

# DISCONTINUOUS GALERKIN FINITE ELEMENT APPROXIMATION OF HAMILTON–JACOBI–BELLMAN EQUATIONS WITH CORDÈS COEFFICIENTS

IAIN SMEARS\* AND ENDRE SÜLI†

**Abstract.** We propose an  $hp$ -version discontinuous Galerkin finite element method for fully nonlinear second-order elliptic Hamilton–Jacobi–Bellman equations with Cordès coefficients. The method is proven to be consistent and stable, with convergence rates that are optimal with respect to mesh size, and suboptimal in the polynomial degree by only half an order. Numerical experiments on problems with strongly anisotropic diffusion coefficients illustrate the accuracy and computational efficiency of the scheme. An existence and uniqueness result for strong solutions of the fully nonlinear problem, and a semismoothness result for the nonlinear operator are also provided.

**Key words.** Hamilton–Jacobi–Bellman equations,  $hp$ -version discontinuous Galerkin finite element methods, Cordès condition, fully nonlinear equations, semismooth Newton methods

**AMS subject classifications.** 65N30, 65N12, 65N15, 35J15, 35J66, 35D35, 49M15, 47J25

**1. Introduction.** We study the numerical analysis of fully nonlinear second-order elliptic Hamilton–Jacobi–Bellman (HJB) equations of the form

$$\sup_{\alpha \in \Lambda} [L^\alpha u - f^\alpha] = 0 \quad \text{in } \Omega, \quad (1.1)$$

where  $\Omega$  is a convex domain in  $\mathbb{R}^n$ ,  $n \geq 2$ ,  $\Lambda$  is a compact metric space, and the  $L^\alpha$ ,  $\alpha \in \Lambda$ , are elliptic operators of the form

$$L^\alpha v = \sum_{i,j=1}^n a_{ij}^\alpha v_{x_i x_j} + \sum_{i=1}^n b_i^\alpha v_{x_i} - c^\alpha v. \quad (1.2)$$

HJB equations characterise the value functions of stochastic control problems, which arise from applications in engineering, physics, economics, and finance [10]. The solution of (1.1) leads to the best choices of controls from the set  $\Lambda$  for steering a stochastic process towards optimising the expected value of a functional. We are interested in consistent, stable, convergent and high-order methods for multidimensional uniformly elliptic HJB equations with anisotropic diffusions.

Discrete state Markov chain approximations to the underlying stochastic dynamics were amongst the earliest computational approaches to these problems [18]. Alongside the advent of the notion of a viscosity solution to a fully nonlinear second-order equation [6], it became apparent that these Markov chain approximations admit equivalent interpretations as *monotone* finite difference methods (FDM) [5, 10], i.e. that satisfy a discrete maximum principle. These methods feature a general convergence theory due to Barles and Souganidis [1], and are capable of approximating non-smooth viscosity solutions of certain degenerate problems.

Various authors have commented on the necessarily low-order convergence rates of monotone schemes [22], and on the restrictions that the choice of stencil imposes on the set of problems amenable to discretization by monotone FDM [7, 16]. For an

---

\*Mathematical Institute, University of Oxford, 24-29 St. Giles', Oxford OX1 3LB, UK, iain.smears@maths.ox.ac.uk

†Mathematical Institute, University of Oxford, 24-29 St. Giles', Oxford OX1 3LB, UK, suli@maths.ox.ac.uk

analysis of convergence rates, see [9] and the references therein. Motzkin and Wasow [21] found that for any choice of stencil, there exists a uniformly elliptic operator with no consistent and monotone discretisation; yet, for any set of non-degenerate diffusion coefficients with uniformly bounded ellipticity constants, there is a stencil providing a monotone and consistent discretisation. Kocan studied the minimum size of such a stencil as a function of the ellipticity constant in [16]. Conversely, Bonnans and Zidani [5] examined 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. An algorithm was developed in [4] to compute a monotone discretisation of two dimensional problems, with a consistency error depending on the stencil width.

Whilst the above considerations concern the notion of consistency of FDM, convergent monotone methods for fully nonlinear problems can also employ notions of consistency from finite element methods (FEM); see [15] for the first convergent monotone FEM for viscosity solutions of parabolic HJB equations. Böhmer proposed in [2] a non-monotone  $H^2$ -conforming FEM for fully nonlinear PDE with linearisations in divergence form; yet linearisations of the HJB operator are usually in non-divergence form with discontinuous coefficients, and cannot be recast into divergence form.

Discontinuous Galerkin finite element methods (DGFEM) allow the approximate solution to be discontinuous between elements of the mesh, with the continuity conditions being enforced only weakly through the discretised problem [8]. This facilitates  $hp$ -refinement, which varies both mesh size and polynomial degree, thereby allowing for exponential convergence rates, even for problems with non-smooth solutions [28]. For problems in non-divergence form, a challenge in the design of DGFEM is to obtain stable inter-element communication. Nevertheless, the authors found new techniques in [26] to obtain stable discretisations of certain linear non-divergence form equations with discontinuous coefficients, and these techniques are taken further in this work.

We consider here uniformly elliptic HJB equations that satisfy the Cordès condition: we provide a concise and accessible proof of existence and uniqueness of a strong solution of equation (1.1) associated to a homogeneous Dirichlet boundary condition. Then, we construct a stable, consistent and convergent  $hp$ -version DGFEM, for which we prove convergence rates in a discrete  $H^2$ -type norm that are optimal with respect to mesh size, and suboptimal in the polynomial degree by only half an order. As opposed to the monotone methods considered above, our method is consistent regardless of the choice of mesh, thereby permitting  $hp$ -refinement on very general shape-regular sequences of meshes. Our experiments below show the gains in computational efficiency, flexibility, and accuracy over existing monotone methods.

The Cordès condition, defined in §2 below, encompasses a large range of applications. For example, in two spatial dimensions, the condition amounts to simply requiring uniform ellipticity of the diffusion coefficient and coercivity of the lower order terms, see Examples 1 and 2 of §2. Let us now recount how the motivation for the Cordès condition stems from genuine PDE-theoretic considerations. There is a famous solution algorithm for (1.1), due to Bellman and Howard [3, 23], that may be understood as follows. Given an approximate solution  $u^k$ ,  $k \in \mathbb{N}$ , to (1.1), one finds for each  $x \in \Omega$  an  $\Lambda \ni \alpha_k(x) = \operatorname{argmax}_\alpha (L^\alpha u^k - f^\alpha)(x)$ . A new approximation  $u^{k+1}$  is sought as the solution of  $L^{\alpha_k} u^{k+1} = f^{\alpha_k}$ , where  $f^{\alpha_k} : x \mapsto f^{\alpha_k(x)}(x)$ , and where the coefficients of the linear operator  $L^{\alpha_k}$  are similarly defined; formally, a solution of (1.1) is a fixed point of this iteration. It has long been known that this method is in fact a Newton method for a non-differentiable operator [3, 23], and we

contribute to its analysis by showing the semismoothness in function spaces [27] of the HJB operator. The question of the well-posedness of the linear PDE to be solved at each iteration is instructive: these are non-divergence form elliptic equations *with discontinuous coefficients*, and it is known that well-posedness in the strong sense is not guaranteed by uniform ellipticity alone [11, 19, 25], although it is recovered under the Cordès condition [19]. Importantly, we show here that well-posedness of strong solutions extends to HJB equations, under the same condition. Inspired by the analysis of the PDE, the stability of our method is obtained by relating the residual of the equation to terms measuring the lack of  $H^2$ -conformity of the numerical solution.

The structure of this article is as follows. After defining the problem in §2, we prove its well-posedness in §3. The  $hp$ -version DGFEM framework is prepared in §4 and is followed by the definition and consistency analysis of the method in §5. We establish the stability of the scheme in §6 and we determine its convergence rates in §7. Section 8 analyses a superlinearly convergent semismooth Newton method used to solve the discrete problem, and §9 presents the results of numerical experiments that demonstrate the high accuracy and computational efficiency of the method.

**2. Statement of the problem.** Let  $\Omega$  be a bounded convex polytopal open set in  $\mathbb{R}^n$ ,  $n \geq 2$ , and let  $\Lambda$  be a compact metric space. It will always be assumed that  $\Omega$  and  $\Lambda$  are nonempty. Convexity of  $\Omega$  implies that the boundary  $\partial\Omega$  is Lipschitz; see [12]. Let the real-valued functions  $a_{ij} = a_{ji}$ ,  $b_i$ ,  $c$  and  $f$  belong to  $C(\overline{\Omega} \times \Lambda)$  for all  $i, j = 1, \dots, n$ . For each  $\alpha \in \Lambda$ , we consider the function  $a_{ij}^\alpha: x \mapsto a_{ij}(x, \alpha)$ ,  $x \in \overline{\Omega}$ . The functions  $b_i^\alpha$ ,  $c^\alpha$  and  $f^\alpha$  are defined in a similar way. Define the matrix-valued functions  $a^\alpha := (a_{ij}^\alpha)$  and define the vector-valued functions  $b^\alpha := (b_1^\alpha, \dots, b_n^\alpha)$ , where  $\alpha \in \Lambda$ . The bounded linear operators  $L^\alpha: H^2(\Omega) \rightarrow L^2(\Omega)$  are defined by

$$L^\alpha v := \sum_{i,j=1}^n a_{ij}^\alpha v_{x_i x_j} + \sum_{i=1}^n b_i^\alpha v_{x_i} - c^\alpha v, \quad v \in H^2(\Omega), \alpha \in \Lambda. \quad (2.1)$$

Compactness of  $\Lambda$  and continuity of the coefficients  $a$ ,  $b$ ,  $c$  and  $f$  imply that the fully nonlinear operator  $F$ , defined by

$$F: v \mapsto F[v] := \sup_{\alpha \in \Lambda} [L^\alpha v - f^\alpha], \quad (2.2)$$

is well-defined as a mapping from  $H^2(\Omega)$  to  $L^2(\Omega)$ . The problem considered is to find  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  that is a strong solution of the HJB equation subject to a homogeneous Dirichlet boundary condition

$$\begin{aligned} F[u] &= 0 && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned} \quad (2.3)$$

Well-posedness of problem (2.3) is established in §3 under the following hypotheses. It is assumed that there exist positive constants  $\nu \leq \bar{\nu}$  such that

$$\nu |\xi|^2 \leq \sum_{i,j=1}^n a_{ij}^\alpha(x) \xi_i \xi_j \leq \bar{\nu} |\xi|^2 \quad \forall \xi \in \mathbb{R}^n, \forall x \in \Omega, \forall \alpha \in \Lambda. \quad (2.4)$$

The function  $c^\alpha$  is supposed to be non-negative on  $\overline{\Omega}$ , for each  $\alpha \in \Lambda$ . We assume *the Cordès condition*: there exist  $\lambda > 0$  and  $\varepsilon \in (0, 1)$  such that, for each  $\alpha \in \Lambda$ ,

$$\frac{|a^\alpha|^2 + |b^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2}{(\text{Tr } a^\alpha + c^\alpha/\lambda)^2} \leq \frac{1}{n + \varepsilon} \quad \text{in } \overline{\Omega}, \quad (2.5)$$

where  $|\cdot|$  represents the Euclidian norm for vectors and the Frobenius norm for matrices. In the special case  $b^\alpha \equiv 0$  and  $c^\alpha \equiv 0$  for each  $\alpha \in \Lambda$ , we set  $\lambda = 0$  and the Cordès condition (2.5) is replaced by: there exists  $\varepsilon \in (0, 1)$  such that, for each  $\alpha \in \Lambda$ ,

$$\frac{|a^\alpha|^2}{(\text{Tr } a^\alpha)^2} \leq \frac{1}{n-1+\varepsilon} \quad \text{in } \bar{\Omega}. \quad (2.6)$$

Conditions (2.5) and (2.6) are related through the observation that the term  $c^\alpha/\lambda$  may be viewed as the  $(n+1, n+1)$  entry of an  $(n+1) \times (n+1)$  matrix with principal  $n \times n$  sub-matrix  $a^\alpha$ , which explains the difference in the right hand sides of the inequalities in (2.5) and (2.6). The parameter  $\lambda$  serves to make the Cordès condition invariant under rescaling the coordinates. It will be seen below that it is often easy to choose an appropriate value for  $\lambda$ .

