Discontinuous Galerkin finite element approximation of Hamilton–Jacobi–Bellman equations with Cordes coefficients

Iain Smears

Worcester College

University of Oxford

A thesis submitted for the examination of

Doctor of Philosophy

Trinity 2015

Abstract

We propose a discontinuous Galerkin finite element method (DGFEM) for fully nonlinear elliptic Hamilton–Jacobi–Bellman (HJB) partial differential equations (PDE) of second order with Cordes coefficients. Our analysis shows that the method is both consistent and stable, with arbitrarily high-order convergence rates for sufficiently regular solutions. Error bounds for solutions with minimal regularity show that the method is generally convergent under suitable choices of meshes and polynomial degrees. The method allows for a broad range of *hp*-refinement strategies on unstructured meshes with varying element sizes and orders of approximation, thus permitting up to exponential convergence rates, even for nonsmooth solutions. Numerical experiments on problems with nonsmooth solutions and strongly anisotropic diffusion coefficients demonstrate the significant gains in accuracy and computational efficiency over existing methods.

We then extend the DGFEM for elliptic HJB equations to a space-time DGFEM for parabolic HJB equations. The resulting method is consistent and unconditionally stable for varying time-steps, and we obtain error bounds for both rough and regular solutions, which show that the method is arbitrarily high-order with optimal convergence rates with respect to the mesh size, time-step size, and temporal polynomial degree, and possibly suboptimal by an order and a half in the spatial polynomial degree. Exponential convergence rates under combined *hp*- and τq-refinement are obtained in numerical experiments on problems with strongly anisotropic diffusion coefficients and early-time singularities.

Finally, we show that the combination of a semismooth Newton method with nonoverlapping domain decomposition preconditioners leads to efficient solvers for the discrete nonlinear problems. The semismooth Newton method has a superlinear convergence rate, and performs very effectively in computations. We analyse the spectral bounds of nonoverlapping domain decomposition preconditioners for a model problem, where we establish sharp bounds that are explicit in both the mesh sizes and polynomial degrees. We then go beyond the model problem and show computationally that these algorithms lead to efficient and competitive solvers in practical applications to fully nonlinear HJB equations.

To my parents

Acknowledgements

My deepest thanks go to my supervisor, Endre Süli, whose kind support, clear guidance and limitless enthusiasm have made these years engaging, stimulating and rewarding beyond words. It has been my most heartfelt pleasure to learn from you.

I wish to express my greatest gratitude to Max Jensen, as your friendship and encouragement have truly opened up this path.

Tim Barth, you have my sincere thanks for your support and warm welcome during my visit to NASA Ames Research Center. You made this experience uniquely enjoyable and memorable.

Ingrid von Glehn, thank you for carefully reading my thesis, but above all, thank you for your constant support and encouragement.

The Numerical Analysis Group and the Mathematical Institute in Oxford have provided the most wonderful atmosphere during my studies. I have been very lucky to count many of you as my close friends.

Contents

Chapter 1

Introduction

We consider the numerical solution of a class of fully nonlinear second-order partial differential equations (PDE) called Hamilton–Jacobi–Bellman (HJB) equations. These PDE are named after Sir William Rowan Hamilton (1805–1865), Carl Gustav Jacobi (1804–1851), and Richard Bellman (1920–1984). HJB equations arise from models for the optimal control of stochastic processes.

In this chapter, we briefly describe the relation between HJB equations and control problems, followed by the main PDE-theoretic considerations in section 1.2. In sections 1.3 and 1.4, we review the current state of the art of numerical methods for HJB equations and highlight some key challenges. We then introduce in section 1.7 our main contributions that will be detailed in this thesis.

1.1 Optimal control of stochastic processes

Stochastic optimal control problems describe the time evolution of a state vector $X\colon t \mapsto X_t \in \mathbb{R}^d$ subject to a control process $\alpha(\cdot)\colon t \mapsto \alpha_t \in \Lambda$, where Λ is the set of available controls. Specifically, the state vector X obeys a given stochastic differential equation, whose drift and volatility terms are functions of $\alpha \in \Lambda$. Note that α denotes an element of Λ, whereas $\alpha(\cdot)$ denotes a Λ-valued function of time. The aim is to determine a control process that either minimises a given cost functional, or maximises a given utility functional. Typically, the control problem is terminated at a possibly random time, for instance a specified final time T, or at the time τ_{exit} of first exit of X_t from a bounded domain $\Omega \subset \mathbb{R}^d$.

For example, let f and c be real-valued functions on $\overline{\Omega} \times \Lambda$, with $c \geq 0$, and consider the stochastic control problem, subject to a stochastic differential equation, given by

$$\min_{\alpha(\cdot)\in\mathcal{C}} J^{\alpha(\cdot)}(x), \quad J^{\alpha(\cdot)}(x) \coloneqq \mathbb{E}\int_0^{\tau_{\text{exit}}} f^{\alpha_t}(X_t)\exp\left(-\int_0^t c^{\alpha_s}(X_s)\,\mathrm{d}s\right)\mathrm{d}t, \tag{1.1}$$

$$\mathrm{d}X_t = b^{\alpha_t}(X_t)\,\mathrm{d}t + \sigma^{\alpha_t}(X_t)\,\mathrm{d}B_t \quad \text{for } t > 0, \quad X_0 = x, \tag{1.2}$$

where B_t denotes a k-dimensional Brownian motion; $\mathbb{E}$ denotes the expected value at time $t = 0$; the matrix function σ takes values in $\mathbb{R}^{d\times k}$; the function b takes values in $\mathbb{R}^d$; and where $\mathcal{C}$ is a given set[1] of functions $\alpha(\cdot)\colon [0,\infty) \to \Lambda$, with Λ a given compact metric space. Here, we use the notational convention of denoting the dependence of functions on $\alpha \in \Lambda$ and functionals on $\alpha(\cdot) \in \mathcal{C}$ through a superscript, e.g. $f\colon (x,\alpha) \mapsto f^\alpha(x)$.

The control problem (1.1) leads to an elliptic HJB equation as follows: define the $d \times d$ matrix function $a^\alpha \coloneqq \sigma^\alpha(\sigma^\alpha)^\top/2$; the function a^α is usually referred to as the *diffusion coefficient*. The function $u\colon \Omega \to \mathbb{R}$ defined by

$$u(x) \coloneqq -\inf_{\alpha(\cdot)\in\mathcal{C}} J^{\alpha(\cdot)}(x), \quad x \in \Omega, \tag{1.3}$$

solves the HJB equation

$$\begin{aligned} \sup_{\alpha\in\Lambda} [L^\alpha u - f^\alpha] = 0 \quad &\text{in } \Omega, \\ u = 0 \quad &\text{on } \partial\Omega, \end{aligned} \tag{1.4}$$

where the nondivergence form elliptic operators L^α are defined by

$$L^\alpha v \coloneqq a^\alpha : D^2 v + b^\alpha \cdot \nabla v - c^\alpha v, \quad v \in H^2(\Omega),\ \alpha \in \Lambda. \tag{1.5}$$

The derivation of the HJB equation from the stochastic control problem essentially hinges on Bellman's dynamic programming principle and Dynkin's formula [63] for the infinitesimal generator of the stochastic process. It turns out that this derivation points to a solution strategy for the control problem: if the solution of the HJB equation is available, then optimal controls can often be computed as the maximisers of the expression appearing on the left-hand side of (1.4). The main advantage of this approach is that these optimisation problems are posed over Λ rather than $\mathcal{C}$, and can be solved independently for varying $x \in \Omega$. Thus, these local optimisation problems are tractable in many cases once the solution u and its partial derivatives are known.

However, it is clear that the function u and its partial derivatives cannot generally be computed from (1.3). Therefore, the success of this approach does depend on the availability and effectiveness of numerical methods for computing the solution of the HJB equation. For further results concerning the relationship between HJB equations and stochastic control problems, we refer the reader to [36], which provides a rigorous account of the dynamic programming principle and more general control problems.

[1] More precisely, the set $\mathcal{C}$ is required to be a subset of the set of progressively measurable functions with respect to the filtration of the Brownian motion, with possible further restrictions determined by the particular problem being modelled; see [36] for further details.

1.2 The notion of solution and its regularity

The equation (1.4) is called *uniformly elliptic* if there exist positive constants $0 < \nu \leq \overline{\nu}$ such that

$$\nu\,|\xi|^2 \leq \xi^\top a^\alpha(x)\,\xi \leq \overline{\nu}\,|\xi|^2 \quad \forall\,\xi \in \mathbb{R}^d,\ \forall\,x \in \Omega,\ \forall\,\alpha \in \Lambda. \tag{1.6}$$

The regularity theory centered around the celebrated Evans–Krylov Theorem establishes interior $C^{2,\beta}$-regularity of the solution of fully nonlinear uniformly elliptic and parabolic equations with convex nonlinearities, under standard regularity assumptions on the coefficients a, b, c and f [19, 20, 32, 48, 77]; this applies to HJB equations of the form (1.4). In this case the solution can be understood to satisfy the equation in a classical sense.

However, if the uniform ellipticity assumption is relaxed and the diffusion is allowed to become *degenerate*, with $\nu = 0$, then the solution, as defined by (1.3), is typically Lipschitz continuous at best. The eikonal equation with zero Dirichlet boundary datum is a representative example of this situation, the solution being the distance function to the boundary of the domain.

Therefore, in the degenerate case, the solution cannot be understood in the classical sense. Instead, the appropriate notion of solution is that of a *viscosity solution*; for an introductory exposition to viscosity solutions, see [24]. We note that the notion of a viscosity solution will not be required for understanding the analysis of this work. However, we recall several key aspects of viscosity solutions that are important for understanding the current state of the art in numerical methods for fully nonlinear PDE.

The first consideration is that the definition of viscosity solutions requires *a priori* only continuity of the solution, and the well-posedness theory of viscosity solutions is centred around the maximum principle. The notion of viscosity solution is applicable in both the degenerate and uniformly elliptic cases, although in the latter case the regularity theory mentioned above establishes higher regularity of the solution.

The second point is that another important notion of solution of elliptic PDE, namely that of *weak solutions*, is not applicable to fully nonlinear PDE such as HJB equations. This is simply because the second-order derivatives of the unknown solution appear under the nonlinearity, and thus no integration by parts is possible to pass partial derivatives onto test functions. Fully nonlinear PDE are therefore very different from linear, semilinear or quasilinear equations in divergence form. This explains why the development of Galerkin-type numerical methods for these problems has been particularly difficult, despite the successes of these numerical methods for other PDE.

In summary, the regularity of the solution depends on the uniform ellipticity or degeneracy of the problem appearing through a^α, with clearly important consequences at a computational level. However, for computational purposes, it is also important to consider the case where the diffusion coefficient a^α is *strongly anisotropic*, a typical example being

when the diffusion is dominant in certain directions, and the eigenvectors of a^α are not well-aligned with the computational grid. Strongly anisotropic problems can occur in both the uniformly elliptic or degenerate cases, and often occur as nearly degenerate problems with $\nu \ll \overline{\nu}$. As we shall see below, anisotropic diffusion leads to substantial challenges for the practical application of many numerical methods.

1.3 Monotone methods

Some of the earliest computational methods for HJB equations and stochastic control problems were based on approximating the underlying SDE by a discrete Markov chain [52]. Alongside the advent of the notion of viscosity solution [24], it became apparent that these Markov chain approximations admit equivalent interpretations as *monotone* finite difference methods (FDM) [16, 36], i.e. that satisfy a discrete maximum principle.

Broadly stated, these methods approximate the linear operators L^α by linear finite difference operators L_h^α on a computational grid of grid-size h, and solve the discrete problem

$$\sup_{\alpha\in\Lambda} \left[L_h^\alpha u_h - f^\alpha\right](x_i) = 0$$

for each grid point x_i. The scheme is said to be *monotone* provided that, for any grid-function v_h that has a nonnegative local maximum at x_i of the grid, one has $L_h^\alpha v_h(x_i) \leq 0$. This corresponds to requiring that all diagonal entries of the matrix representing L_h^α be negative, with only nonnegative off-diagonal entries. A classical example of a monotone scheme is the Kushner–Dupuis method [52, p. 1012], which is only applicable to rather isotropic problems with diagonally dominant diffusion coefficients a^α.

A central reason for the interest in monotone methods is that Barles and Souganidis provided in [13] a general convergence theory that is applicable to a broad class of possibly degenerate fully nonlinear elliptic and parabolic PDE. Specifically, provided that the underlying PDE satisfies an appropriate maximum principle and that the FDM is monotone, consistent, and stable in the sense that the sequence of numerical solutions $\{u_h\}_h$ remains bounded in the maximum norm, then it can be shown[2] that $\|u - u_h\|_{L^\infty} \to 0$ as $h \to 0$, without a priori regularity assumptions on the analytical solution u.

However, as we shall see below, the computational practice of monotone schemes has lagged behind their theoretical development, especially for strongly anisotropic problems. This is because monotone schemes suffer from significant drawbacks when used in practice. Indeed, various authors have commented on the necessarily low-order convergence rates of monotone schemes [28, 62], and on the restrictions imposed on the choice of stencil

[2]For degenerate problems with Dirichlet boundary conditions, there are some important technicalities concerning convergence of the numerical solutions at the boundary, which usually need to be analysed by some independent means, see [13, Remark 2.2]. See also the notion of *discontinuous viscosity solutions* in [12].

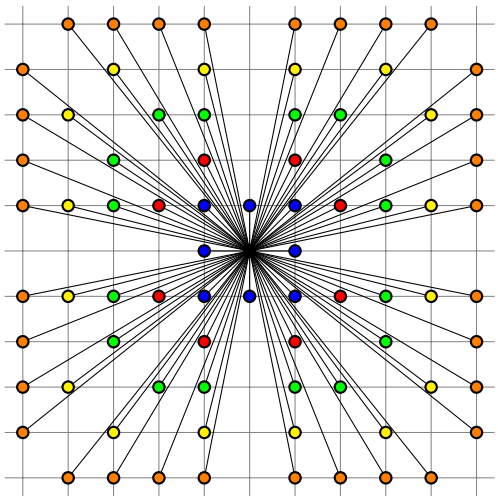

FIGURE 1.1: *In order to discretise strongly anisotropic diffusion coefficients a^α, monotone FDM require the use of wide stencils, such as the one depicted here that involves five layers of neighbouring nodes. The characterisation of the set of operators that can be discretised with this stencil is given in* [16, Section 5]*, where it is shown that it is necessary but not sufficient that $|a^\alpha_{ij}| \leq 5\, a^\alpha_{ii}$ for all $i \neq j$. See our experiments in section* 4.7 *for an example where $|a^\alpha_{ij}| \geq 20\, a^\alpha_{ii}$ for some i and j.*

necessary to achieve monotonicity of the discretisation [25, 47, 50]. It turns out that for strongly anisotropic diffusion coefficients a^α, compact stencils cannot offer a consistent and monotone discretisation, as first shown by Motzkin and Wasow [59]. In this case, a larger stencil is generally needed, as illustrated by Figure 1.1.

It is important to note that contrary to standard finite difference methods, increasing the stencil size in a monotone scheme cannot increase the order of accuracy. Instead, it has been pointed out [16] that increased stencil sizes typically increase the truncation error and thus decrease accuracy. Therefore it is generally desirable to use a minimal stencil that achieves monotonicity and consistency [16].

Kocan showed in [47] that the minimal stencil width required for monotonicity is expected to be of order $\mathcal{E}$ for $d = 2$ dimensions, of order $\mathcal{E}^{5/2}$ for $d = 3$ and of order $\mathcal{E}^{2d-4}$ for $d \geq 4$, where $\mathcal{E} = \overline{\nu}/\nu$ is the ellipticity constant. In the degenerate limit, there are examples of diffusion coefficients where no finite stencil can yield a monotone discretisation [25, 47]; in such a case, the stencil must be continually increased as the mesh is refined. Bonnans and Zidani examined in [16] the conditions that determine the set of problems that can be discretised with various stencils: they found that the number of conditions on the diffusion coefficient grows both with the stencil size and the problem dimension. Their work thus shows that it is generally nontrivial to determine a priori an appropriate stencil.

In terms of potentially usable methods for strongly anisotropic problems, Kuo and Trudinger analysed in [49, 50] a class of wide stencil monotone schemes (which were called therein *multistep schemes of positive type*) for uniformly elliptic HJB and Bellman–Isaacs equations. Moreover, they suggested what appears to be the first potentially applicable

algorithm for constructing the methods in practice[3], although no numerical experiments were presented; to our knowledge, their method has yet to be implemented in practice. Bonnans et al. proposed in [15] an algorithm for computing monotone discretisations of two-dimensional problems with finite stencils, with a consistency error depending on the stencil width. This is achieved by approximating the diffusion coefficient a^α by another coefficient $\tilde{a}^\alpha$ for which a monotone discretisation is available on a user-specified stencil. Convergence is then achieved by increasing the stencil size along with mesh refinement.

Building on the earlier work of Camilli and Falcone [21], Debrabant and Jakobsen [26] developed a semi-Lagrangian framework in which the stencil width continually increases as the mesh is refined. Their framework includes generalisations of the approximation

$$a^\alpha(x) : D^2 v(x) \approx \frac{1}{2} \sum_{p=1}^{d} \frac{\mathcal{I}v(x + k\sigma_p^\alpha) - 2\,\mathcal{I}v(x) + \mathcal{I}v(x - k\sigma_p^\alpha)}{k^2}, \tag{1.7}$$

where we recall that $a^\alpha = \sigma^\alpha(\sigma^\alpha)^\top/2$, k is a mesh-dependent parameter, the operator $\mathcal{I}$ denotes linear or bilinear interpolation onto the grid, the function $v \in C^2$, and where σ_p^α signifies the p-th column of the matrix σ^α. Typically k is of order $\sqrt{h}$, so that the stencil size is of order $1/\sqrt{h}$, thereby leading to first-order accuracy in the truncation error. One advantage of these methods is the guaranteed monotonicity of the discretisation, with consistency achieved as $h \to 0$.

In [26], Jakobsen and Debrabant treat HJB and Bellman–Isaacs equations specifically posed on the entire space $\mathbb{R}^d$ rather than on bounded domains. However, they point to some of the issues arising from the presence of a boundary in [26, Section 6.1], namely a possible loss of accuracy or monotonicity near the boundary. In particular, the finite difference formula (1.7) needs to be modified for points x close to the boundary $\partial\Omega$, possibly by a one-sided asymmetric formula.

In summary, three principal observations arise from the existing literature on monotone methods:

1. In order to treat strongly anisotropic or degenerate problems, large stencils are required to achieve both consistency and monotonicity.

2. Large stencils can have negative consequences for the computational aspects of monotone methods, such as higher truncation errors and increased costs.

3. As a result of these challenges, computational practice of monotone schemes for strongly anisotropic problems has lagged behind theoretical developments, with many practical issues remaining to be investigated.

[3] It is important to note Kocan's comment [47, p. 81] on a mistake in [49], which underestimated the necessary stencil sizes for problems in more than two dimensions.

Monotone finite element methods. The reader will readily notice that the literature described above concerns essentially monotone FDM. This emphasis on FDM stems primarily from the ability to establish monotonicity directly from the definition of the FD approximation. Also, the existing convergence theory of Barles and Souganidis [13] is restricted to the notion of consistency of finite difference methods.

However, Jensen and this author proposed in [45, 46] a *monotone* finite element methods (FEM) for a class of possibly degenerate HJB equations. This method was shown to converge to the viscosity solution in the L^∞-norm, as well as in the H^1-norm if there is a nondegenerate subset of the operators L^α. The method essentially combines mass lumping of the time derivative with nonstandard numerical approximations to the second order elliptic operators L^α, allowing for general implicit and explicit splittings.

To present the main results of [46], it is helpful to consider an example of a scheme covered by their analysis. Consider for example the parabolic HJB equation

$$\partial_t u - \sup_{\alpha\in\Lambda}[L^\alpha u - f^\alpha] = 0 \qquad \text{in } \Omega\times(0,T) \tag{1.8}$$

along with, for example, homogeneous initial time and lateral boundary conditions. In [46], the operators L^α were assumed to be isotropic but possibly degenerate: $a^\alpha \equiv \underline{a}^\alpha \,\mathrm{I}_d$ where $\underline{a}^\alpha$ is a nonnegative scalar function and I_d is the $d\times d$ identity matrix. The case $a^\alpha\equiv 0$ is of course allowed. Let $\mathcal{T}_h$ be a simplicial conforming mesh on Ω of mesh-size h, and let V_h^1 denote the standard $H_0^1(\Omega)$-conforming piecewise linear finite element space on $\mathcal{T}_h$. For $N \coloneqq \dim V_h^1$, let $\{x_h^\ell\}_{\ell=1}^N$ denote the set of interior nodes of the mesh, and let $\hat\phi_h^\ell$ denote the L^1-normalised hat function associated with the node x_h^ℓ, such that $\hat\phi_h^\ell(x_h^m)=0$ if $\ell\neq m$, and $\|\hat\phi_h^\ell\|_{L^1(\Omega)}=1$.

The operator L^α is approximated by a numerical operator $L_h^\alpha\colon V_h^1\to\mathbb{R}^N$ defined by

$$(L_h^\alpha v_h)_\ell \coloneqq -\overline{a}^\alpha(x_i^\ell)\langle\nabla v_h,\nabla\hat\phi_h^\ell\rangle + \langle b^\alpha\cdot\nabla v_h - c^\alpha v_h,\hat\phi_h^\ell\rangle,\quad \ell=1,\dots,N,$$

where $\langle\cdot,\cdot\rangle$ denotes the L^2-inner product, and where the function $\overline{a}^\alpha$ is an approximation to $\underline{a}^\alpha$, allowing for averaging, regularisation, and/or the inclusion of artificial diffusion: in practice, one frequently has $\|\overline{a}^\alpha - \underline{a}^a\|_{L^\infty}\le C\,h$. The numerical scheme is then to find $\{u_h^n\}_{n=1}^{T/\Delta t}\subset V_h^1$ for successive timesteps $\{t_n\}_{n=1}^{T/\Delta t}$ such that

$$\frac{u_h^n(x_h^\ell)-u_h^{n-1}(x_h^\ell)}{\Delta t} - \sup_{\alpha\in\Lambda}\Big[(L_h^\alpha u_h^n)_\ell - \langle f^a,\hat\phi_h^\ell\rangle\Big] = 0 \quad \forall\,\ell=1,\dots,N. \tag{1.9}$$

The method is *monotone*[4] provided that the operators L_h^α possess the property that whenever a function $v_h\in V_h^1$ has a nonnegative local maximum at a node x_h^ℓ, then $(L_h^\alpha v_h)_\ell\le 0$. In practice, this assumption can be guaranteed if, for example in two space dimensions,

[4] This is referred to as the *local monotonicity property* in [46], and is closely related to the weak discrete maximum principle.

the triangles of the mesh $\mathcal{T}_h$ are strictly acute, and the approximation $\overline{a}^\alpha$ includes artificial diffusion of order $h\,\|b^\alpha\|_{L^\infty} + h^2\,\|c^\alpha\|_{L^\infty}$, see [46, Section 8] and the references therein for further details. The construction of similarly monotone finite element schemes for anisotropic diffusion coefficients a^α appears to be more challenging than for the isotropic case, although we note the recent work of Nochetto and Zhang [61].

The FEM is therefore monotone in a similar sense to the monotone FDM considered above, as required by the convergence theory of Barles and Souganidis [13]. However, the FEM does not satisfy the consistency notion required by [13], namely that in general, if $x_h^\ell \to x \in \Omega$ as $h \to 0$ and $v \in C^\infty(\Omega)$, we may have $(L_h^\alpha \mathcal{I}_h v)_\ell \not\to L^\alpha v(x)$, where $\mathcal{I}_h$ is the nodal interpolant into V_h^1. Indeed, a simple (counter)example can be found for the Laplace operator on a square subdivided into four regular simplices [46, p. 146]. Therefore, monotone FEM violate the conditions of the framework of Barles and Souganidis. The challenge of showing convergence to the viscosity solution for monotone FEM was overcome in [46] by employing known $W^{1,\infty}$-norm approximation properties of elliptic projections to permit a modification of the notion of consistency used in the convergence analysis.

1.4 Existing nonmonotone schemes

As a result of the challenges described above, many authors have proposed various nonmonotone methods for various fully nonlinear PDE in order to avoid the stencil restrictions and low-order convergence rates of monotone schemes described above. Further motivation is provided by the success of FEM[5] for many linear, semilinear and quasilinear PDE, where a full analysis is often possible without reliance on monotonicity. As we shall see, a key question here is how to design stable and convergent methods for fully nonlinear equations that do not rely on monotonicity.

The review paper [33] summarises many of the approaches suggested before 2013. For example, one approach due to Feng, Neilan and coworkers is the so-called *vanishing moment method*, involving fourth-order perturbations to the PDE. In order to solve a fully nonlinear PDE of the form $F(x, u, \nabla u, D^2 u) = 0$, where F is a given nonlinear operator, they consider instead the fourth-order semilinear PDE

$$\varepsilon \Delta^2 u_\varepsilon + F(x, u_\varepsilon, \nabla u_\varepsilon, D^2 u_\varepsilon) = 0 \quad \text{in } \Omega,$$

where Δ^2 denotes the biharmonic operator, supplemented by the boundary conditions of the original problem, as well as artificial boundary conditions, since the problem is now of fourth order. The advantage of such an approach is that standard FEM discretisations of fourth-order PDE can be applied to the perturbed problem, although they cannot usually be expected to be robust as $\varepsilon \to 0$. Moreover, the question of the convergence $u_\varepsilon \to u$ as $\varepsilon \to 0$

[5] We include here of course the many variants of classical FEM, such as discontinuous Galerkin FEM (DGFEM).

currently remains open in most cases. From the point of view of computations, it is pointed out in [33] that experiments reveal that the artificial boundary conditions introduced by the vanishing moment method can lead to the appearance of spurious boundary layers in the approximate solutions.

The essential conclusion of [33] and the references therein is that none of the non-monotone methods reviewed there currently offer a satisfactory convergence analysis for fully nonlinear PDE, thus highlighting the many challenges associated with nonmonotone methods. Nevertheless, some methods have offered promising computational results in the absence of theoretical analysis. For instance, Lakkis and Pryer have successfully tested in [53, 54] a FEM using Hessian reconstructions on a broad range of nonlinear elliptic problems[6].

1.5 The Cordes condition

The primary focus of this work is on the development of nonmonotone methods for elliptic and parabolic HJB equations that satisfy *the Cordes condition*. The Cordes condition is an algebraic assumption on the coefficients appearing in the differential operators L^α. Importantly, it encompasses a large range of possibly strongly anisotropic applications; for instance, for elliptic problems in two dimensions without lower-order terms, the Cordes condition is implied by the uniform ellipticity condition (1.6), as shown in the following example.

Example 1.1. Consider the HJB equation (1.4) and assume that $L^\alpha v = a^\alpha : D^2 v$ for all $\alpha \in \Lambda$. In this case, the Cordes condition requires that there exists an $\varepsilon \in (0, 1]$ such that

$$\frac{|a^\alpha|^2}{(\operatorname{Tr} a^\alpha)^2} \leq \frac{1}{d - 1 + \varepsilon} \quad \text{in } \overline{\Omega}, \quad \forall\, \alpha \in \Lambda, \tag{1.10}$$

where $|a^\alpha|$ and $\operatorname{Tr} a^\alpha$ denote respectively the Frobenius norm and the trace of a^α. In two space dimensions $d = 2$, the uniform ellipticity condition (1.6) is sufficient for (1.10). Indeed, for each $\alpha \in \Lambda$, we have $\nu^2 \leq \det a^\alpha$, and $\operatorname{Tr} a^\alpha \leq 2\,\overline{\nu}$. So, for $\varepsilon = \nu^2 / \left(2\,\overline{\nu}^2 - \nu^2\right)$, we have

$$\frac{(a^\alpha_{11})^2 + 2\,(a^\alpha_{12})^2 + (a^\alpha_{22})^2}{(a^\alpha_{11} + a^\alpha_{22})^2} \leq 1 - \frac{2\,\nu^2}{(a^\alpha_{11} + a^\alpha_{22})^2} \leq 1 - \frac{\nu^2}{2\,\overline{\nu}^2} = \frac{1}{1 + \varepsilon}. \tag{1.11}$$

We note that the Cordes condition stated in (1.10) can be generalised to problems with lower-order terms, as shown in Chapter 3. The Cordes condition arises from the literature on nondivergence form PDE. It turns out that the analysis of well-posedness of these problems is more delicate than that of divergence form equations, even in the linear case [23, 38, 56].

[6] It should be noted that the emphasis of these computations was on Monge–Ampère equations rather than HJB equations, thus featuring some similarities but also some notable differences.

For instance, consider the linear problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.12}$$

where $Lv \coloneqq a : D^2 v$ with the matrix-valued function a satisfying the uniform ellipticity condition (1.6). In contrast to the study of divergence form equations, it is usually not possible to define a notion of weak solution to (1.12) when the coefficient a is not sufficiently regular. In the case of a merely continuous coefficient $a \in C(\overline{\Omega})^{d\times d}$, the Calderon–Zygmund theory of strong solutions [38] establishes the well-posedness of the problem, provided that the domain Ω has a $C^{1,1}$ boundary.

However, without additional hypotheses, the well-posedness of (1.12) is generally lost in the case of discontinuous $a \in L^\infty(\Omega)^{d\times d}$. For instance, there is an example in [38, p. 185] of a uniformly elliptic linear operator L with discontinuous a that admits at least two linearly independent strong solutions to the problem $Lu = 0$ in Ω, and $u = 0$ on $\partial\Omega$, where Ω is the unit sphere in $\mathbb{R}^d$, $d \geq 3$. We note that this example does not satisfy (1.10). The benefit of the Cordes condition is that if the possibly discontinuous coefficient a appearing in (1.12) satisfies (1.10), then existence and uniqueness of a solution $u \in H^2(\Omega) \cap H^1_0(\Omega)$, along with continuous dependence on the data, can be shown provided that Ω is convex, see [56]. We note that the convexity assumption on Ω is natural in the present context, since solutions of Poisson's equation may fail to be H^2-regular in nonconvex domains [40].

In this work, we introduce the use of the Cordes condition to the analysis and numerical analysis of fully nonlinear HJB equations. In Chapter 3, we show that, for convex domains, the Cordes condition leads to existence and uniqueness of a function $u \in H^2(\Omega) \cap H^1_0(\Omega)$ that solves pointwise almost everywhere the uniformly elliptic HJB equation (1.4). Further work is currently required to show that this *strong solution* is also the viscosity solution.

The motivation for studying HJB equations under the Cordes condition stems from the fact that HJB equations such as (1.4) are naturally related to linear nondivergence form equations with discontinuous coefficients similar to (1.12). As we now explain, the relation between these linear and nonlinear problems is that nondivergence form linear operators can be viewed as linearisations of the fully nonlinear operator. Indeed, there is a famous iterative algorithm for solving[7] (1.4), originally due to Bellman and Howard [44, 64]. The algorithm has been given various names in the literature, such as Howard's algorithm or policy iteration, although it has long been understood that it admits an interpretation as a Newton method for a nonlinear operator equation [64].

The general principle of the algorithm may be understood as follows. Given an approximate solution u_k, $k \in \mathbb{N}$, to (1.4), one finds for each $x \in \Omega$ an $\alpha_k(x) \in \Lambda$ such that

$$\alpha_k(x) \in \operatorname{argmax}_\alpha \left[(L^\alpha u_k - f^\alpha)(x) \right]. \tag{1.13}$$

[7] In practice, this algorithm is used once the problem has been discretised.

A new approximation u_{k+1} is sought as the solution of the linear problem

$$\begin{aligned} L^{\alpha_k} u_{k+1} &= f^{\alpha_k} \quad \text{in } \Omega, \\ u_{k+1} &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1.14}$$

where $f^{\alpha_k} : x \mapsto f^{\alpha_k(x)}(x)$, and where the coefficients of the linear operator L^{α_k} are similarly defined. Formally, a solution of (1.4) is a fixed point of this iteration. Since the mapping $x \mapsto \alpha_k(x)$ cannot be expected to be continuous in general, the coefficients appearing in the operator L^{α_k} are generally *discontinuous*. Therefore, this establishes the connection between the Cordes condition and fully nonlinear HJB equations.

1.6 Discontinuous Galerkin finite element methods

As explained in section 1.5 above, the analysis of well-posedness of an elliptic HJB equation with Cordes coefficients involves the Sobolev space $H^2(\Omega) \cap H_0^1(\Omega)$. Therefore, conforming finite element discretisations require at least H^2-regularity of the approximation, which in practice amounts to a C^1-continuity requirement on the finite element space. Note that the conformity requirements in the situation of fully nonlinear HJB equations are thus different to those occuring in the case of linear, semilinear or quasilinear divergence form elliptic equations; this difference is naturally related to the fact that fully nonlinear HJB equations do not admit weak formulations.

The demanding implementational aspects of H^2-conforming finite element spaces on general meshes therefore motivate the application of nonconforming methods, either through continuous classical finite element spaces, which are H^1-conforming but not H^2-conforming, or discontinuous finite element spaces. In this work, we study the latter choice, although we note that the majority of our results carry over to continuous FEM, except for certain rather technical details in the error bounds for parabolic problems in Chapter 4, and the nonoverlapping domain decomposition preconditioners studied in Chapter 5.

Discontinuous Galerkin finite element methods (DGFEM) allow the approximation of the solution to be discontinuous between elements, with the continuity conditions being enforced only weakly through the discretised problem. These methods have been analysed and applied to a large range of problems [7, 43, 60]; see also the book [27] by Di Pietro and Ern for a broader introduction and further references.

The lack of inter-element continuity requirements on the discontinuous finite element space facilitates the implementation of *hp*-refinement, where one varies both mesh size and polynomial degree. Roughly speaking, typical *hp*-refinement methods use small elements with low polynomial degrees in regions where the smoothness of the solution is limited, and use large elements with high polynomial degrees in regions of higher smoothness. In the context of continuous Galerkin finite element methods and DGFEM, *hp*-refinement has been used to obtain exponential convergence rates for problems with nonsmooth solutions,

boundary layers or other localised features that would be difficult to resolve with uniform mesh or polynomial refinement; see [11, 42, 57, 68, 78] for further theory, examples and references on *hp*-version FEM and DGFEM.

1.7 Contributions

In this thesis, we present the main contributions of our papers [69, 70, 71, 72] concerning *hp*-version discontinuous Galerkin finite element methods (DGFEM) for uniformly elliptic and uniformly parabolic HJB equations with Cordes coefficients. As we shall see below, the introduction of the Cordes condition has enabled the development of an *hp*-version DGFEM that has a complete theoretical analysis in terms of consistency, stability and error bounds. To our knowledge, this is the first nonmonotone method for fully nonlinear problems that permits such an extensive analysis, and it is the first method that has been shown to exhibit exponential convergence rates for fully nonlinear PDE under *hp*-refinement, even in the strongly anisotropic setting. The main contributions of our work are as follows.

Analysis of the continuous problem. Although the HJB equation (1.4) does not admit a weak formulation, it does admit an equivalent formulation as a variational problem of the form $A(u;v) = 0$ for all $v \in H^2(\Omega) \cap H_0^1(\Omega)$, with A a nonlinear form detailed in Chapter 3. The Cordes condition leads to a key stability result in the form of a *strong monotonicity bound*[8]:

$$\|u-v\|^2_{H^2(\Omega)} \lesssim A(u;u-v) - A(v;u-v) \quad \forall\, u,v \in H^2(\Omega) \cap H_0^1(\Omega). \tag{1.15}$$

As first shown in [71], this enables the application of the theory of strongly monotone operators, in particular the Browder–Minty theorem [65], to prove existence and uniqueness of a solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of (1.4). The analysis of the elliptic problem (1.4) can also be extended to the parabolic setting, as shown in Chapter 4.

Stable and consistent discretisation. After defining the discontinuous Galerkin finite element space $V_{h,\mathbf{p}}$, we construct a discrete analogue A_h of the nonlinear form A, leading to the numerical scheme of finding $u_h \in V_{h,\mathbf{p}}$ such that $A_h(u_h;v_h) = 0$ for all $v_h \in V_{h,\mathbf{p}}$. The method is *consistent* in the usual sense of Galerkin-type methods, i.e. a sufficiently smooth solution u of (1.4) also satisfies $A_h(u;v_h) = 0$ for all $v_h \in V_{h,\mathbf{p}}$. This shows an essential difference between our method and inconsistent methods such as the vanishing moment method described in section 1.4.

However, the main challenge in the design of nonmonotone schemes is to achieve *stability*, in this case through a discrete analogue of the strong monotonicity bound (1.15). This is

[8] To avoid confusion between monotone methods and strong monotonicity of nonlinear operators, see Remark 1.1.

accomplished by combining the Cordes condition with the original idea of relating the residual of the equation to additional terms measuring the lack of H^2-conformity of the numerical solution. In the parabolic setting, we propose a consistent and stable space-time DGFEM, which permits high-order approximation in both space and time.

Remark 1.1. We wish to emphasize that *strong monotonicity* of nonlinear operators is unrelated to the notion of *monotone methods* discussed in section 1.3. Unfortunately, it is hard to avoid these terminologies since they are rather settled in the literature.

High-order convergence rates. The stability and consistency properties enable the derivation of error bounds. In the special case of quasi-uniform meshes of size h and quasi-uniform polynomial degrees p, provided that the solution has elementwise H^s-regularity for some $s > 5/2$, the error bound for the elliptic problem is of the form

$$\|u - u_h\| \lesssim \frac{h^{\min(s,\,p+1)-2}}{p^{s-5/2}} \|u\|_{H^s(\Omega)}. \tag{1.16}$$

where the norm $\|\cdot\|$ is an H^2-type mesh-dependent norm specified later. If instead the solution has only minimal regularity, then there holds a best approximation property

$$\|u - u_h\| \lesssim \inf\left\{\|u - z_h\| : z_h \in V_{h,\mathbf{p}} \cap H^2(\Omega) \cap H_0^1(\Omega)\right\}, \tag{1.17}$$

This bound shows that the method is convergent when the finite element space $V_{h,\mathbf{p}}$ is sufficiently rich. Importantly, these bounds do not depend on the anisotropy of the problem, except through the constants appearing in the Cordes condition. Therefore, our method is able to exploit any available regularity of the solution and yield high-order accuracy for strongly anisotropic problems, as shown by our numerical experiments in subsequent chapters. If *hp*-refinement is employed, it is even possible to achieve convergence rates of the form

$$\|u - u_h\| \lesssim \exp\Big(-c\,\sqrt[3]{\mathrm{DoF}}\Big), \tag{1.18}$$

where DoF is the number of degrees of freedom, even for low-regularity solutions with singularities in parts of the domain.

Robust discrete solvers. In practical terms, the implementation and algorithmic aspects of the method, such as memory costs, are the same as those of usual DGFEM for elliptic and parabolic problems. The fast and scalable solution of the discrete nonlinear problem is obtained by a combination of nonoverlapping domain decomposition preconditioners and the discrete version of the semismooth Newton method that was described in section 1.5 above. In Chapter 3, we show that the semismooth Newton method converges superlinearly and that in practice it remains robust under mesh refinement. In Chapter 5, we

consider the nonoverlapping domain decomposition preconditioners studied in [69], where computations show that they lead to efficient preconditioned GMRES solvers for the linear systems encountered at each step of the Newton method.

Extension to parabolic problems. The numerical scheme can be extended to the parabolic setting in a natural way. In [72], we propose a space-time DGFEM allowing very general meshes, time steps and polynomial degrees, which allows high-order accuracy in both space and time. Essentially, the space-time DGFEM employs a tensor product space consisting of standard piecewise polynomials over a spatial mesh and temporal polynomials of degree q over a time interval of length τ. The resulting method is then consistent and stable in a discrete $L^2(H^2(\Omega)) \cap H^1(L^2(\Omega))$-type Bochner norm over a time interval $(0, T)$. For instance, when employing quasi-uniform meshes and uniform polynomial degrees p in space, and uniform time steps τ with uniform temporal polynomial degree q in time, the error bound is of the form

$$\begin{aligned}(1.19)\quad \|u - u_h\| \lesssim \frac{h^{\min(s,p+1)-2}}{p^{s-7/2}}\|u\|_{L^2(H^s(\Omega))} + \frac{h^{\min(\bar{s},p+1)}}{p^{\bar{s}}}\|u\|_{H^1(H^{\bar{s}}(\Omega))} \\ + \frac{h^{\min(\tilde{s},p+1)-1}}{p^{\tilde{s}-3/2}}\|u(0)\|_{H^{\tilde{s}}(\Omega)} + p^{3/2}\sum_{\ell\in\{0,2\}}\frac{\tau^{\min(\sigma_\ell,q+1)-1+\ell/2}}{q^{\sigma_\ell-1+\ell/2}}\|u\|_{H^{\sigma_\ell}(H^\ell(\Omega))},\end{aligned}$$

where $\|\cdot\|$ now denotes a discrete $L^2(H^2(\Omega)) \cap H^1(L^2(\Omega))$-type norm over $(0, T)$, and where we assume that $s > 5/2$, $\bar{s} > 0$, $\tilde{s} > 3/2$, and $\sigma_\ell \geq 1$ for $\ell \in \{0, 2\}$. The bound (1.19) shows that the method has optimal convergence rates with respect to h, τ and q, and is possibly suboptimal in p by three half-orders.

Although other time discretisations could be considered, a key advantage of space-time DGFEM is that these schemes have the potential for exponential convergence rates, even for low-regularity solutions. Indeed, a combination of τq-refinement in time and hp-refinement in space can lead to a rate

$$(1.20)\qquad \|u - u_h\| \lesssim \exp\Big(-c_1\sqrt[3]{\mathrm{DoF}_x}\Big) + \exp\Big(-c_2\sqrt{\mathrm{DoF}_\tau}\Big),$$

where DoF_x and DoF_τ are respectively the number of spatial and temporal degrees of freedom in the finite element space. Exponential convergence rates of this form were first shown by Schötzau and Schwab in [67] in the context of linear divergence form parabolic problems. Therefore, we show in the numerical experiment of Chapter 4 that our method retains this quality.

Strongly anisotropic problems. The focus of this work is on elliptic and parabolic HJB equations that satisfy the Cordes condition. As shown in Example 1.1 for two space dimensions, the Cordes condition covers many problems with strongly anisotropic diffusion coefficients. Although the Cordes condition does not encompass all strongly anisotropic

problems, we have sought to test our method on such problems whenever possible. Thus the numerical experiments frequently emphasise the computational performance of our method on challenging problems with strongly anisotropic diffusion coefficients.

For instance, in Chapter 3, we apply the numerical method to a strongly anisotropic problem with a solution featuring a sharp boundary layer. Approximating such a solution is challenging for many of the monotone schemes described in section 1.3, because these schemes have been mostly developed for uniform or quasi-uniform grids or triangulations of the domain, which are inefficient at approximating solutions with strongly localised features. It would be highly desirable for monotone schemes to permit the use of graded meshes that are more efficient at approximating the solution, yet this would involve further complications with regards to guaranteeing the monotonicity and consistency of the scheme.

A key advantage of our method for problems of this kind is that it is not restricted by the inherent limitations of monotone methods that were described in section 1.3. Instead, our method can employ rather general meshes with standard compact "stencils" for strongly anisotropic problems. This added flexibility enables us to obtain greater efficiency in approximating the solution of the problem by using highly graded meshes combined with p- and hp-refinement.

The numerical experiments of this work demonstrate the gains in efficiency and performance of the method on a range of strongly anisotropic problems that satisfy the Cordes condition. However, it is clear that not all strongly anisotropic problems are encompassed by the Cordes condition, especially in high dimensions. Therefore, many interesting open and challenging problems remain to be solved in the area of numerical analysis for strongly anisotropic fully nonlinear partial differential equations.

Chapter 2

Nondivergence form elliptic equations

In Chapter 1, it was seen that fully nonlinear HJB equations are related to nondivergence form linear equations with discontinuous coefficients. It is therefore natural to first consider our numerical method in the simpler context of linear nondivergence form elliptic PDE. In this chapter, which is based on our paper [70], we present many of the key ideas that are extended to elliptic and parabolic HJB equations in Chapters 3 and 4.

Consider the linear boundary value problem

$$\begin{aligned} Lu &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $f \in L^2(\Omega)$ and L is a second-order elliptic operator in nondivergence form, i.e. the leading term of L is of the form $a : D^2u$, with coefficients $a \in L^\infty(\Omega)^{d\times d}$. To keep the exposition clear, we focus on operators L without lower-order terms, leaving the extension to problems with lower-order terms to subsequent chapters. The assumption of a homogeneous Dirichlet boundary condition is not essential, and the treatment of nonhomogeneous boundary conditions is discussed in section 2.7 below.

As stated in the introduction, if the diffusion coefficient a is discontinuous, then well-posedness of the solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of (2.1) is known provided that Ω is convex and that a satisfies the Cordes condition, which we recall in (2.5) below. The purpose of this chapter is to show how this condition plays a central role in the analysis of the numerical method proposed in this work.

A key question addressed here is that of specifying a stable discretisation scheme for the boundary value problem (2.1), since the lack of smoothness of a prevents the use of a weak formulation to exhibit the underlying coercive structure of the differential operator. Instead, stability of the numerical method is achieved by coupling the residual of the differential equation to terms measuring the lack of H^2-conformity of the numerical solution. We will see that the choice of bilinear form draws upon a discrete analogue of an identity that is central to the analysis of well-posedness of elliptic problems on convex domains [40, 56].

Section 2.1 defines the problem considered and the notation used in this chapter. This is followed by the definition of the scheme and the analysis of consistency in section 2.3. Stability and well-posedness of the numerical method are proved in section 2.5, followed by the a priori error analysis in section 2.6, where it is found that the convergence rates in a broken H^2-type norm are optimal with respect to the mesh size and suboptimal with respect to the polynomial degree by only half an order. Section 2.7 presents numerical experiments testing the accuracy and robustness of the scheme: the first experiment verifies the predicted convergence rates, and the second experiment gives an example of exponential accuracy under appropriate *hp*-refinement for a problem featuring both discontinuity of the coefficients and nonsmoothness of the solution.

2.1 Analysis of the continuous problem

Let Ω be a bounded convex polyhedral domain in $\mathbb{R}^d$, $d \geq 2$. Note that the convexity assumption implies that Ω has a Lipschitz boundary $\partial\Omega$; see [40].

Let the bounded operator $L\colon H^2(\Omega) \to L^2(\Omega)$ be defined by

$$Lv \coloneqq a : D^2 v, \quad v \in H^2(\Omega), \tag{2.2}$$

where $D^2 v$ denotes the Hessian of v, and $a \in L^\infty(\Omega)^{d\times d}$ is a symmetric matrix. We assume that L is uniformly elliptic, i.e. there exist constants $0 < \nu \leq \overline{\nu}$ such that

$$\nu\,|\xi|^2 \leq \xi^\top a(x)\,\xi \leq \overline{\nu}\,|\xi|^2 \quad \forall\, \xi \in \mathbb{R}^d, \text{ a.e. } x \text{ in } \Omega. \tag{2.3}$$

We consider the following problem: for a given $f \in L^2(\Omega)$, find a strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of the boundary-value problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.4}$$

In this section, we show under suitable conditions on the coefficient a in (2.2) that (2.4) possesses a unique strong solution in the space $H^2(\Omega) \cap H_0^1(\Omega)$. Although this result is already known [56], the formulation of the proof given here is original in order to highlight the key ingredients required for developing consistent, stable and high-order nonmonotone numerical methods. As we will see, these ingredients include the following:

- the Cordes condition, which is an algebraic assumption on the coefficients of the problem, and is therefore preserved at the discrete level;
- the Miranda–Talenti inequality, stated in Theorem 2.2 below. This inequality for the function space $H^2(\Omega) \cap H_0^1(\Omega)$ on convex domains Ω cannot be preserved in the discrete setting except for H^2-conforming discretisations. Therefore, our numerical

scheme will include additional terms to weakly reproduce this inequality on the finite element space consisting of discontinuous piecewise polynomial functions.

The Cordes condition. As explained in section 1.5, in the case of discontinuous diffusion coefficients $a \in L^\infty(\Omega)^{d\times d}$, the uniform ellipticity assumption (2.3) is generally not sufficient to obtain well-posedness of the problem (2.4). We assume *the Cordes condition*: there is a constant $\varepsilon \in (0, 1]$ such that

$$\frac{|a(x)|^2}{(\operatorname{Tr} a(x))^2} \leq \frac{1}{d-1+\varepsilon} \quad \text{for a.e. } x \text{ in } \Omega, \tag{2.5}$$

where $|a(x)|$ and $\operatorname{Tr} a(x)$ denote respectively the Frobenius norm and the trace of $a(x)$. For problems in two dimensions, uniform ellipticity implies the Cordes condition, as shown in Example 1.1, thus demonstrating that our results are relevant to a very broad class of problems, including some that require large stencils for monotone FDM; significant further evidence for this observation is found in the numerical experiments of section 2.7.

Let the function $\gamma \in L^\infty(\Omega)$ be defined by

$$\gamma \coloneqq \frac{\operatorname{Tr} a}{|a|^2}. \tag{2.6}$$

The uniform ellipticity assumption on the operator L implies that there is a $\gamma_0 > 0$ such that $\gamma \geq \gamma_0$ a.e. in Ω. The Cordes condition implies the following inequality that will be central to the subsequent analysis. This result is well-known, see for instance [56] for a proof.

Lemma 2.1. *Let the linear operator L defined by* (2.2) *satisfy the uniform ellipticity condition* (2.3) *and the Cordes condition* (2.5) *and let $\gamma \in L^\infty(\Omega)$ be defined by* (2.6)*. Then, for any open set $U \subset \Omega$ and $v \in H^2(U)$, we have*

$$|\gamma Lv - \Delta v| \leq \sqrt{1-\varepsilon}\, |D^2 v| \quad \text{a.e. in } U, \tag{2.7}$$

where $\varepsilon \in (0, 1]$ is as in (2.5)*.*

Proof. Let $v \in H^2(U)$. Then,

$$|\gamma Lv - \Delta v| = |(\gamma a - \mathrm{I}_d) : D^2 v| \leq |\gamma a - \mathrm{I}_d|\, |D^2 v|,$$

where I_d denotes the $d \times d$ identity matrix. Now, by expanding the square and using the definition of γ from (2.6), we find that

$$|\gamma a - \mathrm{I}_d|^2 = \sum_{i,j=1}^{d} |\gamma a_{ij} - \delta_{ij}|^2 = d - \frac{(\operatorname{Tr} a)^2}{|a|^2} \leq 1 - \varepsilon.$$

Therefore, the Cordes condition (2.5) implies that $|\gamma a - \mathrm{I}_d| \leq \sqrt{1-\varepsilon}$, thereby showing inequality (2.7). □

Remark 2.1. Lemma 2.1 shows that the Cordes condition requires that, after renormalisation by the scalar function γ, the operator L must be sufficiently close to the Laplacian. In particular, the fact that the constant appearing in the right-hand side of (2.7) is strictly less than one will be used in an essential way in the following.

Remark 2.2. Sometimes, a change of coordinates can transform a problem that fails to satisfy (2.5) into an equivalent problem that does satisfy (2.5). For example, assume that $a \in L^\infty(\Omega)^{d\times d}$ is any *constant* uniformly elliptic tensor, not necessarily satisying (2.5). Then, there exists an affine map $F\colon \Omega \to F(\Omega)$ defining a new set of coordinates $\hat{x} = F(x)$ such that the PDE in (2.4) becomes $\Delta u = f$ in $F(\Omega)$. The transformed domain $F(\Omega)$ is moreover convex since Ω is convex and F is affine. Therefore, the change of coordinates defines an equivalent problem that satisfies (2.5) in the new coordinate system.

Example 2.1. In the general case, the Cordes condition becomes increasingly restrictive as the dimension d increases. For example, consider a tridiagonal diffusion coefficient a with on-diagonal entries 1 and off-diagonal entries with magnitudes less than or equal to δ, with $\delta > 0$. Then, it can be shown that the Cordes condition is satisfied provided that

$$\varepsilon \leq \frac{d - 2\,\delta\,(d-1)^2}{d\,(1+2\,\delta) - 2\,\delta}, \quad \delta < \frac{d}{2\,(d-1)^2}.$$

Therefore, asymptotically, the off-diagonal terms $\delta \sim 1/2\,d$.

Miranda–Talenti inequality. We follow [56] in naming the following result the Miranda–Talenti inequality.

Theorem 2.2 (Miranda–Talenti)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain. Then, for any $u \in H^2(\Omega) \cap H^1_0(\Omega)$,*

$$|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}, \tag{2.8a}$$

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2(\Omega)}, \tag{2.8b}$$

where C is a constant depending only on d and $\operatorname{diam}\Omega$.

For a proof of Theorem 2.2, we refer the reader to [40, Chapter 3], where the result is proved for convex domains with $C^{1,1}$ boundaries, and to Appendix A for the generalisation to nonsmooth convex domains.

Well-posedness. The proof of well-posedness of (2.4) essentially relies on its reformulation as a variational problem of the form $A(u,v) = \ell(v)$ for all $v \in H^2(\Omega)\cap H^1_0(\Omega)$, where A

is a suitably chosen bilinear form and ℓ is a suitable linear functional. After showing boundedness of A, we employ the Cordes condition and the Miranda–Talenti inequality to show that A is coercive; the well-posedness of (2.4) then follows directly from the Lax–Milgram theorem [31]. The proof we give here thus differs from the one given in [56], which is based on contractive mappings and Banach's fixed point theorem.

Theorem 2.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, and let the linear operator L defined by* (2.2) *satisfy the uniform ellipticity condition* (2.3) *and the Cordes condition* (2.5). *Then, for any given $f \in L^2(\Omega)$, there exists a unique strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of* (2.4). *Moreover, we have $\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}$, where C depends only on d,* $\operatorname{diam}\Omega$, *$\nu$, $\overline{\nu}$, and ε.*

Proof. Let γ be defined by (2.6). Define $H \coloneqq H^2(\Omega) \cap H_0^1(\Omega)$. Theorem 2.2 shows that the bilinear form $\langle \cdot, \cdot \rangle_\Delta \colon H \times H \to \mathbb{R}$, $\langle u, v \rangle_\Delta \coloneqq \int_\Omega \Delta u \, \Delta v \, \mathrm{d}x$, defines an inner product on H, and it follows that $(H, \langle \cdot, \cdot \rangle_\Delta)$ is a Hilbert space. Let $\|\cdot\|_\Delta$ denote the norm induced by the inner product on H. Define the bilinear form $A \colon H \times H \to \mathbb{R}$ by

$$A(u, v) \coloneqq \int_\Omega \gamma L u \, \Delta v \, \mathrm{d}x, \quad u, v \in H. \tag{2.9}$$

Since $a \in L^\infty(\Omega)^{d \times d}$, Theorem 2.2 shows that A is bounded: for all u, $v \in H$, $|A(u, v)| \leq C\|u\|_\Delta \|v\|_\Delta$. We claim that A is coercive on H. Indeed, using (2.7),

$$A(u, u) = \langle u, u \rangle_\Delta - \int_\Omega (\Delta - \gamma L)\, u \, \Delta u \, \mathrm{d}x \geq \|u\|_\Delta^2 - \sqrt{1 - \varepsilon} |u|_{H^2(\Omega)} \|u\|_\Delta.$$

By Theorem 2.2, we have $|u|_{H^2(\Omega)} \leq \|u\|_\Delta$, so

$$A(u, u) \geq \left(1 - \sqrt{1 - \varepsilon}\right) \|u\|_\Delta^2, \tag{2.10}$$

and hence A is coercive.

Given $f \in L^2(\Omega)$, define $\ell \colon H \to \mathbb{R}$ by $\ell(v) \coloneqq \int_\Omega \gamma f \, \Delta v \, \mathrm{d}x$, for $v \in H$. Then ℓ is a bounded linear functional on H. The Lax–Milgram theorem shows existence and uniqueness of $u \in H$ such that $A(u, v) = \ell(v)$ for all $v \in H$. We claim that $Lu = f$ pointwise a.e. in Ω. For any $g \in L^2(\Omega)$, there is $v \in H$ such that $\Delta v = g$. So

$$\int_\Omega \gamma L u \, g \, \mathrm{d}x = \int_\Omega \gamma f \, g \, \mathrm{d}x \quad \forall\, g \in L^2(\Omega).$$

This implies that $\gamma L u = \gamma f$ a.e. in Ω. Since $\gamma > 0$ a.e. in Ω, we deduce that u is a strong solution of (2.4). Finally, we have

$$\|u\|_{H^2(\Omega)} \leq C\|u\|_\Delta \leq C \frac{\|\gamma\|_{L^\infty(\Omega)}}{1 - \sqrt{1 - \varepsilon}} \|f\|_{L^2(\Omega)},$$

where the constant C from (2.8b) depends only on d and $\operatorname{diam}\Omega$. □

2.2 Definitions

For real numbers a and b, we shall write $a \lesssim b$ to signify that there is a positive constant C such that $a \leq Cb$, where C is independent of the quantities of interest, such as the element sizes and polynomial degrees, but possibly dependent on other quantities, such as the mesh regularity parameters. Furthermore, we write $a \simeq b$ if $a \lesssim b$ and $b \lesssim a$.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded convex polytopal domain. Note that convexity of Ω implies that the boundary $\partial\Omega$ of Ω is Lipschitz [40]. Let $\{\mathcal{T}_h\}_h$ be a sequence of shape-regular meshes on Ω, consisting of simplices or parallelepipeds. For each element $K \in \mathcal{T}_h$, let $h_K \coloneqq \operatorname{diam} K$. It is assumed that $h = \max_{K\in\mathcal{T}_h} h_K$ for each mesh $\mathcal{T}_h$. Let $\mathcal{F}_h^i$ denote the set of interior faces of the mesh $\mathcal{T}_h$ and let $\mathcal{F}_h^b$ denote the set of boundary faces. The set of all faces of $\mathcal{T}_h$ is denoted by $\mathcal{F}_h^{i,b} \coloneqq \mathcal{F}_h^i \cup \mathcal{F}_h^b$. Since each element has piecewise flat boundary, the faces may be chosen to be flat.

Mesh conditions. The meshes are allowed to be irregular, i.e. there may be hanging nodes. We assume that there is a uniform upper bound on the number of faces composing the boundary of any given element; in other words, there is a constant $c_{\mathcal{F}} > 0$, independent of h, such that

$$\max_{K\in\mathcal{T}_h} \operatorname{card}\{F \in \mathcal{F}_h^{i,b} \colon F \subset \partial K\} \leq c_{\mathcal{F}} \quad \forall\, K \in \mathcal{T}_h. \tag{2.11}$$

It is also assumed that any two elements sharing a face have commensurate diameters, i.e. there is a constant $c_{\mathcal{T}} \geq 1$, independent of h, such that

$$\max(h_K, h_{K'}) \leq c_{\mathcal{T}} \min(h_K, h_{K'}) \tag{2.12}$$

for any K and K' in $\mathcal{T}_h$ that share a face. For each h, let $\mathbf{p} \coloneqq (p_K \colon K \in \mathcal{T}_h)$ be a vector of positive integers; note that this requires $p_K \geq 1$ for all $K \in \mathcal{T}_h$. We make the assumption that $\mathbf{p}$ has *local bounded variation*: there is a constant $c_{\mathcal{P}} \geq 1$, independent of h, such that

$$\max(p_K, p_{K'}) \leq c_{\mathcal{P}} \min(p_K, p_{K'}) \tag{2.13}$$

for any K and K' in $\mathcal{T}_h$ that share a face.

Function spaces. For each $K \in \mathcal{T}_h$, let $\mathcal{P}_{p_K}(K)$ be the space of all real-valued polynomials in $\mathbb{R}^d$ with either total or partial degree at most p_K. In particular, we allow the combination of spaces of polynomials of fixed total degree on some parts of the mesh with spaces of polynomials of fixed partial degree on the remainder. We also allow the use of the space of polynomials of total degree at most p_K even when K is a parallelepiped. Throughout this work, we will use $p_K \geq 2$ in practice.

The discontinuous finite element space $V_{h,\mathbf{p}}$ is defined by

$$V_{h,\mathbf{p}} \coloneqq \left\{ v \in L^2(\Omega) \colon v|_K \in \mathcal{P}_{p_K}(K) \,\forall\, K \in \mathcal{T}_h \right\}. \tag{2.14}$$

Let $\mathbf{s} \coloneqq (s_K \colon K \in \mathcal{T}_h)$ denote a vector of non-negative real numbers. The broken Sobolev space $H^{\mathbf{s}}(\Omega; \mathcal{T}_h)$ is defined by

$$H^{\mathbf{s}}(\Omega; \mathcal{T}_h) \coloneqq \left\{ v \in L^2(\Omega) \colon v|_K \in H^{s_K}(K) \,\forall\, K \in \mathcal{T}_h \right\}. \tag{2.15}$$

For $s \geq 0$, we set $H^s(\Omega; \mathcal{T}_h) \coloneqq H^{\mathbf{s}}(\Omega; \mathcal{T}_h)$, where $s_K = s$ for all $K \in \mathcal{T}_h$. The norm $\|\cdot\|_{H^{\mathbf{s}}(\Omega;\mathcal{T}_h)}$ and seminorm $|\cdot|_{H^{\mathbf{s}}(\Omega;\mathcal{T}_h)}$ are defined on $H^{\mathbf{s}}(\Omega; \mathcal{T}_h)$ as

$$\|v\|^2_{H^{\mathbf{s}}(\Omega;\mathcal{T}_h)} \coloneqq \sum_{K \in \mathcal{T}_h} \|v\|^2_{H^{s_K}(K)}, \quad |v|^2_{H^{\mathbf{s}}(\Omega;\mathcal{T}_h)} \coloneqq \sum_{K \in \mathcal{T}_h} |v|^2_{H^{s_K}(K)}. \tag{2.16}$$

For a $v_h \in V_{h,\mathbf{p}}$, we shall denote its element-wise broken gradient by ∇v_h, even though the discontinuous function v_h need not have weak derivatives on Ω.

Traces. It will be helpful to briefly review the construction of certain traces [40]. For each face $F \in \mathcal{F}_h^{i,b}$, let $n_F \in \mathbb{R}^d$ denote a *fixed* choice of a unit normal vector to F. Since F is flat, n_F is constant over F. Let K be an element of $\mathcal{T}_h$ for which $F \subset \partial K$; then n_F is either inward or outward pointing with respect to K. Since n_F is constant over F, n_F extends trivially as a constant vector field over $\overline{K}$. Let $\tau_F \colon H^s(K) \to H^{s-1/2}(F)$, $s > 1/2$, denote the trace operator from K to F. The trace operator τ_F is extended componentwise to vector-valued functions. Then, for $v \in H^s(K)$, $s > 3/2$, the normal derivative of v on F is defined by

$$\tau_F \frac{\partial v}{\partial n_F} \coloneqq \tau_F \left(\nabla v \cdot n_F \right), \tag{2.17}$$

where we use the fact that, after extending n_F to a constant vector field on $\overline{K}$, the function $\nabla v \cdot n_F \in H^{s-1}(K)$ belongs to the domain of τ_F.

