DISCRETE MIRANDA–TALENTI ESTIMATES AND APPLICATIONS TO LINEAR AND NONLINEAR PDES

by

Mohan Wu

University of Pittsburgh, 2019

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Mohan Wu

It was defended on

July 25th 2019

and approved by

Prof. Michael Neilan, Associate Professor, Colloquium Chair

Prof. Ivan Yotov, Professor

Prof. Noel J. Walkington, Professor

Prof. William J. Layton, Professor

Dissertation Director: Prof. Michael Neilan, Associate Professor, Colloquium Chair

DISCRETE MIRANDA–TALENTI ESTIMATES AND APPLICATIONS TO LINEAR AND NONLINEAR PDES

Mohan Wu, PhD

University of Pittsburgh, 2019

In this thesis, we construct simple and convergent finite element methods for linear and nonlinear elliptic differential equations in non-divergence form with discontinuous coefficients. The methods are based on a discrete Miranda-Talenti estimate, which relates the H^2 semi-norm of piecewise polynomials with the L^2 norm of its Laplacian on convex domains. We develop a stability and convergence theory of the proposed methods, and back up the theory with numerical experiments. Furthermore, we construct a finite element method for the Monge-Ampère problem by using an equivalent Hamilton-Jacobi-Bellman formulation.

TABLE OF CONTENTS

LIST OF FIGURES

1.0 INTRODUCTION

In this thesis, we consider finite element methods for linear and nonlinear second order elliptic problems in non-divergence form. A prototypical (linear) example is given by

$$A(x) : D^2u(x) = g(x) \qquad \text{in } \Omega \subset \mathbb{R}^d, \tag{1.0.1}$$

accompanied with Dirichlet boundary conditions. Here, $g \in L^2(\Omega)$ is a given source term, and $A : \Omega \to \mathbb{R}^{d\times d}$ is symmetric and uniformly positive definite on a bounded domain Ω. Such problems naturally appear in stochastic optimal control in the form of the Hamilton–Jacobi–Bellman equation, and they also arise as linearizations of fully nonlinear second-order PDEs.

This thesis focuses on the case in which the coefficient matrix is not differentiable, in particular, integration-by-parts cannot be performed on (1.0.1), and weak solutions based on variational principles are not applicable. In this setting, there are some distinct theories and solution concepts concerning the well-posedness of the problem, each depending on the regularity of the matrix A. For example:

- If A is Hölder continuous and if the boundary of the domain is sufficiently smooth, then there exists a classical solution satisfying the PDE pointwise in Ω [15, Chapter 6].

- If A is uniformly continuous on Ω or if A has vanishing mean oscillation, then there exists a unique strong solution $u \in W^{2,p}(\Omega)$ to the problem, i.e., u satisfies the PDE a.e. in Ω [4, 15].
- If A is essentially bounded and it satisfies the so-called Cordes condition (cf. Definition 3), and if the domain Ω is convex, then there exists a strong solution $u \in H^2(\Omega)$ [22, 30].

In this thesis we focus on the third case which assumes the least regularity conditions on the coefficient matrix, i.e., we assume that the matrix is possibly discontinuous but satisfies the Cordes condition. The Cordes condition, a type of anisotropy condition on A, is a crucial assumption to establish the well-posedness of the elliptic problem (1.0.1); counterexamples in $d \geq 3$ show that solutions to (1.0.1) are not unique for general discontinuous A (cf. Example 1). Another key ingredient to show the well-posedness of (1.0.1) is the Miranda-Talenti estimate, which relates the H^2 semi-norm of a function with the L^2 norm of its Laplacian on convex domains.

Lemma 1 (Miranda-Talenti inequality [16, 22])**.** *Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded convex domain. Then there holds*

$$\|D^2 v\|_{L^2(\Omega)} \leq \|\Delta v\|_{L^2(\Omega)} \qquad \forall v \in H^2(\Omega) \cap H_0^1(\Omega). \tag{1.0.2}$$

The important feature of the estimate (1.0.2), and crucial to the analysis of problem (1.0.1), is that the equivalence constant is exactly one on convex domains.

A goal of this thesis is to develop finite element methods for PDEs in non-divergence form and a convergence theory by extending Lemma 1 to piecewise polynomial functions. In particular, we shall prove the following discrete Miranda-Talenti inequality. A more detailed explanation of the notation is given in subsequent chapters.

Theorem 1 (Discrete Miranda-Talenti inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a convex polytope. Let $V_h \subset H_0^1(\Omega)$ denote the kth degree Lagrange finite element space with respect to a simplicial mesh $\mathcal{T}_h$. Then for any $v_h \in V_h$, we have*

$$\|D^2 v_h\|_{L^2(\mathcal{T}_h)} \leq \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + C_\dagger \left(\sum_{f \in \mathcal{F}_h^I} h_f^{-1} \|[\![\partial v_h / \partial n_f]\!]\|_{L^2(f)}^2 \right)^{1/2}, \tag{1.0.3}$$

where the constant $C_\dagger > 0$ is independent of h and v_h.

We shall show that the estimate (1.0.3) naturally leads to simple and efficient finite element methods for linear and fully nonlinear problems in non-divergence form as well as a stability and convergence theory.

