergence in each case. Furthermore, the number of DoFs and runtimes at each mesh refinement are provided in Tables 8.12 and 8.13, we observe that method B has roughly double the degrees of freedom of method A, and also takes roughly twice as long to implement (i.e., the runtimes are roughly double those of method A).

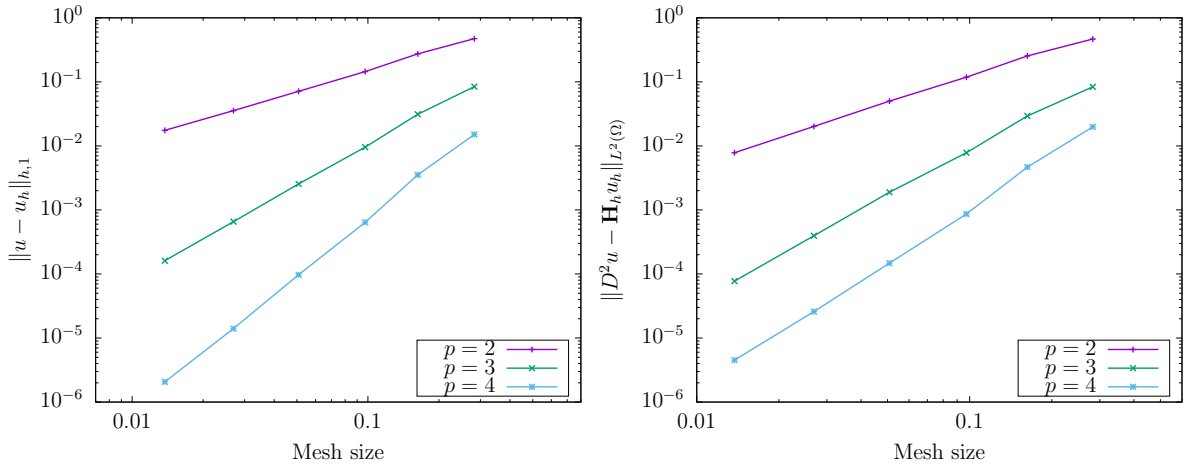


Figure 8.13: Plot of convergence rates $\|u - u_{h,A}\|_{h,1}$ (Left) and $\|D^2u - \mathbf{H}_h u_{h,B}\|_{L^2(\Omega)}$ (right), where $u_{h,A}$ is the numerical solution of method A, $u_{h,B}$ is the numerical solution of method A, and u is the true solution of (8.7.6). We observe the optimal rate of convergence in both cases, that is, the convergence rate is of order h^{p-1}.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	4.73×10^{-1}		8.44×10^{-2}		1.51×10^{-2}	
0.1627	2.73×10^{-1}	(0.99)	3.12×10^{-2}	(1.80)	3.53×10^{-3}	(2.63)
0.0973	1.45×10^{-1}	(1.24)	9.55×10^{-3}	(2.30)	6.40×10^{-4}	(3.32)
0.0508	7.11×10^{-2}	(1.09)	2.55×10^{-3}	(2.03)	9.71×10^{-5}	(2.90)
0.0269	3.55×10^{-2}	(1.09)	6.55×10^{-4}	(2.13)	1.41×10^{-5}	(3.04)
0.0138	1.75×10^{-2}	(1.06)	1.61×10^{-4}	(2.09)	2.07×10^{-6}	(2.86)

Table 8.10: The error values $\|u - u_{h,A}\|_{h,1}$ and EOCs for the semismooth Newton's method given by Algorithm 1 applied to problem (8.7.6).

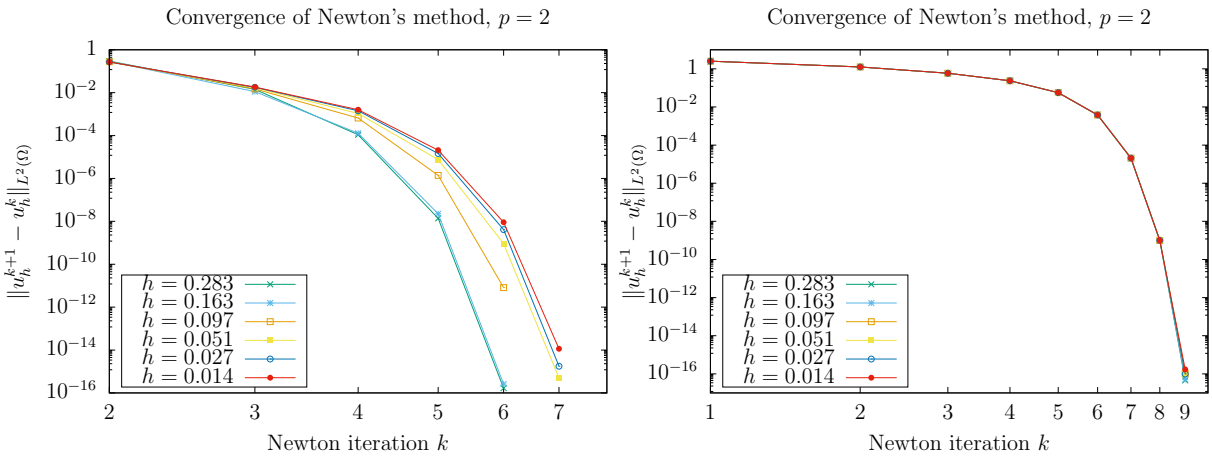


Figure 8.14: Convergence of the semismooth Newton's method given by Algorithm 1 (left) and of the Newton's method given by (8.4.3) applied to problem (8.7.6), with $p = 2$.

Mesh size	$p = 2$		$p = 3$		$p = 4$	
0.2828	4.67×10^{-1}		8.37×10^{-2}		1.98×10^{-2}	
0.1627	2.54×10^{-1}	(1.10)	2.93×10^{-2}	(1.90)	4.66×10^{-3}	(2.62)
0.0973	1.18×10^{-1}	(1.49)	7.82×10^{-3}	(2.57)	8.63×10^{-4}	(3.28)
0.0508	5.01×10^{-2}	(1.32)	1.89×10^{-3}	(2.18)	1.48×10^{-4}	(2.72)
0.0269	2.02×10^{-2}	(1.43)	3.96×10^{-4}	(2.46)	2.59×10^{-5}	(2.74)
0.0138	7.82×10^{-3}	(1.41)	7.73×10^{-5}	(2.44)	4.52×10^{-6}	(2.60)

Table 8.11: The error values $\|D^2u - \mathbf{H}_h u_{h,B}\|_{L^2(\Omega)}$ and EOCs for the Newton's method given by (8.4.3) applied to problem (8.7.6).

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.2828	2.06	2.21	2.79	384	640	960
0.1627	1.85	2.32	2.64	1044	1740	2610
0.0973	2.49	3.16	4.50	3420	5700	8550
0.0508	4.39	10.42	23.28	13920	23200	34800
0.0269	16.23	55.93	149.51	52476	87460	131190
0.0138	112.31	398.01	1199.67	205848	343080	514620

Table 8.12: Runtimes and number of DoFs for for the semismooth Newton's method given by Algorithm 1 applied to problem (8.7.6).

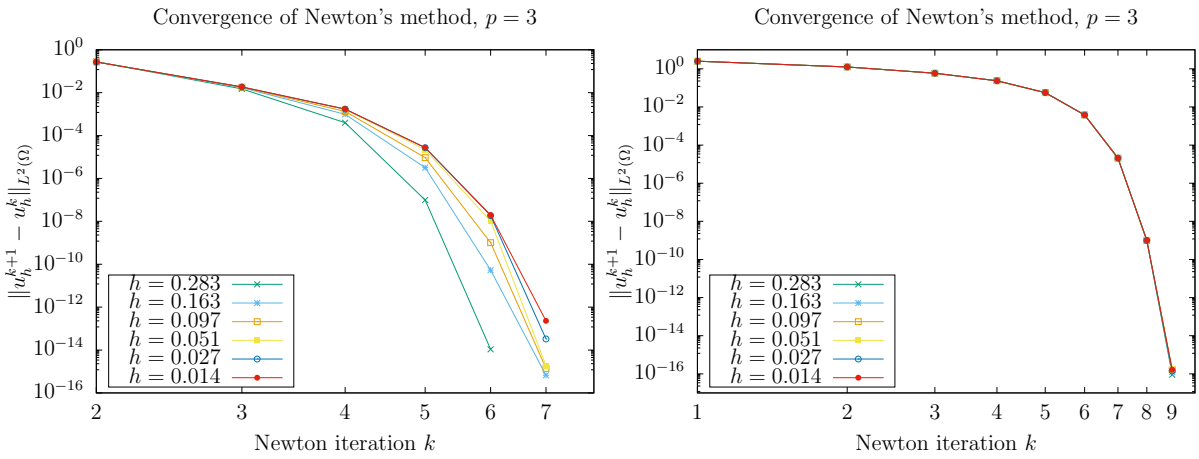


Figure 8.15: Convergence of the semismooth Newton's method given by Algorithm 1 (left) and of the Newton's method given by (8.4.3) applied to problem (8.7.6), with $p = 3$.

	Runtime (seconds)			Number of DoFs		
Mesh size	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
0.2828	13.11	17.65	1.85	580	1252	2180
0.1627	1.35	2.34	3.90	1508	3304	5796
0.0973	3.99	7.28	12.75	4772	10576	18660
0.0508	17.50	32.98	61.73	18980	42388	75076
0.0269	75.81	171.67	354.12	70788	158656	281508
0.0138	396.48	993.69	2180.71	276084	619972	1101092

Table 8.13: Runtimes and number of DoFs for the Newton's method given by (8.4.3) applied to problem (8.7.6).

8.8 Concluding remarks

The nonvariational finite element method proposed in [79] is applicable to MA problem of the form (3.2.1), with $f = f(x)$, and in [49] a proof of existence of a numerical solution was provided the case that $d = 2$ and the polynomial degree $p \geq 3$. We have extended these results, allowing for right-hand side nonlinearities of the form $f(x, z, q) = f_1(x, z)/f_2(x, q)$. We have also defined a further extension of this method, defined by (8.2.1), which encompasses a wider class of nonlinear elliptic equations, with Experiment 8.7.3 demonstrating its applicability to the Weingarten equation. The proof of existence of a numerical solution used was a fixed point argument,

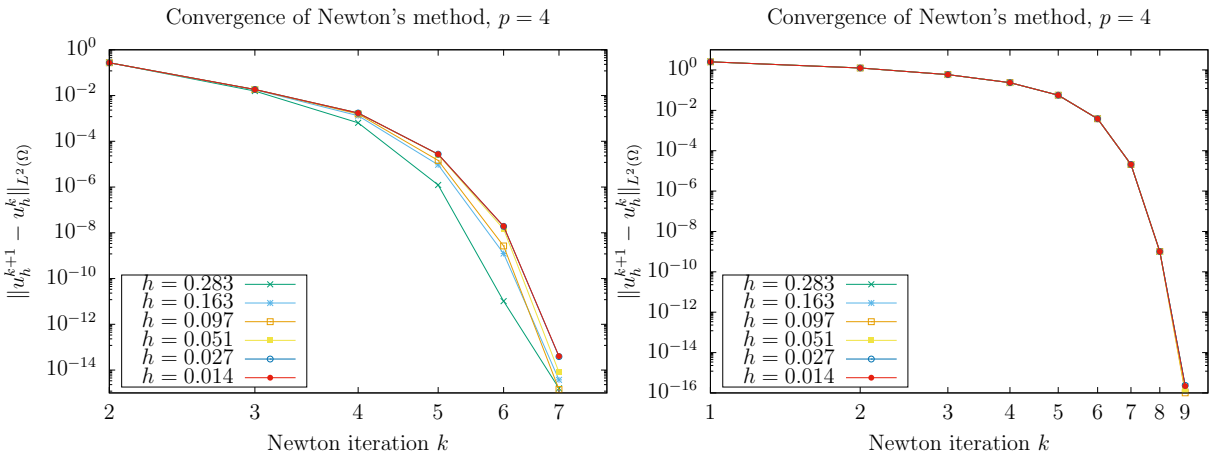


Figure 8.16: Convergence of the semismooth Newton's method given by Algorithm 1 (left) and of the Newton's method given by (8.4.3) applied to problem (8.7.6), with $p = 4$.

which also provides a priori error estimates in a H^1-type norm, however, though the convergence rate is optimal with respect to the mesh size, the uniqueness of the numerical solution, guaranteed by Theorem 8.2.3, holds in a ball of radius $h^{2+\alpha}$ for some $\alpha > 0$, and thus only the numerical experiments of Section 8.7 indicate that this method does not exhibit nonuniqueness on the same scale as discussed in Section 3.5.

In the case that the right-hand side function of (3.2.1) is of the form $f = f(x)$, the method (8.2.1) proposed in this Chapter is comparable with the DGFEM of Chapter 7, given by (7.3.2). In Experiment 8.7.4 we applied both methods to the same test problem, with the goal of determining which method may have more desirable attributes. The conclusion of this experiment is that the methods are similar in term of convergence rates, with the method proposed in this Chapter outperforming the method of Chapter 7 for polynomial degrees $p = 2, 3$, and vice versa for $p = 4$, in terms of convergence rates. However, when considering the computational size of two approaches, and the resulting computation times, the method of Chapter 7 turns out to be more efficient, with roughly half the degrees of freedom and computation times.

Chapter 9

A CGFEM for the MA optimal transport problem

9.1 New contributions and existing methods

In this chapter we present the contributions of the collaborative work [72]. This includes an adaptation of the nonvariational finite element method (NVFEM) proposed in [78, 79], providing a new finite element method for the approximation of solutions to the MA optimal transport problem. We also implement a global gradient recovery scheme, resulting in optimal convergence of the scheme for $\mathbb{P}^1$ finite elements.

Existing methods:

- In [90] the authors propose a finite element element method, also based on the formulation [79], solving the MA optimal transport on the surface of a sphere, and planar domains without boundary. A key difference between the method of [90], and the method that we propose in this chapter is that we deal directly with the optimal transport boundary condition (for further details see Section 9.5.3).

9.2 Set-up

In this chapter, we design a CG finite element method for the approximation of solutions to the MAOT problem (3.2.2), with right-hand side function $f(x, z, q) = f_1(x)/f_2(q)$, that is $u : \Omega \to \mathbb{R}$ satisfies

$$\det D^2 u(x) = \frac{f_1(x)}{f_2(\nabla u(x))}, \quad x \in \Omega, \tag{9.2.1}$$

along with the *second* boundary condition:

$$\nabla u(\Omega) = \Upsilon, \tag{9.2.2}$$

where $\Omega, \Upsilon \subset \mathbb{R}^d$ are two uniformly convex, $C^{2,1}$ domains, and $f_1 : \Omega \to \mathbb{R}^+$, $f_2 : \Upsilon \to \mathbb{R}^+$ are two uniformly positive, uniformly $C^{1,1}$ functions satisfying the preservation of mass condition:

$$\int_\Omega f_1 = \int_\Upsilon f_2. \tag{9.2.3}$$

Recall that under these assumptions, Theorem 3.2.4 guarantees the existence of a solution $u \in C^{3,\alpha}(\Omega) \cap C^{2,\alpha}(\overline{\Omega})$, for any $\alpha \in (0,1)$ of (9.2.1)–(9.2.2), that is unique up to a constant, among convex solutions.

9.3 Linear nonvariational oblique derivative problem

Let Ω be a uniformly convex $C^{2,1}$ domain, and recall that we denote its unit outer normal by $n_{\partial\Omega}$. Assume that $A \in L^\infty(\Omega; \mathbb{R}^{d\times d}_{\text{sym}})$, is uniformly elliptic, i.e., there exist constants $0 < \mu_1 \leq \mu_2 < \infty$ such that

$$\mu_1|\xi|^2 \leq \xi^T A(x)\xi \leq \mu_2|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega, \tag{9.3.1}$$

and let the oblique vector $\beta \in L^\infty(\partial\Omega; \mathbb{S}^{d-1})$. We also assume that $a \in L^\infty(\Omega; \mathbb{R}^d)$, $r \in L^2(\Omega)$, and $s \in H^{1/2}(\partial\Omega)$, and consider the following oblique boundary-value problem: find $u : \Omega \to \mathbb{R}$ that satisfies

$$\begin{cases} A : D^2u + a \cdot \nabla u = r & \text{in } \Omega, \\ \beta \cdot \nabla u = s & \text{on } \partial\Omega. \end{cases} \tag{9.3.2}$$

Furthermore, if $d \geq 3$, we assume that there exists a constant $\delta > 0$ such that

$$\beta \cdot n_{\partial\Omega} \geq \delta \quad \text{a.e. on } \partial\Omega. \tag{9.3.3}$$

Remark 9.3.1 *The problem given above is an oblique derivative problem, though one can see that it is posed in contrast to the oblique boundary-value problem (6.2.1), in that we have not allowed for a free constant in the boundary condition, the boundary condition is inhomogeneous (if $s \neq 0$), and for $d \geq 3$, we have imposed the assumption (9.3.3) upon the oblique vector. One can also see that first order derivatives appear in the PDE. These structural choices are particular to the MAOT problem, for the following reasons:*

1. *The functions f_1 and f_2 defining the right-hand side of the MAOT problem (9.2.1) are assumed to satisfy the compatibility condition (3.2.7), negating the necessity for free constant in the boundary condition;*

2. *The PDE (9.3.2) represents the general linear elliptic equation arising when applying Newton's method to the nonlinear problem (9.2.1), and thus first order derivatives appear in the PDE.*

9.4 Nonvariational finite element method (NVFEM) for the oblique derivative problem

Throughout this chapter, we do not employ curved finite elements (as in Chapters 5 to 8). To this end, let $(\mathscr{T}_h)_{h>0}$ be a shape-regular quasi-uniform family of affine triangulations that approximate Ω. Indeed, since Ω is assumed to be uniformly convex, such triangulations can be defined in the following manner: for each $h > 0$ by taking a collection of $N_h \in \mathbb{N}$ points on $\partial\Omega$, and take their closed convex hull, the interior of which we shall denote by Ω_h (note that as Ω is convex, $\Omega_h \subset \Omega$). We then simply define each triangulation $\mathscr{T}_h$, $h > 0$, to be a shape-regular quasi-uniform triangulation of Ω_h. We then define the following finite element spaces:

$$\mathbb{V}_{h,p} := \{v \in C^0(\overline{\Omega}) : v|_K \in \mathbb{P}^p(K)\, \forall K \in \mathscr{T}_h\}, \tag{9.4.1}$$

$$\mathring{\mathbb{V}}_{h,p} := \mathbb{V}_{h,p} \cap H_0^1(\Omega), \tag{9.4.2}$$

$$\mathbb{Z}_{h,p} := [\mathbb{V}_{h,p}]^d, \tag{9.4.3}$$

$$\mathbb{W}_{h,p} := [\mathbb{V}_{h,p}]^{d(d+1)/2}. \tag{9.4.4}$$

With these definitions in place it is possible to design a continuous Galerkin finite element method for the approximation of solutions of problem (9.3.2) as follows: find $(U, H, c) \in \mathbb{V}_{h,p} \times \mathbb{W}_{h,p} \times \mathbb{R}$ such that

$$\begin{aligned}
&\langle H, \Phi\rangle_{\Omega_h} + \langle \nabla U (\nabla \Phi)^T\rangle_{\Omega_h} - \langle \nabla U n_{\partial\Omega_h}^T \Phi\rangle_{\partial\Omega_h} = 0,\\
&\langle A : H + a \cdot \nabla U, \Phi\rangle_{\Omega_h} + \langle \beta \cdot \nabla U, \Phi\rangle_{\partial\Omega_h} + \langle U, \lambda\rangle_{\Omega_h} + \langle c, \Phi\rangle_{\Omega_h} = \langle r, \Phi\rangle_{\Omega_h} + \langle s, \Phi\rangle_{\partial\Omega_h}
\end{aligned} \tag{9.4.5}$$

for all $\Phi \in \mathbb{V}_{h,p}, \lambda \in \mathbb{R}$.

The nil sum constraint on u, the exact solution of (9.3.2), needed to ensure its uniqueness is discretised by seeking an additional unknown scalar (instead of directly

including this condition in the finite element space) c as a Lagrange multiplier, implemented by the inclusion of the following sum

$$\langle U, \lambda \rangle_{\Omega_h} + \langle c, \Phi \rangle_{\Omega_h} = 0 \tag{9.4.6}$$

in (9.4.5). Setting $\Phi = 0$ in (9.4.5) gives us

$$\int_{\Omega_h} U = 0. \tag{9.4.7}$$

Then, upon choosing $\Phi \in \mathbb{V}_{h,p} \cap H_0^1(\Omega_h)$, we obtain

$$\langle c, \Phi \rangle_{\Omega_h} = \langle r - A : H - a \cdot \nabla U - cU, \Phi \rangle_{\Omega_h} \tag{9.4.8}$$

for all $\Phi \in \mathbb{V}_{h,p} \cap H_0^1(\Omega_h)$, which tells us that c is in fact the $L^2(\Omega_h)$-projection of

$$r - A : H - a \cdot \nabla U - cU \tag{9.4.9}$$

onto $\Phi \in \mathbb{V}_{h,p} \cap H_0^1(\Omega_h)$. Since c is a constant, and the only constant in $\Phi \in \mathbb{V}_{h,p} \cap H_0^1(\Omega_h)$ is zero, we deduce that both integrals must be zero.

Note that the upper equation in (9.4.5) is equivalent to a system of d^2 equations, which, thanks to the symmetry of the finite element Hessian, can be reduced to $d(d+1)/2$ equations; it is equivalent to

$$H = \mathbf{H}_h U. \tag{9.4.10}$$

The NVFEM, whose details for the Dirichlet boundary-value problem are given by [78], can be viewed as a mixed method, where we compute both the numerical solution U and its finite element Hessian $H = \mathbf{H}_h U$, as an auxiliary variable. We stress, however, that the variable H becomes essential in nonlinear problems where the nonlinearity depends on the Hessian. In fact, not only is accessing the finite element Hessian necessary for the internal NVFEM algorithm, but as we see in Section 9.5, it plays a crucial role in the nonlinear solver.

9.5 A Newton's method for the Monge–Ampère optimal transport problem

In order to approximate u satisfying the nonlinear problem (9.2.1), we first apply Newton's method to the nonlinear problem, resulting in a sequence of problems in the form of (9.3.2) with u replaced by u_n. As discovered by Loeper & Rapetti [85],

Newton's method (applied to the MA problem, with periodic boundary conditions), with a damped stepsize converges to the exact solution at the continuous level. The main difficulty is to show that the convexity of the Newton iterate, u_n, is preserved with respect to n. The preservation of convexity leads to a sequence of well-posed elliptic problems. As in the experiments of Chapter 8, we observe that the NVFEM also inherits this property. The reader must note, however, that a proof of the fact that Newton's method preserves convexity on the numerical level for the methods we propose is currently an open problem, and that it is currently only indicated by numerical evidence. We now recap the results of [79], which apply to elliptic Dirichlet boundary-value problems, and then adapt them to problem (9.2.1).

9.5.1 Elliptic operators

The method proposed in [79] is applicable to nonlinear *elliptic* problems. As such, it is pertinent at this point to provide a definition of *ellipticity*, in the sense of nonlinear operators (note that the operators associated to the equation of prescribed Gaussian curvature (8.7.3), and prescribed Weingarten curvature (8.7.5), are examples of smooth elliptic operators). Consider a general (possibly nonlinear) operator of the form

$$v \mapsto \mathscr{F}[v], \tag{9.5.1}$$

where

$$\mathscr{F}[v(x)] := F(x, v(x), \nabla v(x), D^2 v(x)),$$

which is well defined for functions $v \in C^2(\Omega)$, for some given (possibly nonlinear) function

$$F : \Omega \times \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}. \tag{9.5.2}$$

Following [28], for an open set $\mathscr{C} \subset \mathbb{R}^{d\times d}_{\text{sym}}$, the operator $\mathscr{F}[\cdot]$ is called *elliptic* on $\mathscr{C}$ if and only if for each $(x, M) \in \Omega \times \mathscr{C}$ there exist $\lambda_\flat(x, M) \leq \lambda_\sharp(x, M)$ in $\mathbb{R}^+$, such that

$$\lambda_\flat(x, M)|N| \leq F(x, M + N) - F(x, M) \leq \lambda_\sharp(x, M)|N| \quad \forall N \in \mathbb{R}^{d\times d}_{\text{sym}}, \tag{9.5.3}$$

where the matrix norm $|M|$ indicates the Euclidean-induced operator norm (although the definition is independent of the choice of norm except for the values of $\lambda_\flat$ and $\lambda_\sharp$).

If the largest possible set $\mathscr{C}$ for which (9.5.3) is satisfied is a proper subset of $\mathbb{R}^{d\times d}_{\text{sym}}$ we say that the operator $\mathscr{F}$ is *conditionally elliptic*. The operator $\mathscr{F}[\cdot]$ is called *uniformly elliptic on* $\mathscr{C} \subseteq \mathbb{R}^{d\times d}_{\text{sym}}$ if and only if

$$0 < \inf_{\Omega\times\mathbb{R}^{1+d}\times\mathscr{C}} \lambda_\flat, \text{ and } \sup_{\Omega\times\mathbb{R}^{1+d}\times\mathscr{C}} \lambda_\sharp < \infty; \tag{9.5.4}$$

the extrema defined by (9.5.4) are called *lower* and *upper uniform ellipticity constants*. If the infimum in (9.5.4) is zero the operator is called *degenerate elliptic on* $\mathscr{C}$.

9.5.2 Smooth elliptic operators

If F is differentiable (9.5.4) can be obtained from properties of the derivative of F. A generic $M \in \mathbb{R}^{d\times d}$ being written as

$$M = \begin{bmatrix} m_{1,1} & \dots & m_{1,d} \\ \vdots & \ddots & \vdots \\ m_{d,1} & \dots & m_{d,d} \end{bmatrix}, \tag{9.5.5}$$

the derivative of F at M in the direction N is represented by $\nabla_M F(x, M)$, with respect to the Frobenius inner product. Namely,

$$D_M F(x, M)N =: \nabla_M F(x, M) : N \quad \forall N \in \mathbb{R}^{d\times d} \tag{9.5.6}$$

for some matrix $\nabla_M F(x, M)$, where we have

$$\nabla_M F(\cdot, M) = \begin{bmatrix} \partial_{m_{1,1}} F(\cdot, M) & \dots & \partial_{m_{1,d}} F(\cdot, M) \\ \vdots & \ddots & \vdots \\ \partial_{m_{d,1}} F(\cdot, M) & \dots & \partial_{m_{d,d}} F(\cdot, M) \end{bmatrix}. \tag{9.5.7}$$

Usually, the function F (and its gradient) are restricted to the linear subspace $\mathbb{R}^{d\times d}_{\text{sym}} \subset \mathbb{R}^{d\times d}$ in the 4th argument. Therefore, if F is differentiable then (9.5.3) is satisfied for all $M \in \mathscr{C}$ if and only if for each $M \in \mathscr{C}$ the matrix $\nabla_M F(\cdot, M)$ is (symmetric) positive definite, i.e.,

$$\xi^T \nabla_M F(x, M)\xi \geq \lambda_\flat(x, M)|\xi|^2 \quad \forall \xi \in \mathbb{R}^d. \tag{9.5.8}$$

Furthermore $\mathscr{C} = \mathbb{R}^{d\times d}_{\text{sym}}$ and $\lambda_\flat$ is independent of M if and only if the infimum condition in (9.5.4) is satisfied.