Jump and average operators. For each face F, define the jump operator $[\![\cdot]\!]$ and the average operator $\{\cdot\}$ by

$$\begin{aligned}
&[\![\phi]\!] \coloneqq \tau_F\left(\phi|_{K_{\text{ext}}}\right) - \tau_F\left(\phi|_{K_{\text{int}}}\right), && \{\phi\} \coloneqq \tfrac{1}{2}\tau_F\left(\phi|_{K_{\text{ext}}}\right) + \tfrac{1}{2}\tau_F\left(\phi|_{K_{\text{int}}}\right), && \text{if } F \in \mathcal{F}_h^i,\\
&[\![\phi]\!] \coloneqq \tau_F\left(\phi|_{K_{\text{ext}}}\right), && \{\phi\} \coloneqq \tau_F\left(\phi|_{K_{\text{ext}}}\right), && \text{if } F \in \mathcal{F}_h^b,
\end{aligned}$$

where ϕ is a sufficiently regular scalar- or vector-valued function, and K_{ext} and K_{int} are the elements to which F is a face, i.e. $F = \partial K_{\text{ext}} \cap \partial K_{\text{int}}$. Here, the labelling is chosen so that n_F is outward pointing with respect to K_{ext} and inward pointing with respect to K_{int}, see Figure 2.1. In the case of an interior face, the jump and average are independent

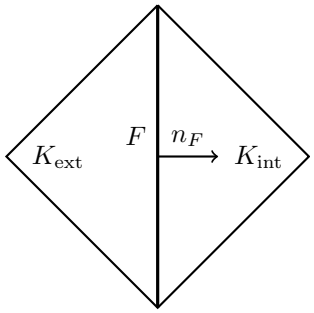

FIGURE 2.1: *Diagram for the notation of jump and average operators. For a face $F \in \mathcal{F}_h^i$, and a chosen normal vector n_F, K_{int} is the element for which n_F is inward pointing, and K_{ext} is the element for which n_F is outward pointing.*

of the choice of n_F. Using this notation, the jump and average of scalar-valued functions, respectively vector-valued functions, are also scalar-valued, respectively vector-valued.

Tangential differential operators. For $F \in \mathcal{F}_h^{i,b}$, let $H_{\mathrm{T}}^s(F)$ denote the space of H^s-regular tangential vector fields on F, thus $H_{\mathrm{T}}^s(F) \coloneqq \{v \in H^s(F)^d \colon v \cdot n_F = 0 \text{ on } F\}$. We define the tangential gradient $\nabla_{\mathrm{T}} \colon H^s(F) \to H_{\mathrm{T}}^{s-1}(F)$ and the tangential divergence $\operatorname{div}_{\mathrm{T}} \colon H_{\mathrm{T}}^s(F) \to H^{s-1}(F)$, where $s \geq 1$, following [40]. Let $\{t_i\}_{i=1}^{d-1} \subset \mathbb{R}^d$ be an orthonormal coordinate system on F. Then, for $u \in H^s(F)$ and $v \in H_{\mathrm{T}}^s(F)$ such that $v = \sum_{i=1}^{d-1} v_i \, t_i$, with $v_i \in H^s(F)$ for $i = 1, \dots, d-1$, we define

$$\nabla_{\mathrm{T}} u \coloneqq \sum_{i=1}^{d-1} t_i \frac{\partial u}{\partial t_i}, \quad \operatorname{div}_{\mathrm{T}} v \coloneqq \sum_{i=1}^{d-1} \frac{\partial v_i}{\partial t_i}. \tag{2.18}$$

The next lemma implies that traces and tangential differential operators commute.

Lemma 2.4. *Let Ω be a bounded polytopal domain, and let $\mathcal{T}_h$ be a mesh on Ω consisting of simplices or parallelepipeds. Then, for each $K \in \mathcal{T}_h$ and each face $F \subset \partial K$, the following identities hold:*

$$\tau_F(\nabla v) = \nabla_{\mathrm{T}}(\tau_F v) + \left(\tau_F \frac{\partial v}{\partial n_F}\right) n_F \qquad \forall\, v \in H^s(K),\ s > \tfrac{3}{2}, \tag{2.19}$$

$$\tau_F(\Delta v) = \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}}(\tau_F v) + \tau_F \frac{\partial}{\partial n_F}(\nabla v \cdot n_F) \qquad \forall\, v \in H^s(K),\ s > \tfrac{5}{2}. \tag{2.20}$$

Proof. First, observe that the terms in (2.19) and (2.20) are independent of the choice of n_F, since a reversal in the sign of n_F leaves the right-hand sides of these equations unchanged. Recall that F is flat, so, after a suitable change of coordinate system, we may assume without loss of generality that $K \subset \mathbb{R}^d_- \coloneqq \{(x, x') \colon x \in \mathbb{R}^{d-1},\ x' \leq 0\}$ and that $F \subset \partial\mathbb{R}^d_- = \{(x, 0) \colon x \in \mathbb{R}^{d-1}\}$. Since the identities (2.19) and (2.20) are independent of the choice of unit normal n_F, we may assume that $n_F = e_d = (0, \dots, 0, 1)$.

Let $s > 3/2$; for $i \in \{1, \dots, d-1\}$ we have the identity

$$\tau_F \frac{\partial v}{\partial x_i} = \frac{\partial}{\partial x_i}(\tau_F v) \quad \forall\, v \in H^s(K). \tag{2.21}$$

Indeed, this identity is valid for a smooth function v and thus extends to general $v \in H^s(K)$, for $s > 3/2$, by construction of the trace operator. So, $v \in H^s(K)$ satisfies

$$\nabla v = \sum_{i=1}^{d-1} \frac{\partial v}{\partial x_i} e_i + \frac{\partial v}{\partial x_d} e_d = \nabla_{\mathrm{T}}\, v + (\nabla v \cdot n_F)\, n_F \quad \text{in } K,$$

and we use the linearity of the trace operator with (2.21) to obtain

$$\tau_F\, (\nabla v) = \nabla_{\mathrm{T}}\, (\tau_F v) + \left(\tau_F \frac{\partial v}{\partial n_F}\right) n_F,$$

thus establishing (2.19). Similarly, for $v \in H^{s+1}(K)$, we write

$$\Delta v = \sum_{i=1}^{d-1} \frac{\partial^2 v}{\partial x_i^2} + \frac{\partial^2 v}{\partial x_d^2} = \mathrm{div}_{\mathrm{T}}\, \nabla_{\mathrm{T}}\, v + n_F \cdot \nabla(\nabla v \cdot n_F) \quad \text{in } K,$$

where the last equality follows from the fact that n_F is constant over K. Then, (2.20) is found by applying the trace operator to both sides of the previous identity and repeatedly applying (2.21) to v and its first tangential derivatives. □

Extensions of the results of this work to meshes with curved elements may make use of generalisations of Lemma 2.4 found in [40, p. 136].

Mesh-dependent norms. For two matrices $A, B \in \mathbb{R}^{d\times d}$, we set $A : B \coloneqq \sum_{i,j=1}^{d} A_{ij} B_{ij}$. For an element K, we define the bilinear form $\langle \cdot, \cdot \rangle_K$ by

$$\langle u, v \rangle_K \coloneqq \begin{cases} \int_K u\, v \,\mathrm{d}x & \text{if } u, v \in L^2(K), \\ \int_K u \cdot v \,\mathrm{d}x & \text{if } u, v \in L^2\left(K; \mathbb{R}^d\right), \\ \int_K u : v \,\mathrm{d}x & \text{if } u, v \in L^2\left(K; \mathbb{R}^{d\times d}\right). \end{cases} \tag{2.22}$$

The abuse of notation will be resolved by the arguments of the bilinear form. The bilinear forms $\langle \cdot, \cdot \rangle_{\partial K}$ and $\langle \cdot, \cdot \rangle_F$, $F \in \mathcal{F}_h^{i,b}$, are defined in a similar way.

For face-dependent positive real numbers μ_F and η_F to be specified later, let the jump stabilisation bilinear form $J_h \colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}}$ be defined by

$$\begin{aligned} J_h(u_h, v_h) \coloneqq & \sum_{F \in \mathcal{F}_h^i} \mu_F \langle [\![\nabla u_h \cdot n_F]\!], [\![\nabla v_h \cdot n_F]\!] \rangle_F \\ & + \sum_{F \in \mathcal{F}_h^{i,b}} \left[\mu_F \langle [\![\nabla_{\mathrm{T}}\, u_h]\!], [\![\nabla_{\mathrm{T}}\, v_h]\!] \rangle_F + \eta_F \langle [\![u_h]\!], [\![v_h]\!] \rangle_F \right]. \end{aligned} \tag{2.23}$$

We define the jump seminorm $|\cdot|_{\mathrm{J}}$ and the mesh-dependent norms $\|\cdot\|_{h,\theta}$, $\theta \in [0,1]$, by

$$|v_h|_{\mathrm{J}}^2 \coloneqq J_h(v_h, v_h), \quad \|v_h\|_{h,\theta}^2 \coloneqq \sum_{K\in\mathcal{T}_h} \Big[\theta |v_h|_{H^2(K)}^2 + (1-\theta)\|\Delta v_h\|_{L^2(K)}^2\Big] + |v_h|_{\mathrm{J}}^2. \tag{2.24}$$

It is straightforward to show that $\|\cdot\|_{h,\theta}$ defines a norm on $V_{h,\mathbf{p}}$ for any $\theta \in [0,1]$; see [70] for details.

Remark 2.3. In the following analysis, we will frequently use the norms $\|\cdot\|_{h,\theta}$ with the choice $\theta = 1/2$ and $\theta = 1$, denoted respectively by $\|\cdot\|_{h,1/2}$ and $\|\cdot\|_{h,1}$. We wish to emphasise that these norms should not be confused with the broken Sobolev norms for $H^{1/2}(\Omega;\mathcal{T}_h)$ and $H^1(\Omega;\mathcal{T}_h)$, which are denoted respectively by $\|\cdot\|_{H^{1/2}(\Omega;\mathcal{T}_h)}$ and $\|\cdot\|_{H^1(\Omega;\mathcal{T}_h)}$.

For each face $F \in \mathcal{F}_h^{i,b}$, define

$$\tilde{h}_F \coloneqq \begin{cases} \min(h_K, h_{K'}) & \text{if } F \in \mathcal{F}_h^i, \\ h_K & \text{if } F \in \mathcal{F}_h^b, \end{cases} \qquad \tilde{p}_F \coloneqq \begin{cases} \max(p_K, p_{K'}) & \text{if } F \in \mathcal{F}_h^i, \\ p_K & \text{if } F \in \mathcal{F}_h^b, \end{cases} \tag{2.25}$$

where K and K' are such that $F = \partial K \cap \partial K'$ if $F \in \mathcal{F}_h^i$ or $F \subset \partial K \cap \partial\Omega$ if $F \in \mathcal{F}_h^b$.

The assumptions on the mesh and the polynomial degrees, in particular (2.12) and (2.13), show that if F is a face of an element K, then

$$h_K \le c_{\mathcal{T}}\, \tilde{h}_F \quad \text{and} \quad \tilde{p}_F \le c_{\mathcal{P}}\, p_K. \tag{2.26}$$

Lemma 2.5. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex polytopal domain and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of meshes satisfying* (2.11)*. Then, for any $\theta \in [0,1]$, we have*

$$\|v_h\|_{H^1(\Omega;\mathcal{T}_h)}^2 \lesssim \|v_h\|_{h,\theta}^2 \quad \forall\, v_h \in V_{h,\mathbf{p}}, \tag{2.27}$$

whenever

$$\mu_F \gtrsim \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad \eta_F \gtrsim \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad \forall\, F \in \mathcal{F}_h^{i,b}. \tag{2.28}$$

Proof. First, it is sufficient to show (2.27) for $\theta = 0$ since $\|\Delta v_h\|_{L^2(K)} \lesssim \|D^2 v_h\|_{L^2(K)}$ for any $v_h \in V_{h,\mathbf{p}}$. Now, let $v_h \in V_{h,\mathbf{p}}$ be arbitrary, and recall the broken Poincaré inequality

$$\|v_h\|_{L^2(\Omega)}^2 \lesssim \sum_{K\in\mathcal{T}_h} \|\nabla v_h\|_{L^2(K)}^2 + \sum_{F\in\mathcal{F}_h^{i,b}} \frac{1}{\tilde{h}_F} \|[\![v_h]\!]\|_{L^2(F)}^2. \tag{2.29}$$

Note that the convexity assumption on Ω allows a simple proof of (2.29) by a duality argument. Therefore, the hypothesis on η_F implies that

$$\|v_h\|_{H^1(\Omega;\mathcal{T}_h)}^2 = \sum_{K\in\mathcal{T}_h} \|v_h\|_{H^1(K)}^2 \lesssim \sum_{K\in\mathcal{T}_h} \|\nabla v_h\|_{L^2(K)}^2 + |v_h|_{\mathrm{J}}^2. \tag{2.30}$$

Integration by parts gives

$$\sum_{K\in\mathcal{T}_h}\|\nabla v_h\|^2_{L^2(K)} = \sum_{K\in\mathcal{T}_h}\langle v_h, -\Delta v_h\rangle_K + \sum_{F\in\mathcal{F}_h^{i,b}}\langle[\![v_h]\!],\{\nabla v_h\cdot n_F\}\rangle_F + \sum_{F\in\mathcal{F}_h^{i}}\langle\{v_h\},[\![\nabla v_h\cdot n_F]\!]\rangle_F. \tag{2.31}$$

Hence, the trace and inverse inequalities imply that

$$\|v_h\|^2_{H^1(\Omega;\mathcal{T}_h)} \lesssim \sum_{K\in\mathcal{T}_h}\|\Delta v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^{i}}\frac{\tilde{p}_F^2}{\tilde{h}_F}\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)} + \sum_{F\in\mathcal{F}_h^{i,b}}\frac{\tilde{p}_F^2}{\tilde{h}_F}\|[\![v_h]\!]\|^2_{L^2(F)}. \tag{2.32}$$

Therefore, we have (2.27) whenever $\mu_F \gtrsim \tilde{p}_F^2/\tilde{h}_F$ and $\eta_F \gtrsim \tilde{p}_F^2/\tilde{h}_F$ for all $F \in \mathcal{F}_h^{i,b}$. □

Remark 2.4. We point out that the assumptions on the stabilisation parameters μ_F and η_F required by (2.28) are not restrictive in practice, since μ_F and η_F are user-defined parameters in the numerical scheme. We refer the reader to the numerical experiments of section 2.7 for examples of specific choices of these parameters in practice.

The main consequence of Lemma 2.5 is that for appropriately chosen η_F and μ_F, the norm $\|\cdot\|_{h,\theta}$ bounds the broken H^1-norm for any $\theta \in [0,1]$, and in the case of $\theta > 0$, the norm $\|\cdot\|_{h,\theta}$ also bounds the broken H^2-norm, i.e. we have $\|v_h\|_{H^2(\Omega;\mathcal{T}_h)} \lesssim \|v_h\|_{h,\theta}$ for any $\theta \in (0,1]$ and any $v_h \in V_{h,\mathbf{p}}$.

2.3 Numerical scheme

Section 2.1 shows that the analysis of the continuous problem essentially rests upon the Cordes condition and the Miranda–Talenti inequality, and it suggests the possibility of discretising the bilinear form of (2.9). However, the Miranda–Talenti inequality is not applicable for discretisation spaces that are not H^2-conforming, as is the case for $V_{h,\mathbf{p}}$. As a result, the design of the numerical scheme must resolve the key question of achieving a discrete analogue of coercivity bound (2.10) to guarantee stability.

Fortunately, the proof of the Miranda–Talenti inequality, as found for instance in [40, Theorem 3.1.1.1], is based on an integration-by-parts identity for the Laplacian of a sufficiently smooth function. Keeping to the principle that nonconforming methods such as DGFEM usually achieve stability by weakly enforcing properties of the approximated function space through stabilisation terms, it is therefore natural to weakly enforce a discrete version of this identity by adding its residual to the numerical scheme.

First, we establish the relevant discrete identity and its consistency in section 2.4. Then, we show in section 2.5 that including its residual in the numerical method results in an appropriate discrete coercivity bound. It is in this sense that the method is stable. Section 2.6 then uses the consistency and stability properties of the method for obtaining error bounds.

To define the numerical scheme, we use the following auxiliary bilinear forms. First, let $B_{h,*}\colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ be defined by

$$\begin{aligned} B_{h,*}(u_h, v_h) \coloneqq & \sum_{K\in\mathcal{T}_h} \langle D^2 u_h, D^2 v_h\rangle_K \\ & + \sum_{F\in\mathcal{F}_h^i} \left[\langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{u_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F + \langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{v_h\}, [\![\nabla u_h \cdot n_F]\!]\rangle_F\right] \\ & - \sum_{F\in\mathcal{F}_h^{i,b}} \left[\langle \nabla_{\mathrm{T}}\{\nabla u_h \cdot n_F\}, [\![\nabla_{\mathrm{T}} v_h]\!]\rangle_F + \langle \nabla_{\mathrm{T}}\{\nabla v_h \cdot n_F\}, [\![\nabla_{\mathrm{T}} u_h]\!]\rangle_F\right], \end{aligned} \tag{2.33}$$

where u_h, v_h will denote functions in $V_{h,\mathbf{p}}$ throughout this work, and $D^2 u_h$ denotes the broken Hessian of u_h. For each $\theta \in [0,1]$, define the bilinear form $B_{h,\theta}\colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ by

$$B_{h,\theta}(u_h, v_h) = \theta\, B_{h,*}(u_h, v_h) + (1-\theta) \sum_{K\in\mathcal{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K + J_h(u_h, v_h), \tag{2.34}$$

where we recall that the bilinear form J_h is defined in (2.23).

The bilinear form $A_h\colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ is defined by

$$A_h(u_h, v_h) \coloneqq \sum_{K\in\mathcal{T}_h} \langle \gamma L u_h, \Delta v_h\rangle_K + B_{h,1/2}(u_h, v_h) - \sum_{K\in\mathcal{T}_h} \langle \Delta u_h, \Delta v_h\rangle_K. \tag{2.35}$$

The numerical scheme for approximating the solution of (2.4) is to find $u_h \in V_{h,\mathbf{p}}$ such that

$$A_h(u_h, v_h) = \sum_{K\in\mathcal{T}_h} \langle \gamma f, \Delta v_h\rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{2.36}$$

We note that the original idea of including the last two terms in (2.35) was first proposed in our paper [70]. It is also in this respect that our method distinguishes itself from previous nonmonotone methods.

If $\mathbf{p} \equiv 1$, i.e. $p_K = 1$ for all $K \in \mathcal{T}_h$, then all terms in $A_h(u_h, v_h)$ vanish except for the jump stabilisation terms of $J_h(u_h, v_h)$. In this case, the numerical solution is $u_h \equiv 0$. This suggests that at least quadratic polynomials ought to be employed. Nevertheless, this still compares favourably with conforming elements, because, for instance, Argyris elements require at least polynomials of degree five on simplicial meshes in two dimensions [17].

2.4 Consistency

We turn to the question of consistency of the scheme (2.36) with respect to the original problem (2.4). It will be seen below that a discrete analogue of the identity of [40, Theorem 3.1.1.1] is central to the analysis of the numerical scheme. The following original result, first shown in [70], establishes the broken form of this identity.

Lemma 2.6. *Let Ω be a bounded Lipschitz polytopal domain, and let $\mathcal{T}_h$ be a simplicial or parallelepipedal mesh. Let $w \in H^s(\Omega;\mathcal{T}_h) \cap H^2(\Omega) \cap H^1_0(\Omega)$, with $s > 5/2$. Then, for every $v_h \in V_{h,\mathbf{p}}$, we have the identities*

$$B_{h,*}(w,v_h) \coloneqq \sum_{K\in\mathcal{T}_h} \langle \Delta w, \Delta v_h\rangle_K \quad \textit{and} \quad J_h(w,v_h) = 0. \tag{2.37}$$

Proof. Let w satisfy the above assumptions and let $v_h \in V_{h,\mathbf{p}}$. Then the second statement in (2.37) is trivial. Now, consider an element $K \in \mathcal{T}_h$, and let $\overline{n}$ be the piecewise constant outward normal on ∂K, momentarily assuming that $w \in H^3(K)$. Then, for $1 \le i,j \le d$, integration by parts gives

$$\begin{aligned}\int_K w_{x_ix_j}(v_h)_{x_ix_j}\,\mathrm{d}x &= \int_{\partial K} w_{x_ix_j}\overline{n}_i(v_h)_{x_j}\,\mathrm{d}s - \int_K w_{x_ix_jx_i}(v_h)_{x_j}\,\mathrm{d}x \\ &= \int_K w_{x_ix_i}(v_h)_{x_jx_j}\,\mathrm{d}x - \int_{\partial K}\left[w_{x_ix_i}\overline{n}_j(v_h)_{x_j} - w_{x_ix_j}\overline{n}_i(v_h)_{x_j}\right]\mathrm{d}s.\end{aligned} \tag{2.38}$$

Summing (2.38) over i,j and using the fact that $\overline{n}$ is piecewise constant over ∂K, we obtain

$$\langle D^2 w, D^2 v_h\rangle_K + \langle \Delta w, \nabla v_h\cdot\overline{n}\rangle_{\partial K} - \langle \nabla(\nabla w\cdot\overline{n}), \nabla v_h\rangle_{\partial K} = \langle \Delta w, \Delta v_h\rangle_K. \tag{2.39}$$

A density argument shows that (2.39) holds for $w \in H^s(K)$, $s > 5/2$. Note that for each face $F \subset \partial K$, $\overline{n} = \pm n_F$ on F. Also, for each face $F \subset \partial K$, identity (2.19) gives

$$\begin{aligned}\langle \nabla(\nabla w\cdot n_F), \nabla v_h\rangle_F &= \int_F \tau_F(\nabla(\nabla w\cdot n_F))\cdot\tau_F(\nabla v_h)\,\mathrm{d}s \\ &= \int_F \nabla_{\mathrm{T}}(\tau_F(\nabla w\cdot n_F))\cdot\nabla_{\mathrm{T}}(\tau_F v_h) + \left(\tau_F\frac{\partial}{\partial n_F}(\nabla w\cdot n_F)\right)\left(\tau_F\frac{\partial v_h}{\partial n_F}\right)\mathrm{d}s.\end{aligned} \tag{2.40}$$

For each face $F \subset \partial K$, identity (2.20) gives

$$\begin{aligned}\langle \Delta w, \nabla v_h\cdot n_F\rangle_F &= \int_F \tau_F(\Delta w)\,\tau_F(\nabla v_h\cdot n_F)\,\mathrm{d}s \\ &= \int_F \left(\mathrm{div}_{\mathrm{T}}\,\nabla_{\mathrm{T}}(\tau_F w) + \tau_F\frac{\partial}{\partial n_F}(\nabla w\cdot n_F)\right)\left(\tau_F\frac{\partial v_h}{\partial n_F}\right)\mathrm{d}s.\end{aligned} \tag{2.41}$$

Substituting (2.40) and (2.41) into (2.39) and summing over all $K \in \mathcal{T}_h$ shows that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\langle \Delta w, \Delta v_h\rangle_K &= \sum_{K\in\mathcal{T}_h}\langle D^2 w, D^2 v_h\rangle_K \\ &\quad + \sum_{F\in\mathcal{F}_h^{i,b}}\int_F [\![(\mathrm{div}_{\mathrm{T}}\,\nabla_{\mathrm{T}}\,w)\,(\nabla v_h\cdot n_F) - \nabla_{\mathrm{T}}(\nabla w\cdot n_F)\cdot\nabla_{\mathrm{T}}\,v_h]\!]\,\mathrm{d}s.\end{aligned} \tag{2.42}$$

For an interior face $F \in \mathcal{F}_h^i$ and for $w \in H^s(\Omega;\mathcal{T}_h)\cap H^2(\Omega)$, we use the facts that the trace operator commutes with tangential differential operators and that $[\![w]\!] = 0$ on the interior

face F to obtain

$$[\![\operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} w]\!] = \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} [\![w]\!] = 0 \quad \text{on } F.$$

Furthermore, $w \in H^2(\Omega)$ implies $[\![\nabla w]\!] = 0$ on the interior face F; therefore,

$$[\![\nabla_{\mathrm{T}}(\nabla w \cdot n_F)]\!] = \nabla_{\mathrm{T}} [\![\nabla w \cdot n_F]\!] = 0 \quad \text{on } F.$$

So, for any interior face $F \in \mathcal{F}_h^i$, it is found that

$$\begin{aligned}[\![(\operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} w)\,(\nabla v_h \cdot n_F) - \nabla_{\mathrm{T}}(\nabla w \cdot n_F) \cdot \nabla_{\mathrm{T}} v_h]\!] \\ = (\operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} \{w\})\,[\![\nabla v_h \cdot n_F]\!] - \nabla_{\mathrm{T}}\{\nabla w \cdot n_F\} \cdot [\![\nabla_{\mathrm{T}} v_h]\!].\end{aligned}$$

For a boundary face $F \in \mathcal{F}_h^b$, the trace $\tau_F w = 0$ on F because $w \in H_0^1(\Omega)$, and thus $\operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} w = 0$ on F. As a result,

$$\begin{aligned}[\![(\operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} w)\,(\nabla v_h \cdot n_F) - \nabla_{\mathrm{T}}(\nabla w \cdot n_F) \cdot \nabla_{\mathrm{T}} v_h]\!] &= -\nabla_{\mathrm{T}}\big(\tau_F(\nabla w \cdot n_F)\big) \cdot \nabla_{\mathrm{T}}\,(\tau_F v_h) \\ &= -\nabla_{\mathrm{T}}\{\nabla w \cdot n_F\} \cdot [\![\nabla_{\mathrm{T}} v_h]\!].\end{aligned}$$

Substituting the above simplifications into (2.42) shows that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \langle D^2 w, D^2 v_h\rangle_K + \sum_{F\in\mathcal{F}_h^i} \langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{w\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F \\ - \sum_{F\in\mathcal{F}_h^{i,b}} \langle \nabla_{\mathrm{T}}\{\nabla w \cdot n_F\}, [\![\nabla_{\mathrm{T}} v_h]\!]\rangle_F = \sum_{K\in\mathcal{T}_h} \langle \Delta w, \Delta v_h\rangle_K.\end{aligned} \tag{2.43}$$

It follows from the hypotheses on w that $[\![\nabla w \cdot n_F]\!]$ vanishes on any interior face and that $[\![\nabla_{\mathrm{T}} w]\!]$ vanishes on any face. Therefore,

$$\sum_{F\in\mathcal{F}_h^i} \langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{v_h\}, [\![\nabla w \cdot n_F]\!]\rangle - \sum_{F\in\mathcal{F}_h^{i,b}} \langle \nabla_{\mathrm{T}}\{\nabla v_h \cdot n_F\}, [\![\nabla_{\mathrm{T}} w]\!]\rangle_F = 0. \tag{2.44}$$

Identity (2.37) then follows from (2.43) and (2.44). □

Recalling the definition of $B_{h,\theta}$ in (2.34), it is clear that if a function w satisfies the hypotheses of Lemma 2.6, then, for any $\theta \in [0,1]$, we have

$$B_{h,\theta}(w, v_h) = \sum_{K\in\mathcal{T}_h} \langle \Delta w, \Delta v_h\rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{2.45}$$

Therefore, recalling the definition of A_h in (2.35), we obtain the following consistency result first given in [70].

Corollary 2.7. *Let Ω be a bounded convex polytopal domain, let $\mathcal{T}_h$ be a simplicial or parallelepipedal mesh, and let $u \in H^2(\Omega) \cap H_0^1(\Omega)$ be the unique solution of (2.4). If the*

solution $u \in H^s(\Omega; \mathcal{T}_h)$, $s > 5/2$, then we have

$$A_h(u, v_h) = \sum_{K \in \mathcal{T}_h} \langle \gamma f, \Delta v_h \rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{2.46}$$

An important implication of the consistency of the proposed method is that it does not introduce artificial fourth-order perturbations, as opposed to the vanishing moment method described in section 1.4.

2.5 Stability

In this section, we show that the choice of discrete bilinear form in (2.35) reproduces the coercivity of the continuous bilinear form of (2.9). This guarantees the stability of the numerical scheme, including well-posedness of the discrete problem (2.36).

Lemma 2.8. *Let Ω be a bounded convex polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11)*. Then, for each constant $\kappa > 1$, there exists a positive constant c_μ, independent of h, $\mathbf{p}$ and θ, such that*

$$B_{h,\theta}(v_h, v_h) \geq \frac{\theta}{\kappa} \sum_{K \in \mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + (1-\theta) \sum_{K \in \mathcal{T}_h} \|\Delta v_h\|^2_{L^2(K)} + \frac{1}{2}|v_h|^2_{\mathrm{J}} \tag{2.47}$$

for any $v_h \in V_{h,\mathbf{p}}$ and any $\theta \in [0,1]$, whenever

$$\mu_F = c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F} \quad \textit{and} \quad \eta_F > 0 \quad \forall\, F \in \mathcal{F}_h^{i,b}. \tag{2.48}$$

Proof. Let $v_h \in V_{h,\mathbf{p}}$, then we have

$$B_{h,\theta}(v_h, v_h) = \theta \sum_{K \in \mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + (1-\theta) \sum_{K \in \mathcal{T}_h} \|\Delta v_h\|^2_{L^2(K)} + |v_h|^2_{\mathrm{J}} + \theta \sum_{i=1}^{2} I_i, \tag{2.49}$$

where the quantities I_i are defined by

$$I_1 \coloneqq 2 \sum_{F \in \mathcal{F}_h^i} \langle \mathrm{div_T}\, \nabla_{\mathrm{T}} \{v_h\}, [\![\nabla v_h \cdot n_F]\!] \rangle_F, \quad I_2 \coloneqq 2 \sum_{F \in \mathcal{F}_h^{i,b}} \langle \nabla_{\mathrm{T}} \{\nabla v_h \cdot n_F\}, [\![\nabla_{\mathrm{T}}\, v_h]\!] \rangle_F.$$

For some $\delta > 0$ to be chosen below, the Cauchy–Schwarz inequality with a parameter gives

$$\begin{aligned} |I_1| &\leq 2 \sqrt{\sum_{F \in \mathcal{F}_h^i} \frac{\delta \tilde{h}_F}{\tilde{p}_F^2} \|\mathrm{div_T}\, \nabla_{\mathrm{T}} \{v_h\}\|^2_{L^2(F)}} \sqrt{\sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}} \\ &\leq \delta \sum_{F \in \mathcal{F}_h^i} \frac{\tilde{h}_F}{\tilde{p}_F^2} \|\mathrm{div_T}\, \nabla_{\mathrm{T}} \{v_h\}\|^2_{L^2(F)} + \sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}. \end{aligned}$$

Since the tangential differential operators commute with the trace operator, for each face $F = \partial K \cap \partial K'$, Young's inequality yields

$$\|\mathrm{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{v_h\}\|^2_{L^2(F)} \leq \frac{1}{2}\|\mathrm{div}_{\mathrm{T}} \nabla_{\mathrm{T}} v_h|_K\|^2_{L^2(F)} + \frac{1}{2}\|\mathrm{div}_{\mathrm{T}} \nabla_{\mathrm{T}} v_h|_{K'}\|^2_{L^2(F)}.$$

Therefore, the trace and inverse inequalities give

$$\delta \sum_{F\in\mathcal{F}_h^i} \frac{\tilde{h}_F}{\tilde{p}_F^2}\|\mathrm{div}_{\mathrm{T}} \nabla_{\mathrm{T}}\{v_h\}\|^2_{L^2(F)} \leq \delta \sum_{F\in\mathcal{F}_h^i} \frac{\tilde{h}_F}{\tilde{p}_F^2} \sum_{\substack{K\\ F\subset\partial K}} C_d\, C_{\mathrm{Tr}}\, \frac{p_K^2}{h_K}\|D^2 v_h\|^2_{L^2(K)},$$

where C_d is a constant depending only on the dimension d and where C_{Tr} is the constant of the trace and inverse inequality, which depends on the shape-regularity of $\mathcal{T}_h$. Since each element has at most $c_{\mathcal{F}}$ faces by (2.11), a counting argument shows that

$$\sum_{F\in\mathcal{F}_h^i} \sum_{\substack{K\\ F\subset\partial K}} \|D^2 v_h\|^2_{L^2(K)} \leq c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)}.$$

We then use the definitions of $\tilde{p}_F$ and $\tilde{h}_F$ from (2.25) to obtain

$$|I_1| \leq \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta\tilde{h}_F}\|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}. \tag{2.50}$$

A similar analysis shows that

$$|I_2| \leq \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^{i,b}} \frac{\tilde{p}_F^2}{\delta\tilde{h}_F}\|[\![\nabla_{\mathrm{T}} v_h]\!]\|^2_{L^2(F)}, \tag{2.51}$$

where C_d is a constant depending only on d. Inequalities (2.50) and (2.51) imply that

$$\begin{aligned} B_{h,\theta}(v_h, v_h) \geq{}& \theta(1 - 2\delta C(d) C_{\mathrm{Tr}} c_{\mathcal{F}}) \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + (1-\theta) \sum_{K\in\mathcal{T}_h} \|\Delta v_h\|^2_{L^2(K)} \\ &+ \sum_{F\in\mathcal{F}_h^i} \left(\mu_F - \frac{\theta\tilde{p}_F^2}{\delta\tilde{h}_F}\right)\|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)} + \sum_{F\in\mathcal{F}_h^{i,b}} \left(\mu_F - \frac{\theta\tilde{p}_F^2}{\delta\tilde{h}_F}\right)\|[\![\nabla_{\mathrm{T}} v_h]\!]\|^2_{L^2(F)} \\ &+ \sum_{F\in\mathcal{F}_h^{i,b}} \eta_F \|[\![v_h]\!]\|^2_{L^2(F)} \end{aligned}$$

Let $\kappa > 1$ be given. Then, since $\kappa^{-1} < 1$, there exists a $\delta > 0$ sufficiently small such that $(1 - 2\delta C_d C_{\mathrm{Tr}} c_{\mathcal{F}}) > \kappa^{-1}$. Then, we choose $c_\mu = 2\delta^{-1}$ and $\mu_F = c_\mu \tilde{p}_F^2/\tilde{h}_F$. Therefore $\mu_F - \theta\,\tilde{p}_F^2/\delta\tilde{h}_F \geq \mu_F/2$ for any $\theta \in [0,1]$, thereby completing the proof of (2.47). □

Lemma 2.8 ensures that it is possible to choose μ_F and η_F such that (2.47) holds for some $\kappa < (1-\varepsilon)^{-1}$, because $(1-\varepsilon)^{-1} > 1$. In this case, it is seen from the proof of

Lemma 2.8 that $\delta \lesssim \varepsilon$ and therefore the constant c_μ can be chosen to be of order $1/\varepsilon$.

Theorem 2.9. *Under the hypotheses of Lemma* 2.8*, let μ_F and η_F be chosen so that* (2.47) *and* (2.48) *hold with $\kappa < (1-\varepsilon)^{-1}$. Let the linear operator L defined by* (2.2) *satisfy the uniform ellipticity condition* (2.3) *and the Cordes condition* (2.5)*. Then, the bilinear form A_h is coercive on $V_{h,\mathbf{p}}$ with respect to the norm $\|\cdot\|_{h,1}$. In particular, for any $v_h \in V_{h,\mathbf{p}}$, there holds*

$$\|v_h\|_{h,1}^2 \leq \frac{2\kappa}{1-\kappa\,(1-\varepsilon)} A_h(v_h, v_h). \tag{2.52}$$

Therefore, there exists a unique solution $u_h \in V_{h,\mathbf{p}}$ of the numerical scheme (2.36)*. Moreover, u_h satisfies*

$$\|u_h\|_{h,1} \leq \frac{2\,\kappa\,\sqrt{d}\,\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa\,(1-\varepsilon)} \|f\|_{L^2(\Omega)}. \tag{2.53}$$

Proof. Let $v_h \in V_{h,\mathbf{p}}$ and note that (2.7) implies that

$$\begin{aligned}\langle \gamma L v_h, \Delta v_h\rangle_K - \langle \Delta v_h, \Delta v_h\rangle_K &= \langle (\gamma L - \Delta)\, v_h, \Delta v_h\rangle_K \\ &\leq \|(\gamma L - \Delta)\, v_h\|_{L^2(K)} \|\Delta v_h\|_{L^2(K)} \\ &\leq \sqrt{1-\varepsilon}\|D^2 v_h\|_{L^2(K)} \|\Delta v_h\|_{L^2(K)}.\end{aligned}$$

We use the Cauchy–Schwarz inequality with a parameter, together with the fact that (2.47) holds with $\kappa < (1-\varepsilon)^{-1}$, to get

$$\begin{aligned}A_h(v_h, v_h) \geq \frac{1}{2} \sum_{K\in\mathcal{T}_h} \left[\frac{1}{\kappa}\|D^2 v_h\|_{L^2(K)}^2 + \|\Delta v_h\|_{L^2(K)}^2\right] + \frac{1}{2}|v_h|_{\mathrm{J}}^2 \\ - \sum_{K\in\mathcal{T}_h} \sqrt{1-\varepsilon}\,\|D^2 v_h\|_{L^2(K)} \|\Delta v_h\|_{L^2(K)}.\end{aligned} \tag{2.54}$$

By applying the Cauchy inequality

$$\sqrt{1-\varepsilon}\|D^2 v_h\|_{L^2(K)} \|\Delta v_h\|_{L^2(K)} \leq \frac{1-\varepsilon}{2}\|D^2 v_h\|_{L^2(K)}^2 + \frac{1}{2}\|\Delta v_h\|_{L^2(K)}^2,$$

we thereby obtain from (2.54)

$$A_h(v_h, v_h) \geq \frac{1-\kappa\,(1-\varepsilon)}{2\,\kappa} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|_{L^2(K)}^2 + \frac{1}{2}|v_h|_{\mathrm{J}}^2 \geq \frac{1-\kappa\,(1-\varepsilon)}{2\,\kappa}\|v_h\|_{h,1}^2. \tag{2.55}$$

The previous inequality completes the proof of the coercivity bound (2.52), which in turn proves that there exists a unique solution $u_h \in V_{h,\mathbf{p}}$ of (2.36). Then, applying (2.52) to the

numerical solution u_h shows that

$$\|u_h\|_{h,1}^2 \leq \frac{2\,\kappa}{1-\kappa\,(1-\varepsilon)} A_h(u_h,u_h) \leq \frac{2\,\kappa}{1-\kappa\,(1-\varepsilon)} \sum_{K\in\mathcal{T}_h} |\langle \gamma f, \Delta u_h\rangle_K|.$$

Since $\|\Delta u_h\|_{L^2(K)} \leq \sqrt{d}\,\|D^2 u_h\|_{L^2(K)}$, it is found that (2.53) follows from

$$\|u_h\|_{h,1}^2 \leq \frac{2\,\kappa\,\sqrt{d}\,\|\gamma\|_{L^\infty(\Omega)}}{1-\kappa\,(1-\varepsilon)} \|f\|_{L^2(\Omega)} \|u_h\|_{h,1}. \qquad \square$$

2.6 Error analysis

The above stability results make use of the strict positivity requirement on the jump penalty terms η_F, $F \in \mathcal{F}_h^{i,b}$, as given in (2.48). In the following, we require the upper bound

$$\eta_F \leq c_\eta \frac{\tilde{p}_F^4}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}, \tag{2.56}$$

with $c_\eta > 0$ a fixed constant. As explained in Remark 2.4, this is simply a condition on the choice of the user-defined parameters η_F, and is not restrictive in practice. By combining (2.28), (2.48) and (2.56), it is seen that these parameters should be chosen to satisfy

$$\mu_F = c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad c_\eta \frac{\tilde{p}_F^2}{\tilde{h}_F} \leq \eta_F \leq c_\eta \frac{\tilde{p}_F^4}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}, \tag{2.57}$$

with c_μ and c_η user-defined constants to be chosen sufficiently large.

The consistency result of Corollary 2.7 involves a regularity assumption on the solution; for problems that satisfy this assumption, we shall speak of *sufficiently regular* solutions. In this case, we obtain error bounds that are explicit in the approximation orders in section 2.6.1. However, we remark from the onset that this regularity assumption is not necessary for convergence of the numerical method, as shown in section 2.6.2. Indeed, the numerical solution guarantees a *best approximation property* with respect to the H^2-conforming subspace of $V_{h,\mathbf{p}}$. In a nutshell, this property shows that our method is at least as accurate in H^2-type norms as an H^2-conforming method on the same mesh with the same polynomial degrees. Furthermore, we note that computations for problems that fail to satisfy the regularity assumption also indicate that the method is generally convergent under a wide range of choices of meshes and polynomial degrees: see the experiment of section 3.7.3.

2.6.1 Error bound for solutions with sufficient regularity

Theorem 2.10. *Let Ω be a bounded convex polytopal domain, and let the shape-regular sequence of simplicial or parallelepipedal meshes $\{\mathcal{T}_h\}_h$ satisfy* (2.11) *and* (2.12)*, with* $\mathbf{p}$

satisfying (2.13) *for each* h. *Let the linear operator* L *defined by* (2.2) *satisfy the uniform ellipticity condition* (2.3) *and the Cordes condition* (2.5). *Let* $u \in H^2(\Omega) \cap H_0^1(\Omega)$ *be the unique solution of* (2.4), *and assume that* $u \in H^{\mathbf{s}}(\Omega; \mathcal{T}_h)$, *with* $s_K > 5/2$ *for each* $K \in \mathcal{T}_h$. *Let* μ_F *and* η_F *be chosen as in Theorem* 2.9 *and as in* (2.56) *for all* $F \in \mathcal{F}_h^{i,b}$. *Then, there exists a constant* $C > 0$, *independent of* h, $\mathbf{p}$, *and* u, *but depending on* $\max_K s_K$, *such that*

$$\|u - u_h\|_{h,1}^2 \leq C \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K - 4}}{p_K^{2s_K - 5}} \|u\|_{H^{s_K}(K)}^2, \tag{2.58}$$

where $t_K = \min(s_K, p_K + 1)$ *for each* $K \in \mathcal{T}_h$.

Note that for the special case of quasi-uniform meshes and uniform polynomial degrees, the error bound (2.58) simplifies to

$$|u - u_h|_{H^2(\Omega;\mathcal{T}_h)} \leq \|u - u_h\|_{h,1} \lesssim \frac{h^{\min(s, p+1)-2}}{p^{s-5/2}} \|u\|_{H^s(\Omega;\mathcal{T}_h)}.$$

Thus it is seen that the rates are optimal with respect to the mesh size and suboptimal in the polynomial degree only by half an order. We remark that the standard analysis of *hp*-version DGFEM for divergence form elliptic equations leads to a similar suboptimality, and that optimal rates were recovered in [37] through considerations of regularity of the solution in augmented Sobolev spaces.

Proof. Theorem C.6 implies[1] that there exists a $z_h \in V_{h,\mathbf{p}}$, and a constant C, independent of u, h_K, and p_K, but dependent on $\max_K s_K$, such that, for each $K \in \mathcal{T}_h$, each nonnegative integer $j \leq s_K$, and for each multi-index β with $|\beta| < s_K - 1/2$, we have

$$\|u - z_h\|_{H^j(K)} \leq C \frac{h_K^{t_K - j}}{p_K^{s_K - j}} \|u\|_{H^{s_K}(K)}, \tag{2.59}$$

$$\|D^\beta (u - z_h)\|_{L^2(\partial K)} \leq C \frac{h_K^{t_K - |\beta| - 1/2}}{p_K^{s_K - |\beta| - 1/2}} \|u\|_{H^{s_K}(K)}. \tag{2.60}$$

Now, set $\psi_h = z_h - u_h$ and $\xi_h = z_h - u$. Since u satisfies the hypotheses of Corollary 2.7, it follows that (2.46) holds. So, coercivity of A_h from (2.52) and Corollary 2.7 imply that

$$\begin{aligned} \|\psi_h\|_{h,1}^2 &\lesssim A_h(z_h, \psi_h) - A_h(u_h, \psi_h) \\ &= A_h(z_h, \psi_h) - \sum_{K \in \mathcal{T}_h} \langle \gamma f, \Delta \psi_h \rangle_K = A_h(\xi_h, \psi_h). \end{aligned} \tag{2.61}$$

[1] The parameter m_K appearing in Theorem C.6 can be chosen such that $m_K = s_K - 1$ for each $K \in \mathcal{T}_h$.

Therefore, $\|\psi_h\|_{h,1}^2 \lesssim \sum_{i=1}^8 E_i$, where

$$E_1 \coloneqq \sum_{K\in\mathcal{T}_h} |\langle D^2\xi_h, D^2\psi_h\rangle_K|, \qquad E_5 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} |\langle \nabla_{\mathrm{T}}\{\nabla\xi_h\cdot n_F\}, [\![\nabla_{\mathrm{T}}\,\psi_h]\!]\rangle_F|,$$

$$E_2 \coloneqq \sum_{K\in\mathcal{T}_h} |\langle (\gamma L-\Delta)\xi_h, \Delta\psi_h\rangle_K|, \qquad E_6 \coloneqq \sum_{F\in\mathcal{F}_h^{i}} |\langle \mathrm{div}_{\mathrm{T}}\,\nabla_{\mathrm{T}}\{\xi_h\}, [\![\nabla\psi_h\cdot n_F]\!]\rangle_F|,$$

$$E_3 \coloneqq \sum_{K\in\mathcal{T}_h} |\langle \Delta\xi_h, \Delta\psi_h\rangle_K|, \qquad E_7 \coloneqq \sum_{F\in\mathcal{F}_h^{i}} |\langle \mathrm{div}_{\mathrm{T}}\,\nabla_{\mathrm{T}}\{\psi_h\}, [\![\nabla\xi_h\cdot n_F]\!]\rangle_F|,$$

$$E_4 \coloneqq |J_h(\xi_h,\psi_h)|, \qquad E_8 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} |\langle \nabla_{\mathrm{T}}\{\nabla\psi_h\cdot n_F\}, [\![\nabla_{\mathrm{T}}\,\xi_h]\!]\rangle_F|.$$

It is then deduced that

$$E_1+E_2+E_3 \lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}}\|u\|_{H^{s_K}(K)}^2}\;\|\psi_h\|_{h,1}. \tag{2.62}$$

Now, $E_4 \le |\xi_h|_{\mathrm{J}}|\psi_h|_{\mathrm{J}} \le (e_1+e_2+e_3)^{\frac{1}{2}}\,\|\psi_h\|_{h,1}$, where the quantities e_i are defined by

$$e_1 \coloneqq \sum_{F\in\mathcal{F}_h^{i}} \mu_F\|[\![\nabla\xi_h\cdot n_F]\!]\|_{L^2(F)}^2, \qquad e_2 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F\|[\![\nabla_{\mathrm{T}}\,\xi_h]\!]\|_{L^2(F)}^2,$$

$$e_3 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \eta_F\|[\![\xi_h]\!]\|_{L^2(F)}^2.$$

Recalling (2.25) and (2.48), we use (2.60) to obtain

$$e_1 \lesssim \sum_{F\in\mathcal{F}_h^{i}} \frac{\tilde{p}_F^2}{\tilde{h}_F} \sum_{\substack{K\\ F\subset\partial K}} \|\nabla\xi_h\|_{L^2(\partial K)}^2 \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}}\|u\|_{H^{s_K}(K)}^2.$$

Similarly, we use the hypothesis $\eta_F \lesssim \tilde{p}_F^4/\tilde{h}_F^3$ from (2.56) to find that

$$e_2+e_3 \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}}\|u\|_{H^{s_K}(K)}^2.$$

Therefore,

$$E_4 \lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}}\|u\|_{H^{2}s_K(K)}^2}\;\|\psi_h\|_{h,1}.$$

It is found that

$$E_5 + E_6 \lesssim \sqrt{\sum_{F\in\mathcal{F}_h^{i,b}} \frac{\tilde{h}_F}{\tilde{p}_F^2} \sum_{\substack{K\\ F\subset\partial K}} \|D^2\xi_h\|^2_{L^2(\partial K)}}\, \|\psi_h\|_{h,1}$$
$$\lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}} \|u\|^2_{H^{s_K}(K)}}\, \|\psi_h\|_{h,1}.$$

It follows from the inverse and trace inequalities that

$$E_7 + E_8 \lesssim \sqrt{e_1+e_2}\, \|\psi_h\|_{h,1} \lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|^2_{H^{s_K}(K)}}\, \|\psi_h\|_{h,1}.$$

The above inequalities imply that

$$\|u - z_h\|_{h,1} + \|z_h - u_h\|_{h,1} \lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|^2_{H^{s_K}(K)}}.$$

The triangle inequality and the above inequalities complete the proof of (2.58). □

2.6.2 Error bound for solutions with minimal regularity

In this paragraph, we provide an error bound that is valid for problems with minimal regularity [70], namely $u \in H^2(\Omega)\cap H_0^1(\Omega)$. This result consists of a quasi-optimal approximation property with respect to the $H^2(\Omega)\cap H_0^1(\Omega)$-conforming subspace of $V_{h,\mathbf{p}}$. By restricting the quasi-optimal approximation property to this conforming subspace, we are able to remove the assumption of $u \in H^s(\Omega;\mathcal{T}_h)$, $s > 5/2$, required by Theorem 2.10.

Theorem 2.11. *Let Ω be a bounded convex polytopal domain, and let the shape-regular sequences of simplicial or parallelepipedal meshes $\{\mathcal{T}_h\}_h$ satisfy* (2.11) *and* (2.12)*, with $\mathbf{p}$ satisfying* (2.13) *for each h. Let the linear operator L defined by* (2.2) *satisfy the uniform ellipticity condition* (2.3) *and the Cordes condition* (2.5)*. Let $u \in H^2(\Omega)\cap H_0^1(\Omega)$ be the unique solution of* (2.4)*, and let μ_F and η_F be chosen as in Theorem* 2.9*. Then,*

$$\|u - u_h\|_{h,1} \lesssim \inf\{|u - z_h|_{H^2(\Omega)} \colon z_h \in V_{h,\mathbf{p}} \cap H^2(\Omega)\cap H_0^1(\Omega)\}. \tag{2.63}$$

Proof. If $z_h \in V_{h,\mathbf{p}}\cap H^2(\Omega)\cap H_0^1(\Omega)$, then Lemma 2.6 applies to z_h, because z_h is a piecewise polynomial and thus $z_h \in H^s(\Omega;\mathcal{T}_h)$ for $s > 5/2$. Setting $\psi_h = z_h - u_h \in V_{h,\mathbf{p}}$, coercivity

of A_h gives

$$\|z_h - u_h\|_{h,1}^2 \lesssim A_h(z_h - u_h, \psi_h) = \sum_{K\in\mathcal{T}_h} \langle \gamma L z_h, \Delta\psi_h\rangle_K - \sum_{K\in\mathcal{T}_h} \langle \gamma f, \Delta\psi_h\rangle_K$$
$$= \sum_{K\in\mathcal{T}_h} \langle \gamma L(z_h - u), \Delta\psi_h\rangle_K \lesssim |u - z_h|_{H^2(\Omega)} \|z_h - u_h\|_{h,1}.$$

Thus $\|z_h - u_h\|_{h,1} \lesssim |u - z_h|_{H^2(\Omega)}$. Since the functions u and z_h belong to $H^2(\Omega) \cap H_0^1(\Omega)$, it follows that their jumps in values and gradients vanish on all interior faces, and that their values and tangential gradients vanish on all boundary faces. Therefore, we have the identity $\|u - z_h\|_{h,1} = |u - z_h|_{H^2(\Omega)}$. So, the triangle inequality gives

$$\|u - u_h\|_{h,1} \leq \|u - z_h\|_{h,1} + \|z_h - u_h\|_{h,1} \lesssim |u - z_h|_{H^2(\Omega)}.$$

Since z_h was arbitrary, taking the infimum over all $z_h \in V_{h,\mathbf{p}} \cap H^2(\Omega) \cap H_0^1(\Omega)$ completes the proof. □

The quasi-optimal approximation property indicates that convergence of the method can be guaranteed provided that the approximation space $V_{h,\mathbf{p}}$ is sufficiently rich, although computations suggest this assumption is also not strictly necessary. Although the question of finding the weakest assumptions on $V_{h,\mathbf{p}}$ that are sufficient for convergence remains open, we note that a similar issue in the context of divergence form elliptic equations has been largely resolved by discrete compactness methods; see [27] and the references therein.

A further direct consequence of Theorem 2.11 is that the scheme proposed here is at least as accurate, up to a constant, as any H^2-conforming method on the same mesh with the same polynomial degrees; this is simply because if $u_h^c \in V_{h,\mathbf{p}} \cap H^2(\Omega) \cap H_0^1(\Omega)$ is the numerical approximation of u produced by an arbitrary conforming numerical method, the bound (2.63) implies that $\|u - u_h\|_{h,1} \lesssim |u - u_h^c|_{H^2(\Omega)}$.

2.7 Numerical experiments

This section presents the numerical experiments from our paper [70]. In the first problem, we demonstrate the convergence rates predicted by Theorem 2.10 for a solution with limited global regularity but with sufficient broken regularity, and in the second experiment, we test the scheme under *hp*-refinement on a problem with a singular solution.

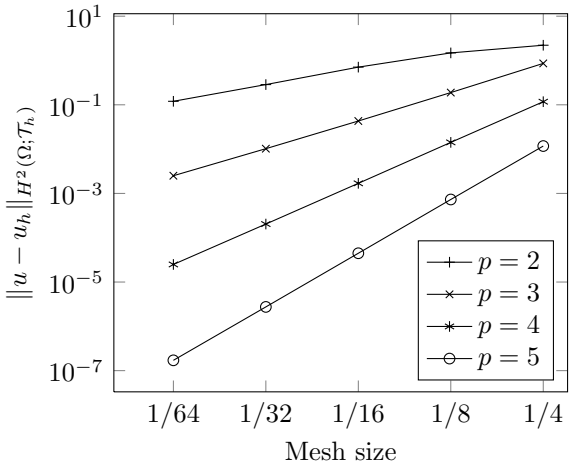

FIGURE 2.2: *Convergence rates for the numerical scheme applied to problem of section* 2.7.1. *The error* $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)}$ *is plotted against mesh size* h *for various polynomial degrees* p. *The optimal convergence rates* $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)} \simeq h^{p-1}$ *are observed.*

2.7.1 First experiment

Let $\Omega = (-1, 1)^2$, and consider the problem

$$\sum_{i,j=1}^{d} (1 + \delta_{ij}) \frac{x_i}{|x_i|} \frac{x_j}{|x_j|} u_{x_i x_j} = f \quad \text{in } \Omega, \tag{2.64}$$

$$u = 0 \quad \text{on } \partial\Omega.$$

Here, f is chosen so that the solution of (2.64) is $u(x, y) = (x\,\mathrm{e}^{1-|x|} - x)(y\,\mathrm{e}^{1-|y|} - y)$. Observe that the Cordes condition (2.5) holds with $\varepsilon = 3/5$ and that the coefficients of the differential operator in (2.64) are discontinuous across the set $\{(x, y) \in \Omega \colon x = 0 \text{ or } y = 0\}$.

We apply the numerical scheme (2.36) on meshes obtained by regular subdivision of Ω into uniform quadrilaterals of side-length $h = 2^{-k}$, $2 \le k \le 6$. It follows that $u \in H^s(\Omega; \mathcal{T}_h)$ for all $s > 5/2$, but $u \notin H^s(\Omega)$ for any $s > 5/2$. The finite element spaces $V_{h,\mathbf{p}}$ are defined by employing the space of polynomials of fixed total degree p on each element. For the choice of penalty parameters, we set $c_\mu = c_\eta = 10$ and set $\eta_F = c_\eta\, \tilde{p}_F^4 / \tilde{h}_F^3$. Figure 2.2 plots the errors for various polynomial degrees p, with $2 \le p \le 5$. The expected optimal convergence rates $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)} \simeq h^{p-1}$ are observed, in accordance with Theorem 2.10.

2.7.2 Second experiment

In this example, we demonstrate the robustness of the scheme by illustrating exponential accuracy for a problem that involves both nonsmoothness of the solution and discontinuity of the coefficients at a corner of the domain. We also show how to apply the numerical scheme to problems with inhomogeneous boundary conditions.

Mesh size	$p=2$		$p=3$		$p=4$		$p=5$	
1/4	2.21		8.60×10^{-1}		1.18×10^{-1}		1.17×10^{-2}	
1/8	1.48	(0.58)	1.89×10^{-1}	(2.18)	1.42×10^{-2}	(3.05)	7.30×10^{-4}	(4.01)
1/16	7.08×10^{-1}	(1.07)	4.31×10^{-2}	(2.13)	1.69×10^{-3}	(3.07)	4.46×10^{-5}	(4.03)
1/32	2.86×10^{-1}	(1.31)	1.02×10^{-2}	(2.07)	2.04×10^{-4}	(3.05)	2.74×10^{-6}	(4.02)
1/64	1.20×10^{-1}	(1.25)	2.51×10^{-3}	(2.02)	2.49×10^{-5}	(3.03)	1.71×10^{-7}	(4.00)

TABLE 2.3: *Errors $\|u-u_h\|_{H^2(\Omega;\mathcal{T}_h)}$ for the numerical scheme applied to problem of section 2.7.1. The estimated orders of convergence between successive mesh refinements are given in parentheses.*

It can be verified that for $\alpha>1$, $u=|x|^\alpha$, $x\in\Omega=(0,1)^2$, solves

$$\sum_{i,j=1}^{d}\left(\delta_{ij}+\frac{x_i\,x_j}{|x|^2}\right)u_{x_ix_j}=c_\alpha|x|^{\alpha-2}=:f\qquad\text{in }\Omega, \tag{2.65}$$

where c_α is a suitable constant depending only on α. Notice that the term $x_i\,x_j/|x|^2$ fails to be continuous at the origin when $i\neq j$. This example draws upon the examples in [38, 56] that illustrate the possibility of ill-posedness of the nondivergence form PDE with discontinous coefficients when the Cordes condition fails. However, the operator in (2.65) satisfies the Cordes condition (2.5) with $\varepsilon=4/5$. In the following, we take $\alpha=1.6$, so $u\in H^{2.6-\delta}(\Omega)$ for arbitrarily small δ.

In order to extend the numerical scheme (2.36) to problems with nonhomogeneous boundary conditions, the right-hand side must be suitably modified as follows. Let g be the restriction of u on $\partial\Omega$. Then the numerical scheme for problem (2.65) is to find $u_h\in V_{h,\mathbf{p}}$ such that for every $v_h\in V_{h,\mathbf{p}}$, there holds

$$\begin{aligned}A_h(u_h,v_h)=&\sum_{K\in\mathcal{T}_h}\langle\gamma\,f,\Delta v_h\rangle_K+\sum_{F\in\mathcal{F}_h^b}\left[\mu_F\langle\nabla_{\mathrm{T}}\,g,\nabla_{\mathrm{T}}\,v_h\rangle_F+\eta_F\langle g,v_h\rangle_F\right]\\&-\frac{1}{2}\sum_{F\in\mathcal{F}_h^b}\left[\langle\mathrm{div}_{\mathrm{T}}\,\nabla_{\mathrm{T}}\,g,\nabla v_h\cdot n_F\rangle_F+\langle\nabla_{\mathrm{T}}\left(\nabla v_h\cdot n_F\right),\nabla_{\mathrm{T}}\,g\rangle_F\right].\end{aligned} \tag{2.66}$$

Following [68], we construct a sequence of geometrically refined meshes as shown in see Figure 2.4. The polynomial degrees associated to the elements of the mesh are increasing linearly with the distance from the origin. Figure 2.5 plots the errors in the broken H^1-norm and H^2-seminorm against $\sqrt[3]{\mathrm{DoF}}$, where DoF is the number of degrees of freedom, and shows that a convergence rate of order $\exp(-c\sqrt[3]{\mathrm{DoF}})$ is achieved.

The optimality of the observed exponential convergence rate can be heuristically explained as follows: for an analytic function v on an element K, we have $\inf_{v_p\in\mathcal{P}_{p_K}}\|v-v_p\|_{H^2(K)}\lesssim\exp(-c\,p_K)$, where $\mathcal{P}_{p_K}$ is the space of polynomials of degree at most p_K. The p-th mesh of the sequence has on the order of p elements, and each element is associated to a polynomial degree of order p. Therefore, the number of degrees of freedom $\mathrm{DoF}\simeq p^3$, and thus the error for analytic v is at best $\|v-v_h\|_{H^2(\Omega;\mathcal{T}_h)}\lesssim\exp(-c\,p)\simeq\exp(-c\,\sqrt[3]{\mathrm{DoF}})$.

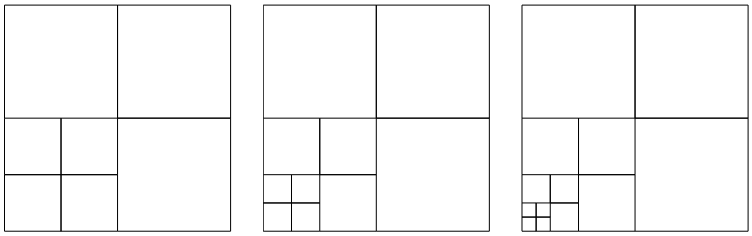

FIGURE 2.4: *Sequence of geometrically graded meshes used for the solution of* (2.65) *on* $\Omega = (0,1)^2$. *The polynomial degrees are chosen to be linearly increasing away from the origin, starting with* $p_K \geq 2$ *on the element closest to the origin. The sequence of meshes is continued by refinement of the element closest to the origin.*

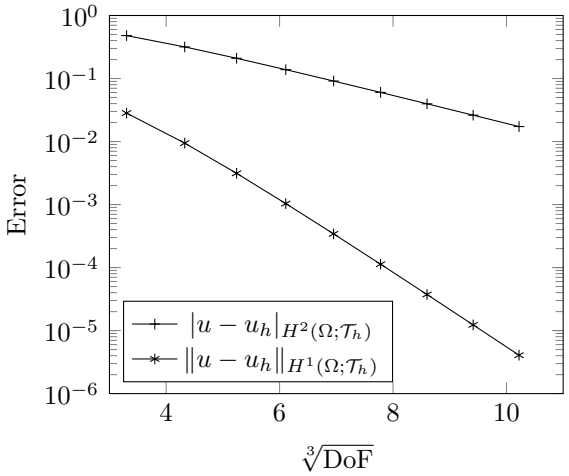

FIGURE 2.5: *Exponential accuracy of the numerical scheme for the problem of section* 2.7.2 *on geometrically graded meshes. The errors in the broken* H^1*-norm and* H^2*-seminorm are plotted against the cube root of the number of degrees of freedom.*

Elements	DoF	$\|u - u_h\|_{L^2(\Omega)}$	$\|u - u_h\|_{H^1(\Omega;\mathcal{T}_h)}$	$\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)}$
4	36	2.349×10^{-3}	2.829×10^{-2}	4.799×10^{-1}
7	81	4.346×10^{-4}	9.439×10^{-3}	3.176×10^{-1}
10	144	8.166×10^{-5}	3.132×10^{-3}	2.096×10^{-1}
13	228	1.491×10^{-5}	1.036×10^{-3}	1.383×10^{-1}
16	336	2.743×10^{-6}	3.426×10^{-4}	9.124×10^{-2}
19	471	4.954×10^{-7}	1.131×10^{-4}	6.020×10^{-2}
22	636	9.840×10^{-8}	3.737×10^{-5}	3.972×10^{-2}
25	834	1.949×10^{-8}	1.233×10^{-5}	2.620×10^{-2}
28	1068	4.799×10^{-9}	4.072×10^{-6}	1.729×10^{-2}

TABLE 2.6: *Errors of the approximations to the solution of problem* (2.65) *on geometrically graded meshes. Exponential convergence rates are observed, with faster rates in lower-order norms.*

Chapter 3

Elliptic Hamilton–Jacobi–Bellman equations

We turn to the study of elliptic Hamilton–Jacobi–Bellman (HJB) equations of the form

$$\sup_{\alpha\in\Lambda}\left[L^\alpha u - f^\alpha\right] = 0 \quad \text{in } \Omega, \tag{3.1}$$

where Ω is a convex domain in $\mathbb{R}^d$, $d \geq 2$, Λ is a compact metric space, and the differential operators L^α, $\alpha \in \Lambda$, are defined by

$$L^\alpha v \coloneqq a^\alpha : D^2 v + b^\alpha \cdot \nabla v - c^\alpha\, v. \tag{3.2}$$

This chapter is largely based on our paper [71], which considers uniformly elliptic HJB equations with lower-order terms; therefore, the Cordes condition presented in earlier chapters is generalised to allow for the lower-order terms. Importantly, we show in section 3.1 that the specific structure of the HJB operator as well as the Cordes condition lead to a straightforward proof of well-posedness that employs the theory of strongly monotone operators. Although these tools of nonlinear functional analysis are commonly used for the study of semilinear or quasilinear equations, to our knowledge this represents their first application to a fully nonlinear PDE.