Despite its non-variational structure, a flurry of finite element methods have recently been developed for problems in non-divergence form (1.0.1). In the case that the coefficient matrix is continuous, finite element methods have been developed in [9, 11, 25]; these methods and their analysis are based on discrete Calderon-Zygmund estimates. The first Galerkin method in the case of discontinuous coefficients was done in [30], where an intricate *hp*-discontinuous Galerkin (DG) method was proposed for elliptic PDEs satisfying the Cordes condition. There, the authors bypass a discrete Miranda–Talenti estimate by adding auxiliary terms in their formulation. This method was extended to the fully nonlinear Hamilton–Jacobi–Bellman equation with continuous coefficients satisfying the Cordes condition in [31, 32]. The method was then extended to Lipschitz continuous domains with piecewise curved boundaries in [20]. A related but simpler DG method for elliptic problems in non-divergence form based on a least-squares formulation is proposed in [24]. However, it is unclear whether this method extends to fully nonlinear problems. A weak Galerkin method was presented in [34], and a mixed discretization based on stable finite element Stokes spaces is proposed in [12]. A finite element method based on

the convolution of finite differences for (1.0.1) was proposed in [27], and the stability of the method was shown via discrete Alexandrov-Bakelman-Pucci estimates. Extensions of these results to fully nonlinear problems was done in [28].

An advantage of the proposed methods is their relative simplicity; the methods can be readily implemented on standard finite element method software packages. Furthermore, in contrast to [9,11,25], the methods are provably convergent for linear problems with discontinuous coefficients satisfying the Cordes condition. Finally, as far as we are aware, the methods have the fewest number of global degrees of freedom on simplicial meshes for problems with discontinuous coefficients. On the other hand, the cost for the simplicity of the methods are restrictions on the finite element spaces and the mesh. For example, in contrast to [30–32], we require that the mesh is simplicial and does not contain hanging nodes, and that the polynomial degree does not vary between elements. Furthermore, we do not track the dependence of the polynomial degree in the stability and error estimates.

The rest of the thesis is organized as follows. In Section 1.1 we establish the notation and state some preliminary results. In Section 2.1 and 2.2 we build an enriching operator that connects the Lagrange finite element space with a cubic spline space in two and three dimensions. With this enriching operator, we prove Theorem 1 in Section 2.3. In Chapter 3 we propose a finite element method for linear PDEs in non-divergence form, prove the well-posedness of the method, and derive optimal order estimates. These results are extended to the fully nonlinear Hamilton–Jacobi–Bellman equation in Chapter 4 and numerical experiments are presented in Section 4.3. Then we discuss the Monge-Ampère problem in Chapter 5, its relation with Hamilton–Jacobi–Bellman representation, and propose a finite element method based on this formulation. Finally in Chapter 6, we make a conclusion of the thesis and discuss possible future research.

1.1 PRELIMINARIES

Let $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) be a bounded, convex polytope, and let $\mathcal{T}_h$ be a conforming and shape–regular simplicial triangulation of Ω without hanging nodes. For each element $T \in \mathcal{T}_h$, let $h_T := \operatorname{diam} T$ and $h := \max_{T\in\mathcal{T}_h} h_T$. We denote by $\mathcal{V}_h$ and $\mathcal{F}_h$ the set of vertices and the set of $(d-1)$–dimensional faces of $\mathcal{T}_h$, respectively. We write $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B$, where $\mathcal{F}_h^I$ denotes the set of interior faces and $\mathcal{F}_h^B$ denotes the set of boundary faces. Likewise, we denote the sets of interior and boundary vertices as $\mathcal{V}_h^I$ and $\mathcal{V}_h^B$, respectively. Let $\mathcal{V}_T$ and $\mathcal{M}_T$ be the set of vertices and (1-dimensional) edge midpoints, respectively, of a simplex $T \in \mathcal{T}_h$. Let $\mathcal{T}_p$ be the set of simplexes in $\mathcal{T}_h$ that share the common vertex $p \in \mathcal{V}_h$. We also denote by $\mathcal{F}_p^I$ (resp., $\mathcal{F}_p^B$) the set of interior (resp., boundary) $(d-1)$–dimensional faces in $\mathcal{F}_h^I$ (resp., $\mathcal{F}_h^B$) that share the common vertex p. The set of interior and boundary edge midpoints in $\mathcal{T}_h$ is denoted by $\mathcal{M}_h^I$ and $\mathcal{M}_h^B$, respectively, and we set $\mathcal{M}_h = \mathcal{M}_h^I \cup \mathcal{M}_h^B$.

For each face $f \in \mathcal{F}_h$, let $n_f \in \mathbb{R}^d$ denote a fixed choice of a (constant) unit normal vector to f. If $f \in \mathcal{F}_h^B$, we assume that n_f coincides with the outward unit normal of $\partial\Omega$ restricted to f. We then define the jump operator $[\![\cdot]\!]$ on f by

$$
\begin{aligned}
[\![v]\!] &:= v|_{T_{out}} - v|_{T_{in}} \quad && \text{if } f = \partial T_{out} \cap \partial T_{in} \in \mathcal{F}_h^I, \\
[\![v]\!] &:= v|_{T_{out}} && \text{if } f = \partial T_{out} \cap \partial\Omega \in \mathcal{F}_h^B,
\end{aligned}
$$

where v is a sufficiently regular scalar valued or vector-valued function. Here, the labeling is chosen so that n_f is outward pointing for T_{out} and inward pointing for T_{in}.