Lemma 9.5.1 (ellipticity of the Monge–Ampère operator)
The Monge–Ampère operator[1]

$$\mathscr{F}[v] := F(x, \nabla v, D^2 v) \text{ with } F(x, p, M) := \det M - \frac{f_1(x)}{f_2(p)} \tag{9.5.9}$$

and f_1, f_2 as described in Subsection 3, is degenerate conditionally elliptic for M in the cone $\mathrm{SPD}(\mathbb{R}^d)$ *of symmetric positive definite linear transformations on* $\mathbb{R}^d$.

[1] Since the function F generating the Monge–Ampère operator $\mathscr{F}$ does not depend on the values of the second variable representing the values of the operand (v or r) we remove the dependence on these variables.

Proof: From the definitions in 9.5.2, we need to show that $v \mapsto \det D^2 v$ is elliptic. Recall the definition of the cofactor matrix, or tensor, of an invertible M:

$$\operatorname{Cof} M := \det(M)(M^{-1})^T \tag{9.5.10}$$

this definition can be extended by uniform continuity to singular matrices. By the definition of matrix invariants [13] we have, for each $M, N \in \mathbb{R}^{d\times d}$ and $\theta \in \mathbb{R}$,

$$\det(M + \theta N) = \det M + \operatorname{Cof} M : N\theta + \rho(\theta) \tag{9.5.11}$$

for a remainder function ρ satisfying

$$|\rho(\theta)| \leq C|M|^d|N|^d\theta^2 \quad \forall \theta \in [0, 1) \tag{9.5.12}$$

for some C, from which we derive Jacobi's formula

$$D\det(M)N = \operatorname{Tr}\operatorname{Cof}(M)N = \operatorname{Cof}(M) : N \quad \forall M, N \in \mathbb{R}^{d\times d}. \tag{9.5.13}$$

Thus, the gradient of F with respect to the Frobenius inner product of matrices is

$$\nabla_M F(x, M) = \operatorname{Cof} M \quad \forall M \in \mathbb{R}^{d\times d}. \tag{9.5.14}$$

This remains true when we restrict F to matrices M (and variations thereof N) in $\mathbb{R}^{d\times d}_{\text{sym}}$, or more specifically $\operatorname{SPD}(\mathbb{R}^d)$. Indeed, if $M \in \operatorname{SPD}(\mathbb{R}^d)$ then it is invertible, furthermore $M^{-1} \in \operatorname{SPD}(\mathbb{R}^d)$, and $\operatorname{Cof} M = \det(M)M^{-1} \in \operatorname{SPD}(\mathbb{R}^d)$. This holds because the eigenvalues of M^{-1} are the reciprocals of the eigenvalues of M, and since M is positive definite, all of its eigenvalues must be strictly positive. Thus for all $\xi \in \mathbb{R}^d$ we have that

$$\begin{aligned} \xi^T \nabla_M F(x, M)\xi &= \det(M)\xi^T M^{-1}\xi \\ &\geq \frac{|\xi|^2 \det M}{\lambda_\sharp}, \end{aligned} \tag{9.5.15}$$

where $\lambda_\sharp$ is the largest eigenvalue of M. Noting that since M is positive definite, its determinant is also strictly positive; it then follows that (9.5.3) is satisfied. Since $\operatorname{SPD}(\mathbb{R}^d)$ is a proper subset of $\mathbb{R}^{d\times d}_{\text{sym}}$ this means that $\mathscr{F}$ is only conditionally elliptic with maximal domain of ellipticity the functions whose Hessian is in $\operatorname{SPD}(\mathbb{R}^d)$, i.e., the uniformly convex functions. Finally noting that

$$\inf_{M\in\operatorname{SPD}(\mathbb{R}^d)} \lambda_\flat(M) = 0, \tag{9.5.16}$$

it follows that F is *degenerate elliptic on* $\operatorname{SPD}(\mathbb{R}^d)$. $\square$

9.5.3 Quantifying the second boundary condition

One of the computational difficulties of the MAOT equation - setting aside the nonlinearity of the PDE - is the *second* boundary condition. Since the condition $\nabla u(\Omega) = \Upsilon$ (i.e., (9.2.2) is not a boundary condition, it is useful to utilise the following equivalent representation proven in [104] (under the weaker assumption that u is uniformly convex function, and Ω and Υ, are simply connected domains) that u satisfies (9.2.1) and (9.2.2) if and only if u satisfies (9.2.1) and

$$\nabla u(\partial\Omega) = \partial\Upsilon. \tag{9.5.17}$$

However, (9.5.17) is only given implicitly, and is not compatible with computations. In order to implement this boundary condition, we consider an idea from [116], that is used to prove the existence and uniqueness (among convex solutions, and up to a constant) of a solution to (9.2.1)–(9.2.2). In particular, the author of [116] considers a uniformly concave defining function $b^{\text{conc}} : \mathbb{R}^d \to \mathbb{R}$ for the target domain, that is,

$$\Upsilon = \{q \in \mathbb{R}^d : b^{\text{conc}}(q) > 0\}.$$

Since one can see that $\partial\Upsilon = \{q \in \mathbb{R}^d : b^{\text{conc}}(q) = 0\}$, the boundary condition can be expressed as follows

$$b^{\text{conc}}(\nabla u(x)) = 0, \quad x \in \partial\Omega.$$

Alternatively, one can consider a convex defining function $b : \mathbb{R}^d \to \mathbb{R}$, for the target domain, i.e., $\Upsilon = \{q \in \mathbb{R}^d : b(q) < 0\}$. One such example is the signed distance function, i.e.,

$$b(q) := \begin{cases} -\operatorname{dist}(q, \partial\Upsilon), & q \in \Upsilon, \\ \operatorname{dist}(q, \partial\Upsilon), & q \in \mathbb{R}^d \setminus \Upsilon. \end{cases} \tag{9.5.18}$$

Thus, in order to capture the transport boundary condition (9.2.2), we introduce the nonlinear operator

$$\mathscr{B}[u] := b(\nabla u). \tag{9.5.19}$$

With the notation from (9.5.9) and (9.5.19), problem (9.2.1)–(9.2.2) consists of finding a function $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} \mathscr{F}[u(x)] = 0, & x \in \Omega, \\ \mathscr{B}[u(x)] = 0, & x \in \partial\Omega. \end{cases} \tag{9.5.20}$$

Remark 9.5.2 (Approach utilised in the literature) *A similar approach is utilised in [15] (where the authors provide a finite difference method for the numerical approximation of solutions to the MA optimal transport problem), to implement the boundary condition (9.2.2). In [15] the authors represent the signed distance function in terms of supporting hyperplanes. This boundary condition is then linearised by iterating over Neumann boundary conditions (see [15] for further details).*

9.5.4 Newton's method at the PDE & FEM level

By Lemma 9.5.1 the operator $\mathscr{F}[\cdot]$ is elliptic on $\mathrm{SPD}(\mathbb{R}^d)$. We introduce the cone of convex functions with zero integral on Ω

$$\mathscr{C} := \left\{ v \in C^2(\overline{\Omega}) : D^2 v(x) \in \mathrm{SPD}(\mathbb{R}^d) \quad \forall x \in \Omega \text{ and } \langle v \rangle_\Omega = 0 \right\}. \tag{9.5.21}$$

To approximate the solution of (9.5.20) we will apply the following *Newton's method:* for each $n \in \mathbb{N}_0$, assuming $u_n \in \mathscr{C}$ is given, find $u_{n+1} \in \mathscr{C}$ satisfying

$$\begin{cases} D\mathscr{F}[u_n(x)](u_{n+1}(x) - u_n(x)) + \mathscr{F}[u_n(x)] = 0, \text{ for } x \in \Omega, \\ D\mathscr{B}[u_n(x)](u_{n+1}(x) - u_n(x)) + \mathscr{B}[u_n(x)] = 0, \text{ for } x \in \partial\Omega, \end{cases} \tag{9.5.22}$$

where the $D\mathscr{F}$ and $D\mathscr{B}$ are the (infinite dimensional) directional derivatives, explicitly calculated as

$$\begin{aligned} D\mathscr{F}[v]w &:= DF(\cdot, \nabla v, D^2 v)(0, \nabla w, D^2 w) \\ &= \mathrm{Cof}(D^2 v) : D^2 w + \frac{f_1}{f_2(\nabla v)^2} Df_2(\nabla v)\nabla w, \end{aligned} \tag{9.5.23}$$

and

$$D\mathscr{B}[v]w := Db(\nabla v)\nabla w. \tag{9.5.24}$$

9.5.5 NVFEM–Newton's method

At the PDE level, for each $n \in \mathbb{N}$, finding each term u_n of the sequence defined by (9.5.22) requires one to solve an oblique derivative boundary-value problem of the form (9.3.2), with the unknown $\theta_{n+1} := u_{n+1} - u_n$, with the following substitutions for the coefficients A, a, β, and source terms r, s:

$$\begin{aligned} A(x) &\leftarrow \mathrm{Cof}\, D^2 u_n(x) && =: \hat{A}(D^2 u_n(x)), \\ a(x) &\leftarrow \frac{f_1(x)}{f_2(\nabla u_n(x))^2} \nabla f_2(\nabla u_n(x)) && =: \hat{a}(x, \nabla u_n(x)), \\ \beta(x) &\leftarrow Db(\nabla u_n(x)) && =: \hat{\beta}(\nabla u_n(x)), \\ r(x) &\leftarrow -\det D^2 u_n(x) + \frac{f_1(x)}{f_2(\nabla u_n(x))} && =: \hat{r}(x, \nabla u_n(x), D^2 u_n(x)), \\ s(x) &\leftarrow -b(\nabla u_n(x)). && =: \hat{s}(\nabla u_n(x)). \end{aligned} \tag{9.5.25}$$

Since (9.4.5) provides us with the NVFEM for the approximation of solutions to problems of this form, we may discretise (9.5.22) by substituting coefficients $\hat{A}, \hat{a}, \hat{\beta}$ and source terms $\hat{r}, \hat{s}$ present in (9.5.25) into (9.4.5).

This results in the following iterative scheme, which we call the NVFEM–Newton's method: for each $n \in \mathbb{N}_0$, assuming $(U_n, H_n) \in \mathbb{V}_{h,p} \times \mathbb{W}_{h,p}$ is given, find $(U_{n+1}, H_{n+1}, c_{n+1}) \in \mathbb{V}_{h,p} \times \mathbb{W}_{h,p} \times \mathbb{R}$ such that

$$\begin{aligned}
&\langle H_{n+1}, \Phi\rangle_{\Omega_h} + \langle \nabla U_{n+1} D\Phi\rangle_{\Omega_h} - \langle \nabla U_{n+1} n^T \Phi\rangle_{\partial\Omega_h} = 0 \quad \forall \Phi \in \mathbb{V}_{h,p}, \\
&\quad \langle \hat{A}(H_n) : (H_{n+1} - H_n) + \hat{a}(\cdot, \nabla U_n) \cdot \nabla[U_{n+1} - U_n] + \hat{r}(\cdot, \nabla U_n, H_n), \Phi\rangle_{\Omega_h} \\
&\qquad + \langle \hat{\beta}(\nabla U_n) \cdot \nabla[U_{n+1} - U_n] + \hat{s}(\nabla U_n(x)), \Phi\rangle_{\partial\Omega_h} + \langle U_{n+1}, \lambda\rangle_{\Omega_h} + \langle c_{n+1}, \Phi\rangle_{\Omega_h} \\
&\qquad\quad = 0 \quad \forall \Phi \in \mathbb{V}_{h,p}, \lambda \in \mathbb{R}.
\end{aligned} \tag{9.5.26}$$

Remark 9.5.3 (Shortcomings of the approach (9.5.26)) *Our numerical experiments show that the algorithm given by (9.5.26) produces sequences that appear to be divergent for $\mathbb{P}^1$ elements. This is to be expected, since by (4.11.3), for a function $v_h \in \mathbb{V}_{h,p}$,*

$$\mathbf{H}_h v_h = \mathcal{P}_{\mathbb{W}_{h,p}}(D_h^2 v_h) + \mathcal{L}(\nabla v_h).$$

But, one can see that for a piecewise $\mathbb{P}^1$ function, its piecewise Hessian is zero, and thus the finite element Hessian of such a function is determined solely by the lift operator, $\mathcal{L}$.

For $\mathbb{P}^p$ elements with $p \geq 2$, the algorithm converges, but, as the numerical experiments in Subsection 9.8.1 show, the convergence rates are suboptimal (in a function approximation sense) in the $L^2(\Omega_h)$ norm. For instance, for $\mathbb{P}^2$ elements, where the expected optimal convergence rate is 3, we observe a convergence rate of order 2 at best.

Remark 9.5.4 (Boundary approximation) *We believe that the suboptimal results mentioned above caused by approximating a curved convex domain by a polytopal mesh. The use of $\mathbb{P}^p$, $p \geq 2$ approximation requires the positioning of degrees of freedom on the approximating boundary that in fact lie in the interior of the true domain. This is why we observe a "cap" on our convergence rates. A solution to this problem, at least from an empirical point of view, based on extensive numerical computation is provided by the use of* gradient recovery*, in the case of $\mathbb{P}^1$ elements (we still observe suboptimal rates in the $L^2(\Omega_h)$ norm for quadratics and higher). These results further support the necessity for the curved boundary approximation, employed in Chapters 5–8.*

Definition 9.5.5 (projection-based gradient recovery) *We define the* projection-based gradient recovery operator

$$\mathbf{G}_h : \mathbb{V}_{h,p} \to \mathbb{Z}_{h,p},$$

by

$$\langle \mathbf{G}_h v - \nabla v, \Phi \rangle_{\Omega_h} = 0 \quad \forall \Phi \in \mathbb{V}_{h,p}. \tag{9.5.27}$$

9.6 FE Hessian with gradient recovery

The standard FE Hessian operator, $\mathbf{H}_h$, defined in (4.11.1) is implemented in the NVFEM–Newton's method by its inclusion in (9.5.26). Now that we are equipped with the gradient recovery operator, $\mathbf{G}_h$, given by (9.5.27), we are inclined to define a new finite element Hessian operator $\widetilde{\mathbf{H}_h}$, where one replaces ∇U in (9.5.26), with the recovered gradient $\mathbf{G}_h U$, resulting in the following definition.

Definition 9.6.1 (finite element Hessian with gradient recovery) *We first define the gradient recovered generalised Hessian $\mathscr{H}$ acting on $v \in H^1(\Omega_h)$ via*

$$\langle [\mathscr{H} v]^i_j | \varphi \rangle := -\left\langle [\mathbf{G}_h v]^i, \frac{\partial \varphi}{\partial x_j} \right\rangle_{\Omega_h} + \langle [\mathbf{G}_h v]^i n_j | \varphi \rangle_{H^{1/2}(\partial\Omega_h) \times H^{1/2}(\partial\Omega_h)'} \quad \forall \varphi \in H^1(\Omega_h), \tag{9.6.1}$$

and all $i, j = 1, \ldots, d$. Then, thanks to finite element conformity $\mathbb{V}_{h,p} \subseteq H^1(\Omega)$, we may define the finite element Hessian with gradient recovery operator $\widetilde{\mathbf{H}_h}$, acting upon $v \in H^1(\Omega)$ as follows:

$$\langle [\widetilde{\mathbf{H}_h} v]^i_j, \Phi \rangle_{\Omega_h} = \langle [\mathscr{H} v]^i_j | \Phi \rangle \quad \forall \Phi \in \mathbb{V}_{h,p},\ i, j = 1, \ldots, d. \tag{9.6.2}$$

Remark 9.6.2 (Gradient recovery for $\mathbb{P}^1$ elements) *Upon applying the gradient recovery operator $\mathbf{G}_h$, defined by (9.5.27), in algorithm (9.5.26) for $\mathbb{P}^1$ element approximation we observe that it does converge. Moreover, we observe optimal convergence results in this case (see the first experiment in Subsection 9.8.1).*

9.7 NVFEM–Newton's method with finite element gradient recovery

We incorporate the gradient recovery operator into our system, by replacing ∇U_{n+1} with $\mathbf{G}_h U_{n+1}$ in (9.5.26). This swap of roles in the discrete gradient operator, implies a

possible swap of the Hessian recovery operator $\mathbf{H}_h$ with the *modified Hessian recovery operator* $\widetilde{\mathbf{H}_h} : \mathbb{V}_{h,p} \to \mathbb{W}_{h,p}$ for any $v_h \in \mathbb{V}_{h,p}$,

$$\langle \widetilde{\mathbf{H}_h} v_h, \Phi \rangle_{\Omega_h} + \langle \mathbf{G}_h v_h \, D\Phi \rangle_{\Omega_h} - \langle \mathbf{G}_h v_h \, n^T \Phi \rangle_{\partial\Omega_h} = 0 \quad \forall \Phi \in \mathbb{V}_{h,p}. \tag{9.7.1}$$

Rewriting the Newton scheme (9.5.26) using $\widetilde{\mathbf{H}_h}$ instead of $\mathbf{H}_h$ in *incremental form* reads as follows, for each $n \in \mathbb{N}_0$,

1. given $(U_n, G_n, H_n) \in \mathbb{V}_{h,p} \times \mathbb{Z}_{h,p} \times \mathbb{W}_{h,p}$, satisfying

$$\begin{aligned} &G_n = \mathbf{G}_h U_n, \quad H_n = \widetilde{\mathbf{H}_h} U_n, \\ &U_n \text{ is uniformly finite element convex (i.e., (4.11.2)), and } \langle U_n \rangle_\Omega = 0, \end{aligned} \tag{9.7.2}$$

2. find $\Theta \in \mathbb{V}_{h,p}$ (along with its recovered gradient $\mathbf{G}_h\Theta =: G_\Theta$, its modified recovered Hessian $\widetilde{\mathbf{H}_h}\Theta =: H_\Theta$ and a scalar c) such that:

$$\begin{aligned} \langle H_\Theta, \Phi \rangle_{\Omega_h} + \langle G_\Theta \, D\Phi \rangle_{\Omega_h} - \langle G_\Theta \, n^T \Phi \rangle_{\partial\Omega_h} = 0 \quad &\forall \Phi \in \mathbb{V}_{h,p}, \\ \langle G_\Theta, \Phi \rangle_{\Omega_h} - \langle \nabla\Theta, \Phi \rangle_{\Omega_h} = 0 \quad &\forall \Phi \in \mathbb{V}_{h,p}, \\ \langle \hat{A}(H_n) : H_\Theta + \hat{b}(G_n) \cdot G_\Theta + F(\cdot, G_n, H_n), \Phi \rangle_{\Omega_h} & \\ + \langle \hat{\beta}(G_n) \cdot G_\Theta + \hat{s}(G_n), \Phi \rangle_{\partial\Omega_h} + \langle \Theta, \lambda \rangle_{\Omega_h} + \langle c, \Phi \rangle_{\Omega_h} = 0 \quad &\forall \Phi \in \mathbb{V}_{h,p}, \lambda \in \mathbb{R}, \end{aligned} \tag{9.7.3}$$

   where the functions $\hat{A}$, $\hat{b}$, $\hat{\beta}$, and $\hat{s}$ are given by (9.5.25), with G_n and H_n in place of ∇U_n and $D^2 u_n$, respectively,

3. define

$$(U_{n+1}, G_{n+1}, \widetilde{H}_{n+1}) := (\Theta, G_\Theta, \widetilde{H}_\Theta) + (U_n, G_n, \widetilde{H}_n). \tag{9.7.4}$$

9.7.1 The linear system

Each step of Newton's method involves solving a linear system (corresponding to a nonvariational linear elliptic equation with oblique boundary conditions) of the form

$$\mathbf{E} \begin{bmatrix} \theta^T & \gamma^T & \delta^T & c \end{bmatrix}^T = \mathbf{F}, \tag{9.7.5}$$

where $\mathbf{E}$ is a $(1 + (1 + d + d(d+1)/2)^2$ block-matrix in $\mathbb{R}^{(1+(1+d+d(d+1)/2)N)^2}$, and

$$\theta \in \mathbb{R}^N, \gamma \in \mathbb{R}^{dN}, \delta \in \mathbb{R}^{Nd(d+1)/2}, c \in \mathbb{R}, \mathbf{F} \in \mathbb{R}^{1+(1+d+d(d+1)/2)N}. \tag{9.7.6}$$

Recall that $N := \dim \mathbb{V}_{h,p}$.

The array $(\theta^T, \gamma^T, \delta^T, c)^T$ represents all of the finite element solution components, i.e., $(\Theta_{n+1}, G_\Theta, H_\Theta, c_{n+1})$, via their finite expansion as finite element functions.

In particular, we recall that $\mathbf{\Phi} = (\Phi_1, \ldots, \Phi_N)^T$ forms a basis of $\mathbb{V}_{h,p}$. θ defines Θ via

$$\Theta := \theta^T \mathbf{\Phi}. \tag{9.7.7}$$

Furthermore,

$$[G_\theta]_\alpha := \gamma_\alpha^T \mathbf{\Phi}, \tag{9.7.8}$$

for $\alpha = 1, \ldots, d$, corresponding to each geometric coordinate.

Penultimately, for the finite element Hessian, we have that $\delta = [\delta_{\alpha,\beta}]$ for upper-triangular indices α, β, and

$$[\widetilde{H}_\Theta]_\alpha^\beta := [\delta_{\alpha,\beta}]^T \mathbf{\Phi}, \quad \alpha = 1, \ldots, d, \ \beta = \alpha \ \ldots, d, \tag{9.7.9}$$

and, of course, the final entry, c defines c_{n+1}.

We are about to present a pseudocode for the Newton's method, but first, for the clarity of the reader, we give an example of the block matrix, $\mathbf{E}$, and vector, $\mathbf{F}$ arising on the left-hand side and right-hand side of the linear system (9.7.5), respectively, in the two-dimensional setting. Note that the block entries of $\mathbf{E}$ are given explicitly in the pseudocode below. In this case, we have that

$$(\theta, \gamma, \delta, c) = (\theta, \gamma_1, \gamma_2, \delta_{1,1}, \delta_{1,2}, \delta_{2,2}, c).$$

Furthermore,

$$\mathbf{E} = \begin{bmatrix} \mathrm{Diag}(d) & C_1 & C_2 & B_{1,1} & B_{1,2} & B_{2,2} & d \\ A_1 & M & 0 & 0 & 0 & 0 & 0 \\ A_2 & 0 & M & 0 & 0 & 0 & 0 \\ 0 & R_1 & 0 & M & 0 & 0 & 0 \\ 0 & R_2 & 0 & 0 & M & 0 & 0 \\ 0 & 0 & R_2 & 0 & 0 & M & 0 \\ d^T & 0^T & 0^T & 0^T & 0^T & 0^T & \mathrm{Tr}(\mathrm{Diag}(d)) \end{bmatrix}, \tag{9.7.10}$$

and

$$\mathbf{F} = \begin{bmatrix} -\langle F(\cdot, G_n, H_n), \mathbf{\Phi} \rangle_{\Omega_h} - \langle b(G_n), \mathbf{\Phi} \rangle_{\partial\Omega_h} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{9.7.11}$$

Algorithm 3 Newton–Raphson–NVFEM pseudocode

Require: $\Omega_h \subset \mathbb{R}^2$, $f_1 : \Omega_h \to \mathbb{R}^+$, $f_2 : \mathbb{R}^2 \to \mathbb{R}^+$, $b : \mathbb{R}^d \to \mathbb{R}$, $\mathrm{tol} \in \mathbb{R}^+$, $\mathrm{itermax} \in \mathbb{N}$, $\mathscr{T}_h$ a mesh on $\overline{\Omega_h}$, $(u_0, c_0) \in \mathbb{V}_{h,p} \cap L^2_0(\Omega_h) \times \mathbb{R}$

$r \leftarrow 1$
$n \leftarrow 0$
$G_0 \leftarrow \mathbf{G}_h U_0$
$\widetilde{H}_0 \leftarrow \widetilde{\mathbf{H}}_h U_0$
$u_0 \leftarrow (U_0, G_0, H_0, c_0)$
$\mathbf{\Phi} \leftarrow (\Phi_1, \ldots, \Phi_N)^T$ (basis of $\mathbb{V}_{h,p}$)
while $n \leq \mathrm{itermax}$ & $r > \mathrm{tol}$ **do**
  for $\alpha = 1, 2$ **do**
    $R_\alpha \leftarrow \langle \mathbf{\Phi}(\partial_\alpha \mathbf{\Phi})^T \rangle_{\Omega_h} - \langle n_\alpha \mathbf{\Phi}\mathbf{\Phi}^T \rangle_{\partial\Omega_h}$
    $A_\alpha \leftarrow \langle \mathbf{\Phi}(\partial_\alpha \mathbf{\Phi})^T \rangle_{\Omega_h}$
    $C_\alpha \leftarrow \langle \frac{f_1}{f_2(G_n)^2} [Df_2(G_n)]^\alpha \mathbf{\Phi}, \mathbf{\Phi}^T \rangle_{\Omega_h} + \langle [Db(G_n)]^\alpha \mathbf{\Phi}, \mathbf{\Phi}^T \rangle_{\partial\Omega_h}$
    for $\alpha \leq \beta \leq 2$ **do**
      $[B]^\beta_\alpha \leftarrow -\langle [\mathrm{Cof}(\widetilde{H}_n)]^\beta_\alpha \mathbf{\Phi}, \mathbf{\Phi}^T \rangle_{\Omega_h}$
    end for
  end for
  $M \leftarrow \langle \mathbf{\Phi}, \mathbf{\Phi}^T \rangle_{\Omega_h}$
  $d \leftarrow \langle \mathbf{\Phi}, 1 \rangle_{\Omega_h}$
  construct E given by (9.7.10)
  construct F given by (9.7.11)
  solve $\mathbf{E} \begin{bmatrix} \theta^T & \gamma_1^T & \gamma_2^T & \delta_{1,1}^T & \delta_{1,2}^T & \delta_{2,2}^T & c \end{bmatrix}^T = \mathbf{F}$
  $\Theta \leftarrow \theta^T \mathbf{\Phi}$
  for $\alpha = 1, 2$ **do**
    $[G_\Theta]^\alpha \leftarrow \gamma^T \mathbf{\Phi}$
    for $\alpha \leq \beta \leq 2$ **do**
      $[\widetilde{H}_\Theta]^\beta_\alpha \leftarrow \delta^T_{\alpha,\beta} \mathbf{\Phi}$
    end for
  end for
  $r \leftarrow \|\Theta\|_{L^\infty(\Omega_h)}$
  $(U_n, G_n, \widetilde{H}_n) \leftarrow (U_n, G_n, \widetilde{H}_n) + (\Theta, G_\Theta, \widetilde{H}_\Theta)$
  $n \leftarrow n + 1$
end while

9.7.2 Implementation

Software and code: The experiments in this Chapter have been implemented in version 1.6 of the FEniCS software [2, 86], which interfaces directly with PETSc [6, 7] running through a Python interface [39, 63]. Two working Firedrake scripts, MA-with-gradient-recovery.py and MA-no-gradient-recovery.py, used to generate the experiments of this Chapter are available in the Github repository:
https://github.com/ekawecki/Monge–Ampere.