Following the analysis of the continuous problem, we propose the numerical scheme in section 3.2 and show its consistency in section 3.3. Building on the discretisation techniques developed in Chapter 2, stability of the numerical method is achieved through a discrete strong monotonicity property that mirrors that of the continuous case. This implies the well-posedness of the discrete problem, as shown in section 3.4. Therefore, our method overcomes the challenges typically encountered by nonmonotone methods[1] that were described in Chapter 1. The combination of strong monotonicity and consistency of the method leads to error bounds that are optimal in h, and suboptimal in p by only half an order.

[1] See Remark 1.1 concerning the difference between monotone methods and strongly monotone operators.

A further contribution, given in section 3.6, is the proof of semismoothness in function spaces of the fully nonlinear operator in (3.1) for a general compact metric space Λ. Similar results appear to have only been known in more restricted cases such as the finite-dimensional setting [14]. The semismoothness of the HJB operator implies that the discrete nonlinear problem can be solved with a superlinearly convergent semismooth Newton method, which is in fact the natural adaptation to the current context of the classical algorithm described in section 1.5.

The numerical experiments of section 3.7 test the method on problems with nonsmooth solutions and strongly anisotropic diffusion coefficients, and demonstrate the gains in accuracy and computational efficiency over existing methods. In particular, the exponential convergence rates obtained by *hp*-refinement remain robust in the nonlinear setting, thereby constituting the first examples of exponential convergence rates for fully nonlinear PDE. We also test the robustness and efficiency of the semismooth Newton method, where we show that it leads to the fast and accurate solution of the nonlinear discrete problems.

3.1 Analysis of the continuous problem

Let Ω be a bounded convex polytopal open set in $\mathbb{R}^d$, $d \geq 2$, and let Λ be a compact metric space. It will always be assumed that Ω and Λ are nonempty. Let the symmetric $\mathbb{R}^{d\times d}$-valued function a, the $\mathbb{R}^d$-valued function b, and scalar-valued functions c and f be continuous on $\overline{\Omega} \times \Lambda$. For each $\alpha \in \Lambda$, we consider the function $a^\alpha \colon x \mapsto a(x,\alpha)$, $x \in \overline{\Omega}$, i.e. the dependence on α is denoted through a superscript. The bounded linear operators $L^\alpha \colon H^2(\Omega) \to L^2(\Omega)$ are defined by (3.2). Compactness of Λ and continuity of the coefficients a, b, c and f imply that the fully nonlinear operator F, defined by

$$F \colon v \mapsto F[v] \coloneqq \sup_{\alpha\in\Lambda} \left[L^\alpha v - f^\alpha\right], \tag{3.3}$$

is well-defined as a mapping from $H^2(\Omega)$ to $L^2(\Omega)$. The problem considered is to find a function $u \in H^2(\Omega) \cap H^1_0(\Omega)$ that is a strong solution of the HJB equation subject to a homogeneous Dirichlet boundary condition

$$\begin{aligned} F[u] &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{3.4}$$

Well-posedness of (3.4) is established in section 3.1 under the following hypotheses. It is assumed that there are positive constants $0 < \nu \leq \overline{\nu}$ such that

$$\nu\,|\xi|^2 \leq \xi^\top a^\alpha(x)\,\xi \leq \overline{\nu}\,|\xi|^2 \quad \forall\, \xi \in \mathbb{R}^d,\ \forall\, x \in \Omega,\ \forall\, \alpha \in \Lambda. \tag{3.5}$$

The function c is assumed to be nonnegative on $\overline{\Omega} \times \Lambda$.

The Cordes condition. For problems with lower-order terms, we assume the Cordes condition: there exist $\lambda > 0$ and $\varepsilon \in (0, 1]$ such that, for each $\alpha \in \Lambda$,

$$\frac{|a^\alpha|^2 + |b^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2}{(\operatorname{Tr} a^\alpha + c^\alpha/\lambda)^2} \leq \frac{1}{d+\varepsilon} \quad \text{in } \overline{\Omega}, \tag{3.6}$$

where $|\cdot|$ represents the Euclidean norm for vectors and the Frobenius norm for matrices. In the special case where $b^\alpha \equiv 0$ and $c^\alpha \equiv 0$ for each $\alpha \in \Lambda$, instead of assuming (3.6), we assume that there exists $\varepsilon \in (0, 1]$ such that, for each $\alpha \in \Lambda$,

$$\frac{|a^\alpha|^2}{(\operatorname{Tr} a^\alpha)^2} \leq \frac{1}{d-1+\varepsilon} \quad \text{in } \overline{\Omega}. \tag{3.7}$$

To cover this special case in the following analysis, we set $\lambda = 0$ in all expressions appearing below if $b^\alpha \equiv 0$ and $c^\alpha \equiv 0$ for all $\alpha \in \Lambda$.

Conditions (3.6) and (3.7) are related through the observation that the term c^α/λ may be viewed as the $(d+1, d+1)$ entry of a $(d+1)\times(d+1)$ matrix with principal $d\times d$ sub-matrix a^α, which explains the difference in the right hand sides appearing in (3.6) and (3.7). The parameter λ serves to make the Cordes condition invariant under rescaling the coordinates. It will be seen below that it is often easy to choose an appropriate value for λ.

Example 3.1. We show how the Cordes condition (3.6) arises in practice in an example based on stochastic control problems [36]. We consider a situation where the controls permit the choice of orientation and angle between two independent Wiener diffusions. Let $\Omega \subset \mathbb{R}^2$ and let $\Lambda = [0, \pi/3]\times \mathrm{SO}(2)$, where $\mathrm{SO}(2)$ is the set of 2×2 rotation matrices. The diffusions act along the directions σ_1^α and σ_2^α, where

$$\sigma^\alpha \coloneqq (\sigma_1^\alpha\ \sigma_2^\alpha) \coloneqq R^\top \begin{pmatrix} 1 & \sin\theta \\ 0 & \cos\theta \end{pmatrix}, \quad \alpha = (\theta, R) \in \Lambda. \tag{3.8}$$

In stochastic control problems, we have $a^\alpha \coloneqq \sigma^\alpha(\sigma^\alpha)^\top/2$ and usually $c^\alpha \equiv c > 0$ is a fixed constant [36]. Then, $\operatorname{Tr} a^\alpha = 1$ and $|a^\alpha|^2 = (1 + \sin^2\theta)/2 \leq 7/8$; so (3.7) holds with $\varepsilon = 1/7$. Momentarily assuming that $b^\alpha \equiv 0$, by choosing the value $\lambda = \frac{8}{7}c$ that minimises the left-hand side in (3.6), we find that (3.6) also holds with $\varepsilon = 1/7$. For non-zero b^α, the Cordes condition holds for $\varepsilon < 1/7$ whenever $|b^\alpha|^2/c$ is sufficiently small, which amounts to a standard coercivity assumption.

Example 3.1 is considered further in the numerical experiments of section 3.7.1. Observe that for any choice of Cartesian coordinates on $\mathbb{R}^2$, for $\theta = \pi/3$ there is an $R \in \mathrm{SO}(2)$ such that a^α is not diagonally dominant. In this case, the classical monotone Kushner–Dupuis FDM is therefore not applicable to the resulting HJB equation [16].

Define the strictly positive function $\gamma\colon \overline{\Omega} \times \Lambda \to \mathbb{R}_{>0}$ by

$$\gamma(x,\alpha) \coloneqq \frac{\operatorname{Tr} a^\alpha(x) + c^\alpha(x)/\lambda}{|a^\alpha(x)|^2 + |b^\alpha(x)|^2/2\lambda + (c^\alpha(x)/\lambda)^2}. \tag{3.9}$$

In the special case $b^\alpha \equiv 0$ and $c^\alpha \equiv 0$ for all $\alpha \in \Lambda$, we take $\lambda = 0$ and define

$$\gamma(x,\alpha) \coloneqq \frac{\operatorname{Tr} a^\alpha(x)}{|a^\alpha(x)|^2}. \tag{3.10}$$

As above, for each $\alpha \in \Lambda$, we define $\gamma^\alpha\colon x \mapsto \gamma(x,\alpha)$, $x \in \overline{\Omega}$. It follows from the continuity assumptions on the coefficients and from the uniform ellipticity condition (3.5) that γ is continuous over $\overline{\Omega} \times \Lambda$. Furthermore, nonnegativity of c, continuity of the coefficients, and (3.5) imply that there is a positive constant $\gamma_0 > 0$ such that $\gamma \geq \gamma_0$ on $\overline{\Omega} \times \Lambda$.

Define the operator $F_\gamma\colon H^2(\Omega) \to L^2(\Omega)$ by

$$F_\gamma[v] \coloneqq \sup_{\alpha\in\Lambda} \left[\gamma^\alpha \left(L^\alpha v - f^\alpha\right)\right]. \tag{3.11}$$

It will be seen in the proof of Theorem 3.4 below that the HJB equation (3.4) is in fact equivalent to the problem $F_\gamma[u] = 0$ in Ω, $u = 0$ on $\partial\Omega$. This implies that it is possible to renormalise the HJB operator within the nonlinearity.

For λ as in (3.6), let the operator L_λ be defined by

$$L_\lambda v \coloneqq \Delta v - \lambda v, \quad v \in H^2(\Omega). \tag{3.12}$$

The following inequality generalises Lemma 2.7 [71].

Lemma 3.1. *Let Ω be a bounded open subset of $\mathbb{R}^d$ and suppose that* (3.5) *holds, and suppose that either* (3.6) *holds with $\lambda > 0$, or that* (3.7) *holds with $b^\alpha \equiv 0$, $c^\alpha \equiv 0$ for all α, and $\lambda = 0$. Then, for any open set $U \subset \Omega$ and u, $v \in H^2(U)$, $w \coloneqq u - v$, the following inequality holds a.e. in U:*

$$|F_\gamma[u] - F_\gamma[v] - L_\lambda(u-v)| \leq \sqrt{1-\varepsilon}\,\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2}. \tag{3.13}$$

Proof. It will be clear how to adapt the following arguments to treat the simpler situation where $b^\alpha \equiv 0$, $c^\alpha \equiv 0$ and $\lambda = 0$. So, we consider the case where (3.6) holds with $\lambda > 0$. First, set $w \coloneqq u - v$. We have the identity

$$F_\gamma[u] - L_\lambda u = \sup_{\alpha\in\Lambda} \left[\gamma^\alpha L^\alpha u - L_\lambda u - \gamma^\alpha f^\alpha\right].$$

Note that $|\sup_\alpha x^\alpha - \sup_\alpha y^\alpha| \leq \sup_\alpha |x^\alpha - y^\alpha|$ for any bounded sets $\{x^\alpha\}_\alpha$, $\{y^\alpha\}_\alpha \subset \mathbb{R}$.

Therefore,

$$\begin{aligned}|F_\gamma[u] - F_\gamma[v] - L_\lambda w| &\le \sup_{\alpha\in\Lambda} |\gamma^\alpha L^\alpha w - L_\lambda w| \\ &\le \sup_{\alpha\in\Lambda} |\gamma^\alpha a^\alpha - \mathrm{I}_d||D^2 w| + |\gamma^\alpha||b^\alpha||\nabla w| + |\lambda - c^\alpha\gamma^\alpha||w|,\end{aligned}$$

where I_d is the $d \times d$ identity matrix. The Cauchy–Schwarz inequality implies that

$$|F_\gamma[u] - F_\gamma[v] - L_\lambda w| \le \sup_{\alpha\in\Lambda} \left[\sqrt{C^\alpha}\right] \sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2 |w|^2},$$

where, for each $\alpha \in \Lambda$,

$$C^\alpha \coloneqq |\gamma^\alpha a^\alpha - \mathrm{I}_d|^2 + \frac{|\gamma^\alpha|^2\,|b^\alpha|^2}{2\lambda} + \frac{|\lambda - c^\alpha\gamma^\alpha|^2}{\lambda^2}. \tag{3.14}$$

Expanding the square terms in (3.14) gives

$$C^\alpha = d + 1 - 2\gamma^\alpha \left(\operatorname{Tr} a^\alpha + \frac{c^\alpha}{\lambda}\right) + |\gamma^\alpha|^2 \left(|a^\alpha|^2 + \frac{|b^\alpha|^2}{2\lambda} + \frac{|c^\alpha|^2}{\lambda^2}\right).$$

The definition of γ in (3.9) and the Cordes condition (3.6) imply that $C^\alpha \le 1 - \varepsilon$ on U for every $\alpha \in \Lambda$, thus completing the proof of (3.13). □

Miranda–Talenti inequality. For $\lambda \ge 0$ as above, define the semi-norm $|\cdot|_{H^2(\Omega),\lambda}$ by

$$|u|^2_{H^2(\Omega),\lambda} \coloneqq |u|^2_{H^2(\Omega)} + 2\lambda|u|^2_{H^1(\Omega)} + \lambda^2\|u\|^2_{L^2(\Omega)}. \tag{3.15}$$

If $\lambda > 0$, then this defines a norm on $H^2(\Omega)$. The following result follows from the Miranda–Talenti inequality given in Theorem 2.2. Recall that $L_\lambda u = \Delta u - \lambda u$.

Theorem 3.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain. Then, for any $\lambda \ge 0$ and any $u \in H^2(\Omega) \cap H^1_0(\Omega)$, the following inequalities hold:*

$$|u|_{H^2(\Omega),\lambda} \le \|L_\lambda u\|_{L^2(\Omega)}, \tag{3.16a}$$

$$\|u\|_{H^2(\Omega)} \le C\|L_\lambda u\|_{L^2(\Omega)}, \tag{3.16b}$$

where C is a positive constant depending only on d and $\operatorname{diam}\Omega$.

Proof. The identity $\int_\Omega u\,\Delta u\,\mathrm{d}x = -\int_\Omega |\nabla u|^2 \mathrm{d}x$, based on integration by parts, gives

$$\|L_\lambda u\|^2_{L^2(\Omega)} = \int_\Omega (\Delta u - \lambda u)^2\,\mathrm{d}x = \|\Delta u\|^2_{L^2(\Omega)} + 2\lambda|u|^2_{H^1(\Omega)} + \lambda^2\|u\|^2_{L^2(\Omega)}. \tag{3.17}$$

The Miranda–Talenti inequality $|u|_{H^2(\Omega)} \le \|\Delta u\|_{L^2(\Omega)}$ thus implies (3.16a). The bound (3.16b) follows from (3.17) and (2.8b). □

Well-posedness. For a Banach space X, we say that an operator $\mathcal{A}\colon X \to X^*$ is *bounded* if $\mathcal{A}$ maps bounded sets in X to bounded sets in X^*. We say that $\mathcal{A}$ is *hemicontinuous* if the mapping $[0,1] \ni t \mapsto \langle \mathcal{A}(t\,u + (1-t)\,v), w\rangle$ is continuous for any u, v and $w \in X$, where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between X^* and X. Also, we say that $\mathcal{A}$ is *strongly monotone* if there exists a positive constant $c > 0$ such that

$$\|u - v\|_X^2 \le c\,\langle \mathcal{A}(u) - \mathcal{A}(v), u - v\rangle \quad \forall\, u, v \in X. \tag{3.18}$$

Theorem 3.3 (Browder–Minty)**.** *Let X be a separable reflexive Banach space and let $\mathcal{A}\colon X \to X^*$ be a bounded, hemicontinuous, and strongly monotone operator. Then, for each $\ell \in X^*$, there exists a unique $u \in X$ such that $\mathcal{A}(u) = \ell$.*

For a proof of Theorem 3.3, see [65, Ch. 10], where the strong monotonicity assumption is somewhat relaxed. In fact, the boundedness assumption can also be removed, as shown in [22, Section 9.14]. The Browder–Minty theorem can be seen as a natural nonlinear generalisation of the Lax–Milgram theorem, with the coercivity assumption of the Lax–Milgram theorem being replaced by the strong monotonicity condition (3.18). The operators encountered in this work are found to be Lipschitz continuous, which is sufficient to guarantee their boundedness and their hemicontinuity.

Theorem 3.4. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, and let Λ be a compact metric space. Let the data a, b, c, f be continuous on $\overline{\Omega} \times \Lambda$ and satisfy* (3.5) *and either* (3.6) *with $\lambda > 0$ or* (3.7) *with $c \equiv 0$, $b \equiv 0$ and $\lambda = 0$. Then, there exists a unique strong solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$ of the HJB equation* (3.4)*. Moreover, u is also the unique solution of $F_\gamma[u] = 0$ in Ω, $u = 0$ on $\partial\Omega$.*

Proof. First, set $H \coloneqq H^2(\Omega) \cap H_0^1(\Omega)$; then H is a separable Hilbert space. The proof consists of showing solvability of the equation $F_\gamma[u] = 0$ in H by the method of Browder and Minty, and establishing its equivalence with the HJB equation (3.4). Let the operator $\mathcal{A}\colon H \to H^*$ be defined by

$$\langle \mathcal{A}(u), v\rangle \coloneqq \int_\Omega F_\gamma[u]\, L_\lambda v \,\mathrm{d}x, \quad u,\, v \in H. \tag{3.19}$$

We claim that $\mathcal{A}$ is Lipschitz continuous and strongly monotone. Indeed, let u, $v \in H$ and set $w \coloneqq u - v$. Then, by adding and subtracting $\|L_\lambda w\|_{L^2(\Omega)}^2$, we get

$$\langle \mathcal{A}(u) - \mathcal{A}(v), u - v\rangle = \|L_\lambda w\|_{L^2(\Omega)}^2 + \int_\Omega \left(F_\gamma[u] - F_\gamma[v] - L_\lambda w\right) L_\lambda w \,\mathrm{d}x. \tag{3.20}$$

Lemma 3.1 and the Cauchy–Schwarz inequality show that

$$\langle \mathcal{A}(u) - \mathcal{A}(v), u - v\rangle \ge \|L_\lambda w\|_{L^2(\Omega)}^2 - \sqrt{1-\varepsilon}\,|w|_{H^2(\Omega),\lambda}\|L_\lambda w\|_{L^2(\Omega)}. \tag{3.21}$$

We then use (3.16a) to obtain $\langle \mathcal{A}(u) - \mathcal{A}(v), u - v\rangle \geq (1 - \sqrt{1-\varepsilon})\, \|L_\lambda w\|^2_{L^2(\Omega)}$, so

$$\|u - v\|^2_{H^2(\Omega)} \lesssim \langle \mathcal{A}(u) - \mathcal{A}(v), u - v\rangle \tag{3.22}$$

as a result of (3.16b), thus showing that $\mathcal{A}$ is strongly monotone.

Compactness of Λ and continuity of the data imply that $\mathcal{A}$ is Lipschitz continuous: to see this, let u, v, $z \in H$. Then, we find that

$$|\langle \mathcal{A}(u) - \mathcal{A}(v), z\rangle| \leq \|F_\gamma[u] - F_\gamma[v]\|_{L^2(\Omega)} \|L_\lambda z\|_{L^2(\Omega)} \lesssim \|u - v\|_{H^2(\Omega)} \|z\|_{H^2(\Omega)},$$

where the constant depends only on λ and on the supremum norms of a, b, c and γ on $\overline{\Omega} \times \Lambda$. Lipschitz continuity and strong monotonicity imply that $\mathcal{A}$ is bounded, continuous, coercive and strongly monotone, so the Browder–Minty theorem shows that there exists a unique function $u \in H$ such that $\mathcal{A}(u) = 0$.

For every $g \in L^2(\Omega)$, there is a $v \in H$ such that $L_\lambda v = g$. Therefore $\mathcal{A}(u) = 0$ implies $\int_\Omega F_\gamma[u]\, g \,\mathrm{d}x = 0$ for all $g \in L^2(\Omega)$, thus showing that $F_\gamma[u] = 0$ a.e. in Ω. We claim that $F_\gamma[u] = 0$ if and only if u solves (3.4). Since γ^α is positive, $\gamma^\alpha(L^\alpha u - f^\alpha) \leq 0$ for all $\alpha \in \Lambda$ is equivalent to $L^\alpha u - f^\alpha \leq 0$ for all $\alpha \in \Lambda$; i.e $F[u] \leq 0$ if and only if $F_\gamma[u] \leq 0$. Compactness of Λ and continuity of a, b, c, f and γ imply that at a.e. point of Ω, the suprema in the definitions of $F[u]$ and $F_\gamma[u]$ are attained by an element of Λ, thereby giving $F[u] \geq 0$ if and only if $F_\gamma[u] \geq 0$. Therefore, existence and uniqueness of the solution u of $F_\gamma[u] = 0$ in Ω is equivalent to existence and uniqueness of a solution of (3.4). □

3.2 Numerical scheme

In the following, we employ the definitions and the notation defined in section 2.2. In order to define the numerical scheme, we first extend the definition of $B_{h,*}$ from (2.33). This extension enables us to treat problems with lower-order terms. For $\lambda \geq 0$ as in (3.6), let

$$\begin{aligned}
B_{h,*}(u_h, v_h) \coloneqq & \sum_{K \in \mathcal{T}_h} \left[\langle D^2 u_h, D^2 v_h\rangle_K + 2\lambda \langle \nabla u_h, \nabla v_h\rangle_K + \lambda^2 \langle u_h, v_h\rangle_K \right] \\
& + \sum_{F \in \mathcal{F}_h^i} \left[\langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} \{u_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F + \langle \operatorname{div}_{\mathrm{T}} \nabla_{\mathrm{T}} \{v_h\}, [\![\nabla u_h \cdot n_F]\!]\rangle_F \right] \\
& - \sum_{F \in \mathcal{F}_h^{i,b}} \left[\langle \nabla_{\mathrm{T}} \{\nabla u_h \cdot n_F\}, [\![\nabla_{\mathrm{T}} v_h]\!]\rangle_F + \langle \nabla_{\mathrm{T}} \{\nabla v_h \cdot n_F\}, [\![\nabla_{\mathrm{T}} u_h]\!]\rangle_F \right] \\
& - \lambda \sum_{F \in \mathcal{F}_h^{i,b}} \left[\langle \{\nabla u_h \cdot n_F\}, [\![v_h]\!]\rangle_F + \langle \{\nabla v_h \cdot n_F\}, [\![u_h]\!]\rangle_F \right] \\
& - \lambda \sum_{F \in \mathcal{F}_h^{i}} \left[\langle \{u_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F + \langle \{v_h\}, [\![\nabla u_h \cdot n_F]\!]\rangle_F \right].
\end{aligned} \tag{3.23}$$

We recall that the term J_h is defined in (2.23), and that the bilinear forms $B_{h,\theta}$, $\theta \in [0,1]$, are defined in (2.34). The nonlinear form $A_h \colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ is defined by

$$A_h(u_h; v_h) \coloneqq \sum_{K \in \mathcal{T}_h} \langle F_\gamma[u_h], L_\lambda v_h \rangle_K + B_{h,1/2}(u_h, v_h) - \sum_{K \in \mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h \rangle_K. \tag{3.24}$$

The form A_h is linear in its second argument but nonlinear in its first argument. The numerical method for approximating the solution of (3.4) is to find $u_h \in V_{h,\mathbf{p}}$ such that

$$A_h(u_h; v_h) = 0 \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{3.25}$$

The choice of nonlinear form in (3.24) is made to mirror the addition–substraction step of (3.20) in the proof of Theorem 3.4.

3.3 Consistency

As explained in section 2.3, the numerical scheme weakly enforces discrete analogues of the identities behind the Miranda–Talenti inequality. The next result extends Lemma 2.6 to include lower-order terms [71].

Lemma 3.5. *Let Ω be a bounded Lipschitz polytopal domain and let $\mathcal{T}_h$ be a simplicial or parallelepipedal mesh on Ω. Let $w \in H^s(\Omega; \mathcal{T}_h) \cap H^2(\Omega) \cap H^1_0(\Omega)$, $s > 5/2$. Then, for every $v_h \in V_{h,\mathbf{p}}$, we have the identities*

$$B_{h,*}(w, v_h) = \sum_{K \in \mathcal{T}_h} \langle L_\lambda w, L_\lambda v_h \rangle_K \quad \textit{and} \quad J_h(w, v_h) = 0. \tag{3.26}$$

Proof. The second part of (3.26) is obvious. We also note that all terms in $B_{h,*}(w, v_h)$ that involve jumps of w or of its first derivates vanish. For the case $\lambda = 0$, the stated result reduces to Lemma 2.6, which treats the consistency of the second order terms. So, for $\lambda > 0$, the identities of (3.26) are deduced from the previous result and from the identities

$$-\lambda \sum_{K \in \mathcal{T}_h} \langle \Delta w, v_h \rangle_K = \lambda \sum_{K \in \mathcal{T}_h} \langle \nabla w, \nabla v_h \rangle_K - \lambda \sum_{F \in \mathcal{F}_h^{i,b}} \langle \{\nabla w \cdot n_F\}, [\![v_h]\!] \rangle_F, \tag{3.27}$$

$$-\lambda \sum_{K \in \mathcal{T}_h} \langle w, \Delta v_h \rangle_K = \lambda \sum_{K \in \mathcal{T}_h} \langle \nabla w, \nabla v_h \rangle_K - \lambda \sum_{F \in \mathcal{F}_h^{i}} \langle \{w\}, [\![\nabla v_h \cdot n_F]\!] \rangle_F, \tag{3.28}$$

for all $v_h \in V_{h,\mathbf{p}}$, where we have used the fact that $w|_F = 0$ for all $F \in \mathcal{F}_h^b$ in (3.28). □

If the function w satisfies the hypotheses of Lemma 3.5, then (3.26) implies that

$$B_{h,\theta}(w, v_h) = \sum_{K \in \mathcal{T}_h} \langle L_\lambda w, L_\lambda v_h \rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}},\ \forall\, \theta \in [0,1]. \tag{3.29}$$

The following consistency result for the scheme follows immediately from Theorem 3.4, from (3.29), and from the definition of A_h in (3.24).

Corollary 3.6. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex polytopal domain, and let $\mathcal{T}_h$ be a simplicial or parallelepipedal mesh. Let $u \in H^2(\Omega) \cap H^1_0(\Omega)$ be the unique solution of (3.4). If the solution $u \in H^s(\Omega; \mathcal{T}_h)$, $s > 5/2$, then we have $A_h(u; v_h) = 0$ for every $v_h \in V_{h,\mathbf{p}}$.*

We remark that the comments from sections 2.4 and 2.6, concerning the regularity assumptions on the solution, carry over to the current setting. We also refer the reader to the numerical experiment of section 3.7.3 for an example showing convergence of the method when this assumption is relaxed.

3.4 Stability

We extend the mesh-dependent norms $\|\cdot\|_{h,\theta}$ defined by (2.24) as follows. For $\lambda \geq 0$ as above, define the seminorms $|\cdot|_{H^2(K),\lambda}$, $K \in \mathcal{T}_h$, and $|\cdot|_{H^2(\Omega;\mathcal{T}_h),\lambda}$ on $H^2(\Omega; \mathcal{T}_h)$ by

$$|v|^2_{H^2(K),\lambda} \coloneqq \|D^2 v\|^2_{L^2(K)} + 2\lambda \|\nabla v\|^2_{L^2(K)} + \lambda^2 \|v\|^2_{L^2(K)}, \tag{3.30}$$

$$|v|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} \coloneqq \sum_{K \in \mathcal{T}_h} |v|^2_{H^2(K),\lambda}. \tag{3.31}$$

For each $\theta \in [0,1]$, define the functional $\|\cdot\|_{h,\theta} \colon V_{h,\mathbf{p}} \to \mathbb{R}_{\geq 0}$ by

$$\|v_h\|^2_{h,\theta} \coloneqq \sum_{K \in \mathcal{T}_h} \Big[\theta |v_h|^2_{H^2(K),\lambda} + (1-\theta)\|L_\lambda v_h\|^2_{L^2(K)}\Big] + |v_h|^2_{\mathrm{J}}. \tag{3.32}$$

For each $\theta \in [0,1]$, the functional $\|\cdot\|_{h,\theta}$ is a norm on $V_{h,\mathbf{p}}$. Indeed, homogeneity and the triangle inequality are clear. If $\|v_h\|_{h,\theta} = 0$, then $v_h \in H^2(\Omega) \cap H^1_0(\Omega)$ since $[\![\nabla v_h]\!] = 0$ for all $F \in \mathcal{F}_h^i$, and $[\![v_h]\!] = 0$ for all $F \in \mathcal{F}_h^{i,b}$. Moreover, $L_\lambda v_h \equiv 0$ (if $\theta = 1$, use $|v_h|_{H^2(K),\lambda} = 0$ $\forall\, K$), so $v_h \equiv 0$ as a result of (3.16b).

Lemma 3.7. *Let Ω be a bounded convex polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying (2.11). Then, for each constant $\kappa > 1$, there exists a positive constant c_μ and c_η, independent of h, $\mathbf{p}$ and θ, such that, for any $v_h \in V_{h,\mathbf{p}}$ and any $\theta \in [0,1]$, we have*

$$B_{h,\theta}(v_h, v_h) \geq \frac{\theta}{\kappa} |v_h|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} + (1-\theta) \sum_{K \in \mathcal{T}_h} \|L_\lambda v_h\|^2_{L^2(K)} + \frac{1}{2}|v_h|^2_{\mathrm{J}} \tag{3.33}$$

whenever

$$\mu_F = c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F} \quad \text{and} \quad \eta_F > \lambda\, c_\eta \frac{\tilde{p}_F^2}{\tilde{h}_F} \quad \forall\, F \in \mathcal{F}_h^{i,b}. \tag{3.34}$$

The strict inequality in the second part of (3.34) serves to cover the case $\lambda = 0$. See also (3.41) below for an upper bound restriction on η_F.

Proof. For $v_h \in V_{h,\mathbf{p}}$, we have

$$B_{h,\theta}(v_h, v_h) = \theta\, |v_h|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} + (1-\theta) \sum_{K\in\mathcal{T}_h} \|L_\lambda v_h\|^2_{L^2(K)} + |v_h|^2_{\mathrm{J}} + \theta \sum_{i=1}^{4} I_i,$$

where

$$\begin{aligned} I_1 &\coloneqq 2 \sum_{F\in\mathcal{F}_h^i} \langle \mathrm{div}_{\mathrm{T}}\, \nabla_{\mathrm{T}} \{v_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F, & I_3 &\coloneqq -2\lambda \sum_{F\in\mathcal{F}_h^i} \langle \{v_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F, \\ I_2 &\coloneqq -2 \sum_{F\in\mathcal{F}_h^{i,b}} \langle \nabla_{\mathrm{T}} \{\nabla v_h \cdot n_F\}, [\![\nabla_{\mathrm{T}}\, v_h]\!]\rangle_F, & I_4 &\coloneqq -2\lambda \sum_{F\in\mathcal{F}_h^{i,b}} \langle \{\nabla v_h \cdot n_F\}, [\![v_h]\!]\rangle_F. \end{aligned}$$

Lemma 2.8 shows that there is a constant C_d depending only on d, such that for any $\delta > 0$,

$$|I_1| \le \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}, \tag{3.35}$$

$$|I_2| \le \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \|D^2 v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^{i,b}} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla_{\mathrm{T}}\, v_h]\!]\|^2_{L^2(F)}, \tag{3.36}$$

where C_{Tr} is the combined constant of the trace and inverse inequalities, and $c_{\mathcal{F}}$ is given by (2.11). The inverse and trace inequalities also show that

$$\begin{aligned} |I_3| &\le 2\lambda \sqrt{\sum_{F\in\mathcal{F}_h^i} \frac{\delta \tilde{h}_F}{\tilde{p}_F^2} \|\{v_h\}\|^2_{L^2(F)}} \sqrt{\sum_{F\in\mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}} \\ &\le \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} \lambda^2 \|v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|[\![\nabla v_h \cdot n_F]\!]\|^2_{L^2(F)}. \end{aligned} \tag{3.37}$$

Similarly, it is found that

$$|I_4| \le \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} \sum_{K\in\mathcal{T}_h} 2\lambda \|\nabla v_h\|^2_{L^2(K)} + \sum_{F\in\mathcal{F}_h^{i,b}} \frac{\lambda \tilde{p}_F^2}{2\delta \tilde{h}_F} \|[\![v_h]\!]\|^2_{L^2(F)}. \tag{3.38}$$

We may take C_d to be the same constant in each of the above inequalities. So,

$$B_{h,\theta}(v_h, v_h) \geq \theta\,(1 - \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}})\,|v_h|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} + (1-\theta)\sum_{K\in\mathcal{T}_h}\|L_\lambda v_h\|^2_{L^2(K)}$$
$$+ \sum_{F\in\mathcal{F}_h^i}\left(\mu_F - \frac{2\theta\tilde{p}_F^2}{\delta\tilde{h}_F}\right)\|[\![\nabla v_h\cdot n_F]\!]\|^2_{L^2(F)} + \sum_{F\in\mathcal{F}_h^{i,b}}\left(\mu_F - \frac{\theta\tilde{p}_F^2}{\delta\tilde{h}_F}\right)\|[\![\nabla_{\mathrm{T}}\, v_h]\!]\|^2_{L^2(F)}$$
$$+ \sum_{F\in\mathcal{F}_h^{i,b}}\left(\eta_F - \frac{\lambda\theta\tilde{p}_F^2}{2\delta\tilde{h}_F}\right)\|[\![v_h]\!]\|^2_{L^2(F)}.$$

For any given $\kappa > 1$, there is a $\delta > 0$ such that $1 - \delta\, C_d\, C_{\mathrm{Tr}}\, c_{\mathcal{F}} > 1/\kappa$. Set $c_\mu = 4/\delta$ and $c_\eta = 1/\delta$, so that (3.33) holds whenever μ_F and η_F satisfy (3.34). □

Theorem 3.8. *Let Ω be a bounded convex polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11)*. Let Λ be a compact metric space and let the data satisfy* (3.5) *and either* (3.6) *or* (3.7) *with $b \equiv 0$, $c \equiv 0$, $\lambda = 0$. Let η_F and μ_F be chosen so that Lemma* 3.7 *holds with $\kappa < (1-\varepsilon)^{-1}$. Then, for every u_h and $v_h \in V_{h,\mathbf{p}}$, we have the strong monotonicity inequality*

$$\|u_h - v_h\|^2_{h,1} \leq \frac{2\,\kappa}{1-\kappa\,(1-\varepsilon)}\big(A_h(u_h; u_h - v_h) - A_h(v_h; u_h - v_h)\big). \tag{3.39}$$

Moreover, A_h is Lipschitz continuous, i.e. for any u_h, v_h and z_h in $V_{h,\mathbf{p}}$, we have

$$|A_h(u_h; z_h) - A_h(v_h; z_h)| \lesssim \|u_h - v_h\|_{h,1}\|z_h\|_{h,1}. \tag{3.40}$$

Therefore, there exists a unique solution $u_h \in V_{h,\mathbf{p}}$ to the numerical scheme (3.25)*.*

Proof. First, note that since $\varepsilon \in (0,1]$, it is possible to choose the terms μ_F and η_F such that $\kappa < (1-\varepsilon)^{-1}$. Let u_h and v_h belong to $V_{h,\mathbf{p}}$ and set $w_h \coloneqq u_h - v_h$. Then,

$$A_h(u_h; w_h) - A_h(v_h; w_h) = B_{h,1/2}(w_h, w_h) + \sum_{K\in\mathcal{T}_h}\langle F_\gamma[u_h] - F_\gamma[v_h] - L_\lambda w_h, L_\lambda w_h\rangle_K.$$

Note that Lemma 3.1 gives

$$\sum_{K\in\mathcal{T}_h}|\langle F_\gamma[u_h] - F_\gamma[v_h] - L_\lambda w_h, L_\lambda w_h\rangle_K| \leq \sqrt{1-\varepsilon}\sum_{K\in\mathcal{T}_h}|w_h|_{H^2(K),\lambda}\|L_\lambda w_h\|_{L^2(K)}$$
$$\leq \frac{1-\varepsilon}{2}|w_h|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} + \frac{1}{2}\sum_{K\in\mathcal{T}_h}\|L_\lambda w_h\|^2_{L^2(K)}.$$

This inequality and Lemma 3.7 show that

$$A_h(u_h; w_h) - A_h(v_h; w_h) \geq \frac{1-\kappa\,(1-\varepsilon)}{2\,\kappa}|w_h|^2_{H^2(\Omega;\mathcal{T}_h),\lambda} + \frac{1}{2}|w_h|^2_{\mathrm{J}} \geq \|w_h\|^2_{h,1}/C,$$

where $C \coloneqq 2\,\kappa/\left(1-\kappa\,(1-\varepsilon)\right)$. Since $\kappa\,(1-\varepsilon)<1$, we obtain (3.39).

Now, let $z_h \in V_{h,\mathbf{p}}$; using the linearity of $B_{h,\theta}$ and inverse inequalities, we find that there exists a constant C depending on the constants appearing in the proof of Lemma 3.7, but not on h or $\mathbf{p}$, such that $|B_{h,1/2}(u_h - v_h, z_h)| \leq C\|u_h - v_h\|_{h,1}\|z_h\|_{h,1}$. Using Lemma 3.1 and the above inequalities, we deduce that

$$\sum_{K\in\mathcal{T}_h} |\langle F_\gamma[u_h] - F_\gamma[v_h] - L_\lambda(u_h - v_h), L_\lambda z_h\rangle_K| \lesssim \|u_h - v_h\|_{h,1}\|z_h\|_{h,1}.$$

It then follows that A_h is Lipschitz continuous, as stated in (3.40). The Browder–Minty theorem along with (3.39) and (3.40) imply that there exists a unique $u_h \in V_{h,\mathbf{p}}$ such that $A_h(u_h; v_h) = 0$ for all $v_h \in V_{h,\mathbf{p}}$. □

Remark 3.1. In the above stability result, it was required that c_μ and c_η be chosen so that Lemma 3.7 holds for some $\kappa < (1-\varepsilon)^{-1}$. It can be seen from the proof of Lemma 3.7 that c_μ and c_η can be chosen to be of order $1/\varepsilon$, and κ can be chosen so that the constant in (3.39) is also of order $1/\varepsilon$ when ε is small.

3.5 Error analysis

In the following, we require that

$$\eta_F \leq c_\eta \max(1,\lambda)\,\frac{\tilde{p}_F^4}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}. \tag{3.41}$$

Therefore, the combined requirements on the user-defined parameters μ_F and η_F given by (2.28), (3.34) and (3.41) imply

$$\mu_F = c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad c_\eta \max(1,\lambda)\frac{\tilde{p}_F^2}{\tilde{h}_F} < \eta_F \leq c_\eta \max(1,\lambda)\frac{\tilde{p}_F^4}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}, \tag{3.42}$$

with c_μ and c_η user-defined constants to be chosen sufficiently large.

3.5.1 Error bound for solutions with sufficient regularity

The consistency and stability of the method lead to the following error bound [71].

Theorem 3.9. *Let Ω be a bounded convex polytopal domain, and let the shape-regular sequence of simplicial or parallelepipedal meshes $\{\mathcal{T}_h\}_h$ satisfy* (2.11) *and* (2.12)*, with $\mathbf{p}$ satisfying* (2.13) *for each h. Let Λ be a compact metric space, and let the data satisfy* (3.5)*, and either* (3.6) *or* (3.7) *when $b \equiv 0$, $c \equiv 0$ and $\lambda = 0$. Let $u \in H^2(\Omega)\cap H_0^1(\Omega)$ be the unique solution of* (3.4)*, and assume that $u \in H^{\mathbf{s}}(\Omega;\mathcal{T}_h)$, with $s_K > 5/2$ for each $K \in \mathcal{T}_h$. Let μ_F and η_F be chosen as in Theorem* 3.8*, and let η_F also satisfy* (3.41)*. Then, there*

exists a positive constant C, independent of h, $\mathbf{p}$ and u, but depending on $\max_K s_K$, such that

$$\|u - u_h\|_{h,1}^2 \leq C \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|_{H^{s_K}(K)}^2, \tag{3.43}$$

where $t_K = \min(s_K, p_K + 1)$ for each $K \in \mathcal{T}_h$.

Note that for the special case of quasi-uniform meshes and uniform polynomial degrees, if $u \in H^s(\Omega)$ with $s > 5/2$, the error bound (3.43) simplifies to

$$\|u - u_h\|_{h,1} \lesssim \frac{h^{\min(s,\, p+1)-2}}{p^{s-5/2}} \|u\|_{H^s(\Omega)}.$$

Therefore, the convergence rates are optimal with respect to the mesh size and suboptimal in the polynomial degree by half an order.

Proof. Theorem C.6 implies that there exists a $z_h \in V_{h,\mathbf{p}}$, and a constant C, independent of u, h_K and p_K, but dependent on $\max_K s_K$, such that, for each $K \in \mathcal{T}_h$, each nonnegative integer $j \leq s_K$, and for each multi-index β with $|\beta| < s_K - 1/2$, we have

$$\|u - z_h\|_{H^j(K)} \leq C \frac{h_K^{t_K-j}}{p_K^{s_K-j}} \|u\|_{H^{s_K}(K)}, \tag{3.44}$$

$$\|D^\beta (u - z_h)\|_{L^2(\partial K)} \leq C \frac{h_K^{t_K-|\beta|-1/2}}{p_K^{s_K-|\beta|-1/2}} \|u\|_{H^{s_K}(K)}. \tag{3.45}$$

Set $\psi_h \coloneqq u_h - z_h$ and $\xi_h \coloneqq u - z_h$. By Corollary 3.6, we have $A_h(u; v_h) = 0$ for all $v_h \in V_{h,\mathbf{p}}$. Strong monotonicity of A_h on $V_{h,\mathbf{p}}$, as shown in Theorem 3.8, yields

$$\|\psi_h\|_{h,1}^2 \lesssim A_h(u_h; \psi_h) - A_h(z_h; \psi_h) = A_h(u; \psi_h) - A_h(z_h; \psi_h). \tag{3.46}$$

By applying the Cauchy–Schwarz inequality to the terms appearing on the right-hand side of (3.46) and applying inverse inequalities to $\psi_h \in V_{h,\mathbf{p}}$, we eventually obtain

$$A_h(u; \psi_h) - A_h(z_h; \psi_h) \leq \sqrt{\textstyle\sum_{i=1}^{10} E_i}\, \|\psi_h\|_{h,1}, \tag{3.47}$$

where the quantities E_i are defined by

$$E_1 \coloneqq \sum_{K\in\mathcal{T}_h} |\xi_h|^2_{H^2(K),\lambda}, \qquad E_2 \coloneqq \sum_{K\in\mathcal{T}_h} \|L_\lambda \xi_h\|^2_{L^2(K)},$$
$$E_3 \coloneqq \sum_{K\in\mathcal{T}_h} \|F_\gamma[u] - F_\gamma[z_h]\|^2_{L^2(K)}, \qquad E_4 \coloneqq \sum_{F\in\mathcal{F}_h^i} \mu_F^{-1}\|\mathrm{div}_{\mathrm{T}}\, \nabla_{\mathrm{T}}\{\xi_h\}\|^2_{L^2(F)},$$
$$E_5 \coloneqq \sum_{F\in\mathcal{F}_h^i} \mu_F\|[\![\nabla\xi_h\cdot n_F]\!]\|^2_{L^2(F)}, \qquad E_6 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F^{-1}\|\nabla_{\mathrm{T}}\{\nabla\xi_h\cdot n_F\}\|^2_{L^2(F)},$$
$$E_7 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F\|[\![\nabla_{\mathrm{T}}\,\xi_h]\!]\|^2_{L^2(F)}, \qquad E_8 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \lambda^2\eta_F^{-1}\|\{\nabla\xi_h\cdot n_F\}\|^2_{L^2(F)},$$
$$E_9 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} (\lambda\mu_F+\eta_F)\|[\![\xi_h]\!]\|^2_{L^2(F)}, \qquad E_{10} \coloneqq \sum_{F\in\mathcal{F}_h^{i}} \lambda^2\mu_F^{-1}\|\{\xi_h\}\|^2_{L^2(F)}.$$

The inequality in (3.44) shows that

$$E_1 + E_2 \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}}\|u\|^2_{H^{s_K}(K)}. \tag{3.48}$$

By compactness of Λ, continuity of the data and (3.5), F_γ is Lipschitz continuous, so

$$E_3 \lesssim \sum_{K\in\mathcal{T}_h} \|\xi_h\|^2_{H^2(K)} \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}}\|u\|^2_{H^{s_K}(K)}. \tag{3.49}$$

We use (2.11), (2.12), (2.13), (3.34) and (3.45) to obtain

$$E_4 + E_6 \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2}\frac{h_K^{2t_K-5}}{p_K^{2s_K-5}}\|u\|^2_{H^{s_K}(K)} = \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}}\|u\|^2_{H^{s_K}(K)}, \tag{3.50}$$

$$E_5 + E_7 \lesssim \sum_{K\in\mathcal{T}_h} \frac{p_K^2}{h_K}\frac{h_K^{2t_K-3}}{p_K^{2s_K-3}}\|u\|^2_{H^{s_K}(K)} = \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}}\|u\|^2_{H^{s_K}(K)}. \tag{3.51}$$

Similarly, we use (3.34) to get

$$E_8 \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2}\frac{h_K^{2t_K-3}}{p_K^{2s_K-3}}\|u\|^2_{H^{s_K}(K)} = \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-2}}{p_K^{2s_K-1}}\|u\|^2_{H^{s_K}(K)}. \tag{3.52}$$

By hypothesis, $\eta_F \lesssim \tilde{p}_F^4/\tilde{h}_F^3$ by (3.41), so (2.12) and (2.13) imply that

$$E_9 \lesssim \sum_{K\in\mathcal{T}_h} \frac{p_K^4}{h_K^3}\frac{h_K^{2t_K-1}}{p_K^{2s_K-1}}\|u\|^2_{H^{s_K}(K)} = \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}}\|u\|^2_{H^{s_K}(K)}. \tag{3.53}$$

Finally, (3.45) yields

$$E_{10} \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2}\frac{h_K^{2t_K-1}}{p_K^{2s_K-1}}\|u\|^2_{H^{s_K}(K)} = \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K}}{p_K^{2s_K+1}}\|u\|^2_{H^{s_K}(K)}. \tag{3.54}$$

The bound (3.43) is then obtained from $\|u-u_h\|_{h,1} \le \|\psi_h\|_{h,1} + \|\xi_h\|_{h,1}$ and the above inequalities. □

3.5.2 Error bound for solutions with minimal regularity

The following result generalises Theorem 2.11, showing a quasi-optimal approximation property of the method with respect to any H^2-conforming approximation on the same mesh with the same polynomial degrees. Therefore, the comments surrounding Theorem 2.11 carry over to the current nonlinear setting.

Theorem 3.10. *Let Ω be a bounded convex polytopal domain, and let the shape-regular sequence of simplicial or parallelepipedal meshes $\{\mathcal{T}_h\}_h$ satisfy* (2.11) *and* (2.12)*, with $\mathbf{p}$ satisfying* (2.13) *for each h. Let Λ be a compact metric space, and let the data satisfy* (3.5)*, and either* (3.6) *or* (3.7) *when $b\equiv 0$, $c\equiv 0$ and $\lambda = 0$. Let $u\in H^2(\Omega)\cap H^1_0(\Omega)$ be the unique solution of* (3.4)*. Let μ_F and η_F be chosen as in Theorem* 3.8*. Then, we have*

$$\|u-u_h\|_{h,1} \lesssim \inf\left\{|u-z_h|_{H^2(\Omega),\lambda} \colon z_h\in V_{h,\mathbf{p}}\cap H^2(\Omega)\cap H^1_0(\Omega)\right\}. \tag{3.55}$$

Proof. Let $z_h\in V_{h,\mathbf{p}}\cap H^2(\Omega)\cap H^1_0(\Omega)$; since z_h is a piecewise polynomial, it follows that $z_h\in H^s(\Omega;\mathcal{T}_h)$ for any $s>5/2$. Therefore, Theorem 3.8, the numerical scheme (3.25), and Lemma 3.5, applied to z_h, show that

$$\|u_h-z_h\|^2_{h,1} \lesssim A_h(u_h;z_h-u_h) - A_h(z_h;z_h-u_h) = -\sum_{K\in\mathcal{T}_h}\langle F_\gamma[z_h], L_\lambda(z_h-u_h)\rangle_K.$$

Since u solves (3.4), we find that

$$\begin{aligned}\|z_h-u_h\|^2_{h,1} &\lesssim \sum_{K\in\mathcal{T}_h}\langle F_\gamma[u]-F_\gamma[z_h], L_\lambda(z_h-u_h)\rangle_K\\ &\lesssim \|F_\gamma[u]-F_\gamma[z_h]\|_{L^2(\Omega)}\,|z_h-u_h|_{H^2(\Omega;\mathcal{T}_h),\lambda},\end{aligned}$$

which implies that $\|z_h-u_h\|_{h,1}\lesssim\|F_\gamma[u]-F_\gamma[z_h]\|_{L^2(\Omega)}$, since $|\psi_h|_{H^2(\Omega;\mathcal{T}_h),\lambda}\le\|\psi_h\|_{h,1}$ for $\psi_h := z_h-u_h\in V_{h,\mathbf{p}}$. Using (3.6), it is found that $\|F_\gamma[u]-F_\gamma[z_h]\|_{L^2(\Omega)}\lesssim|u-z_h|_{H^2(\Omega),\lambda}$. This shows that

$$\|u_h-z_h\|_{h,1}\lesssim|u-z_h|_{H^2(\Omega),\lambda}. \tag{3.56}$$

Since the functions u and z_h belong to $H^2(\Omega)\cap H^1_0(\Omega)$, it follows that their jumps in values and gradients vanish on all interior faces, and that their values and tangential gradients

vanish on all boundary faces. Therefore, we have the identity $\|u - z_h\|_{h,1} = |u - z_h|_{H^2(\Omega),\lambda}$.

So, the triangle inequality and (3.56) imply that

$$\|u - u_h\|_{h,1} \leq |u - z_h|_{H^2(\Omega),\lambda} + \|z_h - u_h\|_{h,1} \lesssim |u - z_h|_{H^2(\Omega),\lambda}. \tag{3.57}$$

Since z_h was arbitrary, taking the infimum over all $z_h \in V_{h,\mathbf{p}} \cap H^2(\Omega) \cap H_0^1(\Omega)$ in (3.57) yields (3.55). □

Remark 3.2. The constants appearing in the error bounds of Theorems 3.9 and 3.10 depend on the constant of the stability bound of Theorem 3.8. Therefore, as explained in Remark 3.1, the behaviour of these constants for problems presenting a small value of ε is typically of order $1/\varepsilon$. It is therefore seen that the parameter ε plays here a similar role as that of the ellipticity constant $\overline{\nu}/\nu$ in the analysis of standard FEM for diffusion equations. The performance of the method for problems with very small values of ε is studied in the numerical experiments of section 3.7.

3.6 Semismooth Newton method

We turn to the analysis of an algorithm for solving the discrete problem (3.25), which can be interpreted as a Newton method for nonsmooth operator equations [64]. After showing that the algorithm is well-posed, we establish its superlinear convergence as a consequence of the semismoothness in function spaces of the HJB operator. The semismoothness of finite-dimensional HJB operators in a different form was studied in [14].

For $1 \leq r \leq \infty$, a function $u \in W^{2,r}(\Omega; \mathcal{T}_h)$ defines a vector-valued function $\mathbf{u} \in L^r(\Omega; \mathbb{R}^m)$ through $\mathbf{u} = (u, \nabla u, D^2 u)$, where ∇u and $D^2 u$ denote the broken gradient and broken Hessian of u, see section 2.2. For a vector $\mathbf{u} = (z, p, M) \in \mathbb{R}^m$, define the function $F_\gamma \colon \Omega \times \mathbb{R}^m \to \mathbb{R}$ by

$$F_\gamma(x, \mathbf{u}) := \sup_{\alpha \in \Lambda} \left[\gamma^\alpha(x) \left(a^\alpha(x) : M + b^\alpha(x) \cdot p - c^\alpha(x)\, z - f^\alpha(x) \right) \right]. \tag{3.58}$$

For each $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, we define $\Lambda(x, \mathbf{u})$ as the set of all $\alpha \in \Lambda$ such that the supremum in (3.58) is attained. This defines a set-valued map $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$.

Lemma 3.11. *Let Ω be a bounded open subset of $\mathbb{R}^d$, let Λ be a compact metric space, let the data a, b, c and f be continuous on $\overline{\Omega} \times \Lambda$, and suppose that (3.5) holds. Then, for each $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, $\Lambda(x, \mathbf{u})$ is a non-empty closed 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 $(y, \mathbf{v}) \in V$.*

We remark that the uniform ellipticity condition (3.5) is only used in Lemma 3.11 to guarantee that $\gamma \in C(\overline{\Omega} \times \Lambda)$.

Proof. For every $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, where $\mathbf{u} = (z, p, M)$, compactness of Λ and continuity of a, b, c, f and γ imply the existence of a maximiser in (3.3); so $\Lambda(x, \mathbf{u})$ is non-empty. The set $\Lambda(x, \mathbf{u})$ is closed: if α is in the closure of $\Lambda(x, \mathbf{u})$, say $\alpha_j \to \alpha$, with $\alpha_j \in \Lambda(x, \mathbf{u})$ for each $j \in \mathbb{N}$, then continuity of the data implies that

$$\gamma^{\alpha}\left(a^{\alpha} \colon M + b^{\alpha} \cdot p - c^{\alpha} z - f^{\alpha}\right)|_x = \lim_{j \to \infty} \gamma^{\alpha_j}\left(a^{\alpha_j} \colon M + b^{\alpha_j} \cdot p - c^{\alpha_j} z - f^{\alpha_j}\right)|_x. \tag{3.59}$$

Since $\alpha_j \in \Lambda(x, \mathbf{u})$ for each $j \in \mathbb{N}$, the right hand side of (3.59) equals $F(x, \mathbf{u})$, thus giving $\alpha \in \Lambda(x, \mathbf{u})$ and showing that $\Lambda(x, \mathbf{u})$ is closed.

We prove upper semicontinuity of $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$ by contradiction. Suppose that there exists an $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$, a neighbourhood U of $\Lambda(x, \mathbf{u})$, and a sequence $\{(x_j, \mathbf{u}_j)\}_{j=1}^{\infty}$, $\mathbf{u}_j = (z_j, p_j, M_j)$, converging to $(x, \mathbf{u})$, together with $\alpha_j \in \Lambda(x_j, \mathbf{u}_j) \setminus U$ for all $j \in \mathbb{N}$. Because Λ is compact and $\Lambda \setminus U$ is closed, there exists a subsequence, to which we pass without change of notation, such that $\alpha_j \to \alpha \in \Lambda \setminus U$. On the one hand, $\Lambda(x, \mathbf{u})$ is non-empty so there is $\beta \in \Lambda(x, \mathbf{u})$. Then, by definition of F_γ,

$$\gamma^{\alpha}\left(a^{\alpha} \colon M + b^{\alpha} \cdot p - c^{\alpha} z - f^{\alpha}\right)|_x \leq F_\gamma(x, \mathbf{u}). \tag{3.60}$$

On the other hand, $\alpha_j \in \Lambda(x_j, \mathbf{u}_j)$ implies that we have, for each $j \in \mathbb{N}$,

$$\gamma^{\alpha_j}\left(a^{\alpha_j} \colon M_j + b^{\alpha_j} \cdot p_j - c^{\alpha_j} z_j - f^{\alpha_j}\right)|_{x_j} \geq \gamma^{\beta}\left(a^{\beta} \colon M_j + b^{\beta} \cdot p_j - c^{\beta} z_j - f^{\beta}\right)|_{x_j}.$$

Taking the limit $j \to \infty$ in the above inequality shows that equality holds in (3.60) because $\beta \in \Lambda(x, \mathbf{u})$. Hence, $\alpha \in \Lambda(x, \mathbf{u})$; however, U is an open neighbourhood of $\Lambda(x, \mathbf{u})$ and $\alpha \in \Lambda \setminus U$, so we have a contradiction. □

The following selection theorem, due to Kuratowski and Ryll-Nardzewski [51], is required for the analysis of the algorithm for solving (3.25). See Appendix B for a proof of this result.

Theorem 3.12. *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 non-empty and closed for every $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$. Then, for any Lebesgue measurable function $\mathbf{u} \colon \Omega \to \mathbb{R}^m$, there exists a Lebesgue measurable selection $\alpha \colon \Omega \to \Lambda$ such that $\alpha(x) \in \Lambda(x, \mathbf{u}(x))$ for a.e. $x \in \Omega$.*

For $u \in W^{2,r}(\Omega; \mathcal{T}_h)$, let $\Lambda[u]$ be the set of all Lebesgue measurable functions $\alpha \colon \Omega \to \Lambda$ such that $\alpha(x) \in \Lambda(x, \mathbf{u}(x))$ for a.e. $x \in \Omega$, where $\mathbf{u} = (u, \nabla u, D^2 u)$. Lemma 3.11 and Theorem 3.12 show that $\Lambda[u]$ is non-empty for each $u \in W^{2,r}(\Omega; \mathcal{T}_h)$. For measurable $\alpha \colon \Omega \to \Lambda$, we define $\gamma^{\alpha} \colon \Omega \to \mathbb{R}_{>0}$ through $\gamma^{\alpha}(x) = \gamma(x, \alpha(x))$, where $\gamma \colon \Omega \times \Lambda \to \mathbb{R}_{>0}$ was defined by (3.9) or (3.10). It follows from uniform continuity of γ over $\Omega \times \Lambda$ that $\gamma^{\alpha} \in L^{\infty}(\Omega)$, with $\|\gamma^{\alpha}\|_{L^{\infty}(\Omega)} \leq \|\gamma\|_{C(\overline{\Omega} \times \Lambda)}$. The functions a^{α}, b^{α}, c^{α} and f^{α} and the

operator L^α are defined in a similar way and are likewise bounded. It is clear that if $\alpha \in \Lambda[u]$, then $F_\gamma[u] = \gamma^\alpha(L^\alpha u - f^\alpha)$ a.e. in Ω.

3.6.1 Algorithm

We now present the definition of the semismooth Newton method for solving (3.25) and state the main result concerning its convergence rate. Choose $u_h^0 \in V_{h,\mathbf{p}}$. Given $u_h^k \in V_{h,\mathbf{p}}$, $k \in \mathbb{N}$, choose $\alpha_k \in \Lambda[u_h^k]$. Then, obtain $u_h^{k+1} \in V_{h,\mathbf{p}}$ satisfying

$$A_h^k(u_h^{k+1}, v_h) = \sum_{K\in\mathcal{T}_h} \langle \gamma^{\alpha_k} f^{\alpha_k}, L_\lambda v_h\rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}, \tag{3.61}$$

where the bilinear form $A_h^k \colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ is defined by

$$A_h^k(w_h, v_h) \coloneqq \sum_{K\in\mathcal{T}_h} \langle (\gamma^{\alpha_k} L^{\alpha_k} w_h, L_\lambda v_h\rangle_K + B_{h,1/2}(w_h, v_h) - \sum_{K\in\mathcal{T}_h} \langle L_\lambda w_h, L_\lambda v_h\rangle_K. \tag{3.62}$$

The fact that $\alpha_k \colon \Omega \to \Lambda$ is measurable ensures that A_h^k is well-defined. As in the proof of Theorem 3.8, it is found that the bilinear forms A_h^k, $k \in \mathbb{N}$, are coercive on $V_{h,\mathbf{p}}$. In fact, for each $k \in \mathbb{N}$, we have

$$\|v_h\|_{h,1}^2 \le \frac{2\,\kappa}{1-\kappa\,(1-\varepsilon)} A_h^k(v_h, v_h) \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{3.63}$$

Therefore, the sequence of iterates $\{u_h^k\}_{k=1}^\infty$ is well-defined by (3.61) and remains bounded in $V_{h,\mathbf{p}}$. The main result of this section is the following.

Theorem 3.13. *Under the hypotheses of Theorem* 3.8*, there exists a constant $R > 0$, possibly depending on h and $\mathbf{p}$, such that if $\|u_h - u_h^0\|_{h,1} < R$, then the sequence $\{u_h^k\}_{k=1}^\infty$ converges to u_h with a superlinear convergence rate.*

The proof of this theorem will be given in the next section.

Remark 3.3. It is known that it is generally not possible to establish mesh-independence of R for the class of operators considered here as a result of the so-called *norm gap* [41, 76]; see also Remark 3.4 below. However, it is seen from the numerical experiments in section 3.7, in particular in Figures 3.3 and 3.6 below, that in practice, the convergence rates of the algorithm depend only weakly on the discretisation parameters.

Since the bilinear forms A_h^k are stable in an H^2-type norm, the condition numbers of the resulting linear systems are typically large, thereby limiting the performance of many iterative solution algorithms. For common choices of basis $\{\phi_i\}$ of $V_{h,\mathbf{p}}$, it can be shown that the condition number $\kappa(\mathbf{A}^k)$ of the matrix $\mathbf{A}^k \coloneqq \big(A_h^k(\phi_i, \phi_j)\big)$ satisfies

$$\kappa(\mathbf{A}^k) \lesssim \max_{K\in\mathcal{T}_h} \frac{p_K^8}{h_K^4} \frac{\max_{K\in\mathcal{T}_h} h_K^d}{\min_{K\in\mathcal{T}_h} h_K^d}. \tag{3.64}$$

Therefore, there can be significant benefits in using preconditioners when solving these linear systems with iterative methods such as GMRES. In Chapter 5, based on our work [69], we develop and analyse a class of nonoverlapping domain decomposition preconditioners to accelerate the iterative solution of the linear problems (3.61). For computational efficiency, it is desirable to find a preconditioner that can be assembled once and used for each of the linear problems encountered in the semismooth Newton method. This requires finding a preconditioner for the linear problems (3.61) that remains robust for a large range of coefficients appearing in the definition of the bilinear forms A_h^k in (3.62).

3.6.2 Semismoothness

The proof of Theorem 3.13 rests upon the notion of semismoothness, as defined in [76]. We recall the definition below. For sets X and Y, we write $G\colon X \rightrightarrows Y$ if G is a set-valued map that maps X into the subsets of Y.

Definition 3.1. Let X and Y be Banach spaces, and let $F\colon U \subset X \to Y$ be a map defined on a non-empty open set U of X. Let $DF\colon U \rightrightarrows \mathcal{L}(X,Y)$ be a set-valued map with non-empty images. For $x \in U$, the map F is called *DF-semismooth at x* if

$$\lim_{\|e\|_X \to 0} \frac{1}{\|e\|_X} \sup_{D \in DF[x+e]} \|F[x+e] - F[x] - De\|_Y = 0. \tag{3.65}$$

The map F is called *DF-semismooth on U* if F is DF-semismooth at x, for every $x \in U$. The set-valued map DF is then called a *generalised differential of F on U*.