Remark 1. *The assumptions of $\mathcal{T}_h$ implies that there is a $c_{\mathcal{T}} > 0$, independent of h, such that*

$$
\max_{T\in\mathcal{T}_h} \operatorname{card}\{T' \in \mathcal{T}_h : T \cap T' \neq \emptyset\} \leq c_{\mathcal{T}} \quad \forall h > 0. \tag{1.1.1}
$$

Now let T, $T' \in \mathcal{T}_h$ such that they share a common vertex p. Since Ω is assumed to be a convex polytope, T and T' can be connected by simplices through faces, i.e., there exists a finite sequence $\{T_j\}_{j=0}^{M} \subset \mathcal{T}_p$ labeled such that $T_0 = T$, $T_M = T'$, and $\partial T_j \cap \partial T_{j+1} \in \mathcal{F}_h$ for all j. By (1.1.1) we have $M \leq c_{\mathcal{T}}$.

Definition 1. *We say that a node $p \in \mathcal{V}_h^B \cup \mathcal{M}_h^B$ is a* flat node *if the normal vectors of all the faces in $\mathcal{F}_p^B$ are parallel. Otherwise we say that p is a* sharp node. *We set $\mathcal{V}_h^B = \mathcal{V}_h^\flat \cup \mathcal{V}_h^\#$ and $\mathcal{M}_h^B = \mathcal{M}_h^\flat \cup \mathcal{M}_h^\#$, where $\mathcal{V}_h^\flat$ and $\mathcal{M}_h^\flat$ denote the set of flat nodes in $\mathcal{V}_h^B$ and $\mathcal{M}_h^B$, respectively, and $\mathcal{V}_h^\#$ and $\mathcal{M}_h^\#$ denote the set of sharp nodes in $\mathcal{V}_h^B$ and $\mathcal{M}_h^B$, respectively.*

Let $\mathbb{P}_k(D)$ denote the space of polynomials of degree less than or equal to k with domain D, and let

$$V_h := \{v \in C(\overline{\Omega}) \cap H_0^1(\Omega) : v|_T \in \mathbb{P}_k(T) \ for \ T \in \mathcal{T}_h\}$$

be the kth–degree Lagrange finite element space associated with $\mathcal{T}_h$.

We define the piecewise defined semi–norms $|\cdot|_{H^2(\mathcal{T}_h)}$ and $\|\cdot\|_{L^2(\mathcal{T}_h)}$ by

$$|v|^2_{H^2(\mathcal{T}_h)} = \sum_{T\in\mathcal{T}_h} |v|^2_{H^2(T)}, \qquad \|v\|^2_{L^2(\mathcal{T}_h)} = \sum_{T\in\mathcal{T}_h} \|v\|^2_{L^2(T)}.$$

And we define $u \in H^2(\Omega)$ if and only if $\|v\|^2_{H^2(\Omega)} := \|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + |v|^2_{H^2(\Omega)} < \infty$. We define $\lesssim$ such that $u \lesssim v$ means that $u \leq Cv$ for some $C > 0$ independent of the mesh size h.

Inverse-type estimates are widely used in the error analysis in this thesis. Classical inverse estimates are of the form

$$\|v\|_{H^s(T)} \lesssim h_T^{-s} \|v\|_{L^2(T)} \quad \forall v \in V_h.$$

The scaling technique is defined as following: Let T be a simplex in a regular mesh. Define $\hat{T} := \{\frac{1}{h_T}x : x \in T\}$, where h_T is the diameter of T. Then for any function $v \in \mathbb{P}_k(T)$, we define $\hat{v}(\hat{x}) = v(x)$, with $\hat{x} = \frac{1}{h_T}x$. We have an affine map F from $\hat{T}$ to T, $F(\hat{x}) = h_T\hat{x} = x$, $|DF| = h_T^d$. Thus, we can compare the norm of v in T and norm of $\hat{v}$ in $\hat{T}$.

$$\|v\|^2_{L^2(T)} = \int_T |v|^2\,dx = \int_{F(\hat{T})} |v|^2\,dx = \int_{\hat{T}} |\hat{v}|^2|DF|\,d\hat{x} = h_T^d\|\hat{v}\|^2_{L^2(\hat{T})},$$

$$\|\nabla v\|^2_{L^2(T)} = \int_T |\nabla v|^2\,dx = \int_{F(\hat{T})} |\nabla v|^2\,dx = \int_{\hat{T}} |\frac{1}{h_T}\nabla\hat{v}|^2|DF|\,d\hat{x} = h_T^{d-2}\|\nabla\hat{v}\|^2_{L^2(\hat{T})}.$$

Definition 2. *We say that u is a strong solution to a PDE with zero Dirichlet boundary condition on Ω if it has regularity*

$$u \in V := H^2(\Omega) \cap H_0^1(\Omega),$$

and satisfies the PDE almost everywhere in Ω.

Theorem 2 (Browder–Minty Theorem)**.** *Let $\langle \mathcal{M}[u], v\rangle$ be a dual pairing between V^* and V. If the operator $\mathcal{M}$ is bounded, continuous, coercive and strongly monotone, then there exist a unique $u \in V$ satisfying $\langle \mathcal{M}[u], v\rangle := 0 \quad \forall v \in V$.*

2.0 DISCRETE MIRANDA-TALENTI INEQUALITY

In this chapter, we develop a proof for the Discrete Miranda-Talenti inequality stated in Theorem 1. First we consider a H^2-conforming Clough-Tocher finite element space, and then build an enriching operator from the this space to Lagrange space V_h. We use the stability properties of this enriching operator and the continuous Miranda-Talenti inequality to give the full proof.