9.8 Experiments

In this subsection we display our experiments for the MAOT problem. In each case we performed several benchmark approximations; allowing us to document the experimental order of convergence. In these examples, the source domain, Ω, is given by the unit disk in $\mathbb{R}^2$, and the target domain, Υ, is either given by the unit disk or an ellipse.

With this information, we observe optimal convergence rates when implementing the $\mathbb{P}^1$ gradient recovery scheme (9.7.2) – (9.7.4), that is,

$$\|u - U_n\|_{L^2(\Omega_h)} \leq Ch_n^2 \text{ and } \|u - U_n\|_{H^1(\Omega_h)} \leq Ch_n. \tag{9.8.1}$$

Furthermore, we observe suboptimal convergence results, when implementing either (9.5.26)–(9.4.6) or (9.7.2)–(9.7.4), when the polynomial degree $p \geq 2$, i.e., we observe the following:

$$\|u - U_n\|_{L^2(\Omega_h)} \leq Ch_n^2 \quad \text{and} \quad |u - U_n|_{H^1(\Omega_h)} \leq Ch_n^2, \tag{9.8.2}$$

in contrast to the optimal rates

$$\|u - U_n\|_{L^2(\Omega_h)} \leq Ch_n^{p+1} \text{ and } \|u - U_n\|_{H^1(\Omega_h)} \leq Ch_n^p, \tag{9.8.3}$$

where the latter are the convergence results one would have expected for an optimal numerical scheme. It is our belief that the suboptimal convergence results are caused by the piecewise linear approximation of domains with curved boundaries.

Another characteristic worth mentioning is that of superconvergence [120]. When implementing (9.7.2) – (9.7.4), in all of the experiments of Subsection 9.8.1 the recovered gradient outperforms the standard gradient; in some cases we observe that the recovered gradient error is consistently close to an entire order higher than that of the standard gradient (see the $\mathbb{P}^1$ approximation of Subsection 9.8.2, for example).

9.8.1 Disk to disk experiments

Below we display the convergence results for the MAOT problem

$$\begin{cases} \det D^2 u = f(x)(1+|\nabla u|^2)^2, & \text{in } \Omega, \\ \nabla u(\Omega) = \Upsilon, \end{cases} \tag{9.8.4}$$

with $\Omega = \Upsilon = \{(x,y) \in \mathbb{R}^2 : |x|^2 < 1\}$. Here, f is chosen so that the true solution u is either given by

$$u(x,y) := \frac{1}{2}(x^2+y^2) - \frac{1}{4},$$

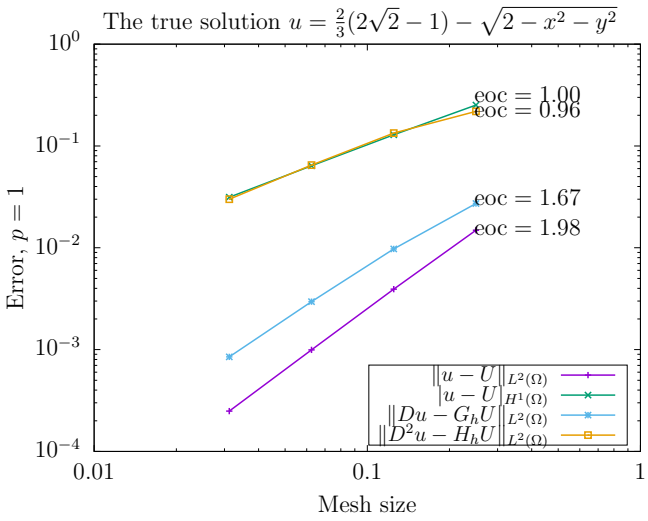
or

$$u(x,y) := \frac{2}{3}(2\sqrt{2}-1) - \sqrt{2-x^2-y^2}.$$

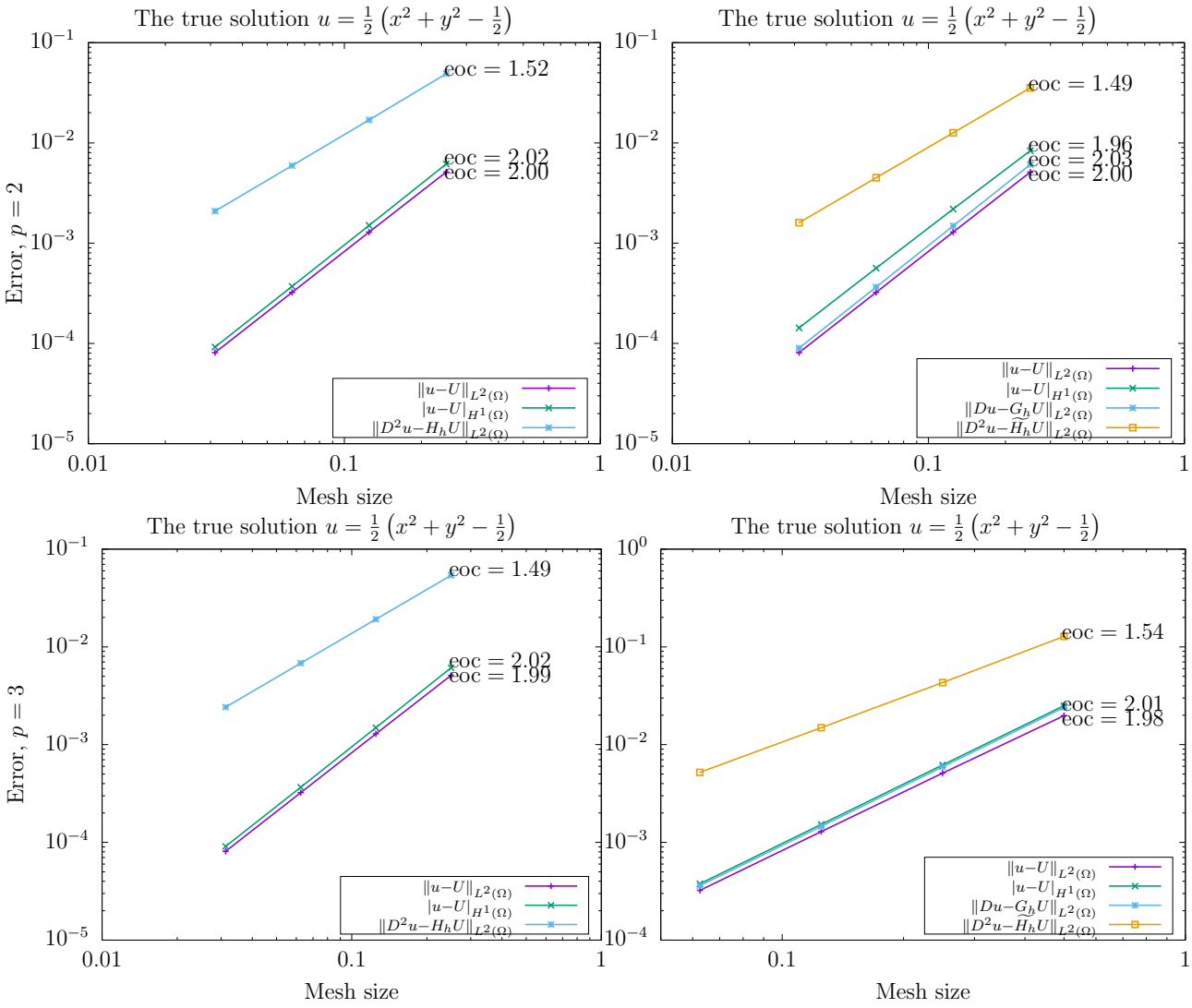
We provide the convergence results for polynomial degree, $p = 1, 2, 3$, and employ both the NVFEM–Newton's method with finite element gradient recovery (9.7.2)–(9.7.4), and the (no gradient recovery) NVFEM–Newton's method (9.5.26).

Without gradient recovery:

With gradient recovery:



The algorithm does not converge in this case



9.8.2 Disk to oval experiments

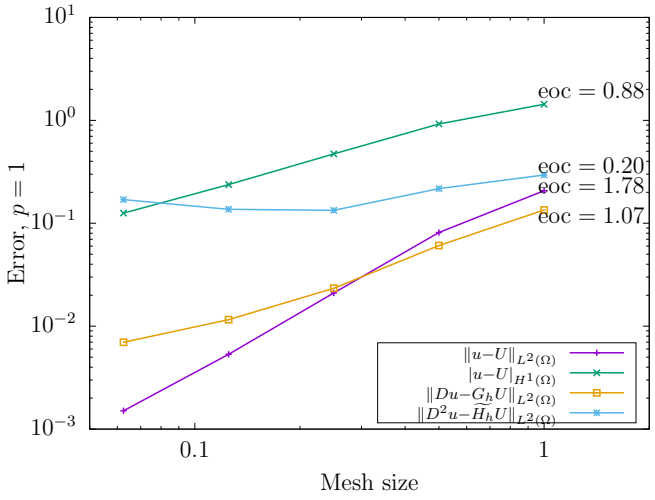
Below we display the convergence results for the MAOT problem (9.8.4), where $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$, $\Upsilon = \{\frac{x^2}{4} + \frac{y^2}{9} < 1\}$, and the function f is chosen so that the true solution, u, is given by

$$u(x, y) := x^2 + \frac{3}{2}y^2 - \frac{1}{\pi}\int_\Omega x^2 + \frac{3}{2}y^2.$$

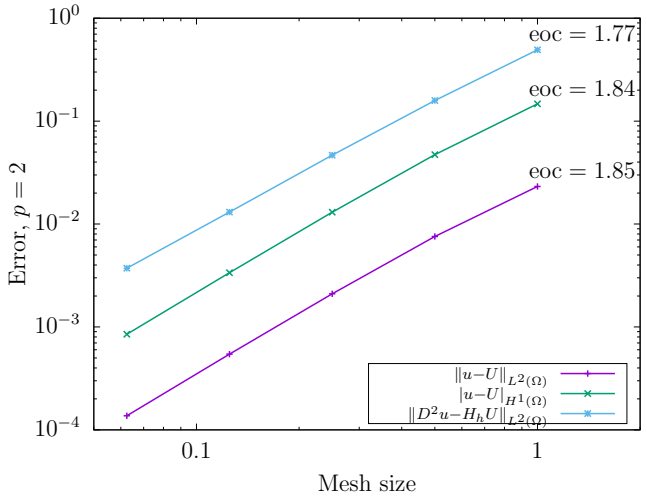
We provide the convergence results for polynomial degree, $p = 1, 2, 3$, and employ both the NVFEM–Newton's method with finite element gradient recovery (9.7.2)–(9.7.4), and the (no gradient recovery) NVFEM–Newton's method (9.5.26).

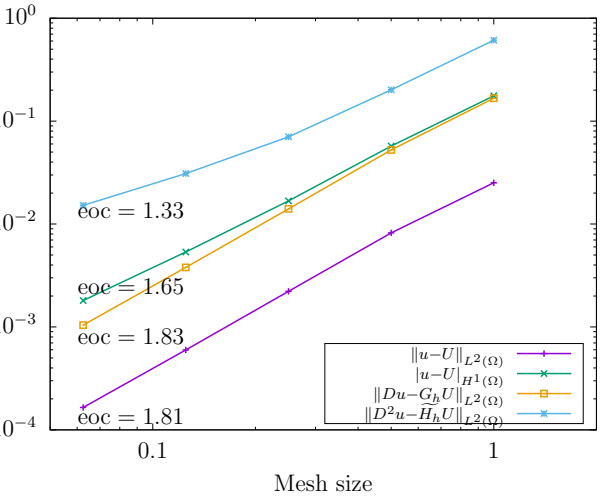
Without gradient recovery:

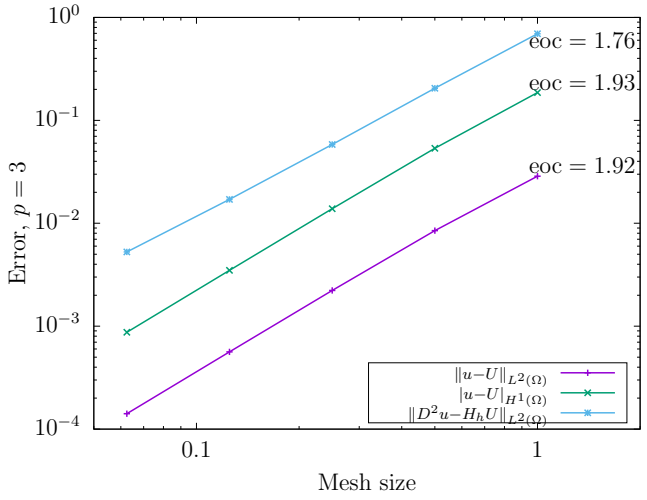
With gradient recovery:

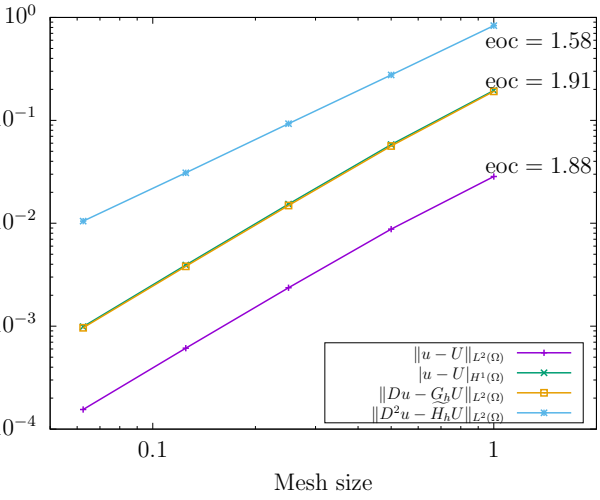


The algorithm does not converge in this case









9.8.3 Image transport experiments

The last numerical examples we present are examples of image intensity transport on one fixed uniform mesh. We transport a bitmap image of Gaspard Monge, between two geometric objects.

We model this problem as a MA optimal transport problem (9.2.1)–(9.2.2), where the the source domain, Ω, is the unit square $(-1/2, 1/2)^2$, which corresponds to the "space" that the original bitmap image of Monge occupies, and the target domain Υ is either given by Ω, or the unit disk. We obtain the density functions f_1, f_2 in (9.2.1) directly from the bitmap image of Monge. Using the functions "imread" and "im2bw" from the Image Processing Matlab toolbox [88] provides one with a matrix $M \in \mathbb{R}^{410\times 410}$, with $M_{ij} \in \{0, 1\}$, $i, j = 1, \ldots, 410$, where the value 0 corresponds to a black pixel, and the value 1 corresponds to a white pixel.

We consider the uniform grid $\{(x_k, y_\ell)\}_{k,\ell=1}^{410} = \{-0.5 + k/410, -0.5 + \ell/410)\}_{k,\ell=0}^{480}$, on the unit square $(-1/2, 1/2)^2$. Such a grid is also represented by 410^2 squares of width $1/410$, that is, 410 in the x-direction and 410 in the y-direction. For each $i, j = 1, \ldots, 410$ we define the square $S_{ij} = [x_{i-1}, x_i] \times [y_{i-1}, y_i]$, and associate with each S_{ij} the entry M_{ij} of M. This provides us with a piecewise constant map on $(-1/2, 1/2)^2$, given by $M := \sum_{i,j=1}^{410} M_{ij}\chi_{S_{ij}}$ (where $\chi_A(x) = 1$ if $x \in A$, and $\chi_A(x) = 0$ otherwise). However, this function is not uniformly positive, and so we obtain a uniformly positive density function by defining $f_1 := M + 1$. Heuristically, this means that

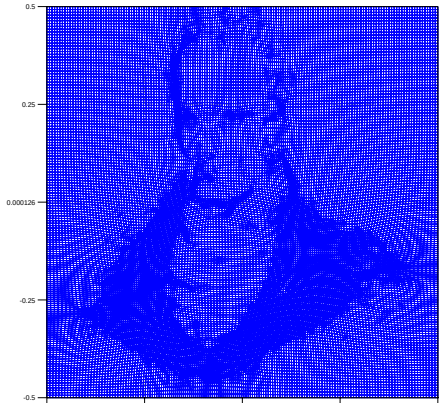
$$f_1 = \begin{cases} 2 \text{ if the pixel is white,} \\ 1 \text{ if the pixel is black.} \end{cases} \tag{9.8.5}$$

The second density function, f_2, is taken to be a constant function defined as follows:

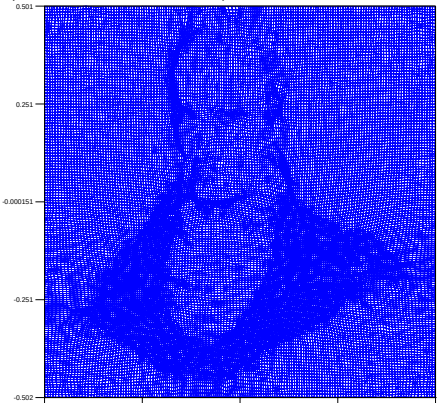
$$f_2 \equiv \frac{1}{|\Upsilon|} \int_\Omega f_1, \tag{9.8.6}$$

so that the mass is preserved, i.e., $\int_\Upsilon f_2 = \int_\Upsilon \left(\frac{1}{|\Upsilon|} \int_\Omega f_1 \right) = \int_\Omega f_1$. The resulting effect is for the white areas elements to be expanded and the black ones to be compressed. Reporting the transformation of a uniform rectangular grid (not the computational grid) under the projected gradient or the recovered gradient map renders the original bitmap using rectangles that are small in areas where the image is black and large where the image is white. The computational mesh is chosen to match the resolution of the bitmap. Although the function f_1 as defined here is discontinuous, this is not an issue as there is only one mesh and we only look at the possible use of MAOT solver as a way to encode image information in a purely discrete fashion.

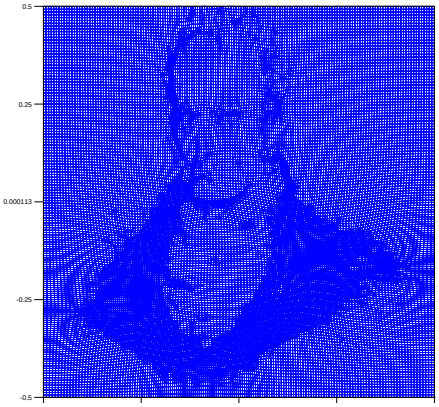
A bitmap of a portrait of Gaspard Monge, Lithography by F.S. Delpech (Public Domain)



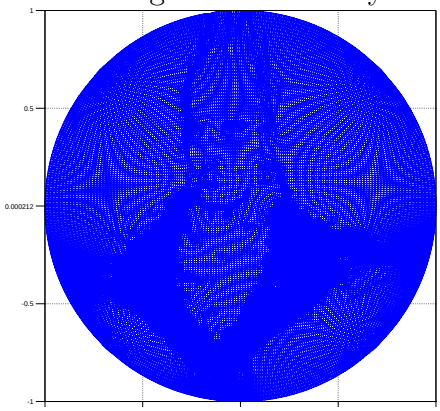
$\mathbb{P}^2$ FE without gradient recovery



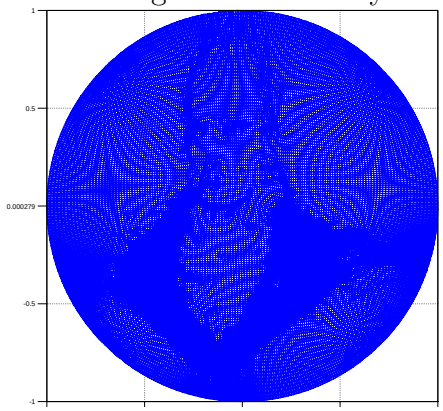
$\mathbb{P}^1$ FE with gradient recovery



$\mathbb{P}^2$ FE with gradient recovery



$\mathbb{P}^2$ FE without gradient recovery



$\mathbb{P}^2$ FE with gradient recovery

Figure 9.1: Gapard Monge's mesh-portrait obtained by mass transporting a uniform rectangular mesh into either a uniform rectangular mesh or a mesh of the unit disk.

9.9 Concluding remarks on this method

We have presented a CG finite element method for the approximation of solutions to the Monge–Ampère optimal transport problem. To our knowledge this is not present in the existing literature to date. We have demonstrated the robustness of the method, as well as its ability to capture optimal error results in the L^2-norm (in the $\mathbb{P}^1$ case), through a series of experiments. This exhibits an advance in the area of mass transportation, and methods for both linear and fully nonlinear elliptic equations with linear or nonlinear oblique boundary conditions, as well as demonstrating the applicability of variants of the nonvariational finite element method introduced in [78, 79].

In terms of future research, the formulation of this method poses the currently open question of existence and uniqueness of a solution to the numerical scheme (9.7.2) –(9.7.4), as well as the question of the derivation of a priori error estimates. In order to achieve optimal error bounds for arbitrary polynomial degree p, a potential avenue would be to incorporate the use of nonaffine approximations of the computational domain, as we have done in Chapters 5 to 8.

Conclusion

In this thesis, we have provided a general PDE theoretical framework for the existence and uniqueness of strong solutions to HJB equations (of which, linear nondivergence form second-order elliptic PDEs are a subclass), in Sobolev spaces, $H \subset H^2(\Omega)$ that satisfy the Miranda–Talenti estimates, and for which, the Laplacian is a surjection from H onto $L^2(\Omega)$. This builds upon the PDE analysis framework of [111], and, in particular, proves the well-posedness of the HJB equation with oblique boundary conditions. We have coupled this PDE framework with a new numerical framework, generalising the discontinuous Galerkin finite element method (DGFEM) of [111] to domains with curved boundaries, for the approximation of strong solutions to the HJB equation, with both Dirichlet, and oblique boundary conditions. This framework includes the existence and uniqueness of numerical solutions satisfying optimal a priori error estimates. In order to achieve these new results, we have also reviewed several key results from the theory of finite elements on polytopal domains, in the context of curved finite elements; namely, trace estimates, discrete Poincaré–Friedrichs' inequalities, a discrete Sobolev inequality, stability of the L^2-projection operator, and optimal interpolation estimates with integer and non integer Sobolev regularity. Furthermore, we have have made two new contributions in this area, one being the proof of inverse estimates in H^k and $W^{k,\infty}$ norms (and seminorms), and curvature bounds for curved simplices. The inverse estimates seem to be available to some extent in the literature, for instance in [33] for quadratic isoparametric Lagrange finite elements. However, the results on curvature bounds for curved simplices do not appear present in the current literature on finite element theory, most likely due to the fact that the polytopal counterpart of this is trivial, since polytopal domains, and their simplicial approximations are not curved. Both of these results build upon the scaling arguments present in [16], and seem to be significant in the design of consistent and stable nonconforming finite element schemes, where the use of integration by parts identities leads to curvature dependent terms and higher order derivatives arising

in the formulation, which must be appropriately bounded, requiring the use of the aforementioned simplicial curvature and inverse estimates, respectively.

We have also provided a variation of the equivalence of the MA Dirichlet problem to a uniformly elliptic HJB Dirichlet boundary-value problem, proven by Krylov in [74]. Krylov's characterisation provides one with a HJB equation, where the set of controls, X, is given by

$$X := \{W \in \mathbb{R}^{d\times d}_{\mathrm{Sym}} : W \geq 0, \operatorname{Tr} W = 1\}.$$

As mentioned in Section 3.5, we wish to consider HJB equations where the control set includes only matrices that satisfy the Cordes condition. However, since X contains matrices that are degenerate (for example $e_i e_i^T$, where $e_i \in \mathbb{R}^d$ is a canonical basis vector), which do not satisfy the Cordes condition, and so we cannot directly use Krylov's characterisation. To this end, under the assumption that the solution of the MA problem is uniformly $C^{2,\alpha}$, $\alpha \in (0,1)$, and uniformly convex, we have proven in the case $d = 2$, that the characterisation still holds with the restricted control set

$$X_\xi := \{W \in X : \det W \geq \xi\},$$

where $0 < \xi \leq 1/4$ is dependent upon the uniform C^2 norm of the true solution (note that one only requires $\xi \leq 1/4$ so that the control set X_ξ is nonempty), for which all of the elements satisfy the Cordes condition. Utilising this equivalence, we have successfully implemented our new numerical framework, providing a new DGFEM for the approximation of classical solutions to the MA Dirichlet problem.

Our efforts have also produced several advances into the study of the nonvariational finite element method, for MA type problems. We have extended the analysis present in [92], proving existence of a numerical solution to the nonvariational finite element method for the MA Dirichlet problem, allowing for more general lower order nonlinearities. In particular, the type of nonlinearities that typically arise in the MA optimal transport problem. This framework includes a priori error estimates in the H^1-seminorm, and a new estimate in a H^2-style seminorm (where the role of the Hessian of the numerical solution is interchanged with the finite element Hessian), and a proof of the convergence of Newton's method for this numerical scheme. However, though the proven convergence rates are optimal with respect to the mesh size, the uniqueness of the numerical solution, guaranteed by Theorem 8.2.3, holds in a ball of radius $h^{2+\alpha}$ for some $\alpha > 0$, and thus only the numerical experiments of Section 8.7 indicate that this method does not exhibit nonuniqueness on the same scale as discussed in Section 3.5. In particular, Newton's method is observed to converge to

either the uniformly convex or uniformly concave solution of the corresponding MA problem, depending on the choice of initial guess (recall the discussion in Section 8.6).

Furthermore, we have provided an extended NVFEM for more general nonlinear elliptic PDEs (generalising the scheme introduced in [79] for the MA equation, which considers nonlinearities of the Hessian), with Experiment 8.7.3 demonstrating its ability to optimally capture smooth solutions of the Weingarten equation, a nonlinear geometric PDE.