For $1 \le q < r \le \infty$, the map $DF_\gamma\colon W^{2,r}(\Omega;\mathcal{T}_h) \rightrightarrows \mathcal{L}\left(W^{2,r}(\Omega;\mathcal{T}_h), L^q(\Omega)\right)$ is defined by

$$DF_\gamma[u] \coloneqq \left\{ \gamma^\alpha L^\alpha \coloneqq \gamma^\alpha \left(a^\alpha\colon D^2 + b^\alpha \cdot \nabla - c^\alpha \right) : \alpha \in \Lambda[u] \right\}. \tag{3.66}$$

Theorem 3.14. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set, let Λ be a compact metric space, let the data a, b, c and f be continuous on $\overline{\Omega} \times \Lambda$, and suppose that* (3.5) *holds. Let $\mathcal{T}_h$ be a mesh on Ω. Then, for any $1 \le q < r \le \infty$, the operator $F_\gamma\colon W^{2,r}(\Omega;\mathcal{T}_h) \to L^q(\Omega)$ defined by $F_\gamma[u] = F_\gamma(\cdot, u, \nabla u, D^2 u)$ is DF_γ-semismooth on $W^{2,r}(\Omega;\mathcal{T}_h)$.*

Proof. Supposing the claim to be false, there exist a function $u \in W^{2,r}(\Omega;\mathcal{T}_h)$, a constant $\rho > 0$, and a sequence $\{e_j\}_{j=0}^\infty \subset W^{2,r}(\Omega;\mathcal{T}_h)$, with $\|e_j\|_{W^{2,r}(\Omega;\mathcal{T}_h)} \to 0$, and $\alpha_j \in \Lambda[u+e_j]$ such that, for each $j \in \mathbb{N}$,

$$\frac{1}{\|e_j\|_{W^{2,r}(\Omega;\mathcal{T}_h)}} \|F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j\|_{L^q(\Omega)} > \rho. \tag{3.67}$$

We will show that there is a subsequence for which (3.67) is violated, and thus obtain a contradiction. Since $\|e_j\|_{W^{2,r}(\Omega;\mathcal{T}_h)} \to 0$, by passing to a subsequence without change of notation, we may assume that e_j and its first and second broken derivatives tend to 0

pointwise a.e. in Ω. The following inequality will help to simplify the argument:

$$|F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j| \lesssim G_j \left(|e_j| + |\nabla e_j| + |D^2 e_j| \right), \tag{3.68}$$

where $G_j \colon \Omega \to \mathbb{R}_{\geq 0}$ is defined by

$$G_j \coloneqq \inf_{\alpha \in \Lambda(\cdot, \mathbf{u}(\cdot))} |\gamma^\alpha a^\alpha - \gamma^{\alpha_j} a^{\alpha_j}| + |\gamma^\alpha b^\alpha - \gamma^{\alpha_j} b^{\alpha_j}| + |\gamma^\alpha c^\alpha - \gamma^{\alpha_j} c^{\alpha_j}|. \tag{3.69}$$

It can be deduced from Lemma 3.11 that G_j is measurable, since it is the composition of a lower semicontinuous function with a measurable function; compactness of Λ and continuity of the data imply that $\|G_j\|_{L^\infty(\Omega)}$ is uniformly bounded for all $j \in \mathbb{N}$.

We prove (3.68): since $\alpha_j \in \Lambda[u+e_j]$, we have a.e. in Ω:

$$F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j = \gamma^{\alpha_j} \left(L^{\alpha_j} u - f^{\alpha_j} \right) - F_\gamma[u] \leq 0. \tag{3.70}$$

Now, for a.e. $x \in \Omega$, and arbitrary $\alpha \in \Lambda(x, \mathbf{u}(x))$, we have

$$\begin{aligned} 0 &\leq F_\gamma[u+e_j] - \gamma^\alpha \left(L^\alpha (u+e_j) - f^\alpha \right) \\ &= \gamma^{\alpha_j} \left(L^{\alpha_j} u - f^{\alpha_j} \right) - F_\gamma[u] + \left(\gamma^{\alpha_j} L^{\alpha_j} - \gamma^\alpha L^\alpha \right) e_j \\ &= F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j + \left(\gamma^{\alpha_j} L^{\alpha_j} - \gamma^\alpha L^\alpha \right) e_j, \end{aligned} \tag{3.71}$$

where it is understood that the above expressions are evaluated at x. Rearranging (3.70) and (3.71) gives $\left(\gamma^\alpha L^\alpha - \gamma^{\alpha_j} L^{\alpha_j} \right) e_j \leq F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j \leq 0$, so

$$|F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j| \leq |\left(\gamma^\alpha L^\alpha - \gamma^{\alpha_j} L^{\alpha_j} \right) e_j|. \tag{3.72}$$

Since (3.72) holds for arbitrary $\alpha \in \Lambda(x, \mathbf{u}(x))$, we readily obtain (3.68).

We claim that $G_j \to 0$ pointwise a.e. in Ω. Recall that $\mathbf{e}_j \coloneqq (e_j, \nabla e_j, D^2 e_j)$ tends to zero pointwise a.e. in Ω. Let $\varrho > 0$ and $x \in \Omega$ be such that $\mathbf{e}_j(x) \to 0$. Then, by continuity of the data on the compact metric space $\overline{\Omega} \times \Lambda$, there is a $\delta > 0$ such that, for any α, $\beta \in \Lambda$ with $\operatorname{dist}(\alpha, \beta) < \delta$,

$$|\gamma^\alpha a^\alpha - \gamma^\beta a^\beta| + |\gamma^\alpha b^\alpha - \gamma^\beta b^\beta| + |\gamma^\alpha c^\alpha - \gamma^\beta c^\beta| < \varrho \quad \text{at } x \in \Omega.$$

Since $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$ is upper-semicontinuous by Lemma 3.11, there is an $N \in \mathbb{N}$ such that for each $j \geq N$, there is an $\alpha \in \Lambda(x, \mathbf{u}(x))$ with $\operatorname{dist}(\alpha, \alpha_j(x)) < \delta$. Therefore $0 \leq G_j(x) < \varrho$ for all $j \geq N$, and hence $G_j \to 0$ pointwise a.e. in Ω.

Because $1 \leq q < r \leq \infty$, setting $s = r/q > 1$ and s' such that $1/s + 1/s' = 1$, we have $1 \leq s' < \infty$. Inequality (3.68) followed by an application of Hölder's inequality shows that

$$\frac{1}{\|e_j\|_{W^{2,r}(\Omega; \mathcal{T}_h)}} \|F_\gamma[u+e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j\|_{L^q(\Omega)} \lesssim \|G_j\|_{L^{qs'}(\Omega)}, \tag{3.73}$$

Since $G_j \to 0$ pointwise a.e. and $\{G_j\}_{j=0}^{\infty}$ is uniformly bounded in $L^\infty(\Omega)$, the dominated convergence theorem implies that $\|G_j\|_{L^{qs'}(\Omega)} \to 0$. Therefore, (3.73) contradicts (3.67), and F_γ is DF_γ-semismooth at u, thus completing the proof. □

Remark 3.4. The restriction $q < r$ in Theorem 3.14 cannot be relaxed in general, as evidenced by the counter-example in [41] involving a special case of the class of operators considered here. This is an example of the so-called *norm gap* described in [41, 76].

Proof of Theorem 3.13. Since $\alpha_k \in \Lambda[u_h^k]$ for each k, we have $F_\gamma[u_h^k] = \gamma^{\alpha_k} L^{\alpha_k} u_h^k - \gamma^{\alpha_k} f^{\alpha_k}$. Therefore, (3.61) is equivalent to

$$A_h^k(u_h^{k+1}, v_h) = \sum_{K\in\mathcal{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h^k - F_\gamma[u_h^k], L_\lambda v_h\rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{3.74}$$

The definition of the numerical scheme (3.25) implies that u_h satisfies

$$A_h^k(u_h, v_h) = \sum_{K\in\mathcal{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h - F_\gamma[u_h], L_\lambda v_h\rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{3.75}$$

After subtracting (3.75) from (3.74), the bound (3.63) then shows that

$$\|u_h^{k+1} - u_h\|_{h,1} \le C_1 \|F_\gamma[u_h^k] - F_\gamma[u_h] - \gamma_k^\alpha L_k^\alpha (u_h^k - u_h)\|_{L^2(\Omega)}, \tag{3.76}$$

where the constant C_1 depends only on κ, ε, γ, and d, but not on k. Fix $r > 2$; since $V_{h,\mathbf{p}}$ is finite-dimensional, there is a constant C_2 depending on h and $\mathbf{p}$ such that $\|v_h\|_{W^{2,r}(\Omega;\mathcal{T}_h)} \le C_2\|v_h\|_{h,1}$ for all $v_h \in V_{h,\mathbf{p}}$. Theorem 3.14 shows that for each $\rho \in (0,1)$, there is a $R_\rho > 0$ such that if $\|w_h - u_h\|_{h,1} < R_\rho$, then, for any $\alpha \in \Lambda[w_h]$,

$$\|F_\gamma[w_h] - F_\gamma[u_h] - \gamma^\alpha L^\alpha (w_h - u_h)\|_{L^2(\Omega)} \le \frac{\rho}{C_1 C_2} \|w_h - u_h\|_{W^{2,r}(\Omega;\mathcal{T}_h)}. \tag{3.77}$$

If $\|u_h^0 - u_h\|_{h,1} < R_\rho$ for some $\rho < 1$, then we use (3.76) and (3.77) to obtain

$$\|u_h^{k+1} - u_h\|_{h,1} \le \rho \|u_h^k - u_h\|_{h,1} \quad \forall\, k \ge 0 \quad \Longrightarrow \lim_{k\to\infty} \|u_h^k - u_h\|_{h,1} = 0.$$

The convergence is superlinear since $\|u_h^k - u_h\|_{h,1} < R_\rho$ is eventually satisfied $\forall\, \rho < 1$. □

3.7 Numerical experiments

We provide the results of two tests of the scheme on problems with strongly anisotropic diffusion coefficients, and an experiment for a solution that does not meet the regularity assumption of the analysis of section 3.5.1.

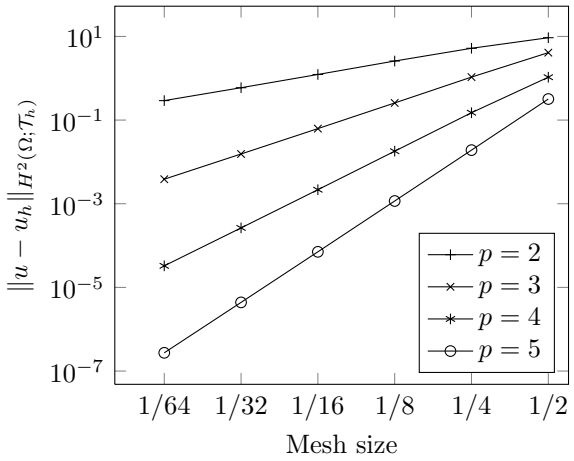

FIGURE 3.1: *The errors in approximating the solution of the problem of section* 3.7.1 *for various mesh sizes and polynomial degrees. The optimal convergence rates* $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)} \simeq h^{p-1}$ *are observed.*

Mesh size	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
1/2	9.31		4.09		1.06		3.19×10^{-1}	
1/4	5.20	(0.84)	1.07	(1.94)	1.50×10^{-1}	(2.82)	1.91×10^{-2}	(4.06)
1/8	2.58	(1.01)	2.57×10^{-1}	(2.05)	1.80×10^{-2}	(3.06)	1.16×10^{-3}	(4.04)
1/16	1.23	(1.07)	6.25×10^{-2}	(2.04)	2.16×10^{-3}	(3.06)	7.09×10^{-5}	(4.04)
1/32	5.94×10^{-1}	(1.05)	1.55×10^{-2}	(2.01)	2.64×10^{-4}	(3.03)	4.38×10^{-6}	(4.02)
1/64	2.92×10^{-1}	(1.02)	3.86×10^{-3}	(2.00)	3.28×10^{-5}	(3.01)	2.73×10^{-7}	(4.00)

TABLE 3.2: *Errors* $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)}$ *for the numerical scheme applied to problem of section* 3.7.1. *The estimated orders of convergence between successive mesh refinements are given in parentheses.*

3.7.1 First experiment

We consider once again Example 3.1 for testing the accuracy of the scheme and the performance of the semismooth Newton method. Recalling that $\Lambda = [0, \pi/3] \times \mathrm{SO}(2)$ and $a^\alpha = \sigma^\alpha(\sigma^\alpha)^\top/2$, with σ^α given by (3.8), let $\Omega = (0,1)^2$, let $b^\alpha \equiv 0$, $c^\alpha \equiv \pi^2$ and choose $f^\alpha \equiv \sqrt{3}\sin^2\theta/\pi^2 + g$, g independent of α, so that the exact solution of the HJB equation (3.4) is $u(x,y) = \exp(xy)\sin(\pi x)\sin(\pi y)$. These choices are made so that the optimal controls vary significantly throughout the domain, and to ensure that the corresponding diffusion coefficient is not diagonally dominant in parts of Ω.

The numerical scheme (3.25) is applied with meshes obtained by regular subdivision of Ω into uniform quadrilateral elements of size $h = 2^{-k}$, $1 \le k \le 6$. The finite element spaces $V_{h,\mathbf{p}}$ are defined by employing the space of polynomials of fixed total degree p on each element, with $2 \le p \le 5$. The penalty parameters are set to $c_\mu = c_\eta = 10$ and $\eta_F = c_\eta \tilde{p}_F^4/\tilde{h}_F^3$. Figure 3.1 confirms the optimal convergence rates with respect to mesh refinement that are predicted by Theorem 3.9. The numerical solutions were obtained by the semismooth Newton method of section 3.6, for which we use a strict convergence criterion by requiring a relative residual below 5×10^{-12} and a step-increment L^2-norm below 1×10^{-11}.

The initial guess used for each computation was $u_h^0 \equiv 0$. Figure 3.3 demonstrates the fast convergence of the algorithm.

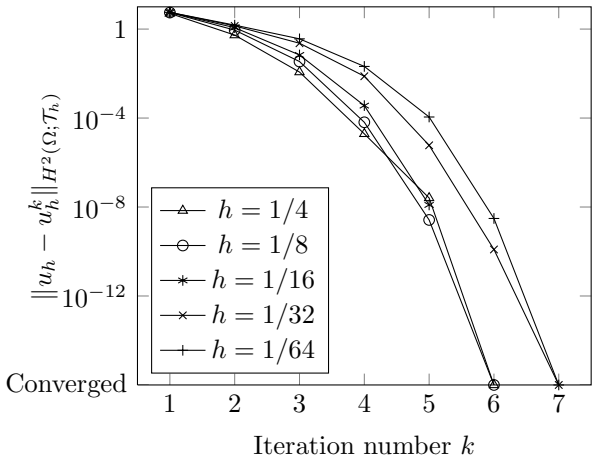

FIGURE 3.3: *Convergence histories of the semismooth Newton method applied to the problem of section* 3.7.1 *on successively refined meshes, with* $p = 4$. *The predicted superlinear convergence rate is observed, and the number of iterations required for convergence varies little under refinement.*

3.7.2 Second experiment

We investigate the robustness of the scheme against a combination of near-degenerate diffusions, nonsmooth solutions and boundary layers. Let $\Omega = (0,1)^2$, $b^\alpha \equiv (0,1)$, $c^\alpha \equiv 10$ and define

$$a^\alpha \coloneqq \alpha^\top \begin{pmatrix} 20 & 1 \\ 1 & 0.1 \end{pmatrix} \alpha, \quad \alpha \in \Lambda \coloneqq \mathrm{SO}(2). \tag{3.78}$$

For $\lambda = 1/2$, the Cordes condition (3.6) holds with $\varepsilon \approx 0.0024$. We choose f^α so that the solution of the corresponding HJB equation is

$$u(x,y) = (2x-1)\left(\mathrm{e}^{1-|2x-1|} - 1\right)\left(y + \frac{1-\mathrm{e}^{y/\delta}}{\mathrm{e}^{1/\delta}-1}\right), \quad \delta > 0. \tag{3.79}$$

We choose $\delta = 5 \times 10^{-3}$ to be of same order as ε, thus leading to a sharp boundary layer in a neighbourhood of $\left\{(x,y) \in \overline{\Omega} \colon y = 1\right\}$. As explained in section 1.3, monotone FDM require very large stencils to discretise this problem. On uniform grids, these low-order methods would require a fine grid to resolve the boundary layer, whilst the use of locally refined grids is complicated by the monotonicity requirements.

Our method features no such constraints, so we are free to take advantage of *hp*-refinement techniques that are capable of delivering highly accurate approximations for a smaller computational cost. Following a suggestion in [57], we perform a sequence of com-

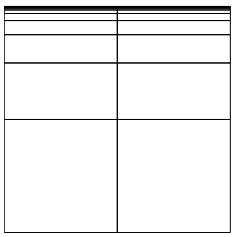

FIGURE 3.4: *Mesh on Ω used for the approximation of* (3.79)*. The origin is at the bottom left corner. The mesh has* 8 *geometrically refined layers with grading factor* $1/2$*.*

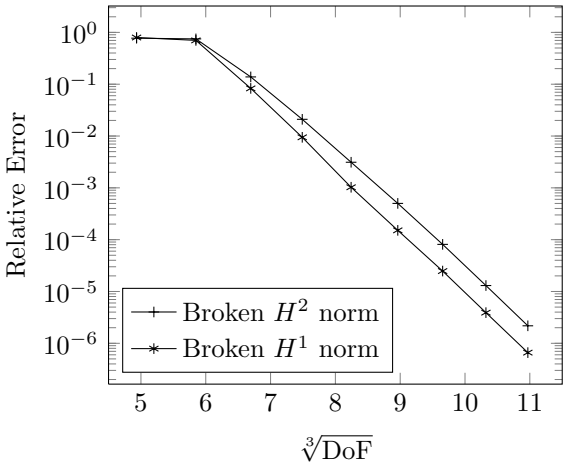

FIGURE 3.5: *Exponential convergence in the broken H^1 and H^2-norms of the approximations to the solution defined by* (3.79)*. The relative errors $\|u - u_h\|/\|u\|$ are plotted against the cube root of the number of degrees of freedom, with each data point corresponding to a computation using a total polynomial degree $p = 2, \dots, 10$.*

putations by increasing the uniform polynomial degrees p from 2 to 10 on a fixed mesh shown in Figure 3.4. The number of degrees of freedom ranges from 120 to 1320. The following results were obtained with $c_\mu = c_\eta = 10$, as in section 3.7.1. Here, we use $\eta_F = \lambda\, c_\eta\, \tilde{p}_F^4/\tilde{h}_F^3$. Figure 3.5 shows that the error converges with a rate of order $\exp\big(-c\sqrt[3]{\text{DoF}}\big)$, as expected from the results in [78], which leads to high accuracy with few degrees of freedom.

To study the convergence behaviour of the semismooth Newton method as applied to this problem, we compute the numerical solution u_h to high accuracy and obtain the relative errors of the successive iterates u_h^k in the broken H^2-norm. The initial guess used for each computation was $u_h^0 \equiv 0$. The results are presented in Figure 3.6, which shows that, for this example, attaining a relative error of 10^{-10} requires one additional Newton step per polynomial degree p. This confirms the observation made in Remark 3.3 that mesh-independence of the superlinear convergence radius cannot be achieved for this class of operators, although the dependence is not severe.

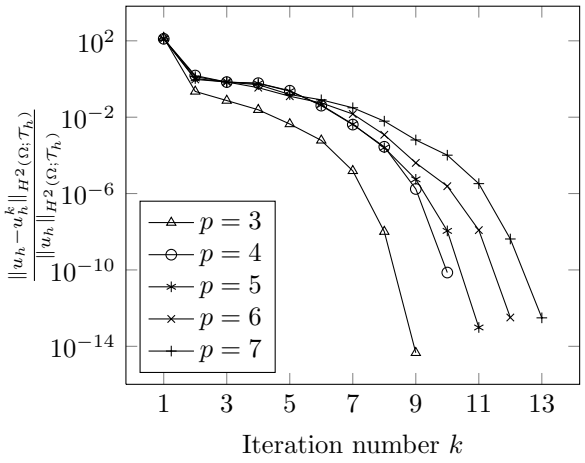

FIGURE 3.6: *Convergence histories of the semismooth Newton method applied to the problem of section* 3.7.2. *One additional Newton step per polynomial degree p is required in order to attain a relative error below 10^{-10}. The number of iterations required for convergence thus increases slowly with the polynomial degree p.*

3.7.3 Third experiment

We consider the convergence of the scheme when relaxing the assumption on the solution of broken H^s-regularity for some $s > 5/2$. We also treat a problem with an inhomogeneous Dirichlet boundary condition, as explained in section 2.7.2.

Consider a hexagonal domain $\Omega \subset \mathbb{R}^2$ with unit face length, as shown in Figure 3.7. Laplace's equation, $\Delta u = 0$ in Ω, is a special case of the HJB equation at hand, and the boundary condition $u = g$ on $\partial\Omega$ is chosen such that $u = r^{3/2}\sin(\frac{3}{2}\theta)$, where r is the distance to the upper vertex of Ω, and θ is the counter-clockwise angle from the upper left face of Ω. It follows that broken H^s-regularity of u fails for $s \geq 5/2$.

The first set of computations use uniform h-refinement, whereas the second uses hp-refinement on geometrically graded meshes with linearly increasing polynomial degrees away from the singularity of the solution. Figure 3.7 illustrates the respective sequences of meshes. We use $c_\mu = c_\eta = 10$ and $\eta_F = c_\eta\, \tilde{p}_F^4/\tilde{h}_F^3$.

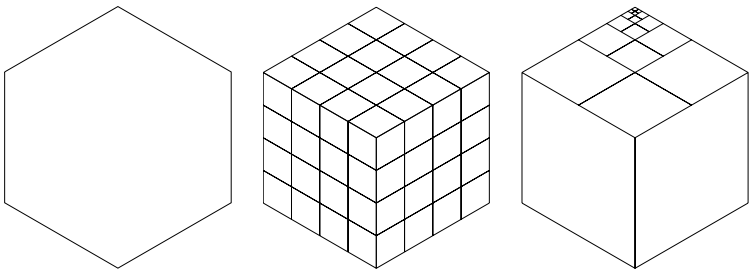

FIGURE 3.7: *The planar hexagonal domain $\Omega \subset \mathbb{R}^2$ considered in the experiment of section* 3.7.3 *with uniformly refined and geometrically graded parallelepipedal meshes.*

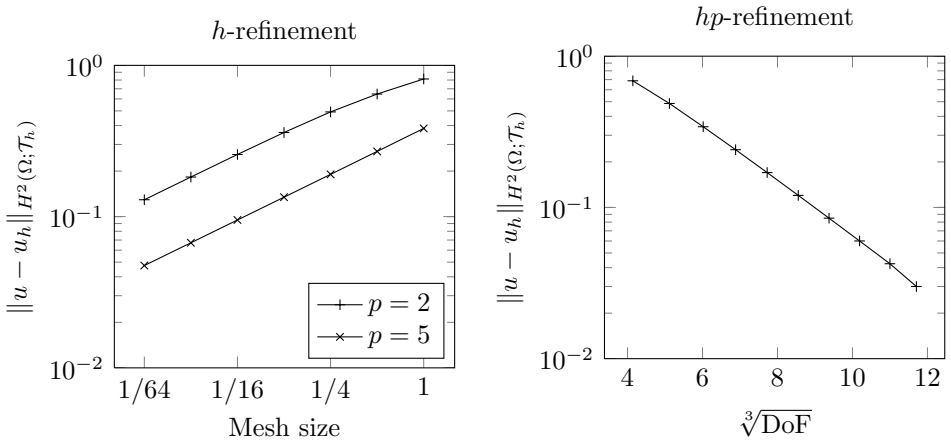

FIGURE 3.8: *Convergence in the broken H^2-norm of the approximations to the singular solution on the hexagonal domain. We observe approximate convergence rates of order $h^{1/2}$ for uniform h-refinement, whilst rates of order* $\exp\big(-c\sqrt[3]{\text{DoF}}\big)$ *are obtained under hp-refinement. The number of degrees of freedom used for the largest computations were* 1604 *(hp-refinement),* 73728 *(h-refinement, $p = 2$) and* 258048 *(h-refinement, $p = 5$), thus showing the efficiency of hp-refinement.*

Figure 3.8 shows that the method is convergent under both refinement strategies. In the case of h-refinement, the convergence rate is approximately of order $h^{1/2}$, irrespective of the polynomial degree, as may be expected given the limited regularity of the solution. In the case of hp-refinement, the rate is of order $\exp\big(-c\sqrt[3]{\text{DoF}}\big)$, where DoF is the number of degrees of freedom.

These results show that the regularity assumption on the solution used in the analysis is not a necessary condition for the convergence of the numerical scheme. Furthermore, the ability to use hp-refinement is a significant advantage for computational efficiency.

Chapter 4

Parabolic Hamilton–Jacobi–Bellman equations

In this chapter, we consider the Cauchy–Dirichlet problem for Hamilton–Jacobi–Bellman (HJB) equations of the form

$$\partial_t u - \sup_{\alpha\in\Lambda}[L^\alpha u - f^\alpha] = 0 \qquad \text{in } \Omega \times I, \tag{4.1}$$

where $I = (0, T)$ is a bounded time interval, and the coefficients of the nondivergence form elliptic operators L^α, as well as f^α, are allowed to depend on space and time.

This chapter extends the spatial discretisation of HJB equations from the earlier chapters by a discontinuous Galerkin (DG) time-stepping scheme [74]. DG time discretisations offer many advantages over other discretisations, since they possess excellent stability properties and they allow very general choices of time steps and orders of accuracy [2, 74]. For instance, they allow arbitrarily high-order approximation without restrictions on the size of the time-step. Moreover, these methods can yield exponential convergence rates for solutions with limited early-time regularity, commonly encountered in parabolic problems, through τq-refinement [67].

A key contribution of this chapter concerns the construction of an appropriate DG time-stepping scheme that addresses a particular difficulty that we now describe. Recall that the approach based on the Cordes condition involves a renormalisation of the HJB operator, as shown in Chapter 3. In the present context, we find that (4.1) is equivalent to

$$\inf_{\alpha\in\Lambda}\left[\gamma^\alpha\left(\partial_t u - L^\alpha u + f^\alpha\right)\right] = 0 \quad \text{in } \Omega \times I, \tag{4.2}$$

where γ^α is an appropriate scalar function. It is clear that the nonlinear operator in (4.2) now includes the time derivative of the solution inside the nonlinearity. Unfortunately, this implies that the standard DG time-stepping method cannot be used, since the time derivative cannot be cast onto a test function.

Keeping to the approach of the earlier chapters, where the choice of discretisation follows closely the analysis of the continuous problem, we first find an appropriate functional setting in which the Cauchy–Dirichlet problem associated to (4.1) is well-posed. It turns out that the Cordes condition leads to well-posedness in a $H^1(L^2) \cap L^2(H^2)$-type Bochner norm. Similarly to Chapter 3, the analysis rests upon the reformulation of the problem as a variational equation involving a strongly monotone operator followed by an application of the Browder–Minty theorem.

The continuous analysis helps to determine the essential components of the time-stepping scheme; we present these ideas first in a semidiscrete context in section 4.2 in order to simplify the presentation. The fully discrete method is then given in section 4.3, and is shown to be consistent in section 4.4. Importantly, we show that in section 4.5 that this nonstandard DG time-stepping method is stable in a discrete $H^1(L^2) \cap L^2(H^2)$-type norm, analogous to the continuous analysis.

The second main contribution of this chapter concerns the error analysis of the proposed method. We note that the techniques of error analysis in the literature on DG time discretisations of parabolic equations often require sufficient smoothness of the solution [2, 67], which, in the present setting, would correspond to assuming $H^1(H^2)$-regularity. For instance, employing an approximation result from [67], we show in section 4.6.1 that for sufficiently smooth solutions, quasi-uniformity of the mesh, time partitions, and polynomial degrees leads to an error bound of the form

$$\begin{aligned}(4.3)\quad \|u - u_h\| \lesssim{}& \frac{h^{\min(s,\,p+1)-2}}{p^{s-7/2}}\|u\|_{L^2(H^s)} + \frac{h^{\min(\bar{s},\,p+1)}}{p^{\bar{s}}}\|u\|_{H^1(H^{\bar{s}})} \\ &+ p^{3/2}\sum_{\ell\in\{0,2\}} \frac{\tau^{\min(\sigma_\ell,\,q+1)-1+\ell/2}}{q^{\sigma_\ell-1+\ell/2}}\|u\|_{H^{\sigma_\ell}(H^\ell)} + \frac{h^{\min(\tilde{s},\,p+1)-1}}{p^{\tilde{s}-3/2}}\|u(0)\|_{H^{\tilde{s}}},\end{aligned}$$

where the norm $\|\cdot\|$ is a discrete $H^1(L^2) \cap L^2(H^2)$-type norm, and we assume that $s > 5/2$, $\bar{s} > 0$, $\tilde{s} > 3/2$, and $\sigma_\ell \geq 1$ for $\ell \in \{0, 2\}$. Note that for $\ell = 2$, this requires at least $H^1(H^2)$-regularity. This bound implies that the method has optimal convergence rates in terms of the mesh size h, time-interval length τ, and temporal polynomial degrees q; the rates in the spatial polynomial degrees p are possibly suboptimal by an order and a half, as is common for DGFEM that are stable in discrete H^2-norms [60], but which is weaker than our results in Chapters 2 and 3. The reason for this difference lies with additional terms required for stability, as shown in section 4.5.

A heuristic argument based on parabolic regularity theory shows that assuming $H^1(H^2)$-regularity is comparable to assuming $L^2(H^4)$-regularity. Therefore, this appears as rather more restrictive than the regularity assumptions of the preceding chapters, which in the present context would amount to assuming $L^2(H^s)$-regularity for $s > 5/2$. Therefore, we use Clément-type projection operators to obtain bounds assuming only $H^\sigma(H^2)$-regularity for any $\sigma \geq 0$.

To distinguish the different techniques of analysis employed to obtain these results, we shall therefore refer to such problems as having *solutions with low regularity*, as opposed to the case of *regular solutions* when $\sigma \geq 1$. In particular, the bounds for low-regularity solutions cover problems with early-time singularities induced by the initial datum.

In section 4.7.1, we test the numerical scheme on a problem with strongly anisotropic diffusion in order to demonstrate the method's accuracy and efficiency. Furthermore, we show in the numerical experiment of section 4.7.2 that our method offers exponential convergence rates for solutions with low regularity under *hp*- and τq-refinement.

4.1 Analysis of the continuous problem

Let Ω be a bounded convex polytopal open set in $\mathbb{R}^d$, $d \geq 2$, let Λ be a compact metric space, and let $I \coloneqq (0,T)$, with $T > 0$. It is assumed that Ω and Λ are non-empty. Convexity of Ω implies that the boundary $\partial\Omega$ of Ω is Lipschitz [40]. Let the symmetric $\mathbb{R}^{d\times d}$-valued function a, the $\mathbb{R}^d$-valued function b, and scalar-valued functions c and f be continuous on $\overline{\Omega} \times \overline{I} \times \Lambda$. For each $\alpha \in \Lambda$, define the functions $a^\alpha \colon (x,t) \mapsto a(x,t,\alpha)$, where $(x,t) \in \overline{\Omega} \times \overline{I}$; the functions b^α, c^α and f^α are similarly defined.

The operators $L^\alpha \colon L^2(I;H^2(\Omega)) \to L^2(I;L^2(\Omega))$ are given by

$$L^\alpha v \coloneqq a^\alpha : D^2 v + b^\alpha \cdot \nabla v - c^\alpha v, \quad v \in L^2(I;H^2(\Omega)),\ \alpha \in \Lambda, \tag{4.4}$$

where $D^2 v$ denotes the Hessian matrix of v. Compactness of Λ and continuity of the functions a, b, c and f imply that the fully nonlinear operator F, given by

$$F \colon v \mapsto F[v] \coloneqq \partial_t v - \sup_{\alpha\in\Lambda} \left[L^\alpha v - f^\alpha\right] = \inf_{\alpha\in\Lambda} \left[\partial_t v - L^\alpha v + f^\alpha\right], \tag{4.5}$$

is well-defined as a mapping from the space $H(I;\Omega)$ into $L^2(I;L^2(\Omega))$, where $H(I;\Omega)$ is defined by

$$H(I;\Omega) \coloneqq L^2(I;H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(I;L^2(\Omega)). \tag{4.6}$$

The problem considered is to find a function $u \in H(I;\Omega)$ that is a strong solution of the parabolic HJB equation subject to Cauchy–Dirichlet boundary conditions:

$$\begin{aligned} F[u] &= 0 \quad && \text{in } \Omega \times I, \\ u &= 0 \quad && \text{on } \partial\Omega \times I, \\ u &= u_0 \quad && \text{on } \Omega \times \{0\}, \end{aligned} \tag{4.7}$$

where $u_0 \in H_0^1(\Omega)$. Note that the lateral condition $u = 0$ on $\partial\Omega \times I$ is incorporated in the function space $H(I;\Omega)$.

The PDE is said to be *uniformly parabolic* if there exist constants $0 < \nu \leq \overline{\nu}$ such that

$$\nu\,|\xi|^2 \leq \xi^\top a^\alpha(x,t)\,\xi \leq \overline{\nu}\,|\xi|^2 \quad \forall\,\xi \in \mathbb{R}^d,\ \forall\,(x,t) \in \Omega \times I,\ \forall\,\alpha \in \Lambda. \tag{4.8}$$

Well-posedness of (4.7) is established in section 4.1 under the following hypotheses. We assume uniform parabolicity, nonnegativity of c, and the Cordes condition [70, 71]: there exist $\varepsilon \in (0,1]$, $\lambda > 0$ and $\omega > 0$ such that

$$\frac{|a^\alpha|^2 + 1/\lambda^2 + 1/\omega^2}{(\operatorname{Tr} a^\alpha + 1/\lambda + 1/\omega)^2} \leq \frac{1}{d+1+\varepsilon} \quad \text{in } \overline{\Omega} \times \overline{I},\ \forall\,\alpha \in \Lambda. \tag{4.9}$$

In the special case where $b \equiv 0$ and $c \equiv 0$, we set $\lambda = 0$ and assume that there exist $\varepsilon \in (0,1]$ and $\omega > 0$ such that

$$\frac{|a^\alpha|^2 + 1/\omega^2}{(\operatorname{Tr} a^\alpha + 1/\omega)^2} \leq \frac{1}{d+\varepsilon} \quad \text{in } \overline{\Omega} \times \overline{I},\ \forall\,\alpha \in \Lambda. \tag{4.10}$$

As explained in previous chapters, the quantities λ and ω are included to make the Cordes condition invariant under rescaling of the spatial and temporal domains.

Given (4.9), by considering substitutions for the solution of the form $u = \mathrm{e}^{\mu t}\tilde{u}$, we can assume without loss of generality that there exist $\varepsilon \in (0,1]$, $\lambda > 0$ and $\omega > 0$ such that

$$\frac{|a^\alpha|^2 + |b^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2 + 1/\omega^2}{(\operatorname{Tr} a^\alpha + c^\alpha/\lambda + 1/\omega)^2} \leq \frac{1}{d+1+\varepsilon} \quad \text{in } \overline{\Omega} \times \overline{I},\ \forall\,\alpha \in \Lambda. \tag{4.11}$$

The relevance of (4.9) is to show that the Cordes condition is essentially independent of the lower-order terms b^α and c^α, although it will be simpler to work with (4.11). Define the strictly positive function $\gamma\colon \Omega \times I \times \Lambda \to \mathbb{R}_{>0}$ by

$$\gamma(x,t,\alpha) \coloneqq \frac{\operatorname{Tr} a^\alpha(x,t) + c^\alpha/\lambda + 1/\omega}{|a^\alpha(x,t)|^2 + |b^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2 + 1/\omega^2}. \tag{4.12}$$

In the case of $b \equiv 0$ and $c \equiv 0$, the function γ is defined by

$$\gamma(x,t,\alpha) \coloneqq \frac{\operatorname{Tr} a^\alpha(x,t) + 1/\omega}{|a^\alpha(x,t)|^2 + 1/\omega^2}. \tag{4.13}$$

Continuity of the data implies that $\gamma \in C(\overline{\Omega} \times \overline{I} \times \Lambda)$, and it follows from (4.8) that there exists a positive constant $\gamma_0 > 0$ such that $\gamma \geq \gamma_0$ on $\overline{\Omega} \times \overline{I} \times \Lambda$. For each $\alpha \in \Lambda$, define $\gamma^\alpha\colon (x,t) \mapsto \gamma(x,t,\alpha)$, and define the operator $F_\gamma\colon H(I;\Omega) \to L^2(I;L^2(\Omega))$ by

$$F_\gamma[v] \coloneqq \inf_{\alpha\in\Lambda} \left[\gamma^\alpha\left(\partial_t v - L^\alpha v + f^\alpha\right)\right]. \tag{4.14}$$

It is important to observe that the operator F_γ includes $\partial_t u$ inside the nonlinearity.

For ω and λ as in (4.11), we introduce the operators L_λ and L_ω defined by

$$L_\lambda v \coloneqq \Delta v - \lambda v \qquad L_\omega v \coloneqq \omega\, \partial_t v - L_\lambda v. \tag{4.15}$$

The following result is similar to Lemma 3.1, so the proof is omitted here.

Lemma 4.1. *Let Ω be a bounded open subset of $\mathbb{R}^d$, let $I = (0,T)$, and suppose that (4.11) holds, or that (4.10) holds if $b \equiv 0$ and $c \equiv 0$. Let $U \subset \Omega$ be an open set, let $J \subset I$ be an open interval, and let the functions u, $v \in L^2(J; H^2(U)) \cap H^1(J; L^2(U))$, and set $w \coloneqq u - v$. Then, the following inequality holds a.e. in U, for a.e. $t \in J$:*

$$|F_\gamma[u] - F_\gamma[v] - L_\omega w| \le \sqrt{1-\varepsilon}\, \sqrt{\omega^2 |\partial_t w|^2 + |D^2 w|^2 + 2\lambda |\nabla w|^2 + \lambda^2 |w|^2}, \tag{4.16}$$

with $\lambda = 0$ if $b \equiv 0$ and $c \equiv 0$.

Gelfand triple. For shorthand, we define the space H by

$$H \coloneqq H^2(\Omega) \cap H_0^1(\Omega).$$

Theorem 3.2 shows that H is a Hilbert space when equipped with the inner-product $\langle u, v \rangle_\Delta \coloneqq \langle L_\lambda u, L_\lambda v \rangle_{L^2(\Omega)}$, where L_λ is from (4.15) and $\lambda \ge 0$ is from (4.11). It is possible to identify H^*, the dual space of H, with $L^2(\Omega)$ through the duality pairing

$$\langle f, v \rangle_{L^2 \times H} \coloneqq \int_\Omega f\, (-L_\lambda v)\, \mathrm{d}x, \quad f \in L^2(\Omega),\ v \in H. \tag{4.17}$$

Indeed, we clearly have $L^2(\Omega) \hookrightarrow H^*$, and H^2-regularity of solutions of Poisson's equation in convex domains [40] shows that this embedding is an isometry: for any $f \in L^2(\Omega)$, we have $\|f\|_{L^2(\Omega)} = \|f\|_{H^*}$. To show the surjectivity of this embedding, note that if $\varphi \in H^*$, then the Riesz representation theorem implies that there is a unique $w \in H$ such that $\langle w, v \rangle_\Delta = \varphi(v)$ for all $v \in H$. Then $f = -L_\lambda w \in L^2(\Omega)$ satisfies $\langle f, v \rangle_{L^2 \times H} = \varphi(v)$ for all $v \in H$. It follows from Poincaré's inequality that the space $H_0^1(\Omega)$ may be equipped with the inner-product

$$\langle u, v \rangle_{H_0^1} \coloneqq \int_\Omega \nabla u \cdot \nabla v + \lambda\, u\, v\, \mathrm{d}x, \tag{4.18}$$

along with associated norm $\|\cdot\|_{H_0^1}$.

The relevance of these choices of duality pairing and inner-products is that the spaces H, $H_0^1(\Omega)$ and $L^2(\Omega)$ form a Gelfand triple as a result of the following integration by parts identity: for any $w \in H_0^1(\Omega)$ and $v \in H$, we have

$$\langle w, v \rangle_{L^2 \times H} = \int_\Omega w\, (-L_\lambda v)\, \mathrm{d}x = \int_\Omega \nabla w \cdot \nabla v + \lambda w v\, \mathrm{d}x = \langle w, v \rangle_{H_0^1}. \tag{4.19}$$

The general theory of Bochner spaces, see for instance [79], yields the following result.

Lemma 4.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain and let $I = (0,T)$. Then,*

$$H \hookrightarrow H_0^1(\Omega) \hookrightarrow L^2(\Omega)$$

form a Gelfand triple [79] *under the inner product $\langle \cdot, \cdot \rangle_{H_0^1}$ and the duality pairing $\langle \cdot, \cdot \rangle_{L^2 \times H}$. The space $H(I;\Omega)$ is continuously embedded in $C(\overline{I}; H_0^1(\Omega))$, and for every u, $v \in H(I;\Omega)$ and any $t \in \overline{I}$, we have*

$$\langle u(t), v(t) \rangle_{H_0^1} = \langle u(0), v(0) \rangle_{H_0^1} + \int_0^t \langle \partial_t u, v \rangle_{L^2 \times H} + \langle \partial_t v, u \rangle_{L^2 \times H} \, \mathrm{d}s. \tag{4.20}$$

Define the norm $\|\cdot\|_{H(I;\Omega)}$ on $H(I;\Omega)$ by

$$\|v\|_{H(I;\Omega)}^2 \coloneqq \int_0^T \omega^2 \|\partial_t v\|_{L^2(\Omega)}^2 + |v|_{H^2(\Omega),\lambda}^2 \, \mathrm{d}t, \quad v \in H(I;\Omega). \tag{4.21}$$

We will make use of the following solvability result for the Cauchy–Dirichlet problem associated to the linear operator L_ω from (4.15).

Theorem 4.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain and let $I = (0,T)$. For each $g \in L^2(I; L^2(\Omega))$ and $v_0 \in H_0^1(\Omega)$, there exists a unique $v \in H(I;\Omega)$ such that*

$$\begin{aligned} L_\omega v &= g \quad \text{a.e. in } \Omega, \text{ for a.e. } t \in I, \\ v(0) &= v_0 \quad \text{in } \Omega. \end{aligned} \tag{4.22}$$

Moreover, the function v satisfies

$$\|v\|_{H(I;\Omega)}^2 + \omega \|v(T)\|_{H_0^1}^2 \leq \|g\|_{L^2(I;L^2(\Omega))}^2 + \omega \|v_0\|_{H_0^1}^2. \tag{4.23}$$

In Theorem 4.3, well-posedness of (4.22) is simply an application of the general theory of Galerkin's method for parabolic equations, see [79]. The bound (4.23) is obtained by combining (4.20), integration by parts and the Miranda–Talenti inequality.

Well-posedness. The following result shows that the methods of analysis used in Chapter 3 generalise to the parabolic setting [72].

Theorem 4.4. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, let $I = (0,T)$, and let Λ be a compact metric space. Let the data a, b, c and f be continuous on $\overline{\Omega} \times \overline{I} \times \Lambda$ and satisfy* (4.8) *and* (4.11)*, or alternatively* (4.10) *in the case where $b \equiv 0$ and $c \equiv 0$. Then, there exists a unique strong solution $u \in H(I;\Omega)$ of the HJB equation* (4.7)*. Moreover, u is also the unique solution of $F_\gamma[u] = 0$ in $\Omega \times I$, $u = 0$ on $\partial\Omega \times I$ and $u = u_0$ on $\Omega \times \{0\}$.*

Proof. Let the operator $\mathcal{A}\colon H(I;\Omega) \to H(I;\Omega)^*$ be defined by

$$\langle \mathcal{A}(u), v\rangle \coloneqq \int_I \int_\Omega F_\gamma[u]\, L_\omega v \,\mathrm{d}x\,\mathrm{d}t + \omega \langle u(0) - u_0, v(0)\rangle_{H_0^1}. \tag{4.24}$$

Compactness of Λ and continuity of the data imply that $\mathcal{A}$ is Lipschitz continuous. Indeed, letting u, v and $z \in H(I;\Omega)$, we find that

$$\begin{aligned} |\langle \mathcal{A}(u) - \mathcal{A}(v), z\rangle| &\le \|F_\gamma[u] - F_\gamma[v]\|_{L^2(I;L^2(\Omega))} \|L_\omega z\|_{L^2(I;L^2(\Omega))} \\ &\quad + \omega \|u(0) - v(0)\|_{H_0^1} \|z(0)\|_{H_0^1} \le C\|u - v\|_{H(I;\Omega)} \|z\|_{H(I;\Omega)}, \end{aligned} \tag{4.25}$$

where the constant C depends only on the dimension d, ω, T, and on the supremum norms of a, b, c and f and γ over $\overline{\Omega} \times \overline{I} \times \Lambda$. We also claim that $\mathcal{A}$ is strongly monotone. Define $w \coloneqq u - v$. Addition and subtraction of $\int_{I_n} \langle L_\omega w, L_\omega w\rangle_{L^2}\,\mathrm{d}t$ shows that

$$\begin{aligned} \langle \mathcal{A}(u) - \mathcal{A}(v), w\rangle &= \|L_\omega w\|^2_{L^2(I;L^2(\Omega))} + \omega \|w(0)\|^2_{H_0^1} \\ &\quad + \int_I \int_\Omega \left(F_\gamma[u] - F_\gamma[v] - L_\omega w\right) L_\omega w \,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{4.26}$$

Lemma 4.1, the bound (4.23) and the Cauchy–Schwarz inequality show that

$$\begin{aligned} \langle \mathcal{A}(u) - \mathcal{A}(v), w\rangle &\ge \frac{1}{2}\|L_\omega w\|^2_{L^2(I;L^2(\Omega))} + \omega \|w(0)\|^2_{H_0^1} - \frac{1-\varepsilon}{2}\|w\|^2_{H(I;\Omega)} \\ &\ge \frac{\varepsilon}{2}\|w\|^2_{H(I;\Omega)} + \frac{\omega}{2}\|w(T)\|^2_{H_0^1} + \frac{\omega}{2}\|w(0)\|^2_{H_0^1}. \end{aligned} \tag{4.27}$$

The inequalities (4.25) and (4.27) imply that $\mathcal{A}$ is a bounded, continuous, coercive and strongly monotone operator, so the Browder–Minty theorem [65] shows that there exists a unique $u \in H(I;\Omega)$ such that $\mathcal{A}(u) = 0$.

Theorem 4.3 shows that for each $g \in L^2(I;L^2(\Omega))$, there exists a $v \in H(I;\Omega)$ such that $L_\omega v = g$ and $v(0) = 0$. So, $\mathcal{A}(u) = 0$ implies that $\int_I \int_\Omega F_\gamma[u]\, g \,\mathrm{d}x\,\mathrm{d}t = 0$ for all $g \in L^2(I;L^2(\Omega))$, and since $F_\gamma[u] \in L^2(I;L^2(\Omega))$, we obtain $F_\gamma[u] = 0$. Theorem 4.3 also shows that $\langle u(0), v\rangle_{H_0^1} = \langle u_0, v\rangle_{H_0^1}$ for all $v \in H_0^1(\Omega)$, hence $u(0) = u_0$.

We claim that $u \in H(I;\Omega)$ solves $F_\gamma[u] = 0$, $u(0) = u_0$, if and only if u solves (4.7). Since γ^α is positive, $\gamma^\alpha(\partial_t u - L^\alpha u + f^\alpha) \ge 0$ for all $\alpha \in \Lambda$ is equivalent to $\partial_t u - L^\alpha u + f^\alpha \ge 0$ for all $\alpha \in \Lambda$, so $F_\gamma[u] \ge 0$ is equivalent to $F[u] \ge 0$. Compactness of Λ and continuity of the data imply that for a.e. $t \in I$, for a.e. point of Ω, the extrema in the definitions of $F_\gamma[u]$ and $F[u]$ are attained by some elements of Λ, thereby giving $F_\gamma[u] \le 0$ if and only if $F[u] \le 0$. Therefore, existence and uniqueness in $H(I;\Omega)$ of a solution of $F_\gamma[u] = 0$ is equivalent to existence and uniqueness of a solution of (4.7). □

4.2 Temporal semi-discretisation

In this section, we explore some of the general principles underlying the numerical scheme for the parabolic problem (4.7). Before presenting the fully discrete scheme in section 4.3, we briefly consider in this section the temporal semi-discretisation of parabolic HJB equations, so as to highlight some key ideas in the derivation and analysis of a stable method.

Let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of partitions of $(0,T)$ into half-intervals $I_n := (t_{n-1}, t_n] \in \mathcal{J}_\tau$, with $1 \le n \le N = N(\tau)$. We say that $\mathcal{J}_\tau$ is *regular* provided that

$$[0,T] = \bigcup_{I_n \in \mathcal{J}_\tau} \overline{I_n}, \quad 0 = t_0 \le t_{n-1} < t_n \le t_N = T, \quad \forall\, n \le N,\ \forall\, \tau. \tag{4.28}$$

For each interval $I_n \in \mathcal{J}_\tau$, let $\tau_n := |t_n - t_{n-1}|$. It is assumed that $\tau = \max_{1\le n\le N} \tau_n$. For each τ, let $\mathbf{q} = (q_1, \dots, q_N)$ be a vector of positive integers, so $q_n \ge 1$ for all $I_n \in \mathcal{J}_\tau$. For a vector space V and $I_n \in \mathcal{J}_\tau$, let $\mathcal{Q}_{q_n}(V)$ denote the space of V-valued univariate polynomials of degree at most q_n. Recalling that $H := H^2(\Omega) \cap H_0^1(\Omega)$, we define the semidiscrete DG finite element space $V^{\tau,\mathbf{q}}$ by

$$V^{\tau,\mathbf{q}} := \left\{ v \in L^2(I;H) : v|_{I_n} \in \mathcal{Q}_{q_n}(H) \ \ \forall\, I_n \in \mathcal{J}_\tau \right\}. \tag{4.29}$$

Functions from $V^{\tau,\mathbf{q}}$ are taken to be left-continuous, but are generally discontinuous at the partition points $\{t_n\}_{n=1}^{N-1}$. We denote the right-limit of $v \in V^{\tau,\mathbf{q}}$ at t_n by $v(t_n^+)$, where $0 \le n < N$. The jump operators $\llparenthesis \cdot \rrparenthesis_n$ and average operators $\langle \cdot \rangle_n$, $0 \le n \le N$, are defined by

$$\begin{aligned}
&\llparenthesis v \rrparenthesis_n := -v(0^+), && \langle v \rangle_n := v(0^+), && \text{if } n = 0,\\
&\llparenthesis v \rrparenthesis_n := v(t_n) - v(t_n^+), && \langle v \rangle_n := \tfrac{1}{2} v(t_n) + \tfrac{1}{2} v(t_n^+), && \text{if } 1 \le n < N,\\
&\llparenthesis v \rrparenthesis_n := v(T), && \langle v \rangle_n := v(T), && \text{if } n = N.
\end{aligned} \tag{4.30}$$

Define the nonlinear form $A_\tau \colon V^{\tau,\mathbf{q}} \times V^{\tau,\mathbf{q}} \to \mathbb{R}$ by

$$\begin{aligned}
A_\tau(u_\tau; v_\tau) := &\sum_{n=1}^{N} \int_{I_n} \langle F_\gamma[u_\tau], L_\omega v_\tau \rangle_{L^2(\Omega)} \,\mathrm{d}t \\
&- \omega \sum_{n=0}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \langle v_\tau \rangle_n \rangle_{H_0^1} + \frac{\omega}{2} \sum_{n=1}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \llparenthesis v_\tau \rrparenthesis_n \rangle_{H_0^1}.
\end{aligned} \tag{4.31}$$

We note that $\frac{1}{2}\llparenthesis v \rrparenthesis_n - \langle v \rangle_n = v(t_n^+)$ for $1 \le n < N$. The semidiscrete scheme consists of finding $u_\tau \in V^{\tau,\mathbf{q}}$ such that

$$A_\tau(u_\tau; v_\tau) = \omega \langle u_0, v_\tau(0^+) \rangle_{H_0^1} \qquad \forall\, v_\tau \in V^{\tau,\mathbf{q}}. \tag{4.32}$$

Since the solution $u \in H(I;\Omega)$ of (4.7) belongs to $C(\overline{I}; H_0^1(\Omega))$, it is clear that $A_\tau(u; v_\tau) = \omega \langle u_0, v_\tau(0^+) \rangle_{H_0^1}$ for all $v_\tau \in V^{\tau,\mathbf{q}}$, so the scheme is consistent.

By considering test functions v_τ that have support on successive intervals $\overline{I_n} \in \mathcal{J}_\tau$, it is easily seen that $u_\tau|_{I_n}$ is determined only by the data and by $u(t_{n-1})$, thus (4.32) is a time-stepping scheme.

The main ingredients required to show that the above scheme is stable are as follows. We introduce the bilinear form $C_\tau \colon V^{\tau,\mathbf{q}} \times V^{\tau,\mathbf{q}} \to \mathbb{R}$ defined by

$$C_\tau(u_\tau, v_\tau) \coloneqq \sum_{n=1}^{N} \int_{I_n} \langle L_\omega u_\tau, L_\omega v_\tau \rangle_{L^2(\Omega)} \,\mathrm{d}t - \omega \sum_{n=0}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \langle v_\tau \rangle_n \rangle_{H^1_0} + \frac{\omega}{2} \sum_{n=1}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \llparenthesis v_\tau \rrparenthesis_n \rangle_{H^1_0}. \tag{4.33}$$

Integration by parts shows that for any u_τ, $v_\tau \in V^{\tau,\mathbf{q}}$, we have

$$C_\tau(u_\tau, v_\tau) = \sum_{n=1}^{N} \int_{I_n} \omega^2 \langle \partial_t u_\tau, \partial_t v_\tau \rangle_{L^2(\Omega)} + \langle L_\lambda u_\tau, L_\lambda v_\tau \rangle_{L^2(\Omega)} \,\mathrm{d}t + \omega \sum_{n=1}^{N} \langle \langle u_\tau \rangle_n , \llparenthesis v_\tau \rrparenthesis_n \rangle_{H^1_0} + \frac{\omega}{2} \sum_{n=1}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \llparenthesis v_\tau \rrparenthesis_n \rangle_{H^1_0}. \tag{4.34}$$

Combining (4.33) and (4.34) reveals the stability properties of C_τ when re-written as

$$C_\tau(u_\tau, v_\tau) = \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} \omega^2 \langle \partial_t u_\tau, \partial_t v_\tau \rangle_{L^2} + \langle L_\lambda u_\tau, L_\lambda v_\tau \rangle_{L^2} + \langle L_\omega u_\tau, L_\omega v_\tau \rangle_{L^2} \,\mathrm{d}t + \frac{\omega}{2} \sum_{n=1}^{N} \langle \langle u_\tau \rangle_n , \llparenthesis v_\tau \rrparenthesis_n \rangle_{H^1_0} - \frac{\omega}{2} \sum_{n=0}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \langle v_\tau \rangle_n \rangle_{H^1_0} + \frac{\omega}{2} \sum_{n=1}^{N-1} \langle \llparenthesis u_\tau \rrparenthesis_n, \llparenthesis v_\tau \rrparenthesis_n \rangle_{H^1_0}. \tag{4.35}$$

Indeed, it follows from (4.35) and the Miranda–Talenti inequality that, for any $u_\tau \in V^{\tau,\mathbf{q}}$,

$$C_\tau(u_\tau, u_\tau) \geq \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} \omega^2 \|\partial_t u\|^2_{L^2(\Omega)} + |u_\tau|^2_{H^2(\Omega),\lambda} + \|L_\omega u_\tau\|^2_{L^2(\Omega)} \mathrm{d}t + \frac{\omega}{2} \|u_\tau(T)\|^2_{H^1_0} + \frac{\omega}{2} \|u_\tau(0^+)\|^2_{H^1_0} + \frac{\omega}{2} \sum_{n=1}^{N-1} \|\llparenthesis u_\tau \rrparenthesis_n\|^2_{H^1_0}. \tag{4.36}$$

The key observation here is that the antisymmetric terms in (4.35) cancel in $C_\tau(u_\tau, u_\tau)$, and this technique will be used again in section 4.5 for the analysis of stability of the fully discrete scheme. Stability of the scheme is then obtained as follows: (4.33) implies that

$$A_\tau(u_\tau; v_\tau) = \sum_{n=1}^{N} \int_{I_n} \langle F_\gamma[u_\tau] - L_\omega u_\tau, L_\omega v_\tau \rangle_{L^2(\Omega)} \,\mathrm{d}t + C_\tau(u_\tau, v_\tau) \quad \forall\, u_\tau,\, v_\tau \in V^{\tau,\mathbf{q}};$$

which mirrors the addition-subtraction step of the proof of Theorem 4.4.

Then, we use (4.36) to show that A_τ is strongly monotone: for any u_τ, $v_\tau \in V^{\tau,\mathbf{q}}$,

$w_\tau \coloneqq u_\tau - v_\tau$, we have

$$A_\tau(u_\tau; w_\tau) - A_\tau(v_\tau; w_\tau) \geq \frac{\varepsilon}{2} \sum_{n=1}^{N} \int_{I_n} \omega^2 \|\partial_t w_\tau\|^2_{L^2(\Omega)} + |w_\tau|^2_{H^2(\Omega),\lambda} \mathrm{d}t + \frac{\omega}{2} \sum_{n=0}^{N} \|(\!|w_\tau|\!)_n\|^2_{H^1_0}.$$

Therefore, the well-posedness of the semidiscrete scheme can be shown by an induction argument, based on the Browder–Minty Theorem, that is similar to the one given in the proof of Theorem 4.8 below, concerning the well-posedness of the fully discrete scheme. Instead of pursuing the analysis of the semidiscrete scheme further, we now turn towards the fully discrete method.

4.2.1 Comparison with the standard DG time-stepping method

The time semi-discretisation presented above introduces a DG time-stepping method that is different to the standard method. For example, the standard method [2, 74] for solving the *linear* problem (4.22) consists of finding $v_\tau \in \tilde{V}^{\tau,\mathbf{q}}$ such that

$$\begin{aligned} (4.37) \quad & \sum_{n=1}^{N} \int_{I_n} \left[\omega \langle \partial_t v_\tau, w_\tau \rangle_{L^2} + \langle v_\tau, w_\tau \rangle_{H^1_0} \right] \mathrm{d}t - \omega \sum_{n=0}^{N-1} \langle (\!|v_\tau|\!)_n, \langle w_\tau \rangle_n \rangle_{L^2} \\ & + \frac{\omega}{2} \sum_{n=1}^{N-1} \langle (\!|v_\tau|\!)_n, (\!|w_\tau|\!)_n \rangle_{L^2} = \sum_{n=1}^{N} \int_{I_n} \langle g, w_\tau \rangle_{L^2} \, \mathrm{d}t + \omega \langle v_0, w_\tau(0^+) \rangle_{L^2} \qquad \forall\, w \in \tilde{V}^{\tau,\mathbf{q}}, \end{aligned}$$

where $\tilde{V}^{\tau,\mathbf{q}} \coloneqq \{v \in L^2(I; H^1_0(\Omega)) : v|_{I_n} \in \mathcal{Q}_{q_n}(H^1_0(\Omega)) \,\forall\, I_n \in \mathcal{J}_\tau\}$. Notice that the function spaces $V^{\tau,\mathbf{q}}$ and $\tilde{V}^{\tau,\mathbf{q}}$ differ substantially.

Thus, in the linear case, the standard method (4.37) relies on the weak form of the operator L_λ in the PDE, and is based on the Gelfand triple $H^1_0(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$. One of the reasons for considering an alternative time discretisation method is that the fully nonlinear HJB equation does not admit a weak formulation, and cannot be formulated in terms of this Gelfand triple. This helps to explain the unusual choice of Gelfand triple $H \hookrightarrow H^1_0(\Omega) \hookrightarrow L^2(\Omega)$ used in the formulation of our scheme (4.32).

One of the main structural differences between the two DG time-stepping methods is that our scheme (4.32) will be shown to be both strongly monotone and Lipschitz continuous in the same Bochner norm, whereas it is well-known that the standard scheme (4.37) fails to yield coercivity and continuity in the same norm as a result of the absence of terms explicitly controlling the $L^2(H^{-1})$-norm of the time derivative [74]. As shown in the following analysis, this additional property of the scheme (4.32) facilitates the analysis of stability and convergence rates, especially with regards to the results of section 4.6.2. Both schemes remain comparable in terms of computational cost for those problems to which both are applicable.

4.3 Numerical scheme

Function spaces. In addition to the spatial finite element space $V_{h,\mathbf{p}}$ defined in section 2.2, we define the space-time discontinuous Galerkin finite element space $V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ over a regular partition $\mathcal{J}_\tau$ by

$$V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \coloneqq \left\{ v \in L^2\left(I; V_{h,\mathbf{p}}\right) : v|_{I_n} \in \mathcal{Q}_{q_n}\left(V_{h,\mathbf{p}}\right) \ \ \forall\, I_n \in \mathcal{J}_\tau \right\}. \tag{4.38}$$

As in section 4.2, we consider a function $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ to be left-continuous. The support of v_h, denoted by $\operatorname{supp} v_h$, is a subset of $\overline{I}$, and is understood to be the support of $v_h \colon I \to V_{h,\mathbf{p}}$, i.e. when viewing v_h as a mapping from I into $V_{h,\mathbf{p}}$.

Bilinear and nonlinear forms. In addition to the bilinear forms $B_{h,\theta} \colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ defined in Chapter 3, we define the bilinear form $a_h \colon V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \to \mathbb{R}$ by

$$\begin{aligned} a_h(u_h, v_h) \coloneqq \sum_{K \in \mathcal{T}_h} \langle \nabla u_h, \nabla v_h \rangle_K + \lambda \langle u_h, v_h \rangle_K - \sum_{F \in \mathcal{F}_h^{i,b}} \langle \{\nabla u_h \cdot n_F\}, [\![v_h]\!] \rangle_F \\ - \sum_{F \in \mathcal{F}_h^{i,b}} \langle \{\nabla v_h \cdot n_F\}, [\![u_h]\!] \rangle_F + \sum_{F \in \mathcal{F}_h^{i,b}} \mu_F \langle [\![u_h]\!], [\![v_h]\!] \rangle_F. \end{aligned} \tag{4.39}$$

Observe that the bilinear form a_h corresponds precisely to the standard symmetric interior penalty discretisation of the operator $-L_\lambda$, and its symmetry plays an important role in the subsequent analysis.