2.1 THE CLOUGH-TOCHER FINITE ELEMENT

To prove the discrete Miranda-Talenti inequality (1.0.3), we first introduce the (auxiliary) Clough-Tocher finite element space, a d–dimensional C^1 piecewise cubic polynomial space defined on a simplicial mesh [5, 35] . Its construction is done on a split of each simplex of the mesh obtained as follows:

- For $d = 2$, the simplex is split by connecting the vertices to the barycenter of the simplex (cf. Figure 2.1 left).
- For $d = 3$, each of the 2–dimensional faces of the simplex is split by connecting the vertices to the barycenter of the face. Then the vertices and the barycenters of the faces are connected to the barycenter of the tetrahedron (cf. Figure 2.1

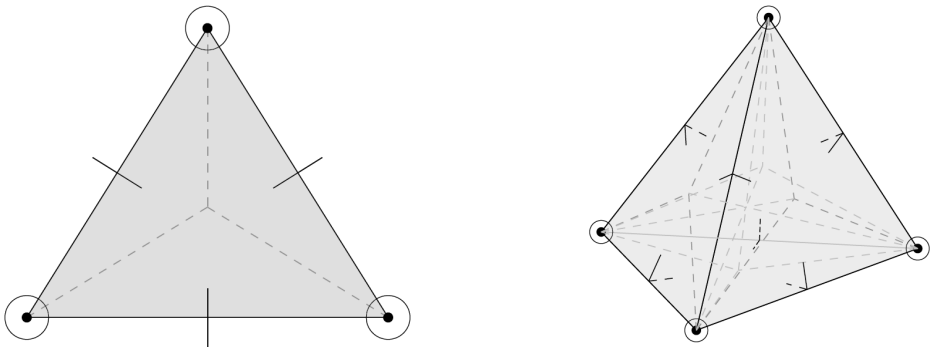

Figure 1: The two-dimensional and three-dimensional Clough-Tocher element. Solid circles indicate function evaluation, large circles indicate derivative evaluation, and straight lines indicate directional derivative evaluation.

right).

Thus we see that each simplex is split into $N_d := (d+1)!/2$ subsimplices. We denote the set of subsimplices of this split by $T_r := \{T_i\}_{i=1}^{N_d}$ and define

$$\mathbb{P}_k(T_r) := \prod_{i=1}^{N_d} \mathbb{P}_k(T_i),$$

to be the space of (local) piecewise polynomials of degree k with respect to the split. A unisolvent set of degrees of freedom which induces a globally C^1 piecewise polynomial space is given in the next proposition. We refer to [35] for a proof.

Proposition 1. *For a one-dimensional edge e, denote by $\{s_i^e\}_{i=1}^{d-1} \subset \mathbb{R}^d$ a set of unit vectors such that, together with the direction determined by the edge, they provide a basis for $\mathbb{R}^d$. Then*

1. *The dimension of $C^1(T) \cap \mathbb{P}_3(T_r)$ is $\frac{1}{2}(d+1)(d^2+d+2)$.*

2. *A function* $v_h \in C^1(T) \cap \mathbb{P}_3(T_r)$ *is uniquely determined by*

   (a) *the values* $v_h(p)$ *and* $\nabla v_h(p)$ *for all* $p \in \mathcal{V}(T)$*;*

   (b) *the values* $\nabla v_h(p) \cdot s_i^e$ *for all* $p \in \mathcal{M}_T$ *and* $i \in \{1, \ldots, d-1\}$ *such that* p *is the edge midpoint of* e*.*

Remark 2. *For the convenience of the proofs below, we set the direction vectors* $\{s_i^e\}_{i=1}^{d-1}$ *associated with a (boundary) flat edge midpoint as follows: In two and three dimensions, we take* s_1^e *to be the common normal vector given in Definition 1. In three dimensions, we take* s_2^e *to be orthogonal to both the common normal and the edge. Note that* s_2^e *is tangent to the boundary faces associated with the edge. Furthermore, we omit the superscript* e *in the notation when the context is clear.*

The degrees of freedom given in Proposition 1 induce a global piecewise cubic space

$$\tilde{V}_h := \{v_h \in H^2(\Omega) : \ v_h|_T \in C^1(T) \cap \mathbb{P}_3(T_r), \ \forall T \in \mathcal{T}_h\}.$$

A characterization of the associated space with zero Dirichlet boundary conditions with respect to the degrees of freedom is summarized in the next lemma.

Lemma 2. *A function* $v_h \in \tilde{V}_h$ *satisfies* $v_h \in \tilde{V}_{h,0} := \tilde{V}_h \cap H_0^1(\Omega)$ *if and only if (i)* $v_h(p) = 0$ *for* $p \in \mathcal{V}_h^B$*; (ii)* $\frac{\partial v_h}{\partial t}(p) = 0$ *for all* $p \in \mathcal{V}_h^\flat$ *and tangent vectors* t *with respect to faces/edges in* $\mathcal{F}_p^B$*; (iii)* $\nabla v_h(p) = 0$ *for all* $p \in \mathcal{V}_h^\# \cup \mathcal{M}_h^\#$*; and in three dimensions (iv)* $\frac{\partial v_h}{\partial s_2}(p) = 0$ *for all* $p \in \mathcal{M}_h^\flat$*.*

Proof. It is clear that if $v_h \in \tilde{V}_{h,0}$ then (i) is satisfied.

Let $p \in \mathcal{V}_h^\flat$ be a flat vertex, and denote the common normal vector at p by n. If $v_h \in \tilde{V}_{h,0}$, then $v_h = 0$ on $\mathcal{F}_p^B$, and so the tangential derivatives of v_h are zero along the boundary faces in $\mathcal{F}_p^B$; thus (ii) and (iv) hold. Note that the derivative of v_h in the direction of n is not restricted on $\mathcal{V}_h^\flat \cup \mathcal{M}_h^\flat$.