EXAMPLE 1. *We show how the Cordès condition (2.5) arises in practice in an example from stochastic control problems [10]. We consider a problem where the controls permit the choice of orientation and angle between two Wiener diffusions. Let  $\Omega$  be a domain in  $\mathbb{R}^2$  and let  $\Lambda = [0, \pi/3] \times \text{SO}(2)$ , where  $\text{SO}(2)$  is the set of  $2 \times 2$  rotation matrices. The diffusions act along the directions  $\sigma_1^\alpha$  and  $\sigma_2^\alpha$ , where*

$$\sigma^\alpha := (\sigma_1^\alpha \ \sigma_2^\alpha) := R^\text{T} \begin{pmatrix} 1 & \sin \theta \\ 0 & \cos \theta \end{pmatrix}, \quad \alpha = (\theta, R) \in \Lambda. \quad (2.7)$$

*In stochastic control problems, we have  $a^\alpha := \sigma^\alpha(\sigma^\alpha)^\text{T}/2$  and usually  $c^\alpha \equiv c_0 > 0$  is a fixed constant [10]. Then,  $\text{Tr } a^\alpha = 1$  and  $|a^\alpha|^2 = (1 + \sin^2 \theta)/2 \leq 7/8$ ; so condition (2.6) holds with  $\varepsilon = 1/7$ . Momentarily assuming that  $b^\alpha \equiv 0$ , by choosing the value  $\lambda = \frac{8}{7}c_0$  that minimises the left-hand side in (2.5), we find that condition (2.5) also holds with  $\varepsilon = 1/7$ . For non-zero  $b^\alpha$ , the Cordès condition holds for  $\varepsilon < 1/7$  whenever  $|b^\alpha|^2/c_0$  is sufficiently small; this amounts to a standard coercivity assumption.*

Example 1 is considered further in the numerical experiments of §9.1. Observe that for any choice of Cartesian coordinates on  $\mathbb{R}^2$ , for  $\theta = \pi/3$  there is an  $R \in \text{SO}(2)$  such that  $a^\alpha$  is not diagonally dominant. Therefore, the classical monotone Kushner–Dupuis FDM is not applicable here [5].

EXAMPLE 2. *For problems in two dimensions, i.e.  $n = 2$ , the uniform ellipticity condition (2.4) is sufficient for the Cordès condition (2.6). Indeed, for each  $\alpha \in \Lambda$ , we have  $\nu^2 \leq \det a^\alpha$ , and  $a_{11}^\alpha + a_{22}^\alpha \leq 2\bar{\nu}$ , so, for  $\varepsilon = \nu^2/(2\bar{\nu}^2 - \nu^2)$ , we get*

$$\frac{(a_{11}^\alpha)^2 + 2(a_{12}^\alpha)^2 + (a_{22}^\alpha)^2}{(a_{11}^\alpha + a_{22}^\alpha)^2} \leq 1 - \frac{2\nu^2}{(a_{11}^\alpha + a_{22}^\alpha)^2} \leq 1 - \frac{\nu^2}{2\bar{\nu}^2} = \frac{1}{1+\varepsilon}. \quad (2.8)$$

The above examples demonstrate that the results of this paper 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 §9. Define the strictly positive function  $\gamma: \bar{\Omega} \times \Lambda \rightarrow \mathbb{R}_{>0}$  by

$$\gamma(x, \alpha) := \frac{\text{Tr } a^\alpha(x) + c^\alpha(x)/\lambda}{|a^\alpha(x)|^2 + |b^\alpha(x)|^2/2\lambda + (c^\alpha(x)/\lambda)^2}. \quad (2.9)$$

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

$$\gamma(x, \alpha) := \frac{\text{Tr } a^\alpha(x)}{|a^\alpha(x)|^2}. \quad (2.10)$$

As above, for each  $\alpha \in \Lambda$ , we define  $\gamma^\alpha : x \mapsto \gamma(x, \alpha)$ ,  $x \in \overline{\Omega}$ . It follows from the continuity assumptions on the coefficients and from the uniform ellipticity condition (2.4) that  $\gamma \in C(\overline{\Omega} \times \Lambda)$ . Furthermore, non-negativity of  $c$ , continuity of the coefficients and (2.4) imply that there is a positive constant  $\gamma_0 > 0$  such that  $\gamma \geq \gamma_0$  on  $\overline{\Omega} \times \Lambda$ . Define the operator  $F_\gamma : H^2(\Omega) \rightarrow L^2(\Omega)$  by

$$F_\gamma[v] := \sup_{\alpha \in \Lambda} [\gamma^\alpha (L^\alpha v - f^\alpha)]. \quad (2.11)$$

It will be seen below that the HJB equation (2.3) is in fact equivalent to the problem  $F_\gamma[u] = 0$  in  $\Omega$ ,  $u = 0$  on  $\partial\Omega$ . For  $\lambda$  as above, let the operator  $L_\lambda$  be defined by

$$L_\lambda v := \Delta v - \lambda v, \quad v \in H^2(\Omega). \quad (2.12)$$

The following inequality generalises results in [19, 26] that were used to analyse linear PDE satisfying the Cordès condition. It is key to our analysis of HJB equations.

LEMMA 1. *Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$  and suppose that (2.4) holds, and suppose that either (2.5) holds with  $\lambda > 0$ , or that (2.6) holds with  $b^\alpha \equiv 0$ ,  $c^\alpha \equiv 0$  for all  $\alpha$ , and  $\lambda = 0$ . Then, for any open set  $U \subset \Omega$  and  $u, v \in H^2(U)$ ,  $w := u - v$ , the following inequality holds a.e. in  $U$ :*

$$|F_\gamma[u] - F_\gamma[v] - L_\lambda(u - v)| \leq \sqrt{1 - \varepsilon} \sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2}. \quad (2.13)$$

*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 (2.5) holds with  $\lambda > 0$ . First, set  $w := u - v$ . Note that we have the identity  $F_\gamma[u] - L_\lambda u = \sup_{\alpha \in \Lambda} [\gamma^\alpha L^\alpha u - L_\lambda u - \gamma^\alpha f^\alpha]$ . Also, for bounded sets of real numbers,  $\{x^\alpha\}_\alpha$  and  $\{y^\alpha\}_\alpha$ , we have  $|\sup_\alpha x^\alpha - \sup_\alpha y^\alpha| \leq \sup_\alpha |x^\alpha - y^\alpha|$ . Therefore,

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

where  $I_n$  is the  $n \times n$  identity matrix. The Cauchy–Schwarz inequality with a parameter gives

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

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

$$C^\alpha := |\gamma^\alpha a^\alpha - I_n|^2 + |\gamma^\alpha|^2 \frac{|b^\alpha|^2}{2\lambda} + \frac{|\lambda - c^\alpha \gamma^\alpha|^2}{\lambda^2}. \quad (2.14)$$

Expanding the square terms in (2.14) gives

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

The definition of  $\gamma$  in (2.9) and the Cordès condition (2.5) imply that  $C^\alpha \leq 1 - \varepsilon$  on  $U$  for every  $\alpha \in \Lambda$ , thus completing the proof of (2.13).  $\square$

In the following analysis, we shall write  $a \lesssim b$  for  $a, b \in \mathbb{R}$  to signify that there exists a constant  $C$  such that  $a \leq Cb$ , where  $C$  is independent of the mesh size and polynomial degrees used to define the finite element spaces below, but otherwise possibly dependent on other fixed quantities, such as the constants in (2.4) and (2.5) or the shape-regularity parameters of the mesh, for example.

**3. Analysis of the PDE.** For  $\lambda \geq 0$  as above, define the semi-norm  $|\cdot|_{H^2(\Omega),\lambda}$  on  $H^2(\Omega)$  by

$$|u|_{H^2(\Omega),\lambda}^2 := |u|_{H^2(\Omega)}^2 + 2\lambda|u|_{H^1(\Omega)}^2 + \lambda^2\|u\|_{L^2(\Omega)}^2. \quad (3.1)$$

If  $\lambda > 0$ , then this defines a norm on  $H^2(\Omega)$ . The following result follows from the Miranda–Talenti estimate; see [12, 19, 26]. Recall that  $L_\lambda u = \Delta u - \lambda u$ .