Define the bilinear forms $C_h^{\mathcal{F}}$ and $C_h \colon V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \times V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \to \mathbb{R}$ by

$$\begin{aligned} C_h^{\mathcal{F}}(u_h, v_h) \coloneqq \omega \sum_{n=1}^{N} \int_{I_n} \sum_{F \in \mathcal{F}_h^i} \langle [\![\nabla u_h \cdot n_F]\!], \{\partial_t v_h\} \rangle_F \, \mathrm{d}t \\ + \omega \sum_{n=1}^{N} \int_{I_n} \sum_{F \in \mathcal{F}_h^{i,b}} \left[\mu_F \langle [\![u_h]\!], [\![\partial_t v_h]\!] \rangle_F - \langle [\![u_h]\!], \{\nabla \partial_t v_h \cdot n_F\} \rangle_F \right] \mathrm{d}t, \end{aligned} \tag{4.40}$$

$$\begin{aligned} C_h(u_h, v_h) \coloneqq \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h} \langle L_\omega u_h, L_\omega v_h \rangle_K \, \mathrm{d}t + C_h^{\mathcal{F}}(u_h, v_h) \\ + \sum_{n=1}^{N} \int_{I_n} B_{h,1/2}(u_h, v_h) - \sum_{K \in \mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h \rangle_K \, \mathrm{d}t \\ - \omega \sum_{n=0}^{N-1} a_h(\langle\!| u_h |\!\rangle_n, \langle v_h \rangle_n) + \frac{\omega}{2} \sum_{n=1}^{N-1} a_h(\langle\!| u_h |\!\rangle_n, \langle\!| v_h |\!\rangle_n). \end{aligned} \tag{4.41}$$

Define the nonlinear form $A_h\colon V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \times V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \to \mathbb{R}$ by

$$A_h(u_h; v_h) \coloneqq \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} [\langle F_\gamma[u_h], L_\omega v_h\rangle_K - \langle L_\omega u_h, L_\omega v_h\rangle_K]\,\mathrm{d}t + C_h(u_h, v_h). \tag{4.42}$$

The form A_h is linear in its second argument, but it is nonlinear in its first argument. Supposing that u_0 is sufficiently regular, such as $u_0 \in H^s(\Omega;\mathcal{T}_h)$, with $s > 3/2$, the numerical scheme is to find $u_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ such that

$$A_h(u_h; v_h) = \omega\, a_h(u_0, v_h(0^+)) \qquad \forall\, v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}. \tag{4.43}$$

If u_0 fails to be sufficiently regular, then u_0 can be replaced in the right-hand side of (4.43) by a suitable projection into $V_{h,\mathbf{p}}$, at the expense of introducing a consistency error that vanishes in the limit. By testing with functions $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ that are supported on $\overline{I_n}$ and by adopting the convention that $u_h(t_0) \coloneqq u_0$, it is found that (4.43) is equivalent to

$$\begin{aligned}\int_{I_n} \sum_{K\in\mathcal{T}_h} \langle F_\gamma[u_h], L_\omega v_h\rangle_K + B_{h,1/2}(u_h, v_h) - \sum_{K\in\mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h\rangle_K\,\mathrm{d}t \\ + \omega \int_{I_n} \sum_{F\in\mathcal{F}_h^i} \langle [\![\nabla u_h\cdot n_F]\!], \{\partial_t v_h\}\rangle_F + \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \langle [\![u_h]\!], [\![\partial_t v_h]\!]\rangle_F\,\mathrm{d}t \\ - \omega \int_{I_n} \sum_{F\in\mathcal{F}_h^{i,b}} \langle [\![u_h]\!], \{\nabla\partial_t v_h\cdot n_F\}\rangle_F\,\mathrm{d}t + \omega\, a_h(u_h(t_{n-1}^+), v_h(t_{n-1}^+)) \\ = \omega\, a_h(u_h(t_{n-1}), v_h(t_{n-1}^+)) \quad \forall\, v_h \in \mathcal{Q}_{q_n}(V_{h,\mathbf{p}}).\end{aligned} \tag{4.44}$$

Therefore, (4.43) defines a time-stepping scheme, and it is (4.44) that is solved in practice.

4.4 Consistency

Lemma 4.5. *Let Ω be a bounded Lipschitz polytopal domain, let $\mathcal{T}_h$ be a simplicial or parallelepipedal mesh on Ω. Let $I = (0,T)$ and let $\mathcal{J}_\tau = \{I_n\}_{n=1}^N$ be a regular partition of I. Suppose that $u_0 \in H_0^1(\Omega) \cap H^r(\Omega;\mathcal{T}_h)$ with $r > 3/2$. Then, for any $w \in H(I;\Omega) \cap L^2(I; H^s(\Omega;\mathcal{T}_h))$, with $s > 5/2$, such that $w(0) = u_0$, we have*

$$C_h(w, v_h) = \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle L_\omega w, L_\omega v_h\rangle_K\,\mathrm{d}t + \omega\, a_h(u_0, v_h(0^+)) \qquad \forall\, v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}. \tag{4.45}$$

Proof. Let the function w be as above, so that $w(t) \in H^2(\Omega) \cap H_0^1(\Omega) \cap H^s(\Omega;\mathcal{T}_h)$ for a.e. $t \in I$. Lemma 3.5 shows that

$$\int_{I_n} B_{h,1/2}(w, v_h)\,\mathrm{d}t = \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle L_\lambda w, L_\lambda v_h\rangle_K\,\mathrm{d}t \quad \forall\, I_n \in \mathcal{J}_\tau,\ \forall\, v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}.$$

The spatial regularity of w also implies that $[\![\nabla w(t) \cdot n_F]\!]$ vanishes for all $F \in \mathcal{F}_h^i$ and a.e. $t \in I$, whilst $[\![w(t)]\!]$ and $[\![\nabla_{\mathrm{T}} w(t)]\!]$ vanish for all $F \in \mathcal{F}_h^{i,b}$ and a.e. $t \in I$. Therefore we have $C_h^{\mathcal{F}}(w, v) = 0$ for all $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$. Finally, since $H(I;\Omega) \hookrightarrow C(\bar{I}; H_0^1(\Omega))$ by Lemma 4.2, the jump $(\!|w|\!)_n = 0$ for each $0 < n < N$, and thus $a_h((\!|w|\!)_n, v_h) = 0$ for all $v_h \in V_{h,\mathbf{p}}$, $0 < n < N$. The above identities and the definition of C_h in (4.41) imply (4.45). □

Lemma 4.5 and the definition of the nonlinear form A_h in (4.42) immediately imply the following consistency result for the numerical scheme.

Corollary 4.6. *Under the hypotheses of Lemma* 4.5*, suppose that the solution* $u \in H(I;\Omega)$ *of* (4.7) *belongs to* $L^2(I; H^s(\Omega; \mathcal{T}_h))$*, with* $s > 5/2$*. Then,* u *satisfies*

$$A_h(u; v_h) = \omega\, a_h(u_0, v_h(0^+)) \quad \forall\, v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}. \tag{4.46}$$

4.5 Stability

The quantities μ_F and η_F may be chosen as in Lemma 3.7 whilst also guaranteeing the standard discrete Poincaré inequality for the bilinear form a_h:

$$\sum_{K\in\mathcal{T}_h} \|v_h\|_{H^1(K)}^2 + \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \|[\![v_h]\!]\|_{L^2(F)}^2 \lesssim a_h(v_h, v_h) \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{4.47}$$

This implies that the symmetric bilinear form a_h is coercive on $V_{h,\mathbf{p}}$, and thus defines an inner-product on $V_{h,\mathbf{p}}$, with an associated norm defined by $\|v_h\|_{a_h}^2 \coloneqq a_h(v_h, v_h)$ for $v_h \in V_{h,\mathbf{p}}$.

In the subsequent analysis, we shall choose μ_F and η_F to be given by

$$\mu_F \coloneqq c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad \eta_F \coloneqq \max(1,\lambda)\, c_\eta \frac{\tilde{p}_F^6}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}, \tag{4.48}$$

where c_μ and c_η are constants chosen so that both (4.47) and Lemma 3.7 hold, for some $\kappa < (1-\varepsilon)^{-1}$. Note that these orders of penalisation are the strongest that remain consistent with the discrete H^2-type norm appearing in the analysis of this work; see [60] for an example of a scheme for the biharmonic equation using the same penalisation orders. These penalisation orders are stronger than those required in the analysis of Chapters 2 and 3; the reason for this choice will become apparent in this section.

For each $\theta \in [0,1]$, we introduce the functional $\|\cdot\|_{h,\theta} \colon V_{h,\mathbf{p}}^{\tau,\mathbf{q}} \to \mathbb{R}$ defined by

$$\begin{aligned}\|v_h\|_{h,\theta}^2 \coloneqq \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \theta \left[\omega^2 \|\partial_t v_h\|_{L^2(K)}^2 + |v_h|_{H^2(K),\lambda}^2\right] + |v_h|_{\mathrm{J}}^2 \,\mathrm{d}t \\ + \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} (1-\theta)\, \|L_\omega v_h\|_{L^2(K)}^2 \mathrm{d}t + \omega \sum_{n=0}^{N} \|(\!|v_h|\!)_n\|_{a_h}^2.\end{aligned} \tag{4.49}$$

To verify that the functional $\|\cdot\|_{h,\theta}$ defines a norm on $V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, suppose that $\|v_h\|_{h,\theta} = 0$ for some $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$. Then, the jumps of v_h vanish across the mesh faces and across time intervals and, therefore, $v_h \in H(I;\Omega)$ with $v_h(0) = 0$. The fact that the volume terms in $\|v_h\|_{h,\theta}$ also vanish shows that $L_\omega v_h = 0$, so it follows from (4.23) that $v_h \equiv 0$. Hence, the functional $\|\cdot\|_{h,\theta}$ defines a norm on $V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$.

Recall the definition of the $|\cdot|_{H^2(K),\lambda}$ for $K \in \mathcal{T}_h$ from (3.30).

Lemma 4.7. *Under the hypotheses of Lemma* 3.7*, let $I = (0,T)$ and $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of I. Let μ_F and η_F satisfy* (4.48) *for each face F, chosen so that Lemma* 3.7 *holds for a given $\kappa > 1$. Then, for every $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, we have*

$$C_h(v_h, v_h) \geq \frac{1}{2}\sum_{n=1}^{N}\int_{I_n}\sum_{K\in\mathcal{T}_h}\omega^2\|\partial_t v_h\|_{L^2(K)}^2 + \frac{1}{\kappa}|v_h|_{H^2(K),\lambda}^2 + |v_h|_{\mathrm{J}}^2\,\mathrm{d}t + \frac{1}{2}\sum_{n=1}^{N}\int_{I_n}\sum_{K\in\mathcal{T}_h}\|L_\omega v_h\|_{L^2(K)}^2\mathrm{d}t + \frac{\omega}{2}\sum_{n=0}^{N}\|\langle\!|v_h|\!\rangle_n\|_{a_h}^2. \tag{4.50}$$

Proof. We begin by showing that, for any u_h, $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, the bilinear form C_h satisfies the following identity:

$$C_h(u_h, v_h) = \sum_{n=1}^{N}\int_{I_n}\sum_{K\in\mathcal{T}_h}\omega^2\langle\partial_t u_h, \partial_t v_h\rangle_K + B_{h,1/2}(u_h, v_h)\,\mathrm{d}t - C_h^{\mathcal{F}}(v_h, u_h) + \omega\sum_{n=1}^{N} a_h(\langle u_h\rangle_n, \langle\!|v_h|\!\rangle_n) + \frac{\omega}{2}\sum_{n=1}^{N-1} a_h(\langle\!|u_h|\!\rangle_n, \langle\!|v_h|\!\rangle_n). \tag{4.51}$$

The first step in deriving (4.51) is to show that for any u_h, $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, we have

$$\sum_{n=1}^{N}\int_{I_n}\sum_{K\in\mathcal{T}_h}\langle\omega\,\partial_t u_h, -L_\lambda v_h\rangle_K + \langle\omega\,\partial_t v_h, -L_\lambda u_h\rangle_K\,\mathrm{d}t = \omega\sum_{n=1}^{N} a_h(\langle u_h\rangle_n, \langle\!|v_h|\!\rangle_n) + \omega\sum_{n=0}^{N-1} a_h(\langle\!|u_h|\!\rangle_n, \langle v_h\rangle_n) - C_h^{\mathcal{F}}(u_h, v_h) - C_h^{\mathcal{F}}(v_h, u_h). \tag{4.52}$$

Indeed, integration by parts over $\mathcal{T}_h$ shows that, for any $I_n \in \mathcal{J}_\tau$ and a.e. $t \in I_n$,

$$\sum_{K\in\mathcal{T}_h}\langle\omega\,\partial_t u_h, -L_\lambda v_h\rangle_K = \omega\sum_{K\in\mathcal{T}_h}\langle\nabla\partial_t u_h, \nabla v_h\rangle_K + \lambda\langle\partial_t u_h, v_h\rangle_K - \omega\sum_{F\in\mathcal{F}_h^i}\langle\{\partial_t u_h\}, [\![\nabla v_h\cdot n_F]\!]\rangle_F - \omega\sum_{F\in\mathcal{F}_h^{i,b}}\langle[\![\partial_t u_h]\!], \{\nabla v_h\cdot n_F\}\rangle_F. \tag{4.53}$$

Therefore, it is found that, for any $I_n \in \mathcal{J}_\tau$ and a.e. $t \in I_n$,

$$
\begin{aligned}
(4.54)\quad & \sum_{K\in\mathcal{T}_h} \langle \omega\, \partial_t u_h, -L_\lambda v_h\rangle_K + \langle \omega\, \partial_t v_h, -L_\lambda u_h\rangle_K \\
&\quad = \omega \frac{\mathrm{d}}{\mathrm{d}t}\, a_h(u_h, v_h) - \omega \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \left[\langle [\![\partial_t u_h]\!], [\![v_h]\!]\rangle_F + \langle [\![u_h]\!], [\![\partial_t v_h]\!]\rangle_F\right] \\
&\qquad - \omega \sum_{F\in\mathcal{F}_h^{i}} \left[\langle \{\partial_t u_h\}, [\![\nabla v_h \cdot n_F]\!]\rangle_F + \langle \{\partial_t v_h\}, [\![\nabla u_h \cdot n_F]\!]\rangle_F\right] \\
&\qquad\qquad + \omega \sum_{F\in\mathcal{F}_h^{i,b}} \left[\langle [\![v_h]\!], \{\nabla \partial_t u_h \cdot n_F\}\rangle_F + \langle [\![u_h]\!], \{\nabla \partial_t v_h \cdot n_F\}\rangle_F\right].
\end{aligned}
$$

We obtain (4.52) upon integration and summation of (4.54) over all time intervals.

The second step towards (4.51) is to use (4.52) to find that

$$
\begin{aligned}
\sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle L_\omega u_h, L_\omega v_h\rangle_K \mathrm{d}t &= \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \omega^2 \langle \partial_t u_h, \partial_t v_h\rangle_K + \langle L_\lambda u_h, L_\lambda v_h\rangle_K \mathrm{d}t \\
&\quad + \omega \sum_{n=1}^{N} a_h(\langle u_h\rangle_n, (\!|v_h|\!)_n) + \omega \sum_{n=0}^{N-1} a_h((\!|u_h|\!)_n, \langle v_h\rangle_n) - C_h^{\mathcal{F}}(u_h, v_h) - C_h^{\mathcal{F}}(v_h, u_h).
\end{aligned}
$$

The proof of (4.51) is then completed by substituting the above identity in the definition of C_h from (4.41).

Expanding C_h with both (4.41) and (4.51) shows that

$$
\begin{aligned}
(4.55)\quad C_h(u_h, v_h) &= \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \omega^2 \langle \partial_t u_h, \partial_t v_h\rangle_K + B_{h,1}(u_h, v_h) + J_h(u_h, v_h)\, \mathrm{d}t \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle L_\omega u_h, L_\omega v_h\rangle_K \, \mathrm{d}t + \frac{1}{2}\, C_h^{\mathcal{F}}(u_h, v_h) - \frac{1}{2}\, C_h^{\mathcal{F}}(v_h, u_h) \\
&\quad + \frac{\omega}{2} \sum_{n=1}^{N} a_h(\langle u_h\rangle_n, (\!|v_h|\!)_n) - \frac{\omega}{2} \sum_{n=0}^{N-1} a_h((\!|u_h|\!)_n, \langle v_h\rangle_n) + \frac{\omega}{2} \sum_{n=1}^{N-1} a_h((\!|u_h|\!)_n, (\!|v_h|\!)_n).
\end{aligned}
$$

Note that to get (4.55), we have used the identity

$$
B_{h,1/2}(u_h, v_h) - \frac{1}{2} \sum_{K\in\mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h\rangle_K = \frac{1}{2} B_{h,1}(u_h, v_h) + \frac{1}{2} J_h(u_h, v_h).
$$

To show (4.50), we substitute $u_h = v_h$ in (4.55) and first observe that the flux terms

involving $C_h^{\mathcal{F}}$ cancel. Furthermore, the symmetry of the bilinear form a_h implies that

$$\sum_{n=1}^{N} a_h(\langle v_h\rangle_n, \langle\!\langle v_h\rangle\!\rangle_n) - \sum_{n=0}^{N-1} a_h(\langle\!\langle v_h\rangle\!\rangle_n, \langle v_h\rangle_n) + \sum_{n=1}^{N-1} \|\langle\!\langle v_h\rangle\!\rangle_n\|_{a_h}^2$$
$$= a_h(v_h(T), v_h(T)) + a_h(v_h(0^+), v_h(0^+)) + \sum_{n=1}^{N-1} \|\langle\!\langle v_h\rangle\!\rangle_n\|_{a_h}^2 = \sum_{n=0}^{N} \|\langle\!\langle v_h\rangle\!\rangle_n\|_{a_h}^2.$$

Then, we apply Lemma 3.7 for $\theta = 1$ to get $B_{h,1}(v_h, v_h) \geq \kappa^{-1} \sum_{K\in\mathcal{T}_h} |v_h|_{H^2(K),\lambda}^2$, thereby yielding (4.50). □

Recall that for a function $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, the support of v_h is a subset of $\overline{I}$, since v_h is viewed as a mapping from I into $V_{h,\mathbf{p}}$.

Theorem 4.8. *Let Ω be a bounded convex polytopal domain and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of meshes satisfying (2.11). Let $I = (0,T)$ and let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of I. Let Λ be a compact metric space and let the data a, b, c and f be continuous on $\overline{\Omega}\times\overline{I}\times\Lambda$ and satisfy (4.8) and (4.11), or alternatively (4.10) in the case where $b \equiv 0$ and $c \equiv 0$. Assume that the initial data $u_0 \in H_0^1(\Omega) \cap H^s(\Omega;\mathcal{T}_h)$ with $s > 3/2$. Let μ_F and η_F satisfy (4.48), with c_μ and c_η chosen so that Lemmas 3.7 and 4.7 hold with $\kappa < (1-\varepsilon)^{-1}$. Then, for every z_h, $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, we have*

$$\|z_h - v_h\|_{h,1}^2 \leq \frac{2\,\kappa}{1-\kappa\,(1-\varepsilon)} \left(A_h(z_h; z_h - v_h) - A_h(v_h; z_h - v_h)\right). \tag{4.56}$$

Moreover, A_h is interval-wise Lipschitz continuous, in the sense that, for any $I_n \in \mathcal{J}_\tau$, any u_h, v_h and $z_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ with support contained in $\overline{I_n}$, we have

$$|A_h(u_h; z_h) - A_h(v_h; z_h)| \lesssim \|u_h - v_h\|_{h,1}\|z_h\|_{h,1}. \tag{4.57}$$

Therefore, there exists a unique solution $u_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ of the numerical scheme (4.43).

Proof. We begin by showing strong monotonicity of the nonlinear form A_h. Let z_h, $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ and set $w_h \coloneqq z_h - v_h$. Then, by (4.42) and Lemma 4.7, we have

$$A_h(z_h; w_h) - A_h(v_h; w_h) = C_h(w_h, w_h) + \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle F_\gamma[z_h] - F_\gamma[v_h] - L_\omega w_h, L_\omega w_h\rangle_K \,\mathrm{d}t.$$

Lemma 4.1 and Young's inequality show that

$$\sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} |\langle F_\gamma[z_h] - F_\gamma[v_h] - L_\omega w_h, L_\omega w_h\rangle_K| \mathrm{d}t \leq \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \|L_\omega w_h\|_{L^2(K)}^2 \mathrm{d}t$$
$$+ \frac{1-\varepsilon}{2} \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \omega^2 \|\partial_t w_h\|_{L^2(K)}^2 + |w_h|_{H^2(K),\lambda}^2 \mathrm{d}t.$$

Since $1 < \kappa < (1-\varepsilon)^{-1}$, Lemma 4.7 implies that

$$A_h(z_h; w_h) - A_h(v_h; w_h) \geq \frac{1}{C} \sum_{n=1}^{N} \int_{I_n} \sum_{K \in \mathcal{T}_h} \omega^2 \|\partial_t w_h\|_{L^2(K)}^2 + |w_h|_{H^2(K),\lambda}^2 \mathrm{d}t + \frac{1}{2} \sum_{n=1}^{N} \int_{I_n} |w_h|_{\mathrm{J}}^2 \,\mathrm{d}t + \frac{\omega}{2} \sum_{n=0}^{N} \|(\!|w_h|\!)_n\|_{a_h}^2, \tag{4.58}$$

where $C = 2\,\kappa/(1-\kappa\,(1-\varepsilon)) \geq 2$, thus showing (4.56).

To show (4.57), consider u_h, v_h and $z_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ that all have support in $\overline{I_n}$, and set $w_h \coloneqq u_h - v_h$. It then follows from $\operatorname{supp} v_h \subset \overline{I_n}$ that

$$\|v_h\|_{h,1}^2 = \int_{I_n} \sum_{K \in \mathcal{T}_h} \left[\omega^2 \|\partial_t v_h\|_{L^2(K)}^2 + |v_h|_{H^2(K),\lambda}^2\right] + |v_h|_{\mathrm{J}}^2 \,\mathrm{d}t + \omega \|v_h(t_n)\|_{a_h}^2 + \omega \|v_h(t_{n-1}^+)\|_{a_h}^2,$$

and similarly for u_h and z_h. We also have

$$\begin{aligned} A_h(u_h; z_h) - A_h(v_h; z_h) = {} & \int_{I_n} \sum_{K \in \mathcal{T}_h} \langle F_\gamma[u_h] - F_\gamma[v_h], L_\omega z_h \rangle_K \,\mathrm{d}t + C_h^{\mathcal{F}}(w_h, z_h) \\ & + \int_{I_n} B_{h,1/2}(w_h, z_h) - \sum_{K \in \mathcal{T}_h} \langle L_\lambda w_h, L_\lambda z_h \rangle_K \,\mathrm{d}t + \omega\, a_h(w_h(t_{n-1}^+), z_h(t_{n-1}^+)). \end{aligned}$$

Lipschitz continuity of F_γ implies that

$$\int_{I_n} \sum_{K \in \mathcal{T}_h} |\langle F_\gamma[u_h] - F_\gamma[v_h], L_\omega z_h \rangle_K| \,\mathrm{d}t \lesssim \|w_h\|_{h,1} \|z_h\|_{h,1}.$$

Furthermore, we have $|C_h^{\mathcal{F}}(w_h, z_h)| \leq E_1 + E_2$, where

$$\begin{aligned} E_1 &\coloneqq \omega \int_{I_n} \sum_{F \in \mathcal{F}_h^i} |\langle [\![\nabla w_h \cdot n_F]\!], \{\partial_t z_h\} \rangle_F| \,\mathrm{d}t, \\ E_2 &\coloneqq \omega \int_{I_n} \sum_{F \in \mathcal{F}_h^{i,b}} \mu_F |\langle [\![w_h]\!], [\![\partial_t z_h]\!] \rangle_F| + |\langle [\![w_h]\!], \{\nabla \partial_t z_h \cdot n_F\} \rangle_F| \,\mathrm{d}t. \end{aligned}$$

The shape-regularity of the meshes $\{\mathcal{T}\}_h$, the mesh assumption (2.11) and the trace and inverse inequalities show that

$$\begin{aligned} E_1 &\lesssim \sqrt{\int_{I_n} \sum_{K \in \mathcal{T}_h} \omega^2 \,\|\partial_t z_h\|_{L^2(K)}^2 \mathrm{d}t} \sqrt{\int_{I_n} \sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|[\![\nabla w_h \cdot n_F]\!]\|_{L^2(F)}^2 \mathrm{d}t}, \\ E_2 &\lesssim \sqrt{\int_{I_n} \sum_{K \in \mathcal{T}_h} \omega^2 \,\|\partial_t z_h\|_{L^2(K)}^2 \mathrm{d}t} \sqrt{\int_{I_n} \sum_{F \in \mathcal{F}_h^{i,b}} \frac{\tilde{p}_F^6}{\tilde{h}_F^3} \|[\![w_h]\!]\|_{L^2(F)}^2 \mathrm{d}t}. \end{aligned}$$

Since μ_F and η_F satisfy (4.48), we conclude that

$$|C_h^{\mathcal{F}}(w_h, z_h)| \lesssim \|w_h\|_{h,1} \, \|z_h\|_{h,1}.$$

By applying trace and inverse inequalities on the flux terms of the bilinear form $B_{h,*}$, it is found that

$$\int_{I_n} |B_{h,1/2}(w_h, z_h)| + \sum_{K \in \mathcal{T}_h} |\langle L_\lambda w_h, L_\lambda z_h \rangle_K| \, \mathrm{d}t \lesssim \|u_h - v_h\|_{h,1} \, \|z_h\|_{h,1},$$

thus completing the proof of (4.57). Since the numerical scheme (4.43) is equivalent to solving (4.44) for each $I_n \in \mathcal{J}_\tau$, and since A_h is strongly monotone and Lipschitz continuous on the subspace of $V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ of functions with support in $\overline{I_n}$, for each $I_n \in \mathcal{J}_\tau$, repeated applications of the Browder–Minty theorem show that there exists a unique $u_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ that solves (4.43). □

It is clear from the proof of Theorem 4.8 that the choice of penalisation orders in (4.48) was made to treat the term E_2 defined above. As we shall see in the following section, this leads to a suboptimality in p by one order and a half.

4.6 Error analysis

In the first part of this section, we present error bounds for regular solutions, i.e. when the solution belongs to $H^1(I_n; H)$ for each $I_n \in \mathcal{J}_\tau$. It is found that the method has convergence orders that are optimal with respect to h, τ and $\mathbf{q}$, and that are possibly suboptimal with respect to $\mathbf{p}$ by an order and a half. In a second part, we use Clément quasi-interpolants in Bochner spaces to extend the analysis under weaker regularity assumptions, in order to cover the case where $u \notin H^1(I_n; H)$.

There are two reasons for presenting the error analysis in two parts. First, the error analysis for regular solutions is simpler and permits the use of known approximation theory from [67], whereas the case of solutions with lower regularity requires the additional construction of a Clément quasi-interpolation operator. Second, the Clément operator is generally suboptimal by one order in τ when applied to very regular solutions. Thus, the results given here for regular and solutions with low regularity are complementary to each other. Since error bounds for solutions with minimal spatial regularity were given in Chapter 3, we choose to focus here instead on the temporal regularity of the solution. Prioritising the treatment of temporal regularity is well-justified in the present context since parabolic regularity theory for model problems shows that $H^1(H^2)$-regularity is comparable to $L^2(H^4)$-regularity.

We will present error bounds in the norm $\|\cdot\|_h$ defined by

$$\|v\|_h^2 \coloneqq \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \left[\omega^2 \|\partial_t v\|_{L^2(K)}^2 + |v|_{H^2(K),\lambda}^2\right] + |v|_{\mathrm{J}}^2 \,\mathrm{d}t + \omega \sum_{n=0}^{N-1} \|\langle\!\langle v\rangle\!\rangle_n\|_{a_h}^2. \tag{4.59}$$

We remark that for $v_h \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$, we have $\|v_h\|_{h,1}^2 = \|v_h\|_h^2 + \omega\|\langle\!\langle v_h\rangle\!\rangle_N\|_{a_h}^2$. Error bounds in the norm $\|\cdot\|_{h,1}$ can be shown under additional regularity assumptions for the solution at the final time T. To simplify the notation in this section, let

$$X_0 \coloneqq L^2(\Omega), \quad X_1 \coloneqq H_0^1(\Omega), \quad X_2 \coloneqq H = H^2(\Omega) \cap H_0^1(\Omega). \tag{4.60}$$

Similarly to the definition of the broken Sobolev spaces $H^s(\Omega;\mathcal{T}_h)$, for a Hilbert space X, we define the broken Bochner space $H^\sigma(I;X;\mathcal{J}_\tau)$ to be the space of functions $u \in L^2(I;X)$ with restrictions $u|_{I_n} \in H^\sigma(I_n;X)$ for each $I_n \in \mathcal{J}_\tau$. We equip $H^\sigma(I;X;\mathcal{J}_\tau)$ with the obvious norm.

4.6.1 Error bound for regular solutions

If the solution u of (4.7) belongs to $H^1(I;H,\mathcal{J}_\tau)$, then the error analysis may be based on the following approximation result, found for instance in [67], albeit presented here in a form amenable to our purposes.

Theorem 4.9. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, and let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of $I = (0,T)$. For each τ, let $\mathbf{q} = (q_1,\dots,q_N)$ be a vector of positive integers. Then, for each τ, there exists a linear operator $\Pi_\tau^{\mathbf{q}}\colon H(I;\Omega)\cap H^1(I;H;\mathcal{J}_\tau) \to V^{\tau,\mathbf{q}}$ such that the following holds. The operator $\Pi_\tau^{\mathbf{q}}$ is an interpolant at the interval endpoints, i.e. for any $u \in H(I;\Omega)\cap H^1(I;H;\mathcal{J}_\tau)$, we have $\Pi_\tau^{\mathbf{q}}u(t_n) = \Pi_\tau^{\mathbf{q}}u(t_n^+) = u(t_n)$ for each $0 \le n \le N$. For any $I_n \in \mathcal{J}_\tau$, any $\ell \in \{0,1,2\}$, any real number $\sigma_{n,\ell} \ge 1$ and any $j \in \{0,1\}$, we have*

$$\|u - \Pi_\tau^{\mathbf{q}}u\|_{H^j(I_n;X_\ell)} \lesssim \frac{\tau_n^{\varrho_{n,\ell}-j}}{q_n^{\sigma_{n,\ell}-j}} \|u\|_{H^{\sigma_{n,\ell}}(I_n;X_\ell)} \qquad \forall\, u \in H^{\sigma_{n,\ell}}(I_n;X_\ell), \tag{4.61}$$

where $\varrho_{n,\ell} \coloneqq \min(\sigma_{n,\ell}, q_n+1)$, and where the constant depends only on $\sigma_{n,\ell}$ and $\max\tau$.

The construction of $\Pi_\tau^{\mathbf{q}}$ in the proof of Theorem 4.9 involves the truncated Legendre series of $\partial_t u$ and the values of u at the partition points. Therefore, the requirement of $H^1(I;H;\mathcal{J}_\tau)$ regularity is used to ensure that $\Pi_\tau^{\mathbf{q}}|_{I_n}$ maps into $\mathcal{Q}_{q_n}(H)$. A different approximation operator is used in section 4.6.2 to perform an analysis under weaker regularity assumptions.

Theorem 4.10. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex polytopal domain and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11), (2.12) *and* (2.13).

Let $I = (0,T)$ and let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of I, and, for each τ, let $\mathbf{q} = (q_1, \dots, q_N)$ be a vector of positive integers. Let Λ be a compact metric space and let the data a, b, c and f be continuous on $\overline{\Omega} \times \overline{I} \times \Lambda$ and satisfy (4.8) and (4.11), or alternatively (4.10) in the case where $b \equiv 0$ and $c \equiv 0$. Let μ_F and η_F satisfy (4.48), with c_μ and c_η chosen so that Lemmas 3.7 and 4.7 hold with $\kappa < (1-\varepsilon)^{-1}$.

Let $u \in H(I;\Omega)$ be the unique solution of the HJB equation (4.7), and assume that $u \in L^2(I; H^{\mathbf{s}}(\Omega;\mathcal{T}_h))$ and $\partial_t u \in L^2(I; H^{\overline{\mathbf{s}}}(\Omega,\mathcal{T}_h))$ for each h, with $s_K > 5/2$ and $\overline{s}_K > 0$ for each $K \in \mathcal{T}_h$. Suppose also that, for each τ, each $\ell \in \{0,2\}$ and each $I_n \in \mathcal{J}_\tau$, the function $u|_{I_n} \in H^{\sigma_{n,\ell}}(I_n; X_\ell)$ for some $\sigma_{n,\ell} \geq 1$. Assume that $u_0 \in H^1_0(\Omega) \cap H^{\tilde{\mathbf{s}}}(\Omega;\mathcal{T}_h)$ with $\tilde{s}_K > 3/2$ for each $K \in \mathcal{T}_h$. Then, we have

$$\begin{aligned}(4.62)\quad \|u-u_h\|_h^2 \lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-7}}\|u\|^2_{H^{s_K}(K)} + \frac{h_K^{2\bar{t}_K}}{p_K^{2\overline{s}_K}}\|\partial_t u\|^2_{H^{\overline{s}_K}(K)}\mathrm{d}t \\ + \max_{K\in\mathcal{T}_h} p_K^3 \sum_{n=1}^N \sum_{\ell\in\{0,2\}} \frac{\tau_n^{2\varrho_{n,\ell}-2+\ell}}{q_n^{2\sigma_{n,\ell}-2+\ell}}\|u\|^2_{H^{\sigma_{n,\ell}}(I_n;X_\ell)} + \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\tilde{t}_K-2}}{p_K^{2\tilde{s}_K-3}}\|u_0\|^2_{H^{\tilde{s}_K}(K)},\end{aligned}$$

with a constant independent of u, h, $\mathbf{p}$, τ and $\mathbf{q}$, where

$$\begin{aligned} t_K &\coloneqq \min(s_K, p_K+1), & \bar{t}_K &\coloneqq \min(\overline{s}_K, p_K+1),\\ \tilde{t}_K &\coloneqq \min(\tilde{s}_K, p_K+1), & \varrho_{n,\ell} &\coloneqq \min(\sigma_{n,\ell}, q_n+1),\end{aligned}$$

for each $K \in \mathcal{T}_h$, each $1 \leq n \leq N$ and each $\ell \in \{0,2\}$.

It is seen that the error bound is optimal with respect to h, τ and $\mathbf{q}$, but is suboptimal with respect to $\mathbf{p}$ by an order and a half. We remark that since Theorem 4.10 assumes $u \in H^1(I_1;H)$, the initial data satisfies $u_0 \in H$, so we may take $\tilde{s}_K \geq 2$ for each $K \in \mathcal{T}_h$.

Proof. The approximation theory for *hp*-version discontinuous Galerkin finite element spaces shows that there exists a sequence of linear projection operators $\{\Pi_h^{\mathbf{p}}\}_h$, with $\Pi_h^{\mathbf{p}} \colon L^2(\Omega) \to V_{h,\mathbf{p}}$ and such that for each $K \in \mathcal{T}_h$, for each nonnegative real number $r_K \leq \max(s_K, \overline{s}_K, \tilde{s}_K)$ and for each nonnegative integer $j \leq r_K$, and if $r_K > 1/2$, for each multi-index β such that $|\beta| < r_K - 1/2$, we have

$$(4.63)\qquad \|u - \Pi_h^{\mathbf{p}} u\|_{H^j(K)} \lesssim \frac{h_K^{\min(r_K,p_K+1)-j}}{(p_K+1)^{r_K-j}}\|u\|_{H^{r_K}(K)} \qquad \forall\, u \in H^{r_K}(K),$$

$$(4.64)\qquad \|D^\beta(u - \Pi_h^{\mathbf{p}} u)\|_{L^2(\partial K)} \lesssim \frac{h_K^{\min(r_K,p_K+1)-|\beta|-1/2}}{(p_K+1)^{r_K-|\beta|-1/2}}\|u\|_{H^{r_K}(K)} \qquad \forall\, u \in H^{r_K}(K),$$

where the constant is independent of r_K, h_K, p_K but possibly dependent on s_K, $\overline{s}_K$ and $\tilde{s}_K$. The technical form of this approximation result expresses the optimality and stability of $\Pi_h^{\mathbf{p}}$ for functions in $H^{r_K}(K)$, $0 \leq r_K \leq \max(s_K, \overline{s}_K, \tilde{s}_K)$. In particular, we will use the fact that $\Pi_h^{\mathbf{p}}$ is elementwise L^2-stable, H^1-stable and H^2-stable in the analysis below.

For each h and τ, let $z_\tau \coloneqq \Pi_\tau^{\mathbf{q}} u \in V^{\tau,\mathbf{q}}$, and let $z_h \coloneqq \Pi_h^{\mathbf{p}} z_\tau \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$. Continuity of z_τ implies continuity of z_h, so that $(\!|z_h|\!)_n = 0$ for each $1 \leq n < N$. Furthermore, we have $z_\tau(0^+) = u_0$, so $z_h(0^+) = \Pi_h^{\mathbf{p}} u_0$. Let $\xi_h \coloneqq u - z_h$ and let $\psi_h \coloneqq u_h - z_h$, so that $u - u_h = \xi_h - \psi_h$. Recall that $\|\psi_h\|_h \leq \|\psi_h\|_{h,1}$. Theorem 4.8, the scheme (4.43) and Corollary 4.6 show that

$$
\begin{aligned}
(4.65)\quad \|\psi_h\|_{h,1}^2 &\lesssim A_h(u_h;\psi_h) - A_h(z_h;\psi_h) = A_h(u;\psi_h) - A_h(z_h;\psi_h) \\
&= \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle F_\gamma[u] - F_\gamma[z_h], L_\omega \psi_h\rangle_K + B_{h,1/2}(\xi_h,\psi_h)\,\mathrm{d}t \\
&\quad - \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \langle L_\lambda \xi_h, L_\lambda \psi_h\rangle_K\,\mathrm{d}t + C_h^{\mathcal{F}}(\xi_h,\psi_h) + \omega\, a_h(\xi_h(t_0^+),\psi_h(t_0^+)).
\end{aligned}
$$

Therefore $\|\psi_h\|_h^2 \leq \|\psi_h\|_{h,1}^2 \leq \sum_{i=1}^4 D_i$, where the quantities D_i, $1 \leq i \leq 4$, are defined by

$$
D_1 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} |\langle F_\gamma[u] - F_\gamma[z_h], L_\omega\psi_h\rangle_K| + |\langle L_\lambda\xi_h, L_\lambda\psi_h\rangle_K|\mathrm{d}t,
$$

$$
D_2 \coloneqq \sum_{n=1}^N \int_{I_n} |B_{h,1/2}(\xi_h,\psi_h)|\mathrm{d}t, \quad D_3 \coloneqq |C_h^{\mathcal{F}}(\xi_h,\psi_h)|, \quad D_4 \coloneqq \omega|a_h(\xi_h(0^+),\psi_h(0^+))|.
$$

Lipschitz continuity of F_γ implies that $D_1 \lesssim \sqrt{E_1+E_2}\ \|\psi_h\|_{h,1}$, where E_1 and E_2 are defined by

$$
E_1 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|\partial_t \xi_h\|_{L^2(K)}^2 \mathrm{d}t, \quad E_2 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|\xi_h\|_{H^2(K)}^2 \mathrm{d}t.
$$

Since the sequence of meshes $\{\mathcal{T}_h\}_h$ is shape-regular and since $\psi_h|_{I_n} \in \mathcal{Q}_{q_n}(V_{h,\mathbf{p}})$ for each $I_n \in \mathcal{J}_\tau$, the use of trace and inverse inequalities on the flux terms appearing in $B_{h,1/2}(\xi_h,\psi_h)$ yields $D_2 \lesssim \sqrt{\sum_{i=2}^6 E_i}\ \|\psi_h\|_{h,1}$, where the quantities E_i, $3 \leq i \leq 5$, are defined by

$$
E_3 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{F\in\mathcal{F}_h^i} \mu_F^{-1} \|\mathrm{div}_{\mathrm{T}}\, \nabla_{\mathrm{T}}\{\xi_h\}\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F^{-1}\|\nabla_{\mathrm{T}}\{\nabla\xi_h\cdot n_F\}\|_{L^2(F)}^2 \mathrm{d}t,
$$

$$
E_4 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{F\in\mathcal{F}_h^{i,b}} \eta_F^{-1} \|\{\nabla\xi_h\cdot n_F\}\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_h^{i}} \mu_F^{-1}\|\{\xi_h\}\|_{L^2(F)}^2 \mathrm{d}t,
$$

$$
E_5 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{F\in\mathcal{F}_h^{i}} \mu_F \|[\![\nabla\xi_h\cdot n_F]\!]\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F\|[\![\nabla_{\mathrm{T}}\,\xi_h]\!]\|_{L^2(F)}^2 \mathrm{d}t,
$$

$$
E_6 \coloneqq \sum_{n=1}^N \int_{I_n} \sum_{F\in\mathcal{F}_h^{i,b}} \eta_F \|[\![\xi_h]\!]\|_{L^2(F)}^2 \mathrm{d}t.
$$

Note that $\partial_t \psi_h|_{I_n} \in \mathcal{Q}_{q_n-1}(V_{h,\mathbf{p}})$ for each $I_n \in \mathcal{J}_\tau$. Thus, similarly to the proof of Theorem 4.8, the use of trace and inverse inequalities leads to $D_3 \lesssim \sqrt{E_4 + E_5}\, \|\psi_h\|_{h,1}$. It follows from (4.47) that we have $D_4 \lesssim \sqrt{E_6 + E_7 + E_8}\, \|\psi_h\|_{h,1}$, where the quantities E_i, $7 \le i \le 9$, are defined by

$$E_7 \coloneqq \sum_{K\in\mathcal{T}_h} \|u_0 - \Pi_h^{\mathbf{p}} u_0\|^2_{H^1(K)}, \qquad E_8 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \|u_0 - \Pi_h^{\mathbf{p}} u_0\|^2_{L^2(F)},$$
$$E_9 \coloneqq \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F^{-1} \|\{\nabla(u_0 - \Pi_h^{\mathbf{p}} u_0)\cdot n_F\}\|^2_{L^2(F)}.$$

Therefore, (4.65) implies that $\|\psi_h\|_h^2 \lesssim \sum_{i=1}^9 E_i$. The properties of the operator $\Pi_h^{\mathbf{p}}$, namely its linearity, L^2-stability and approximation properties (4.63), together with (4.61), imply that

$$\begin{aligned} E_1 &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|\partial_t u - \Pi_h^{\mathbf{p}} \partial_t u\|^2_{L^2(K)} + \|\Pi_h^{\mathbf{p}}(\partial_t u - \partial_t z_\tau)\|^2_{L^2(K)} \mathrm{d}t \\ &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|\partial_t u - \Pi_h^{\mathbf{p}} \partial_t u\|^2_{L^2(K)} \mathrm{d}t + \sum_{n=1}^N \|u - z_\tau\|^2_{H^1(I_n;X_0)} \\ &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\bar{t}_K}}{p_K^{2\bar{s}_K}} \|\partial_t u\|^2_{H^{\bar{s}_K}(K)} \mathrm{d}t + \sum_{n=1}^N \frac{\tau_n^{2\varrho_{n,0}-2}}{q_n^{2\sigma_{n,0}-2}} \|u\|^2_{H^{\sigma_{n,0}}(I_n;X_0)}. \end{aligned} \tag{4.66}$$

Since the operator $\Pi_h^{\mathbf{p}}$ is elementwise H^2-stable, it is found that

$$\begin{aligned} E_2 &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|u - \Pi_h^{\mathbf{p}} u\|^2_{H^2(K)} + \|\Pi_h^{\mathbf{p}}(u - z_\tau)\|^2_{H^2(K)} \mathrm{d}t \\ &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \|u - \Pi_h^{\mathbf{p}} u\|^2_{H^2(K)} \mathrm{d}t + \sum_{n=1}^N \|u - z_\tau\|^2_{L^2(I_n;X_2)} \\ &\lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^N \frac{\tau_n^{2\varrho_{n,2}}}{q_n^{2\sigma_{n,2}}} \|u\|^2_{H^{\sigma_{n,2}}(I_n;X_2)}. \end{aligned} \tag{4.67}$$

The mesh assumptions (2.11), (2.12) and (2.13), the bound (4.64), and the application of

trace and inverse inequalities on $\Pi_h^{\mathbf{p}}(u - z_\tau)|_{I_n} \in \mathcal{Q}_{q_n}(V_{h,\mathbf{p}})$, imply that

$$\begin{aligned} E_3 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2} \|D^2(u - \Pi_h^{\mathbf{p}} z_\tau)\|^2_{L^2(\partial K)} \,\mathrm{d}t \\ &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K}{p_K^2} \left[\|D^2(u - \Pi_h^{\mathbf{p}} u) + D^2 \Pi_h^{\mathbf{p}}(u - z_\tau)\|^2_{L^2(\partial K)} \right] \mathrm{d}t \\ &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}} \|u\|^2_{H^{s_K}(K)} + \sum_{K\in\mathcal{T}_h} \|u - z_\tau\|^2_{H^2(K)} \mathrm{d}t \\ &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \frac{\tau_n^{2\varrho_{n,2}}}{q_n^{2\sigma_{n,2}}} \|u\|^2_{H^{\sigma_{n,2}}(I_n;X_2)}. \end{aligned} \tag{4.68}$$

Similarly to E_3, we find that

$$E_4 \lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K}}{p_K^{2s_K+1}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \frac{\tau_n^{2\varrho_{n,0}}}{q_n^{2\sigma_{n,0}}} \|u\|^2_{H^{\sigma_{n,0}}(I_n;X_0)}. \tag{4.69}$$

The spatial regularity of u and z_τ imply that

$$\begin{aligned} E_5 = \sum_{n=1}^{N} \int_{I_n} \sum_{F\in\mathcal{F}_h^i} \mu_F \|[\![\nabla \left[u - \Pi_h^{\mathbf{p}} u + \Pi_h^{\mathbf{p}}(u - z_\tau) - (u - z_\tau)\right] \cdot n_F]\!]\|^2_{L^2(F)} \mathrm{d}t \\ + \sum_{n=1}^{N} \int_{I_n} \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \|[\![\nabla_{\mathrm{T}} \left[u - \Pi_h^{\mathbf{p}} u + \Pi_h^{\mathbf{p}}(u - z_\tau) - (u - z_\tau)\right]]\!]\|^2_{L^2(F)} \mathrm{d}t. \end{aligned}$$

Therefore, the mesh assumptions (2.11), (2.12) and (2.13) and (4.64) yield

$$\begin{aligned} E_5 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{p_K^2}{h_K} \|\nabla(u - \Pi_h^{\mathbf{p}} u) + \nabla \left[u - z_\tau - \Pi_h^{\mathbf{p}}(u - z_\tau)\right]\|^2_{L^2(\partial K)} \mathrm{d}t \\ &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|^2_{H^{s_K}(K)} + \sum_{K\in\mathcal{T}_h} p_K \|u - z_\tau\|^2_{H^2(K)} \mathrm{d}t \\ &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \max_{K\in\mathcal{T}_h} p_K \sum_{n=1}^{N} \frac{\tau_n^{2\varrho_{n,2}}}{q_n^{2\sigma_{n,2}}} \|u\|^2_{H^{\sigma_{n,2}}(I_n;X_2)}. \end{aligned} \tag{4.70}$$

Likewise, it follows from the spatial regularity of z_τ, the mesh assumptions, and the ap-

proximation bound (4.64) that

$$
\begin{aligned}
E_6 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{p_K^6}{h_K^3} \|u - \Pi_h^{\mathbf{p}} u + \Pi_h^{\mathbf{p}}(u - z_\tau) - (u - z_\tau)\|_{L^2(\partial K)}^2 \mathrm{d}t \\
&\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-7}} \|u\|_{H^{s_K}(K)}^2 + \sum_{K\in\mathcal{T}_h} p_K^3 \|u - z_\tau\|_{H^2(K)}^2 \mathrm{d}t \qquad (4.71) \\
&\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-7}} \|u\|_{H^{s_K}(K)}^2 \mathrm{d}t + \max_{K\in\mathcal{T}_h} p_K^3 \sum_{n=1}^{N} \frac{\tau_n^{2\varrho_{n,2}}}{q_n^{2\sigma_{n,2}}} \|u\|_{H^{\sigma_{n,2}}(I_n;X_2)}^2.
\end{aligned}
$$

Finally, it is readily shown that

$$
\sum_{i=7}^{9} E_i \lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\tilde{t}_K-2}}{p_K^{2\tilde{s}_K-3}} \|u_0\|_{H^{\tilde{s}_K}(K)}^2. \qquad (4.72)
$$

Since $\|\xi_h\|_h^2 \leq \sum_{i=1}^{9} E_i$, the above bounds and the triangle inequality $\|u - u_h\|_h \leq \|\xi_h\|_h + \|\psi_h\|_h$ complete the proof of (4.62). □

4.6.2 Error bound for solutions with low regularity

The proof of Theorem 4.10 depends on the approximation result from Theorem 4.9, which requires that the solution u belongs to $H^1(I;H;\mathcal{J}_\tau)$. In this section, we relax this condition by using a Clément quasi-interpolation result instead of Theorem 4.9.

For $\mathcal{J}_\tau$ a regular partition of $(0,T)$, let $\{\phi_m\}_{m=0}^N$ denote the set of hat functions of $\mathcal{J}_\tau$, i.e. ϕ_m is the unique piecewise-affine function on $\mathcal{J}_\tau$ such that $\phi_m(t_n) = \delta_{nm}$ for $0 \leq n, m \leq N$. For $0 \leq m \leq N$, let $J_m \coloneqq \operatorname{supp} \phi_m$, and note that $J_m = \overline{I_m} \cup \overline{I_{m+1}}$ for $1 \leq m < N$, whilst $J_0 = \overline{I_1}$ and $J_N = \overline{I_N}$.

Theorem 4.11. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, and let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of $I = (0,T)$. For each τ, let $\mathbf{q} = (q_1, \ldots, q_N)$ be a vector of positive integers. Suppose that there exist positive constants c_τ and c_q such that, for each τ, we have*

$$
\frac{1}{c_\tau} \leq \frac{\tau_{n-1}}{\tau_n} \leq c_\tau, \quad \frac{1}{c_q} \leq \frac{q_{n-1}}{q_n} \leq c_q, \quad 2 \leq n \leq N. \qquad (4.73)
$$

Let $u \in L^2(I;H)$ and suppose that $u|_{J_m} \in H^{\sigma_{m,\ell}}(J_m;X_\ell)$ for some $\sigma_{m,\ell} \in \mathbb{R}_{\geq 0}$ for each $\ell \in \{0,1,2\}$ and each $0 \leq m \leq N$. Then, there exists a sequence of functions $\{z_\tau\}_\tau$, such that $z_\tau \in V^{\tau,\mathbf{q}}$ for each τ, and such that the following properties hold. The functions z_τ are continuous on I, i.e. $(\!|z_\tau|\!)_n = 0$ for each $1 \leq n < N$. For each $\ell \in \{0,1,2\}$ and each $I_n \in \mathcal{J}_\tau$, we have

$$
\|z_\tau\|_{L^2(I_n;X_\ell)} \lesssim \sum_{J_m \supset I_n} \|u\|_{L^2(J_m;X_\ell)}, \qquad (4.74)
$$

where the constant is independent of all other quantities. For each $\ell \in \{0,1,2\}$, each $I_n \in \mathcal{J}_\tau$ and each nonnegative integer $j \le \min_{J_m \supset I_n} \sigma_{m,\ell}$, we have

$$\|u - z_\tau\|_{H^j(I_n;X_\ell)} \lesssim \sum_{J_m \supset I_n} \frac{\tau_n^{\varrho_{m,\ell}-j}}{q_n^{\sigma_{m,\ell}-j}} \|u\|_{H^{\sigma_{m,\ell}}(J_m;X_\ell)}, \tag{4.75}$$

where $\varrho_{m,\ell} \coloneqq \min(\sigma_{m,\ell}, \min_{I_n \subset J_m} q_n)$, and the constant depends only on $\max \sigma_{m,\ell}$, $\max \tau$, c_τ and c_q.

Proof. For $0 \le m \le N$, define $\bar{q}_m \coloneqq \min_{I_n \subset J_m} q_n$, and note $\bar{q}_m \ge 1$ for all m since $q_n \ge 1$ for all n. Since $u \in L^2(J_m; X_2)$ for each m, standard approximation theory for Bochner spaces (see Appendix C) implies that there exist functions $v_m \in \mathcal{Q}_{\bar{q}_m - 1}(H)$, $0 \le m \le N$, with the following properties. For each $\ell \in \{0,1,2\}$, we have $\|v_m\|_{L^2(J_m;X_\ell)} \lesssim \|u\|_{L^2(J_m;X_\ell)}$, with a constant independent of all other quantities. For each $\ell \in \{0,1,2\}$ and each nonnegative integer $j \le \sigma_{m,\ell}$, we have

$$\|u - v_m\|_{H^j(J_m;X_\ell)} \lesssim \frac{|J_m|^{\varrho_{m,\ell}-j}}{\bar{q}_m^{\sigma_{m,\ell}-j}} \|u\|_{H^{\sigma_{m,\ell}}(J_m;X_\ell)}, \tag{4.76}$$

where $\varrho_{m,\ell} \coloneqq \min(\sigma_{m,\ell}, \bar{q}_m)$, where $|J_m|$ is the length of the interval J_m, and where the constant depends only on $\max \sigma_{m,\ell}$ and $\max \tau$.

The hypothesis (4.73) and the bound (4.76) imply that, for each $I_n \subset J_m$, each $\ell \in \{0,1,2\}$ and each nonnegative integer $j \le \sigma_{m,\ell}$,

$$\|u - v_m\|_{H^j(I_n;X_\ell)} \lesssim \frac{\tau_n^{\varrho_{m,\ell}-j}}{q_n^{\sigma_{m,\ell}-j}} \|u\|_{H^{\sigma_{m,\ell}}(J_m;X_\ell)}, \tag{4.77}$$

where the constant depends only on $\max \sigma_{m,\ell}$, $\max \tau$, c_τ and c_q.

Define $z_\tau \coloneqq \sum_{m=0}^N \phi_m v_m$, where ϕ_m is the hat function over the interval J_m. Note that we have $v_m|_{I_n} \in \mathcal{Q}_{q_n - 1}(H)$ for each $I_n \in \mathcal{J}_\tau$ since $\bar{q}_m \le q_n$ for each $I_n \subset J_m$. Since ϕ_m is piecewise affine, it follows that $z_\tau|_{I_n} \in \mathcal{Q}_{q_n}(H)$ for each $I_n \in \mathcal{J}_\tau$, thereby showing that $z_\tau \in V^{\tau,\mathfrak{q}}$. Furthermore, it is clear that z_τ is continuous on I, i.e. $(\!|z_\tau|\!)_n = 0$ for each $1 \le n \le N-1$. The bound (4.74) follows from $\|v_m\|_{L^2(J_m;X_\ell)} \lesssim \|u\|_{L^2(J_m;X_\ell)}$ and from the fact that $\|\phi_m\|_{L^\infty(I)} = 1$ for each $0 \le m \le N$. Since $\{\phi_m\}_{m=0}^N$ forms a partition of unity, the bound (4.77) implies that, for each $I_n \in \mathcal{J}_\tau$ and each $\ell \in \{0,1,2\}$,

$$\begin{aligned}\|u - z_\tau\|_{L^2(I_n;X_\ell)} &\le \sum_{J_m \supset I_n} \|\phi_m(u - v_m)\|_{L^2(I_n;X_\ell)} \\ &\lesssim \sum_{J_m \supset I_n} \|u - v_m\|_{L^2(I_n;X_\ell)} \lesssim \sum_{J_m \supset I_n} \frac{\tau_n^{\varrho_{m,\ell}}}{q_n^{\sigma_{m,\ell}}} \|u\|_{H^{\sigma_{m,\ell}}(J_m;X_\ell)},\end{aligned}$$

and, for each integer $1 \le j \le \min_{J_m \supset I_n} \sigma_{m,\ell}$,

$$|u - z_\tau|_{H^j(I_n;X_\ell)} \le \sum_{J_m \supset I_n} |\phi_m(u - v_m)|_{H^j(I_n;X_\ell)}$$
$$\lesssim \sum_{J_m \supset I_n} |u - v_m|_{H^j(I_n;X_\ell)} + \frac{1}{\tau_n}|u - v_m|_{H^{j-1}(I_n;X_\ell)} \lesssim \sum_{J_m \supset I_n} \frac{\tau_n^{\varrho_{m,\ell}-j}}{q_n^{\sigma_{m,\ell}-j}} \|u\|_{H^{\sigma_{m,\ell}}(J_m;X_\ell)}.$$

This completes the proof of (4.75). □

Theorem 4.12. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex polytopal domain and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11) *and* (2.12) *and* (2.13)*. Let $I = (0,T)$ and let $\{\mathcal{J}_\tau\}_\tau$ be a sequence of regular partitions of I, and, for each τ, let $\mathbf{q}$ be a vector of positive integers such that* (4.73) *holds. Let Λ be a compact metric space and let the data a, b, c and f be continuous on $\overline{\Omega} \times \overline{I} \times \Lambda$ and satisfy* (4.8) *and* (4.11)*, or alternatively* (4.10) *in the case where $b \equiv 0$ and $c \equiv 0$. Let μ_F and η_F satisfy* (4.48)*, with c_μ and c_η chosen so that Lemmas* 3.7 *and* 4.7 *hold with $\kappa < (1-\varepsilon)^{-1}$.*

Let $u \in H(I;\Omega)$ be the unique solution of the HJB equation (4.7)*, and assume that $u \in L^2(I;H^{\mathbf{s}}(\Omega;\mathcal{T}_h))$ and $\partial_t u \in L^2(I;H^{\overline{\mathbf{s}}}(\Omega,\mathcal{T}_h))$ for each h, with $s_K > 5/2$ and $\overline{s}_K > 0$ for each $K \in \mathcal{T}_h$. Suppose also that, for each τ, $\ell \in \{0,1,2\}$, and each $0 \le m \le N$, the function $u|_{J_m} \in H^{\sigma_{m,\ell}}(J_m;X_\ell)$ for some real $\sigma_{m,\ell} \ge 0$, with $\sigma_{m,0} \ge 1$ for all m. Assume that $u_0 \in H^1_0(\Omega) \cap H^{\tilde{\mathbf{s}}}(\Omega;\mathcal{T}_h)$ with $\tilde{s}_K > 3/2$ for each $K \in \mathcal{T}_h$. Then, we have*

$$\|u - u_h\|_h^2 \lesssim \sum_{n=1}^N \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-7}} \|u\|^2_{H^{s_K}(K)} + \frac{h_K^{2\bar{t}_K}}{p_K^{2\overline{s}_K}} \|\partial_t u\|^2_{H^{\overline{s}_K}(K)} \mathrm{d}t$$
$$+ \max_{K\in\mathcal{T}_h} p_K^3 \sum_{n=1}^N \sum_{\ell=0}^2 \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,\ell}-2+\ell}}{q_n^{2\sigma_{m,\ell}-2+\ell}} \|u\|^2_{H^{\sigma_{m,\ell}}(J_m;X_\ell)} + \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\tilde{t}_K-2}}{p_K^{2\tilde{s}_K-3}} \|u_0\|^2_{H^{\tilde{s}_K}(K)}, \tag{4.78}$$

with a constant independent of h, $\mathbf{p}$, τ, $\mathbf{q}$, and u, and where $t_K := \min(s_K, p_K+1)$, $\bar{t}_K := \min(\overline{s}_K, p_K+1)$, and $\tilde{t}_K := \min(\tilde{s}_K, p_K+1)$ for each $K \in \mathcal{T}_h$, and where $\varrho_{m,\ell} := \min(\sigma_{m,\ell}, \min_{I_n \subset J_m} q_n)$ for each $0 \le m \le N$ and each $\ell \in \{0,1,2\}$.

Proof. For each h, let $\Pi_h^{\mathbf{p}} \colon L^2(\Omega) \to V_{h,\mathbf{p}}$ denote the approximation operator of the proof of Theorem 4.10; for each τ, let $z_\tau \in V^{\tau,\mathbf{q}}$ denote the approximation of u given by Theorem 4.11; then define $z_h := \Pi_h^{\mathbf{p}} z_\tau \in V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$. The fact that z_τ is continuous on $(0,T)$ implies that z_h is also continuous on $(0,T)$, so $(\!|z_h|\!)_n = 0$ for $1 \le n < N$. Let $\xi_h := u - z_h$ and $\psi_h := u_h - z_h$, so that $u - u_h = \xi_h - \psi_h$. As in the proof of Theorem 4.10, it is found that $\|\psi_h\|_h^2 \le \|\psi_h\|_{h,1}^2 \lesssim \sum_{i=1}^9 E_i$, where the quantities E_i, $1 \le i \le 9$, are defined as before. Note that since $\sigma_{m,0} \ge 1$ for all m, the bound (4.75) is applicable for $j = 1$ and $\ell = 0$. Therefore, the arguments from the proof of Theorem 4.10 and the approximation properties of z_τ from

Theorem 4.11 imply that

$$
\begin{aligned}
E_1 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\bar{t}_K}}{p_K^{2\bar{s}_K}} \|\partial_t u\|^2_{H^{\bar{s}_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{t_n^{2\varrho_{m,0}-2}}{q_n^{2\sigma_{m,0}-2}} \|u\|^2_{H^{\sigma_{m,0}}(J_m;X_0)},\\
E_2 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,2}}}{q_n^{2\sigma_{m,2}}} \|u\|^2_{H^{\sigma_{m,2}}(J_m;X_2)},\\
E_3 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,2}}}{q_n^{2\sigma_{m,2}}} \|u\|^2_{H^{\sigma_{m,2}}(J_m;X_2)},\\
E_4 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K}}{p_K^{2s_K+1}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,0}}}{q_n^{2\sigma_{m,0}}} \|u\|^2_{H^{\sigma_{m,0}}(J_m;X_0)},\\
E_5 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \max_{K\in\mathcal{T}_h} p_K \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,2}}}{q_n^{2\sigma_{m,2}}} \|u\|^2_{H^{\sigma_{m,2}}(J_m;X_2)},\\
E_6 &\lesssim \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-7}} \|u\|^2_{H^{s_K}(K)} \mathrm{d}t + \max_{K\in\mathcal{T}_h} p_K^3 \sum_{n=1}^{N} \sum_{J_m \supset I_n} \frac{\tau_n^{2\varrho_{m,2}}}{q_n^{2\sigma_{m,2}}} \|u\|^2_{H^{\sigma_{m,2}}(J_m;X_2)}.
\end{aligned}
$$

Using inverse inequalities and H^1-stability of $\Pi_h^{\mathbf{p}}$, we find that

$$
\begin{aligned}
E_7 + E_8 &= \sum_{K\in\mathcal{T}_h} \|u_0 - \Pi_h^{\mathbf{p}} z_\tau(0^+)\|^2_{H^1(K)} + \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F^{-1} \|\{\nabla(u_0 - \Pi_h^{\mathbf{p}} z_\tau(0^+)) \cdot n_F\}\|^2_{L^2(F)}\\
&\lesssim \sum_{K\in\mathcal{T}_h} \|u_0 - \Pi_h^{\mathbf{p}} u_0\|^2_{H^1(K)} + \|u_0 - z_\tau(0^+)\|^2_{H^1(\Omega)}\\
&\lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\tilde{t}_K-2}}{p_K^{2\tilde{s}_K-2}} \|u_0\|^2_{H^{\tilde{s}_K}(K)} + \|u_0 - z_\tau(0^+)\|^2_{H^1(\Omega)}.
\end{aligned}
$$

Since $z_\tau|_{I_1} \in \mathcal{Q}_{q_n}(H)$, we have $z_\tau(0^+) \in H_0^1(\Omega)$, so

$$
\begin{aligned}
E_9 &= \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \|[\![u_0 - \Pi_h^{\mathbf{p}} z_\tau(0^+)]\!]\|^2_{L^2(F)}\\
&= \sum_{F\in\mathcal{F}_h^{i,b}} \mu_F \|[\![u_0 - \Pi_h^{\mathbf{p}} u_0 + \Pi_h^{\mathbf{p}}(u_0 - z_\tau(0^+)) - (u_0 - z_\tau(0^+))]\!]\|^2_{L^2(F)}\\
&\lesssim \sum_{K\in\mathcal{T}_h} \frac{h_K^{2\tilde{t}_K-2}}{p_K^{2\tilde{s}_K-3}} \|u_0\|^2_{H^{\tilde{s}_K}(K)} + \max_{K\in\mathcal{T}_h} p_K \|u_0 - z_\tau(0^+)\|^2_{H^1(\Omega)}.
\end{aligned}
\tag{4.79}
$$

Poincaré's inequality and (4.75) then show that

$$\begin{aligned}\|u_0 - z_\tau(0^+)\|_{H^1(\Omega)}^2 &\lesssim \|u - z_\tau\|_{L^2(I_1;X_2)}\|u - z_\tau\|_{H^1(I_1;X_0)} + \frac{1}{\tau_1}\|u - z_\tau\|_{L^2(I_1;X_1)}^2 \\ &\lesssim \sum_{J_m \supset I_1} \frac{\tau_1^{2\varrho_{m,2}}}{q_1^{2\sigma_{m,2}}}\|u\|_{H^{\sigma_{m,2}}(J_m;X_2)}^2 + \frac{\tau_1^{2\varrho_{m,0}-2}}{q_1^{2\sigma_{m,0}-2}}\|u\|_{H^{\sigma_{m,0}}(J_m;X_0)}^2 \\ &\qquad + \sum_{J_m \supset I_1} \frac{\tau_1^{2\varrho_{m,1}-1}}{q_1^{2\sigma_{m,1}}}\|u\|_{H^{\sigma_{m,1}}(J_m;X_1)}^2.\end{aligned}$$

Since $\|\xi_h\|_h^2 \lesssim \sum_{i=1}^9 E_i$, the combination of the above bounds with the triangle inequality $\|u - u_h\|_h \le \|\xi_h\|_h + \|\psi_h\|_h$ completes the proof of (4.78). □

4.7 Numerical experiments

In the first experiment, we study the performance of the method on a fully nonlinear problem with strongly anisotropic diffusion coefficients, and observe optimal convergence rates for smooth solutions. In the second experiment, we show that the scheme gives exponential convergence rates when combining hp-refinement and τq-refinement, even for problems with low-regularity solutions.