We now show that if $v_h \in \tilde{V}_{h,0}$ then $\nabla v_h = 0$ on sharp nodes. Let $p \in \mathcal{V}_h^{\#} \cup \mathcal{M}_h^{\#}$ be a sharp node, and denote by $f_1,\ f_2 \in \mathcal{F}_p^B$ two faces with nonparallel unit normal vectors $n_1,\ n_2$. Then there exist two orthogonal bases of $\mathbb{R}^d$, $\{n_1, t_{1,1}, \ldots, t_{1,d-1}\}$ and $\{n_2, t_{2,1}, \ldots, t_{2,d-1}\}$, where $\{t_{1i}\}_{i=1}^{d-1}$ are the tangential vectors of f_1, $\{t_{2,j}\}_{j=1}^{d-1}$ are the tangential vectors of f_2. Therefore, for each $i = 1, \ldots, d-1$, there exist a unique decomposition $t_{1,i} = \sum_{j=1}^{d-1} c_{i,j} t_{2,j} + c_{i,d} n_2$ ($c_{i,j} \in \mathbb{R}$). We claim that there exists $1 \le i \le d-1$ such that $c_{i,d} \neq 0$.

Suppose the claim is not true, i.e., $c_{i,d} = 0$ for all i, which implies that $\operatorname{span}\{t_{1,i}\}_{i=1}^{d-1} \subset \operatorname{span}\{t_{2,j}\}_{j=1}^{d-1}$ for all j. Since the dimensions of the (linearly independent) sets are the same, we conclude that $\operatorname{span}\{t_{1,i}\}_{i=1}^{d-1} = \operatorname{span}\{t_{2,j}\}_{j=1}^{d-1}$, and therefore $n_1 \cdot t_{2,j} = 0$ for all j. Hence, since $\{n_2, t_{2,1}, \ldots, t_{2,d-1}\}$ is a orthogonal basis, we have

$$n_1 = \sum_{j=1}^{d-1} (n_1 \cdot t_{2,j}) t_{2,j} + (n_1 \cdot n_2) n_2 = (n_1 \cdot n_2) n_2,$$

implying that n_1 and n_2 are parallel, a contradiction. Thus there exists $1 \le i \le d-1$ such that $c_{i,d} \neq 0$.

Now since $v_h = 0$ on $\mathcal{F}_p^B$, the tangential derivatives of v_h are zero along the boundary faces f_1 and f_2, i.e., $\frac{\partial v_h}{\partial t_{1,i}}(p) = 0$ and $\frac{\partial v_h}{\partial t_{2,j}}(p) = 0$ for $i, j = 1, \ldots, d-1$. We then have

$$0 = \frac{\partial v_h}{\partial t_{1,i}}(p) = \sum_{j=1}^{d-1} c_{i,j} \frac{\partial v_h}{\partial t_{2,j}}(p) + c_{i,d} \frac{\partial v_h}{\partial n_2}(p) = c_{i,d} \frac{\partial v_h}{\partial n_2}(p) \Rightarrow \frac{\partial v_h}{\partial n_2}(p) = 0.$$

Therefore, the directional derivatives of v_h at p are zero along $\{n_2, t_{2,1}, t_{2,2} ..., t_{2,d-1}\}$, the basis of $\mathbb{R}^d$, and it thus follows that (iii) is satisfied.

Finally, suppose that $v_h \in \tilde{V}_h$ vanishes at the values (i)–(iv). Since v_h, respected to an edge, is a one-dimensional cubic polynomial, we conclude from (i)–(iii) that v_h vanishes on the boundary edges. Therefore in the case $d = 2$, $v_h = 0$ on $\partial\Omega$. In

three dimensions, we use condition (iv) and the two-dimensional unisolvency result in Proposition 1 to conclude that $v_h = 0$ on $\partial\Omega$ when $d = 3$ as well. □

Remark 3. *Let $p \in \mathcal{V}_h^\sharp \cup \mathcal{M}_h^\sharp$, and let $\{t_{1,i}\}_{i=1}^{d-1}$ and $\{t_{2,j}\}_{j=1}^{d-1}$ span the tangent space of some $f_1, f_2 \in \mathcal{F}_p^B$ with nonparallel unit normal vectors. Then the proceeding proof shows that there exists an i such that $\{t_{1,i}, t_{2,1}, \ldots, t_{2,d-1}\}$ forms a basis of $\mathbb{R}^d$.*

2.2 ENRICHING OPERATOR

In this section, we construct a linear operator connecting the Lagrange finite element space to the Clough–Tocher finite element space by averaging. This is done by assigning the values specified in Proposition 1 and Lemma 2.

Let N be any (global) degree of freedom of $\tilde{V}_{h,0}$. If N is an interior degree of freedom, then we set

$$N(E_h v_h) = \frac{1}{|\mathcal{T}_N|} \sum_{T \in \mathcal{T}_N} N(v_T), \tag{2.2.1}$$

where $v_T := v_h|_T$ is the function v_h restricted to the simplex T, $\mathcal{T}_N$ is the set of simplexes in $\mathcal{T}_h$ that share the degree of freedom N, and $|\mathcal{T}_N|$ is the number of elements in $\mathcal{T}_N$.