**THEOREM 2.** *Let  $\Omega$  be a bounded convex open subset of  $\mathbb{R}^n$ . Then, for any  $\lambda \geq 0$  and any  $u \in H^2(\Omega) \cap H_0^1(\Omega)$ , the following inequalities hold:*

$$|u|_{H^2(\Omega),\lambda} \leq \|L_\lambda u\|_{L^2(\Omega)}, \quad (3.2a)$$

$$\|u\|_{H^2(\Omega)} \leq C\|L_\lambda u\|_{L^2(\Omega)}, \quad (3.2b)$$

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

*Proof.* In [26, Theorem 2], it is shown that on bounded convex domains, we have the Miranda–Talenti estimate  $|u|_{H^2(\Omega)} \leq \|\Delta u\|_{L^2(\Omega)}$  for any  $u \in H^2(\Omega) \cap H_0^1(\Omega)$ . The identity  $\int_\Omega u \Delta u \, dx = - \int_\Omega |\nabla u|^2 \, dx$ , based on integration by parts, gives

$$\|L_\lambda u\|_{L^2(\Omega)}^2 = \int_\Omega (\Delta u - \lambda u)^2 \, dx = \|\Delta u\|_{L^2(\Omega)}^2 + 2\lambda|u|_{H^1(\Omega)}^2 + \lambda^2\|u\|_{L^2(\Omega)}^2. \quad (3.3)$$

The Miranda–Talenti estimate and (3.3) give (3.2a). The bound (3.2b) follows from (3.3) and the estimate  $\|u\|_{H^2(\Omega)} \leq C(n, \text{diam } \Omega)\|\Delta u\|_{L^2(\Omega)}$  shown in [26, Thm. 2].  $\square$

**THEOREM 3.** *Let  $\Omega$  be a bounded convex open subset of  $\mathbb{R}^n$ , and let  $\Lambda$  be a compact metric space. Let the data  $a, b, c, f$  be continuous on  $\overline{\Omega} \times \Lambda$  and satisfy (2.4) and either (2.5) with  $\lambda > 0$  or (2.6) 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 (2.3). Moreover,  $u$  is also the unique solution of  $F_\gamma[u] = 0$  in  $\Omega$ ,  $u = 0$  on  $\partial\Omega$ .*

*Proof.* First, set  $H := H^2(\Omega) \cap H_0^1(\Omega)$ ; 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 (2.3). Let the operator  $\mathcal{A}: H \rightarrow H^*$  be defined by

$$\langle \mathcal{A}(u), v \rangle = \int_\Omega F_\gamma[u] L_\lambda v \, dx, \quad u, v \in H. \quad (3.4)$$

We claim that  $\mathcal{A}$  is Lipschitz continuous and strongly monotone. Indeed, let  $u, v \in H$  and set  $w := u - v$ . Then, by adding and subtracting  $L_\lambda w$ , we get

$$\langle \mathcal{A}(u) - \mathcal{A}(v), u - v \rangle = \|L_\lambda w\|_{L^2(\Omega)}^2 + \int_\Omega (F_\gamma[u] - F_\gamma[v] - L_\lambda w) L_\lambda w \, dx. \quad (3.5)$$

Lemma 1 and the Cauchy–Schwarz inequality show that

$$\langle \mathcal{A}(u) - \mathcal{A}(v), u - v \rangle \geq \|L_\lambda w\|_{L^2(\Omega)}^2 - \sqrt{1 - \varepsilon} |w|_{H^2(\Omega),\lambda} \|L_\lambda w\|_{L^2(\Omega)}. \quad (3.6)$$

We then use (3.2a) to obtain  $\langle \mathcal{A}(u) - \mathcal{A}(v), u - v \rangle \geq (1 - \sqrt{1 - \varepsilon}) \|L_\lambda w\|_{L^2(\Omega)}^2$ , so  $\|u - v\|_{H^2(\Omega)}^2 \lesssim \langle \mathcal{A}(u) - \mathcal{A}(v), u - v \rangle$  as a result of (3.2b), thus showing that  $\mathcal{A}$  is strongly monotone. Compactness of  $\Lambda$  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

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

where the constant  $C$  depends only on  $\lambda$  and on the supremum norms of  $a_{ij}$ ,  $b_i$ ,  $c$  and  $\gamma$  over  $\bar{\Omega} \times \Lambda$ , for  $i, j = 1, \dots, n$ . Lipschitz continuity and strong monotonicity imply that  $\mathcal{A}$  is bounded, continuous, coercive and strongly monotone, so the Browder–Minty theorem [24] shows that there exists a unique  $u \in H$  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  $\int_\Omega F_\gamma[u] g \, dx = 0$  for all  $g \in L^2(\Omega)$ , thus showing that  $F_\gamma[u] = 0$  a.e. in  $\Omega$ . We claim that  $F_\gamma[u] = 0$  if and only if  $u$  solves (2.3). Since  $\gamma^\alpha$  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  $\Lambda$  and continuity of  $a$ ,  $b$ ,  $c$ ,  $f$  and  $\gamma$  imply that at a.e. point of  $\Omega$ , the suprema in the definitions of  $F[u]$  and  $F_\gamma[u]$  are attained by an element of  $\Lambda$ , 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  $\Omega$  is equivalent to existence and uniqueness of a solution of (2.3).  $\square$

**4. Finite element spaces.** Let  $\{\mathcal{T}_h\}_h$  be a sequence of shape-regular meshes on  $\Omega$ , consisting of simplices or parallelepipeds. For each element  $K \in \mathcal{T}_h$ , let  $h_K := \text{diam } K$ . 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  $\mathcal{T}_h$ , i.e.  $F \in \mathcal{F}_h^i$  if and only if  $F = \partial K \cap \partial K'$ , for some elements  $K$  and  $K' \in \mathcal{T}_h$ , and  $F$  is the closure of a non-empty smooth connected hypersurface that is open relative to  $\partial K \cap \partial K'$ . Since each element has piecewise flat boundary, it follows that any interior face is flat. Let  $\mathcal{F}_h^b$  denote the set of boundary faces of the mesh, that is,  $F \in \mathcal{F}_h^b$  if and only if  $F$  is the closure of a non-empty smooth connected hypersurface that is a relatively open subset of  $\partial\Omega \cap \partial K$ , for some  $K \in \mathcal{T}_h$ , and  $F$  is maximal in the sense that there is no smooth connected hypersurface containing  $F$  that is also a relatively open subset of  $\partial\Omega \cap \partial K$ . Boundary faces are therefore also flat. The set of all faces is  $\mathcal{F}_h^{i,b} := \mathcal{F}_h^i \cup \mathcal{F}_h^b$ .

*Mesh conditions.* We shall make the following assumptions on the meshes. 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  $c_{\mathcal{F}} > 0$ , independent of  $h$ , such that

$$\max_{K \in \mathcal{T}_h} \text{card} \left\{ F \in \mathcal{F}_h^{i,b} : F \subset \partial K \right\} \leq c_{\mathcal{F}} \quad \forall K \in \mathcal{T}_h, \forall h > 0. \quad (4.1)$$

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

$$\max(h_K, h_{K'}) \leq c_{\mathcal{T}} \min(h_K, h_{K'}), \quad (4.2)$$

for any  $K$  and  $K'$  in  $\mathcal{T}_h$  that share a face. For each  $h$ , let  $\mathbf{p} = (p_K : K \in \mathcal{T}_h)$  be a vector of positive integers. In order to let  $p_K$  appear in the denominator of various expressions, we shall assume that  $p_K \geq 1$  for all  $K \in \mathcal{T}_h$ . We make the assumption that  $\mathbf{p}$  has *local bounded variation* [14]: there is a  $c_{\mathcal{P}} \geq 1$ , independent of  $h$ , such that

$$\max(p_K, p_{K'}) \leq c_{\mathcal{P}} \min(p_K, p_{K'}), \quad (4.3)$$

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 polynomials with either total or partial degree less than or equal to  $p_K$ . The discontinuous Galerkin finite element space  $V_{h,\mathbf{p}}$  is defined by

$$V_{h,\mathbf{p}} := \left\{ v \in L^2(\Omega), v|_K \in \mathcal{P}_{p_K}(K), \forall K \in \mathcal{T}_h \right\}. \quad (4.4)$$

Let  $\mathbf{s} = (s_K : K \in \mathcal{T}_h)$  denote a vector of non-negative real numbers, and let  $r \in [1, \infty]$ . The broken Sobolev space  $W^{\mathbf{s},r}(\Omega; \mathcal{T}_h)$  is defined by

$$W^{\mathbf{s},r}(\Omega; \mathcal{T}_h) := \{v \in L^r(\Omega), v|_K \in W^{s_K,r}(K), \forall K \in \mathcal{T}_h\}. \quad (4.5)$$

For shorthand, define  $H^{\mathbf{s}}(\Omega; \mathcal{T}_h) := W^{\mathbf{s},2}(\Omega; \mathcal{T}_h)$ , and set  $W^{s,r}(\Omega; \mathcal{T}_h) := W^{\mathbf{s},r}(\Omega; \mathcal{T}_h)$ , where  $s_K = s$ ,  $s \geq 0$ , for all  $K \in \mathcal{T}_h$ . For  $v \in W^{1,r}(\Omega; \mathcal{T}_h)$ , let  $\nabla_h v \in L^r(\Omega; \mathbb{R}^n)$  denote the broken gradient of  $v$ , i.e.  $(\nabla_h v)|_K = \nabla(v|_K)$  for all  $K \in \mathcal{T}_h$ . Higher broken derivatives are defined in a similar way. Define a norm on  $W^{\mathbf{s},r}(\Omega; \mathcal{T}_h)$  by

$$\|v\|_{W^{\mathbf{s},r}(\Omega; \mathcal{T}_h)}^r := \sum_{K \in \mathcal{T}_h} \|v\|_{W^{s_K,r}(K)}^r, \quad (4.6)$$

with the usual modification when  $r = \infty$ .

*Jump, average, and tangential operators.* For each face  $F$ , let  $n_F \in \mathbb{R}^n$  denote a fixed choice of a unit normal vector to  $F$ . Since each face  $F$  is flat, the normal  $n_F$  is constant. For an element  $K \in \mathcal{T}_h$  and a face  $F \subset \partial K$ , let  $\tau_F : H^s(K) \rightarrow H^{s-1/2}(F)$ ,  $s > 1/2$ , denote the trace operator from  $K$  to  $F$ . The trace operator  $\tau_F$  is extended componentwise to vector-valued functions. Define the jump operator  $[[\cdot]]$  over  $F$  by

$$[[\phi]] := \begin{cases} \tau_F(\phi|_{K_{\text{ext}}}) - \tau_F(\phi|_{K_{\text{int}}}) & \text{if } F \in \mathcal{F}_h^i, \\ \tau_F(\phi|_{K_{\text{ext}}}) & \text{if } F \in \mathcal{F}_h^b, \end{cases} \quad (4.7)$$

and define  $\{\cdot\}$ , the average operator over  $F$ , by

$$\{\phi\} := \begin{cases} \frac{1}{2}(\tau_F(\phi|_{K_{\text{ext}}}) + \tau_F(\phi|_{K_{\text{int}}})) & \text{if } F \in \mathcal{F}_h^i, \\ \tau_F(\phi|_{K_{\text{ext}}}) & \text{if } F \in \mathcal{F}_h^b, \end{cases} \quad (4.8)$$

where  $\phi$  is a sufficiently regular scalar or vector-valued function, and  $K_{\text{ext}}$  and  $K_{\text{int}}$  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 for  $K_{\text{ext}}$  and inward pointing for  $K_{\text{int}}$ . Using this notation, the jump and average of scalar-valued functions, resp. vector-valued, are scalar-valued, resp. vector-valued. For two matrices  $A, B \in \mathbb{R}^{n \times n}$ , we set  $A : B = \sum_{i,j=1}^n A_{ij} B_{ij}$ . For an element  $K$ , we define the inner product  $\langle \cdot, \cdot \rangle_K$  by

$$\langle u, v \rangle_K := \begin{cases} \int_K u v \, dx & \text{if } u, v \in L^2(K), \\ \int_K u \cdot v \, dx & \text{if } u, v \in L^2(K; \mathbb{R}^n), \\ \int_K u : v \, dx & \text{if } u, v \in L^2(K; \mathbb{R}^{n \times n}). \end{cases} \quad (4.9)$$

The abuse of notation will be resolved by the arguments of the inner product. The inner products  $\langle \cdot, \cdot \rangle_{\partial K}$  and  $\langle \cdot, \cdot \rangle_F$ ,  $F \in \mathcal{F}_h^{i,b}$ , are defined in a similar way. For a face  $F$ , let  $\nabla_T$  and  $\text{div}_T$  denote respectively the tangential gradient and tangential divergence operators on  $F$ ; see [12, 26].