4.7.1 First experiment

We examine the orders of convergence of the method for a problem with strongly anisotropic diffusion coefficients and a smooth solution. Let $\Omega = (0,1)^2$, $I = (0,1)$, let $b^\alpha \equiv 0$, $c^\alpha \equiv 0$ and let the diffusion coefficients a^α be defined by

$$a^\alpha \coloneqq \alpha \begin{pmatrix} 1 & 1/40 \\ 1/40 & 1/800 \end{pmatrix} \alpha^\top, \quad \alpha \in \Lambda \coloneqq \mathrm{SO}(2), \tag{4.80}$$

where $\mathrm{SO}(2)$ is the special orthogonal group of 2×2 matrices. For $\omega = 1$, $\lambda = 0$, it is found that the Cordes condition (4.10) holds with $\varepsilon \approx 1.25 \times 10^{-3}$. We choose f^α so that the exact solution is $u = (1 - \mathrm{e}^{-t})\,\mathrm{e}^{xy}\sin(\pi x)\sin(\pi y)$.

The numerical scheme (4.43) is applied on a sequence of uniform meshes obtained by regular subdivision of Ω into quadrilateral elements of width $h = 2^{-k}$, $1 \le k \le 5$. The corresponding time partitions $\mathcal{J}_\tau$ are obtained by regular subdivision of the time interval $(0,1)$ into intervals of length $\tau = 2^{-k+1}$, $1 \le k \le 5$. The finite element spaces $V_{h,\mathbf{p}}^{\tau,\mathbf{q}}$ are defined using polynomials of total degree p in space and degree $q = p - 1$ in time, with $p \in \{2,3,4\}$. We set the penalty parameters $c_\mu = c_\eta = 5/2$ and $\sigma = 1$ in (4.48). The semismooth Newton method analysed in section 3.6 is used to compute the numerical solution at each timestep.

In order to study the accuracy of the method, we measure the error in the norm $|||\cdot|||_h$

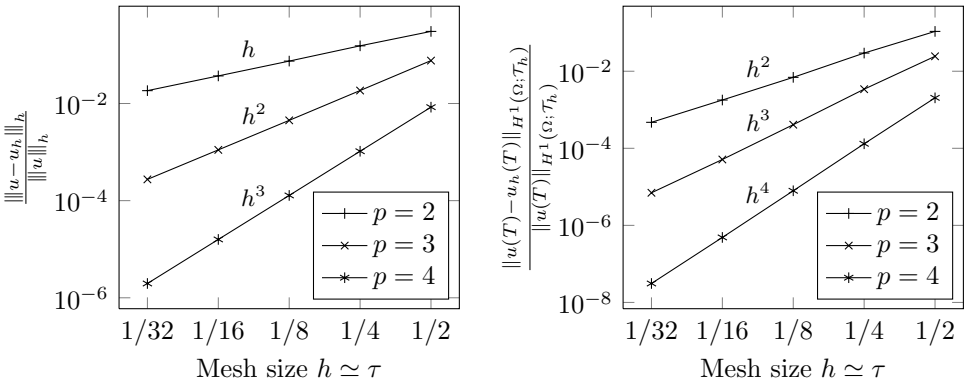

FIGURE 4.1: *Relative errors in approximating the solution of the problem of section* 4.7.1 *using uniform meshes and time partitions with* $\tau \simeq h$ *and* $p = q+1$. *It is seen that the optimal convergence rates* $|||u - u_h|||_h \simeq h^{p-1}+\tau^q$ *are achieved. The final time error, as measured in the broken* H^1*-norm, also converges with the optimal rate* $\|u(T) - u_h(T)\|_{H^1(\Omega;\mathcal{T}_h)} \simeq h^p$.

defined by

$$|||v|||_h^2 \coloneqq \sum_{n=1}^{N} \int_{I_n} \sum_{K\in\mathcal{T}_h} \left[\omega^2 \|\partial_t v\|_{L^2(K)}^2 + \|v\|_{H^2(K)}^2\right] \mathrm{d}t. \tag{4.81}$$

Figure 4.1 presents the global relative errors achieved by the method, where it is seen that the optimal orders of convergence $|||u - u_h|||_h \simeq h^{p-1} + \tau^q$ are achieved. The relative end-time errors, naturally measured in the broken H^1-norm, are also presented in Figure 4.1, which shows the optimal convergence rates $\|u(T)-u_h(T)\|_{H^1(\Omega;\mathcal{T}_h)} \simeq h^p$. These results show that the method can deliver high accuracy despite the strong anisotropy of the problem and the very small value of the constant ε appearing in the Cordes condition.

4.7.2 Second experiment

In section 4.6.2, we considered error bounds for solutions with low regularity. The significance of these results stems from the fact that the solutions of many parabolic HJB equations possess limited regularity as a result of early-time singularities induced by the initial datum.

This difficulty appears even in the simplest special case of the HJB equation (4.7), namely the heat equation: indeed, consider $\partial_t u = \Delta u$ in $\Omega \times (0,T)$, $\Omega = (0,1)^2$, with homogeneous lateral boundary condition $u = 0$ on $\partial\Omega \times (0,T)$ and initial datum $u_0(x,y) \coloneqq x\,(1-x)\sin(\pi y)$. Then, the solution is

$$u(x,y,t) = \frac{4}{\pi^3} \sum_{k=1}^{\infty} \frac{1-(-1)^k}{k^3} \exp(-(k^2+1)\,\pi^2 t) \sin(k\pi x) \sin(\pi y). \tag{4.82}$$

It can be shown that for sufficiently small $t > 0$ and nonnegative integers σ and ℓ such

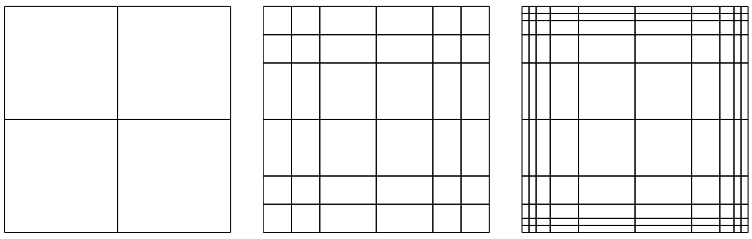

FIGURE 4.2: *Geometrically-graded spatial meshes used in conjunction with the geometrically-graded temporal meshes for the problem of section* 4.7.2. *From left to right, the meshes are those used for the first, third and fifth computations. The corresponding number of spatial degrees of freedom* DoF_x *are respectively* 100*,* 1128*, and* 3980*.*

that $2\sigma + \ell \geq 3$, we have $\|\partial_t^\sigma u\|_{X_\ell}^2 \simeq t^{-(2\sigma+\ell-5/2)}$, with the constants of these lower and upper bounds both depending on σ and ℓ, but not on t. Therefore, $u \notin H^1(I;H)$, rather $u \in H^{7/4-\delta}(I;L^2(\Omega)) \cap H^{5/4-\delta}(I;H_0^1(\Omega)) \cap H^{3/4-\delta}(I;H)$ for arbitrarily small $\delta > 0$. It is noted that a linear problem is chosen here so that the solution may be found explicity through (4.82). Nevertheless, this example exhibits many features that are typical of more general parabolic problems, so that the following results remain relevant to more general HJB equations.

Despite the limited regularity of the solution, accurate results can be obtained by using geometrically-graded time partitions with varying temporal polynomial degrees; see [67]. A combination of τq-refinement in time and hp-refinement in space can lead to a rate

$$|||u - u_h|||_h \lesssim \exp(-c_1 \sqrt[3]{\mathrm{DoF}_x}) + \exp(-c_2 \sqrt{\mathrm{DoF}_\tau}), \tag{4.83}$$

where $\mathrm{DoF}_x \coloneqq \dim V_{h,\mathbf{p}}$, where $\mathrm{DoF}_\tau = \sum_{n=1}^N (q_n + 1)$ is the number of degrees of freedom of the temporal finite element space, and where c_1 and c_2 are positive constants. We give here an experimental confirmation of these expectations.

The method is applied on a sequence of geometrically-graded partitions $\{\mathcal{J}_\tau\}_\tau$ constructed as follows. Let $T = 0.05$, and let $t_n = \sigma^{N-n} T$ for $n = 1, \ldots, N$, for a chosen $\sigma \in (0,1)$, and $N = 2, \ldots, 6$. As suggested in [67], we choose $\sigma = 0.2$. The temporal polynomial degrees are linearly increasing with n, with $q_n \coloneqq n + 1$. We choose T to be small, because in practice it is natural to use τq-refinement on a small initial time segment, and then apply uniform or spectral refinement on the remaining time interval, see [67]. Starting with a partition of Ω into four quadrilateral elements, for each successive computation, we refine the meshes geometrically towards the boundary, thereby yielding the meshes shown in Figure 4.2. The polynomial degrees $p_K \geq 3$ are chosen to be linearly increasing away from the boundary. Figure 4.3 presents the resulting errors in the norms $|||\cdot|||_h$ and $\|\cdot\|_{L^2(I;H^1(\Omega;\mathcal{T}_h))}$, plotted against $\sqrt[3]{\mathrm{DoF}_x}$ and $\sqrt{\mathrm{DoF}_\tau}$. It is found that the convergence rates of (4.83) are attained, with higher accuracies being achieved in lower-order norms. These results show the efficiency of the method for problems with limited regularity.

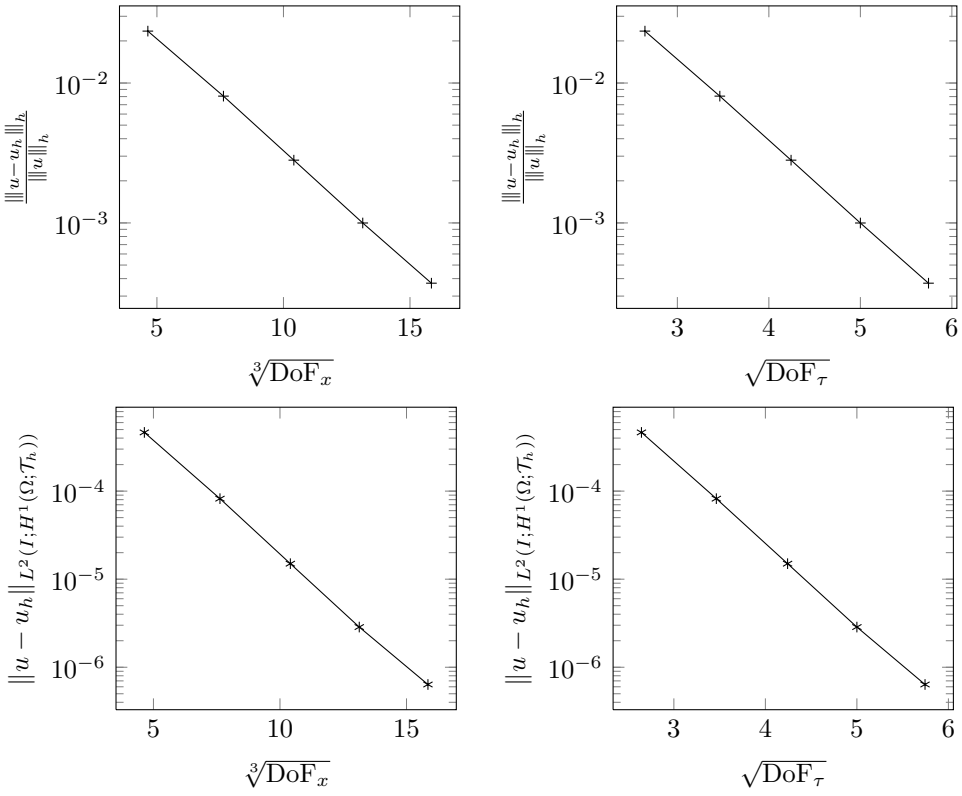

FIGURE 4.3: *Exponential convergence rates under hp-τq refinement for the problem of section* 4.7.2. *The errors in the norms* $|||\cdot|||_h$ *and* $\|\cdot\|_{L^2(I;H^1(\Omega;\mathcal{T}_h))}$ *are plotted against* $\sqrt[3]{\mathrm{DoF}_x}$ *and* $\sqrt{\mathrm{DoF}_\tau}$, *where* DoF_x *is the number of spatial degrees of freedom and* DoF_τ *is the number of temporal degrees of freedom. Exponential convergence rates of the form of* (4.83) *are confirmed.*

Chapter 5

Nonoverlapping domain decomposition preconditioners

It was shown in section 3.6 that the discretised HJB equation can be solved efficiently with a semismooth Newton method. This algorithm defines a sequence of linear problems involving bilinear forms that are stable in the H^2-type norm $\|\cdot\|_{h,1}$. As a result, the condition numbers of the resulting linear systems are typically large, thereby limiting the performance of many iterative solution algorithms.

The purpose of this chapter, which is based on our paper [69], is to study a class of preconditioners for accelerating the iterative solution of these discrete linear systems. The approach we adopt consists of developing preconditioners for the model problem of finding $u_h \in V_{h,\mathbf{p}}$ such that

$$B_{h,1}(u_h, v_h) = \ell_h(v_h) \quad \forall\, v_h \in V_{h,\mathbf{p}}, \tag{5.1}$$

where $\ell_h \colon V_{h,\mathbf{p}} \to \mathbb{R}$ is a bounded linear functional on $V_{h,\mathbf{p}}$. Our choice of model problem essentially relies on the fact that the discrete problems of the semismooth Newton method are uniformly stable in the norm $\|\cdot\|_{h,1}$ appearing in the bound (3.63), which is equivalent to the norm induced by $B_{h,1}$, as shown by Lemma 3.7. Of course, many generalisations are possible, such as preconditioners based on the bilinear forms $B_{h,\theta}$ for $\theta \in [0,1]$, for which we expect similar results. The advantage of this approach is that the preconditioners can be assembled once, and re-used at each iterative step of the semismooth Newton method.

The class of preconditioners we consider here comprise overlapping and nonoverlapping domain decomposition methods, which have been successfully developed for a range of applications of DGFEM by many authors [3, 4, 5, 6, 34, 35, 55]. In order to solve a problem on a fine mesh $\mathcal{T}_h$, these methods combine a coarse space solver, defined on a coarse mesh $\mathcal{T}_H$, with local fine mesh solvers, defined on a subdomain decomposition $\mathcal{T}_S$ of the domain Ω. The discontinuous nature of the finite element space leads to a significant flexibility in the choice of the decomposition $\mathcal{T}_S$, which can either be overlapping or nonoverlapping.

As explained in the above references, these methods possess many advantages in terms of simplicity and applicability, as they allow very general choices of basis functions, non-matching meshes and varying element shapes, and are naturally suited for parallelisation. This flexibility makes this class of methods a natural choice for a first study of preconditioners for the problems at hand. Moreover, it has been pointed out by various authors, such as Lasser and Toselli in [55, p. 1235], that the nonoverlapping methods feature reduced inter-subdomain communication burdens, thus representing a key advantage for scalability in parallel computations.

It is of practical interest for applications to determine the influence of the parameters of the preconditioner on the spectral bounds. For problems involving H^1-type norms, such as interior penalty methods for diffusion problems, nonoverlapping additive Schwarz preconditioners for h-version methods [34] lead to condition numbers of order $1 + H/h$, and overlapping methods lead to a condition number of order $1+H/\delta$, where H is the coarse mesh size, and δ is the size of the overlap of the subdomains. Antonietti and Houston considered the case of nonoverlapping methods for hp-version DGFEM in [5], and they showed a bound of order $1+p^2H/h$; however, their numerical experiments lead to a conjecture of the improved bound of $1 + p^2H/qh$, where q is the coarse space polynomial degree.

For problems involving H^2-type norms, earlier results covered only the case of h-version methods. In [35], Feng and Karakashian considered nonoverlapping preconditioners for h-version discretizations of the biharmonic equation, which result in condition numbers of order $1 + H^3/h^3$. We also mention the related work by Brenner and Wang in [18] for C^0-interior penalty methods for fourth order problems, where a bound of order $1 + H^3/\delta^3$ was derived for an overlapping additive Schwarz method.

As can be seen from the theoretical analysis in the above references, the effectiveness of the preconditioner depends in an essential way on the approximation properties between the coarse and fine spaces. In the analysis of h-version DGFEM, it is sufficient to consider low-order projection operators from the fine space to the coarse space; for example, coarse element mean-value projections are employed in [34] and local first-order elliptic projections are used in [35]. However, low-order projections lead to suboptimal bounds for the condition number in the case of hp-version DGFEM. Therefore, as we explain below, a key contribution of this work is an original high-order approximation result between coarse and fine spaces that is of optimal order in both the mesh sizes and the polynomial degrees; this result thus represents the main ingredient for the sharp analysis of the condition number bounds.

Summary of contributions. We consider the discontinuous finite element space $V_{h,\mathbf{p}}$ of degree p over a fine mesh $\mathcal{T}_h$ on a convex polytope $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, and we equip $V_{h,\mathbf{p}}$ with the H^2-type norm $\|\cdot\|_{h,1}$. where $\mu_F \simeq p^2/h$ and $\eta_F \simeq p^6/h^3$, and where, for simplicity of exposition, we momentarily assume quasi-uniformity of the mesh sizes and polynomial degrees; this assumption is relaxed throughout this chapter.

The central result of this chapter is a sharp bound for the condition number of the preconditioned system obtained by applying a nonoverlapping domain decomposition preconditioner to the model problem (5.1). The bound that we obtain is

$$\kappa(\mathbf{PB}) \lesssim 1 + \frac{p^2 H}{q\, h} + \frac{p^6\, H^3}{q^3\, h^3}, \tag{5.2}$$

where $\mathbf{B}$ denotes the matrix representing the bilinear form in (5.1), $\mathbf{P}$ denotes the matrix of the preconditioner, and $\kappa(\mathbf{PB})$ denotes the condition number of the preconditioned system. Importantly, the bound (5.2) is sharp in both the mesh sizes and the polynomial degrees.

The central original result underpinning our analysis is Theorem 5.5 of section 5.1, which shows that for any $v_h \in V_{h,\mathbf{p}}$, there exists a function $v \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$\|v_h - v\|_{L^2(\Omega)} + \frac{h}{p}\|v_h - v\|_{H^1(\Omega;\mathcal{T}_h)} \lesssim \frac{h^2}{p^2}|v_h|_{\mathrm{J}}, \quad \|v\|_{H^2(\Omega)} \lesssim \|v_h\|_{h,1}. \tag{5.3}$$

Although there does not appear to be any comparable result in the literature, this result is a natural converse to classical direct approximation theory, since, here, the nonsmooth function from the discrete space $V_{h,\mathbf{p}}$ is approximated by a smoother function from an infinite dimensional space. It can be interpreted as a precise form of the statement that the norm $\|\cdot\|_{h,1}$ renders $V_{h,\mathbf{p}}$ close to $H^2(\Omega) \cap H_0^1(\Omega)$.

It follows almost immediately that there exists a function v_H in the coarse space $V_{H,\mathbf{q}}$, of polynomials of degree q on $\mathcal{T}_H$, such that

$$\|v_h - v_H\|_{H^k(\Omega;\mathcal{T}_h)} \lesssim \frac{H^{2-k}}{q^{2-k}}\|v_h\|_{h,1}, \quad k \in \{0,1,2\}, \tag{5.4}$$

thus yielding an approximation result for $V_{h,\mathbf{p}}$ by $V_{H,\mathbf{q}}$ that is optimal with respect to both the mesh size and the polynomial degree. The above approximation results are used to show the bound (5.2).

The first numerical experiment, in section 5.4.1, confirms that (5.2) is sharp with respect to the orders in the polynomial degrees. In the experiment of section 5.4.2, we find that nonoverlapping methods are efficient and competitive with respect to classical overlapping methods. Although the bound (5.2) relates to the model problem (5.1) rather than the linear systems encountered in the semismooth Newton method, we show in the experiment of section 5.4.3, that the preconditioners retain their robustness and efficiency under h-refinement in these more challenging applications to nonsymmetric, fully nonlinear Hamilton–Jacobi–Bellman equations, thereby yielding effective solvers for these problems.

5.1 Approximation of discontinuous functions

Using the definitions of section 2.2, it is assumed henceforth that the parameters μ_F and η_F appearing in (2.23) are given by

$$\mu_F \coloneqq c_\mu \frac{\tilde{p}_F^2}{\tilde{h}_F}, \quad \eta_F \coloneqq c_\eta \frac{\tilde{p}_F^6}{\tilde{h}_F^3} \quad \forall\, F \in \mathcal{F}_h^{i,b}, \tag{5.5}$$

where c_μ and c_η are fixed positive constants independent of h and $\mathbf{p}$. Recall Lemmas 2.5 and 3.7, which establish in the present context that for c_μ and c_η appropriately chosen,

$$\|v_h\|^2_{H^2(\Omega;\mathcal{T}_h)} + |v_h|^2_{\mathrm{J}} \simeq \|v_h\|^2_{h,1} \simeq B_{h,1}(v_h, v_h) \quad \forall\, v_h \in V_{h,\mathbf{p}}. \tag{5.6}$$

In order to simplify the presentation, we will assume throughout this chapter that the parameter $\lambda = 0$ in the definition of $B_{h,1}$, although this restriction is not essential.

As explained above, an optimal analysis of the spectral bounds for the class of preconditioners to be considered in section 5.2 rests upon the optimality of approximation properties between coarse and fine spaces. Therefore, in this section, we first determine how closely a function in $V_{h,\mathbf{p}}$ can be approximated by functions in $H^2(\Omega) \cap H_0^1(\Omega)$. This leads to an approximation result for functions in $V_{h,\mathbf{p}}$ by functions in $V_{H,\mathbf{q}}$ that is of optimal order in both the coarse mesh size and polynomial degree.

Lifting operators. Let $\mathbf{V}_{h,\mathbf{p}}$ denote the space of d-dimensional vector fields with components in $V_{h,\mathbf{p}}$. Let $\mathbf{r}_h \colon L^2(\mathcal{F}_h^{i,b}) \to \mathbf{V}_{h,\mathbf{p}}$ and $r_h \colon L^2(\mathcal{F}_h^i) \to V_{h,\mathbf{p}}$ be defined by

$$\begin{aligned} \sum_{K\in\mathcal{T}_h} \langle \mathbf{r}_h(w), \mathbf{v}_h\rangle_K &= \sum_{F\in\mathcal{F}_h^{i,b}} \langle w, \{\mathbf{v}_h \cdot n_F\}\rangle_F \quad \forall\, \mathbf{v}_h \in \mathbf{V}_{h,\mathbf{p}}, \\ \sum_{K\in\mathcal{T}_h} \langle r_h(w), v_h\rangle_K &= \sum_{F\in\mathcal{F}_h^{i}} \langle w, \{v_h\}\rangle_F \qquad \forall\, v_h \in V_{h,\mathbf{p}}. \end{aligned} \tag{5.7}$$

The following result is well-known; for instance, see [5] for a proof.

Lemma 5.1. *Let Ω be a bounded Lipschitz domain and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of meshes satisfying* (2.11), (2.12) *and* (2.13). *Then, the lifting operators satisfy the following bounds:*

$$\|\mathbf{r}_h(w)\|^2_{L^2(\Omega)} \lesssim \sum_{F\in\mathcal{F}_h^{i,b}} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|w\|^2_{L^2(F)} \qquad \forall\, w \in L^2(\mathcal{F}_h^{i,b}), \tag{5.8a}$$

$$\|r_h(w)\|^2_{L^2(\Omega)} \lesssim \sum_{F\in\mathcal{F}_h^{i}} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|w\|^2_{L^2(F)} \qquad \forall\, w \in L^2(\mathcal{F}_h^{i}). \tag{5.8b}$$

For $v_h \in V_{h,\mathbf{p}}$ and $\mathbf{v}_h \in \mathbf{V}_{h,\mathbf{p}}$, define $G_h(v_h) \in \mathbf{V}_{h,\mathbf{p}}$ and $D_h(\mathbf{v}_h) \in V_{h,\mathbf{p}}$ by

$$G_h(v_h) \coloneqq \nabla v_h - \mathbf{r}_h(\llbracket v_h \rrbracket), \quad D_h(\mathbf{v}_h) \coloneqq \operatorname{div} \mathbf{v}_h - r_h(\llbracket \mathbf{v}_h \cdot n_F \rrbracket). \tag{5.9}$$

The following result was first shown in [69].

Lemma 5.2. *Let Ω be a bounded Lipschitz polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11), (2.12) *and* (2.13). *Let η_F and μ_F satisfy* (5.5) *for all $F \in \mathcal{F}_h^{i,b}$. Then, for any $v_h \in V_{h,\mathbf{p}}$, we have*

$$\sum_{K\in\mathcal{T}_h} \frac{p_K^4}{h_K^2} \|\mathbf{r}_h(\llbracket v_h \rrbracket)\|_{L^2(K)}^2 \lesssim |v_h|_{\mathrm{J}}^2, \tag{5.10a}$$

$$\sum_{K\in\mathcal{T}_h} |\mathbf{r}_h(\llbracket v_h \rrbracket)|_{H^1(K)}^2 + \sum_{F\in\mathcal{F}_h^i} \mu_F \|\llbracket \mathbf{r}_h(\llbracket v_h \rrbracket) \cdot n_F \rrbracket\|_{L^2(F)}^2 \lesssim |v_h|_{\mathrm{J}}^2. \tag{5.10b}$$

Proof. Define the piecewise constant function p^4/h^2 by $p^4/h^2|_K = p_K^4/h_K^2$ for each element $K \in \mathcal{T}_h$; we can then view the function $p^4/h^2\, \mathbf{r}_h(\llbracket v_h \rrbracket)$ as a function in $\mathbf{V}_{h,\mathbf{p}}$. So, the definition of the lifting operator gives

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \frac{p_K^4}{h_K^2} \|\mathbf{r}_h(\llbracket v_h \rrbracket)\|_{L^2(K)}^2 &= \sum_{F\in\mathcal{F}_h^{i,b}} \langle \{p^4/h^2 \mathbf{r}_h(\llbracket v_h \rrbracket) \cdot n_F\}, \llbracket v_h \rrbracket \rangle \\ &\lesssim \sqrt{\sum_{F\in\mathcal{F}_h^{i,b}} \frac{\tilde{h}_F^3}{\tilde{p}_F^6} \frac{\tilde{p}_F^8}{\tilde{h}_F^4} \|\mathbf{r}_h(\llbracket v_h \rrbracket)\|_{L^2(F)}^2}\, |v_h|_{\mathrm{J}}.\end{aligned}$$

The trace and inverse inequalities then yield

$$\sum_{K\in\mathcal{T}_h} \frac{p_K^4}{h_K^2} \|\mathbf{r}_h(\llbracket v_h \rrbracket)\|_{L^2(K)}^2 \lesssim \sqrt{\sum_{K\in\mathcal{T}_h} \frac{p_K^4}{h_K^2} \|\mathbf{r}_h(\llbracket v_h \rrbracket)\|_{L^2(K)}^2}\, |v_h|_{\mathrm{J}},$$

which implies (5.10a). The bound (5.10b) then follows from (5.10a) as a result of the trace and inverse inequalities. □

Corollary 5.3. *Under the hypotheses of Lemma* 5.2, *every $v_h \in V_{h,\mathbf{p}}$ satisfies*

$$\sum_{K\in\mathcal{T}_h} |G_h(v_h)|_{H^1(K)}^2 + \sum_{F\in\mathcal{F}_h^i} \mu_F \|\llbracket G_h(v_h) \cdot n_F \rrbracket\|_{L^2(F)}^2 \lesssim \|v_h\|_{h,1}^2, \tag{5.11}$$

$$\|D_h(G_h(v_h))\|_{L^2(\Omega)} \lesssim \|v_h\|_{h,1}. \tag{5.12}$$

Proof. Inequality (5.11) is an easy consequence of the definition of G_h in (5.9) and of Lemma 5.2. To show (5.12), we consider

$$D_h(G_h(v_h)) = \Delta v_h - \operatorname{div} \mathbf{r}_h(\llbracket v_h \rrbracket) - r_h(\llbracket \nabla v_h \cdot n_F \rrbracket) + r_h(\llbracket \mathbf{r}_h(\llbracket v_h \rrbracket) \cdot n_F \rrbracket). \tag{5.13}$$

In view of (5.8b), it is apparent that the L^2-norms of the first and third terms on the right-hand side of (5.13) are bounded by $\|v_h\|_{h,1}$, whilst the bounds on the L^2-norms of the second and fourth terms follow from (5.10b). □

Approximation by H^2-regular functions. The first step towards the aforementioned approximation result consists of the discrete analogue of the orthogonality of Helmholtz decompositions.

Lemma 5.4. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded Lipschitz polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying* (2.11), (2.12) *and* (2.13). *If μ_F and η_F satisfy* (5.5) *for every face $F \in \mathcal{F}_h^{i,b}$, then, for any $v_h \in V_{h,\mathbf{p}}$ and any $\psi \in H^1(\Omega)^{2d-3}$, we have*

$$\left|\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x\right| + \left|\int_\Omega \nabla v_h\cdot\operatorname{curl}\psi\,\mathrm{d}x\right| \lesssim \max_{K\in\mathcal{T}_h}\frac{h_K}{p_K^{3/2}}|v_h|_{\mathrm{J}}\|\psi\|_{H^1(\Omega)}. \tag{5.14}$$

Proof. It follows from (5.10a) that $\|\nabla v_h - G_h(v_h)\|_{L^2(\Omega)} \lesssim \max_K h_K/p_K^2|v_h|_{\mathrm{J}}$, so it is enough to show that (5.14) is satisfied by $G_h(v_h)$. Consider momentarily $\psi \in H^2(\Omega)^{2d-3}$; then, integration by parts yields

$$\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x = \sum_{F\in\mathcal{F}_h^{i,b}}\langle[\![v_h]\!],\{\operatorname{curl}\psi\cdot n_F\}\rangle_F - \sum_{K\in\mathcal{T}_h}\langle\mathbf{r}_h([\![v_h]\!]),\operatorname{curl}\psi\rangle_K.$$

Therefore, the definitions of the lifting operators $\mathbf{r}_h$ and r_h imply that

$$\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x = \sum_{F\in\mathcal{F}_h^{i,b}}\langle[\![v_h]\!],\{\operatorname{curl}(\psi-\psi_h)\cdot n_F\}\rangle_F - \sum_{K\in\mathcal{T}_h}\langle\mathbf{r}_h([\![v_h]\!]),\operatorname{curl}(\psi-\psi_h)\rangle_K$$

for any $\psi_h \in V_{h,\mathbf{p}}$ if $d=2$, or $\psi_h \in \mathbf{V}_{h,\mathbf{p}}$ if $d=3$. Thus, if $\psi \in H^2(\Omega)^{2d-3}$, it is seen from the approximation bounds of Appendix C and from the lifting bound (5.10a) that

$$\left|\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x\right| \lesssim \max_{K\in\mathcal{T}_h}\frac{h_K^2}{p_K^3}|v_h|_{\mathrm{J}}\|\psi\|_{H^2(\Omega)}. \tag{5.15}$$

Now, let $\psi \in H^1(\Omega)^{2d-3}$. We apply [1, Thm. 5.33] to the components of ψ: for each $\varepsilon > 0$, there exists a $\psi_\varepsilon \in C^\infty(\mathbb{R}^d)^{2d-3}$ such that

$$\|\psi-\psi_\varepsilon\|_{L^2(\Omega)} + \varepsilon\|\psi-\psi_\varepsilon\|_{H^1(\Omega)} \lesssim \varepsilon|\psi|_{H^1(\Omega)}, \tag{5.16a}$$

$$\|\psi_\varepsilon\|_{H^2(\Omega)} \lesssim \varepsilon^{-1}\|\psi\|_{H^1(\Omega)}, \tag{5.16b}$$

where, importantly, the constants in (5.16) do not depend on ε. Define $\phi_\varepsilon := \psi - \psi_\varepsilon$, so

that

$$\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x=\int_\Omega G(v_h)\cdot\operatorname{curl}\psi_\varepsilon\,\mathrm{d}x+\int_\Omega G(v_h)\cdot\operatorname{curl}\phi_\varepsilon\,\mathrm{d}x.$$

The bounds (5.15) and (5.16b) show that

$$\left|\int_\Omega G(v_h)\cdot\operatorname{curl}\psi_\varepsilon\,\mathrm{d}x\right|\lesssim\varepsilon^{-1}\max_{K\in\mathcal{T}_h}\frac{h_K^2}{p_K^3}|v_h|_{\mathrm{J}}\|\psi\|_{H^1(\Omega)}. \tag{5.17}$$

Integration by parts yields

$$\int_\Omega G(v_h)\cdot\operatorname{curl}\phi_\varepsilon\,\mathrm{d}x=\sum_{F\in\mathcal{F}_h^{i,b}}\langle[\![\nabla v_h\times n_F]\!],\phi_\varepsilon\rangle_F-\sum_{K\in\mathcal{T}_h}\langle\mathbf{r}_h([\![v_h]\!]),\operatorname{curl}\phi_\varepsilon\rangle_K.$$

Lemma 5.2 and (5.16a) imply that

$$\sum_{K\in\mathcal{T}_h}|\langle\mathbf{r}_h([\![v_h]\!]),\operatorname{curl}\phi_\varepsilon\rangle_K|\lesssim\max_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}|v_h|_{\mathrm{J}}\|\psi\|_{H^1(\Omega)}. \tag{5.18}$$

Recall the continuous trace inequality [58]: for an element K and a face $F\subset\partial K$,

$$\|\phi_\varepsilon\|_{L^2(F)}^2\lesssim|\phi_\varepsilon|_{H^1(K)}\|\phi_\varepsilon\|_{L^2(K)}+\frac{1}{h_K}\|\phi_\varepsilon\|_{L^2(K)}^2\lesssim\frac{h_K}{p_K^2}|\phi_\varepsilon|_{H^1(K)}^2+\frac{p_K^2}{h_K}\|\phi_\varepsilon\|_{L^2(K)}^2.$$

Therefore, the fact that $\mu_F=c_\mu\,\tilde{p}_F^2/\tilde{h}_F$ leads to

$$\begin{aligned}\sum_{F\in\mathcal{F}_h^{i,b}}|\langle[\![\nabla v_h\times n_F]\!],\phi_\varepsilon\rangle_F|&\lesssim\sqrt{\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{p_K^4}|\phi_\varepsilon|_{H^1(K)}^2+\|\phi_\varepsilon\|_{L^2(K)}^2}\,|v_h|_{\mathrm{J}}\\&\lesssim\left(\max_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}|\phi_\varepsilon|_{H^1(\Omega)}+\|\phi_\varepsilon\|_{L^2(\Omega)}\right)|v_h|_{\mathrm{J}},\end{aligned}$$

where we have used the identity $|[\![\nabla v_h\times n_F]\!]|=|[\![\nabla_{\mathrm{T}}\,v_h]\!]|$ for each face F, because $\nabla_{\mathrm{T}}\,v_h$ is the component of ∇v_h that is orthogonal to n_F. Therefore, we deduce from (5.16a) and (5.18) that

$$\left|\int_\Omega G(v_h)\cdot\operatorname{curl}\phi_\varepsilon\,\mathrm{d}x\right|\lesssim\left(\max_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}+\varepsilon\right)|v_h|_{\mathrm{J}}\|\psi\|_{H^1(\Omega)}. \tag{5.19}$$

Combining (5.17) and (5.19) yields

$$\left|\int_\Omega G(v_h)\cdot\operatorname{curl}\psi\,\mathrm{d}x\right|\lesssim\left(\varepsilon^{-1}\max_{K\in\mathcal{T}_h}\frac{h_K^2}{p_K^3}+\max_{K\in\mathcal{T}_h}\frac{h_K}{p_K^2}+\varepsilon\right)|v_h|_{\mathrm{J}}\|\psi\|_{H^1(\Omega)}.$$

The bound (5.14) is then obtained by taking $\varepsilon\coloneqq\max_{K\in\mathcal{T}_h}h_K/p_K^{3/2}$. □

The following original approximation result was first shown in [69].

Theorem 5.5. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded convex polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of meshes satisfying (2.11), (2.12) and (2.13). Let μ_F and η_F satisfy (5.5) for every face $F \in \mathcal{F}_h^{i,b}$. For a given $v_h \in V_{h,\mathbf{p}}$, let $v \in H^2(\Omega) \cap H_0^1(\Omega)$ be the unique solution of the boundary-value problem*

$$\Delta v = D_h(G_h(v_h)) \quad \text{in } \Omega, \tag{5.20a}$$

$$v = 0 \quad \text{on } \partial\Omega. \tag{5.20b}$$

Then, the approximation v to v_h satisfies

$$\|v_h - v\|_{L^2(\Omega)} + \max_{K\in\mathcal{T}_h} \frac{h_K}{p_K} \|v_h - v\|_{H^1(\Omega;\mathcal{T}_h)} \lesssim \max_{K\in\mathcal{T}_h} \frac{h_K^2}{p_K^2} |v_h|_{\mathrm{J}}, \tag{5.21a}$$

$$\|v\|_{H^2(\Omega)} \lesssim \|v_h\|_{h,1}. \tag{5.21b}$$

Remark 5.1. The above result is nearly optimal in the sense that only the jump seminorm $|v_h|_{\mathrm{J}}$ appears on the right-hand side of the error bound (5.21a), and that the correct orders of convergence are established.

Proof. Note that convexity of Ω implies that v is well-defined, see [40], and that (5.21b) holds as a result of Corollary 5.3. First, we show that for any $p \in H^k(\Omega) \cap H_0^1(\Omega)$, $k \in \{1,2\}$, we have

$$\left| \int_\Omega (\nabla v - G_h(v_h)) \cdot \nabla p \,\mathrm{d}x \right| \lesssim \max_{K\in\mathcal{T}_h} \frac{h_K^k}{p_K^k} |v_h|_{\mathrm{J}} \|p\|_{H^k(\Omega)}. \tag{5.22}$$

Indeed, since v solves (5.20), integration by parts yields

$$\int_\Omega (\nabla v - G_h(v_h)) \cdot \nabla p \,\mathrm{d}x = \sum_{K\in\mathcal{T}_h} \langle r_h([\![G_h(v_h)\cdot n_F]\!]), p\rangle_K - \sum_{F\in\mathcal{F}_h^i} \langle [\![G_h(v_h)\cdot n_F]\!], \{p\}\rangle_F.$$

Then, the definition of the lifting operator gives

$$\begin{aligned}\int_\Omega (\nabla v - G_h(v_h)) \cdot \nabla p \,\mathrm{d}x = &\sum_{K\in\mathcal{T}_h} \langle r_h([\![G_h(v_h)\cdot n_F]\!]), p - p_h\rangle_K \\ &- \sum_{F\in\mathcal{F}_h^i} \langle [\![G_h(v_h)\cdot n_F]\!], \{p - p_h\}\rangle_F \quad \forall\, p_h \in V_{h,\mathbf{p}}.\end{aligned} \tag{5.23}$$

Recalling that $p_K \geq 1$ for each element K, it is then seen that (5.22) follows from Corollary 5.3 and from the approximation bounds of Appendix C.

The remainder of the proof makes use of Helmholtz decompositions of vector fields [39]: for any $\mathbf{v} \in L^2(\Omega)^d$, there exist $p \in H_0^1(\Omega)$ and $\psi \in H^1(\Omega)^{2d-3}$, such that

$$\mathbf{v} = \nabla p + \operatorname{curl} \psi \quad \text{in } \Omega. \tag{5.24}$$

Indeed, $p \in H^1_0(\Omega)$ is defined by

$$\int_\Omega \nabla p \cdot \nabla q \,\mathrm{d}x = \int_\Omega \mathbf{v} \cdot \nabla q \,\mathrm{d}x \quad \forall\, q \in H^1_0(\Omega). \tag{5.25}$$

Then, $\mathbf{v} - \nabla p$ is divergence free, thus $\langle (\mathbf{v} - \nabla p) \cdot \mathbf{n}, 1 \rangle_{\partial\Omega} = 0$, where $\mathbf{n}$ is the unit outward normal on $\partial\Omega$. Since the convex domain Ω has a connected boundary, it follows from [39, Thms. 3.1 & 3.4 pp. 37–45] that there exists a $\psi \in H^1(\Omega)^{2d-3}$ such that $\mathbf{v} = \nabla p + \operatorname{curl} \psi$. Moreover, ψ may be chosen so that $\|p\|_{H^1(\Omega)} + \|\psi\|_{H^1(\Omega)} \lesssim \|\mathbf{v}\|_{L^2(\Omega)}$ for some constant independent of $\mathbf{v}$. This is a consequence of the open mapping theorem and the facts that $\mathcal{V} \coloneqq \{\mathbf{v} \in L^2(\Omega)^d \colon \operatorname{div} \mathbf{v} = 0\}$ is a closed subspace of $L^2(\Omega)^d$, and that the mapping $\psi \mapsto \operatorname{curl} \psi$ is a surjective bounded linear mapping from $H^1(\Omega)^{2d-3}$ to $\mathcal{V}$.

Now, observe that $\|\nabla v_h - G_h(v_h)\|_{L^2(\Omega)} \lesssim \max_{K \in \mathcal{T}_h} h_K / p_K^2 \, |v_h|_{\mathrm{J}}$ by (5.10a), so it is enough to consider the $\|G(v_h) - \nabla v\|_{L^2(\Omega)}$ to bound $|v_h - v|_{H^1(\Omega;\mathcal{T}_h)}$. Let $p \in H^1_0(\Omega)$ and $\psi \in H^1(\Omega)^{2d-3}$ satisfy

$$\nabla v - G_h(v_h) = \nabla p + \operatorname{curl} \psi, \tag{5.26}$$

with $\|p\|_{H^1(\Omega)} + \|\psi\|_{H^1(\Omega)} \lesssim \|\nabla v - G_h(v_h)\|_{L^2(\Omega)}$. Then, noting that ∇v and $\operatorname{curl} \psi$ are orthogonal, it is deduced that

$$\|\nabla v - G_h(v_h)\|^2_{L^2(\Omega)} = \int_\Omega (\nabla v - G_h(v_h)) \cdot \nabla p \,\mathrm{d}x - \int_\Omega G_h(v_h) \cdot \operatorname{curl} \psi \,\mathrm{d}x. \tag{5.27}$$

Inequality (5.22) and the bound $\|p\|_{H^1(\Omega)} \lesssim \|\nabla v - G_h(v_h)\|_{L^2(\Omega)}$ give

$$\left| \int_\Omega (\nabla v - G_h(v_h)) \cdot \nabla p \,\mathrm{d}x \right| \lesssim \max_{K \in \mathcal{T}_h} \frac{h_K}{p_K} |v_h|_{\mathrm{J}} \|\nabla v - G_h(v_h)\|_{L^2(\Omega)}.$$

The bounds of Lemma 5.4 show that

$$\left| \int_\Omega G_h(v_h) \cdot \operatorname{curl} \psi \,\mathrm{d}x \right| \lesssim \max_{K \in \mathcal{T}_h} \frac{h_K}{p_K^{3/2}} |v_h|_{\mathrm{J}} \|\nabla v - G_h(v_h)\|_{L^2(\Omega)}.$$

Therefore, equation (5.27) and the above bounds yield

$$\|\nabla v - G_h(v_h)\|_{L^2(\Omega)} \lesssim \max_{K \in \mathcal{T}_h} \frac{h_K}{p_K} |v_h|_{\mathrm{J}}. \tag{5.28}$$

We now consider the error $\|v_h - v\|_{L^2(\Omega)}$. Since the domain Ω is convex, there is a unique function $z \in H^2(\Omega) \cap H^1_0(\Omega)$ that solves $-\Delta z = v_h - v$ in Ω, with $\|z\|_{H^2(\Omega)} \lesssim \|v_h - v\|_{L^2(\Omega)}$.

Then, it is found that

$$\|v_h - v\|_{L^2(\Omega)}^2 = \int_\Omega (G_h(v_h) - \nabla v) \cdot \nabla z \,\mathrm{d}x + \sum_{K\in\mathcal{T}_h} \langle \mathbf{r}_h([\![v_h]\!]), \nabla z\rangle_K - \sum_{F\in\mathcal{F}_h^{i,b}} \langle [\![v_h]\!], \{\nabla z \cdot n_F\}\rangle_F.$$

Applying the bound (5.22) to $z \in H^2(\Omega) \cap H_0^1(\Omega)$ gives

$$\left| \int_\Omega (G_h(v_h) - \nabla v) \cdot \nabla z \,\mathrm{d}x \right| \lesssim \max_{K\in\mathcal{T}_h} \frac{h_K^2}{p_K^2} |v_h|_{\mathrm{J}} \|v_h - v\|_{L^2(\Omega)}.$$

Letting z_h be the projection of z into $V_{h,\mathbf{p}}$ given by Theorem C.6, it is found that

$$\begin{aligned} &\left| \sum_{K\in\mathcal{T}_h} \langle \mathbf{r}_h([\![v_h]\!]), \nabla z\rangle_K - \sum_{F\in\mathcal{F}_h^{i,b}} \langle [\![v_h]\!], \{\nabla z \cdot n_F\}\rangle_F \right| \\ &= \left| \sum_{K\in\mathcal{T}_h} \langle \mathbf{r}_h([\![v_h]\!]), \nabla (z - z_h)\rangle_K - \sum_{F\in\mathcal{F}_h^{i,b}} \langle [\![v_h]\!], \{\nabla (z - z_h)\cdot n_F\}\rangle_F \right| \lesssim \max_K \frac{h_K^2}{p_K^3} |v_h|_{\mathrm{J}} \|v_h - v\|_{L^2(\Omega)}. \end{aligned}$$

Thus, we have shown that

$$\|v_h - v\|_{L^2(\Omega)} \lesssim \max_{K\in\mathcal{T}_h} \frac{h_K^2}{p_K^2} |v_h|_{\mathrm{J}}. \tag{5.29}$$

The bounds (5.28) and (5.29) imply (5.21a). □

Approximation by coarse grid functions. Theorem 5.5 leads to the following result about the approximability of the fine space $V_{h,\mathbf{p}}$ with respect to the coarse space $V_{H,\mathbf{q}}$.

Theorem 5.6. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded convex polytopal domain, and let $\{\mathcal{T}_H\}_H$ and $\{\mathcal{T}_h\}_h$ be nested shape-regular sequences of meshes satisfying* (2.11), (2.12) *and* (2.13). *Let μ_F and η_F satisfy* (5.5) *for every face $F \in \mathcal{F}_h^{i,b}$. Then, for any $v_h \in V_{h,\mathbf{p}}$, there exists a $v_H \in V_{H,\mathbf{q}}$, such that*

$$\|v_h - v_H\|_{H^k(\Omega;\mathcal{T}_h)} \lesssim \left(\max_{D\in\mathcal{T}_H} \frac{H_D}{q_D} \right)^{2-k} \|v_h\|_{h,1}, \quad k \in \{0, 1, 2\}, \tag{5.30a}$$

$$\|v_H\|_{h,1}^2 \lesssim \left(1 + \max_{D\in\mathcal{T}_H} \left[\frac{H_D}{q_D} \max_{K\in\mathcal{T}_h(D)} \frac{p_K^2}{h_K} + \frac{H_D^3}{q_D^3} \max_{K\in\mathcal{T}_h(D)} \frac{p_K^6}{h_K^3} \right] \right) \|v_h\|_{h,1}^2. \tag{5.30b}$$

Proof. Let $v \in H^2(\Omega) \cap H_0^1(\Omega)$ be the approximation to v_h considered in Theorem 5.5. Since $q_D \geq 2$ for all $D \in \mathcal{T}_H$, Theorem C.6 implies that there exists $v_H \in V_{H,\mathbf{q}}$ such that

$$\|v - v_H\|_{H^j(D)} \lesssim \frac{H_D^{2-j}}{q_D^{2-j}} \|v\|_{H^2(D)}, \qquad \|D^\beta (v - v_H)\|_{L^2(\partial D)} \lesssim \frac{H_D^{3/2-|\beta|}}{q_D^{3/2-|\beta|}} \|v\|_{H^2(D)}, \tag{5.31}$$

for each $D \in \mathcal{T}_H$, each nonnegative integer $j \leq 2$, and each multi-index β with $|\beta| \leq 1$. Therefore, it is seen that (5.30a) follows from the triangle inequality in conjunction with (5.21b) and (5.31). In particular, note that $\|v_H\|_{H^2(\Omega;\mathcal{T}_h)} \lesssim \|v\|_{H^2(\Omega)}$, and since Theorem 5.5 implies that $\|v\|_{H^2(\Omega)} \lesssim \|v_h\|_{h,1}$, we obtain $\|v_H\|_{H^2(\Omega;\mathcal{T}_h)} \lesssim \|v_h\|_{h,1}$.

It remains to show (5.30b) by bounding the jump seminorm $|v_H|_{\mathrm{J}}$ as follows. If the face $F \in \mathcal{F}_h^i(D)$ for $D \in \mathcal{T}_H$, then the jumps of v_H and its first derivatives vanish because v_H is a polynomial over D. Since $v \in H^2(\Omega) \cap H_0^1(\Omega)$, $[\![v_H]\!] = [\![v_H - v]\!]$ and $[\![\nabla_{\mathrm{T}} v_H]\!] = [\![\nabla_{\mathrm{T}}(v_H - v)]\!]$ for each face $F \in \mathcal{F}_h^{i,b}(\partial D)$, whilst $[\![\nabla v_H \cdot n_F]\!] = [\![\nabla(v_H - v) \cdot n_F]\!]$ for each face $F \in \mathcal{F}_h^i(\partial D)$. Therefore, it is deduced from the mesh assumptions on $\mathcal{T}_h$ and $\mathcal{T}_H$ and (5.31) that

$$\begin{aligned}\sum_{F \in \mathcal{F}_h^{i,b}} \eta_F \|[\![v_H]\!]\|_{L^2(F)}^2 &\leq \sum_{D \in \mathcal{T}_H} \sum_{F \in \mathcal{F}_h^{i,b}(\partial D)} \eta_F \|[\![v_H - v]\!]\|_{L^2(F)}^2 \\ &\lesssim \sum_{D \in \mathcal{T}_H} \max_{K \in \mathcal{T}_h(D)} \frac{p_K^6}{h_K^3} \|v_H - v\|_{L^2(\partial D)}^2 \lesssim \max_{D \in \mathcal{T}_H} \left[\frac{H_D^3}{q_D^3} \max_{K \in \mathcal{T}_h(D)} \frac{p_K^6}{h_K^3} \right] \|v\|_{H^2(\Omega)}^2.\end{aligned}$$

Similar bounds also yield

$$\begin{aligned}\sum_{F \in \mathcal{F}_h^{i,b}} \mu_F \|[\![\nabla_{\mathrm{T}} v_H]\!]\|_{L^2(F)}^2 + \sum_{F \in \mathcal{F}_h^i} \mu_F \|[\![\nabla v_H \cdot n_F]\!]\|_{L^2(F)}^2 \\ \lesssim \max_{D \in \mathcal{T}_H} \left[\frac{H_D}{q_D} \max_{K \in \mathcal{T}_h(D)} \frac{p_K^2}{h_K} \right] \|v\|_{H^2(\Omega)}^2.\end{aligned}$$

Since $\|v\|_{H^2(\Omega)} \lesssim \|v_h\|_{h,1}$, the proof of (5.30b) is complete. □

Previous results on the approximation of fine mesh functions by coarse mesh functions typically involved lower-order projection operators, which were therefore suboptimal in terms of q in bounds such as (5.30a). The original result of an approximation with optimal orders in both H and q of Theorem 5.6 will lead to a sharp analysis of the nonoverlapping domain decomposition preconditioners in the next section.

5.2 Domain decomposition preconditioners

Throughout this section, we consider the model problem of finding $u_h \in V_{h,\mathbf{p}}$ such that

$$B_{h,1}(u_h, v_h) = \ell_h(v_h) \qquad \forall\, v_h \in V_{h,\mathbf{p}}, \tag{5.32}$$

where $\ell_h \colon V_{h,\mathbf{p}} \to \mathbb{R}$ is a bounded linear functional on $V_{h,\mathbf{p}}$. The resulting preconditioners based on $B_{h,1}$ are applied to HJB equations in the experiments of section 5.4.

The condition number of the linear system of (5.32) depends on the choice of basis for $V_{h,\mathbf{p}}$. However, in practice, the basis is often chosen to be either a nodal basis or a mapped orthonormal basis. For example, let us assume that each basis function ϕ_i of $V_{h,\mathbf{p}}$ has

support in only one element, and is mapped from a member of a set of functions that are L^2-orthonormal on a reference element. Then, it can be shown that the ℓ^2-norm condition number $\kappa(\mathbf{B})$ of the matrix $\mathbf{B} \coloneqq (B_{h,1}(\phi_i, \phi_j))$ satisfies

$$\kappa(\mathbf{B}) \lesssim \max_{K\in\mathcal{T}_h} \frac{p_K^8}{h_K^4} \frac{\max_{K\in\mathcal{T}_h} h_K^d}{\min_{K\in\mathcal{T}_h} h_K^d}, \tag{5.33}$$

where it is recalled that d is the dimension of the domain Ω.

Domain decomposition. Let Ω be partitioned into a set $\mathcal{T}_S \coloneqq \{\Omega_i\}_{i=1}^N$ of nonoverlapping Lipschitz polytopal subdomains Ω_i. The partition $\mathcal{T}_S$ is assumed to be conforming. A coarse simplicial or parallelepipedal mesh $\mathcal{T}_H$ is associated to each fine mesh $\mathcal{T}_h$. Let $H_D \coloneqq \operatorname{diam} D$ for each $D \in \mathcal{T}_H$ and suppose that $H \coloneqq \max_{D\in\mathcal{T}_H} H_D$. It is required that the sequence of meshes $\{\mathcal{T}_H\}_H$ satisfy the mesh conditions of section 2.2. Furthermore, the partitions $\mathcal{T}_S$, $\mathcal{T}_H$ and $\mathcal{T}_h$ are assumed to be *nested*, in the sense that no face of $\mathcal{T}_S$, respectively $\mathcal{T}_H$, cuts the interior of an element of $\mathcal{T}_H$, respectively $\mathcal{T}_h$. Hence, each element $D \in \mathcal{T}_H$ satisfies $\overline{D} = \bigcup \overline{K}$, where the union is over all elements $K \in \mathcal{T}_h$ such that $K \subset D$.

For each mesh $\mathcal{T}_H$, let $\mathbf{q} \coloneqq (q_D \colon D \in \mathcal{T}_H)$ be a vector of *positive* integers; so $q_D \geq 1$ for each element $D \in \mathcal{T}_H$. Assume that $\mathbf{q}$ satisfies the bounded variation property of (2.13), and that $q_D \leq \min_{K\subset D} p_K$ for all $D \in \mathcal{T}_H$. For each $D \in \mathcal{T}_H$, define the sets

$$\begin{aligned} \mathcal{T}_h(D) &\coloneqq \{K \in \mathcal{T}_h : K \subset D\}, & \mathcal{F}_h^i(D) &\coloneqq \left\{F \in \mathcal{F}_h^i : F \subset D\right\}, \\ \mathcal{F}_h^i(\partial D) &\coloneqq \left\{F \in \mathcal{F}_h^i : F \subset \partial D\right\}, & \mathcal{F}_h^{i,b}(\partial D) &\coloneqq \{F \in \mathcal{F}_h^{i,b} : F \subset \partial D\}. \end{aligned} \tag{5.34}$$

Although the sets $\mathcal{F}_h^i(D)$ and $\mathcal{F}_h^{i,b}(D)$ are not disjoint, the above assumptions on the meshes imply that $\mathcal{F}_h^{i,b} = \bigcup_D \mathcal{F}_h^i(D) \cup \mathcal{F}_h^{i,b}(\partial D)$ and that $\mathcal{F}_h^i = \bigcup_D \mathcal{F}_h^i(D) \cup \mathcal{F}_h^i(\partial D)$.

Define the function spaces

$$V_{h,\mathbf{p}}^i \coloneqq \left\{v \in L^2(\Omega_i) : v|_K \in \mathcal{P}_{p_K}(K) \quad \forall\, K \in \mathcal{T}_h, K \subset \Omega_i\right\}, \quad 1 \leq i \leq N, \tag{5.35a}$$

$$V_{H,\mathbf{q}} \coloneqq \left\{v \in L^2(\Omega) : v|_D \in \mathcal{P}_{q_D}(D) \quad \forall\, D \in \mathcal{T}_H\right\}. \tag{5.35b}$$

For convenience of notation, let $V_{h,\mathbf{p}}^0 \coloneqq V_{H,\mathbf{q}}$. It follows from the above conditions on the meshes that every function $v_H \in V_{H,\mathbf{q}}$ also belongs to $V_{h,\mathbf{p}}$, so let $I_0 \colon V_{H,\mathbf{q}} \to V_{h,\mathbf{p}}$ denote the natural imbedding map. For $1 \leq i \leq N$, let $I_i \colon V_{h,\mathbf{p}}^i \to V_{h,\mathbf{p}}$ denote the natural injection operator defined by

$$I_i\, v_i \coloneqq \begin{cases} v_i & \text{on } \Omega_i, \\ 0 & \text{on } \Omega - \Omega_i, \end{cases} \qquad \forall\, v_i \in V_{h,\mathbf{p}}^i. \tag{5.36}$$

Then, any function $v_h \in V_{h,\mathbf{p}}$ can be decomposed as $v_h = \sum_{i=1}^N I_i\, (v_h|_{\Omega_i})$.

Let the bilinear forms $B^i_{h,1}\colon V^i_{h,\mathbf{p}} \times V^i_{h,\mathbf{p}} \to \mathbb{R}$, $0 \le i \le N$, be defined by

$$B^i_{h,1}(u_i, v_i) \coloneqq B_{h,1}(I_i\, u_i, I_i\, v_i) \qquad \forall\, u_i, v_i \in V^i_{h,\mathbf{p}}. \tag{5.37}$$

It is clear that the bilinear forms $B^i_{h,1}$ are symmetric and coercive on $V^i_{h,\mathbf{p}} \times V^i_{h,\mathbf{p}}$. For each $0 \le i \le N$, let $\mathbf{B}_i$ denote the matrix that corresponds to the bilinear form $B^i_{h,1}$ and let $\mathbf{I}_i$ denotes the matrix corresponding to the injection operator I_i. Then, we define the additive Schwarz preconditioner $\mathbf{P}$ by

$$\mathbf{P} \coloneqq \sum_{i=0}^{N} \mathbf{P}_i, \quad \mathbf{P}_i \coloneqq \mathbf{I}_i\, \mathbf{B}_i^{-1}\, \mathbf{I}_i^{\top}. \tag{5.38}$$

The preconditioner $\mathbf{P}$ is symmetric, and it can be employed in the preconditioned conjugate gradient method [30]. Further preconditioners, such as multiplicative, symmetric multiplicative and hybrid methods, are presented in [73, 75] and the references therein.

5.3 Spectral bounds

The general theory of Schwarz methods [73, 75] simplifies the analysis of the preconditioners described above to the verification of three key properties.

Property 5.1. Suppose that there exists a constant c_0 such that each $v_h \in V_{h,\mathbf{p}}$ admits a decomposition $v_h = \sum_{i=0}^{N} I_i\, v_i$, with $v_i \in V^i_{h,\mathbf{p}}$, for each $0 \le i \le N$, with

$$\sum_{i=0}^{N} B^i_{h,1}(v_i, v_i) \le c_0\, B_{h,1}(v_h, v_h). \tag{5.39}$$

Property 5.2. Assume that there exist constants $\varepsilon_{ij} \in [0, 1]$, such that

$$|B_{h,1}(I_i\, v_i, I_j v_j)| \le \varepsilon_{ij} \sqrt{B_{h,1}(I_i\, v_i, I_i\, v_i)\; B_{h,1}(I_j\, v_j, I_j\, v_j)}, \tag{5.40}$$

for all $v_i \in V^i_{h,\mathbf{p}}$ and all $v_j \in V^j_{h,\mathbf{p}}$, $1 \le i, j \le N$. Let $\rho(\mathcal{E})$ denote the spectral radius of the matrix $\mathcal{E} \coloneqq (\varepsilon_{ij})$.

Property 5.3. Suppose that there exists a constant $\omega \in (0, 2)$, such that

$$B_{h,1}(I_i\, v_i, I_i\, v_i) \le \omega\, B^i_{h,1}(v_i, v_i) \quad \forall\, v_i \in V^i_{h,\mathbf{p}},\ 0 \le i \le N. \tag{5.41}$$

Properties 5.1–5.3 are sometimes referred to respectively as the stable decomposition property, the strengthened Cauchy–Schwarz inequality, and local stability. The following theorem from the theory of Schwarz methods is quoted from [75].

Theorem 5.7. *If Properties 5.1–5.3 hold, then the condition number* $\kappa(\mathbf{PB})$ *satisfies*

$$\kappa(\mathbf{PB}) \leq c_0\,\omega\,(\rho(\mathcal{E})+1)\,. \tag{5.42}$$

Remark 5.2. With the above choices of bilinear forms $B^i_{h,1}$ and with the arguments presented in [5], it is seen that (5.41) holds in fact with equality for $\omega = 1$. Also, in (5.40), we can take $\varepsilon_{ij} = 1$ if $\partial\Omega_i \cap \partial\Omega_j \neq \emptyset$, and $\varepsilon_{ij} = 0$ otherwise. Therefore, as explained in [5], $\rho(\mathcal{E}) \leq N_c+1$, where N_c is the maximum number of adjacent subdomains that a given subdomain might have. Therefore, Properties 5.2 and 5.3 hold, and it remains to verify Property 1.

The following theorem determines a bound on the constant appearing in (5.39), which can be used in conjunction with Theorem 5.7 to analyse the properties of the preconditioners. The proof of this result is given in the next section.