Finally, we have introduced a new NVFEM for the MA optimal transport problem, employing a gradient recovery scheme, based on global L^2 recovery. We have observed that this method provides an optimal piecewise-linear approximation of solutions to the MA optimal transport problem. This latter method constitutes a collaborative effort between the author of this thesis, O. Lakkis, and T. Pryer.

Overall, we have proposed four schemes for the finite element approximation of MA type problems proposed in Chapters 7, 8, and 9, and so we shall (where possible) discuss the comparative strengths and weaknesses of these methods. The two methods proposed in Chapter 9 (distinguished by including or not including the gradient recovery operator i.e., (9.4.5) or (9.7.2)–(9.7.4)) are applicable to MA optimal transport problems, and thus cannot be directly compared with the methods of Chapters 7 and 8. However, the method with gradient recovery (9.7.2)–(9.7.4) is advantageous in comparison to the method without gradient recovery (9.4.5), as it is observed to converge and optimally capture solutions using piecewise linear finite elements. As a byproduct of the gradient recovery method, one is also provided with a *continuous* approximation of the gradient, which would be desirable in applications, such as mesh adaptivity (see [26, 90]).

When approximating solutions to the MA equation, with a right-hand side function of the form $f = f(x)$, method (8.2.1) proposed in Chapter 8 is comparable with the DGFEM of Chapter 7, given by (7.3.2). In Experiment 8.7.4 we applied both methods to the same test problem, with the goal of determining which method may have more desirable attributes. The conclusion of this experiment is that the methods are similar in term of convergence rates, with the method proposed in Chapter 8 outperforming the method of Chapter 7 for polynomial degrees $p = 2, 3$, and vice versa for $p = 4$ (in terms of convergence rates). However, when considering the computational size of two approaches, and the resulting computation times, the method of Chapter 7 turns out to be more efficient, with roughly half the degrees of freedom and computation times, thus implying that the DGFEM of Chapter 7 is more computationally efficient.

Furthermore, extrapolating this evidence to higher dimensions would indicate an even greater advantage of the DGFEM of Chapter 7 in terms of computational efficiency, due to the fact that the in the DGFEM the numerical solution only has one component u_h, whereas the method of Chapter 8 has $1 + d(d+2)/2$ components, due to the symmetry of the finite element Hessian, and so the number of degrees of freedom would grow more rapidly in terms of the dimension, d.

This leads us to discuss an advantage and weakness of the DGFEM approach. In particular, as demonstrated in Experiment 7.9.5, by solving 210 different MA type problems, we show that the method is robust with respect to the choice of initial guess for the semismooth Newton's method (Algorithm 1), and upon convergence, the method always converges to the *unique numerical solution*, which is an advantage over the method of Chapter 8 (and indeed other potential numerical methods for MA type problems, where uniqueness can be an issue). In dimension $d = 2$, we require the true solution of the corresponding MA problem to be uniformly convex and uniformly $C^{2,\alpha}$. However, one may conjecture that for $d \geq 3$, generalisations of this method may impose larger constraints upon the class of solutions, if one may hope to provide a HJB equation with Cordes coefficients (in particular, a sufficient condition may be that the cofactor matrix of the true solution satisfies the Cordes condition, which for $d = 2$, only requires uniform convexity, but for $d \geq 3$, as discussed in Remark 3.3.5, the Cordes condition is a stronger requirement).

A final, further advantage of method (8.2.1) is that it is applicable to a wider range of nonlinearities (in particular, we have proven this in the case that for $d = 2$) and its formulation in higher dimensions does not appear to impose further restrictions upon the class of solutions that it may approximate (indeed it can be directly stated).

In terms of future research avenues, one goal is to extend the results present in this thesis to higher dimensions. Currently, due to the nontrivial nature of the problems under consideration, several of the results present in this thesis only apply to the two dimensional setting. There are several interesting difficulties to overcome when the dimension $d \geq 3$. In the analysis of the NVFEM for the MA Dirichlet problem, it is the nature of the determinant present in the MA equation that poses difficulties, since, in two dimensions, the determinant acts as a quadratic map on the space of matrices, which is not the case in higher dimensions.

To develop a DGFEM for the HJB equation with oblique boundary conditions (and extend the corresponding PDE theory framework) in higher dimensions, there is a very interesting open problem that will need to be addressed. One must prove

that the Miranda–Talenti estimate holds for $d \geq 3$, for functions that satisfy $\beta \cdot \nabla u$ is constant on $\partial\Omega$.

Another research goal is to design and analyse a DGFEM for the MA optimal transport problem, this will require incorporating the *nonlinear* oblique boundary condition into the numerical method that currently applies to *linear* oblique boundary-value problems.

A final avenue would be to research the extent to which the new curved finite element theory proven in this thesis may allow for the generalisation of existing interior penalty discontinuous Galerkin finite element methods, and other nonconforming methods, such as C^0 interior penalty methods [25]. Furthermore, we have proven the stability of the DGFEM proposed in Chapter 5 (indeed, this is verified by Experiment 5.9.3), with domain assumptions that are not sufficient to prove existence of a strong solution to the corresponding Dirichlet boundary-value problem. It may be the case that the proposed method could yield a generalisation for the approximation of solutions to fourth-order elliptic problems (for example, the clamped plate problem [17]).

Appendix A

Proofs for the fixed point argument of Chapter 8

We now provide the proofs of the results used in Section 8.3 to prove Theorem 8.2.3.

Lemma A.1 *The norm $\|\cdot\|_h$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$ when restricted to the finite element space $\mathring{\mathbb{V}}_{h,p}$, i.e., there exist constants C_1, C_2, independent of h, such that*

$$C_1\|v\|_{H^1(\Omega)} \leq \|v\|_h \leq C_2\|v\|_{H^1(\Omega)} \quad \forall v \in \mathring{\mathbb{V}}_{h,p}. \tag{A.1}$$

Proof: Let $v \in \mathring{\mathbb{V}}_{h,p}$. Then, since $\mathring{\mathbb{V}}_{h,p} \subset H_0^1(\Omega)$, an application of the Poincaré inequality yields $\|v\|_{H^1(\Omega)} \leq C|v|_{H^1(\Omega)}$, where the constant C is independent of v. Furthermore, by the definition of the norm $\|\cdot\|_h$, we trivially obtain $|v|_{H^1(\Omega)} \leq \|v\|_h$ for all $v \in \mathring{\mathbb{V}}_{h,p}$, and so

$$\|v\|_{H^1(\Omega)} \leq C\|v\|_h,$$

which is the first estimate of (A.1).

We also see that

$$\sum_{F\in\mathscr{E}_h} h_F\|\langle\!\langle \nabla v\rangle\!\rangle\|_{2,F}^2 \leq Ch \sum_{K\in\mathscr{T}_h} \|\nabla v\|_{L^2(\partial K)}^2. \tag{A.2}$$

Furthermore, from (4.6.2) and (4.6.27), we obtain

$$\sum_{K\in\mathscr{T}_h} \|\nabla v\|_{L^2(\partial K)}^2 \leq C(h^{-1}\|\nabla v\|_{L^2(\Omega)}^2 + h\|\nabla v\|_{H^1(\Omega;Th)}^2) \leq Ch^{-1}\|v\|_{H^1(\Omega)}^2.$$

Applying the above to (A.2) yields

$$\sum_{F\in\mathscr{E}_h} h_F\|\langle\!\langle \nabla v\rangle\!\rangle\|_{2,F}^2 \leq C\|v\|_{H^1(\Omega)}^2,$$

and thus

$$\|v\|_h^2 = |v|_{H^1(\Omega)}^2 + \sum_{F\in\mathscr{E}_h} h_F \|\langle\!\langle \nabla v \rangle\!\rangle\|_{2,F}^2 \leq C\|v\|_{H^1(\Omega)}^2.$$

Taking square roots above, we obtain the second estimate of (A.1). $\square$

The proof of the following Lemma is very similar to the proof of Lemma 4.3 in [92], however, the proof requires stability estimates for the L^2-projection operator, which we have proven in the context of curved finite elements in Chapter 4 (see Corollary 4.8.4).

Lemma A.2 *We have that*

$$h^{s-j}\|A_h\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)} + \|A - A_h\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} \lesssim h^{s-j}\|u\|_{W^{s+2,\infty}(\Omega)}, \quad 0 \leq j \leq s \leq p+1. \tag{A.3}$$

Proof: Let $j, s \in \mathbb{N}_0$ such that $0 \leq j \leq s \leq p+1$. Then, we see that

$$\begin{aligned} \|A - A_h\|_{L^\infty(\Omega)} &= \|\operatorname{Cof}(D^2u) - \mathcal{P}_{\mathbb{W}_{h,p}}(\operatorname{Cof}(D^2u))\|_{L^\infty(\Omega)} \\ &\lesssim \inf_{W\in[\mathbb{V}_{h,p}]^{2\times 2}} \|\operatorname{Cof}(D^2u) - W\|_{L^\infty(\Omega)} \\ &\lesssim h^{s-j}\|u\|_{W^{s+2,\infty}(\Omega)}, \end{aligned} \tag{A.4}$$

where the first estimate follows from (4.8.12), and the final estimate follows from (4.5.16), since $u \in C^{p+2,\alpha}(\overline{\Omega})$, $\alpha \in (0,1)$, and thus $\operatorname{Cof}(D^2u)) \in [C^{p,\alpha}(\overline{\Omega})]^{2\times 2} \subset [W^{p+1,\infty}(\Omega)]^{2\times 2}$. Now, by (4.6.27), (4.5.16), and (A.4)

$$\begin{aligned} &h^{s-j}\|A_h\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)} + \|A - A_h\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} \\ &\leq h^{s-j}\|A - A_h\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)} + \|A - A_h\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} + h^{s-j}\|A\|_{W^{s,\infty}(\Omega)} \\ &\leq h^{s-j}(\|A - \pi_h(A)\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)} + \|\pi_h(A) - A_h\|_{W^{s,\infty}(\Omega;\mathscr{T}_h)}) \\ &\qquad + \|\pi_h(A) - A\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} + \|\pi_h(A) - A_h\|_{W^{j,\infty}(\Omega;\mathscr{T}_h)} + h^{s-j}\|A\|_{W^{s,\infty}(\Omega)} \\ &\lesssim h^{-j}\|\pi_h(A) - A_h\|_{L^\infty(\Omega)} + h^{s-j}\|A\|_{W^{s,\infty}(\Omega)} \\ &\leq h^{-j}\|\pi_h(A) - A\|_{L^\infty(\Omega)} + h^{-j}\|A - A_h\|_{L^\infty(\Omega)} + h^{s-j}\|A\|_{W^{s,\infty}(\Omega)} \\ &\lesssim h^{s-j}\|u\|_{W^{s+2,\infty}(\Omega)}. \qquad \square \end{aligned}$$

Corollary A.3 *Let $\Omega \subset \mathbb{R}^2$ be piecewise C^2, and let $\{\mathscr{T}_h\}_{h>0}$ be a family of regular meshes on $\overline{\Omega}$. Then, for $v \in \mathring{\mathbb{V}}_{h,p}$, we have that*

$$\|v\|_{\infty,\Omega} \lesssim (1 + |\ln h|)^{1/2}\|v\|_{H^1(\Omega)} \lesssim (1 + |\ln h|)^{1/2}\|v\|_h. \tag{A.5}$$

Proof: This is direct consequence of (4.7.15) and (A.1). $\square$

Corollary A.4 *We have the following estimate for u and u_*:*

$$\|u - u_*\|_{2,\Omega} \lesssim h^p. \tag{A.6}$$

Proof: Noting that $(\pi_h(u) - u_*)|_{\partial\Omega} = \phi_h - \phi_h = 0$, the desired estimate is a direct consequence of Poincaré's inequality and (8.3.15). $\square$

Lemma A.5 *For any $v \in \mathring{\mathbb{V}}_{h,p}$, if $h \leq h_0$, with $h_0 \in (0,1)$ sufficiently small, we have that*

$$\|Af_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)} \lesssim h^{2-m}\|\nabla v\|_{2,\Omega}, \quad m = 0, 1, \tag{A.7}$$

where $C > 0$ depends on the shape-regularity of the mesh and $\|u\|_{W^{p+3,\infty}(\Omega)}$.

Proof: For the sake of presentation, we will omit the $\mathbb{W}_{h,p}$ subscript of the projection operator throughout this proof. For $m \in \{0, 1\}$, we see that

$$\begin{aligned}\|Af_2(\nabla u)v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)} \leq{}& \|(Af_2(\nabla u) - \mathcal{P}(Af_2(\nabla u)))v\|_{H^m(\Omega;\mathscr{T}_h)} \\ &+ \|\mathcal{P}(Af_2(\nabla u)))v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)}.\end{aligned}$$

Furthermore, by (4.5.15), we have that

$$\begin{aligned}\|(Af_2(\nabla u) - \mathcal{P}(Af_2(\nabla u)))v\|_{H^m(\Omega;\mathscr{T}_h)} &\leq \|Af_2(\nabla u) - \mathcal{P}(Af_2(\nabla u))\|_{W^{1,\infty}(\Omega)}\|v\|_{H^m(\Omega)} \\ &\lesssim h^{2-m}\|\nabla v\|_{L^2(\Omega)}.\end{aligned}$$

Thus,

$$\begin{aligned}&\|Af_2(\nabla u)v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)} \lesssim \\ &\qquad h^{2-m}\|\nabla v\|_{L^2(\Omega)} + \|\mathcal{P}(Af_2(\nabla u)))v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)}.\end{aligned}$$

Now, noting that (4.6.27) holds for any *continuous piecewise polynomial*, we obtain

$$\begin{aligned}&\|Af_2(\nabla u)v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{H^m(\Omega;\mathscr{T}_h)} \lesssim \\ &\qquad h^{2-m}\|\nabla v\|_{L^2(\Omega)} + h^{-m}\|\mathcal{P}(Af_2(\nabla u)))v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{L^2(\Omega)}.\end{aligned}$$

From (4.5.15) and Hölder's inequality, we obtain

$$\begin{aligned}&\|\mathcal{P}(Af_2(\nabla u))v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{L^2(\Omega)} \leq \\ &\quad \|A_h f_2(\nabla u)v - \mathcal{P}(A_h f_2(\nabla u)v)\|_{L^2(\Omega)} + \|(A_h f_2(\nabla u) - \mathcal{P}(Af_2(\nabla u)))v\|_{L^2(\Omega)} \\ &\lesssim h^{p+1}\|A_h f_2(\nabla u)v\|_{H^{p+1}(\Omega;\mathscr{T}_h)} + \|(A_h f_2(\nabla u) - \mathcal{P}(A_h f_2(\nabla u)))v\|_{L^2(\Omega)} \\ &\qquad + \|(\mathcal{P}(Af_2(\nabla u)) - \mathcal{P}(A_h f_2(\nabla u)))v\|_{L^2(\Omega)} \\ &\lesssim h^{p+1}\|A_h\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)}\|f_2(\nabla u)\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)}\|v\|_{H^{p+1,\infty}(\Omega;\mathscr{T}_h)} \\ &+ \|A_h f_2(\nabla u) - \mathcal{P}(A_h f_2(\nabla u))\|_{L^\infty(\Omega)}\|v\|_{L^2(\Omega)} + \|\mathcal{P}((A - A_h)f_2(\nabla u))\|_{L^2(\Omega)}\|v\|_{L^\infty(\Omega)}.\end{aligned} \tag{A.8}$$

Inequality (A.7) now follows from showing that the preceding estimate is bounded by $h^2\|\nabla v\|_{L^2(\Omega)}$ (up to a h-independent constant).

Since $v \in \mathring{\mathbb{V}}_{h,p}$ and $A_h \in \mathbb{W}_{h,p}$, it follows that $\|v\|_{H^{p+1}(\Omega;\mathscr{T}_h)} = \|v\|_{H^p(\Omega;\mathscr{T}_h)}$, and $\|A_h\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)} = \|A_h\|_{W^{p,\infty}(\Omega;\mathscr{T}_h)}$. Furthermore, $f_2 \in C^{p,\alpha}(\mathbb{R}^2;\mathbb{R}^+)$, and $u \in C^{p+2,\alpha}(\overline{\Omega})$, and so there exists $K \geq 0$, such that $|\nabla u| \leq K$ for all $x \in \overline{\Omega}$, thus $\|f_2(\nabla u)\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)}$ is uniformly bounded, independently of h. From this, applying (A.3) and (4.6.27) yields

$$h^{p+1}\|A_h\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)}\|f_2(\nabla u)\|_{W^{p+1,\infty}(\Omega;\mathscr{T}_h)}\|v\|_{H^{p+1,\infty}(\Omega;\mathscr{T}_h)} \lesssim$$
$$h^{p+1}\|A_h\|_{W^{p,\infty}(\Omega;\mathscr{T}_h)}\|v\|_{H^{p,\infty}(\Omega;\mathscr{T}_h)} \lesssim h^{p+1}h^{-(p-1)}\|v\|_{H^1(\Omega)} \lesssim h^2\|\nabla v\|_{L^2(\Omega)}.$$

Similarly,

$$\begin{aligned}\|A_h f_2(\nabla u) - \mathcal{P}(A_h f_2(\nabla u))\|_{L^\infty(\Omega)}\|v\|_{L^2(\Omega)} &\lesssim h^2\|A_h f_2(\nabla u)\|_{W^{3,\infty}(\Omega;\mathscr{T}_h)}\|\nabla v\|_{L^2(\Omega)}\\ &\lesssim h^2\|\nabla v\|_{L^2(\Omega)}.\end{aligned}$$

Finally, by utilising the definition of the projection operator, and applying (4.5.15) in conjunction with (A.5), we obtain

$$\begin{aligned}\|\mathcal{P}((A - A_h)f_2(\nabla u))\|_{L^2(\Omega)}\|v\|_{L^\infty(\Omega)} &\lesssim (1 + |\ln h|)^{\frac{1}{2}}h^3\|u\|_{W^{p+3}(\Omega;\mathscr{T}_h)}\|\nabla v\|_{L^2(\Omega)}\\ &\lesssim h^2\|\nabla v\|_{L^2(\Omega)},\end{aligned}$$

as long as $h \leq h_0$, with $h_0 \in (0,1)$ sufficiently small. This concludes the proof. □

Lemma A.6 *We have the following identity for the operator $L_u : V \to \mathring{V}'$. For all $w \in V$ and $v \in \mathring{V}$,*

$$\begin{aligned}\langle L_u[w], v\rangle = &\int_\Omega f_2(\nabla u)A\nabla w \cdot \nabla v - \nabla w \cdot (\nabla_h \cdot (Af_2(\nabla u)v - \mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v)))\\ &+ \frac{1}{2}\int_\Omega (D^2u + \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u)) : (D^2u - \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u))(D_q f_2(\nabla u)\cdot\nabla w)v\\ &+ \int_\Omega D_z f_1(x,u)wv - \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![(\mathcal{P}_{\mathbb{W}_{h,p}}(A_h f_2(\nabla u)v) - Af_2(\nabla u)v)\langle\!\langle\nabla w\rangle\!\rangle]\cdot n_F]\!].\end{aligned}$$

Proof: For ease of notation, we will omit the $\mathbb{W}_{h,p}$ subscript of $\mathcal{P}_{\mathbb{W}_{h,p}}$ throughout this proof. Let us first state the following algebraic identity:

$$ac - bd = \frac{1}{2}(a+b)(c-d) + \frac{1}{2}(a-b)(c+d).$$

From this, we see that for all $\Phi \in \mathbb{W}_{h,p},\ w \in \mathring{V}$ we have

$$\begin{aligned}\sum_{K\in\mathscr{T}_h}\int_{\partial K}(\Phi\nabla w)\cdot n_{\partial K} &= \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![(\Phi\nabla w)\cdot n_F]\!] \\ &= \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![(\langle\!\langle\Phi\rangle\!\rangle\nabla w)\cdot n_F]\!] + \sum_{F\in\mathscr{E}_h^{b}}\int_F [\![(\Phi\langle\!\langle\nabla w\rangle\!\rangle)\cdot n_F]\!],\end{aligned} \tag{A.9}$$

where n_F is a *fixed* choice of unit normal to $F \in \mathscr{E}_h^{i,b}$. Firstly, by (4.11.3) and (4.11.4), we see that

$$\begin{aligned}&\int_\Omega A_h : (\mathbf{H}_h w) f_2(\nabla u) v \\ &\quad= \int_\Omega \mathbf{H}_h w : \mathcal{P}(A_h f_2(\nabla u) v) \\ &\quad= \int_\Omega (D_h^2 w + \mathcal{L}(\nabla w)) : \mathcal{P}(A_h f_2(\nabla u) v) \\ &\quad= \sum_{K\in\mathscr{T}_h}\int_K D^2 w : \mathcal{P}(A_h f_2(\nabla u) v) - \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![(\langle\!\langle\mathcal{P}(A_h f_2(\nabla u) v)\rangle\!\rangle\nabla w)\cdot n_F]\!] \\ &\quad= \sum_{K\in\mathscr{T}_h}\int_K (A : D^2 w) f_2(\nabla u) v - \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![(\langle\!\langle\mathcal{P}(A_h f_2(\nabla u) v)\rangle\!\rangle\nabla w)\cdot n_F]\!] \\ &\qquad+ \sum_{K\in\mathscr{T}_h}\int_K D^2 w : (\mathcal{P}(A_h f_2(\nabla u) v) - A f_2(\nabla u) v)\end{aligned} \tag{A.10}$$

Note that $\nabla \cdot A = \nabla \cdot \mathrm{Cof}(D^2 u) = 0$ in Ω. Applying this to (A.10), integrating by parts, and recalling that $v|_{\partial\Omega} = 0$, yields

$$\begin{aligned}&\int_\Omega A_h : (\mathbf{H}_h w) f_2(\nabla u) v \\ &= \sum_{K\in\mathscr{T}_h}\int_K \nabla\cdot(A\nabla w) f_2(\nabla u) v - \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![(\langle\!\langle\mathcal{P}(A_h f_2(\nabla u) v)\rangle\!\rangle\nabla w)\cdot n_F]\!] \\ &\qquad+ \sum_{K\in\mathscr{T}_h}\int_K D^2 w : (\mathcal{P}(A_h f_2(\nabla u) v) - A f_2(\nabla u) v) \\ &= \sum_{K\in\mathscr{T}_h}\int_{\partial K}(A\nabla w f_2(\nabla u) v)\cdot n_{\partial K} - \int_K (A\nabla w)\cdot\nabla(f_2(\nabla u) v) \\ &\qquad- \sum_{F\in\mathscr{E}_h^{i}}\int_F [\![(\langle\!\langle\mathcal{P}(A_h f_2(\nabla u) v)\rangle\!\rangle\nabla w)\cdot n_F]\!] \\ &\qquad+ \sum_{K\in\mathscr{T}_h}\int_K D^2 w : (\mathcal{P}(A_h f_2(\nabla u) v) - A f_2(\nabla u) v) \\ &= -\sum_{K\in\mathscr{T}_h}\int_K (A\nabla w)\cdot\nabla(f_2(\nabla u) v) - D^2 w : ((A f_2(\nabla u) v) - \mathcal{P}(A_h f_2(\nabla u) v))\end{aligned}$$

$$- \sum_{F \in \mathscr{E}_h^i} \int_F [\![(\langle\!\langle \mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v \rangle\!\rangle \nabla w) \cdot n_F]\!] \tag{A.11}$$

Now, from (A.11), a further application of integration by parts, and (A.9), we obtain

$$\begin{aligned}
&\int_\Omega A_h : (\mathbf{H}_h w) f_2(\nabla u)v \\
&= -\int_\Omega A\nabla w \cdot \nabla(f_2(\nabla u)v) + [\nabla \cdot (\mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v)] \cdot \nabla w \\
&\quad - \sum_{F \in \mathscr{E}_h^i} \int_F [\![(\langle\!\langle \mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v \rangle\!\rangle \nabla w) \cdot n_F]\!] \\
&\quad + \sum_{K \in \mathscr{T}_h} \int_{\partial K} [(\mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v) \nabla w] \cdot n_{\partial K} \\
&= -\int_\Omega A\nabla w \cdot \nabla(f_2(\nabla u)v) + [\nabla \cdot (\mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v)] \cdot \nabla w \\
&\quad + \sum_{F \in \mathscr{E}_h^{i,b}} \int_F [\![[(\mathcal{P}(A_h f_2(\nabla u)v) - A f_2(\nabla u)v) \langle\!\langle \nabla w \rangle\!\rangle] \cdot n_F]\!].
\end{aligned} \tag{A.12}$$

By the chain rule, and the symmetry of the Hessian, it is clear that

$$\nabla f_2(\nabla u) = D^2 u \; D_q f_2(\nabla u),$$

and so

$$\begin{aligned}
\int_\Omega A\nabla w \cdot \nabla f_2(\nabla u)v - \det(\mathcal{P}(D^2 u))(D_q f_2(\nabla u) \cdot \nabla w)v = \\
= \int_\Omega A\nabla w \cdot (D^2 u \; D_q f_2(\nabla u))v - \det(D^2 u)(D_q f_2(\nabla u) \cdot \nabla w)v \\
+ \int_\Omega (\det D^2 u - \det \mathcal{P}(D^2 u))(D_q f_2(\nabla u) \cdot \nabla w)v.
\end{aligned}$$

Recalling that $A = \operatorname{Cof} D^2 u = \det D^2 u (D^2 u)^{-1}$, we obtain

$$\begin{aligned}
A\nabla w \cdot (D^2 u \; D_q f_2(\nabla u)) &= (D^2 u \; D_q f_2(\nabla u))^T A \nabla w \\
&= D_q f_2(\nabla u)^T D^2 u (\det(D^2 u)(D^2 u)^{-1}) \nabla w \\
&= \det(D^2 u) D_q f_2(\nabla u) \cdot \nabla w,
\end{aligned}$$

and thus

$$\begin{aligned}
&\int_\Omega A\nabla w \cdot \nabla f_2(\nabla u)v - \det \mathcal{P}(D^2 u) D_q f_2(\nabla u) \cdot \nabla w v = \\
&\quad = \int_\Omega (\det D^2 u - \det \mathcal{P}(D^2 u))(D_q f_2(\nabla u) \cdot \nabla w)v \\
&\quad = \frac{1}{2} \int_\Omega (D^2 u + \mathcal{P}(D^2 u)) : (D^2 u - \mathcal{P}(D^2 u))(D_q f_2(\nabla u) \cdot \nabla w)v.
\end{aligned} \tag{A.13}$$