If N corresponds to a function evaluation at a boundary vertex $p \in \mathcal{V}_h^B$, we set $N(E_h v_h) = 0$. If N is a boundary degree of freedom corresponding to the function gradient at a flat vertex $p \in \mathcal{V}_h^\flat$, let the common unit normal vector of faces in $\mathcal{F}_h^b$ be n, and set

$$N(E_h v_h) = \frac{1}{|\mathcal{T}_N|} \sum_{T \in \mathcal{T}_N} (N(v_T) \cdot n) n. \tag{2.2.2}$$

Thus, $N(E_h v_h)$ is a vector with direction n and magnitude $\frac{1}{|\mathcal{T}_N|} \sum_{T \in \mathcal{T}_N} \frac{\partial v_T}{\partial n}(p)$.

If N is a boundary degree of freedom corresponding to a function directional derivative at $p \in \mathcal{M}_h^\flat$ with direction unit vector s_i, let the common unit normal vector be n, and set

$$N(E_h v_h) = \frac{1}{|\mathcal{T}_N|} \sum_{T \in \mathcal{T}_N} (s_i \cdot n) \frac{\partial v_T}{\partial n}(p). \tag{2.2.3}$$

Finally, if N is a boundary degree of freedom corresponding to the function derivative or directional derivative at some $p \in \mathcal{V}_h^\# \cup \mathcal{M}_h^\#$, we set $N(E_h v_h) = 0$. Note that this construction and Lemma 2 show that $E_h v_h \in \tilde{V}_{h,0}$.

Lemma 3. *For $k = 2$ or 3, the map E_h satisfies the estimate*

$$|v_h - E_h v_h|^2_{H^2(\mathcal{T}_h)} \lesssim \sum_{f \in \mathcal{F}_h^I} h_f^{-1} \left\| [\![\partial v_h / \partial n_f]\!] \right\|^2_{L^2(f)} \quad \forall v_h \in V_h. \tag{2.2.4}$$

Proof. The proof of (2.2.4) in the two dimensional setting is given in [1, 13], thus it suffices to prove the result when $d = 3$.

Let $v_h \in V_h$ be arbitrary and set $w_h = v_h - E_h v_h$. Fix $T \in \mathcal{T}_h$ and set $w_T = w_h|_T$. Let $\hat{T} := \frac{1}{h_T} T$, define $\hat{x} \in \hat{T}$ by $\hat{x} = \frac{1}{h_T} x$ for $x \in T$ and define $\hat{w}$ by $\hat{w}(\hat{x}) = w(x)$ for any $x \in T$. From Proposition 1, the inclusion $\mathbb{P}_3(T) \subset \mathbb{P}_3(T_r) \cap C^1(T)$, scaling and shape regularity, and since $w_h(p) = 0$ for all $p \in \mathcal{V}_h$, we have

$$\begin{aligned}
\|v_h - E_h v_h\|^2_{L^2(T)} &= \|w\|^2_{L^2(T)} \\
&= h_T^3 \|\hat{w}\|^2_{L^2(\hat{T})} \\
&\lesssim \sum_{\hat{p} \in \mathcal{V}_{\hat{T}}} (h_{\hat{T}}^3 |\hat{w}(\hat{p})|^2 + h_{\hat{T}}^3 |\hat{\nabla} \hat{w}_{\hat{T}}(\hat{p})|^2) + \sum_{\hat{m} \in \mathcal{M}_{\hat{T}}} \sum_{i=1}^{2} h_{\hat{T}}^3 \left| \frac{\partial \hat{w}_{\hat{T}}}{\partial s_i}(\hat{m}) \right|^2 \\
&\lesssim \sum_{p \in \mathcal{V}_T} (h_T^3 |w(p)|^2 + h_T^5 |\nabla w_T(p)|^2) + \sum_{m \in \mathcal{M}_T} \sum_{i=1}^{2} h_T^5 \left| \frac{\partial w_T}{\partial s_i}(m) \right|^2 \\
&= \sum_{p \in \mathcal{V}_T} h_T^5 |\nabla w_T(p)|^2 + \sum_{m \in \mathcal{M}_T} \sum_{i=1}^{2} h_T^5 \left| \frac{\partial w_T}{\partial s_i}(m) \right|^2 .
\end{aligned} \tag{2.2.5}$$

By (2.2.1), for an interior point p (i.e., a point that is not on $\partial\Omega$), we have

$$\begin{aligned} |\nabla w_T(p)|^2 &= \Big(\frac{1}{|\mathcal{T}_p|}\sum_{T'\in\mathcal{T}_p}|\nabla v_T(p)-\nabla v_{T'}(p)|\Big)^2 \\ &\lesssim \sum_{T'\in\mathcal{T}_p}|\nabla v_T(p)-\nabla v_{T'}(p)|^2. \end{aligned} \tag{2.2.6}$$

For any $T'\in\mathcal{T}_p$, there exist a finite sequence of simplices $\{T_j\}_{j=0}^M\subset\mathcal{T}_p$ labeled such that $T_0=T$, $T_M=T'$, and $\partial T_j\cap\partial T_{j+1}\in\mathcal{F}_h^I$. We emphasize that M is is bounded uniformly in h by the shape regularity of $\mathcal{T}_h$ (cf. Remark 1). Hence, by an inverse estimate, and since v_h is continuous across the faces,

$$\begin{aligned} |\nabla v_T(p)-\nabla v_{T'}(p)| &\le \sum_{j=0}^{M-1}|\nabla v_{T_j}(p)-\nabla v_{T_{j+1}}(p)| \\ &\le \sum_{f\in\mathcal{F}_p^I}\big\|[\![\nabla v_h]\!]\big\|_{L^\infty(f)} \\ &\lesssim \sum_{f\in\mathcal{F}_p^I}h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}. \end{aligned} \tag{2.2.7}$$