Theorem 5.8. *Let* $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, *be a bounded convex polytopal domain, and let* $\mathcal{T}_S$, $\{\mathcal{T}_H\}_H$ *and* $\{\mathcal{T}_h\}_h$ *be successively nested shape-regular sequences of meshes, with* $\mathcal{T}_S$ *conforming, and* $\{\mathcal{T}_H\}_H$ *and* $\{\mathcal{T}_h\}_h$ *satisfying* (2.11), (2.12) *and* (2.13). *Let* μ_F *and* η_F *satisfy* (5.5) *for each face* F, *with* c_μ *and* c_η *so that* (5.6) *holds. Then, each* $v_h \in V_{h,\mathbf{p}}$ *admits a decomposition* $v_h = \sum_{i=0}^N I_i\,v_i$, *with* $v_i \in V^i_{h,\mathbf{p}}$, $0 \leq i \leq N$, *such that*

$$\sum_{i=0}^{N} B^i_{h,1}(v_i,v_i) \lesssim \tilde{c}_0\,B_{h,1}(v_h,v_h), \tag{5.43}$$

where the constant $\tilde{c}_0$ *is given by*

$$\tilde{c}_0 \coloneqq 1 + \max_{D\in\mathcal{T}_H}\left[\frac{q_D}{H_D}\max_{K\in\mathcal{T}_h(D)}\frac{p_K^2}{h_K}\right]\max_{D\in\mathcal{T}_H}\frac{H_D^2}{q_D^2} + \max_{D\in\mathcal{T}_H}\left[\frac{q_D}{H_D}\max_{K\in\mathcal{T}_h(D)}\frac{p_K^6}{h_K^3}\right]\max_{D\in\mathcal{T}_H}\frac{H_D^4}{q_D^4}. \tag{5.44}$$

It follows from Theorems 5.7 and 5.8 that the condition number of **PB** satisfies

$$\kappa(\mathbf{PB}) \lesssim \tilde{c}_0\,(N_c+2)\,, \tag{5.45}$$

where $\tilde{c}_0$ is as above, and N_c is the maximum number of adjacent subdomains that a given subdomain from $\mathcal{T}_S$ might have. If the sequence of coarse spaces $\{V_{H,\mathbf{q}}\}_H$ satisfy the assumption that $H_D/q_D \lesssim \min_{D\in\mathcal{T}_H} H_D/q_D$ for all $D \in \mathcal{T}_H$, then the constant $\tilde{c}_0$ in the above proposition simplifies to

$$\tilde{c}_0 \simeq 1 + \max_{D\in\mathcal{T}_H}\left[\frac{H_D}{q_D}\max_{K\in\mathcal{T}_h(D)}\frac{p_K^2}{h_K} + \frac{H_D^3}{q_D^3}\max_{K\in\mathcal{T}_h(D)}\frac{p_K^6}{h_K^3}\right]. \tag{5.46}$$

Moreover, if the sequences of meshes $\{\mathcal{T}_H\}_H$ and $\{\mathcal{T}_h\}_h$ are quasiuniform, and if the polynomial degrees are also quasiuniform in the sense that $q \coloneqq \max_D q_D \lesssim q_D$ for all $D \in \mathcal{T}_H$ and $p \coloneqq \max_K p_K \lesssim p_K$ for all $K \in \mathcal{T}_h$, then the condition number of the preconditioned

matrix **PB** satisfies the bound

$$\kappa\left(\mathbf{PB}\right) \lesssim \left(N_c + 2\right)\left(1 + \frac{p^2\, H}{q\, h} + \frac{p^6\, H^3}{q^3\, h^3}\right). \tag{5.47}$$

It is well-known that the above bound is optimal in terms of H and h, see [18, 35].

The numerical experiments of section 5.4 show that the bound (5.47) is also sharp in terms of the orders of p and q. Choosing the coarse space such that $H \simeq h$ and $q \simeq p$ implies that $\kappa(\mathbf{PB}) \lesssim p^3$, which shows that the preconditioner is not robust with respect to p, yet constitutes nonetheless a significant improvement over the condition number of order p^8/h^4 for the unpreconditioned matrix.

Stable decomposition property. The following lemma, due to Feng and Karakashian in [34], provides a trace inequality for the boundaries ∂D of elements $D \in \mathcal{T}_H$. However, the inequality is not written there in the form that is required for our purposes. So, we present again the proof, with some variations from the arguments in [34].

Lemma 5.9. *Let $\{\mathcal{T}_H\}_H$ and $\{\mathcal{T}_h\}_h$ be shape-regular sequences of nested simplicial or parallelepipedal meshes satisfying the conditions* (2.11) *and* (2.12)*, and let* $\mathbf{p}$ *satisfy* (2.13)*. Let $v \in L^2(D)$ belong to $\mathcal{P}_{p_K}(K)$ for each $K \subset D$. Then, we have*

$$\|v\|^2_{L^2(\partial D)} \lesssim \sum_{K \in \mathcal{T}_h(D)} |v|_{H^1(K)} \|v\|_{L^2(K)} + \frac{1}{H_D}\|v\|^2_{L^2(D)} + \sqrt{\sum_{F \in \mathcal{F}_h^i(D)} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|[\![v]\!]\|^2_{L^2(F)}}\, \|v\|_{L^2(D)}. \tag{5.48}$$

Proof. As shown in [34], since each element $D \in \mathcal{T}_H$ is an affine image of a convex reference element, it follows that there is a point $x_0 \in D$, such that $(x - x_0) \cdot n_{\partial D} \gtrsim H_D$ for each $x \in \partial D$, where $n_{\partial D}$ is the unit outward normal vector to ∂D. Therefore,

$$\|v\|^2_{L^2(\partial D)} \lesssim \frac{1}{H_D} \int_{\partial D} |v|^2\, (x - x_0) \cdot n_{\partial D}\, \mathrm{d}s. \tag{5.49}$$

Integration by parts shows that

$$\int_{\partial D} |v|^2\, (x - x_0) \cdot n_{\partial D}\, \mathrm{d}s = \sum_{K \in \mathcal{T}_h(D)} \int_K \Big[\mathrm{div}\,(x - x_0)\, |v|^2 + 2v\, \nabla v \cdot (x - x_0)\Big]\, \mathrm{d}x - \sum_{F \in \mathcal{F}_h^i(D)} \langle [\![v^2]\!], \{(x - x_0) \cdot n_F\}\rangle_F.$$

Since $[\![v^2]\!] = 2[\![v]\!]\{v\}$, it is found that

$$\int_{\partial D} |v|^2 (x - x_0) \cdot n_{\partial D} \, \mathrm{d}s \lesssim H_D \sum_{K \in \mathcal{T}_h(D)} |v|_{H^1(K)} \|v\|_{L^2(K)} + \|v\|^2_{L^2(D)} + H_D \sqrt{\sum_{F \in \mathcal{F}_h^i(D)} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|[\![v]\!]\|^2_{L^2(F)}} \sqrt{\sum_{F \in \mathcal{F}_h^i(D)} \frac{\tilde{h}_F}{\tilde{p}_F^2} \|\{v\}\|^2_{L^2(F)}}.$$

The inverse and trace inequalities imply that

$$\sum_{F \in \mathcal{F}_h^i(D)} \frac{\tilde{h}_F}{\tilde{p}_F^2} \|\{v\}\|^2_{L^2(F)} \lesssim \|v\|^2_{L^2(D)}.$$

Therefore, (5.48) follows from (5.49) and the above bounds. □

Equipped with the approximation result of Theorem 5.6, it is now possible to prove Theorem 5.8 using a similar approach to [5, 34, 35].

Proof of Theorem 5.8. Let v_H be given as in Theorem 5.6, set $v_0 \coloneqq v_H$, and denote by $v_i \in V^i_{h,\mathbf{p}}$ the restriction of $v_h - v_H$ to Ω_i, $1 \le i \le N$; hence $v_h = \sum_{i=0}^N I_i v_i$. Then, we write

$$\sum_{i=0}^N B^i_{h,1}(v_i, v_i) = B_{h,1}(v_H, v_H) + B_{h,1}(v_h - v_H, v_h - v_H) - \sum_{\substack{i,\, j=1 \\ i \neq j}}^N B_{h,1}(I_i v_i, I_j v_j). \tag{5.50}$$

Observe that the constant appearing on the right-hand side of (5.30b) can be bounded in terms of $\tilde{c}_0$, which was defined in (5.44). So, Theorem 5.6 and (5.6) imply

$$B_{h,1}(v_H, v_H) \simeq \|v_H\|^2_{h,1} \lesssim \tilde{c}_0 \|v_h\|^2_{h,1} \simeq \tilde{c}_0 \, B_{h,1}(v_h, v_h), \tag{5.51a}$$

$$B_{h,1}(v_h - v_H, v_h - v_H) \lesssim \|v_h\|^2_{h,1} + \|v_H\|^2_{h,1} \lesssim \tilde{c}_0 \, B_{h,1}(v_h, v_h). \tag{5.51b}$$

It remains to bound the last term in (5.50), which concerns the interface flux and jump terms at the boundaries of the subdomains of $\mathcal{T}_S$. Expanding this term and using the triangle inequality leads to

$$\sum_{\substack{i,\, j=1 \\ i \neq j}}^N |B_{h,1}(I_i v_i, I_j v_j)| \le \sum_{k=1}^5 E_k, \tag{5.52}$$

where the quantities E_k are defined by

$$E_1 \coloneqq \sum_{\substack{i,j=1\\ i\neq j}}^{N} \sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}} \eta_F \,|\langle (v_h - v_H)|_{\Omega_i}, (v_h - v_H)|_{\Omega_j}\rangle_F|, \tag{5.53a}$$

$$E_2 \coloneqq \sum_{\substack{i,j=1\\ i\neq j}}^{N} \sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}} \mu_F \,|\langle \nabla_{\mathrm{T}}(v_h - v_H)|_{\Omega_i}, \nabla_{\mathrm{T}}(v_h - v_H)|_{\Omega_j}\rangle_F|, \tag{5.53b}$$

$$E_3 \coloneqq \sum_{\substack{i,j=1\\ i\neq j}}^{N} \sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}} \mu_F \,|\langle \nabla(v_h - v_H)|_{\Omega_i}\cdot n_F, \nabla(v_h - v_H)|_{\Omega_j}\cdot n_F\rangle_F|, \tag{5.53c}$$

$$E_4 \coloneqq \sum_{\substack{i,j=1\\ i\neq j}}^{N} \sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}} |\langle \mathrm{div}_{\mathrm{T}}\, \nabla_{\mathrm{T}}(v_h - v_H)|_{\Omega_i}, \nabla(v_h - v_H)|_{\Omega_j}\cdot n_F\rangle_F|, \tag{5.53d}$$

$$E_5 \coloneqq \sum_{\substack{i,j=1\\ i\neq j}}^{N} \sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}} |\langle \nabla_{\mathrm{T}}(\nabla(v_h - v_H)|_{\Omega_i}\cdot n_F), \nabla_{\mathrm{T}}(v_h - v_H)|_{\Omega_j}\rangle_F|. \tag{5.53e}$$

Note that in (5.53), we have made use of the symmetry of the sum over i, j, $i\neq j$, and the fact that any face $F\subset\partial\Omega_i\cap\partial\Omega_j$ must be an interior face.

Defining $\eta_D \coloneqq \max_{K\in\mathcal{T}_h(D)} p_K^6/h_K^3$ for each $D\in\mathcal{T}_H$, the hypotheses (2.12) and (2.13) and the nestedness of the meshes imply that

$$E_1 \lesssim \sum_{D\in\mathcal{T}_H} \eta_D \|v_h - v_H\|_{L^2(\partial D)}^2.$$

Therefore, using the trace inequality of Lemma 5.9, we find that

$$E_1 \lesssim \sum_{D\in\mathcal{T}_H} \eta_D \Bigg[\frac{H_D}{q_D} \sum_{K\in\mathcal{T}_h(D)} |v_h - v_H|_{H^1(K)}^2 + \frac{H_D}{q_D} \sum_{F\in\mathcal{F}_h^i(D)} \frac{\tilde{p}_F^2}{\tilde{h}_F} \|[\![v_h]\!]\|_{L^2(F)}^2 + \frac{q_D}{H_D} \sum_{K\in\mathcal{T}_h(D)} \|v_h - v_H\|_{L^2(K)}^2 \Bigg].$$

Notice that the jumps $[\![v_H]\!]$ vanish for faces $F\in\mathcal{F}_h^i(D)$. Therefore, the approximation bound of Theorem 5.6 gives

$$E_1 \lesssim \max_{D\in\mathcal{T}_H}\left[\eta_D \frac{H_D}{q_D}\right] \max_{D\in\mathcal{T}_H} \frac{H_D^2}{q_D^2} \|v_h\|_{h,1}^2 + \max_{D\in\mathcal{T}_H}\left[\eta_D \frac{H_D}{q_D} \max_{F\in\mathcal{F}_h^i(D)} \frac{\tilde{h}_F^2}{\tilde{p}_F^4}\right] |v_h|_{\mathrm{J}}^2 + \max_{D\in\mathcal{T}_H}\left[\eta_D \frac{q_D}{H_D}\right] \max_{D\in\mathcal{T}_H} \frac{H_D^4}{q_D^4} \|v_h\|_{h,1}^2, \tag{5.54}$$

and thus it follows from (2.12) and (2.13) and (5.6) that

$$E_1 \lesssim \max_{D\in\mathcal{T}_H}\left[\frac{q_D}{H_D}\max_{K\in\mathcal{T}_h(D)}\frac{p_K^6}{h_K^3}\right]\max_{D\in\mathcal{T}_H}\frac{H_D^4}{q_D^4}\,B_{h,1}(v_h,v_h). \tag{5.55}$$

Remark that in going from (5.54) to (5.55), we have used the bounds

$$\frac{H_D}{q_D}\lesssim\frac{q_D}{H_D}\max_{D\in\mathcal{T}_H}\frac{H_D^2}{q_D^2},\qquad\qquad \frac{H_D}{q_D}\max_{F\in\mathcal{F}_h^i(D)}\frac{\tilde{h}_F^2}{\tilde{p}_F^4}\lesssim\frac{q_D}{H_D}\max_{D\in\mathcal{T}_H}\frac{H_D^4}{q_D^4}.$$

This is done because it is currently not possible to improve the last term in (5.54), as a consequence of the nonlocal form of the bounds in Theorems 5.5 and Theorem 5.6.

The Cauchy–Schwarz inequality with a parameter and the symmetry of the sum over i, j, $j\neq i$, imply that

$$\sum_{k=2}^{5}E_k\lesssim\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}}\mu_F^{-1}\|D^2(v_h-v_H)|_{\Omega_i}\|_{L^2(F)}^2+\mu_F\|\nabla(v_h-v_H)|_{\Omega_j}\|_{L^2(F)}^2. \tag{5.56}$$

Since $\mathcal{T}_S$ is conforming, each face F may appear at most twice in the above sum, and thus the trace and inverse inequalities imply that

$$\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}}\mu_F^{-1}\|D^2(v_h-v_H)|_{\Omega_i}\|_{L^2(F)}^2\lesssim\sum_{K\in\mathcal{T}_h}\|v_h-v_H\|_{H^2(K)}^2\lesssim\tilde{c}_0\,B_{h,1}(v_h,v_h). \tag{5.57}$$

Defining $\mu_D\coloneqq\max_{K\in\mathcal{T}_h(D)}p_K^2/h_K$, we apply Lemma 5.9 componentwise to the gradient of v_h-v_H to find that

$$\begin{aligned}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\substack{F\in\mathcal{F}_h^i\\ F\subset\partial\Omega_i\cap\partial\Omega_j}}\mu_F\|\nabla(v_h-v_H)|_{\Omega_j}\|_{L^2(F)}^2&\lesssim\sum_{D\in\mathcal{T}_H}\mu_D\|\nabla(v_h-v_H)\|_{L^2(\partial D)}^2\\ &\lesssim\sum_{D\in\mathcal{T}_H}\mu_D\bigg[\frac{H_D}{q_D}\sum_{K\in\mathcal{T}_h(D)}|v_h-v_H|_{H^2(K)}^2+\frac{H_D}{q_D}\sum_{F\in\mathcal{F}_h^i(D)}\frac{\tilde{p}_F^2}{\tilde{h}_F}\|[\![\nabla v_h]\!]\|_{L^2(F)}^2\\ &\qquad+\frac{q_D}{H_D}\sum_{K\in\mathcal{T}_h(D)}|v_h-v_H|_{H^1(K)}^2\bigg].\end{aligned} \tag{5.58}$$

It is important to observe that only terms involving interior faces of the mesh $\mathcal{T}_h$ appear on the right-hand side of the above inequality, so that

$$\|[\![\nabla v_h]\!]\|_{L^2(F)}^2=\|[\![\nabla_{\mathrm{T}}\,v_h]\!]\|_{L^2(F)}^2+\|[\![\nabla v_h\cdot n_F]\!]\|_{L^2(F)}^2\quad\forall\,F\in\mathcal{F}_h^i(D).$$

So, we deduce that

$$\sum_{D\in\mathcal{T}_H} \mu_D \|\nabla(v_h - v_H)\|^2_{L^2(\partial D)} \lesssim \max_{D\in\mathcal{T}_H}\left[\mu_D \frac{H_D}{q_D}\right] \|v_h - v_H\|^2_{H^2(\Omega;\mathcal{T}_h)} + \max_{D\in\mathcal{T}_H}\left[\mu_D \frac{H_D}{q_D}\right] |v_h|^2_{\mathrm{J}} + \max_{D\in\mathcal{T}_H}\left[\mu_D \frac{q_D}{H_D}\right] \|v_h - v_H\|^2_{H^1(\Omega;\mathcal{T}_h)},$$

and thus Theorem 5.6 and (5.6) show that

$$\sum_{D\in\mathcal{T}_H} \mu_D \|\nabla(v_h - v_H)\|^2_{L^2(\partial D)} \lesssim \max_{D\in\mathcal{T}_H}\left[\frac{q_D}{H_D} \max_{K\in\mathcal{T}_h(D)} \frac{p_K^2}{h_K}\right] \max_{D\in\mathcal{T}_H} \frac{H_D^2}{q_D^2}\, B_{h,1}(v_h, v_h). \tag{5.59}$$

Therefore, the inequalities (5.56), (5.57) and (5.59) show that

$$\sum_{k=2}^{5} E_k \lesssim \max_{D\in\mathcal{T}_H}\left[\frac{q_D}{H_D} \max_{K\in\mathcal{T}_h(D)} \frac{p_K^2}{h_K}\right] \max_{D\in\mathcal{T}_H} \frac{H_D^2}{q_D^2}\, B_{h,1}(v_h, v_h). \tag{5.60}$$

In summary, combining the inequalities (5.51), (5.55) and (5.60) implies that

$$\sum_{i=0}^{N} B^i_{h,1}(I_i v_i, I_i v_i) \lesssim \tilde{c}_0\, B_{h,1}(v_h, v_h) + \sum_{k=1}^{5} E_k \lesssim \tilde{c}_0\, B_{h,1}(v_h, v_h), \tag{5.61}$$

which completes the proof of the stable decomposition property of Theorem 5.8. □

The proof of Theorem 5.8 completes the verification of Properties 5.1–5.3, and thus gives the bound (5.45) for the condition number of the preconditioned system.

5.4 Numerical experiments

In this section, we confirm the sharpness of the spectral bounds of section 5.2 and we investigate the performance and competitiveness of the preconditioners in practical applications. The implementation of the numerical experiments below employed direct factorisations to construct the coarse mesh and local solvers.

5.4.1 First experiment

Since the bound (5.45) is the first to be explicit in both coarse and fine mesh polynomial degrees, it is important to ascertain its sharpness.

Let $\Omega = (0,1)^2$, and let the fixed meshes $\mathcal{T}_H = \mathcal{T}_S$ be obtained by a uniform subdivision of Ω into 4 squares, and let $\mathcal{T}_h$ be obtained by uniform subdivision of Ω into 16 squares. We consider the sequence of spaces $V_{h,\mathbf{p}}$ of piecewise polynomials on $\mathcal{T}_h$ with total degree p, where $p = 2, \ldots, 12$, and the coarse spaces $V_{H,\mathbf{q}}$ of piecewise polynomials on $\mathcal{T}_H$ with

$\kappa(\mathbf{PB})$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$	q rate
$p=2$	2.16×10^{1}					
$p=3$	3.34×10^{2}	6.71×10^{1}				
$p=4$	1.94×10^{3}	3.16×10^{2}	1.35×10^{2}			
$p=5$	7.22×10^{3}	1.43×10^{3}	4.11×10^{2}	2.10×10^{2}		
$p=6$	2.12×10^{4}	4.40×10^{3}	1.31×10^{3}	6.44×10^{2}	3.03×10^{2}	3.60
$p=7$	5.31×10^{4}	1.10×10^{4}	3.50×10^{3}	1.70×10^{3}	8.97×10^{2}	3.35
$p=8$	1.18×10^{5}	2.46×10^{4}	7.91×10^{3}	4.27×10^{3}	2.10×10^{3}	3.25
$p=9$	2.38×10^{5}	4.88×10^{4}	1.61×10^{4}	8.68×10^{3}	4.55×10^{3}	3.10
$p=10$	4.48×10^{5}	9.17×10^{4}	3.00×10^{4}	1.64×10^{4}	8.86×10^{3}	3.00
$p=11$	7.92×10^{5}	1.61×10^{5}	5.29×10^{4}	2.90×10^{4}	1.58×10^{4}	2.97
$p=12$	1.33×10^{6}	2.71×10^{5}	8.89×10^{4}	4.87×10^{4}	2.66×10^{4}	2.97
p rate	5.97	5.94	5.96	5.97	6.03	

TABLE 5.1: *The dependence of the condition number $\kappa(\mathbf{PB})$ on the coarse and fine mesh polynomial degrees for experiment of section* 5.4.1. *The asymptotic rates are computed by regression on the last three entries of each column for p and each row for q. It is found that $\kappa(\mathbf{PB})$ is of order $1+p^6/q^3$, as predicted in section* 5.2.

total degree q, where $q=2,\ldots,6$. We apply the additive Schwarz preconditioner defined in section 5.2 to the bilinear form $B_{h,1}$, with $c_\mu = c_\eta = 10$.

These choices are made to ensure that the resulting number of degrees of freedom is small, being at most equal to 1456 in the case of $p=12$, thereby facilitating the accurate computation of the condition numbers $\kappa(\mathbf{PB})$ of the preconditioned matrix $\mathbf{PB}$. The resulting condition numbers are given in Table 5.1 and represented in Figure 5.2, which show that $\kappa(\mathbf{PB})$ is of order $1+p^6/q^3$, in agreement with the results of section 5.2 and in particular with the bound (5.47). This confirms that the predicted rates with respect to the polynomial degrees are optimal.

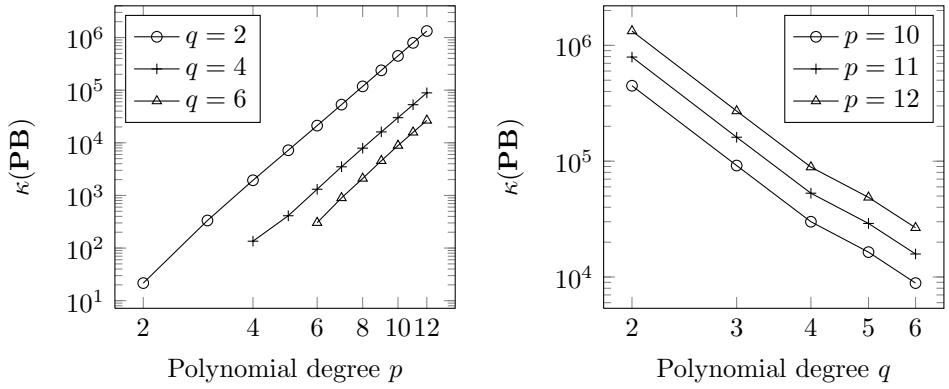

FIGURE 5.2: *The dependence of the condition number $\kappa(\mathbf{PB})$ on the coarse and fine mesh polynomial degrees for experiment of section* 5.4.1. *The condition numbers of Table* 5.1 *are plotted for representative degrees p and q, showing the predicted rates $\kappa(\mathbf{PB}) \simeq 1+p^6/q^3$.*

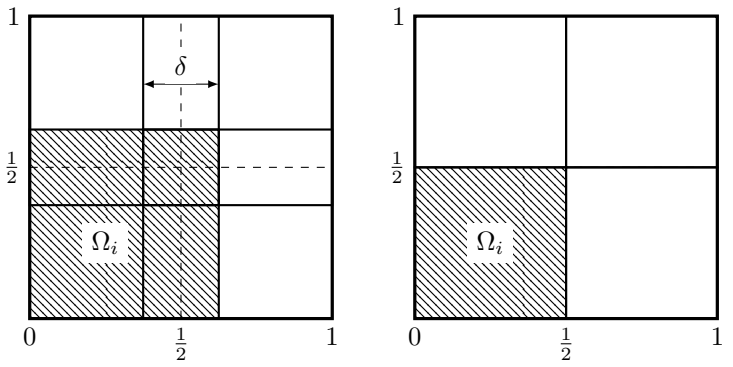

FIGURE 5.3: *Overlapping and nonoverlapping decompositions of* $\Omega = (0,1)^2$ *used in the experiment of section* 5.4.2. *Four subdomains are used for both the overlapping and nonoverlapping methods, with the overlap size* δ *defined as the length shown above.*

5.4.2 Second experiment

In this section, we compare the efficiency of nonoverlapping methods with the closely related overlapping methods. It is found that the methods achieve similar performances in terms of iteration counts, although nonoverlapping methods are often faster as a result of lower computational costs.

Let $\Omega \coloneqq (0,1)^2$, and let $\mathcal{T}_h$ be obtained by uniform subdivision of Ω into squares of size $h = 2^{-k}$, $k = 3, \ldots, 8$. Let $V_{h,\mathbf{p}}$ consist of the space of polynomials of fixed partial degree $p = 2$ on each element $K \in \mathcal{T}_h$. Consider the model problem (5.32), where the linear functional ℓ_h is chosen so that the solution u_h of (5.32) approximates the function $u(x,y) \coloneqq \mathrm{e}^{xy}\sin(\pi x)\sin(\pi y)$; specifically, we define

$$\ell_h(v_h) \coloneqq \sum_{K \in \mathcal{T}_h} \langle \Delta u, \Delta v_h \rangle_K \quad \forall\, v_h \in V_{h,\mathbf{p}}.$$

Using the results of [70], it can then be shown that $\|u - u_h\|_{H^2(\Omega;\mathcal{T}_h)} \lesssim h^{p-1}$. The penalty parameters c_μ and c_η are chosen so that $\mu_F = 10/\tilde{h}_F$ and $\eta_F = 10/\tilde{h}_F^3$.

Overlapping domain decomposition. Let $\delta \in (0,1)$ and let Ω be divided into overlapping subdomains $\mathcal{T}_S = \{\Omega_i\}_{i=1}^4$, as shown in the left-hand side diagram of Figure 5.3. This yields an overlapping decomposition of Ω with overlap δ; here, we use $\delta \in \{1/4, 1/8, 1/16\}$. Let $\mathcal{T}_H$ be a coarse mesh consisting of a uniform subdivision of Ω into 4 squares, thus yielding the ratios $H/\delta \in \{2, 4, 8\}$, and let $V_{H,\mathbf{q}}$ consist of the space of polynomials of fixed partial degree $q = 2$ on each element $D \in \mathcal{T}_H$. The local spaces $V_{h,\mathbf{p}}^i$ with associated solvers $B_{h,1}^i$, $1 \le i \le 4$, are defined analogously to the nonoverlapping case, described in section 5.2. The additive Schwarz preconditioner is also defined analogously to section 5.2.

Nonoverlapping domain decomposition. The domain Ω is partitioned into four subdomains $\mathcal{T}_S = \{\Omega_i\}_{i=1}^4$, as shown in the right-hand side diagram of Figure 5.3. We consider three sequences of coarse meshes $\mathcal{T}_H$, also obtained by uniform subdivision of Ω into squares of size $H = 2^{-m}$, $m = 1, \ldots, k-1$, so that $H/h \in \{2, 4, 8\}$. The nonoverlapping additive Schwarz preconditioner is defined as in section 5.2.

Results. The implementations of the overlapping and nonoverlapping methods were the same, except for the required difference in handling the subdomains. Since the parallelisations of overlapping and nonoverlapping methods differ, our implementation was in serial in order to permit a more straightforward comparison.

Table 5.4 gives the number of iterations required to reduce the Euclidean norm of the residual by a factor of 10^{-6}. The results for both methods are comparable to those in the literature: see for instance [3, 5, 18, 55]. Table 5.4 also presents a representative sample of the CPU times required for the assembly of the preconditioner and the application of the preconditioned conjugate gradient (PCG) method. The assembly timing strictly includes the time spent on assembling and factorising the coarse and local mesh solvers, whereas the solver time strictly includes the time spent on applying the PCG method. These timings are meant to provide a relative comparison of the methods, with better absolute timings achievable by parallelisation.

For the same iteration count, the nonoverlapping methods are generally faster in both assembly and solution. This advantage in efficiency is essentially the result of the smaller dimension of the subdomain solvers. The nonoverlapping method is also generally cheaper in terms of memory costs. Observe also that the method yielding the lowest iteration count is not necessarily the fastest: in this example the nonoverlapping method with $H = 4h$ offers the fastest total solution time. Remarkably, the cheapest nonoverlapping method with $H = 8h$ is slightly faster than the most expensive overlapping method with $H = 2\delta$, even if it requires more than twice the total number of iterations. Overall, our results show that both methods are efficient, with low iteration counts that remain bounded as H/δ or H/h is held fixed.

5.4.3 Third experiment

We will now consider applications of the preconditioning methods to problems of practical interest, namely fully nonlinear HJB equations. This introduces several challenges, such as nonsymmetric linear systems that appear in the semismooth Newton method. Nevertheless, it is found that nonoverlapping methods in particular remain robust and lead to efficient solvers for these problems. This example is closely related to the experiment of section 3.7.1.

		PCG Iteration count					
		Overlapping			Nonoverlapping		
DoF	h	$H = 2\delta$	$H = 4\delta$	$H = 8\delta$	$H = 2h$	$H = 4h$	$H = 8h$
144	1/4				20		
576	1/8	18			22	29	
2304	1/16	18	24		22	30	43
9216	1/32	18	25	37	20	32	52
36864	1/64	18	25	41	18	30	50
147456	1/128	18	26	41	17	27	48
589824	1/256	18	26	42	17	25	40

	Timing					
	Overlapping			Nonoverlapping		
$h = 1/128$	$H = 2\delta$	$H = 4\delta$	$H = 8\delta$	$H = 2h$	$H = 4h$	$H = 8h$
Assembly time	18.6s	14.5s	13.0s	14.0s	11.9s	11.6s
Solver time	8.39s	9.56s	13.3s	6.51s	8.62s	14.4s

TABLE 5.4: *Number of preconditioned conjugate gradient (PCG) iterations required to reduce the residual norm by a factor of 10^{-6} for overlapping and nonoverlapping methods, in the experiment of section 5.4.2, along with sample timings for assembly and timings of the PCG algorithm. The methods yield similar iteration counts for similar ratios of H/δ or H/h, but the nonoverlapping method is faster to assemble and apply, as a result of the smaller number of degrees of freedom in the local solvers. In practice, the best choice of preconditioner involves a tradeoff between iteration counts and computational costs.*

Consider the boundary-value problem

$$\begin{aligned} \sup_{\alpha\in\Lambda} [L^\alpha u - f^\alpha] = 0 \quad &\text{in } \Omega, \\ u = 0 \quad &\text{on } \partial\Omega, \end{aligned} \tag{5.62}$$

where $\Omega = (0,1)^2$, $\Lambda \coloneqq [0,\pi/3] \times \mathrm{SO}(2)$, and where $L^\alpha v \coloneqq a^\alpha : D^2 v$, with

$$a^\alpha \coloneqq \frac{1}{2} R \begin{pmatrix} 1+\sin^2\theta & \sin\theta\,\cos\theta \\ \sin\theta\,\cos\theta & \cos^2\theta \end{pmatrix} R^\top, \quad \alpha = (\theta, R) \in \Lambda. \tag{5.63}$$

The source terms f^α, $\alpha \in \Lambda$, are chosen so that the solution is $u(x,y) = \mathrm{e}^{xy}\sin(\pi x)\sin(\pi y)$, whilst yielding large variations in the values of α that attain the supremum in (5.62).

The numerical scheme (3.25) is applied on a sequence of fine meshes $\mathcal{T}_h$ obtained by uniform subdivision of Ω into squares of size $h = 2^{-k}$, $k = 3, \ldots, 7$. Let $V_{h,\mathbf{p}}$ consist of the space of polynomials of fixed partial degree $p = 2$ on each element $K \in \mathcal{T}_h$. To solve the nonlinear problem (3.25), we use the semismooth Newton method detailed in section 3.6. Each iteration of the semismooth Newton method leads to a nonsymmetric but positive definite linear system [71], which we solve using the left-preconditioned GMRES method, with the current approximate solution used as the starting guess. The preconditioners employed are the nonoverlapping and overlapping preconditioners from the experiment of

		Average GMRES iterations (Newton steps)					
		Overlapping			Nonoverlapping		
DoF	h	$H = 2\delta$	$H = 4\delta$	$H = 8\delta$	$H = 2h$	$H = 4h$	$H = 8h$
144	1/4				18.3 (5)		
576	1/8	21.2 (5)			19.0 (5)	23.2 (5)	
2304	1/16	24.0 (5)	25.4 (5)		18.2 (5)	24.0 (5)	31.0 (5)
9216	1/32	28.5 (6)	29.7 (6)	38.2 (6)	19.0 (6)	24.5 (6)	35.2 (6)
36864	1/64	32.2 (6)	33.3 (6)	44.5 (6)	19.7 (6)	23.7 (6)	31.8 (6)
147456	1/128	34.2 (6)	35.3 (6)	48.7 (6)	20.0 (6)	23.2 (6)	29.8 (6)

TABLE 5.5: *Average number of left-preconditioned GMRES iterations per Newton step required to reduce the (preconditioned) residual by a factor of* 10^{-6}*, with total number of Newton steps in parentheses, for the problem of section* 5.4.3*. The iteration counts for the nonoverlapping methods remain low and bounded for fixed values of* H/h*.*

section 5.4.2, and we emphasize that they are based on preconditioning the bilinear form $B_{h,1}$, implying that no additional factorisations are required between Newton steps, thus constituting a significant advantage in terms of efficiency.

As explained in section 3.7.1, a key challenge in this example is that the diffusion coefficient a^α is highly anisotropic for θ near $\pi/3$, and the rotation matrices R may lead to large variations in the resulting diffusions across the domain and between Newton steps. As a result, we may expect significant anisotropic variations in the resulting linearisations encountered in the application of the semismooth Newton method. It is thus of interest to determine if the above preconditioners remain robust in this context, even though they are built from the isotropic bilinear form $B_{h,1}$.

Table 5.5 gives the average number of GMRES iterations per Newton step required to reduce the Euclidean norm of the (preconditioned) residual by a factor of 10^{-6}, along with the total number of Newton steps required to achieve convergence, defined here as a step-increment L^2-norm below 10^{-6}. As in the case of section 5.4.2, the nonoverlapping methods are generally faster than the overlapping methods in terms of assembly and application.

These results show that nonoverlapping methods remain robust in face of the anisotropy, lack of symmetry and nonlinearity of the problem. We comment that right-preconditioning in the GMRES method yields similar results when used with an appropriate scaling of the unpreconditioned residual Euclidean norm. We thereby conclude that these preconditioning methods lead to effective solvers for these challenging problems.

Conclusion

In this thesis, we have proposed the first arbitrarily high-order numerical method for a broad class of fully nonlinear HJB PDE that is supported by a complete analysis in terms of consistency, stability and error bounds. Our approach, based on the Cordes condition, overcomes many long-standing difficulties in the analysis of high-order numerical methods for fully nonlinear PDE. The Cordes condition also encompasses a broad range of HJB PDE with strongly anisotropic diffusion coefficients that present a considerable challenge to existing monotone schemes.

Our numerical experiments on challenging problems with strongly anisotropic diffusion coefficients, nonsmooth solutions, boundary layers and early-time singularities demonstrate the robustness, flexibility and accuracy of the numerical method. In particular, we have provided the first examples of exponential convergence rates under hp- and τq-refinement for fully nonlinear PDE.

We have shown that the semismooth Newton method combined with nonoverlapping domain decomposition preconditioners enables the fast and robust iterative solution of the discretised problems. Using original approximation results for discontinuous finite element spaces, we obtained sharp spectral bounds for these preconditioners applied to a symmetric model problem, thus constituting the first result for this class of preconditioners that is explicit in all discretisation parameters. Applications of these preconditioners to fully nonlinear HJB equations demonstrate the robustness and competitiveness of these algorithms in computations.

The approach to HJB PDE introduced in this thesis opens up the possibility of exploiting many advanced computational techniques that improve the accuracy and efficiency of computations. For instance, since strongly anisotropic problems may feature solutions with highly localised features, it is of significant interest to develop adaptive algorithms that automatically adapt the mesh and polynomial degrees to the problem at hand. However, the numerical examples given in this thesis consisted of meshes and polynomial degrees that were chosen a priori. Nevertheless, the flexibility in mesh-choice and polynomial degrees of our method lends itself well to adaptive mesh refinement algorithms driven by a posteriori error analysis. Therefore, an interesting direction for further work involves developing a posteriori error bounds and adaptive algorithms for these equations.

An additional goal for further research is to improve on the preconditioning results of Chapter 5, since it was seen there that these preconditioners are not robust with respect to the polynomial degree. However, there exists a significant literature on preconditioning p- and hp-version FEM, mostly in the context of H^1-type norms. Therefore, it is of interest to consider the possibility of extending these results to the current context of H^2-type norms. Moreover, there is the interesting question of extending the current two-level preconditioners to more general multilevel algorithms in order to decrease the computational cost of the coarse mesh and subdomain solvers.

Appendix A

Miranda–Talenti inequality

The purpose of the following appendices is to provide the proofs of several key results used in this thesis. Although the results given here are for the most part already well-known, we have chosen to include original proofs for several reasons. In some cases, the proofs are hard to find in the literature, or they are given in forms seemingly different to those used in this work; in some cases, the proofs in the literature contain significant gaps. Thus, our aim is to present these results as completely as possible in the form most suited to our needs. In all cases, we attempt to indicate as clearly as possible which parts of the proofs are our own work, and which parts follow the literature.

In this first appendix, we establish a result similar to [40, Lemma 3.2.2.1], concerning the approximation of the convex domain Ω by domains with at least $C^{1,1}$ boundaries. This result represents a key ingredient in the proof of the Miranda–Talenti inequality. Our approximation result is slightly weaker than [40, Lemma 3.2.2.1], which was claimed without proof; the difference is that we show approximation by domains with $C^{1,1}$ boundaries rather than C^2 boundaries. Although it is stated in [40] that [40, Lemma 3.2.2.1] "*follows easily from the results*" in [29], it is found that [29] treats approximation of convex sets by "regular convex sets" signifying therein approximation by a convex set such that the set intersects each of its supporting hyperplanes in at most one point and such that each of its boundary points lies on only one supporting hyperplane; this notion of regularity is therefore insufficient for $C^{1,1}$-boundary regularity required here, as shown by counterexamples.

The distance between two subsets A and B of $\mathbb{R}^d$ is defined by

$$\operatorname{dist}(A,B) \coloneqq \sup_{x\in A}\inf_{y\in B}|x-y| + \sup_{y\in B}\inf_{x\in A}|x-y|. \tag{A.1}$$

Here we choose $|\cdot|$ to denote specifically the Euclidean norm on $\mathbb{R}^d$.

Theorem A.1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open convex set. Then, for each $\varepsilon > 0$, there exist open convex sets U_ε and V_ε with $C^{1,1}$ boundaries such that $\operatorname{dist}(\Omega, U_\varepsilon) < \varepsilon$, $\operatorname{dist}(\Omega, V_\varepsilon) < \varepsilon$ and $U_\varepsilon \subset \Omega \subset V_\varepsilon$.*

The proof of Theorem A.1 is split into two lemmas: in Lemma A.2, we show that convex sets can be approximated from the outside by convex sets with $C^{1,1}$ boundaries, and in Lemma A.3, we show how to construct inner approximations by translation and rescaling of outer approximations.

Lemma A.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open convex set. Then, for each $\delta > 0$, there exists an open convex set Ω_δ with $C^{1,1}$ boundary such that $\Omega \subset \Omega_\delta$ and* $\operatorname{dist}(\Omega, \Omega_\delta) = \delta$.

Proof. Define

$$\Omega_\delta \coloneqq \bigcup_{x \in \Omega} B(x, \delta),$$

where $B(x, \delta)$ is the open Euclidean ball around x of radius δ. Clearly, $\Omega \subset \Omega_\delta$ and Ω_δ is open. It is also straightforward to check that Ω_δ is convex and that $\operatorname{dist}(\Omega, \Omega_\delta) = \delta$.

We begin by noting the fact that Ω_δ satisfies *the uniform interior sphere condition with radius δ*: for every $x \in \partial\Omega_\delta$, there exists $y \in \overline{\Omega}$ such that $B(y, \delta) \subset \Omega_\delta$ and $|x - y| = \delta$. To see this, suppose that $x \in \partial\Omega_\delta$. Then we must have $\inf_{y \in \Omega} |x - y| \geq \delta$ for otherwise x would belong to Ω_δ. Now let $\{x_n\}_{n=1}^\infty \subset \Omega_\delta$ be a sequence converging to x; there is a sequence $\{y_n\}_{n=1}^\infty \subset \Omega$ such that $|x_n - y_n| < \delta$ for all $n \in \mathbb{N}$. Since Ω is precompact, there exists a subsequence of $\{y_n\}_{n=1}^\infty$, to which we pass without change of notation, and $y \in \overline{\Omega}$, such that $y_n \to y$. Therefore $|x - y| = \lim_{n \to \infty} |x_n - y_n| \leq \delta$ and hence $|x - y| = \delta$. We must have $B(y, \delta) \subset \Omega_\delta$ since for any point $\tilde{x} \in B(y, \delta)$, we use $y \in \overline{\Omega}$ to see that there is $\tilde{y} \in \Omega$ such that $|\tilde{x} - \tilde{y}| < \delta$, thus $\tilde{x} \in \Omega_\delta$. This shows that Ω_δ satisfies the uniform interior sphere condition with radius δ.

We claim that Ω_δ has a $C^{1,1}$ boundary in the sense of [40, Definition 1.2.1.1]. Let $x_0 \in \partial\Omega_\delta$ and let V be a neighbourhood of x_0 such that under a new orthonormal coordinate system, $V = \{(y_1, \ldots, y_d) : -a_i < y_i < a_i,\ 1 \leq i \leq d\}$. Let $V' = \{(y_1, \ldots, y_{d-1}) : -a_i < y_i < a_i,\ 1 \leq i \leq d-1\}$ and let $\varphi \colon V' \to \mathbb{R}$ be such that $|\varphi(y')| \leq a_d/2$ for each $y' \in V'$ and such that $\Omega_\delta \cap V = \{(y', y_d) \in V : \varphi(y') < y_d\}$ and $\partial\Omega_\delta \cap V = \{(y', y_n) \in V : \varphi(y') = y_d\}$. With these definitions, φ is a convex function. Moreover, by [40, Corollary 1.2.2.3], which states that Ω_δ has a Lipschitz boundary, it follows that $\varphi \in C^{0,1}(\bar{V'})$, after possibly shrinking V'.

Since V' has Lipschitz boundary, if we show that $\varphi \in W^{2,\infty}(V')$, then it will follow that $\varphi \in C^{1,1}\left(\bar{V'}\right)$ and the proof will be complete. Following the arguments in [19, Section 1.2], to prove that $\varphi \in W^{2,\infty}(V')$ it is enough to show that there exists a constant $C > 0$ such that, for each compact subset $K \subset\subset V'$ and each unit vector $v \in \mathbb{R}^{d-1}$, we have

$$\lim_{h \to 0} \sup_{y' \in K} \frac{|\varphi(y' + hv) + \varphi(y' - hv) - 2\varphi(y')|}{h^2} \leq C. \tag{A.2}$$

It follows from the interior sphere condition and the supporting hyperplane theorem that for each point $y \in \partial\Omega_\delta \cap V$, there exists a unique inward-pointing unit normal vector to $\partial\Omega_\delta$. This defines a vector field $\nu \colon V' \to \mathbb{R}^d$ of inward-pointing unit normal vectors, and we also define $\nu' \coloneqq (\nu_1, \cdots, \nu_{d-1})$. Since $\varphi \in C^{0,1}(V')$, we have $\nu_d > 0$ for each $y' \in V'$.

It follows from geometry that for any $y' \in V'$, any unit vector $v \in \mathbb{R}^{d-1}$, and any h sufficiently small,

$$\varphi(y') - \frac{h\, v \cdot \nu'}{\nu_d} \le \varphi(y' + hv) \le \varphi(y') + \delta\nu_d - \sqrt{\delta^2\nu_d^2 - h^2 + 2\delta h\, v \cdot \nu'}, \tag{A.3}$$

where we omit the dependence of ν on y' for simplicity of notation. Therefore, it is found that φ is differentiable at y', for every $y' \in V'$, and that $D\varphi = -\nu'/\nu_d$, or equivalently

$$\nu(y') = \frac{(-\varphi_{x_1}(y'), \dots, -\varphi_{x_{d-1}}(y'), 1)}{\sqrt{1 + |D\varphi(y')|^2}}. \tag{A.4}$$

Therefore, we have

$$\varphi(y' + hv) \le \varphi(y') + \frac{\delta}{\sqrt{1 + |D\varphi(y')|^2}} - \sqrt{\frac{\delta^2}{1 + |D\varphi(y')|^2} - h^2 + 2\delta h\, v \cdot \nu'}. \tag{A.5}$$

It follows from (A.5) and convexity of φ that

$$\begin{aligned} 0 &\le \varphi(y' + hv) + \varphi(y' - hv) - 2\varphi(y') \\ &\le \frac{2\delta}{\sqrt{1 + |D\varphi(y')|^2}} - \sum_{\sigma \in \{-1,1\}} \sqrt{\frac{\delta^2}{1 + |D\varphi(y')|^2} - h^2 + 2\sigma\delta h\, v \cdot \nu'}. \end{aligned} \tag{A.6}$$

The bound (A.2) with $C = C\Big(d, |\varphi|_{C^{0,1}(\bar{V}')}, \delta\Big)$ is then obtained from (A.6), thus showing that $\varphi \in C^{1,1}(\bar{V}')$. □

Lemma A.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open convex set. For every $\varepsilon > 0$, there exists $\delta > 0$, such that if V is an open set containing Ω with $\operatorname{dist}(\Omega, V) < \delta$, then there is an open set $U \subset \Omega$ such that U is similar to V and $\operatorname{dist}(\Omega, U) < \varepsilon$. In particular, the boundary of U is of the same class as the boundary of V and U is convex if and only if V is also convex.*

Proof. By hypothesis, Ω is a bounded open convex set, and without loss of generality we may assume that $0 \in \Omega$. Let $g_\Omega \colon \mathbb{R}^d \to \mathbb{R}$ denote the *gauge functional* of Ω, defined by

$$g_\Omega(x) = \inf\{t > 0 : x/t \in \Omega\}. \tag{A.7}$$

It is well-known that g_Ω is a sublinear functional on $\mathbb{R}^d$, and that

$$\Omega = \{x \in \mathbb{R}^d : g_\Omega(x) < 1\}.$$

Since Ω is open and bounded, there exist positive constants r and R such that

$$r\,|x| \le g_\Omega(x) \le R\,|x| \quad \forall\, x \in \mathbb{R}^d.$$

Also, by boundedness of Ω, there is $K > 0$ sufficiently large such that Ω is contained in the ball of radius K about the origin. Let $\varepsilon > 0$ be fixed; without loss of generality we may assume that $\varepsilon < K$. Now suppose that V is an open set containing Ω such that $\operatorname{dist}(\Omega, V) < \delta$ and

$$\delta < \frac{(1-\varepsilon/K)^{-1} - 1}{R}. \tag{A.8}$$

Since g_Ω is sublinear, it follows that $g_\Omega(x) \le 1 + R\,\delta$ for all $x \in V$. Indeed, if $x \in V$, then there exists a $y \in \Omega$ such that $|x - y| \le \delta$. Therefore,

$$g_\Omega(x) \le g_\Omega(y) + g_\Omega(x-y) \le 1 + R\,\delta.$$

Now let $\lambda \in (0,1)$ and define

$$U = \{\lambda\, x : x \in V\}.$$

If $\lambda < (1 + R\,\delta)^{-1}$, then $g_\Omega(x) < 1$ for all $x \in U$ by sublinearity of g_Ω; hence $U \subset \Omega$. In this case, $\operatorname{dist}(\Omega, U) = \sup_{x\in\Omega} \operatorname{dist}(x, U)$, which we estimate as follows. Let $x \in \Omega \subset V$. Then $\lambda x \in U$, and thus $\operatorname{dist}(x, U) \le (1-\lambda)|x| \le (1-\lambda)K$. So $\operatorname{dist}(\Omega, U) \le (1-\lambda)K < \varepsilon$ provided that $\lambda > 1 - \varepsilon/K$. It is then clear from (A.8) that the interval $\big(1-\varepsilon/K, (1+R\,\delta)^{-1}\big)$ is non-empty, thereby allowing λ to be chosen so that U satisfies the claims stated above. □

Proof of Theorem A.1. Let $\varepsilon > 0$ be given, and let $\delta > 0$ be as in Lemma A.3. Without loss of generality, assume that $\delta < \varepsilon$. Then let the open convex set Ω_δ with $C^{1,1}$ boundary be as in Lemma A.2, so that $\Omega \subset \Omega_\delta$ and $\operatorname{dist}(\Omega, \Omega_\delta) = \delta < \varepsilon$. Lemma A.3 shows that there exists an open convex set $U_\varepsilon \subset \Omega$, such that $\operatorname{dist}(\Omega, U_\varepsilon) < \varepsilon$ and U_ε has a $C^{1,1}$ boundary. Then U_ε and $V_\varepsilon := \Omega_\delta$ satisfy the claim of Theorem A.1. □

Lemma A.4. *Let Ω be a bounded open convex set in $\mathbb{R}^d$ and suppose that $\{U_\varepsilon\}_\varepsilon$ is a sequence of open convex subsets of Ω such that $\operatorname{dist}(\Omega, U_\varepsilon) < \varepsilon$ for every $\varepsilon > 0$. Then, for any compact set $K \subset \Omega$, there exists an $\varepsilon_0 > 0$ such that $K \subset U_\varepsilon$ for all $\varepsilon < \varepsilon_0$.*

Proof. Since K is a compact subset of the open set Ω, there exists a $\delta > 0$ such that for every $x \in K$, $B(x, \delta) \subset \Omega$. Set $\varepsilon_0 < \delta$, and assume that there is an $\varepsilon < \varepsilon_0$ such that $K \not\subset U_\varepsilon$, i.e. there is an $x \in K \setminus U_\varepsilon$. Since U_ε is an open convex set, the supporting hyperplane theorem shows that there exists a bounded linear functional ℓ on $\mathbb{R}^d$ such that $\ell(x) > \ell(y)$ for every $y \in U_\varepsilon$. This implies that $\|\ell\| > 0$, and hence after renormalising, we may assume that $\|\ell\| = 1$. Moreover, there is a unit vector $v \in \mathbb{R}^d$ such that $\ell(v) = 1$. Then, set $z = x + \varepsilon\, v$ and note that $z \in \Omega$ since $\varepsilon < \delta$. Furthermore, the hypothesis on U_ε implies that there is a point $y \in U_\varepsilon$ such that $|z - y| < \varepsilon$. Therefore,

$$\ell(y) = \ell(y - z) + \ell(x + \varepsilon\, v) = \ell(y - z) + \varepsilon + \ell(x) \ge -|y - z| + \varepsilon + \ell(x) \ge \ell(x).$$

However, this contradicts $\ell(x) > \ell(y)$ for all $y \in U_\varepsilon$, therefore $K \subset U_\varepsilon$. □

For completeness, we now show how Theorem A.1 is used in the proof of the Miranda–Talenti inequality for a general convex domain Ω with possibly nonsmooth boundary. The proof given here follows closely [40] and [56].

Theorem A.5 (Miranda–Talenti)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain. Then, for any $u \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}, \tag{A.9a}$$

$$\|u\|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2(\Omega)}, \tag{A.9b}$$

where C is a constant depending only on d and $\operatorname{diam}\Omega$*.*

Proof. The results of [38, 40, 56] imply that, for any convex domain Ω with $C^{1,1}$ boundary and for any $u \in H^2(\Omega) \cap H_0^1(\Omega)$, there holds

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

The generalisation to nonsmooth convex domains is largely based on the proof of H^2-regularity of solutions of Poisson's equation on convex domains [40, Theorem 3.2.1.2]. Let Ω be a convex domain, with possibly nonsmooth boundary; Theorem A.1 shows that for any $\varepsilon > 0$, there is an open convex subset $U_\varepsilon \subset \Omega$ such that U_ε has a $C^{1,1}$ boundary Γ_ε, and such that $\operatorname{dist}(U_\varepsilon, \Omega) < \varepsilon$. Select a sequence of such sets U_ε with $\varepsilon \to 0$.

For a function $u \in H^2(\Omega) \cap H_0^1(\Omega)$, set $f = \Delta u \in L^2(\Omega)$. Then, define $u_\varepsilon \in H^2(U_\varepsilon) \cap H_0^1(U_\varepsilon)$ as the strong solution of

$$\begin{aligned} \Delta u_\varepsilon &= f \quad \text{in } U_\varepsilon, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma_\varepsilon. \end{aligned} \tag{A.10}$$

Since Γ_ε is of class $C^{1,1}$, existence and uniqueness of $u_\varepsilon \in H^2(U_\varepsilon) \cap H_0^1(U_\varepsilon)$ is shown in [38]. The extensions by zero of u_ε, also denoted u_ε, belong to $H_0^1(\Omega)$ for all $\varepsilon > 0$, and there exists a constant C depending only on $\operatorname{diam}\Omega$ such that $\|u_\varepsilon\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}$; thus u_ε is uniformly bounded in $H_0^1(\Omega)$. Additionally, since the Miranda–Talenti estimate holds for domains of class $C^{1,1}$, it is found that $|u_\varepsilon|_{H^2(U_\varepsilon)} \leq \|\Delta u\|_{L^2(\Omega)}$ for all ε. Let v_{ij}^ε denote the zero-extension of $(u_\varepsilon)_{x_i x_j}$ onto Ω. We claim that there is a subsequence, to which we pass without change of notation, such that $u_\varepsilon \rightharpoonup u$ in $H_0^1(\Omega)$, with $v_{ij}^\varepsilon \rightharpoonup u_{x_i x_j}$ in $L^2(\Omega)$, where $u_{x_i x_j}$ is the ij-th weak derivative of u on Ω. These results are shown by the arguments in [40], yet a proof is given here for completeness.

First, from the bounds $\|u_\varepsilon\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}$ and $\|v_{ij}^\varepsilon\|_{L^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}$ for all ε, there is a subsequence, to which we pass without change of notation, such that $u_\varepsilon \rightharpoonup \tilde{u}$ in $H^1(\Omega)$ for some $\tilde{u} \in H_0^1(\Omega)$, and such that $v_{ij}^\varepsilon \rightharpoonup v_{ij}$ in $L^2(\Omega)$ for some functions $v_{ij} \in L^2(\Omega)$,

$1 \le i, j \le d$. First, we show that $u = \tilde{u}$. The hypothesis that $u \in H^2(\Omega)$ and the definition of weak derivatives implies that, for any $\phi \in C_0^\infty(\Omega)$, we have

$$\int_\Omega \nabla u \cdot \nabla \phi \, \mathrm{d}x = -\int_\Omega \Delta u \, \phi \, \mathrm{d}x = -\int_\Omega f \, \phi \, \mathrm{d}x.$$

Since ϕ is compactly supported in Ω and since U_ε is convex, Lemma A.4 shows that there is an $\varepsilon_0 > 0$ such that $\operatorname{supp} \phi \subset U_\varepsilon$ for all $\varepsilon < \varepsilon_0$. Since $u_\varepsilon \in H^2(U_\varepsilon) \cap H_0^1(U_\varepsilon)$ solves (A.10) and $u_\varepsilon \rightharpoonup \tilde{u}$ in $H^1(\Omega)$, we get

$$\begin{aligned}\int_\Omega \nabla \tilde{u} \cdot \nabla \phi \, \mathrm{d}x &= \lim_{\varepsilon \to 0} \int_\Omega \nabla u_\varepsilon \cdot \nabla \phi \, \mathrm{d}x = \lim_{\varepsilon \to 0} \int_{U_\varepsilon} \nabla u_\varepsilon \cdot \nabla \phi \, \mathrm{d}x \\ &= \lim_{\varepsilon \to 0} -\int_{U_\varepsilon} f \, \phi \, \mathrm{d}x = -\int_\Omega f \, \phi \, \mathrm{d}x.\end{aligned}$$

Since $\phi \in C_0^\infty(\Omega)$ is arbitrary and $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, the equality $u = \tilde{u}$ follows from the fact that $u, \tilde{u} \in H_0^1(\Omega)$ satisfy

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega \nabla \tilde{u} \cdot \nabla v \, \mathrm{d}x \quad \forall \, v \in H_0^1(\Omega).$$

Then, since $u_\varepsilon \rightharpoonup u$ as $\varepsilon \to 0$ and since $v_{ij}^\varepsilon \rightharpoonup v_{ij}$, we find that, for any $\phi \in C_0^\infty(\Omega)$,

$$\begin{aligned}\int_\Omega u \, \phi_{x_i x_j} \, \mathrm{d}x &= \lim_{\varepsilon \to 0} \int_{U_\varepsilon} u_\varepsilon \, \phi_{x_i x_j} \, \mathrm{d}x = \lim_{\varepsilon \to 0} \int_{U_\varepsilon} (u_\varepsilon)_{x_i x_j} \, \phi \, \mathrm{d}x \\ &= \lim_{\varepsilon \to 0} \int_\Omega v_{ij}^\varepsilon \, \phi \, \mathrm{d}x = \int_\Omega v_{ij} \, \phi \, \mathrm{d}x.\end{aligned}$$

So $v_{ij} = u_{x_i x_j}$ is the ij-th weak derivative of u on Ω, as claimed. Finally, lower semi-continuity of the L^2-norm with respect to weak convergence implies that

$$|u|_{H^2(\Omega)}^2 = \sum_{i,j=1}^d \|v_{ij}\|_{L^2(\Omega)}^2 \le \liminf_{\varepsilon \to 0} \sum_{i,j=1}^d \|v_{ij}^\varepsilon\|_{L^2(\Omega)}^2 = \liminf_{\varepsilon \to 0} |u_\varepsilon|_{H^2(U_\varepsilon)}^2 \le \|\Delta u\|_{L^2(\Omega)}^2.$$

This proves inequality (A.9a). Using the hypothesis that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ to approximate the function u by an H^1-convergent sequence of functions of class $C_0^\infty(\Omega)$, we find that

$$|u|_{H^1(\Omega)}^2 \le \|u\|_{L^2(\Omega)} \|\Delta u\|_{L^2(\Omega)}. \tag{A.11}$$

Indeed, since $u \in H_0^1(\Omega)$, there exists a sequence $\{u_\rho\}_{\rho > 0} \subset C_0^\infty(\Omega)$ such that $u_\rho \to u$ in $H^1(\Omega)$ as $\rho \to 0$. Then, since $u \in H^2(\Omega)$, the definition of weak derivatives implies that

$$\int_\Omega \nabla u \cdot \nabla u \, \mathrm{d}x = \lim_{\rho \to 0} \int_\Omega \nabla u \cdot \nabla u_\rho \, \mathrm{d}x = -\lim_{\rho \to 0} \int_\Omega \Delta u \, u_\rho \, \mathrm{d}x \le \|\Delta u\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)},$$

where we have used the fact that $\|u_\rho\|_{L^2(\Omega)} \to \|u\|_{L^2(\Omega)}$ as $\rho \to 0$. Then, inequality (A.9b) is obtained from Poincaré's inequality together with inequality (A.11). □

Appendix B

Kuratowski–Ryll-Nardzewski theorem

In this appendix, we provide a proof of Theorem 3.12 that was used in the construction and analysis of the semismooth Newton method of section 3.6. The proof essentially relies on the application of a general measurable selection theorem due to Kuratowski and Ryll-Nardzewski [51], for which a readily available source is [8, p. 90]. We give in Theorem B.1 below a statement and proof of this key result; there are two main reasons for its inclusion here. The first reason is that the result given in [8] is restricted to the case where the topological space X appearing below is a subinterval of $\mathbb{R}$, whereas for our purposes it is necessary to consider a more general setting. The second reason is that the selection in [8] is shown to be Lebesgue measurable, whereas we show and make essential use of its Borel-measurability.

For a set-valued function $F\colon X \rightrightarrows Y$ and a subset $U \subset Y$, where X and Y are sets, the preimage of U under F is defined as $F^{-1}(U) \coloneqq \{x \in X : F(x) \cap U \neq \emptyset\}$. For a metric space Y and $U \subset Y$, we denote $B(U, \varepsilon) \coloneqq \cup_{x \in U} B(x, \varepsilon)$, where $B(x, \varepsilon)$ denotes the open ball of radius ε in the metric of Y.

Theorem B.1. *Let X be a topological space and let $\mathcal{B}$ denote the Borel σ-algebra of X. Let Λ be a compact metric space, and let $F\colon X \rightrightarrows \Lambda$ be a set-valued function, such that $F(x)$ is non-empty and closed in Λ for all $x \in X$, and $F^{-1}(U) \in \mathcal{B}$ for every open set $U \subset \Lambda$. Then, there exists a Borel measurable selection $\alpha\colon X \to \Lambda$ from F, i.e. $\alpha(x) \in F(x)$ for all $x \in \Lambda$ and $\alpha^{-1}(U) \in \mathcal{B}$ for every open set $U \subset \Lambda$.*

Proof. Since Λ is a compact metric space, it is complete, separable and has finite diameter strictly less than some number $M < \infty$. Let $C = \{c_i\}_{i=0}^{\infty}$ be a countable dense subset of Λ. For $k \in \mathbb{N}$, define $\varepsilon_k \coloneqq M/2^k$. We construct a sequence of mappings $\alpha_k\colon X \to \Lambda$, $k \in \mathbb{N}$, that satisfy the following properties:

Property B.1. The map α_k is Borel measurable for each $k \in \mathbb{N}$.

Property B.2. For every $x \in X$ and $k \in \mathbb{N}$, $\alpha_k(x) \in B\left(F(x), \varepsilon_k\right)$.

Property B.3. For every $x \in X$ and $k \geq 1$, $\alpha_k(x) \in B\left(\alpha_{k-1}(x), \varepsilon_{k-1}\right)$.

Define $\alpha_0(x) \coloneqq c_0$ for every $x \in X$. We check that Property B.1 and Property B.2 hold. For any open set $U \subset \Lambda$, either $c_0 \in U$ and $\alpha_0^{-1}(U) = X$, or $c_0 \notin U$ and $\alpha_0^{-1}(U) = \emptyset$. Either way, $\alpha_0^{-1}(U)$ is a Borel set of X and hence Property B.1 holds. For any $x \in X$, $F(x)$ is non-empty and Λ has diameter less than $M = \varepsilon_0$, therefore we see that $\operatorname{dist}(\alpha_0(x), F(x)) < M$, so Property B.2 holds. Property B.3 holds by definition for $k = 0$.

Now, assume that for a given $k \geq 1$, α_{k-1} has been defined and satisfies the above properties. Define

$$A_j \coloneqq F^{-1}\left(B(c_j, \varepsilon_k)\right) \cap \alpha_{k-1}^{-1}\left(B(c_j, \varepsilon_{k-1})\right). \tag{B.1}$$

Set $E_0 \coloneqq A_0$ and $E_i \coloneqq A_i - \bigcup_{j<i} E_j$ for each $i \in \mathbb{N}$; note that E_i is a Borel set for each $i \in \mathbb{N}$ as a consequence of Property B.1 for α_{k-1} and the hypothesis that the preimages under F of open sets are Borel sets.

We claim that $X = \bigcup_{i=0}^{\infty} E_i$. Let $x \in X$ be arbitrary. Since α_{k-1} satisfies Property B.2 and $F(x)$ is non-empty, there is a $\beta \in F(x)$ such that

$$\rho \coloneqq \operatorname{dist}(\alpha_{k-1}(x), \beta) < \varepsilon_{k-1}. \tag{B.2}$$

By density of C, there is a $c_j \in C$ such that $\operatorname{dist}(\beta, c_j) < \min(\varepsilon_k, \varepsilon_{k-1} - \rho)$. Recalling that

$$F^{-1}\left(B(c_j, \varepsilon_k)\right) \coloneqq \left\{y \in X : F(y) \cap B(c_j, \varepsilon_k) \neq \emptyset\right\},$$

we see that $x \in F^{-1}\left(B(c_j, \varepsilon_k)\right)$. Additionally, $\operatorname{dist}(\alpha_{k-1}(x), c_j) < \varepsilon_{k-1}$, showing that $x \in \alpha_{k-1}^{-1}\left(B(c_j, \varepsilon_{k-1})\right)$. This shows that $x \in A_j$, and thus by construction of $\{E_i\}_{i=0}^{\infty}$, $x \in E_i$ for some $i \leq j$. Therefore, $X = \bigcup_{i=0}^{\infty} E_i$ as claimed.