From (A.12) and (A.13) we obtain

$$\begin{aligned}\langle L_u[w], v\rangle &= \int_\Omega (D_z f_1(x,u)w - f_2(\nabla u)A_h : \mathbf{H}_h[w] - \det(\mathcal{P}(D^2u))D_q f_2(\nabla u)\cdot\nabla w)v \\ &= \int_\Omega f_2(\nabla u)A\nabla w\cdot\nabla v - \nabla w\cdot\nabla_h\cdot(Af_2(\nabla u)v - \mathcal{P}(A_h f_2(\nabla u)v)) \\ &+ \frac{1}{2}\int_\Omega (D^2u + \mathcal{P}(D^2u)) : (D^2u - \mathcal{P}(D^2u))(D_q f_2(\nabla u)\cdot\nabla w)v \\ &+ \int_\Omega D_z f_1(x,u)wv - \sum_{F\in\mathscr{E}_h^{i,b}}\int_F [\![(\mathcal{P}(A_h f_2(\nabla u)v) - Af_2(\nabla u)v)\langle\!\langle\nabla w\rangle\!\rangle]\cdot n_F]\!]. \qquad \square\end{aligned}$$

Corollary A.7 *We have the following estimate for u and u_*:*

$$\|u - u_*\|_{W^{1,\infty}(\Omega)} \lesssim (1 + |\ln h|)^{1/2} h^{p-1}. \tag{A.14}$$

Proof: Noting that $(\pi_h(u) - u_*)|_{\partial\Omega} = \phi_h - \phi_h = 0$, we see that

$$\begin{aligned}\|u - u_*\|_{W^{1,\infty}(\Omega)} &\lesssim \|u - \pi_h(u)\|_{W^{1,\infty}(\Omega)} + \|\pi_h(u) - u_*\|_{W^{1,\infty}(\Omega)} \\ &\lesssim h^p\|u\|_{W^{p+1,\infty}(\Omega)} + h^{-1}\|\pi_h(u) - u_*\|_{\infty,\Omega} \\ &\lesssim h^p + (1 + |\ln h|)^{1/2}h^{-1}|\pi_h(u) - u_*|_{H^1(\Omega)} \\ &\le h^p + (1 + |\ln h|)^{1/2}h^{-1}(|\pi_h(u) - u|_{H^1(\Omega)} + |u - u_*|_{H^1(\Omega)}) \\ &\lesssim h^p + (1 + |\ln h|)^{1/2}h^{-1}(h^p\|u\|_{H^{p+1}(\Omega)} + \|u - u_*\|_h) \\ &\lesssim h^p + (1 + |\ln h|)^{1/2}h^{-1}h^p \\ &\lesssim (1 + |\ln h|)^{1/2}h^{p-1}. \qquad \square\end{aligned}$$

Lemma A.8 *Let $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u := \{v - u : v \in \mathbb{B}_{h^{2+\alpha}}(u_*)\}$, for some $\alpha > 0$. Then, if $h \le h_0$, for some $h_0 \in (0,1)$, we have the following estimate:*

$$\|R[w_1] - R[w_2]\|_{-1,h} \lesssim \sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}\|w_1 - w_2\|_h. \tag{A.15}$$

Furthermore, for $w \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, we have

$$\|R[w]\|_{-1,h} \lesssim h^{2+\alpha}\sum_{i=1}^{3} h^{i(1+\alpha)-1}(1 + |\ln h|)^{\frac{i}{2}}. \tag{A.16}$$

Proof: First, let us define $R_i : V \to \mathring{\mathbb{V}}'_{h,p}$, $i = 1, \ldots, 5$ by

$$\langle R_1[w], v\rangle = \int_\Omega \gamma_1(w)v \quad \forall w \in V,\ \forall v \in \mathring{\mathbb{V}}_{h,p},$$

$$
\begin{aligned}
\langle R_2[w], v\rangle &= -\int_\Omega (\det \mathcal{P}_{\mathbb{W}_{h,p}}(D^2u) + A_h : \mathbf{H}_h[w])\gamma_2(\nabla w)v \quad \forall w \in V,\ \forall v \in \mathring{\mathbb{V}}_{h,p},\\
\langle R_3[w], v\rangle &= -\int_\Omega (f_2(\nabla u) + D_q f_2(\nabla u)\cdot\nabla w)\det(\mathbf{H}_h[w])v \quad \forall w \in V,\ \forall v \in \mathring{\mathbb{V}}_{h,p},\\
\langle R_4[w], v\rangle &= -\int_\Omega (A_h : \mathbf{H}_h[w] D_q f_2(\nabla u)\cdot\nabla w)v \quad \forall w \in V,\ \forall v \in \mathring{\mathbb{V}}_{h,p},\\
\langle R_5[w], v\rangle &= -\int_\Omega (\det(\mathbf{H}_h[w])\gamma_2(\nabla w))v \quad \forall w \in V,\ \forall v \in \mathring{\mathbb{V}}_{h,p},
\end{aligned}
\tag{A.17}
$$

where we are denoting:

$$
\gamma_1(w) := R^{f_1,u}(w) = f_1(x, u+w) - f_1(x,u) - D_z f_1(x,u)w,
$$

and

$$
\gamma_2(\nabla w) := R^{f_2,\nabla u}(\nabla w) = f_2(\nabla u + \nabla w) - f_2(\nabla u) - D_q f_2(\nabla u)\cdot\nabla w,
$$

where $R^{f_1,u}$ and $R^{f_2,\nabla u}$ are the quadratic remainder terms of f_1 and f_2 associated with (x,u) and ∇u respectively, given by Definition 2.2.10. It is clear that $R = \sum_{i=1}^5 R_i$.

Our goal is to prove contraction estimates for R, which, by virtue of the triangle inequality, is achieved by proving contraction estimates for $R_1, \dots, R_5$. In order to do this, we must appeal to the abstract calculus structure of Lemma 2.2.11 and Corollary 2.2.12, which allows us to quantify the error arising from the quadratic Taylor expansion that lead to the definition of R. However, as we can see in Lemma 2.2.11, we are required to identify a set U on which f_1, and f_2 are bounded in the $C^{2,1}$-norm. Since f_1, f_2 are at least p-times (with $p \geq 3$) continuously differentiable on $\overline{\Omega}\times\mathbb{R}$, $\mathbb{R}^2$, respectively, their restrictions to compact subsets are indeed bounded in the $C^{2,1}$-norm.

Furthermore, in order to apply Corollary 2.2.12, the arguments of γ_1 and γ_2 must lie in a ball with radius $r \in [0,1)$ that is also contained in U. In order to justify that we may suitably bound our expansions with $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, we require the following argument: let $w_i \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, $i = 1,2$, then $w_i = v_i - u$ for some $v_i \in \mathbb{B}_{h^{2+\alpha}}(u_*)$, and so

$$
\begin{aligned}
\|w_i\|_{W^{1,\infty}(\Omega)} &\leq \|v_i - u_*\|_{W^{1,\infty}(\Omega)} + \|u_* - u\|_{W^{1,\infty}(\Omega)}\\
&\leq C(1 + |\ln h|)^{\frac{1}{2}} h^{-1} h^{2+\alpha} + \|u_* - u\|_{W^{1,\infty}(\Omega)}\\
&\leq C(1 + |\ln h|)^{\frac{1}{2}} h^{1+\alpha}.
\end{aligned}
\tag{A.18}
$$

Note that the final inequality follows from (A.14). Since $p \geq 3$, we can choose $h_0 \in (0,1)$ such that if $0 < h \leq h_0$, then

$$\|w_i\|_{W^{1,\infty}(\Omega)} \leq C(1 + |\ln h|)^{\frac{1}{2}} h^{1+\alpha} \leq r \in [0,1).$$

Furthermore, since $u \in C^{p+2}(\overline{\Omega})$, there exists a constant $K > 0$ such that $\|u\|_{W^{1,\infty}(\Omega)} \leq K$. Let us define $U_j := \{x \in \mathbb{R}^j : |x| < \max\{K, 1\}\}$, $j = 1, 2$. Then $U_j \supset B_{r,j} := \{x \in \mathbb{R}^j : |x| < r\}$, $j = 1, 2$, and we have that $\|f_2\|_{C^{2,1}(\overline{U_2})} < \infty$, and, for all $x \in \overline{\Omega}$, $\|f_1(x, \cdot)\|_{C^{2,1}(\overline{U_1})} \leq \|f_1\|_{C^{2,1}(\overline{\Omega} \times \overline{U_1})} < \infty$.

Thus, since $w_i(\Omega) \subseteq B_{r,1}$, $\nabla w_i(\Omega) \subset B_{r,2}$, $i = 1, 2$, $u(\Omega) \subset U_1$, $\nabla u(\Omega) \subset U_2$, by Corollary 2.2.12 we have that

$$\begin{cases} |\gamma_1(w_1) - \gamma_1(w_2)| \leq C(1,1)\|f_1\|_{C^{2,1}(\overline{\Omega} \times \overline{U_1})}(|w_1| + |w_2|)|w_1 - w_2| & \text{in } \Omega, \\ |\gamma_2(\nabla w_1) - \gamma_2(\nabla w_2)| \leq C(2,1)\|f_2\|_{C^{2,1}(\overline{U_2})}(|\nabla w_1| + |\nabla w_2|)|\nabla w_1 - \nabla w_2| & \text{a.e. in } \Omega. \end{cases}$$

The above estimates will be used several times in the following analysis. It is also useful to note that since $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, we may represent $w_i = v_i - u$, $i = 1, 2$, for some $v_i \in \mathbb{B}_{h^{2+\alpha}}(u_*)$. Moreover, we see that $w_1 - w_2, v_i - u_* \in \mathring{\mathbb{V}}_{h,p}$, $i = 1, 2$.

The following estimates will also be applied several times:

$$\|w_i\|_h \leq \|v_i - u_*\|_h + \|u_* - u\|_h \lesssim h^{2+\alpha}, \tag{A.19}$$

$$\begin{aligned} \|\mathbf{H}_h w_i\|_{2,\Omega} &\leq \|\mathbf{H}_h[v_i - u_*]\|_{2,\Omega} + \|\mathbf{H}_h[u_* - u]\|_{2,\Omega} \\ &\leq Ch^{-1}\|v_i - u_*\|_h + h^{p-1} \\ &\lesssim h^{1+\alpha}, \end{aligned} \tag{A.20}$$

note that the first inequality in (A.20) follows from (4.11.5) and (8.3.15).

Let us consider the first remainder term, R_1. For $w_1, w_2 \in \mathbb{B}_{h^{2+\alpha}}(u_*) - u$, $v \in \mathring{\mathbb{V}}_{h,p}$, that

$$\begin{aligned} \langle R_1[w_1] - R_1[w_2], v\rangle &= \int_\Omega (\gamma_1(w_1) - \gamma_1(w_2))v \\ &\lesssim \int_\Omega (|w_1| + |w_2|)|w_1 - w_2||v| \\ &\lesssim (\|w_1\|_{2,\Omega} + \|w_2\|_{2,\Omega})\|w_1 - w_2\|_{2,\Omega}\|v\|_{\infty,\Omega}. \end{aligned}$$

Thus, by (A.5) and (A.6)

$$\begin{aligned} \langle R_1[w_1] - R_1[w_2], v\rangle &= \langle R_1[v_1 - u] - R_1[v_2 - u], v\rangle \\ &\lesssim (\|v_1 - u\|_{2,\Omega} + \|v_2 - u\|_{2,\Omega})\|w_1 - w_2\|_{2,\Omega}\|v\|_{\infty,\Omega} \\ &\lesssim (1 + |\ln h|)^{1/2}(\|v_1 - u\|_{2,\Omega} + \|v_2 - u\|_{2,\Omega})\|w_1 - w_2\|_{2,\Omega}\|v\|_h \\ &\lesssim (1 + |\ln h|)^{1/2}\sum_{i=1}^{2}[\|v_i - u_*\|_{2,\Omega} + \|u - u_*\|_{2,\Omega}]\|w_1 - w_2\|_{2,\Omega}\|v\|_h \\ &\lesssim (1 + |\ln h|)^{1/2}h^{2+\alpha}\|w_1 - w_2\|_h\|v\|_h. \end{aligned}$$

Thus, it follows that

$$\|R_1[w_1] - R_1[w_2]\|_{-1,h} \lesssim (1 + |\ln h|)^{1/2} h^{2+\alpha} \|w_1 - w_2\|_h. \tag{A.21}$$

We also see that

$$\begin{aligned}
\langle R_2[w_1] - R_2[w_2], v\rangle &= -\int_\Omega \det \mathcal{P}_{\mathbb{W}_{h,p}}(D^2 u)(\gamma_2(\nabla w_1) - \gamma_2(\nabla w_2))v \\
&\quad - \frac{1}{2}\int_\Omega (A_h : \mathbf{H}_h[w_1 + w_2](\gamma_2(\nabla w_1) - \gamma_2(\nabla w_2)))v \\
&\quad - \frac{1}{2}\int_\Omega (\gamma_2(\nabla w_1) + \gamma_2(\nabla w_2))(A_h : \mathbf{H}_h[w_1 - w_2])v \\
&\lesssim \int_\Omega (|\nabla w_1| + |\nabla w_2|)|\nabla w_1 - \nabla w_2||v| \\
&\quad + \int_\Omega (|\mathbf{H}_h[w_1]| + |\mathbf{H}_h[w_2]|)(|\nabla w_1| + |\nabla w_2|)|\nabla w_1 - \nabla w_2||v| \\
&\quad + \int_\Omega (|\nabla w_1|^2 + |\nabla w_2|^2)|\mathbf{H}_h[w_1 - w_2]|)|v| \\
&\lesssim (\|w_1\|_h + \|w_2\|_h)\|w_1 - w_2\|_h \|v\|_{\infty,\Omega} \\
&\quad + (\|\mathbf{H}_h[w_1]\|_{2,\Omega} + \|\mathbf{H}_h[w_2]\|_{2,\Omega})(\|w_1\|_h + \|w_2\|_h)\|w_1 - w_2\|_{W^{1,\infty}(\Omega)}\|v\|_{\infty,\Omega} \\
&\quad + (\|w_1\|_h + \|w_2\|_h)(\|w_1\|_{W^{1,\infty}(\Omega)} + \|w_2\|_{W^{1,\infty}(\Omega)})\|H[w_1 - w_2]\|_{2,\Omega}\|v\|_{\infty,\Omega} \\
&\lesssim (1 + |\ln h|)^{1/2}(\|w_1\|_h + \|w_2\|_h)\|w_1 - w_2\|_h\|v\|_h \\
&\quad + (1 + |\ln h|)h^{-1}\sum_{i=1}^{2}\|\mathbf{H}_h w_i\|_{2,\Omega}(\|w_1\|_h + \|w_2\|_h)\|w_1 - w_2\|_h\|v\|_h \\
&\quad + (1 + |\ln h|)^{1/2}h^{-1}(\|w_1\|_h + \|w_2\|_h)(\|w_1\|_{W^{1,\infty}(\Omega)} + \|w_2\|_{W^{1,\infty}(\Omega)})\|w_1 - w_2\|_h\|v\|_h,
\end{aligned}$$

where the final inequality follows from (4.6.27), (4.11.5), and (A.5).

Then, by (A.18), (A.19), and (A.20), it then follows that

$$\langle R_2[w_1] - R_2[w_2], v\rangle \lesssim ((1 + |\ln h|)^{1/2}h^{1+\alpha} + (1 + |\ln h|)h^{2+2\alpha})\|w_1 - w_2\|_h\|v\|_h,$$

which results in

$$\|R_2[w_1] - R_2[w_2]\|_{-1,h}((1 + |\ln h|)^{1/2}h^{1+\alpha} + (1 + |\ln h|)h^{2+2\alpha})\|w_1 - w_2\|_h. \tag{A.22}$$

For the third remainder term, we have

$$\begin{aligned}
\langle R_3[w_1] - R_3[w_2], v\rangle &= \int_\Omega (f_2(\nabla u)(\det(\mathbf{H}_h[w_1]) - \det(\mathbf{H}_h[w_2])) \\
&\quad + \frac{1}{2}\int_\Omega D_q f_2(\nabla u)\cdot\nabla(w_1 + w_2)(\det(\mathbf{H}_h[w_1]) - \det(\mathbf{H}_h[w_2])))v \\
&\quad + \frac{1}{2}\int_\Omega (\det(\mathbf{H}_h[w_1]) + \det(\mathbf{H}_h[w_2]))D_q f_2(\nabla u)\cdot\nabla(w_1 - w_2)v
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \int_\Omega (|\mathbf{H}_h[w_1]| + |\mathbf{H}_h[w_2]|)|\mathbf{H}_h[w_1 - w_2]||v| \\
&\quad + \int_\Omega (|\nabla w_1| + |\nabla w_2|)(|\mathbf{H}_h[w_1]| + |\mathbf{H}_h[w_2]|)|\mathbf{H}_h[w_1 - w_2]||v| \\
&\quad + \int_\Omega (|\mathbf{H}_h[w_1]|^2 + |\mathbf{H}_h[w_2]|^2)|\nabla w_1 - \nabla w_2||v| \\
&\lesssim (\|\mathbf{H}_h[w_1]\|_{2,\Omega} + \|\mathbf{H}_h[w_2]\|_{2,\Omega})\|\mathbf{H}_h[w_1 - w_2]\|_{2,\Omega}\|v\|_{\infty,\Omega} \\
&\quad + \left(\sum_{i=1}^2 \|\mathbf{H}_h[w_i]\|_{2,\Omega}\right)\left(\sum_{i=1}^2 \|w_i\|_{W^{1,\infty}(\Omega)}\right)\|\mathbf{H}_h[w_1 - w_2]\|_{2,\Omega}\|v\|_{\infty,\Omega} \\
&\quad + \left(\sum_{i=1}^2 \|\mathbf{H}_h[w_i]\|_{2,\Omega}\right)^2 \|w_1 - w_2\|_{W^{1,\infty}(\Omega)}\|v\|_{\infty,\Omega} \\
&\lesssim ((1 + |\ln h|)^{1/2}h^\alpha + (1 + |\ln h|)h^{1+2\alpha})\|w_1 - w_2\|_h\|v\|_h,
\end{aligned}
$$

where the final inequality follows from (4.11.5), (A.5), (A.18), and (A.20). Thus,

$$
\|R_3[w_1] - R_3[w_2]\|_{-1,h} \lesssim ((1 + |\ln h|)^{1/2}h^{1+\alpha} + (1 + |\ln h|)h^{1+2\alpha})\|w_1 - w_2\|_h. \quad \text{(A.23)}
$$

For the fourth remainder term, we have

$$
\begin{aligned}
\langle R_4[w_1] - R_4[w_2], v\rangle &= -\frac{1}{2}\int_\Omega A_h : \mathbf{H}_h[w_1 + w_2]D_q f_2(\nabla u)\cdot\nabla(w_1 - w_2)v \\
&\quad + \frac{1}{2}\int_\Omega A_h : \mathbf{H}_h[w_1 - w_2]D_q f_2(\nabla u)\cdot\nabla(w_1 + w_2)v \\
&\lesssim \left(\left(\sum_{i=1}^2 \|\mathbf{H}_h[w_i]\|_{2,\Omega}\right)\|w_1 - w_2\|_h + \left(\sum_{i=1}^2 \|w_i\|_h\right)\|\mathbf{H}_h[w_1 - w_2]\|_{2,\Omega}\right)\|v\|_{\infty,\Omega} \\
&\lesssim (1 + |\ln h|)^{1/2}h^{1+\alpha}\|w_1 - w_2\|_h\|v\|_h,
\end{aligned}
$$

where the final inequality follows from (4.11.5), (A.5), (A.19), and (A.20), and so

$$
\|R_4[w_1] - R_4[w_2]\|_{-1,h} \lesssim (1 + |\ln h|)^{1/2}h^{1+\alpha}\|w_1 - w_2\|_h. \quad \text{(A.24)}
$$

For the fifth, and final remainder term, we have

$$
\begin{aligned}
\langle R_5[w_1] - R_5[w_2], v\rangle &= -\frac{1}{2}\int_\Omega ((\det(\mathbf{H}_h[w_1]) + \det(\mathbf{H}_h[w_2]))(\gamma_2(\nabla w_1) - \gamma_2(\nabla w_2)))v \\
&\quad - \frac{1}{2}\int_\Omega (\det(\mathbf{H}_h[w_1]) - \det(\mathbf{H}_h[w_2]))(\gamma_2(\nabla w_1) + \gamma_2(\nabla w_2))v \\
&\lesssim \int_\Omega (|\mathbf{H}_h[w_1]|^2 + |\mathbf{H}_h[w_2]|^2)(|\nabla w_1| + |\nabla w_2|)|\nabla w_1 - \nabla w_2||v| \\
&\quad + \int_\Omega (|\mathbf{H}_h[w_1]| + |\mathbf{H}_h[w_2]|)(|\nabla w_1|^2 + |\nabla w_2|^2)|\mathbf{H}_h[w_1 - w_2]||v| \\
&\lesssim \left(\sum_{i=1}^2 \|\mathbf{H}_h[w_i]\|_{2,\Omega}\right)^2 \left(\sum_{i=1}^2 \|w_i\|_{W^{1,\infty}(\Omega)}\right)\|w_1 - w_2\|_{W^{1,\infty}(\Omega)}\|v\|_{\infty,\Omega}
\end{aligned}
$$

$$+\left(\sum_{i=1}^{2}\|\mathbf{H}_h[w_i]\|_{2,\Omega}\right)\left(\sum_{i=1}^{2}\|w_i\|_{W^{1,\infty}(\Omega)}\right)^2\|\mathbf{H}_h[w_1-w_2]\|_{2,\Omega}\|v\|_{\infty,\Omega}$$
$$\lesssim (1+|\ln h|)^{3/2}h^{2+3\alpha}\|w_1-w_2\|_h\|v\|_h,$$

where the final inequality follows from (4.6.27), (4.11.5), (A.5), (A.19) and (A.20), and so

$$\|R_5[w_1]-R_5[w_2]\|_{-1,h}\lesssim (1+|\ln h|)^{3/2}h^{2+3\alpha}\|w_1-w_2\|_h. \tag{A.25}$$

Since $R=\sum_{i=1}^{5}R_i$, using the triangle inequality, we can obtain an upper bound for $\|R[w_1]-R[w_2]\|_{-1,h}$ by summing our bounds for $R_1,\ldots,R_5$, i.e., from (A.21)–(A.25) we obtain (noting that $h\le h_0<1$)

$$\begin{aligned}
&\|R[w_1]-R[w_2]\|_{-1,h}\le\sum_{i=1}^{5}\|R_i[w_1]-R_i[w_2]\|_{-1,h}\\
&\lesssim\Big\{(1+|\ln h|)^{\frac12}h^{2+\alpha} && \text{by } (A.21)\\
&\quad+(1+|\ln h|)^{\frac12}h^{1+\alpha}+(1+|\ln h|)h^{2+2\alpha} && \text{by } (A.22)\\
&\quad+(1+|\ln h|)^{\frac12}h^{\alpha}+(1+|\ln h|)h^{1+2\alpha} && \text{by } (A.23)\\
&\quad+(1+|\ln h|)^{\frac12}h^{1+\alpha} && \text{by } (A.24)\\
&\quad+(1+|\ln h|)^{\frac32}h^{2+3\alpha}\Big\}\times\|w_1-w_2\|_h && \text{by } (A.25)\\
&\lesssim (h^{\alpha}(1+|\ln h|)^{\frac12}+h^{1+2\alpha}(1+|\ln h|)+h^{2+3\alpha}(1+|\ln h|)^{\frac32})\|w_1-w_2\|_h\\
&=\sum_{i=1}^{3}h^{i(1+\alpha)-1}(1+|\ln h|)^{\frac{i}{2}}\|w_1-w_2\|_h.
\end{aligned}$$

Thus we have obtained (A.15). The derivation of (A.16) is analogous; we simply appeal to Corollary 2.2.13, i.e., we let $w_1:=w\in\mathbb{B}_{h^{2+\alpha}}(u_*)-u$, and $w_2\equiv 0$.