Applying (2.2.7) to (2.2.6), we find that

$$|\nabla w(p)|^2\lesssim\sum_{f\in\mathcal{F}_p^I}h_f^{-2}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2. \tag{2.2.8}$$

Using similar arguments, we have for any interior midpoint m and $i\in\{1,2\}$,

$$\Big|\frac{\partial v_T}{\partial s_i}(m)-\frac{\partial v_{T'}}{\partial s_i}(m)\Big|\le\sum_{f\in\mathcal{F}_m^I}\big\|[\![\nabla v_h]\!]\big\|_{L^\infty(f)}\lesssim\sum_{f\in\mathcal{F}_m^I}h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)},$$

and therefore,

$$\Big|\frac{\partial w_T}{\partial s_i}(m)\Big|^2=\Big|\frac{1}{|\mathcal{T}_m|}\sum_{T'\in\mathcal{T}_m}|\frac{\partial v_T}{\partial s_i}(m)-\frac{\partial v_{T'}}{\partial s_i}(m)|\Big|^2\lesssim\sum_{f\in\mathcal{F}_m^I}h_f^{-2}\,\|[\![\partial v/\partial n_f]\!]\|_{L^2(f)}^2\,. \tag{2.2.9}$$

At a sharp vertex p, $\nabla E_h v_h$ vanishes, and thus

$$|\nabla w_T(p)|^2 = |\nabla v_T(p)|^2.$$

Since p is a sharp vertex, there exist two simplexes T', $T'' \in \mathcal{T}_p$, boundary faces $f_1 \subset \partial T' \cap \partial\Omega$, $f_2 \subset \partial T'' \cap \partial\Omega$, and f_1, f_2 do not have a common normal vector. Hence, by Remark 3, there exist a tangential vector $t_{1,i}$ of f_1 and two tangential vectors $\{t_{2,1}, t_{2,2}\}$ of f_2 such that together, the three vectors form a basis of $\mathbb{R}^3$.

By connecting T through a sequence of simplex in $\mathcal{T}_p$ to T', we have

$$\begin{aligned}\left|\frac{\partial v_T}{\partial t_{1,i}}(p)\right|^2 &\lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|[\![\partial v_h/\partial t_{1,i}]\!]\right\|_{L^2(f)}^2 + h_{f_1}^{-2}\left\|\partial v_{T'}/\partial t_{1,i}\right\|_{L^2(f_1)}^2 \\ &\lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|[\![\partial v_h/\partial n_f]\!]\right\|_{L^2(f)}^2\end{aligned} \tag{2.2.10}$$

because the tangential derivatives of v vanish on $\partial\Omega$.

Similarly, by connecting T through a sequence of simplex in $\mathcal{T}_p$ to T'', we have

$$\left|\frac{\partial v_T}{\partial t_{2,j}}(p)\right|^2 \lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|[\![\partial v_h/\partial n_f]\!]\right\|_{L^2(f)}^2 \quad \text{for } j = 1, 2,$$

and therefore

$$|\nabla w_T(p)|^2 \lesssim \left|\frac{\partial v_T}{\partial t_{1,i}}(p)\right|^2 + \sum_{j=1}^{2}\left|\frac{\partial v_T}{\partial t_{2,j}}(p)\right|^2 \lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|\partial v_h/\partial n_f\right\|_{L^2(f)}^2 \qquad \forall p \in \mathcal{V}_T \cap \mathcal{V}_h^\sharp. \tag{2.2.11}$$

Next, for a boundary flat vertex $p \in \mathcal{V}_h^\flat$ with common unit normal vector n, we first write

$$|\nabla w_T(p)|^2 \lesssim \left|\frac{\partial v_T}{\partial n}(p)\, n - \nabla E_h v_h(p)\right|^2 + \left|\nabla v_T(p) - \frac{\partial v_T}{\partial n}(p)\, n\right|^2. \tag{2.2.12}$$

By (2.2.2) and by applying similar steps in (2.2.6), (2.2.7), and (2.2.8), we have

$$
\begin{aligned}
|\frac{\partial v_T}{\partial n}(p)\, n - \nabla E_h v_h|^2 &\lesssim \sum_{T'\in\mathcal{T}_p} \Big|\big(\frac{\partial v_T}{\partial n}(p) - \nabla v_{T'}(p)\cdot n\big)n\Big|^2 \\
&= \sum_{T'\in\mathcal{T}_p} \Big|\frac{\partial v_T}{\partial n}(p) - \frac{\partial v_{T'}}{\partial n}(p)\Big|^2 \\
&\lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\, \|[\![\partial v_h/\partial n_f]\!]\|^2_{L^2(f)}\,.
\end{aligned}
\tag{2.2.13}
$$

Since p is a flat vertex, there exist a simplex T' with a boundary face $f_3 \in \mathcal{F}_p^B$. Let $\{t_{3,1}, t_{3,2}\}$ denote an orthonormal basis of f_3. Then by marching to the boundary (as in (2.2.10)), we have

$$
\begin{aligned}
|\nabla v_T(p) - \frac{\partial v_T}{\partial n}(p)\, n|^2 &= |\sum_{i=1}^{2} \frac{\partial v_T}{\partial t_{3,i}}(p)\, t_{3i}|^2 \\
&\lesssim \sum_{i=1}^{2} |\frac{\partial v_T}{\partial t_{3,i}}(p)|^2 \lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\, \|[\![\partial v/\partial n_f]\!]\|^2_{L^2(f)}\,.
\end{aligned}
\tag{2.2.14}
$$