**5. Numerical scheme.** The definition of the numerical scheme requires the following bilinear and nonlinear forms. First, for  $\lambda \geq 0$  as above, the symmetric



bilinear form  $B_{h,*}: V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \rightarrow \mathbb{R}$  is defined by

$$\begin{aligned}
B_{h,*}(u_h, v_h) &:= \sum_{K \in \mathcal{T}_h} [\langle D^2 u_h, D^2 v_h \rangle_K + 2\lambda \langle \nabla u_h, \nabla v_h \rangle_K + \lambda^2 \langle u_h, v_h \rangle_K] \\
&+ \sum_{F \in \mathcal{F}_h^i} [\langle \operatorname{div}_T \nabla_T \{u_h\}, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F + \langle \operatorname{div}_T \nabla_T \{v_h\}, \llbracket \nabla u_h \cdot n_F \rrbracket \rangle_F] \\
&- \sum_{F \in \mathcal{F}_h^{i,b}} [\langle \nabla_T \{ \nabla u_h \cdot n_F \}, \llbracket \nabla_T v_h \rrbracket \rangle_F + \langle \nabla_T \{ \nabla v_h \cdot n_F \}, \llbracket \nabla_T u_h \rrbracket \rangle_F] \\
&- \lambda \sum_{F \in \mathcal{F}_h^{i,b}} [\langle \{ \nabla u_h \cdot n_F \}, \llbracket v_h \rrbracket \rangle_F + \langle \{ \nabla v_h \cdot n_F \}, \llbracket u_h \rrbracket \rangle_F] \\
&- \lambda \sum_{F \in \mathcal{F}_h^i} [\langle \{ u_h \}, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F + \langle \{ v_h \}, \llbracket \nabla u_h \cdot n_F \rrbracket \rangle_F],
\end{aligned}$$

where  $u_h$  and  $v_h$  will denote functions in  $V_{h,\mathbf{p}}$  throughout this work. Then, for positive face-dependent quantities  $\mu_F$  and  $\eta_F$  to be specified later, the jump stabilisation bilinear form  $J_h: V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \rightarrow \mathbb{R}$  is defined by

$$\begin{aligned}
J_h(u_h, v_h) &:= \sum_{F \in \mathcal{F}_h^{i,b}} [\mu_F \langle \llbracket \nabla_T u_h \rrbracket, \llbracket \nabla_T v_h \rrbracket \rangle_F + \eta_F \langle \llbracket u_h \rrbracket, \llbracket v_h \rrbracket \rangle_F] \\
&+ \sum_{F \in \mathcal{F}_h^i} \mu_F \langle \llbracket \nabla u_h \cdot n_F \rrbracket, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F. \tag{5.1}
\end{aligned}$$

For each  $\theta \in [0, 1]$ , define the bilinear form  $B_{h,\theta}: V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \rightarrow \mathbb{R}$  by

$$B_{h,\theta}(u_h, v_h) := \theta B_{h,*}(u_h, v_h) + (1 - \theta) \sum_{K \in \mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h \rangle_K + J_h(u_h, v_h). \tag{5.2}$$

The nonlinear form  $A_h: V_{h,\mathbf{p}} \times V_{h,\mathbf{p}} \rightarrow \mathbb{R}$  is defined by

$$A_h(u_h; v_h) := \sum_{K \in \mathcal{T}_h} \langle F_\gamma[u_h], L_\lambda v_h \rangle_K + B_{h,\frac{1}{2}}(u_h, v_h) - \sum_{K \in \mathcal{T}_h} \langle L_\lambda u_h, L_\lambda v_h \rangle_K. \tag{5.3}$$

The form  $A_h$  is linear in its second argument but nonlinear in its first argument. The numerical scheme for approximating the solution of (2.3) is to find  $u_h \in V_{h,\mathbf{p}}$  such that

$$A_h(u_h; v_h) = 0 \quad \forall v_h \in V_{h,\mathbf{p}}. \tag{5.4}$$

The choice of nonlinear form in (5.3) is made to mirror the addition–subtraction step of (3.5) in the proof of Theorem 3. It will be seen below that the last two terms of (5.3) vanish when the first argument of the form is smooth, and it is in this sense that this method relates the residual of the numerical solution to its lack of smoothness.

**5.1. Consistency.** The next result shows that the bilinear form  $B_{h,\theta}$  is obtained from a discrete analogue of the identities that underpin Theorem 2.

**LEMMA 4.** *Let  $\Omega$  be a bounded Lipschitz polytopal domain and let  $\mathcal{T}_h$  be a simplicial or parallelepipedal mesh on  $\Omega$ . Let the function  $w$  belong either to  $C^1(\overline{\Omega}) \cap H^s(\Omega; \mathcal{T}_h)$  or to  $H^s(\Omega)$ ,  $s > 5/2$ , and suppose that  $w|_F = 0$  in the trace sense for every boundary face  $F \in \mathcal{F}_h^b$ . Then, for every  $v_h \in V_{h,\mathbf{p}}$ , we have the identities*

$$B_{h,*}(w, v_h) = \sum_{K \in \mathcal{T}_h} \langle L_\lambda w, L_\lambda v_h \rangle_K \quad \text{and} \quad J_h(w, v_h) = 0. \tag{5.5}$$

*Proof.* The second part of (5.5) is obvious. We also note that all terms in  $B_{h,*}(w, v_h)$  that involve jumps of  $w$  or of its first derivatives vanish. For the case  $\lambda = 0$ , the stated result reduces to [26, Lemma 5], which treats the consistency of the second order terms, namely  $B_{h,*}(w, v_h) = \sum_K \langle \Delta w, \Delta v_h \rangle_K$  for all  $v_h \in V_{h,\mathbf{p}}$ . So, for  $\lambda > 0$ , the identities of (5.5) are deduced from the previous result and from the identities

$$-\lambda \sum_{K \in \mathcal{T}_h} \langle \Delta w, v_h \rangle_K = \lambda \sum_{K \in \mathcal{T}_h} \langle \nabla w, \nabla v_h \rangle_K - \lambda \sum_{F \in \mathcal{F}_h^{i,b}} \langle \{\nabla w \cdot n_F\}, \llbracket v_h \rrbracket \rangle_F, \quad (5.6)$$

$$-\lambda \sum_{K \in \mathcal{T}_h} \langle w, \Delta v_h \rangle_K = \lambda \sum_{K \in \mathcal{T}_h} \langle \nabla w, \nabla v_h \rangle_K - \lambda \sum_{F \in \mathcal{F}_h^i} \langle \{w\}, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F, \quad (5.7)$$

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 (5.7).  $\square$

If the function  $w$  satisfies the hypotheses of Lemma 4, then (5.5) implies that

$$B_{h,\theta}(w, v_h) = \sum_{K \in \mathcal{T}_h} \langle L_\lambda w, L_\lambda v_h \rangle_K \quad \forall v_h \in V_{h,\mathbf{p}}, \quad \forall \theta \in [0, 1]. \quad (5.8)$$

The following consistency result for the scheme (5.4) follows immediately from Theorem 3, (5.8) and from the definition of  $A_h$  in (5.3).

**COROLLARY 5.** *Let  $\Omega$  be a bounded convex polytopal domain, let  $\mathcal{T}_h$  be a simplicial or parallelepipedal mesh on  $\Omega$  and let  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  be the unique solution of problem (2.3). If  $u \in H^s(\Omega)$  or if  $u \in C^1(\bar{\Omega}) \cap H^s(\Omega; \mathcal{T}_h)$ ,  $s > 5/2$ , then  $u$  satisfies  $A_h(u; v_h) = 0$  for every  $v_h \in V_{h,\mathbf{p}}$ .*

**6. Stability.** 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

$$|v|_{H^2(K),\lambda}^2 := \|D^2 v\|_{L^2(K)}^2 + 2\lambda \|\nabla v\|_{L^2(K)}^2 + \lambda^2 \|v\|_{L^2(K)}^2, \quad (6.1)$$

$$|v|_{H^2(\Omega; \mathcal{T}_h),\lambda}^2 := \sum_{K \in \mathcal{T}_h} |v|_{H^2(K),\lambda}^2. \quad (6.2)$$

For a positive constant  $c_*$ , independent of  $h$  and to be specified later, and  $\theta \in [0, 1]$ , define the functionals  $\|\cdot\|_{\text{DG}(\theta)}: V_{h,\mathbf{p}} \rightarrow \mathbb{R}_{\geq 0}$  by

$$\|v_h\|_{\text{DG}(\theta)}^2 := \sum_{K \in \mathcal{T}_h} \left[ \theta |v_h|_{H^2(K),\lambda}^2 + (1-\theta) \|L_\lambda v_h\|_{L^2(K)}^2 \right] + c_* J_h(v_h, v_h). \quad (6.3)$$

For each  $\theta \in [0, 1]$ ,  $\|\cdot\|_{\text{DG}(\theta)}$  is a norm on  $V_{h,\mathbf{p}}$ . Indeed, homogeneity and the triangle inequality are clear. If  $\|v_h\|_{\text{DG}(\theta)} = 0$ , then  $v_h \in H^2(\Omega) \cap H_0^1(\Omega)$  since  $\llbracket \nabla v_h \rrbracket = 0$  for all  $F \in \mathcal{F}_h^i$ , and  $\llbracket v_h \rrbracket = 0$  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 (3.2b). For each face  $F \in \mathcal{F}_h^{i,b}$ , define

$$\tilde{h}_F := \begin{cases} \min(h_K, h_{K'}), & \text{if } F \in \mathcal{F}_h^i, \\ h_K, & \text{if } F \in \mathcal{F}_h^b, \end{cases} \quad \tilde{p}_F := \begin{cases} \max(p_K, p_{K'}), & \text{if } F \in \mathcal{F}_h^i, \\ p_K, & \text{if } F \in \mathcal{F}_h^b, \end{cases} \quad (6.4)$$

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 (4.2) and (4.3), show that if  $F$  is a face of  $K$ , then

$$h_K \leq c_{\mathcal{T}} \tilde{h}_F \quad \text{and} \quad \tilde{p}_F \leq c_{\mathcal{P}} p_K. \quad (6.5)$$

LEMMA 6. *Let  $\Omega$  be a bounded convex polytopal domain and let  $\{\mathcal{T}_h\}_h$  be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying (4.1). Then, for each constant  $\kappa > 1$ , there exist positive constants  $c_{\text{stab}}$  and  $c_*$ , independent of  $h$ ,  $\mathbf{p}$  and  $\theta$ , such that*

$$\|v_h\|_{\text{DG}(\theta)}^2 \leq \kappa B_{h,\theta}(v_h, v_h) \quad \forall v_h \in V_{h,\mathbf{p}}, \forall \theta \in [0, 1], \quad (6.6)$$

whenever, for any fixed constant  $\sigma \geq 1$ ,

$$\mu_F = \sigma c_{\text{stab}} \frac{\tilde{p}_F^2}{\tilde{h}_F} \quad \text{and} \quad \eta_F > \sigma \lambda c_{\text{stab}} \frac{\tilde{p}_F^2}{\tilde{h}_F}. \quad (6.7)$$

The strict inequality in the second part of (6.7) serves to cover the case  $\lambda = 0$ .