Taking $w\in\mathbb{B}_{h^{2+\alpha}}(u_*)-u$ and $v\in\mathring{\mathbb{V}}_{h,p}$, by (A.6) and (A.5), we obtain

$$\langle R_1[w],v\rangle\lesssim\|w\|_{2,\Omega}^2\|v\|_{\infty,\Omega}\lesssim(1+|\ln h|)^{\frac12}h^{4+2\alpha}\|v\|_h. \tag{A.26}$$

By (A.5), (A.18), (A.19), and (A.20), we obtain

$$\begin{aligned}
\langle R_2[w],v\rangle&\lesssim(\|w\|_h^2+\|\mathbf{H}_hw\|_{2,\Omega}\|w\|_h\|w\|_{W^{1,\infty}(\Omega)})\|v\|_{\infty,\Omega}\\
&\lesssim(h^{4+2\alpha}(1+|\ln h|)^{\frac12}+h^{4+3\alpha}(1+|\ln h|))\|v\|_h.
\end{aligned} \tag{A.27}$$

By (A.5), (A.19), and (A.20), we obtain

$$\begin{aligned}
\langle R_3[w],v\rangle&\lesssim(\|\mathbf{H}_hw\|_{2,\Omega}^2+\|\mathbf{H}_hw\|_{2,\Omega}^2\|w\|_{W^{1,\infty}(\Omega)})\|v\|_{\infty,\Omega}\\
&\lesssim((1+|\ln h|)^{\frac12}h^{2+2\alpha}+(1+|\ln h|)h^{3+3\alpha})\|v\|_h.
\end{aligned} \tag{A.28}$$

By (A.5), (A.19), and (A.20), we obtain

$$\langle R_4[w], v\rangle \lesssim \|\mathbf{H}_h w\|_{2,\Omega}\|w\|_h\|v\|_{\infty,\Omega} \lesssim (1+|\ln h|)^{\frac{1}{2}}h^{3+2\alpha}\|v\|_h. \tag{A.29}$$

Finally, by (A.5), (A.19), and (A.20),

$$\langle R_5[w], v\rangle \lesssim \|\mathbf{H}_h w\|_{2,\Omega}^2\|w\|_{W^{1,\infty}(\Omega)}^2\|v\|_{\infty,\Omega} \lesssim (1+|\ln h|)^{\frac{3}{2}}h^{4+4\alpha}\|v\|_h. \tag{A.30}$$

Thus, it follows that

$$\begin{aligned}
\langle R[w], v\rangle &= \sum_{i=1}^{5}\langle R_i[w], v\rangle \lesssim \Big\{(1+|\ln h|)^{\frac{1}{2}}h^{2p} && \text{by } (A.26)\\
&\quad + h^{4+2\alpha}(1+|\ln h|)^{\frac{1}{2}} + h^{4+3\alpha}(1+|\ln h|) && \text{by } (A.27)\\
&\quad + (1+|\ln h|)^{\frac{1}{2}}h^{2+2\alpha} + (1+|\ln h|)h^{3+3\alpha} && \text{by } (A.28)\\
&\quad + (1+|\ln h|)^{\frac{1}{2}}h^{3+2\alpha} && \text{by } (A.29)\\
&\quad + (1+|\ln h|)^{\frac{3}{2}}h^{4+4\alpha}\ \Big\} \times \|v\|_h && \text{by } (A.30)\\
&\lesssim (h^{2+2\alpha}(1+|\ln h|)^{\frac{1}{2}} + h^{3+3\alpha}(1+|\ln h|) + h^{4+4\alpha}(1+|\ln h|)^{\frac{3}{2}})\|v\|_h\\
&= h^{2+\alpha}\sum_{i=1}^{3}h^{i(1+\alpha)-1}(1+|\ln h|)^{\frac{i}{2}}\|v\|_h.
\end{aligned}$$

By the definition of $\|\cdot\|_{-1,h}$, we obtain (A.15). $\square$

Appendix B

Data for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
1	3089599.32	0.50	6
2	7579655.15	0.28	6
3	14760336.26	0.16	7
4	3150458.84	0.10	7
5	218268867.53	0.05	7
6	542447450.07	0.03	8
7	910192579.44	0.01	7
8	5018313.99	0.50	6
9	9444236.63	0.28	6
10	2538421.34	0.16	7
11	48667241.55	0.10	7
12	159481809.63	0.05	7
13	195170239.59	0.03	7
14	682066167.58	0.01	8
15	1973988.67	0.50	7
16	2642355.08	0.28	6
17	6427946.40	0.16	7
18	73594883.72	0.10	7
19	75116712.66	0.05	7
20	338418648.81	0.03	7
21	581873560.10	0.01	7
22	4197895.27	0.50	6
23	20468545.12	0.28	6
24	37391881.80	0.16	7
25	57306326.45	0.10	6
26	53861186.13	0.05	7

Table B.1: Initial distance, mesh size, and number of Newton iterations for the index i in the range $1, \ldots, 26$, for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
27	163842183.95	0.03	7
28	937562497.74	0.01	7
29	637895.22	0.50	6
30	734122.33	0.28	6
31	34369881.29	0.16	7
32	853023.83	0.10	7
33	137520631.22	0.05	7
34	48235590.71	0.03	8
35	299685506.11	0.01	7
36	1838915.77	0.50	6
37	7940868.05	0.28	7
38	29510546.36	0.16	6
39	98593989.70	0.10	7
40	165143046.65	0.05	7
41	333014524.71	0.03	8
42	421591682.84	0.01	7
43	6789010.68	0.50	6
44	4960417.55	0.28	6
45	28440054.28	0.16	7
46	57074843.12	0.10	6
47	8635511.59	0.05	7
48	355319787.08	0.03	7
49	1416399903.32	0.01	7
50	1713380.42	0.50	6
51	15069373.92	0.28	6
52	21635067.13	0.16	7
53	108190210.72	0.10	7
54	397829085.23	0.05	7
55	474884727.96	0.03	7
56	1756918495.70	0.01	14
57	10784290.47	0.50	6
58	14068302.92	0.28	6
59	39923076.96	0.16	7
60	10593511.55	0.10	7
61	42788765.19	0.05	7
62	330653407.31	0.03	6
63	2491935658.34	0.01	10
64	7493685.75	0.50	6
65	28499629.84	0.28	7
66	25697621.10	0.16	7
67	64908529.70	0.10	7

Table B.2: Initial distance, mesh size, and number of Newton iterations for the index i in the range $27, \ldots, 67$, for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
68	108240288.36	0.05	7
69	176137840.81	0.03	7
70	1096993542.64	0.01	9
71	526722.54	0.50	6
72	1773479.64	0.28	7
73	10021648.81	0.16	7
74	55443360.42	0.10	7
75	176847481.92	0.05	7
76	908318797.18	0.03	7
77	129483993.57	0.01	7
78	7254721.16	0.50	7
79	5744392.50	0.28	7
80	15983317.32	0.16	7
81	24125153.01	0.10	7
82	35438987.78	0.05	7
83	675332847.84	0.03	8
84	1207828497.98	0.01	7
85	5578622.54	0.50	7
86	16065342.24	0.28	7
87	6418264.96	0.16	7
88	15775962.92	0.10	7
89	134452747.06	0.05	7
90	27432274.48	0.03	7
91	362473195.52	0.01	7
92	12810457.97	0.50	7
93	19406994.54	0.28	7
94	22053708.99	0.16	7
95	31367673.06	0.10	7
96	318871390.88	0.05	7
97	622929439.43	0.03	7
98	809463625.45	0.01	7
99	4243041.73	0.50	6
100	13479197.55	0.28	7
101	23438909.13	0.16	7
102	32624787.32	0.10	7
103	175534181.58	0.05	7
104	60959806.27	0.03	7
105	813407855.49	0.01	10
106	2352216.95	0.50	6
107	8166806.59	0.28	7

Table B.3: Initial distance, mesh size, and number of Newton iterations for the index i in the range $68, \ldots, 107$, for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
108	12569701.31	0.16	7
109	19900537.47	0.10	7
110	54693296.32	0.05	7
111	213612002.70	0.03	9
112	719013991.36	0.01	7
113	5711448.60	0.50	6
114	11654458.28	0.28	6
115	567854.04	0.16	7
116	63743138.85	0.10	7
117	7054061.57	0.05	7
118	102127534.53	0.03	19
119	313292289.39	0.01	8
120	944401.17	0.50	6
121	7694254.40	0.28	7
122	3221090.85	0.16	7
123	44146683.84	0.10	6
124	102475312.96	0.05	6
125	158849320.41	0.03	7
126	246371471.83	0.01	7
127	8148145.82	0.50	6
128	10044970.32	0.28	6
129	47611485.42	0.16	7
130	49659558.67	0.10	7
131	99121122.56	0.05	7
132	439207168.70	0.03	8
133	1585653747.74	0.01	7
134	3072861.68	0.50	6
135	7618492.84	0.28	6
136	46914578.83	0.16	7
137	27035953.55	0.10	7
138	86220739.85	0.05	7
139	185949027.79	0.03	7
140	842031032.01	0.01	8
141	5764053.42	0.50	6
142	13356469.27	0.28	6
143	14741742.64	0.16	6
144	12951501.79	0.10	7
145	223184282.30	0.05	7
146	346455255.04	0.03	7
147	378000720.64	0.01	7

Table B.4: Initial distance, mesh size, and number of Newton iterations for the index i in the range $108, \ldots, 147$, for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
148	2804013.28	0.50	6
149	1036112.56	0.28	6
150	26470433.46	0.16	7
151	91154602.91	0.10	7
152	158010996.03	0.05	7
153	938290654.09	0.03	7
154	2014600226.27	0.01	19
155	6289996.68	0.50	6
156	17908638.22	0.28	6
157	56651789.04	0.16	7
158	136696155.85	0.10	7
159	37600873.58	0.05	7
160	560568443.13	0.03	7
161	2448171028.96	0.01	9
162	7735930.83	0.50	6
163	7462547.55	0.28	7
164	34141576.58	0.16	7
165	33455516.54	0.10	7
166	358848015.56	0.05	7
167	273208888.42	0.03	7
168	2237495837.81	0.01	9
169	6750450.64	0.50	6
170	1760267.08	0.28	7
171	17867487.48	0.16	7
172	52319234.84	0.10	7
173	127110321.28	0.05	7
174	159856103.50	0.03	10
175	56080718.04	0.01	7
176	1636408.85	0.50	6
177	837819.82	0.28	6
178	11029934.26	0.16	7
179	24302687.12	0.10	7
180	135533658.72	0.05	7
181	499987268.86	0.03	7
182	1396629275.90	0.01	8
183	7289680.78	0.50	6
184	7860897.27	0.28	6

Table B.5: Initial distance, mesh size, and number of Newton iterations for the index i in the range $148, \ldots, 184$, for Experiment 7.9.5

MA prob. no.	$\|u_h^N - u_h^0\|_{h,1}$	Mesh size	Newton steps
185	28481642.41	0.16	7
186	88640530.34	0.10	7
187	164788478.27	0.05	7
188	83372130.31	0.03	7
189	941098199.79	0.01	9
190	6844494.09	0.50	6
191	3236549.98	0.28	7
192	30720530.34	0.16	7
193	37141731.59	0.10	7
194	118102856.57	0.05	7
195	336339460.33	0.03	7
196	1406694946.76	0.01	8
197	6190489.55	0.50	6
198	23284302.19	0.28	6
199	867900.07	0.16	6
200	44430657.10	0.10	7
201	198020607.68	0.05	7
202	293099439.70	0.03	7
203	2426174117.25	0.01	8
204	7149318.05	0.50	7
205	4946365.25	0.28	7
206	9328663.18	0.16	7
207	61837627.48	0.10	7
208	131489337.47	0.05	8
209	510652297.89	0.03	7
210	165901781.45	0.01	7

Table B.6: Initial distance, mesh size, and number of Newton iterations for the index i in the range $185, \ldots, 210$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{2,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
1	255430.77	−131701.89	98512.63	34565.22	−29042.31	114814.29	159726.75	176806.91	69865.86	210157.99
2	37460.12	−275197.18	141906.52	385283.33	−352000.88	144620.31	−354511.03	−303849.38	11069.03	−126292.11
3	−192233.90	−202339.05	−53148.39	240395.48	338058.51	467443.97	−141661.25	485312.63	524739.66	−35007.08
4	14362.22	−6888.40	−10229.51	27008.45	−4217.29	15042.04	−11546.33	−12318.45	−14355.00	−2172.08
5	72501.31	−341769.62	−170296.80	416022.48	−204848.32	−455712.27	530687.50	85340.19	−597693.10	−548500.24
6	−310569.18	−71625.94	−566552.96	−505645.18	246202.31	−245728.29	−539293.39	−431653.59	538379.37	27772.88
7	182383.10	276290.33	−208048.98	−268620.45	−153937.37	154322.85	189041.19	−108461.51	254015.16	223687.99
8	−270438.81	−278690.24	304149.39	1402.38	−265242.55	209171.32	342916.00	172163.80	29179.59	271764.67
9	448829.42	275321.92	−396826.91	−426839.04	−587734.50	−47551.98	439342.65	−177477.82	−509092.67	−90333.17
10	16968.11	−8096.82	−51186.22	32989.14	57119.39	−39427.93	−55042.00	2326.92	39237.54	32311.98
11	−349835.78	−166545.11	286004.43	−378285.21	314443.03	373768.09	307028.74	−65304.78	−50920.43	268171.39
12	−587079.32	373163.07	−384201.55	−574951.01	377961.02	372231.34	−300593.90	617046.02	37549.96	772097.64
13	179279.13	−38588.27	−24920.97	341123.54	−287353.02	166347.94	239738.50	172134.68	−25680.03	−80481.08
14	107970.57	−110207.18	51562.91	−72506.04	132694.21	155185.32	131982.74	66788.51	−105921.69	42128.86
15	22944.74	−46025.51	−236920.49	217201.63	−2853.28	−90618.46	113134.14	−181359.11	−100011.18	49653.87
16	−8726.86	−42082.85	76619.43	−107342.67	88180.01	24021.67	87945.23	−86232.54	−87914.55	−98174.14
17	−182571.46	−232713.98	−23784.20	−140007.00	172572.66	121381.66	−76386.18	−110286.66	150439.53	145211.32
18	−376732.99	30426.45	380965.02	286321.28	260638.23	−561708.36	33492.35	−681593.90	213637.20	−31912.06
19	−47454.09	382533.73	−92877.41	−326185.50	−195811.84	206824.08	88204.08	24568.89	−457970.05	352297.22
20	−178186.46	167952.20	54071.09	229427.08	−107834.31	−307291.57	139327.29	−319178.54	−56600.53	75908.78
21	−97924.75	151504.52	1425.62	−65898.11	−114963.53	−2652.84	22466.96	19203.58	121723.71	−142211.67
22	−213168.37	−102202.63	119163.01	−218582.40	−226676.95	−248671.45	134507.94	−201889.79	−185303.42	−152448.78
23	364841.54	−249970.65	−731971.15	344798.30	−365456.88	514811.38	−47407.35	−128948.10	234202.27	−523827.02
24	265571.60	301942.03	−494273.99	−240802.81	148366.71	484304.31	−432781.04	−91249.17	−32567.47	−359935.73
25	27860.85	−570386.69	139847.68	602171.79	264110.23	−659711.66	−261613.00	−637544.40	580537.41	−164251.99
26	309082.52	241722.80	106752.40	205336.35	−110592.82	−317300.57	28355.76	259343.13	−349421.92	−99572.58
27	−163968.95	−149075.28	−58382.73	26693.32	60668.63	152753.68	133718.19	−18816.07	−139402.07	−50194.83
28	−548555.83	−378929.99	245748.08	−251489.31	798272.87	47338.90	607657.77	−251964.89	327018.37	−510233.08
29	−73695.71	−39588.89	−43765.22	−38490.73	1624.25	78673.42	49206.02	67368.65	14436.27	−4595.90
30	12708.53	8827.60	16128.81	−352.85	−49051.18	29106.51	7256.50	20255.72	−17053.91	−30789.84
31	−639375.69	298886.60	3137.52	204639.97	268196.33	−268215.66	−212538.33	475302.93	290540.42	44772.15
32	6766.50	−310.12	−4358.21	10906.09	−11413.83	−8515.76	6708.47	−4721.66	4981.53	−5929.12
33	−65388.83	495131.82	−52259.70	−75488.64	292868.64	−377911.97	234456.53	−496888.09	−128046.63	−559833.50
34	47051.49	14489.19	72462.93	6060.58	−48144.01	−51166.71	17322.88	15801.64	−76268.90	−7030.48
35	51593.37	78041.40	−72676.71	−8574.81	−51801.10	55886.62	52544.66	95747.62	72595.06	101067.44
36	89528.79	−91236.23	−174181.41	−179359.29	56455.87	1180076.43	−70789.55	−9315.36	57644.22	63920.91
37	2043.26	133771.49	−409496.73	421249.85	115493.84	−4919.37	552681.48	380631.00	85077.94	−79288.52
38	−394485.49	341708.06	−604396.69	343908.53	618389.49	367642.95	9320.13	3510097.57	425311.84	56579.27
39	−627732.32	−140621.59	338472.54	−70182.83	28617.31	−520400.04	253771.71	214618.83	388430.92	−395360.64
40	232188.29	−332471.89	−320363.23	251520.00	172255.32	18811.06	−270445.76	−65721.46	−209909.00	51381.57

Table B.7: Initial guess random coefficients for the index i in the range $1, \ldots, 40$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{0,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
41	−167859.06	−159100.23	−218443.87	−106404.70	119082.53	−164588.21	−52365.09	−168857.33	−191479.84	−159589.48
42	−26131.35	226173.60	257035.19	−27700.22	−37127.13	−14866.76	−1766.66	392618.99	−188802.81	−332600.60
43	−57477.24	489387.81	−165355.12	−503547.08	463369.57	499600.43	−270250.84	318094.72	−495009.72	480331.25
44	−39337.85	78435.83	−132584.96	148875.10	62640.76	−102668.11	34248.77	−163884.39	120370.76	−167480.47
45	−149859.93	−450927.31	−190582.21	−304577.65	518698.02	−290286.33	−99243.76	456011.47	−452453.06	−215226.79
46	673225.31	−666698.79	−208826.16	−106040.35	−414930.10	−465479.75	−241108.20	205395.74	94595.47	748270.26
47	−14422.83	3443.44	−12514.42	4473.75	−9085.14	−13461.03	−15356.50	2302.46	−15916.63	−1990.02
48	125248.12	−118791.08	−441484.82	−392021.14	−364303.43	29567.90	−355094.67	223355.82	580375.15	207302.98
49	−407029.00	190949.90	9051.97	−420037.28	−125772.18	−153976.07	−142936.37	−404830.47	−337666.06	−341100.29
50	102910.65	117896.68	−38024.85	200059.24	−60930.23	−138461.90	−139465.84	140448.92	22218.10	29341.21
51	372151.08	84737.55	623034.24	588707.30	−529029.27	−929459.31	700157.87	452023.44	−589170.30	−894444.59
52	373164.04	345970.90	301788.53	−149585.71	335505.86	−232756.18	377549.67	−289643.13	−310274.89	−420895.83
53	−54656.51	−543460.87	539495.17	−78772.84	−821663.72	−438239.59	911850.14	−882639.24	−99874.69	−509755.03
54	−721838.41	−116651.21	651052.74	352012.43	−595011.16	−157195.48	599012.96	13674.82	521737.26	277145.34
55	143882.66	87215.01	333591.08	−223642.12	161395.86	−244651.45	394486.92	211492.53	−323156.92	449229.33
56	433505.05	75646.78	732202.55	364777.33	−68957.25	359756.55	−16340.51	−34084.89	−352709.63	−172184.60
57	605099.18	414858.12	261122.62	−738145.54	603443.59	339291.23	534729.66	−608095.73	−643011.65	−688647.56
58	−413291.44	125049.87	307907.16	−441273.65	−503021.82	303536.37	−226046.56	144233.66	−223430.08	229477.09
59	−104671.68	11359.52	−542620.78	402044.21	538861.91	−410799.05	−588977.28	792923.11	83299.60	−622032.49
60	−59267.29	35207.42	−109481.64	−96482.62	−30082.39	123246.07	112510.56	−139143.05	106959.63	128353.06
61	−92809.35	−110686.22	−111295.11	−11084.06	−15533.89	88608.74	27425.17	62856.41	−55354.77	−14329.52
62	106780.90	−241848.44	239608.16	421344.03	−391699.35	252502.01	−43217.95	380776.97	139700.87	257297.75
63	580256.70	50371.59	−209063.07	−457754.13	414918.68	461694.37	−80346.29	−542808.97	317070.75	−165747.00
64	206872.60	499536.00	−350492.17	45677.95	−26443.27	409204.83	−572230.01	−217312.31	328662.47	−306690.20
65	−790355.08	−290787.50	814994.85	389884.11	−147153.11	407482.82	553045.22	−250459.16	−730452.99	657247.08
66	350769.02	86938.28	127613.79	−517103.16	−33820.66	−25812.26	693899.00	420290.24	222234.60	330786.24
67	302411.07	461052.90	324728.56	−340870.89	−45703.65	−355540.70	489867.26	11633.12	−13269.53	285913.65
68	−204293.88	−236800.77	52579.10	179057.60	22206.64	−200874.92	30518.19	−43137.44	86586.95	178481.84
69	165778.26	−168260.72	147624.57	139372.76	−119922.69	−33739.81	141667.92	9523.02	−116416.85	−191685.69
70	448697.00	123296.05	719612.38	−212554.50	−209028.08	−894960.66	−800216.05	−760860.99	6275534.63	−557425.33
71	−16194.85	−47500.42	34503.19	45038.18	9758.04	−29209.50	44377.82	38237.82	10818.01	13331.89
72	−69904.75	−65664.93	−176014.46	−75771.87	9731.90	65108.85	185540.11	−93154.84	82264.17	133008.29
73	−99060.42	67366.31	16512.74	−139670.82	148623.92	237943.95	−197981.74	64667.23	209216.29	177335.46
74	452383.48	234957.98	31898.88	706713.96	−486564.01	186594.67	−95728.49	−587649.34	208496.41	−119483.66
75	273102.18	−344408.08	−276405.29	−17863.67	73180.88	420294.79	107386.11	33053.79	−59965.97	367892.34
76	436873.28	−366856.50	693223.07	569573.68	−213913.55	−334853.48	−194740.23	−588722.78	−654343.94	670076.89
77	40467.61	−45391.29	−5656.48	−54340.88	−34133.12	−42371.96	573.80	52818.90	−47588.53	−13906.75
78	−294064.46	−275189.08	439996.93	112717.57	260258.33	−460391.78	−206771.45	272364.11	−518664.46	187640.66
79	−249583.48	−72310.56	325609.37	−136795.95	131607.54	287073.26	340471.36	407758.63	−64055.85	−176968.45
80	131996.27	−132464.62	−16866.74	−252919.91	219264.19	182608.93	206120.02	290096.10	64977.84	217676.21

Table B.8: Initial guess random coefficients for the index i in the range $41, \ldots, 80$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{0,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
81	49916.14	32710.67	−99828.04	−177404.40	47817.30	193278.54	175419.85	−138550.48	−212364.36	128052.82
82	96604.27	−66748.53	104211.91	72633.90	88945.60	−134184.62	77365.19	138549.62	38811.94	−138038.44
83	427802.04	501324.91	356442.17	503791.20	−594218.36	−545358.83	−138183.58	−744808.07	318445.85	804216.10
84	45777.73	150554.55	391739.76	−531798.36	428196.22	−285664.06	−100686.61	−503651.72	25271.52	233024.03
85	−274348.34	380671.65	320951.57	189675.96	315648.05	688486.13	134037.14	−75258.26	184399.80	−387327.70
86	−195332.90	160026.78	123886.82	892653.97	−433664.83	−78948.09	−576850.46	131021.81	−570181.86	−758882.48
87	−10723.31	−108099.80	200827.67	115608.49	65503.77	−152647.86	132661.86	120599.51	6131.96	−213428.85
88	98812.03	17924.60	17221.20	20446.65	−45292.09	57733.33	−136343.23	−38396.85	100523.40	124387.10
89	−202849.50	−114063.62	−50772.27	279611.98	77806.35	−261315.78	232555.20	271801.31	−458174.36	12968.76
90	−15066.29	−27857.39	13021.92	23228.58	6791.56	−6307.85	−21276.69	−28418.77	−6724.56	−5428.44
91	4785.06	−17737.02	−93249.32	−26374.19	174113.14	−17185.92	−186125.04	−186168.86	−91603.47	166933.67
92	−286009.31	−777286.72	782705.67	622396.79	420670.23	236599.82	29996.84	−648944.01	−730216.39	553195.76
93	−652462.53	−321848.52	539806.25	557823.56	586412.63	−517756.56	7522.56	709526.62	298204.13	304212.53
94	−230673.11	−74107.82	−307016.09	−137437.65	483126.98	−437134.98	471199.39	13524.98	−356801.10	−57075.75
95	−133893.51	135961.56	112395.30	−81601.48	−171613.56	161709.91	−17596.80	97920.53	148256.45	76696.55
96	142019.86	839238.18	829575.94	493769.48	290654.21	−849518.51	10780.75	−87283.61	17666.31	348464.57
97	339246.18	−769343.11	−213687.79	773023.80	646841.60	−374921.40	−324942.52	684550.79	191540.68	9765.47
98	−357765.33	205031.02	−140475.58	87924.61	607798.39	480644.75	−750762.78	−522822.91	−125819.35	−38409.74
99	53061.50	−196652.72	618007.11	−342427.48	−138596.24	−92539.36	−586582.83	418981.42	−21989.81	14288.85
100	−138453.25	−159868.40	543132.76	505949.59	−19704.29	59383.04	−603551.50	148131.73	−424188.39	328926.47
101	165476.40	11070.83	−728785.79	164265.02	378197.82	351907.96	806103.77	177371.68	−130524.92	557230.57
102	−88370.55	−199620.48	−36516.31	187267.58	−131405.52	−230727.67	21982.87	146570.25	2241.00	40042.79
103	−341553.24	−570562.44	−349331.07	605016.54	532248.43	−478220.96	560777.59	468399.11	435647.10	−172990.90
104	56292.11	43541.85	48007.41	13616.97	15722.20	9538.99	−16318.22	−17312.49	1025.09	−56960.60
105	−91803.69	234824.59	175715.06	265259.14	−97671.94	−134259.70	−10080.50	−40889.42	−44449.48	198658.61
106	−41673.46	131241.27	47793.42	−118360.74	198407.22	64313.07	−187799.85	−86504.72	92264.59	166458.62
107	−25843.53	203264.79	−98004.11	−119750.11	−263618.30	−264050.09	107206.41	283211.08	121745.79	239420.39
108	−15365.88	−20602.92	−4040.98	438503.45	−267768.71	−99883.70	288569.46	−347453.17	−8438.51	10618.99
109	40764.13	−88943.44	−121763.97	134067.61	−66895.92	−16502.48	−115911.83	−138784.21	64886.73	−110231.90
110	126084.93	442975.12	350555.23	−83749.15	−79265.25	23555.83	−11432.79	63585.17	−451812.75	−475225.48
111	−128825.40	−33041.14	−106898.38	32261.51	133382.70	−240568.08	123170.09	−173306.76	15682.21	−122043.54
112	154157.50	−308405.97	279019.63	−183802.65	226953.21	−332723.83	−267595.59	−87193.88	54281.33	−253523.90
113	−468039.72	465300.52	−436589.24	422269.73	388614.49	117966.28	−256792.45	361692.77	−576376.79	213070.92
114	−236048.83	−370583.10	365429.34	371924.52	363161.96	787454.08	−227048.45	308014.89	46605.06	191954.39
115	−4695.39	104.27	−1065.72	−8299.59	−11306.67	5126.46	−5381.90	24.89	10538.88	11471.18
116	−6680.54	62193.08	−345195.68	444652.42	182531.64	−674052.68	−25684.63	15944.47	47209.30	−289664.94
117	−865.04	−24016.77	−21872.95	33976.41	5103.93	−32744.40	7302.84	30178.36	19149.69	30239.68
118	77921.44	−31806.66	40160.99	70401.40	31041.52	5006.07	−30353.65	−29513.11	46394.62	93146.34
119	8293.85	95383.28	−80849.55	50211.88	84709.03	10604.96	−22226.43	−106081.68	4980.42	−83010.48
120	74647.56	−71070.14	82131.95	−76272.28	−117224.91	−58987.68	−21272.13	22680.14	76462.05	4671.27