Because $\{E_i\}_{i=0}^{\infty}$ is a collection of mutually disjoint Borel subsets of X, we may define $\alpha_k \colon X \mapsto \Lambda$ by $\alpha_k(x) \coloneqq c_i$ if $x \in E_i$. The map α_k is well-defined, and satisfies Properties B.2 and B.3 because for any $x \in X$, there is an $i \in \mathbb{N}$ for which $x \in E_i \subset A_i$. Furthermore, the map α_k is Borel measurable, because for any open set $U \subset \Lambda$, the set $\{c_{i_j}\}_{j=0}^{\infty} \coloneqq U \cap C$ is a countable subset of C, and $\alpha_k(x) \in U$ if and only if $\alpha_k(x) \in \{c_{i_j}\}_{j=0}^{\infty}$. Therefore, we have $\alpha_k^{-1}(U) = \bigcup_{j=0}^{\infty} E_{i_j}$, which shows that $\alpha_k^{-1}(U)$ is a countable union of Borel sets, and is thus a Borel set: this proves that α_k satisfies Property B.1.

By induction, the sequence $\{\alpha_k\}_{k=0}^{\infty}$ is well-defined, and α_k satisfies properties B.1, B.2, and B.3 for all $k \geq 0$. Now, Property B.3 implies that for every $x \in X$, $\{\alpha_k(x)\}_{k=0}^{\infty}$ is a Cauchy sequence in Λ, since $\operatorname{dist}\left(\alpha_{k+n}(x), \alpha_k(x)\right) < M/2^{k-1}$ for all n and $k \in \mathbb{N}$. By completeness of Λ, there exists $\alpha(x) \coloneqq \lim_{k\to\infty} \alpha_k(x)$ for all $x \in X$. The hypothesis that $F(x)$ is closed implies that $\alpha(x) \in F(x)$, for otherwise there would exist a k such that $\alpha_k(x) \notin B\left(F(x), \varepsilon_k\right)$. Finally, the function $\alpha \colon X \to \Lambda$ is Borel measurable, because it is the pointwise everywhere[1] limit of Borel measurable functions mapping into a metric space. □

[1] It is important to note that pointwise everywhere convergence is crucial here, as otherwise the function α may fail to be Borel measurable: see the remarks in [66, Section 18].

Lemma B.2. *Let X be a topological space, let Λ be a metric space, and let $F\colon X \rightrightarrows \Lambda$ be an upper semicontinuous set-valued function. Then the preimages under F of closed sets are closed sets and the preimages under F of open sets are Borel sets.*

Proof. Let $A \subset \Lambda$ be closed. Recall that $F^{-1}(A) = \{x \in X : F(x) \cap A \neq \emptyset\}$. If $x \notin F^{-1}(A)$, then A^{c} is an open neighbourhood of $F(x)$. Since F is upper-semicontinuous, there is a neighbourhood V of x such that $F(y) \subset A^{\mathsf{c}}$ for all $y \in V$, or equivalently $V \subset F^{-1}(A)^{\mathsf{c}}$; thus $F^{-1}(A)^{\mathsf{c}}$ is open and $F^{-1}(A)$ is closed. Let $U \subset \Lambda$ be open: since Λ is a metric space, U is a countable union of closed sets. It follows that $F^{-1}(U)$ is also a countable union of closed sets and so is a Borel set. □

Lemma B.3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set, let Λ be a compact metric space and let the set-valued function $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$, defined on $\Omega \times \mathbb{R}^m$, be upper semicontinuous. Let $X \subset \Omega$ and let the function $\mathbf{u}\colon X \mapsto \mathbb{R}^m$ be continuous with respect to the inherited topology on X. Then, the set-valued function $x \mapsto \Lambda\left(x, \mathbf{u}(x)\right)$ is upper semicontinuous on X.*

Proof. Let $x \in X$ and let U be an open neighbourhood of $\Lambda\left(x, \mathbf{u}(x)\right)$. Then, by upper semicontinuity of $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$, there are open neighbourhoods $V_x \subset \Omega$ and $V_u \subset \mathbb{R}^m$, respectively of x and $\mathbf{u}(x)$, such that $\Lambda(y, v) \subset U$ for all $y \in V_x$, $v \in V_u$. Since $\mathbf{u}\colon X \mapsto \mathbb{R}^m$ is continuous with respect to the inherited topology of X, there is a set $V \subset X$, open relative to X, such that $\mathbf{u}(y) \in V_u$ for all $y \in V$. Hence for all $y \in V \cap V_x \subset X$, $\Lambda\left(y, \mathbf{u}(y)\right) \subset U$. We finally note that $V \cap V_x$ is open relative to X, thus showing that $x \mapsto \Lambda\left(x, \mathbf{u}(x)\right)$ is upper semicontinuous. □

Proof of Theorem 3.12. Let $\mathbf{u} = (u_1, \ldots, u_m)$ be as above—after excising a set of measure zero, $\mathbf{u}$ may be taken to be finite everywhere in Ω. For each $\varepsilon > 0$, by Lusin's theorem, there exist measurable sets $E^i_\varepsilon \subset \Omega$, $i = 1, 2, \ldots, m$, such that $u_i\colon E^i_\varepsilon \mapsto \mathbb{R}$ is continuous and $\operatorname{meas}\left(\Omega - E^i_\varepsilon\right) < \varepsilon/k$. Setting $E_\varepsilon = E^1_\varepsilon \cap \cdots \cap E^m_\varepsilon$, we have $\mathbf{u} \in C(E_\varepsilon; \mathbb{R}^m)$ and $\operatorname{meas}\left(\Omega - E_\varepsilon\right) < \varepsilon$.

Define $A_0 \coloneqq \Omega - \bigcup_{n\in\mathbb{N}} E_{1/n}$, $A_1 \coloneqq E_1$ and $A_i \coloneqq E_{1/i} - \bigcup_{k<i} A_k$. Then, $\{A_i\}_{i=0}^\infty$ is a collection of disjoint measurable sets, $\mathbf{u} \in C\left(A_i, \mathbb{R}^m\right)$ for each $i \geq 1$ and $\operatorname{meas}(A_0) = 0$. For each $i \geq 1$, it follows from Lemma B.3 that the set-valued map $F_i\colon A_i \rightrightarrows \Lambda$, $x \mapsto \Lambda(x, \mathbf{u}(x))$ is upper semicontinuous, and hence by Lemma B.2, the preimages under F_i of open sets in Λ are Borel sets of A_i, with respect to the inherited topology of A_i. Therefore, F_i satisfies the assumptions of Theorem B.1, so we deduce that there exists a Borel measurable function $\alpha_i\colon A_i \mapsto \Lambda$ with $\alpha_i(x) \in F_i(x) = \Lambda(x, \mathbf{u}(x))$ for all $x \in A_i$. Since the sets A_i are disjoint and $\Omega = \bigcup_{i=0}^\infty A_i$, we may define $\alpha\colon \Omega \to \Lambda$ by $\alpha(x) \coloneqq \alpha_i(x)$ if $x \in A_i$, $i \geq 1$, and $\alpha(x) \coloneqq \beta$ if $x \in A_0$, for some fixed $\beta \in \Lambda$. Because A_0 has measure zero, $\alpha(x) \in \Lambda(x, \mathbf{u}(x))$ for almost every $x \in \Omega$. Note that each A_i is Lebesgue measurable, therefore the Borel subsets of the topological subspace A_i are also Lebesgue measurable as subsets of $\mathbb{R}^d$. Thus, for any open set $U \subset \Lambda$, we have $\alpha^{-1}(U) = \bigcup_{i=0}^\infty \alpha_i^{-1}(U)$, so $\alpha^{-1}(U)$ is Lebesgue measurable. □

Appendix C

Approximation theory

In this appendix, we establish the approximation theory for discontinuous finite element spaces that is used throughout this work. In particular, we show how to construct the approximation operators for Sobolev and Bochner spaces that satisfy the specific stability properties that were employed in Chapter 4.

We begin by showing a trace theorem for the Besov space $B^{1/2}_{2,1}$ that may be employed elementwise on the mesh. Then, we construct approximation operators for Sobolev spaces that are L^2-stable and that present optimal convergence rates in both the mesh size and the polynomial degree. Our approach is largely based on ideas due to Babuška and Suri in [9, 10], although our construction features modifications required to correct a significant gap contained in [9, Lemma 4.1]. The trace theorem for Besov spaces is then used to obtain optimal convergence rates for the traces of the approximation operators. Finally, we show how to extend the approximation operators from Sobolev spaces to Bochner spaces, whilst retaining the required stability properties.

Trace theorem for Besov spaces

We will show that, for a suitable domain $K \subset \mathbb{R}^d$, functions in the Besov space $B^{1/2}_{2,1}(K)$ have traces in $L^2(\partial K)$. Recall the discrete form of the J-method of interpolation of function spaces [1]: a function $u \in L^2(K)$ belongs to $B^{1/2}_{2,1}(K)$ if and only if there exists a sequence $\{u_i\}_{i\in\mathbb{Z}} \subset H^1(K)$, such that $u = \sum_{i\in\mathbb{Z}} u_i$, where the series converges absolutely in $L^2(K)$, and such that the sequence $\{2^{-i/2}J(2^i, u_i)\}_{i\in\mathbb{Z}} \in \ell^1$, where

$$J(t, v) \coloneqq \max\left[\|v\|_{L^2(K)},\, t\|v\|_{H^1(K)}\right]. \tag{C.1}$$

Moreover, we may define a norm on $B^{1/2}_{2,1}(K)$ by

$$\|u\|_{B^{1/2}_{2,1}(K)} \coloneqq \inf\left\{\|\{2^{-i/2}J(2^i, u_i)\}_{i\in\mathbb{Z}}\|_{\ell^1} : u = \sum\nolimits_{i\in\mathbb{Z}} u_i,\ u_i \in H^1(K)\right\}. \tag{C.2}$$

Also, for any such sequence, we have

$$\lim_{m\to\infty}\|u-\textstyle\sum_{|i|\le m}u_i\|_{B^{1/2}_{2,1}(K)}\le\lim_{m\to\infty}\sum_{|i|>m}2^{-i/2}J(2^i,u_i)=0. \tag{C.3}$$

Hence $H^1(K)$ is dense in $B^{1/2}_{2,1}(K)$.

It is sometimes problematic to work with the infinite series representation of a function in the Besov space $B^{1/2}_{2,1}(K)$, as a result of questions concerning convergence of the series in appropriate norms. The following lemma shows that it is possible to work with representations by finite sums of functions in the dense subspace $H^1(K)$. This result is a key ingredient of our proof of the trace theorem, and it was obtained independently; to the best of our knowledge, it appears to be original to this work.

Lemma C.1. *Let $K\subset\mathbb{R}^d$ be a domain. Then, for each $u\in H^1(K)$, there exists a positive integer m and a finite set $\{u_i\}_{|i|\le m}\subset H^1(K)$, such that*

$$u=\sum_{|i|\le m}u_i,\quad\sum_{|i|\le m}2^{-i/2}J(2^i,u_i)\le\left(2+\sqrt{2}\right)\|u\|_{B^{1/2}_{2,1}(K)}, \tag{C.4}$$

Proof. Since the case $u=0$ is trivial, we assume that $u\neq 0$. Since $H^1(K)$ is embedded in the space $B^{1/2}_{2,1}(K)$, there exists a sequence $\{v_i\}_{i\in\mathbb{Z}}\subset H^1(K)$ such that $u=\sum_{i\in\mathbb{Z}}v_i$, and such that

$$\|\{2^{-i/2}J(2^i,v_i)\}_i\|_{\ell^1}=\sum_{i\in\mathbb{Z}}2^{-i/2}\,J(2^i,v_i)\le\sqrt{2}\,\|u\|_{B^{1/2}_{2,1}(K)}.$$

The series $\sum_{i\in\mathbb{Z}}v_i$ converges absolutely to u in $L^2(K)$, since

$$\sum_{i\in\mathbb{Z}}\|v_i\|_{L^2(K)}\le\sum_{i\in\mathbb{Z}}2^{-i/2}J(2^i,v_i)\le\sqrt{2}\,\|u\|_{B^{1/2}_{2,1}(K)}.$$

Let $m\ge 1$ be the smallest integer such that $\|u\|_{H^1(K)}\le 2^{m/2}\,\|u\|_{B^{1/2}_{2,1}(K)}$; note that m exists since $u\neq 0$, which implies that $\|u\|_{H^1(K)}$ and $\|u\|_{B^{1/2}_{2,1}(K)}$ are strictly positive.

It is then found that

$$\|u-\textstyle\sum_{|i|<m}v_i\|_{L^2(K)}\le\sum_{|i|\ge m}\|v_i\|_{L^2(K)}\le 2^{-m/2}\sum_{|i|\ge m}2^{-i/2}J(2^i,v_i), \tag{C.5}$$

$$\sum_{|i|<m}\|v_i\|_{H^1(K)}\le 2^{(m-1)/2}\sum_{|i|<m}2^{i/2}\,\|v_i\|_{H^1(K)}\le 2^{(m-1)/2}\sum_{|i|<m}2^{-i/2}\,J(2^i,v_i). \tag{C.6}$$

Now, define $u_i\coloneqq v_i$ for $|i|<m$, and $u_{-m}\coloneqq u-\sum_{|i|<m}u_i$, whilst $u_i\coloneqq 0$ otherwise. By hypothesis, $u\in H^1(K)$, so $u_{-m}\in H^1(K)$, and we have $u=\sum_{|i|\le m}u_i$.

It follows from (C.5) that

$$2^{m/2}\|u_{-m}\|_{L^2(K)}=2^{m/2}\|u-\textstyle\sum_{|i|<m}v_i\|_{L^2(K)}\le\sqrt{2}\,\|u\|_{B^{1/2}_{2,1}(K)}. \tag{C.7}$$

Moreover, the choice of the integer m and (C.6) imply that

$$2^{-m/2}\|u_{-m}\|_{H^1(K)} \leq 2^{-m/2}\left(\|u\|_{H^1(K)} + \sum\nolimits_{|i|<m}\|v_i\|_{H^1(K)}\right) \leq 2\,\|u\|_{B^{1/2}_{2,1}(K)}.$$

Therefore, $2^{m/2}J(2^{-m}, u_{-m}) \leq 2\,\|u\|_{B^{1/2}_{2,1}(K)}$, and we find that

$$\sum_{|i|\leq m} 2^{-i/2}J(2^i, u_i) \leq 2^{m/2}J(2^{-m}, u_m) + \sum_{|i|<m} 2^{-i/2}J(2^i, v_i) \leq \left(2+\sqrt{2}\right)\|u\|_{B^{1/2}_{2,1}(K)}. \tag{C.8}$$

Therefore, (C.4) holds, and thus the sequence $\{u_i\}_{|i|\leq m}$ fulfills all of the above claims. □

A significant aspect of Lemma C.1 is that the constant appearing in the inequality of (C.4) is entirely independent of all other quantities, including u, m, and K.

Theorem C.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes on Ω. Then, for each $\mathcal{T}_h$ and each $K \in \mathcal{T}_h$, the trace operator $\gamma\colon H^1(K) \to L^2(\partial K)$ has a unique extension to a bounded linear operator on $B^{1/2}_{2,1}(K)$, and there holds*

$$\|\gamma u\|_{L^2(\partial K)} \lesssim \|u\|_{B^{1/2}_{2,1}(K)} + h_K^{-1/2}\|u\|_{L^2(K)} \qquad \forall\, u \in B^{1/2}_{2,1}(K). \tag{C.9}$$

Proof. For an element $K \in \mathcal{T}_h$, let $\gamma\colon H^1(K) \to L^2(\partial K)$ denote the trace operator. First, we claim that

$$\|\gamma u\|_{L^2(\partial K)} \lesssim \|u\|_{B^{1/2}_{2,1}(K)} + h_K^{-1/2}\|u\|_{L^2(K)} \quad \forall\, u \in H^1(K). \tag{C.10}$$

For a given $u \in H^1(K)$, Lemma C.1 shows that there exists a finite set $\{u_i\}_{|i|\leq m} \subset H^1(K)$ such that (C.4) holds. Since $\{\mathcal{T}_h\}_h$ is a shape-regular sequence of simplicial or parallelepipedal meshes, we have the multiplicative trace inequality (c.f. [27, 58])

$$\|\gamma u\|_{L^2(\partial K)} \lesssim \left(|u|_{H^1(K)} + h_K^{-1}\|u\|_{L^2(K)}\right)^{1/2}\|u\|^{1/2}_{L^2(K)} \quad \forall\, u \in H^1(K), \tag{C.11}$$

where the constant depends only the dimension d and the shape-regularity of $\{\mathcal{T}_h\}_h$. We remark that the multiplicative trace inequality was proven for the case of triangles in two dimensions in [58], and can be extended to simplices and parallelepipeds in $\mathbb{R}^d$, see [27].

Let $\overline{u}$ denote the mean-value of u over K, and note that $\|u - \overline{u}\|_{L^2(K)} \lesssim h_K|u|_{H^1(K)}$.

Then, $u - \overline{u} = \sum_{|i|\leq m}(u_i - \overline{u}_i)$, and (C.11) implies that

$$\begin{aligned}\|\gamma(u-\overline{u})\|_{L^2(\partial K)} &\lesssim \sum_{|i|\leq m} \left(|u_i|_{H^1(K)} + h_K^{-1}\|u_i - \overline{u}_i\|_{L^2(K)}\right)^{1/2} \|u_i - \overline{u}_i\|_{L^2(K)}^{1/2} \\ &\lesssim \sum_{|i|\leq m} |u_i|_{H^1(K)}^{1/2} \|u_i\|_{L^2(K)}^{1/2} \\ &\lesssim \sum_{|i|\leq m} 2^{-i/2}\|u_i\|_{L^2(K)} + 2^{i/2}\|u_i\|_{H^1(K)} \\ &\lesssim \sum_{|i|\leq m} 2^{-i/2} J(2^i, u_i) \lesssim \|u\|_{B^{1/2}_{2,1}(K)}.\end{aligned} \tag{C.12}$$

It is also easily found that $\|\gamma\overline{u}\|_{L^2(\partial K)} \lesssim h_K^{-1/2}\|u\|_{L^2(K)}$. Therefore, the bound (C.10) follows from the above bounds and the triangle inequality. Thus, the trace operator γ is uniformly bounded in the norm of $B^{1/2}_{2,1}(K)$ over the space $H^1(K)$. Since $H^1(K)$ is densely embedded in $B^{1/2}_{2,1}(K)$, it follows that γ has a unique extension to a bounded linear operator $\gamma\colon B^{1/2}_{2,1}(K) \to L^2(\partial K)$, and that (C.9) holds. □

In the following, we will often omit any explicit reference to the trace operator γ. For example, we shall write $\|u\|_{L^2(\partial K)}$ rather than $\|\gamma u\|_{L^2(\partial K)}$.

Polynomial approximation in Sobolev spaces

For a positive integer d and a nonnegative integer p, let $\mathcal{P}_p$ denote the space of real valued polynomials on $\mathbb{R}^d$ with either partial or total degree at most p. The following result is well-known, although we include it for completeness.

Lemma C.3. *For a nonnegative integer p and $\rho \in \mathbb{R}_{>0}$, a function $u\colon (-\rho, \rho) \to \mathbb{R}$ is an algebraic polynomial of degree at most p if and only if the function $V\colon \xi \mapsto u(\rho \sin\xi)$ is a trigonometric polynomial of degree at most p.*

Proof. Suppose that u is an algebraic polynomial of degree at most p. Then it is easily found that V is a trigonometric polynomial of degree at most p. To show the converse, suppose that V is a trigonometric polynomial of degree at most p. Observe that V is necessarily symmetric about $\pm\pi/2$, and thus we have, for any $k \geq 0$,

$$\int_{-\pi}^{\pi} V(\xi)\sin(2k\xi)\,\mathrm{d}\xi = 0, \qquad \int_{-\pi}^{\pi} V(\xi)\cos((2k+1)\,\xi)\,\mathrm{d}\xi = 0. \tag{C.13}$$

Indeed, the first identity in (C.13) is found by writing

$$\begin{aligned}\int_{-\pi}^{\pi} V(\xi)\sin(2k\xi)\,\mathrm{d}\xi &= \int_0^{\pi} (V(\xi) - V(-\xi))\sin(2k\xi)\,\mathrm{d}\xi \\ &= (-1)^k \int_{-\pi/2}^{\pi/2} (V(\tfrac{\pi}{2}+\delta) - V(-\tfrac{\pi}{2}+\delta))\sin(2k\delta)\,\mathrm{d}\delta,\end{aligned} \tag{C.14}$$

and by noting that the right-hand side of (C.14) is the integral of an odd function over an interval centred about $\delta = 0$, as a result of the symmetry of V. The proof of the second identity in (C.13) is analogous.

Since V is a trigonometric polynomial of degree at most p, it follows from (C.13) that

$$V(\xi) = \sum_{1 \le 2k+1 \le p} a_k \sin((2k+1)\,\xi) + \sum_{0 \le 2k \le p} b_k \cos(2k\xi).$$

For $x \in (-\rho, \rho)$ and $k \ge 0$, define $P_{2k+1}(x) \coloneqq \sin((2k+1)\arcsin(x/\rho))$ and $Q_{2k}(x) \coloneqq \cos(2k \arcsin(x/\rho))$. So, for example, $Q_0(x) = 1$, $P_1(x) = x$, and $Q_2(x) = 1 - 2x^2$. Therefore, u may be written as $u(x) = \sum_{1 \le 2k+1 \le p} a_k\, P_{2k+1}(x) + \sum_{0 \le 2k \le p} b_k\, Q_{2k}(x)$. The recurrence relations $P_{2k+1}(x) = P_{2k-1}(x) + 2\,x\,Q_{2k}(x)$ and $Q_{2k+2}(x) = 2\,Q_2(x)\,Q_{2k}(x) - Q_{2k-2}(x)$, for all $k \ge 1$, allow us to deduce that $P_{2k+1} \in \mathcal{P}_{2k+1}$ and that $Q_{2k} \in \mathcal{P}_{2k}$ for each $k \ge 0$, where $\mathcal{P}_p$ denotes here the space of univariate polynomials of degree at most p. It then follows that $u \in \mathcal{P}_p$. □

Theorem C.4. *Let $Q \subset [-1,1]^d$ be either the unit hypercube or the unit simplex in $\mathbb{R}^d$, $d \ge 1$. For each integer $p \ge 0$, there exists a linear operator $\Pi^p \colon L^2(Q) \to \mathcal{P}_p$, with the following properties. There is a constant C, independent of p, such that*

$$\|\Pi^p u\|_{L^2(Q)} \le C \|u\|_{L^2(Q)} \quad \forall\, u \in L^2(Q). \tag{C.15}$$

For nonnegative integers $j \le s$, there is a constant C, independent of p but dependent on s, such that

$$\|u - \Pi^p u\|_{H^j(Q)} \le C (p+1)^{-(s-j)} \|u\|_{H^s(Q)} \quad \forall\, u \in H^s(Q). \tag{C.16}$$

Although the proof of Theorem C.4 is similar to the proof of [10, Lemma 3.1], we include it here in order to highlight a surprising property: we show below that generally $u \ne \Pi^p u$, even if $u \in \mathcal{P}_p$, contrary to what is claimed in [9, Lemma 4.1]. This implies that the approximation operator of [9, 10] does not preserve polynomials, and thus cannot be employed in conjunction with the Bramble–Hilbert lemma to obtain optimal error bounds for *hp*-version finite element spaces.

Proof. First, we momentarily assume that $\mathcal{P}_p$ denotes the space of polynomials of partial degree at most p. Since Q is a Lipschitz domain, the Stein extension theorem [1] shows that there exists a linear total extension operator $E \colon L^2(Q) \to L^2(\mathbb{R}^d)$, such that, for each nonnegative integer s, $\|Eu\|_{H^s(\mathbb{R}^d)} \lesssim \|u\|_{H^s(Q)}$ for all $u \in H^s(Q)$.

For $\rho \in \mathbb{R}_{>0}$, let $Q(\rho) \coloneqq [-\rho, \rho]^d$. Without loss of generality, we may assume that $\operatorname{supp} Eu \subset Q(3/2)$ for every $u \in L^2(Q)$. Let Φ be the diffeomorphism from $Q(\pi/2)$ to $Q(2)$ defined by $\Phi(\xi) \coloneqq (2\sin\xi_1, \dots, 2\sin\xi_d)$. For $u \in L^2(Q)$, let $V(\xi) \coloneqq Eu(\Phi(\xi))$ for $\xi \in \mathbb{R}^d$. It follows that V is a 2π-periodic function that is symmetric about each hyperplane $\xi_i = \pm\pi/2$,

i.e. for any $\xi \in \mathbb{R}^d$ such that $\xi_i = \pm\pi/2$ and any $\delta \in \mathbb{R}$, we have $V(\xi + \delta\, e_i) = V(\xi - \delta\, e_i)$, where e_i is the i-th unit vector. Since $\operatorname{supp} Eu \subset Q(3/2)$, we may use the symmetry of V to show that, for any integer $s \geq 0$ and any $u \in H^s(Q)$, we have

$$\|V\|^2_{H^s(Q(\pi))} = 2^d \|V\|^2_{H^s(Q(\pi/2))} = 2^d \|V\|^2_{H^s(\Phi^{-1}(Q(3/2)))}.$$

Therefore, we deduce that $\|V\|_{H^s(Q(\pi))} \lesssim \|u\|_{H^s(Q)}$ for all $u \in H^s(Q)$ and all integers $s \geq 0$.

Since the function $V \in L^2(Q(\pi))$, it admits the Fourier expansion

$$V(\xi) = \sum_{k \in \mathbb{Z}^d} a_k \, \mathrm{e}^{\mathrm{i}\, k \cdot \xi} \quad \text{for a.e. } \xi \in Q(\pi), \tag{C.17}$$

where the coefficients $a_k \in \mathbb{C}$ satisfy $\overline{a_k} = a_{-k}$, for each $k \in \mathbb{Z}^d$, because V is real-valued. For an integer $p \geq 0$, define the trigonometric polynomial V_p by $V_p(\xi) \coloneqq \sum_{|k|_\infty \leq p} a_k \, \mathrm{e}^{\mathrm{i}\, k \cdot \xi}$. The relation $\overline{a_k} = a_{-k}$ shows that

$$V_p(\xi) = a_0 + \sum_{\substack{k \in \mathbb{N}^d \setminus \{0\} \\ |k|_\infty \leq p}} \tfrac{1}{2}(a_k + \overline{a_k})(\mathrm{e}^{\mathrm{i}\, k \cdot \xi} + \mathrm{e}^{-\mathrm{i}\, k \cdot \xi}) + \tfrac{1}{2}(a_k - \overline{a_k})(\mathrm{e}^{\mathrm{i}\, k \cdot \xi} - \mathrm{e}^{-\mathrm{i}\, k \cdot \xi}),$$

thus implying that V_p is real-valued. For any integers $j \leq s$, and any $u \in H^s(Q)$,

$$\begin{aligned} |V - V_p|^2_{H^j(Q(\pi))} &\lesssim \sum_{|k|_\infty > p} |k|_\infty^{2j} |a_k|^2 \lesssim (p+1)^{-2(s-j)} \sum_{k \in \mathbb{Z}^d} |k|_\infty^{2s} |a_k|^2 \\ &\lesssim (p+1)^{-2(s-j)} |V|^2_{H^s(Q(\pi))} \lesssim (p+1)^{-2(s-j)} \|u\|^2_{H^s(Q)}, \end{aligned} \tag{C.18}$$

where the constants are independent of u and p.

Define the linear map $\Pi^p \colon L^2(Q) \to L^2_{\mathrm{loc}}(Q(2))$ by $\Pi^p u \coloneqq V_p \circ \Phi^{-1}$. Since the mapping $\Phi \colon Q(\pi/2) \to Q(2)$ is a diffeomorphism, and since Q is compactly contained in $Q(2)$, we find that, for any $u \in L^2(Q)$,

$$\|\Pi^p u\|_{L^2(Q)} \lesssim \|V_p\|_{L^2(Q(\pi/2))} \leq \|V\|_{L^2(Q(\pi))} \lesssim \|u\|_{L^2(Q)},$$

where the constants are independent of u and p, thus giving (C.15). Likewise, (C.16) follows from (C.18) and from $\|u - \Pi^p u\|_{H^j(Q)} \lesssim \|V - V_p\|_{H^j(Q(\pi))}$.

In order to show that $\Pi^p u$ is a polynomial of partial degree at most p, it is enough to show that the univariate functions $x_i \mapsto \Pi^p u(x_1, \ldots, x_i, \ldots, x_d)$ are polynomials of degree at most p, for each $x \in Q(2)$. However, this follows from Lemma C.3 because the trigonometric polynomial $V_p = \Pi^p u \circ \Phi$ has partial degree at most p. This completes the proof for the case where $\mathcal{P}_p$ denotes the space of polynomials of partial degree at most p.

The above results extend to the case where $\mathcal{P}_p$ denotes the space of polynomials of total degree p, since the space of polynomials of partial degree at most k is contained in the space of polynomials of total degree at most p whenever $k \leq p/d$. So, we may

choose $k \leq p/d \leq k+1$, and we find that the projector Π^k defined above has the required properties. □

We now show that Π^p is *inexact* when applied to polynomials: in general, $u \neq \Pi^p u$ is possible for $u \in \mathcal{P}_p$. To show this, consider the special case where $d = 1$ and $u \equiv 1$. Since Eu is compactly supported on $Q(3/2)$ and is not identically zero, Eu is necessarily not a polynomial of finite degree on $Q(2)$. Since $V(\xi) = Eu(2\sin\xi)$, Lemma C.3 shows that V is not a trigonometric polynomial of finite degree, and we also have $\|V-1\|_{L^2(Q(\pi))} > 0$. By convergence of Fourier series, there exists a $p_0 \geq 0$ such that for all $p \geq p_0$, we have $\|V - V_p\|_{L^2(Q(\pi))} < \frac{1}{2}\|V-1\|_{L^2(Q(\pi))}$, so that

$$\|V_p - 1\|_{L^2(Q(\pi))} > \tfrac{1}{2}\|V-1\|_{L^2(Q(\pi))} > 0. \tag{C.19}$$

Since nonzero trigonometric polynomials have at most finitely many roots, V_p cannot be identically equal to 1 on any open subset of $Q(\pi)$, because otherwise V_p would have to be identically equal to 1 on $Q(\pi)$, thereby contradicting (C.19). Therefore, $V_p \not\equiv 1 \equiv V$ on $\Phi^{-1}(Q)$, and thus $u \neq \Pi^p u$ on Q.

We note that the polynomial inexactness of the Babuška–Suri projector, as defined in [9, 10], is independent of the choice of the extension operator, since it results from the requirement that the extended functions have compact support. This requirement is not easily avoided, since it is used to obtain the bound $\|V\|_{H^s(Q(\pi))} \lesssim \|u\|_{H^s(Q)}$.

Lemma C.5. *Let $Q \subset [-1,1]^d$ be either the unit hypercube or the unit simplex in $\mathbb{R}^d$, $d \geq 1$. For each pair of nonnegative integers p and m, there exists a linear operator $\Pi^{m,p}\colon L^2(Q) \to \mathcal{P}_p$, the space of polynomials with partial degree at most p, such that $\Pi^{m,p}$ has the following properties. If u is a polynomial of total degree at most $\min(m,p)$, then $\Pi^{m,p}u = u$. There exists a constant C, independent of p and m, such that*

$$\|\Pi^{m,p}u\|_{L^2(Q)} \leq C\|u\|_{L^2(Q)} \qquad \forall\, u \in L^2(Q). \tag{C.20}$$

For any nonnegative integer s, there is a constant C, independent of p but dependent on s and m, such that for each nonnegative integer $j \leq s$,

$$\|u - \Pi^{m,p}u\|_{H^j(Q)} \leq C(p+1)^{-(s-j)} \sum_{r=t}^{s} |u|_{H^r(Q)} \quad \forall\, u \in H^s(Q), \tag{C.21}$$

where $t \coloneqq \min(s, p+1, m+1)$.

Proof. For nonnegative integers m and p, let Π^p be the Babuška–Suri projector as given by Theorem C.4, and let $\Pi_{L^2}^{\min(m,p)}\colon L^2(Q) \to \mathcal{P}_{\min(m,p)}$ denote the L^2 projection into the space of polynomials of total degree at most $\min(m,p)$. Then, define

$$\Pi^{m,p}u \coloneqq \Pi_{L^2}^{\min(m,p)}u + \Pi^p\left(u - \Pi_{L^2}^{\min(m,p)}u\right), \quad u \in L^2(Q). \tag{C.22}$$

It follows that $\Pi^{m,p}$ is a well-defined linear operator mapping $L^2(Q)$ into $\mathcal{P}_p$. Since Π^p is a linear operator, we see that $\Pi^{m,p}$ is exact on the space of polynomials of total degree at most $\min(m,p)$. To show (C.20), we use the triangle inequality

$$\|\Pi^{m,p}u\|_{L^2(Q)} \leq \|\Pi_{L^2}^{\min(m,p)}u\|_{L^2(Q)} + \|\Pi^p\|_{L^2(Q)\to L^2(Q)}\|u - \Pi_{L^2}^{\min(m,p)}u\|_{L^2(Q)}, \tag{C.23}$$

and we note that, by (C.15), $\|\Pi^p\|_{L^2(Q)\to L^2(Q)} \leq C$, with C independent of p, and that $\|\Pi_{L^2}^{\min(m,p)}\|_{L^2(Q)\to L^2(Q)} \leq 1$. Now, let $j \leq s$ be nonnegative integers, and apply (C.16) to obtain

$$\|u - \Pi^{m,p}u\|_{H^j(Q)} \leq C(p+1)^{-(s-j)}\|u - \Pi_{L^2}^{\min(m,p)}u\|_{H^s(Q)} \quad \forall\, u \in H^s(Q), \tag{C.24}$$

where C is independent of p and m but dependent on s. Since Q is the unit simplex or unit hypercube, the Bramble–Hilbert lemma [17] shows that

$$\|u - \Pi_{L^2}^{\min(m,p)}u\|_{H^s(Q)} \leq C\sum_{r=t}^{s}|u|_{H^r(Q)} \quad \forall\, u \in H^s(Q), \tag{C.25}$$

where $t \coloneqq \min(s, \min(m,p)+1)$ and C depends on s, $\min(m,p)$ and on Q. Moreover, by considering separately the cases $p < m$ and $p \geq m$, it is seen that we may choose the constant in (C.25) to depend only on m, and not on p. We thus obtain (C.21) by combining (C.24) and (C.25), and noting that the constant may be chosen to be independent of p. □

Definition of fractional order Sobolev spaces. For a domain K and a real number $s > 0$ such that $s \in (r, r+1)$ for a nonnegative integer r, we define

$$H^s(K) \coloneqq \Big(H^r(K), H^{r+1}(K)\Big)_{s-r,2;J}. \tag{C.26}$$

Here, we use the standard norm on $H^r(K)$ when r is an integer. It follows from the equivalence theorem [1] that $H^s(K) = \big(H^r(K), H^{r+1}(K)\big)_{s-r,2;K}$, where the constant in the equivalence of norms depends only on s. Also, in view of the reiteration theorem, we note that

$$\Big(H^r(K), H^{r+1}(K)\Big)_{s-r,1;J} \hookrightarrow H^s(K) \hookrightarrow \Big(H^r(K), H^{r+1}(K)\Big)_{s-r,\infty;K}, \tag{C.27}$$

where the embedding constants depend only on s, see [1, Thm. 7.16, Cor. 7.20]. We remark that it is important in the following that these constants are independent of the domain K.

Theorem C.6. *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz polytopal domain, and let $\{\mathcal{T}_h\}_h$ be a shape-regular sequence of simplicial or parallelepipedal meshes on Ω. For each mesh $\mathcal{T}_h$, suppose that $h = \max_{K\in\mathcal{T}_h} h_K$, where $h_K \coloneqq \operatorname{diam} K$ for all $K \in \mathcal{T}_h$. For each mesh $\mathcal{T}_h$, let $\mathbf{m} = (m_K;\ K \in \mathcal{T}_h)$ and $\mathbf{p} = (p_K;\ K \in \mathcal{T}_h)$ be vectors of nonnegative integers. Then,*

there exists a sequence of linear operators $\{\Pi_h^{\mathbf{m},\mathbf{p}}\}_h$*, such that* $\Pi_h^{\mathbf{m},\mathbf{p}}\colon L^2(\Omega) \to V_{h,\mathbf{p}}$*, with* $\Pi_h^{\mathbf{m},\mathbf{p}} u|_K = u|_K$ *if* $u|_K$ *is a polynomial of total degree at most* $\min(m_K, p_K)$*, and such that, for each* $K \in \mathcal{T}_h$*,*

$$\|\Pi_h^{\mathbf{m},\mathbf{p}} u\|_{L^2(K)} \lesssim \|u\|_{L^2(K)} \qquad \forall\, u \in L^2(K). \tag{C.28}$$

Also, for each $K \in \mathcal{T}_h$*,* $s_K \in \mathbb{R}_{\geq 0}$*, each nonnegative integer* $j \leq s_K$ *and, if* $s_K > 1/2$*, for each multi-index* β*, with* $|\beta| < s_K - 1/2$*, we have*

$$\|u - \Pi_h^{\mathbf{m},\mathbf{p}} u\|_{H^j(K)} \lesssim \frac{h_K^{t_K - j}}{(p_K+1)^{s_K - j}} \|u\|_{H^{s_K}(K)} \qquad \forall\, u \in H^{s_K}(K), \tag{C.29}$$

$$\|D^\beta (u - \Pi_h^{\mathbf{m},\mathbf{p}} u)\|_{L^2(\partial K)} \lesssim \frac{h_K^{t_K - |\beta| - 1/2}}{(p_K+1)^{s_K - |\beta| - 1/2}} \|u\|_{H^{s_K}(K)} \qquad \forall\, u \in H^{s_K}(K), \tag{C.30}$$

where $t_K \coloneqq \min(s_K, p_K + 1, m_K + 1)$*.*

Proof. Since the meshes $\{\mathcal{T}_h\}$ consist of simplices or parallelepipeds, each element K is affine-equivalent to the unit simplex or unit hypercube, with a corresponding affine mapping $F_K\colon K \to Q$. For each $K \in \mathcal{T}_h$, define $\hat{u} = u \circ F_K^{-1}$ and $\Pi_h^{\mathbf{m},\mathbf{p}} u|_K = (\Pi^{m_K, p_K} \hat{u}) \circ F_K \in \mathcal{P}_{p_K}$, where Π^{m_K, p_K} is the operator given by Lemma C.5. The stability bound (C.28) then follows from the shape-regularity of the mesh and from the bound (C.20) of Lemma C.5.

Also, for any nonnegative integers $j \leq s_K$, we have

$$\|u - \Pi_h^{\mathbf{m},\mathbf{p}} u\|_{H^j(K)} \lesssim \frac{h_K^{t_K - j}}{(p_K+1)^{s_K - j}} \|u\|_{H^{s_K}(K)} \quad \forall\, u \in H^{s_K}(K), \tag{C.31}$$

where $t_K = \min(s_K, p_K + 1, m_K + 1)$ and where the constant depends only on s_K, m_K, on $\max h$ the maximum mesh size over all meshes, on the reference element and on the shape-regularity of $\{\mathcal{T}_h\}$. We remark that the additional dependence on $\max h$ stems from the fact that we use the bound $h_K^{t_K - i} \leq \max h^{j-i}\, h_K^{t_K - j}$, $i \leq j$, to obtain (C.31). The exact interpolation theorem [1] shows that (C.31) extends to each nonnegative integer j and each nonnegative real number s_K such that $j \leq s_K$, thus giving (C.29).

We now show (C.30). Let $s_K > 1/2$ and β be a multi-index with $|\beta| < s_K - 1/2$. First, consider the case where $|\beta| \leq s_K - 1$. Then, (C.30) follows from (C.29) and from the multiplicative trace inequality (C.11). Now, consider the case where $s_K - |\beta| \in (\frac{1}{2}, 1)$. Theorem C.2 shows that, for any $u \in H^{s_K}(K)$,

$$\|D^\beta (u - \Pi_h^{\mathbf{m},\mathbf{p}} u)\|_{L^2(\partial K)} \lesssim \|D^\beta (u - \Pi_h^{\mathbf{m},\mathbf{p}} u)\|_{B^{1/2}_{2,1}(K)} + h_K^{-1/2} \|u - \Pi_h^{\mathbf{m},\mathbf{p}} u\|_{H^{|\beta|}(K)}.$$

Given (C.29) for the case $j = |\beta|$, we can obtain (C.30) provided that we can show that,

for any $u \in H^{s_K}(K)$,

$$\|D^\beta(u - \Pi_h^{\mathbf{m},\mathbf{p}} u)\|_{B^{1/2}_{2,1}(K)} \lesssim \frac{h_K^{t_K - |\beta| - 1/2}}{(p_K+1)^{s_K - |\beta| - 1/2}} \|u\|_{H^{s_K}(K)}. \tag{C.32}$$

The exact interpolation theorem and (C.31) show that $\|u - \Pi_h^{\mathbf{m},\mathbf{p}} u\|_{H^{s_K}(K)} \lesssim \|u\|_{H^{s_K}(K)}$ for any $u \in H^{s_K}(K)$. The reiteration theorem [1] shows that

$$B^{1/2}_{2,1}(K) = \left(L^2(K), H^{s_K - |\beta|}(K)\right)_{\lambda,1;J},$$

where $\lambda \coloneqq \frac{1}{2(s_K - |\beta|)}$, and where the constant in the equivalence of norms depends only on $s_K - |\beta|$. Therefore, for any $u \in H^{s_K}(K)$, there holds

$$\|D^\beta(u - \Pi_h^{\mathbf{m},\mathbf{p}} u)\|_{B^{1/2}_{2,1}(K)} \lesssim \left(\frac{h_K^{t_K - |\beta|}}{(p_K+1)^{s_K - |\beta|}}\right)^{1-\lambda} \|u\|_{H^{s_K}(K)}.$$

Since $t_K \le s_K$, we have $(t_K - |\beta|)(1-\lambda) \ge t_K - |\beta| - 1/2$, and therefore we deduce (C.32) and (C.30). □

Polynomial approximation in Bochner spaces

To simplify the notation in the following approximation results, let the spaces $\{X_\ell\}_{\ell=0}^2$ be defined by

$$X_0 \coloneqq L^2(\Omega), \quad X_1 \coloneqq H_0^1(\Omega), \quad X_2 \coloneqq H = H^2(\Omega) \cap H_0^1(\Omega).$$

The approximation theory for Sobolev spaces can be extended to Bochner spaces as follows.

Lemma C.7. *Let I be an open interval and let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain. Let $\{\psi_k\}_{k=1}^\infty \subset H \coloneqq H^2(\Omega) \cap H_0^1(\Omega)$ be an orthonormal basis of $L^2(\Omega)$, such that $\{\psi_k\}_{k=1}^\infty$ is also an orthogonal basis of $H_0^1(\Omega)$ and of H, which satisfies*

$$\int_\Omega \psi_k \, \psi_j \, \mathrm{d}x = \delta_{kj}, \quad \int_\Omega \nabla\psi_k \cdot \nabla\psi_j \, \mathrm{d}x = \lambda_k \, \delta_{kj}, \quad \int_\Omega \Delta\psi_k \, \Delta\psi_j \, \mathrm{d}x = \lambda_k^2 \, \delta_{kj},$$

where $\lambda_k > 0$ for eack $k \in \mathbb{N}$. Then, for any $\ell \in \{0,1,2\}$, and any $u \in L^2(I; X_\ell)$, we have $u = \sum_{k=1}^\infty u_k \, \psi_k$, where $u_k(t) \coloneqq \langle u(t), \psi_k \rangle_{L^2(\Omega)}$, and where the series converges in $L^2(I; X_\ell)$. For any integer $s \ge 0$, any $u \in H^s(I; X_\ell)$, we have the generalised Parseval identity

$$|u|^2_{H^s(I;X_\ell)} = \sum_{k=1}^\infty \lambda_k^\ell \, |u_k|^2_{H^s(I)}. \tag{C.33}$$

Proof. Let $\ell \in \{0,1,2\}$ and let the function $u \in L^2(I; X_\ell)$. Then, u_k defined above is a measurable real-valued function, and $\|u_k(t)\|_{L^2(I)} \le \|u\|_{L^2(I;X_0)}$ for each $k \in \mathbb{N}$. For each

$m \in \mathbb{N}$, define the function $v_m \in L^2(I; X_2)$ by $v_m \coloneqq \sum_{k=1}^m u_k \psi_k$. Then, orthogonality of the $\{\psi_k\}_{k=1}^\infty$ in X_ℓ implies the Bessel inequality

$$\sum_{k=1}^m \lambda_k^\ell \|u_k\|_{L^2(I)}^2 = \|v_m\|_{L^2(I;X_\ell)}^2 \leq \|u\|_{L^2(I;X_\ell)}^2. \tag{C.34}$$

It can then be shown that $\{v_m\}_{m=1}^\infty$ is a Cauchy sequence in $L^2(I; X_\ell)$, with the limit denoted by v. Moreover, there exists a subsequence of $\{v_m\}_{m=1}^\infty$ which converges to v in X_ℓ pointwise almost everywhere on I. Thus, it follows from the definition of the functions v_m that $\langle v(t), \psi_k\rangle_{L^2(\Omega)} = u_k(t) = \langle u(t), \psi_k\rangle_{L^2(\Omega)}$ for each $k \in \mathbb{N}$, for a.e. $t \in I$, which shows that $v = u$, since $\{\psi_k\}_{k=1}^\infty$ is an orthonormal basis of $L^2(\Omega)$. This proves that $u = \sum_{k=1}^\infty u_k \psi_k$ and shows Parseval's identity (C.33) for the case $s = 0$.

Now, let $s \geq 1$ be an integer, and suppose $u \in H^s(I; X_\ell)$ for some $\ell \in \{0, 1, 2\}$. Let $\phi \in C_0^\infty(I)$, and compute

$$\int_I u_k \, \partial_t^s \phi \, \mathrm{d}t = \int_I \langle u, \partial_t^s(\phi \, \psi_k)\rangle_{L^2(\Omega)} \, \mathrm{d}t = (-1)^s \int_I \langle \partial_t^s u, \psi_k\rangle_{L^2(\Omega)} \, \phi \, \mathrm{d}t. \tag{C.35}$$

Therefore, the weak derivative $\partial_t^s u_k$ exists in $L^2(I)$ and $\partial_t^s u_k = \langle \partial_t^s u, \psi_k\rangle_{L^2(\Omega)}$. So, the generalised Parseval identity (C.33) for integer $s \geq 1$ is found by applying (C.33) for $s = 0$ to the function $\partial_t^s u$. □

Recall that for a Banach space X and a nonnegative integer q, the space of univariate X-valued polynomials of degree at most q is denoted by $\mathcal{Q}_q(X)$.

Lemma C.8. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain, let I be an open interval of length τ_0, and let r and q be nonnegative integers. Then, for each open interval $J \subset I$ of length $\tau \leq \tau_0$, there exists a linear operator $\Pi_\tau^{r,q}$ defined on $L^2(J; L^2(\Omega))$ with the following properties. The operator $\Pi_\tau^{r,q} \colon L^2(J; X_\ell) \to \mathcal{Q}_q(X_\ell)$ for each $\ell \in \{0, 1, 2\}$, with $\Pi_\tau^{r,q} u = u$ if $u \in \mathcal{Q}_{\min(r,q)}(X_\ell)$. Furthermore,*

$$\|\Pi_\tau^{r,q} u\|_{L^2(J;X_\ell)} \lesssim \|u\|_{L^2(J;X_\ell)} \quad \forall \, u \in L^2(J; X_\ell), \tag{C.36}$$

where the constant is independent of all quantities. For any real $\sigma \geq 0$ and any nonnegative integer $j \leq \sigma$,

$$|u - \Pi_\tau^{r,q} u|_{H^j(J;X_\ell)} \lesssim \frac{\tau^{\varrho - j}}{(q+1)^{\sigma - j}} \|u\|_{H^\sigma(J;X_\ell)} \quad \forall \, u \in H^\sigma(J; X_\ell), \tag{C.37}$$

where $\varrho \coloneqq \min(\sigma, r+1, q+1)$, and where the constant depends only on τ_0, σ and r.

Proof. Let $u \in L^2(J; L^2(\Omega))$ and define u_k, $k \in \mathbb{N}$, as in Lemma C.7. Let F denote the affine mapping from the reference element $(-1, 1)$ to J. Then, for each $k \in \mathbb{N}$, define the univariate real-valued polynomial $\Pi_\tau^{r,q} u_k \coloneqq (\Pi^{r,q} \hat{u}_k) \circ F^{-1}$, where $\hat{u}_k \coloneqq u_k \circ F$, and

where $\Pi^{r,q}$ is the approximation operator on the reference element given by Lemma C.5 for $d = 1$. For each $k \in \mathbb{N}$, $\Pi_\tau^{r,q} u_k$ has degree at most q. It follows from Lemma C.5 that $\|\Pi_\tau^{r,q} u_k\|_{L^2(J)} \lesssim \|u_k\|_{L^2(J)}$, where the constant is independent of all other quantities. Therefore, Lemma C.7 implies that $\Pi_\tau^{r,q} u \coloneqq \sum_{k=1}^\infty \Pi_\tau^{r,q} u_k \, \psi_k$ is well-defined in $L^2(J, L^2(\Omega))$. Furthermore, if $u \in L^2(J; X_\ell)$ for some $\ell \in \{0, 1, 2\}$, then Lemma C.7 shows that

$$\|\Pi_\tau^{r,q} u\|_{L^2(J;X)}^2 = \sum_{k=1}^\infty \lambda_k^\ell \|\Pi_\tau^{r,q} u_k\|_{L^2(J)}^2 \lesssim \|u\|_{L^2(J;X_\ell)}^2,$$

where the constant is independent of all quantities, thereby showing (C.36). This also implies that $\Pi_\tau^{\mathbf{r},\mathbf{q}} \colon L^2(J; X_\ell) \to \mathcal{Q}_q(X_\ell)$ for each $\ell \in \{0, 1, 2\}$. Moreover, if $u \in \mathcal{Q}_{\min(r,q)}(X_\ell)$, then $\Pi_\tau^{r,q} u_k = u_k$ for each $k \in \mathbb{N}$ by Lemma C.5, which implies that $\Pi_\tau^{r,q} u = u$ by Lemma C.7.

Let $j \le \sigma$ be nonnegative integers and let $u \in H^\sigma(J; X_\ell)$ for some $\ell \in \{0, 1, 2\}$. Then, Lemmas C.5 and C.7 imply that

$$\begin{aligned} |u - \Pi_\tau^{r,q} u|_{H^j(J;X_\ell)}^2 &= \sum_{k=1}^\infty \lambda_k^\ell \, |u_k - \Pi_\tau^{r,q} u_k|_{H^j(J)}^2 \\ &\lesssim \sum_{\nu=\varrho}^\sigma \frac{\tau^{2(\nu-j)}}{(q+1)^{2(\sigma-j)}} \sum_{k=1}^\infty \lambda_k^\ell \, |u_k|_{H^\nu(J)}^2 \lesssim \frac{\tau^{2(\varrho-j)} \max\left(1, \tau_0^{2(\sigma-\varrho)}\right)}{(q+1)^{2(\sigma-j)}} \sum_{\nu=\varrho}^\sigma |u|_{H^\nu(J;X_\ell)}^2, \end{aligned} \tag{C.38}$$

where the constant depends only on σ and r, thereby giving the bound (C.37) for the case where σ is an integer. Therefore, the bound (C.37) for general $\sigma \in \mathbb{R}_{\ge 0}$ follows from (C.38) and the theory of interpolation of function spaces. □

References

[1] R. A. Adams and J. F. Fournier, *Sobolev spaces*, vol. 140 of Pure Appl. Math., Elsevier, second ed., 2003.

[2] G. Akrivis and C. Makridakis, *Galerkin time-stepping methods for nonlinear parabolic equations*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 261–289.

[3] P. F. Antonietti and B. Ayuso, *Schwarz domain decomposition preconditioners for discontinuous Galerkin approximations of elliptic problems: non-overlapping case*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 21–54.

[4] ——, *Multiplicative Schwarz methods for discontinuous Galerkin approximations of elliptic problems*, M2AN Math. Model. Numer. Anal., 42 (2008), pp. 443–469.

[5] P. F. Antonietti and P. Houston, *A class of domain decomposition preconditioners for hp-discontinuous Galerkin finite element methods*, J. Sci. Comput., 46 (2011), pp. 124–149.

[6] P. F. Antonietti and E. Süli, *Domain decomposition preconditioning for discontinuous Galerkin approximations of convection-diffusion problems*, in Domain decomposition methods in science and engineering XVIII, vol. 70 of Lect. Notes Comput. Sci. Eng., Springer, 2009, pp. 259–266.

[7] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1749–1779.

[8] J.-P. Aubin and A. Cellina, *Differential inclusions*, vol. 264 of Grundlehren Math. Wiss., Springer-Verlag, 1984.

[9] I. Babuška and M. Suri, *The h-p version of the finite element method with quasi-uniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.

[10] ——, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[11] I. Babuška and M. Suri, *The p and h-p versions of the finite element method, basic principles and properties*, SIAM Rev., 36 (1994), pp. 578–632.

[12] G. Barles, *Solutions de viscosité des équations de Hamilton–Jacobi*, vol. 17 of Math. Appl., Springer-Verlag, 1994.

[13] G. Barles and P. Souganidis, *Convergence of approximation schemes for fully nonlinear second-order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[14] O. Bokanowski, S. Maroso, and H. Zidani, *Some convergence results for Howard's algorithm*, SIAM J. Numer. Anal., 47 (2009), pp. 3001–3026.

[15] J. F. Bonnans, É. Ottenwaelter, and H. Zidani, *A fast algorithm for the two dimensional HJB equation of stochastic control*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 723–735.

[16] J. F. Bonnans and H. Zidani, *Consistency of generalized finite difference schemes for the stochastic HJB equation*, SIAM J. Numer. Anal., 41 (2003), pp. 1008–1021.

[17] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts Appl. Math., Springer, third ed., 2008.

[18] S. C. Brenner and K. Wang, *Two-level additive Schwarz preconditioners for C^0 interior penalty methods*, Numer. Math., 102 (2005), pp. 231–255.

[19] L. A. Caffarelli and X. Cabré, *Fully nonlinear elliptic equations*, vol. 43 of Amer. Math. Soc. Colloq. Publ., American Mathematical Society, 1995.

[20] L. A. Caffarelli and L. Silvestre, *On the Evans–Krylov Theorem*, Proc. Amer. Math. Soc., 138 (2009), pp. 263–265.

[21] F. Camilli and M. Falcone, *An approximation scheme for the optimal control of diffusion processes.*, RAIRO Modél. Math. Anal. Numér. 29 (1995), pp. 97–122.

[22] P. G. Ciarlet, *Linear and nonlinear functional analysis with applications*, Society for Industrial and Applied Mathematics, 2013.

[23] H. O. Cordes, *Über die erste Randwertaufgabe bei quasilinearen Differentialgleichungen zweiter Ordnung in mehr als zwei Variablen*, Math. Ann., 131 (1956), pp. 278–312.

[24] 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., 27 (1992), pp. 1–67.

[25] M. G. Crandall and P.-L. Lions, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–41.

[26] K. Debrabant and E. R. Jakobsen, *Semi-Lagrangian schemes for linear and fully nonlinear diffusion equations*, Math. Comp., 82 (2013), pp. 1433–1462.

[27] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69 of Math. Appl., Springer, 2012.

[28] H. Dong and N. V. Krylov, *The rate of convergence of finite-difference approximations for parabolic Bellman equations with Lipschitz coefficients in cylindrical domains*, Appl. Math. Optim., 56 (2007), pp. 37–66.

[29] H. G. Eggleston, *Convexity*, vol. 47 of Camb. Tracts Math. Phys., Cambridge University Press, 1958.

[30] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers*, Numer. Math. Sci. Comput., Oxford University Press, 2005.

[31] L. C. Evans, *Partial differential equations*, vol. 19 of Grad. Stud. Math., American Mathematical Society, 1998.

[32] L. C. Evans, *Classical solutions of the Hamilton–Jacobi–Bellman equation for uniformly elliptic operators*, Trans. Amer. Math. Soc., 275 (2008), pp. 245–255.

[33] X. Feng, R. Glowinski, and M. Neilan, *Recent developments in numerical methods for fully nonlinear second order partial differential equations*, SIAM Rev., 55 (2013), pp. 205–267.

[34] X. Feng and O. A. Karakashian, *Two-level additive Schwarz methods for a discontinuous Galerkin approximation of second order elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1343–1365.

[35] ——, *Two-level non-overlapping Schwarz preconditioners for a discontinuous Galerkin approximation of the biharmonic equation*, J. Sci. Comput., 22/23 (2005), pp. 289–314.

[36] W. H. Fleming and H. M. Soner, *Controlled Markov processes and viscosity solutions*, vol. 25 of Stoch. Model. Appl. Probab., Springer, second ed., 2006.

[37] E. H. Georgoulis and E. Süli, *Optimal error estimates for the hp-version interior penalty discontinuous galerkin finite element method*, IMA J. Numer. Anal., 25 (2005), pp. 205–220.

[38] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Classics Math., Springer-Verlag, 2001.

[39] V. Girault and P.-A. Raviart, *Finite element methods for Navier-Stokes equations*, vol. 5 of Springer Ser. Comput. Math., Springer-Verlag, 1986.

[40] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 69 of Classics Appl. Math., Society of Industrial and Applied Mathematics, 2011.

[41] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), p. 865.

[42] P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a posteriori error estimation for mixed discontinuous galerkin approximations of the stokes problem*, J. Sci. Comput., 22/23 (2005), pp. 347–370.

[43] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[44] R. A. Howard, *Dynamic programming and Markov processes*, The Technology Press of M.I.T., 1960.

[45] M. Jensen and I. Smears, *Finite element methods with artificial diffusion for Hamilton–Jacobi–Bellman equations*, in Numerical Mathematics and Advanced Applications 2011, Springer, 2013, pp. 267–274.

[46] ——, *On the convergence of finite element methods for Hamilton–Jacobi–Bellman equations*, SIAM J. Numer. Anal., 51 (2013), pp. 137–162.

[47] M. Kocan, *Approximation of viscosity solutions of elliptic partial differential equations on minimal grids*, Numer. Math., 72 (1995), pp. 73–92.

[48] N. V. Krylov, *Boundedly inhomogeneous elliptic and parabolic equations*, Izv. Akad. Nauk SSSR Ser. Mat., 46 (1982), pp. 487–523, 670.

[49] H. J. Kuo and N. S. Trudinger, *Linear elliptic difference inequalities with random coefficients*, Math. Comp., 55 (1990), pp. 37–53.

[50] ——, *Discrete methods for fully nonlinear elliptic equations*, SIAM J. Numer. Anal., 29 (1992), pp. 123–135.

[51] K. Kuratowski and C. Ryll-Nardzewski, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[52] H. J. Kushner, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.

[53] O. Lakkis and T. Pryer, *A finite element method for second order nonvariational elliptic problems*, SIAM J. Sci. Comput., 33 (2011), pp. 786–801.

[54] ——, *A finite element method for nonlinear elliptic problems*, SIAM J. Sci. Comput., 35 (2013), pp. A2025–A2045.

[55] C. Lasser and A. Toselli, *An overlapping domain decomposition preconditioner for a class of discontinuous Galerkin approximations of advection-diffusion problems*, Math. Comp., 72 (2003), pp. 1215–1238.

[56] A. Maugeri, D. K. Palagachev, and L. G. Softova, *Elliptic and parabolic equations with discontinuous coefficients*, vol. 109 of Math. Res., Wiley-VCH Verlag, 2000.

[57] J. M. Melenk, *hp-finite element methods for singular perturbations*, vol. 1796 of Lecture Notes in Mathematics, Springer-Verlag, 2002.

[58] P. Monk and E. Süli, *The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals*, SIAM J. Numer. Anal., 36 (1999), pp. 251–274.

[59] T. S. Motzkin and W. Wasow, *On the approximation of linear elliptic differential equations by difference equations with positive coefficients*, J. Math. Physics, 31 (1953), pp. 253–259.

[60] I. Mozolevski, E. Süli, and P. R. Bösing, *hp-version a priori error analysis of interior penalty discontinuous Galerkin finite element approximations to the biharmonic equation*, J. Sci. Comput., 30 (2007), pp. 465–491.

[61] R. H. Nochetto and W. Zhang, *Discrete ABP estimate and convergence rates for linear elliptic equations in nondivergence form*, ArXiv e-prints, (2014).

[62] A. M. Oberman, *Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems*, SIAM J. Numer. Anal., 44 (2006), pp. 879–895.

[63] B. Øksendal, *Stochastic differential equations*, Universitext, Springer-Verlag, sixth ed., 2003.

[64] M. L. Puterman and S. L. Brumelle, *On the convergence of policy iteration in stationary dynamic programming*, Math. Oper. Res., 4 (1979), pp. 60–69.

[65] M. Renardy and R. C. Rogers, *An introduction to partial differential equations*, vol. 13 of Texts Appl. Math., Springer-Verlag, second ed., 2004.

[66] H. Royden and P. Fitzpatrick, *Real Analysis*, Prentice Hall, fourth ed., 2010.

[67] D. Schötzau and C. Schwab, *Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method*, SIAM J. Numer. Anal., 38 (2000), pp. 837–875.

[68] C. Schwab, *p- and hp-finite element methods*, Numer. Math. Sci. Comput., Oxford University Press, 1998.

[69] I. Smears, *Nonoverlapping domain decomposition preconditioners for discontinuous Galerkin finite element methods in H^2-type norms*, ArXiv e-prints, (2014).

[70] I. Smears and E. Süli, *Discontinuous Galerkin finite element approximation of non-divergence form elliptic equations with Cordes coefficients*, SIAM J. Numer. Anal., 51 (2013), pp. 2088–2106.

[71] I. Smears and E. Süli, *Discontinuous Galerkin finite element approximation of Hamilton–Jacobi–Bellman equations with Cordes coefficients*, SIAM J. Numer. Anal., 52 (2014), pp. 993–1016.

[72] I. Smears and E. Süli, *Discontinuous Galerkin finite element methods for time-dependent Hamilton–Jacobi–Bellman equations with Cordes coefficients*, to appear in Numerische Mathematik, (2015).

[73] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain decomposition*, Cambridge University Press, 1996.

[74] V. Thomée, *Galerkin finite element methods for parabolic problems*, vol. 25 of Springer Ser. Comp. Math., Springer-Verlag, second ed., 2006.

[75] A. Toselli and O. Widlund, *Domain decomposition methods*, vol. 34 of Springer Ser. Comp. Math., Springer-Verlag, 2005.

[76] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2002), pp. 805–842 (2003).

[77] L. Wang, *On the regularity theory of fully nonlinear parabolic equations. I*, Comm. Pure Appl. Math., 45 (1992), pp. 27–76.

[78] T. P. Wihler, P. Frauenfelder, and C. Schwab, *Exponential convergence of the hp-DGFEM for diffusion problems*, Comput. Math. Appl., 46 (2003), pp. 183–205.

[79] J. Wloka, *Partial differential equations*, Cambridge University Press, 1987.