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

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

where

$$\begin{aligned} I_1 &:= 2 \sum_{F \in \mathcal{F}_h^i} \langle \text{div}_T \nabla_T \{v_h\}, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F, & I_3 &:= -2\lambda \sum_{F \in \mathcal{F}_h^i} \langle \{v_h\}, \llbracket \nabla v_h \cdot n_F \rrbracket \rangle_F, \\ I_2 &:= -2 \sum_{F \in \mathcal{F}_h^{i,b}} \langle \nabla_T \{ \nabla v_h \cdot n_F \}, \llbracket \nabla_T v_h \rrbracket \rangle_F, & I_4 &:= -2\lambda \sum_{F \in \mathcal{F}_h^{i,b}} \langle \{ \nabla v_h \cdot n_F \}, \llbracket v_h \rrbracket \rangle_F. \end{aligned}$$

In [26, Lemma 7], it is shown that there is a constant  $C(n)$  depending only on  $n$ , such that for any  $\delta > 0$ ,

$$|I_1| \leq \delta C(n) C_{\text{Tr}} c_{\mathcal{F}} \sum_{K \in \mathcal{T}_h} \|D^2 v_h\|_{L^2(K)}^2 + \sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|\llbracket \nabla v_h \cdot n_F \rrbracket\|_{L^2(F)}^2, \quad (6.8)$$

$$|I_2| \leq \delta C(n) C_{\text{Tr}} c_{\mathcal{F}} \sum_{K \in \mathcal{T}_h} \|D^2 v_h\|_{L^2(K)}^2 + \sum_{F \in \mathcal{F}_h^{i,b}} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|\llbracket \nabla_T v_h \rrbracket\|_{L^2(F)}^2, \quad (6.9)$$

where  $C_{\text{Tr}}$  is the combined constant of the trace and inverse inequalities, and  $c_{\mathcal{F}}$  is given by (4.1). The inverse and trace inequalities also show that

$$\begin{aligned} |I_3| &\leq 2\lambda \sqrt{\sum_{F \in \mathcal{F}_h^i} \frac{\delta \tilde{h}_F}{\tilde{p}_F^2} \|\{v_h\}\|_{L^2(F)}^2} \sqrt{\sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|\llbracket \nabla v_h \cdot n_F \rrbracket\|_{L^2(F)}^2} \\ &\leq \delta C(n) C_{\text{Tr}} c_{\mathcal{F}} \sum_{K \in \mathcal{T}_h} \lambda^2 \|v_h\|_{L^2(K)}^2 + \sum_{F \in \mathcal{F}_h^i} \frac{\tilde{p}_F^2}{\delta \tilde{h}_F} \|\llbracket \nabla v_h \cdot n_F \rrbracket\|_{L^2(F)}^2. \end{aligned} \quad (6.10)$$

Similarly, it is found that

$$|I_4| \leq \delta C(n) C_{\text{Tr}} c_{\mathcal{F}} \sum_{K \in \mathcal{T}_h} 2\lambda \|\nabla v_h\|_{L^2(K)}^2 + \sum_{F \in \mathcal{F}_h^{i,b}} \frac{\lambda \tilde{p}_F^2}{2\delta \tilde{h}_F} \|\llbracket v_h \rrbracket\|_{L^2(F)}^2. \quad (6.11)$$

We may take  $C(n)$  to be the same constant in each of the above estimates. So,

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

For  $\kappa > 1$ , there is a  $\delta > 0$  such that  $1 - \delta C(n)C_{\text{Tr},c\mathcal{F}} > \kappa^{-1}$ . Set  $c_{\text{stab}} = 4/\delta$  and  $c_* = \kappa/2$  so that (6.6) holds when  $\mu_F$  and  $\eta_F$  are chosen in accordance with (6.7).  $\square$

**THEOREM 7.** *Let  $\Omega$  be a bounded convex polytopal domain and let  $\{\mathcal{T}_h\}_h$  be a shape-regular sequence of simplicial or parallelepipedal meshes satisfying (4.1). Let  $\Lambda$  be a compact metric space and let the data satisfy (2.4) and either (2.5) or (2.6) with  $b \equiv 0$ ,  $c \equiv 0$ ,  $\lambda = 0$ . Let  $c_{\text{stab}}$ ,  $c_*$ ,  $\eta_F$  and  $\mu_F$  be chosen so that Lemma 6 holds with  $\kappa < (1 - \varepsilon)^{-1/2}$ . Then, for every  $u_h, v_h \in V_{h,\mathbf{p}}$ , we have*

$$\|u_h - v_h\|_{\text{DG}(1)}^2 \leq C(A_h(u_h; u_h - v_h) - A_h(v_h; u_h - v_h)), \quad (6.12)$$

where the constant  $C := 2\kappa/(1 - \kappa^2(1 - \varepsilon))$ . Moreover, there exists a constant  $C$  independent of  $h$  and  $\mathbf{p}$  such that, for any  $u_h, v_h$  and  $z_h$  in  $V_{h,\mathbf{p}}$ ,

$$|A_h(u_h; z_h) - A_h(v_h; z_h)| \leq C\|u_h - v_h\|_{\text{DG}(1)}\|z_h\|_{\text{DG}(1)}. \quad (6.13)$$

Therefore, there exists a unique solution  $u_h \in V_{h,\mathbf{p}}$  to the numerical scheme (5.4). We have the bound

$$\|u_h\|_{\text{DG}(1)} \leq \frac{2\kappa\sqrt{n+1}\|\gamma\|_{C(\bar{\Omega} \times \Lambda)}}{1 - \kappa^2(1 - \varepsilon)} \|\sup_{\alpha \in \Lambda} |f^\alpha|\|_{L^2(\Omega)}. \quad (6.14)$$

*Proof.* Let  $u_h$  and  $v_h$  belong to  $V_{h,\mathbf{p}}$  and set  $w_h := u_h - v_h$ . Then, we have

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

Note that Lemma 1 gives

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

This estimate and Lemma 6 show that

$$\begin{aligned} A_h(u_h; w_h) - A_h(v_h; w_h) &\geq \frac{1 - \kappa^2(1 - \varepsilon)}{2\kappa} |w_h|_{H^2(\Omega;\mathcal{T}_h),\lambda}^2 + \frac{c_*}{\kappa} J_h(w_h, w_h) \\ &\geq \frac{1 - \kappa^2(1 - \varepsilon)}{2\kappa} \|w_h\|_{\text{DG}(1)}^2. \end{aligned}$$

Since  $\kappa^2(1 - \varepsilon) < 1$ , we obtain (6.12). Now let  $z_h \in V_{h,\mathbf{p}}$ . Then, using 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 6, but not on  $h$  or  $\mathbf{p}$ , such that  $|B_{h,\frac{1}{2}}(u_h - v_h, z_h)| \leq C \|u_h - v_h\|_{\text{DG}(1)} \|z_h\|_{\text{DG}(1)}$ . Using Lemma 1 and the above estimates, we deduce that there is a constant  $C$  depending only on  $n$  and  $\varepsilon$  such that

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

It then follows that  $A_h$  is Lipschitz continuous, as stated in (6.13). The Browder–Minty theorem [24] with (6.12) and (6.13) imply that there exists a unique  $u_h \in V_{h,\mathbf{p}}$  such that  $A_h(u_h; v_h) = 0$  for all  $v_h \in V_{h,\mathbf{p}}$ . By taking  $v_h = 0$  in (6.12), we find that

$$\begin{aligned} \|u_h\|_{\text{DG}(1)}^2 &\leq C |A_h(0; u_h)| \leq C \sum_{K \in \mathcal{T}_h} |\langle \sup_{\alpha \in \Lambda} [-\gamma^\alpha f^\alpha], L_\lambda u_h \rangle_K| \\ &\leq C \|\gamma\|_{C(\bar{\Omega} \times \Lambda)} \|\sup_{\alpha \in \Lambda} |f^\alpha|\|_{L^2(\Omega)} \sqrt{n+1} \|u_h\|_{\text{DG}(1)}, \end{aligned}$$

where  $C = 2\kappa / (1 - \kappa^2(1 - \varepsilon))$ , thus showing the bound (6.14).  $\square$

**7. Error analysis.** The good stability properties of the proposed method make it possible to obtain the following *a priori* error bound.

**THEOREM 8.** *Let  $\Omega$  be a bounded convex polytopal domain, let the shape-regular sequence of simplicial or parallelepipedal meshes  $\{\mathcal{T}_h\}_h$  satisfy (4.1) and (4.2), with  $\mathbf{p}$  satisfying (4.3) for each  $h$ . Let  $\Lambda$  be a compact metric space and let the data satisfy (2.4), and either (2.5) or (2.6) 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 (2.3). Assume either that  $u \in H^s(\Omega)$ ,  $s > 5/2$ , or that  $u \in C^1(\bar{\Omega})$ , and assume in addition that  $u \in H^s(\Omega; \mathcal{T}_h)$ , with  $s_K \geq 3$  for all  $K \in \mathcal{T}_h$ . Let  $c_{\text{stab}}$ ,  $c_*$ ,  $\mu_F$  and  $\eta_F$  be chosen as in Theorem 7 and choose  $\eta_F \lesssim \tilde{p}_F^4 / \tilde{h}_F^3$  for all  $F \in \mathcal{F}_h^{i,b}$ . Then, there exists a positive constant  $C$ , independent of  $h$ ,  $\mathbf{p}$  and  $u$ , but depending on  $\max_K s_K$ , such that*

$$\|u - u_h\|_{\text{DG}(1)}^2 \leq C \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K - 4}}{p_K^{2s_K - 5}} \|u\|_{H^{s_K}(K)}^2, \quad (7.1)$$

where  $t_K = \min(p_K + 1, s_K)$  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  $s \geq 3$ , the *a-priori* estimate (7.1) simplifies to

$$\|u - u_h\|_{\text{DG}(1)} \leq C \frac{h^{\min(p+1, s) - 2}}{p^{s-2.5}} \|u\|_{H^s(\Omega)}.$$

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

*Proof.* Since the sequence of meshes is shape-regular, there is 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

$$\|u - z_h\|_{H^q(K)} \leq C \frac{h_K^{t_K - q}}{p_K^{s_K - q}} \|u\|_{H^{s_K}(K)}, \quad 0 \leq q \leq s_K, \quad (7.2)$$

$$\|D^\beta(u - z_h)\|_{L^2(\partial K)} \leq C \frac{h_K^{t_K - q - 1/2}}{p_K^{s_K - q - 1/2}} \|u\|_{H^{s_K}(K)}, \quad |\beta| = q; \quad 0 \leq q \leq s_K - 1. \quad (7.3)$$