Table B.9: Initial guess random coefficients for the index i in the range $81, \ldots, 120$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{0,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
121	−70788.25	−324322.53	141579.54	−427560.54	415722.43	−17481.66	184309.80	−188395.20	91793.83	−20101.75
122	−76502.98	20260.61	21654.19	15094.86	24792.65	21986.80	62426.31	50507.92	−90019.10	−87393.30
123	−65396.04	−45529.21	261253.04	612834.07	105042.96	341141.63	−224972.49	791865.57	−295505.82	−642225.37
124	−98562.30	182480.68	−452486.64	309765.82	526940.66	−137898.98	271448.64	−102756.17	−235547.53	129339.77
125	−301736.50	307291.80	−4317.39	−37899.06	216314.51	108441.99	9323.41	115106.35	−244968.50	−61932.58
126	25999.68	17740.49	107353.07	−40076.42	−138268.77	−64520.17	−53415.24	156146.45	−140875.44	−165399.76
127	345691.81	−619211.06	434338.14	474774.79	340335.51	764343.33	−271946.74	602501.63	340779.97	−237639.89
128	−348331.74	57198.94	129727.99	171882.89	−60574.65	474405.04	191494.84	357232.22	256609.58	344736.45
129	−576169.59	−504690.92	−556858.56	424443.00	−54123.56	−111416.23	693194.42	−104092.10	−686476.62	−300639.69
130	−23562.44	−164.25	298230.38	−456935.73	67243.00	−179028.84	282332.90	−37087.35	−480481.55	89045.21
131	282291.73	294921.70	219107.50	111481.64	−36393.13	−280306.38	301004.87	−59003.46	126059.46	−207655.86
132	−301620.88	319388.42	127471.11	−368602.33	178958.43	−269341.88	325857.85	313428.82	−256225.60	327944.84
133	−453163.22	−508888.00	−4471.91	−129485.06	−182086.77	−63039.61	−233904.28	283899.70	−45311.07	294453.83
134	27860.22	−43933.27	−179682.86	89820.67	−164474.00	237958.74	109979.43	−73107.28	126018.85	−208942.76
135	−6068.77	168709.68	−290702.09	−138390.30	267451.70	−247220.18	85395.54	−135615.75	293392.42	290996.09
136	−449657.03	−913522.40	−121981.49	−931319.19	−33328.84	−460390.87	302206.99	317332.84	−200657.35	−429051.42
137	48814.80	146229.73	153125.58	−75898.28	9837.70	−58288.54	110239.04	136445.26	46897.20	123673.70
138	47366.33	155416.80	154694.69	110971.05	243270.01	−185451.09	−64394.16	175305.98	30651.22	130977.59
139	86066.16	−94904.01	7698.13	121186.52	59580.76	87441.54	33474.80	140962.13	−133603.12	167033.76
140	56823.65	109231.14	278351.38	296459.58	−298152.22	−318387.54	204858.64	−196513.00	288246.07	−297915.54
141	432597.09	−213901.44	−193544.86	−356541.60	122989.91	268048.25	−234560.86	297616.40	−111465.32	410043.59
142	−267656.45	−446633.31	−148736.35	−118649.39	−451903.28	−123224.88	315205.99	−243111.13	8680.87	287782.22
143	87623.92	197171.35	26712.02	88053.51	−192092.94	377616.87	−63490.62	−228925.83	−286027.61	−344897.86
144	−73833.66	26097.16	58399.14	130658.41	119941.36	−2404.96	75504.04	126003.53	67662.74	12532.59
145	−141785.88	−717589.62	−131706.58	52739.78	354535.90	51859.82	−623929.11	−382643.40	−275072.64	413820.71
146	−84014.57	−203858.03	243028.47	205044.61	−145206.26	160853.61	84661.54	282266.86	−111891.13	332927.16
147	191501.58	35723.06	11208.55	−39582.31	−76075.85	−48839.45	42974.59	−149063.68	155272.46	−43989.08
148	96180.63	228310.33	−8144.02	−170610.20	−58888.98	199084.46	−105722.94	147574.85	−148216.39	−195542.60
149	26004.14	19940.35	26948.09	−56305.18	27908.19	−5788.79	−17690.02	45467.56	−40956.29	−7889.73
150	−507341.08	319194.45	502273.54	−259417.30	199290.15	227599.83	−32893.26	192423.66	−368730.31	312115.90
151	−59087.83	199948.23	−272938.41	146515.59	37869.46	810531.63	−743536.66	713077.18	−558847.00	−646177.34
152	366006.70	204269.21	−98994.76	−164817.31	−78794.56	358843.12	286713.73	−250750.49	−287262.35	57978.18
153	−671144.07	−222585.30	−354271.65	−60838.01	−292899.63	−50662.77	−613720.87	−239303.22	640624.61	−764462.52
154	507891.55	18897.80	501389.22	147316.18	313671.01	20061.93	430024.17	59783.83	33399.15	202273.82
155	326873.19	−494158.85	163907.79	416375.70	452922.95	−166048.10	−335912.80	542689.95	−71979.76	266866.10
156	−860366.36	677034.49	−57104.11	−740375.57	729248.03	836061.49	29859.13	−350091.15	617199.25	180200.27
157	592372.22	838960.57	623697.52	−689334.73	321294.83	220287.76	7957.63	−527247.13	795647.45	−489096.05
158	649153.46	516714.47	−666387.19	−377077.06	288754.11	575924.09	−214655.30	142817.75	−313030.74	−174256.85
159	−134229.34	48523.93	54979.46	−150401.09	169711.52	219783.13	−358086.99	62170.34	−52166.30	120199.34
160	−449828.19	−414252.98	505495.74	−15525.71	398624.25	−412898.26	−555483.37	204546.14	432391.78	279146.93

Table B.10: Initial guess random coefficients for the index i in the range $121, \ldots, 160$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{0,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
161	196223.02	−238618.77	−870808.23	69115.07	481078.23	−173332.76	174612.31	−349000.38	86126.07	−664970.24
162	−872690.77	−669876.71	−567574.05	−207529.08	419268.74	503497.89	−524488.93	715625.19	−120672.56	634326.02
163	271942.07	−239150.19	109045.47	−363154.06	54135.86	−125135.49	−253806.17	399061.77	495258.40	19404.68
164	483869.79	445313.48	−311977.27	573733.65	378534.98	−414911.02	−236491.36	−327560.51	375590.06	13695.84
165	113242.49	153926.04	18365.51	−330033.17	42876.46	143597.76	314065.13	322431.37	−252635.71	−333633.54
166	−136871.51	−52544.59	−871417.62	714097.74	−9310.27	759794.41	−477026.07	−391405.38	−74407.98	−789463.15
167	157678.86	353465.26	120869.51	144824.55	−300294.21	−180972.08	−256124.65	−308911.04	−420203.83	476470.46
168	−687474.06	−436258.19	−770134.13	−208183.72	422403.83	−233499.65	−82047.87	269566.76	144585.43	−268067.88
169	−62363.26	−603728.82	264063.18	−476613.00	−472376.75	−448288.01	−91680.40	465034.12	−32564.67	287767.16
170	45209.67	−57082.56	−53060.42	29751.30	−54595.35	54215.20	49.48	58223.64	−20691.79	−4511.99
171	181554.04	4627.43	154576.94	106342.82	154936.41	267188.39	−203102.18	−359143.40	6042.51	58304.20
172	−23956.01	573510.01	−154507.52	117378.56	236133.76	298490.21	−217806.62	−407264.19	−425601.16	−477427.21
173	−286657.52	16091.01	167608.26	43446.92	−383977.32	337010.61	285862.66	361556.78	111755.23	−290123.34
174	−39963.86	46871.60	69212.58	−250905.68	−143794.88	83466.52	−132064.59	5736.51	−197041.96	146050.32
175	−13102.64	−20711.81	6531.39	−16544.59	18951.44	−3598.16	23799.76	−26385.75	1118.24	−23469.62
176	−128532.06	96485.40	−100737.01	−16854.62	52474.33	−51761.91	−32902.62	70611.48	67265.56	−42170.89
177	−21719.57	−41559.62	−17965.02	32227.67	35592.88	−17863.93	−11809.65	53907.64	−4124.61	−4440.76
178	−75640.24	−136077.77	−1104.33	38639.13	−152236.29	−146730.45	6256.26	120257.25	−62819.96	−107277.54
179	56840.04	−90045.94	−146057.99	−54626.85	−94674.95	11832.93	−160503.96	−212032.93	−130438.23	−5680.15
180	−295945.01	−321083.02	232000.46	−138096.82	−277735.67	274310.97	57067.11	−32739.03	−63438.97	−153688.29
181	−285976.00	−301464.98	−287660.38	254687.87	−281377.53	15308.57	−99742.68	−179226.35	−235360.84	3624.10
182	−360250.20	160832.69	−381276.70	508298.98	−414450.87	430970.88	−292436.47	−526437.77	−437489.44	111730.23
183	−264563.22	452813.17	−202396.48	130506.24	−116302.76	274642.52	13404.25	264502.81	453634.08	−431689.38
184	−182893.52	−87102.94	−133452.84	−443521.79	93366.63	226791.78	−24405.70	230871.41	438997.15	−224436.22
185	−364948.49	−132553.95	366401.40	−442266.14	26376.62	−345442.72	−615817.54	614238.16	211439.98	230952.72
186	−568650.41	−456900.30	423679.71	−401055.45	−490822.02	646367.29	568007.77	642199.98	−357643.32	−439368.80
187	90567.13	799486.69	218335.87	−519112.84	−35600.60	263011.03	−757933.89	36831.96	−216386.51	−133780.56
188	−57251.75	59469.51	51508.58	−16364.61	46150.17	4402.48	58202.70	59442.23	30034.89	39828.92
189	−487937.97	−107516.18	371867.67	74116.86	501319.06	212038.70	−396262.29	245344.21	−574373.51	−452462.70
190	−91322.65	511192.73	−502599.20	−32035.74	−436842.31	−539495.04	821158.05	682993.69	−166176.93	265874.07
191	−180397.53	55137.27	−49105.01	66772.83	−14575.36	132246.15	27176.65	−173184.79	−131675.46	−19185.49
192	483037.52	266273.50	−480244.50	−661587.74	427974.31	−839671.78	−507926.69	660391.28	−274215.26	898954.13
193	−484105.24	−196052.93	−139258.70	−171523.15	180637.58	456816.43	402283.57	−431909.36	263410.28	−322376.54
194	−38698.70	−130240.84	52438.27	−278533.98	−419232.64	−43644.84	−409014.81	451778.78	394765.18	−199647.46
195	141928.57	50720.05	−180285.24	−383866.88	−286770.72	60322.96	411251.29	−264186.94	185595.75	−492612.84
196	312157.02	27261.58	328779.75	314419.31	116053.49	340266.36	232041.22	168314.15	216837.37	−201195.04
197	288340.98	161945.23	−449583.72	107467.75	−404431.34	187063.75	194977.72	−242461.73	−235578.68	−455208.82
198	−604580.85	−158786.73	241900.94	366884.35	−453473.43	−677686.47	−361000.88	384619.47	345551.34	−500750.10
199	7927.29	−26668.82	10940.82	−19751.11	−19728.99	−7981.99	−7286.34	14734.73	11000.04	12793.26
200	310176.01	58246.42	−328350.94	162187.26	−80371.70	133643.17	133487.18	283710.85	−128123.31	8375.83

Table B.11: Initial guess random coefficients for the index i in the range $161, \ldots, 200$, for Experiment 7.9.5

MA prob. no.	$u^i_{0,0}$	$u^i_{1,0}$	$u^i_{0,1}$	$u^i_{1,1}$	$u^i_{2,0}$	$u^i_{0,2}$	$u^i_{2,1}$	$u^i_{1,2}$	$u^i_{3,0}$	$u^i_{0,3}$
201	140909.46	352024.45	−623051.43	183334.03	165822.86	−586224.63	426680.95	−190358.88	589525.22	199268.74
202	−92871.13	214089.53	16260.06	206285.76	−323311.59	−56270.46	−281372.22	251884.89	−124336.60	244229.53
203	410786.23	−352119.27	455833.03	−216496.45	273940.70	623619.42	−207125.95	−525517.99	711768.73	463786.40
204	484947.89	−499415.72	704118.61	445763.80	−138307.37	−558428.55	602505.62	−5649.03	352117.51	−738828.15
205	44095.11	150734.38	−159495.40	138818.00	130702.68	25713.16	10270.63	9422.65	101025.85	−23538.15
206	67038.77	190226.36	−113047.96	−61422.80	−167947.60	−63737.44	−66631.88	84405.23	86982.25	48939.85
207	−51559.80	395478.70	−188279.51	33813.15	−25833.22	−120322.50	225656.45	−458842.78	550848.88	540112.33
208	−515055.02	128748.60	−23051.95	379878.96	283465.57	118739.87	−381371.32	74266.80	121626.31	460890.77
209	−157978.28	84325.11	4148.11	−79280.31	−390075.93	−80880.19	773797.50	127906.45	−292875.81	622494.32
210	42420.31	47623.40	27715.47	−26111.04	−4198.54	42413.37	−44026.78	−47019.37	−2728.34	9672.08

Table B.12: Initial guess random coefficients for the index i in the range $201, \ldots, 210$, for Experiment 7.9.5

MA prob. no.	$f^i_{0,0}$	$f^i_{1,0}$	$f^i_{0,1}$	$f^i_{1,1}$	$f^i_{2,0}$	$f^i_{0,2}$	$f^i_{2,1}$	$f^i_{1,2}$	$f^i_{3,0}$	$f^i_{0,3}$
1	158.62	35.50	3.80	409.75	295.53	403.96	409.99	381.37	184.37	437.33
2	92.44	50.46	10.27	110.45	209.24	104.88	358.64	43.13	338.99	313.79
3	574.49	609.31	117.87	492.95	831.64	371.98	329.49	326.94	68.19	175.55
4	172.32	153.52	486.59	566.47	331.72	222.86	468.59	23.25	21.54	498.48
5	206.58	20.71	612.01	702.98	174.80	464.03	20.14	738.28	269.29	34.58
6	531.52	654.23	134.30	430.33	762.73	792.70	706.13	636.10	930.37	242.83
7	105.67	106.97	56.87	10.53	79.13	4.28	25.98	42.26	93.07	71.43
8	24.26	99.05	63.35	142.29	103.21	124.51	12.31	134.03	116.17	9.12
9	169.20	35.80	84.64	48.74	46.73	16.11	150.69	32.00	162.04	70.81
10	119.49	575.52	538.94	492.63	491.43	528.00	139.01	1.96	38.66	496.90
11	83.24	212.05	290.74	302.28	130.08	346.61	296.83	7.21	270.79	419.89
12	125.76	366.95	630.61	305.79	643.76	1.54	590.34	764.57	469.77	732.51
13	298.46	762.57	843.07	856.58	654.49	393.79	885.11	852.44	197.88	766.44
14	27.75	231.32	406.63	135.85	376.73	626.87	491.82	293.28	573.23	50.14
15	19.98	73.55	260.82	219.82	47.45	142.39	86.53	245.14	197.75	66.06
16	500.66	588.30	668.38	205.46	648.69	159.81	747.28	408.94	657.84	123.36
17	430.62	136.09	534.80	409.38	433.91	440.07	332.91	548.38	24.58	473.36
18	38.07	376.88	185.48	304.05	220.96	326.20	90.73	97.97	394.19	321.41
19	218.69	774.58	531.47	741.75	677.31	529.50	632.27	336.45	208.36	630.89
20	89.73	274.96	184.78	269.57	60.28	225.06	162.75	64.31	98.33	139.16
21	244.30	119.78	21.56	140.30	2.96	73.33	254.48	185.12	176.65	240.07
22	205.07	638.35	626.12	677.74	47.94	409.09	536.15	661.31	354.38	226.01
23	486.23	235.19	735.54	130.77	157.35	454.23	243.20	124.25	15.72	277.07
24	306.30	59.46	711.32	121.68	207.63	148.93	213.61	509.54	294.22	639.03
25	178.07	83.53	184.53	171.63	57.55	94.61	205.67	78.53	129.94	11.60
26	799.91	665.84	888.81	514.93	693.47	589.14	709.94	883.22	54.99	39.40
27	49.60	52.36	141.62	158.68	58.19	138.05	51.97	172.49	142.90	90.29
28	136.42	402.63	126.85	255.23	150.85	149.73	318.12	389.50	432.10	149.36
29	498.92	193.97	652.77	669.44	60.63	345.17	245.41	690.82	520.84	734.12
30	245.00	394.67	6.54	122.20	125.58	278.22	146.48	269.93	425.11	260.21
31	674.61	573.22	547.21	177.02	393.38	328.70	757.56	216.11	528.31	730.97
32	271.22	207.36	237.13	398.25	175.17	14.07	479.97	136.89	542.80	404.63
33	83.66	79.01	120.34	119.64	39.22	82.12	16.53	158.40	74.42	9.47
34	199.38	45.35	521.32	242.58	817.67	290.51	512.81	706.80	424.11	104.34
35	49.13	107.37	64.16	73.26	42.59	47.69	91.83	112.46	80.55	12.96
36	298.78	12.46	27.50	139.04	164.62	75.87	273.42	289.33	69.03	127.82
37	159.64	823.97	619.15	825.23	634.12	500.89	737.50	90.81	560.78	201.06
38	609.19	198.99	18.13	545.37	267.57	21.70	839.77	360.07	272.78	90.16
39	14.54	28.23	102.43	40.53	42.86	115.71	20.52	81.90	62.86	25.28
40	292.52	246.78	409.85	3.13	696.81	922.76	636.00	333.47	277.52	825.50
41	615.71	266.90	753.81	144.76	206.52	597.40	764.22	651.67	774.53	88.77
42	90.22	138.66	26.50	4.41	129.99	108.78	98.74	189.19	135.35	56.24
43	187.92	92.01	521.46	68.30	408.16	141.20	36.37	601.99	159.41	251.86
44	165.15	83.84	537.77	63.44	176.77	165.17	77.31	172.55	456.77	222.44
45	42.82	705.87	431.02	232.74	376.19	703.29	523.65	567.69	315.08	174.45
46	567.60	94.56	387.32	186.71	534.04	330.06	566.30	499.97	234.22	2.75
47	119.98	69.28	20.19	101.29	111.93	72.50	54.13	111.94	140.20	141.40
48	595.26	280.67	730.99	455.99	145.74	255.18	675.70	749.74	273.47	551.97
49	51.52	318.82	323.96	44.53	48.87	136.70	272.45	17.66	222.65	300.11
50	692.82	693.81	47.17	738.71	550.58	0.29	343.28	712.64	421.78	731.26
51	145.26	124.65	159.08	107.77	104.69	115.29	126.21	92.33	5.52	67.70
52	336.15	208.34	135.13	289.18	197.25	331.42	232.98	125.81	339.58	84.27
53	25.27	39.32	1.40	76.55	79.33	40.43	53.58	30.42	72.68	5.66
54	48.35	668.37	620.59	445.19	207.39	747.86	732.46	762.93	118.13	231.16
55	70.70	378.17	457.44	344.19	151.65	183.44	76.37	140.93	299.65	88.17
56	7.52	571.96	903.98	228.10	743.04	496.78	484.58	619.51	674.27	509.26
57	367.34	51.21	272.02	273.81	48.69	125.64	39.31	74.41	395.66	101.07
58	428.39	273.94	110.62	285.70	153.59	257.28	681.85	466.55	249.08	339.02
59	363.49	295.00	478.85	262.36	341.09	334.73	74.85	140.83	170.37	190.47
60	647.84	598.41	74.55	473.70	288.19	23.93	464.60	612.82	247.65	249.42

Table B.13: Right-hand side random coefficients for the index i in the range $1, \ldots, 60$, for Experiment 7.9.5

MA prob. no.	$f^i_{0,0}$	$f^i_{1,0}$	$f^i_{0,1}$	$f^i_{1,1}$	$f^i_{2,0}$	$f^i_{0,2}$	$f^i_{2,1}$	$f^i_{1,2}$	$f^i_{3,0}$	$f^i_{0,3}$
61	212.46	109.35	143.92	28.87	48.89	144.50	146.30	23.98	15.18	189.11
62	178.21	303.55	137.61	81.34	34.30	112.52	286.52	18.15	132.51	42.83
63	182.64	33.50	581.15	945.55	912.11	962.76	307.66	297.89	351.51	366.23
64	187.77	239.26	204.38	51.14	18.33	567.16	284.37	43.00	578.61	225.67
65	132.89	779.06	39.21	582.13	254.47	320.85	120.15	129.26	282.03	439.34
66	147.06	443.70	170.50	572.41	376.39	186.01	194.90	159.03	490.38	400.28
67	300.14	373.00	462.67	153.81	272.78	366.47	408.39	322.45	37.53	64.54
68	78.75	178.62	62.43	89.91	175.51	29.90	107.06	167.47	105.98	46.81
69	24.22	432.57	226.72	413.89	73.29	401.22	283.02	610.57	87.12	497.76
70	452.14	186.46	443.59	351.54	435.02	221.05	202.69	253.50	219.14	280.36
71	51.05	144.64	41.75	74.54	129.76	106.89	43.51	75.08	106.85	137.48
72	16.25	4.22	148.22	269.86	228.68	207.10	253.90	89.43	195.99	275.79
73	275.22	769.24	17.44	497.22	830.51	347.99	282.02	467.71	309.33	184.53
74	86.77	9.15	142.36	6.80	11.64	32.13	154.36	102.79	88.50	114.03
75	181.82	74.66	110.19	734.01	650.24	298.26	243.37	824.90	593.03	799.64
76	332.34	319.86	198.86	369.94	183.95	128.50	243.53	20.73	84.46	306.97
77	40.73	119.50	86.25	116.83	53.12	13.76	160.83	206.55	164.39	171.69
78	72.56	7.32	98.58	334.57	496.34	139.33	45.20	627.67	123.72	499.23
79	442.62	75.63	838.20	878.81	709.83	872.15	101.57	500.45	82.87	653.13
80	13.84	4.32	56.46	81.88	83.57	133.00	148.63	52.42	131.87	127.08
81	260.02	134.46	381.47	112.35	477.18	487.82	151.19	317.05	58.96	444.79
82	164.65	29.18	88.95	177.66	25.02	177.84	52.00	107.43	168.01	178.93
83	444.20	79.86	760.89	279.42	237.50	219.43	24.78	226.72	360.88	230.58
84	170.48	177.00	317.41	22.15	550.45	296.79	36.34	343.73	574.85	161.64
85	101.97	385.60	311.60	212.03	306.95	67.49	84.00	65.69	433.05	178.81
86	221.44	72.62	232.10	3.30	542.27	237.56	463.90	244.97	534.76	183.02
87	225.52	44.80	101.36	240.83	101.76	172.76	42.31	9.94	323.08	133.07
88	572.87	133.45	916.97	759.53	189.85	266.82	829.90	652.07	173.08	579.68
89	74.12	10.46	259.58	135.30	111.31	313.21	430.06	15.30	339.10	358.25
90	7.28	65.74	74.54	48.40	129.58	115.55	49.21	17.73	93.04	161.81
91	134.99	149.30	81.25	86.08	117.53	154.58	174.67	92.17	14.12	8.54
92	100.28	527.50	411.44	95.90	13.49	360.63	422.43	56.69	577.02	256.55
93	7.90	617.00	804.74	54.22	39.26	129.34	850.69	134.39	211.90	812.29
94	78.94	111.57	164.74	82.00	130.90	101.07	207.90	387.96	159.75	69.16
95	506.95	547.79	297.17	512.59	338.67	420.81	215.69	106.93	352.05	383.96
96	44.51	410.15	340.40	438.26	15.01	108.96	596.75	4.14	637.89	23.41
97	218.52	174.74	410.91	58.07	107.20	462.13	403.17	79.76	4.43	259.53
98	59.44	81.89	91.40	154.92	62.90	17.60	149.46	8.28	126.80	47.75
99	743.92	151.15	225.71	520.23	720.77	486.79	247.20	409.25	464.81	774.93
100	40.91	286.49	346.15	392.69	95.86	374.86	173.07	318.36	327.01	491.79
101	542.67	200.32	402.23	421.01	707.71	478.48	86.57	281.35	730.96	607.39
102	561.42	835.82	152.60	313.13	573.34	739.17	340.51	256.13	6.05	53.73
103	560.69	542.43	105.27	493.43	18.26	100.51	197.07	365.39	600.24	485.03
104	72.47	296.36	13.22	437.78	90.44	347.01	70.42	62.37	310.26	111.17
105	374.79	75.80	390.87	408.04	515.72	5.62	222.46	562.26	410.45	331.43
106	379.43	10.13	59.18	395.75	161.25	33.92	177.82	607.64	658.78	580.92
107	437.16	923.99	197.46	718.25	744.14	263.04	440.87	444.30	43.57	13.82
108	82.82	454.28	563.81	277.50	773.07	596.50	725.67	528.21	633.85	580.79
109	486.76	244.37	554.88	257.45	640.88	210.21	200.11	28.45	320.97	242.11
110	398.71	203.55	127.16	212.30	441.75	186.76	164.44	80.62	289.17	150.01
111	382.22	68.91	730.09	522.59	239.45	492.02	315.90	456.11	420.94	248.48
112	16.54	174.01	157.09	253.13	58.82	229.73	6.47	102.43	253.67	115.33
113	716.57	42.12	35.08	133.37	396.99	312.93	20.64	425.05	500.15	561.61
114	217.05	367.92	237.57	227.82	155.00	103.56	150.06	125.81	276.48	367.85
115	593.61	919.39	627.92	338.30	544.23	575.49	619.43	498.13	196.16	378.47
116	440.93	328.24	348.19	192.01	342.49	40.74	534.40	341.38	226.05	133.69
117	271.92	445.03	26.70	456.22	418.62	310.08	288.03	318.82	344.85	300.31
118	827.86	496.15	711.56	657.01	542.36	63.26	32.93	316.80	71.95	41.93
119	584.83	607.98	67.05	182.26	289.81	248.40	186.24	56.82	189.10	614.11
120	787.20	517.86	283.04	447.79	221.93	1.81	12.73	734.64	855.64	456.73

Table B.14: Right-hand side random coefficients for the index i in the range $61, \ldots, 120$, for Experiment 7.9.5

MA prob. no.	$f^i_{0,0}$	$f^i_{1,0}$	$f^i_{0,1}$	$f^i_{1,1}$	$f^i_{2,0}$	$f^i_{0,2}$	$f^i_{2,1}$	$f^i_{1,2}$	$f^i_{3,0}$	$f^i_{0,3}$
121	205.63	230.48	253.16	82.23	668.77	358.61	118.17	556.18	307.85	216.70
122	96.91	165.53	139.37	373.33	141.53	344.72	129.18	133.88	351.42	109.11
123	193.79	16.29	121.26	129.78	78.88	141.62	111.20	114.83	18.35	26.40
124	128.17	117.67	113.32	94.77	49.17	2.37	44.64	117.35	32.94	61.24
125	31.95	19.69	16.09	95.16	6.09	48.67	86.22	106.17	69.77	20.86
126	173.96	541.56	139.31	488.78	388.98	456.48	201.66	162.68	522.02	8.68
127	427.75	602.08	22.77	510.16	315.07	814.04	364.35	92.01	439.49	923.26
128	453.32	583.86	314.83	502.75	231.95	578.31	168.01	169.15	286.33	144.71
129	198.86	576.62	824.99	847.22	6.25	658.43	260.12	387.35	75.09	847.30
130	139.96	169.32	56.47	111.74	153.48	176.59	111.77	55.21	153.44	179.81
131	440.16	394.52	63.98	378.08	67.74	349.48	401.11	268.36	338.33	368.32
132	499.91	428.29	557.33	630.27	355.04	233.24	551.36	216.98	624.86	734.45
133	190.41	372.51	153.77	616.50	86.43	622.69	346.06	362.97	362.23	261.00
134	57.73	118.96	55.71	93.51	44.39	19.39	37.27	5.39	71.12	58.97
135	131.20	142.83	316.02	252.47	66.50	344.13	223.29	42.56	430.32	413.05
136	569.79	303.81	115.20	19.32	928.94	505.32	889.86	940.44	904.19	830.54
137	307.46	94.43	182.07	155.23	481.44	443.70	663.77	490.95	161.05	230.10
138	201.24	73.63	35.21	64.90	51.81	258.91	158.83	115.35	77.08	264.34
139	396.15	270.98	567.83	449.19	32.25	7.78	350.22	495.29	280.17	380.98
140	3.09	90.92	624.51	136.63	55.07	543.28	288.86	23.50	514.76	462.58
141	139.31	146.00	42.65	122.91	48.26	91.70	180.21	78.39	33.64	122.54
142	50.38	17.76	45.06	46.73	19.98	94.22	99.10	51.70	48.07	20.06
143	71.54	30.26	54.74	55.73	81.23	59.01	73.51	98.03	32.34	14.39
144	81.13	88.15	5.19	44.93	8.23	16.42	43.42	5.58	175.72	112.77
145	90.79	784.18	380.46	730.43	733.42	611.19	520.23	535.89	343.92	127.03
146	263.63	32.27	297.92	90.36	117.94	240.60	89.38	159.16	184.99	80.81
147	166.62	95.57	23.46	47.09	116.82	15.35	157.76	60.76	65.12	213.40
148	56.79	38.02	73.26	25.56	32.32	14.84	92.90	86.44	85.12	101.55
149	367.18	562.90	281.80	195.30	617.28	270.11	137.77	120.09	19.78	3.93
150	433.62	206.50	106.91	418.22	190.80	227.76	209.67	220.65	167.21	584.41
151	65.14	47.31	214.36	286.19	125.16	200.59	196.07	181.62	164.53	80.11
152	5.50	332.08	497.81	769.72	751.48	758.51	139.70	671.99	589.87	375.00
153	92.05	106.40	250.29	231.75	273.82	195.26	53.83	192.63	324.59	267.21
154	823.40	594.75	25.18	563.40	533.71	213.82	723.41	727.76	389.83	401.46
155	130.47	374.50	214.12	392.47	82.97	204.67	315.87	417.00	98.41	454.84
156	499.15	356.38	41.96	301.51	84.62	575.43	206.14	374.60	188.99	297.16
157	56.64	88.33	12.55	42.85	68.50	28.38	92.25	7.24	122.78	82.91
158	127.06	169.58	648.71	143.24	344.22	310.44	277.96	140.70	233.11	321.46
159	170.65	80.84	151.11	95.85	25.58	75.93	6.59	80.47	205.83	81.33
160	453.83	472.61	168.51	345.65	286.35	5.87	395.97	483.37	289.55	377.65
161	276.23	455.96	854.12	256.61	552.01	499.59	815.76	370.92	278.98	195.87
162	311.93	59.74	225.36	181.89	143.39	305.46	234.09	256.82	201.58	343.79
163	172.18	629.94	250.04	745.83	613.46	540.77	362.30	517.60	702.20	117.88
164	138.55	273.05	309.78	229.33	333.08	105.24	739.03	277.65	267.61	119.27
165	201.40	255.06	309.73	372.10	142.29	223.79	133.62	49.89	86.58	229.26
166	4.50	373.84	412.87	332.30	178.98	101.72	642.52	50.39	167.64	387.88
167	17.07	92.36	89.40	113.71	63.63	92.07	11.65	141.16	135.55	79.23
168	115.27	922.93	557.14	577.17	851.21	513.70	603.63	752.31	193.15	283.96
169	373.35	630.46	408.01	162.30	55.63	648.54	929.26	773.87	275.44	895.91
170	32.18	137.87	92.80	189.41	125.05	196.83	23.47	193.89	93.90	169.43
171	57.14	120.88	171.82	93.90	36.93	137.08	57.00	243.03	157.27	363.56
172	323.98	445.47	38.12	26.49	198.01	595.68	655.05	637.19	43.89	785.15
173	551.08	473.88	234.96	345.69	220.30	152.22	176.74	376.45	455.84	574.21
174	582.35	726.62	361.41	540.24	129.77	502.32	553.35	246.20	587.03	523.52
175	64.98	66.26	5.33	81.50	167.63	232.89	24.87	25.77	162.89	99.65
176	94.00	33.09	168.12	29.26	65.06	74.03	25.76	97.69	179.36	75.45
177	933.43	717.72	875.16	779.99	542.79	81.38	197.73	20.46	741.17	624.55
178	83.22	47.02	143.04	107.45	194.10	79.64	180.29	19.86	42.11	42.68
179	3.71	580.69	483.87	491.38	730.29	481.82	252.99	650.62	406.35	227.74
180	25.69	355.44	372.46	470.72	686.37	170.52	598.39	206.42	93.03	580.28

Table B.15: Right-hand side random coefficients for the index i in the range $121, \ldots, 180$, for Experiment 7.9.5

MA prob. no.	$f^i_{0,0}$	$f^i_{1,0}$	$f^i_{0,1}$	$f^i_{1,1}$	$f^i_{2,0}$	$f^i_{0,2}$	$f^i_{2,1}$	$f^i_{1,2}$	$f^i_{3,0}$	$f^i_{0,3}$
181	449.98	428.95	457.80	46.94	601.79	98.36	235.79	595.06	431.05	249.31
182	245.05	268.32	52.89	254.54	66.46	407.99	35.79	422.65	319.85	56.98
183	481.39	110.93	406.81	228.16	289.45	257.08	422.32	255.85	10.84	175.40
184	181.37	46.93	155.54	68.93	129.16	42.05	204.50	216.12	59.92	145.74
185	121.80	242.44	564.01	286.81	561.79	335.97	334.36	679.10	176.68	359.01
186	652.81	650.55	675.16	191.79	74.42	613.07	239.07	504.61	823.88	847.22
187	157.52	184.61	88.07	6.77	30.74	96.35	150.99	225.15	249.65	238.31
188	318.99	732.25	286.43	208.31	65.59	529.84	543.65	750.98	404.92	804.27
189	207.94	433.00	448.16	566.05	132.58	131.34	455.96	339.36	302.22	177.06
190	242.49	254.77	481.76	171.03	114.35	358.00	303.11	227.35	370.39	194.98
191	35.25	223.82	241.57	210.20	208.80	86.38	165.28	55.65	17.65	203.51
192	151.82	213.31	123.64	228.96	228.12	101.34	99.02	117.96	6.71	78.93
193	137.13	43.74	43.93	170.34	17.55	152.98	162.22	13.68	202.20	80.30
194	161.91	64.59	142.70	323.16	40.42	320.29	75.86	258.28	85.19	204.52
195	555.40	242.77	823.12	608.42	340.58	330.92	689.51	246.53	155.50	406.02
196	332.79	442.95	367.42	266.12	670.34	695.30	53.37	239.41	689.58	379.45
197	511.22	193.50	652.55	524.73	25.68	17.85	41.60	273.90	88.90	32.40
198	116.42	68.59	34.34	102.13	57.04	124.60	143.90	34.66	60.08	27.46
199	618.35	633.87	68.85	415.38	174.41	140.26	106.69	24.63	114.45	262.58
200	502.04	383.57	216.78	389.88	130.69	519.94	52.91	287.37	553.92	205.00
201	233.68	227.09	120.44	120.12	150.70	183.21	239.89	69.89	39.78	245.50
202	509.72	120.08	331.17	83.01	663.01	404.67	258.08	354.38	633.04	305.31
203	155.99	14.49	244.01	562.69	584.80	516.07	492.48	99.09	293.39	134.70
204	1.61	53.80	130.76	397.45	318.99	114.68	812.01	536.17	28.30	766.74
205	17.04	408.95	96.24	299.43	4.72	18.80	94.33	41.40	382.45	352.82
206	100.99	58.77	95.59	74.98	141.17	194.98	167.16	223.25	12.06	140.47
207	271.66	781.21	394.26	104.94	0.52	855.59	922.36	407.78	322.48	440.75
208	660.53	374.24	459.23	131.22	480.20	497.40	652.55	540.89	570.28	176.71
209	41.13	173.65	256.05	14.40	463.72	492.77	172.96	37.74	167.44	587.77
210	66.97	121.36	163.48	96.94	132.08	69.13	118.63	141.91	127.23	37.78

Table B.16: Right-hand side random coefficients for the index i in the range $181, \ldots, 210$, for Experiment 7.9.5

References

[1] Yves Achdou, Francisco J. Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Partial differential equation models in macroeconomics. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2028):20130397, 19, 2014.

[2] Martin S. Alnæs, Jan Blechta, Johan Hake, August Johansson, Benjamin Kehlet, Anders Logg, Chris Richardson, Johannes Ring, Marie E. Rognes, and Garth N. Wells. The fenics project version 1.5. *Archive of Numerical Software*, 3(100), 2015.

[3] G. Awanou. Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: mixed methods (2014).

[4] G. Awanou and H. Li. Error analysis of a mixed finite element method for the Monge–Ampère equation. *Int. J. Num. Analysis and Modeling*, 11:745–761, 2014.

[5] L. Babuška and M. Suri. The optimal convergence rate of the p-version of the finite element method. *SIAM Journal on Numerical Analysis*, 24(4):750–776, 1987.

[6] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.9, Argonne National Laboratory, 2018.

[7] Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

[8] G. Barles. Convergence of numerical schemes for degenerate parabolic equations arising in finance theory. In *Numerical methods in finance*, volume 13 of *Publ. Newton Inst.*, pages 1–21. Cambridge Univ. Press, Cambridge, 1997.

[9] G. Barles and E. R. Jakobsen. On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations. *M2AN Math. Model. Numer. Anal.*, 36(1):33–54, 2002.

[10] G. Barles and E. R. Jakobsen. Error bounds for monotone approximation schemes for parabolic Hamilton–Jacobi–Bellman equations. *Math. Comp.*, 76(260):1861–1893, 2007.

[11] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. In *29th IEEE Conference on Decision and Control*, pages 2347–2349 vol.4, Dec 1990.

[12] J. W. Barrett and C. M. Elliott. Fixed mesh finite element approximations to a free boundary problem for an elliptic equation with an oblique derivative boundary condition. *Comput. Math. Appl.*, 11(4):335–345, 1985.

[13] R. Bellman. *Introduction to matrix analysis*, volume 19 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. Reprint of the second (1970) edition, With a foreword by Gene Golub.

[14] J-D. Benamou, B. D. Froese, and A. M. Oberman. Two numerical methods for the elliptic Monge-Ampère equation. *M2AN Math. Model. Numer. Anal.*, 44(4):737–758, 2010.

[15] J-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampère equation. *J. Comput. Phys.*, 260:107–126, 2014.

[16] C. Bernardi. Optimal finite-element interpolation on curved domains. *SIAM J. Numer. Anal.*, 26(5):1212–1240, 1989.

[17] Daniele Boffi, Franco Brezzi, and Michel Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2013.

[18] O. Bokanowski, A. Picarelli, and C. Reisinger. Stability and convergence of second order backward differentiation schemes for parabolic Hamilton–Jacobi–Bellman equations. *arXiv preprint arXiv:1802.07146*, 2018.

[19] J. F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numer. Anal.*, 41(3):1008–1021, 2003.

[20] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring. I. *Math. Comp.*, 47(175):103–134, 1986.

[21] S. C. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[22] S. C. Brenner and M. Neilan. Finite element approximations of the three dimensional Monge-Ampère equation. *ESAIM Math. Model. Numer. Anal.*, 46(5):979–1001, 2012.

[23] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[24] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[25] S. C. Brenner and L-Y. Sung. C^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. *J. Sci. Comput.*, 22/23:83–118, 2005.

[26] C. J. Budd, R. D. Russell, and E. Walsh. The geometry of r-adaptive meshes generated using optimal transport methods. *J. Comput. Phys.*, 282:113–137, 2015.

[27] L. A. Caffarelli. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2)*, 144(3):453–496, 1996.

[28] L. A. Caffarelli and X. Cabré. *Fully nonlinear elliptic equations*, volume 43 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1995.

[29] L. A. Caffarelli, L. Nirenberg, and J. Spruck. The Dirichlet problem for nonlinear second-order elliptic equations. I. Monge-Ampère equation. *Comm. Pure Appl. Math.*, 37(3):369–402, 1984.

[30] S. Čanič, B. L. Keyfitz, and G. M. Lieberman. A proof of existence of perturbed steady transonic shocks via a free boundary problem. *Comm. Pure Appl. Math.*, 53(4):484–511, 2000.

[31] Pierre Cardaliaguet. Notes on mean field games. Technical report, Technical report, 2010.

[32] Erin Carson and Nicholas J Higham. A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. *SIAM Journal on Scientific Computing*, 39(6):A2834–A2856, 2017.

[33] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

[34] P. G. Ciarlet and P.-A. Raviart. General Lagrange and Hermite interpolation in $\mathbf{R}^n$ with applications to finite element methods. *Arch. Rational Mech. Anal.*, 46:177–199, 1972.

[35] P. G. Ciarlet and P.-A. Raviart. Interpolation theory over curved elements, with applications to finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 1:217–249, 1972.

[36] Ph. Clément. Approximation by finite element functions using local regularization. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér.*, 9(R-2):77–84, 1975.

[37] M. G. Crandall, H. Ishii, and P-L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc. (N.S.)*, 27(1):1–67, 1992.

[38] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 7(R-3):33–75, 1973.

[39] Lisandro D. Dalcin, Rodrigo R. Paz, Pablo A. Kler, and Alejandro Cosimo. Parallel distributed computing using Python. *Advances in Water Resources*, 34(9):1124–1139, 2011. New Computational Methods and Software Tools.

[40] M. H. A. Davis and A. R. Norman. Portfolio selection with transaction costs. *Math. Oper. Res.*, 15(4):676–713, 1990.

[41] K. Debrabant and E. R. Jakobsen. Semi-Lagrangian schemes for linear and fully non-linear diffusion equations. *Math. Comp.*, 82(283):1433–1462, 2013.

[42] P. Delanoë. Classical solvability in dimension two of the second boundary-value problem associated with the Monge-Ampère operator. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 8(5):443–457, 1991.

[43] G. Devillanova and F. Pugliese. A variant on Miranda-Talenti estimate. *Matematiche (Catania)*, 54(1):91–97 (2000), 1999.

[44] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bull. Sci. Math.*, 136(5):521–573, 2012.

[45] J. Douglas, Jr., T. Dupont, and L. Wahlbin. The stability in L^q of the L^2-projection into finite element function spaces. *Numer. Math.*, 23:193–197, 1974/75.

[46] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[47] Z. Fašková, R. Čunderlík, and K. Mikula. Finite element method for solving geodetic boundary value problems. *Journal of geodesy*, 84(2):135–144, 2010.

[48] X. Feng. Convergence of the vanishing moment method for the Monge-Ampére equations in two spatial dimensions. *Trans. AMS (submitted).*

[49] X. Feng, R. Glowinski, and M. Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Rev.*, 55(2):205–267, 2013.

[50] X. Feng and M. Jensen. Convergent semi-Lagrangian methods for the Monge–Ampère equation on unstructured grids. *SIAM J. Numer. Anal.*, 55(2):691–712, 2017.

[51] X. Feng and 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.

[52] X. Feng and M. Neilan. Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. *J. Sci. Comput.*, 38(1):74–98, 2009.

[53] W. H. Fleming and H. M. Soner. *Controlled Markov processes and viscosity solutions*, volume 25 of *Stochastic Modelling and Applied Probability*. Springer, New York, second edition, 2006.

[54] B. D. Froese and A. M. 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, 2011.

[55] Dietmar Gallistl. Numerical approximation of planar oblique derivative problems in nondivergence form. *Math. Comp*, 2018.

[56] C Geuzaine and JF Remacle. Gmsh: A three-dimensional finite element mesh generator with built-in pre-and post-processing facilities, version 2.2. 4, 2008.

[57] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[58] Diogo A. Gomes, Levon Nurbekyan, and Edgard A. Pimentel. *Economic models and mean-field games theory*. Publicações Matemáticas do IMPA. [IMPA Mathematical Publications]. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 2015. 30o Colóquio Brasileiro de Matemática. [30th Brazilian Mathematics Colloquium].

[59] P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics]*. Masson, Paris; Springer-Verlag, Berlin, 1992.

[60] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.

[61] Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. Mean field games and applications. In *Paris-Princeton Lectures on Mathematical Finance 2010*, volume 2003 of *Lecture Notes in Math.*, pages 205–266. Springer, Berlin, 2011.

[62] Q. Han and J-X. Hong. *Isometric embedding of Riemannian manifolds in Euclidean spaces*, volume 130 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2006.

[63] Bruce Hendrickson and Robert Leland. A multilevel algorithm for partitioning graphs. In *Supercomputing '95: Proceedings of the 1995 ACM/IEEE Conference on Supercomputing (CDROM)*, page 28, New York, 1995. ACM Press.

[64] Jia Xing Hong. Realization in $\mathbf{R}^3$ of complete Riemannian manifolds with negative curvature. *Comm. Anal. Geom.*, 1(3-4):487–514, 1993.

[65] N. M. Ivochkina. Classical solvability of the Dirichlet problem for the Monge-Ampère equation. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 131:72–79, 1983. Questions in quantum field theory and statistical physics, 4.

[66] M. Jensen. Numerical Solution of the Simple Monge–Ampére Equation with Non-convex Dirichlet Data on Non-convex Domains. *arXiv preprint arXiv:1705.04653*, 2017.

[67] M. Jensen and I. Smears. Finite element methods with artificial diffusion for Hamilton-Jacobi-Bellman equations. In *Numerical mathematics and advanced applications 2011*, pages 267–274. Springer, Heidelberg, 2013.

[68] M. Jensen and I. Smears. On the convergence of finite element methods for Hamilton-Jacobi-Bellman equations. *SIAM J. Numer. Anal.*, 51(1):137–162, 2013.

[69] Boško S. Jovanović and Endre Süli. *Analysis of finite difference schemes*, volume 46 of *Springer Series in Computational Mathematics*. Springer, London, 2014. For linear partial differential equations with generalized solutions.

[70] E. Kawecki. A DGFEM for Nondivergence Form Elliptic Equations with Cordes Coefficients on Curved Domains. *arXiv preprint arXiv:1708.05028*, 2017.

[71] E. Kawecki. A DGFEM for Uniformly Elliptic Two Dimensional Oblique Boundary Value Problems. *arXiv preprint arXiv:1711.01836*, 2017.

[72] E. Kawecki, O. Lakkis, and T. Pryer. A finite element method for the Monge–Ampère equation with transport boundary conditions. *arXiv preprint arXiv:1807.03535*, 2018.

[73] N. V. Krylov. Boundedly nonhomogeneous elliptic and parabolic equations in a domain. *Mathematics of the USSR-Izvestiya*, 22(1):67, 1984.

[74] N. V. Krylov. *Nonlinear elliptic and parabolic equations of the second order*, volume 7 of *Mathematics and its Applications (Soviet Series)*. D. Reidel Publishing Co., Dordrecht, 1987. Translated from the Russian by P. L. Buzytsky [P. L. Buzytskiĭ].

[75] K. Kuratowski and C. Ryll-Nardzewski. A general theorem on selectors. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 13:397–403, 1965.

[76] H. J. Kushner. Numerical methods for stochastic control problems in continuous time. *SIAM Journal on Control and Optimization*, 28(5):999–1048, 1990.

[77] Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation research part B: methodological*, 45(10):1572–1589, 2011.

[78] O. Lakkis and T. Pryer. A finite element method for second order nonvariational elliptic problems. *SIAM J. Sci. Comput.*, 33(2):786–801, 2011.

[79] O. Lakkis and T. Pryer. A finite element method for nonlinear elliptic problems. *SIAM J. Sci. Comput.*, 35(4):A2025–A2045, 2013.

[80] A Lasis and E Süli. Poincaré-Type inequalities for Broken Sobolev spaces, Isaac Newton Institute for Mathematical Sciences. *Preprint No. NI03067-CPD*, 2003.

[81] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Jpn. J. Math.*, 2(1):229–260, 2007.

[82] M. Lenoir. Optimal isoparametric finite elements and error estimates for domains involving curved boundaries. *SIAM J. Numer. Anal.*, 23(3):562–580, 1986.

[83] G. M. Lieberman. Two-dimensional nonlinear boundary value problems for elliptic equations. *Trans. Amer. Math. Soc.*, 300(1):287–295, 1987.

[84] G. M. Lieberman. *Oblique derivative problems for elliptic equations*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2013.

[85] G. Loeper and F. Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Math. Acad. Sci. Paris*, 340(4):319–324, 2005.

[86] Anders Logg, Kent-Andre Mardal, Garth N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[87] Fabio Luporini, Ana Lucia Varbanescu, Florian Rathgeber, Gheorghe-Teodor Bercea, J. Ramanujam, David A. Ham, and Paul H. J. Kelly. Cross-loop optimization of arithmetic intensity for finite element local assembly. *ACM Transactions on Architecture and Code Optimization*, 11(4):57:1–57:25, 2015.

[88] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[89] A. Maugeri, D. K. Palagachev, and L. G. Softova. *Elliptic and parabolic equations with discontinuous coefficients*, volume 109 of *Mathematical Research*. Wiley-VCH Verlag Berlin GmbH, Berlin, 2000.

[90] Andrew T. T. McRae, Colin J. Cotter, and Chris J. Budd. Optimal-transport-based mesh adaptivity on the plane and sphere using finite elements. *SIAM J. Sci. Comput.*, 40(2):A1121–A1148, 2018.

[91] Robert C Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *The review of Economics and Statistics*, pages 247–257, 1969.

[92] M. Neilan. Finite element methods for fully nonlinear second order PDEs based on a discrete Hessian with applications to the Monge-Ampère equation. *J. Comput. Appl. Math.*, 263:351–369, 2014.

[93] M. Neilan, A. J. Salgado, and W. Zhang. Numerical analysis of strongly nonlinear PDEs. *Acta Numer.*, 26:137–303, 2017.

[94] Bernt Ø ksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[95] A. M. Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. *Discrete Contin. Dyn. Syst. Ser. B*, 10(1):221–238, 2008.

[96] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[97] D. K. Palagachev. The Poincaré problem in L^p-Sobolev spaces. I. Codimension one degeneracy. *J. Funct. Anal.*, 229(1):121–142, 2005.

[98] A. Picarelli, C. Reisinger, and J. R. Arto. Error bounds for monotone schemes for parabolic Hamilton–Jacobi–Bellman equations in bounded domains. *arXiv preprint arXiv:1710.11284*, 2017.

[99] A. V. Pogorelov. *Monge-Ampère equations of elliptic type*. Translated from the first Russian edition by Leo F. Boron with the assistance of Albert L. Rabenstein and Richard C. Bollinger. P. Noordhoff, Ltd., Groningen, 1964.

[100] H. Poincaré. Lecons de Méchanique Céleste, Tome III, Théorie de Marées. *Gauthiers–Villars, Paris*, 1910.

[101] D. M. Pooley, P. A. Forsyth, and K. R. Vetzal. Numerical convergence properties of option pricing PDEs with uncertain volatility. *IMA J. Numer. Anal.*, 23(2):241–267, 2003.

[102] Peter R. Popivanov and Dian K. Palagachev. *The degenerate oblique derivative problem for elliptic and parabolic equations*, volume 93 of *Mathematical Research*. Akademie Verlag, Berlin, 1997.

[103] C. R. Prins. *Inverse methods for illumination optics*. PhD thesis, Technische Univeristeit Eindhoven, 2014.

[104] C. R. Prins, T. Boonkkamp, J. van Roosmalen, WL IJzerman, and TW Tukker. A numerical method for the design of free-form reflectors for lighting applications. 2013.

[105] Florian Rathgeber, David A. Ham, Lawrence Mitchell, Michael Lange, Fabio Luporini, Andrew T. T. McRae, Gheorghe-Teodor Bercea, Graham R. Markall, and Paul H. J. Kelly. Firedrake: automating the finite element method by composing abstractions. *ACM Trans. Math. Softw.*, 43(3):24:1–24:27, 2016.

[106] R. C. Reilly. Mean curvature, the Laplacian, and soap bubbles. *Amer. Math. Monthly*, 89(3):180–188, 197–198, 1982.

[107] M. Renardy and R. C. Rogers. *An introduction to partial differential equations*, volume 13 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2004.

[108] M. V. Safonov. Nonuniqueness for second-order elliptic equations with measurable coefficients. *SIAM J. Math. Anal.*, 30(4):879–895, 1999.

[109] R. Scott. *Finite-element techniques for curved boundaries*. ProQuest LLC, Ann Arbor, MI, 1973. Thesis (Ph.D.)–Massachusetts Institute of Technology.

[110] I. Smears and E. Süli. Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients. *SIAM J. Numer. Anal.*, 51(4):2088–2106, 2013.

[111] I. Smears and E. Süli. Discontinuous Galerkin finite element approximation of Hamilton-Jacobi-Bellman equations with Cordès coefficients. *SIAM J. Numer. Anal.*, 52(2):993–1016, 2014.

[112] I. Smears and E. Süli. Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordes coefficients. *Numerische Mathematik*, 133(1):141–176, 2016.

[113] Endre Süli and David F. Mayers. *An introduction to numerical analysis*. Cambridge University Press, Cambridge, 2003.

[114] Endre Süli and Igor Mozolevski. hp-version interior penalty dgfems for the biharmonic equation. *Computer methods in applied mechanics and engineering*, 196(13-16):1851–1863, 2007.

[115] M. Tsuji. Monge-Ampère equations and surfaces with negative Gaussian curvature. In *Symplectic singularities and geometry of gauge fields (Warsaw, 1995)*, volume 39 of *Banach Center Publ.*, pages 161–170. Polish Acad. Sci. Inst. Math., Warsaw, 1997.

[116] J. Urbas. On the second boundary value problem for equations of Monge-Ampère type. *J. Reine Angew. Math.*, 487:115–124, 1997.

[117] J. Urbas. Oblique boundary value problems for equations of Monge-Ampère type. *Calc. Var. Partial Differential Equations*, 7(1):19–39, 1998.

[118] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[119] G. C. Wen and C. J. Yang. Finite element solutions of the oblique derivative problem for elliptic complex equations of second order. *Sichuan Shifan Daxue Xuebao Ziran Kexue Ban*, 17(3):20–28, 1994.

[120] Z. Zhang and A. Naga. A new finite element gradient recovery method: superconvergence property. *SIAM J. Sci. Comput.*, 26(4):1192–1213 (electronic), 2005.

[121] Y. Zheng. A global solution to a two-dimensional Riemann problem involving shocks as free boundaries. *Acta Math. Appl. Sin. Engl. Ser.*, 19(4):559–572, 2003.

[122] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique. *Internat. J. Numer. Methods Engrg.*, 33(7):1331–1364, 1992.

[123] C. Zuily. Existence locale de solutions C^∞ pour des équations de Monge-Ampère changeant de type. *Comm. Partial Differential Equations*, 14(6):691–697, 1989.