Note that the hypothesis  $s_K \geq 3$  allows the choice of  $q = 2$  in (7.3). Set  $\psi_h := u_h - z_h$  and  $\xi_h := u - z_h$ . By Corollary 5, 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 7, yields

$$\|\psi_h\|_{\text{DG}(1)}^2 \lesssim A_h(u_h; \psi_h) - A_h(z_h; \psi_h) = A_h(u; \psi_h) - A_h(z_h; \psi_h). \quad (7.4)$$

By applying the Cauchy–Schwarz inequality to the terms appearing on the right-hand side of (7.4) and applying inverse inequalities to  $\psi_h \in V_{h,\mathbf{p}}$ , we eventually obtain

$$A_h(u; \psi_h) - A_h(z_h; \psi_h) \leq \sqrt{\sum_{i=1}^{10} E_i} \|\psi_h\|_{\text{DG}(1)}, \quad (7.5)$$

where the quantities  $E_i$  are defined by

$$\begin{aligned} E_1 &:= \sum_{K \in \mathcal{T}_h} |\xi_h|_{H^2(K), \lambda}^2, & E_2 &:= \sum_{K \in \mathcal{T}_h} \|L_\lambda \xi_h\|_{L^2(K)}^2, \\ E_3 &:= \sum_{K \in \mathcal{T}_h} \|F_\gamma[u] - F_\gamma[z_h]\|_{L^2(K)}^2, & E_4 &:= \sum_{F \in \mathcal{F}_h^i} \mu_F^{-1} \|\text{div}_T \nabla_T \{\xi_h\}\|_{L^2(F)}^2, \\ E_5 &:= \sum_{F \in \mathcal{F}_h^i} \mu_F \|\llbracket \nabla \xi_h \cdot n_F \rrbracket\|_{L^2(F)}^2, & E_6 &:= \sum_{F \in \mathcal{F}_h^{i,b}} \mu_F^{-1} \|\nabla_T \{\nabla \xi_h \cdot n_F\}\|_{L^2(F)}^2, \\ E_7 &:= \sum_{F \in \mathcal{F}_h^{i,b}} \mu_F \|\llbracket \nabla_T \xi_h \rrbracket\|_{L^2(F)}^2, & E_8 &:= \sum_{F \in \mathcal{F}_h^{i,b}} \lambda^2 \eta_F^{-1} \|\{\nabla \xi_h \cdot n_F\}\|_{L^2(F)}^2, \\ E_9 &:= \sum_{F \in \mathcal{F}_h^{i,b}} (\lambda \mu_F + \eta_F) \|\llbracket \xi_h \rrbracket\|_{L^2(F)}^2, & E_{10} &:= \sum_{F \in \mathcal{F}_h^i} \lambda^2 \mu_F^{-1} \|\{\xi_h\}\|_{L^2(F)}^2. \end{aligned}$$

The estimate (7.2) shows that

$$E_1 + E_2 \lesssim \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}} \|u\|_{H^{s_K}(K)}^2. \quad (7.6)$$

By compactness of  $\Lambda$ , continuity of the data and (2.4),  $F_\gamma$  is Lipschitz continuous, so

$$E_3 \lesssim \sum_K \|\xi_h\|_{H^2(K)}^2 \lesssim \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-4}} \|u\|_{H^{s_K}(K)}^2. \quad (7.7)$$

We use (4.1), (4.2), (4.3), (6.7) and (7.3) to obtain

$$E_4 + E_6 \lesssim \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \frac{h_K^{2t_K-5}}{p_K^{2s_K-5}} \|u\|_{H^{s_K}(K)}^2 = \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-3}} \|u\|_{H^{s_K}(K)}^2, \quad (7.8)$$

$$E_5 + E_7 \lesssim \sum_{K \in \mathcal{T}_h} \frac{p_K^2}{h_K} \frac{h_K^{2t_K-3}}{p_K^{2s_K-3}} \|u\|_{H^{s_K}(K)}^2 = \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|_{H^{s_K}(K)}^2. \quad (7.9)$$

Similarly, we use (6.7) to get

$$E_8 \lesssim \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \frac{h_K^{2t_K-3}}{p_K^{2s_K-3}} \|u\|_{H^{s_K}(K)}^2 = \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-2}}{p_K^{2s_K-1}} \|u\|_{H^{s_K}(K)}^2. \quad (7.10)$$

By hypothesis,  $\eta_F \lesssim \tilde{p}_F^4 / \tilde{h}_F^3$ , so (4.2) and (4.3) imply that

$$E_9 \lesssim \sum_{K \in \mathcal{T}_h} \frac{p_K^4}{h_K^3} \frac{h_K^{2t_K-1}}{p_K^{2s_K-1}} \|u\|_{H^{s_K}(K)}^2 = \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K-4}}{p_K^{2s_K-5}} \|u\|_{H^{s_K}(K)}^2. \quad (7.11)$$

Finally, (7.3) yields

$$E_{10} \lesssim \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \frac{h_K^{2t_K-1}}{p_K^{2s_K-1}} \|u\|_{H^{s_K}(K)}^2 = \sum_{K \in \mathcal{T}_h} \frac{h_K^{2t_K}}{p_K^{2s_K+1}} \|u\|_{H^{s_K}(K)}^2. \quad (7.12)$$

The *a-priori* bound (7.1) is obtained from  $\|u - u_h\|_{\text{DG}(1)} \leq \|\psi_h\|_{\text{DG}(1)} + \|\xi_h\|_{\text{DG}(1)}$  and the above estimates.  $\square$

**8. Semismooth Newton method.** We turn to the analysis of an algorithm for solving the discrete problem (5.4), which can be interpreted as a Newton method for non-smooth operator equations [23]. After showing that the algorithm is well-posed, we obtain and then use a semismoothness result for HJB operators in function spaces to establish its superlinear convergence. The semismoothness of finite dimensional HJB operators in a different form was studied in [3].

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  $\mathbf{u} = (u, \nabla_h u, D_h^2 u)$ , where  $\nabla_h u$  and  $D_h^2 u$  denote the broken gradient and broken Hessian of  $u$ , see §4. For a vector  $\mathbf{u} = (z, p, M) \in \mathbb{R}^m$ , define the function  $F_\gamma: \Omega \times \mathbb{R}^m \rightarrow \mathbb{R}$  by

$$F_\gamma(x, \mathbf{u}) := \sup_{\alpha \in \Lambda} [\gamma^\alpha (a^\alpha: M + b^\alpha \cdot p - c^\alpha z - f^\alpha)|_x]. \quad (8.1)$$

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 (8.1) is attained. This defines a set-valued map  $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$ .

**LEMMA 9.** *Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$ , let  $\Lambda$  be a compact metric space, let the data  $a, b, c$  and  $f$  be continuous on  $\bar{\Omega} \times \Lambda$  and suppose that (2.4) holds. Then, for each  $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$ ,  $\Lambda(x, \mathbf{u})$  is a non-empty closed subset of  $\Lambda$ . 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 (2.4) is only used in Lemma 9 to guarantee that  $\gamma \in C(\bar{\Omega} \times \Lambda)$ .

*Proof.* For every  $(x, \mathbf{u}) \in \Omega \times \mathbb{R}^m$ ,  $u = (z, p, M)$ , compactness of  $\Lambda$  and continuity of  $a, b, c, f$  and  $\gamma$  imply the existence of a maximiser in (2.2); so  $\Lambda(x, \mathbf{u})$  is non-empty. The set  $\Lambda(x, \mathbf{u})$  is closed: if  $\alpha$  is in the closure of  $\Lambda(x, u)$ , say  $\alpha_j \rightarrow \alpha$ , with  $\alpha_j \in \Lambda(x, u)$  for each  $j \in \mathbb{N}$ , then continuity of the data implies that

$$\gamma^\alpha (a^\alpha: M + b^\alpha \cdot p - c^\alpha z - f^\alpha)|_x = \lim_{j \rightarrow \infty} \gamma^{\alpha_j} (a^{\alpha_j}: M + b^{\alpha_j} \cdot p - c^{\alpha_j} z - f^{\alpha_j})|_x. \quad (8.2)$$

Since  $\alpha_j \in \Lambda(x, \mathbf{u})$  for each  $j \in \mathbb{N}$ , the right hand side of (8.2) 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  $\Lambda$  is compact and  $\Lambda \setminus U$  is closed, there exists a subsequence,

to which we pass without change of notation, such that  $\alpha_j \rightarrow \alpha \in \Lambda \setminus U$ . On the one hand,  $\Lambda(x, \mathbf{u})$  is non-empty so there is  $\beta \in \Lambda(x, \mathbf{u})$ . Then, by definition of  $F$ ,

$$\gamma^\alpha (a^\alpha : M + b^\alpha \cdot p - c^\alpha z - f^\alpha)|_x \leq F(x, \mathbf{u}). \quad (8.3)$$

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

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

Taking the limit  $j \rightarrow \infty$  in the above inequality shows that equality holds in (8.3) 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.  $\square$

The following selection theorem, due to Kuratowski and Ryll-Nardzewski [17], is required for the analysis of the algorithm for solving (5.4).

**THEOREM 10.** *Let  $\Omega \subset \mathbb{R}^n$  be a bounded open set, let  $\Lambda$  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  $\Lambda$ , 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 finite a.e. Lebesgue measurable function  $\mathbf{u}: \Omega \rightarrow \mathbb{R}^m$ , there exists a Lebesgue measurable selection  $\alpha: \Omega \rightarrow \Lambda$  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  $\alpha: \Omega \rightarrow \Lambda$  such that  $\alpha(x) \in \Lambda(x, \mathbf{u}(x))$  for a.e.  $x \in \Omega$ , where  $\mathbf{u} = (u, \nabla_h u, D_h^2 u)$ . Lemma 9 and Theorem 10 show that  $\Lambda[u]$  is non-empty for each  $u \in W^{2,r}(\Omega; \mathcal{T}_h)$ . For measurable  $\alpha: \Omega \rightarrow \Lambda$ , we define  $\gamma^\alpha: \Omega \rightarrow \mathbb{R}_{>0}$  through  $\gamma^\alpha(x) = \gamma(x, \alpha(x))$ , where  $\gamma: \Omega \times \Lambda \rightarrow \mathbb{R}_{>0}$  was defined by (2.9) or (2.10). It follows from uniform continuity of  $\gamma$  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  $a^\alpha$ ,  $b^\alpha$ ,  $c^\alpha$  and  $f^\alpha$  and the operator  $L^\alpha$  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  $\Omega$ .

**8.1. Algorithm.** We now present the definition of the semismooth Newton method for solving (5.4) 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

$$A_h^k(u_h^{k+1}, v_h) = \sum_{K \in \mathcal{T}_h} \langle \gamma^{\alpha_k} f^{\alpha_k}, L_\lambda v_h \rangle_K \quad \forall v_h \in V_{h,\mathbf{p}}, \quad (8.4)$$

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

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

The fact that  $\alpha_k: \Omega \rightarrow \Lambda$  is measurable ensures that  $A_h^k$  is well-defined. As in the proof of Theorem 7, 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

$$\|v_h\|_{\text{DG}(1)}^2 \leq \frac{2\kappa}{1 - \kappa^2(1 - \varepsilon)} A_h^k(v_h, v_h) \quad \forall v_h \in V_{h,\mathbf{p}}. \quad (8.5)$$

Therefore, the sequence of iterates  $\{u_h^k\}_{k=1}^\infty$  is well-defined by (8.4) and remains bounded in  $V_{h,\mathbf{p}}$ . The main result of this section is the following.



**THEOREM 11.** *Under the hypotheses of Theorem 7, there exists a constant  $R > 0$ , possibly depending on  $h$  and  $\mathbf{p}$ , such that if  $\|u_h - u_h^0\|_{\text{DG}(1)} < R$ , where  $u_h$  solves (5.4), 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. Despite the possible dependence of  $R$  on  $h$  and  $p$  in the above theorem, it is seen from the numerical experiments in §9, in particular in Figure 2 below, that in practice, the convergence rates of the algorithm depend only weakly on the discretisation parameters.

**8.2. Semismoothness of HJB operators.** The proof of Theorem 11 rests upon the notion of semismoothness, as defined in [27]. We recall the definition below. For sets  $X$  and  $Y$ , we write  $G: X \rightrightarrows Y$  if  $G$  is a set-valued map that maps  $X$  into the subsets of  $Y$ .

**DEFINITION 12.** *Let  $X$  and  $Y$  be Banach spaces, and let  $F: U \subset X \rightarrow Y$  be a map defined on a non-empty open set  $U$  of  $X$ . Let  $DF: U \rightrightarrows \mathcal{L}(X, Y)$  be a set-valued map with non-empty images. For  $x \in U$ , the map  $F$  is called  $DF$ -semismooth at  $x$  if*

$$\lim_{\|e\|_X \rightarrow 0} \frac{1}{\|e\|_X} \sup_{D \in DF[x+e]} \|F[x+e] - F[x] - De\|_Y = 0. \quad (8.6)$$

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

$$DF_\gamma[u] := \{\gamma^\alpha L^\alpha := \gamma^\alpha (a^\alpha : D_h^2 + b^\alpha \cdot \nabla_h - c^\alpha) : \alpha \in \Lambda[u]\}. \quad (8.7)$$

**THEOREM 13.** *Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$ , let  $\Lambda$  be a compact metric space, let the data  $a, b, c$  and  $f$  be continuous on  $\bar{\Omega} \times \Lambda$  and suppose that (2.4) holds. Let  $\mathcal{T}_h$  be a mesh on  $\Omega$ . Then, for any  $1 \leq q < r \leq \infty$ , the operator  $F_\gamma: W^{2,r}(\Omega; \mathcal{T}_h) \rightarrow L^q(\Omega)$  defined by  $F_\gamma[u] = F_\gamma(\cdot, u, \nabla_h u, D_h^2 u)$  is  $DF_\gamma$ -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  $\|e_j\|_{W^{2,r}(\Omega; \mathcal{T}_h)} \rightarrow 0$ , and  $\alpha_j \in \Lambda[u + e_j]$  such that, for each  $j \in \mathbb{N}$ ,

$$\frac{1}{\|e_j\|_{W^{2,r}(\Omega; \mathcal{T}_h)}} \|F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j\|_{L^q(\Omega)} > \rho. \quad (8.8)$$

We will show that there is a subsequence for which (8.8) is violated, and thus obtain a contradiction. Since  $\|e_j\|_{W^{2,r}(\Omega; \mathcal{T}_h)} \rightarrow 0$ , 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  $\Omega$ . The following inequality will help to simplify the argument:

$$|F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j| \lesssim G_j (|e_j| + |\nabla_h e_j| + |D_h^2 e_j|), \quad (8.9)$$

where  $G_j: \Omega \rightarrow \mathbb{R}_{\geq 0}$  is defined by

$$G_j := \inf_{\alpha \in \Lambda(\cdot, \mathbf{u}(\cdot))} |\gamma^\alpha a^\alpha - \gamma^{\alpha_j} a^{\alpha_j}| + |\gamma^\alpha b^\alpha - \gamma^{\alpha_j} b^{\alpha_j}| + |\gamma^\alpha c^\alpha - \gamma^{\alpha_j} c^{\alpha_j}|. \quad (8.10)$$

It can be deduced from Lemma 9 that  $G_j$  is measurable, since it is the composition of a lower semi-continuous function with a measurable function; compactness of  $\Lambda$  and continuity of the data imply that  $\|G_j\|_{L^\infty(\Omega)}$  is uniformly bounded for all  $j \in \mathbb{N}$ .

We prove (8.9): since  $\alpha_j \in \Lambda[u + e_j]$ , we have a.e. in  $\Omega$ :

$$F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j = \gamma^{\alpha_j} (L^{\alpha_j} u - f^{\alpha_j}) - F_\gamma[u] \leq 0. \quad (8.11)$$

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

$$\begin{aligned} 0 &\leq F_\gamma[u + e_j] - \gamma^\alpha (L^\alpha(u + e_j) - f^\alpha) \\ &= \gamma^{\alpha_j} (L^{\alpha_j} u - f^{\alpha_j}) - F_\gamma[u] + (\gamma^{\alpha_j} L^{\alpha_j} - \gamma^\alpha L^\alpha) e_j \\ &= F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j + (\gamma^{\alpha_j} L^{\alpha_j} - \gamma^\alpha L^\alpha) e_j, \end{aligned} \quad (8.12)$$

where it is understood that the above expressions are evaluated at  $x$ . Rearranging (8.11) and (8.12) gives  $(\gamma^\alpha L^\alpha - \gamma^{\alpha_j} L^{\alpha_j}) e_j \leq F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j \leq 0$ , so

$$|F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j| \leq |(\gamma^\alpha L^\alpha - \gamma^{\alpha_j} L^{\alpha_j}) e_j|. \quad (8.13)$$

Since (8.13) holds for arbitrary  $\alpha \in \Lambda(x, \mathbf{u}(x))$ , we readily obtain (8.9).

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

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

Since  $(x, \mathbf{u}) \mapsto \Lambda(x, \mathbf{u})$  is upper-semicontinuous by Lemma 9, 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  $\text{dist}(\alpha, \alpha_j(x)) < \delta$ . Therefore  $0 \leq G_j(x) < \varrho$  for all  $j \geq N$ , and hence  $G_j \rightarrow 0$  pointwise a.e. in  $\Omega$ .

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 (8.9) followed by an application of Hölder's inequality shows that

$$\frac{1}{\|e_j\|_{W^{2,r}(\Omega; \mathcal{T}_h)}} \|F_\gamma[u + e_j] - F_\gamma[u] - \gamma^{\alpha_j} L^{\alpha_j} e_j\|_{L^q(\Omega)} \lesssim \|G_j\|_{L^{qs'}(\Omega)}, \quad (8.14)$$

Since  $G_j \rightarrow 0$  pointwise a.e. and  $\{G_j\}_{j=0}^\infty$  is uniformly bounded in  $L^\infty(\Omega)$ , the dominated convergence theorem implies that  $\|G_j\|_{L^{qs'}(\Omega)} \rightarrow 0$ . Therefore, (8.14) contradicts (8.8), and  $F_\gamma$  is  $DF_\gamma$ -semismooth at  $u$ , thus completing the proof.  $\square$

**REMARK 1.** *The restriction  $q < r$  in Theorem 13 cannot be relaxed in general, as evidenced by the counter-example in [13] involving a special case of the class of operators considered here.*

*Proof of Theorem 11.* 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, (8.4) is equivalent to

$$A_h^k(u_h^{k+1}, v_h) = \sum_{K \in \mathcal{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h^k - F_\gamma[u_h^k], L_\lambda v_h \rangle_K \quad \forall v_h \in V_{h,\mathbf{p}}. \quad (8.15)$$

The definition of the numerical scheme (5.4) implies that  $u_h$  satisfies

$$A_h^k(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \langle \gamma^{\alpha_k} L^{\alpha_k} u_h - F_\gamma[u_h], L_\lambda v_h \rangle_K \quad \forall v_h \in V_{h,\mathbf{p}}. \quad (8.16)$$

After subtracting (8.16) from (8.15), the bound (8.5) then shows that

$$\|u_h^{k+1} - u_h\|_{\text{DG}(1)} \leq C_1 \|F_\gamma[u_h^k] - F_\gamma[u_h] - \gamma_k^\alpha L_k^\alpha (u_h^k - u_h)\|_{L^2(\Omega)}, \quad (8.17)$$

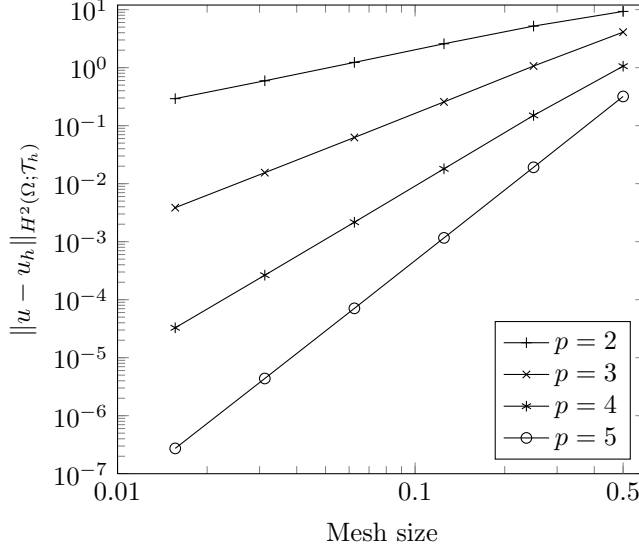


FIG. 1. The errors in approximating the solution of the problem of §9.1 for various mesh sizes and polynomial degrees. The optimal convergence rates  $\|u - u_h\|_{H^2(\Omega; \mathcal{T}_h)} = \mathcal{O}(h^{p-1})$  are observed.

where the constant  $C_1$  depends only on  $\kappa$ ,  $\varepsilon$ ,  $\gamma$ , and  $n$  as in (6.14), 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  $\|v_h\|_{W^{2,r}(\Omega; \mathcal{T}_h)} \leq C_2 \|v_h\|_{\text{DG}(1)}$  for all  $v_h \in V_{h, \mathbf{p}}$ . Theorem 13 shows that for each  $\rho \in (0, 1)$ , there is a  $R_\rho > 0$  such that if  $\|w_h - u_h\|_{\text{DG}(1)} < R_\rho$ , then, for any  $\alpha \in \Lambda[w_h]$ ,

$$\|F_\gamma[w_h] - F_\gamma[u_h] - \gamma^\alpha L^\alpha(w_h - u_h)\|_{L^2(\Omega)} \leq \frac{\rho}{C_1 C_2} \|w_h - u_h\|_{W^{2,r}(\Omega; \mathcal{T}_h)}. \quad (8.18)$$

If  $\|u_h^0 - u_h\|_{\text{DG}(1)} < R_\rho$  for some  $\rho < 1$ , then we use (8.17) and (8.18) to obtain

$$\|u_h^{k+1} - u_h\|_{\text{DG}(1)} \leq \rho \|u_h^k - u_h\|_{\text{DG}(1)} \quad \forall k \geq 0,$$

which yields convergence of  $u_h^k$  to  $u_h$ . For each  $\rho < 1$ ,  $\|u_h^k - u_h\|_{\text{DG}(1)} < R_\rho$  is then eventually satisfied, thus implying a superlinear convergence rate.  $\square$

**9. Numerical experiments.** We provide the results of two tests of the scheme on problems with strongly anisotropic diffusion coefficients.

**9.1. First experiment.** We consider once again Example 1 for testing the accuracy of the scheme and the performance of the semismooth Newton method. Recalling that  $\Lambda = [0, \pi/3] \times \text{SO}(2)$  and  $a^\alpha = \sigma^\alpha (\sigma^\alpha)^\top / 2$ , with  $\sigma^\alpha$  given by (2.7), 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  $\alpha$ , so that the exact solution of the HJB equation (2.3) 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  $\Omega$ .

The numerical scheme (5.4) is applied with meshes obtained by regular subdivision of  $\Omega$  into uniform quadrilateral elements of size  $h = 2^{-k}$ ,  $1 \leq k \leq 6$ . 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  $2 \leq p \leq 5$ . The penalty parameters are set to  $c_{\text{stab}} = 10$  and  $\eta_F = c_{\text{stab}} \tilde{p}_F^4 / \tilde{h}_F^3$ . Figure 1 confirms the optimal convergence rates with respect to mesh refinement that are predicted by Theorem 8.

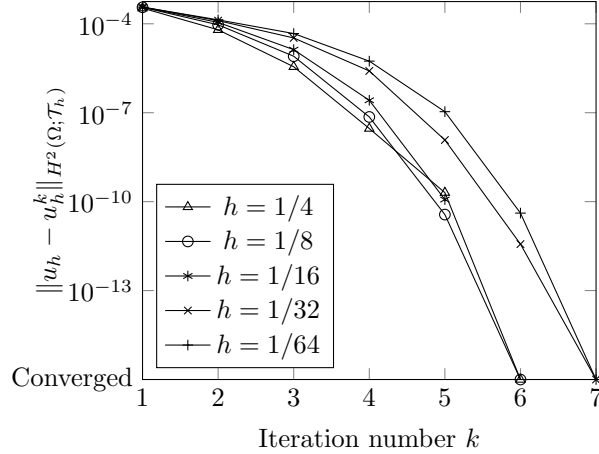


FIG. 2. Convergence histories of the semismooth Newton method applied to the problem of §9.1 on successively refined meshes, with  $p = 4$ . The predicted superlinear convergence rate is observed, and the number of iterations required for convergence shows little variation under refinement.

The numerical solutions were obtained by the semismooth Newton method of §8, for which we use a strict convergence criterion by requiring a relative residual below  $5 \times 10^{-12}$  and a step-increment  $L^2$ -norm below  $1 \times 10^{-11}$ . The initial guess used for each computation was  $u_h^0 \equiv 0$ . The convergence histories shown in Figure 2 demonstrate the fast convergence of the algorithm.

**9.2. Second experiment.** We investigate the robustness of the scheme against a combination of near-degenerate diffusions, non-smooth solutions and boundary layers. This example is to our knowledge the first fully nonlinear second-order problem solved with an exponentially accurate scheme. Let  $\Omega = (0, 1)^2$ ,  $b^\alpha \equiv (0, 1)$ ,  $c^\alpha \equiv 10$  and define

$$a^\alpha := \alpha^T \begin{pmatrix} 20 & 1 \\ 1 & 0.1 \end{pmatrix} \alpha, \quad \alpha \in \Lambda := \text{SO}(2). \quad (9.1)$$

For  $\lambda = 1/2$ , the Cordès condition (2.5) holds with  $\varepsilon \approx 0.0024$ . We choose  $f^\alpha$  so that the solution of the corresponding HJB equation is

$$u(x, y) = (2x - 1) \left( e^{1-|2x-1|} - 1 \right) \left( y + \frac{1 - e^{y/\delta}}{e^{1/\delta} - 1} \right), \quad \delta > 0. \quad (9.2)$$

Note that  $u \in C^1(\bar{\Omega})$  and  $u \notin H^3(\Omega)$ . We choose  $\delta = 0.005$  to be of same order as  $\varepsilon$ , thus leading to a sharp boundary layer in a neighbourhood of  $\{(x, y) \in \bar{\Omega} : y = 1\}$ .

The results of [5] show that a very large stencil would be necessary to obtain a consistent monotone FD discretisation of 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 consistency and 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

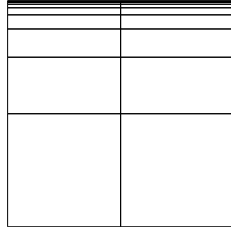


FIG. 3. Mesh on  $\Omega$  used for the approximation of (9.2). The origin is at the bottom left corner. The mesh has 8 geometrically refined layers with grading factor  $1/2$ .

for a smaller computational cost. Following a suggestion in [20], we perform a sequence of computations by increasing the uniform polynomial degrees  $p$  from 2 to 10 on a fixed mesh shown in Figure 3. The number of degrees of freedom ranges from 100 to 1320 and the following results were obtained with  $c_{\text{stab}} = 10$ , as in §9.1. Figure 4 shows that the error converges with a rate of  $\mathcal{O}(\exp(-c\sqrt[3]{\text{DoF}}))$ , which leads to high accuracy with few degrees of freedom.

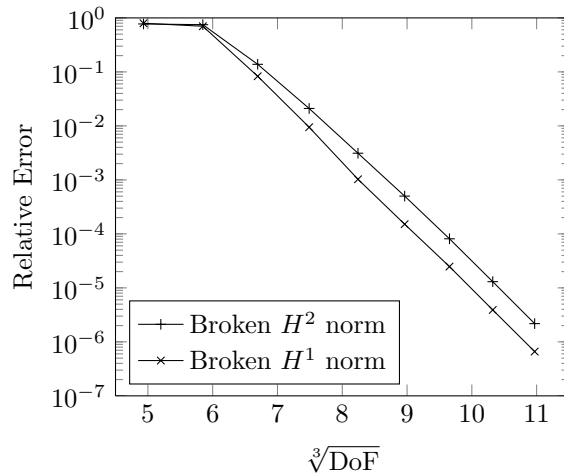


FIG. 4. Exponential convergence in the broken  $H^1$  and  $H^2$  norms of the approximations to the solution defined by (9.2). The relative errors  $\|u - u_h\|/\|u\|$  are plotted against the cube root of the number of degrees of freedom, with each data point corresponding to a computation using a total polynomial degree  $p = 2, \dots, 10$ .

**10. Conclusion.** We have considered the PDE analysis and numerical analysis of HJB equations that satisfy the Cordès condition. Our contributions include an existence and uniqueness result for strong solutions to the fully nonlinear problem, the construction of a consistent and stable  $hp$ -version DGFEM with proven convergence rates, and a study of the semismoothness of HJB operators. The numerical experiments demonstrated the high efficiency and accuracy of the scheme and the fast convergence of the semismooth Newton method, whilst also highlighting the wide applicability of the results of this work.

## REFERENCES

- [1] G. BARLES AND P. SOUGANIDIS, *Convergence of approximation schemes for fully nonlinear second-order equations*, *Asymptotic Anal.*, 4 (1991), pp. 271–283.
- [2] K. BÖHMER, *On finite element methods for fully nonlinear elliptic equations of second order*, *SIAM J. Numer. Anal.*, 46 (2008), pp. 1212–1249.
- [3] O. BOKANOWSKI, S. MAROSO, AND H. ZIDANI, *Some convergence results for Howard’s algorithm*, *SIAM J. Numer. Anal.*, 47 (2009), pp. 3001–3026.
- [4] J. F. BONNANS, É. OTTENWALTER, 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.
- [5] 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.
- [6] M. G. CRANDALL, H. ISHII, AND P.-L. LIONS, *User’s guide to viscosity solutions of second-order partial differential equations*, *Bull. Amer. Math. Soc. (N.S.)*, 27 (1992), pp. 1–67.
- [7] 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.
- [8] D. A. DI PIETRO AND A. ERN, *Mathematical Aspects of Discontinuous Galerkin Methods*, vol. 69 of *Mathématiques & Applications (Berlin)*, Springer, Heidelberg, 2012.
- [9] 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.
- [10] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, vol. 25 of *Stochastic Modelling and Applied Probability*, Springer, New York, 2006.
- [11] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, *Classics in Mathematics*, Springer-Verlag, Berlin, 2001.
- [12] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, vol. 69 of *Classics in Applied Mathematics*, SIAM, Philadelphia, 2011.
- [13] M. HINTERMULLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, *SIAM Journal on Optimization*, 13 (2002), p. 865.
- [14] P. HOUSTON, CH. SCHWAB, AND E. SÜLI, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, *SIAM J. Numer. Anal.*, 39 (2002), pp. 2133–2163.
- [15] M. JENSEN AND I. SMEARS, *On the convergence of finite element methods for Hamilton–Jacobi–Bellman equations*, *SIAM Journal on Numerical Analysis*, 51 (2013), pp. 137–162.
- [16] M. KOCAN, *Approximation of viscosity solutions of elliptic partial differential equations on minimal grids*, *Numer. Math.*, 72 (1995), pp. 73–92.
- [17] 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.
- [18] H. J. KUSHNER, *Numerical methods for stochastic control problems in continuous time*, *SIAM J. Control Optim.*, 28 (1990), pp. 999–1048.
- [19] A. MAUGERI, D. K. PALAGACHEV, AND L. G. SOFTOVA, *Elliptic and Parabolic Equations with Discontinuous Coefficients*, vol. 109 of *Mathematical Research*, Wiley, Berlin, 2000.
- [20] J. M. MELENK, *hp-Finite Element Methods for Singular Perturbations*, vol. 1796 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 2002.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] M. RENARDY AND R. C. ROGERS, *An Introduction to Partial Differential Equations*, vol. 13 of *Texts in Applied Mathematics*, Springer-Verlag, New York, second ed., 2004.
- [25] M. V. SAFONOV, *Nonuniqueness for second-order elliptic equations with measurable coefficients*, *SIAM J. Math. Anal.*, 30 (1999), pp. 879–895.
- [26] I. SMEARS AND E. SÜLI, *Discontinuous Galerkin finite element approximation of non-divergence form elliptic equations with Cordès coefficients*, *Tech. Report NA 12/17*, Univ. of Oxford, 2012. In Review. Available at <http://eprints.maths.ox.ac.uk/1623/>
- [27] M. ULBRICH, *Semismooth Newton methods for operator equations in function spaces*, *SIAM J. Optim.*, 13 (2002), pp. 805–842 (2003).
- [28] T. P. WIHLE, P. FRAUENFELDER, AND CH. SCHWAB, *Exponential convergence of the hp-DGFEM for diffusion problems*, *Comput. Math. Appl.*, 46 (2003), pp. 183–205.