Hence, by (2.2.12)–(2.2.14), we have

$$
|\nabla w(p)|^2 \lesssim \sum_{f\in\mathcal{F}_p^I} h_f^{-2}\, \|[\![\partial v/\partial n_f]\!]\|^2_{L^2(f)} \quad \forall p \in \mathcal{V}_T \cap \mathcal{V}_h^\flat. \tag{2.2.15}
$$

For a boundary flat midpoint m, by (2.2.3) and (2.2.10), we have

$$
\begin{aligned}
\Big|\frac{\partial w}{\partial s_1}(m)\Big|^2 &\lesssim \Big|(s_1\cdot n)\frac{\partial v_T}{\partial n}(m) - \frac{\partial E_h v_h}{\partial s_1}(m)\Big|^2 + \Big|\frac{\partial v_T}{\partial s_1}(m) - (s_1\cdot n)\frac{\partial v_T}{\partial n}(m)\Big|^2 \\
&\lesssim \sum_{T'\in\mathcal{T}_m} \Big|\frac{\partial v_T}{\partial n}(m) - \frac{\partial v_{T'}}{\partial n}(m)\Big|^2 \\
&\lesssim \sum_{f\in\mathcal{F}_m^I} h_f^{1-d}\, \|[\![\partial v/\partial n_f]\!]\|^2_{L^2(f)}\,.
\end{aligned}
\tag{2.2.16}
$$

Likewise, we have that

$$\left|\frac{\partial w}{\partial s_2}(m)\right|^2 \lesssim \left|(s_2\cdot n)\frac{\partial v_T}{\partial n}(m) - \frac{\partial E_h v_h}{\partial s_2}(m)\right|^2 + \left|\frac{\partial v_T}{\partial s_2}(m) - (s_2\cdot n)\frac{\partial v_T}{\partial n}(m)\right|^2$$
$$= \left|\frac{\partial v_T}{\partial s_2}(m)\right|^2 \lesssim \sum_{f\in\mathcal{F}_m^I} h_f^{-2}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2. \tag{2.2.17}$$

Combining (2.2.5), (2.2.8), (2.2.9), (2.2.11), (2.2.15), (2.2.16), and (2.2.17) yields

$$\|v_h - E_h v_h\|_{L^2(T)}^2 \lesssim h_T^5\Big(\sum_{p\in\mathcal{V}_T}\sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2 + \sum_{m\in\mathcal{M}_T}\sum_{f\in\mathcal{F}_m^I} h_f^{-2}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2\Big).$$

Finally, by an inverse estimate and the shape regularity of $\mathcal{T}_h$, we obtain

$$\begin{aligned}|v_h - E_h v_h|_{H^2(\mathcal{T}_h)}^2 &\lesssim \sum_{T\in\mathcal{T}_h} h_T^{-4}\|v_h - E_h v_h\|_{L^2(T)}^2\\ &\lesssim \sum_{T\in\mathcal{T}_h} h_T\Big(\sum_{p\in\mathcal{V}_T}\sum_{f\in\mathcal{F}_p^I} h_f^{-2}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2\\ &\qquad + \sum_{m\in\mathcal{M}_T}\sum_{f\in\mathcal{F}_m^I} h_f^{-2}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2\Big)\\ &\lesssim \sum_{f\in\mathcal{F}_h^I} h_f^{-1}\left\|[\![\partial v/\partial n_f]\!]\right\|_{L^2(f)}^2.\end{aligned}$$

□

Remark 4. *In two dimensions, there exists a family of C^1 Clough-Tocher spaces of degree greater than or equal to three [7]. As a result, the estimate (2.2.4) can be generalized to arbitrary $k \geq 2$ [13]. However, as far as we are aware, degrees of freedom for higher-order Clough-Tocher spaces in three dimensions are not found in the literature; see [33] for partial results. As a result, the estimate (2.2.4) is restricted to $2 \leq k \leq 3$ if $d = 3$.*

Remark 5. *A recent paper by Brenner and Sung [2] shows that the restriction of polynomial degree for $d = 3$ is not necessary. Instead of the Clough-Tocher space, they map V_h to a H^2 conforming virtual element space by a similar enriching operator E_h. Thus we still have the same conclusion without the restriction of polynomial degree for $d = 3$.*

Remark 6. *An operator that maps piecewise polynomials to $H^2(\Omega) \cap H_0^1(\Omega)$ conforming functions has been recently been constructed in [29]. There, the mesh is allowed to have hanging nodes, and the dependence of the polynomial degree is explicitly stated in the estimate. On the other hand, the operator is constructed in a global fashion, and as such, it seems that the mesh must be quasi-uniform in order to get an estimate analogous to (2.2.4) by directly using [29, Theorem 4].*

2.3 PROOF OF DISCRETE MIRANDA-TALENTI INEQUALITY

With the result of Lemma 3, we are able to prove Theorem 1.

Proof. For $v_h \in V_h$, we have $E_h v_h \in H^2(\Omega) \cap H_0^1(\Omega)$, and therefore, by the Miranda-Talenti and triangle inequalities,

$$\begin{aligned} |v_h|_{H^2(\mathcal{T}_h)} &\leq |E_h v_h|_{H^2(\mathcal{T}_h)} + |v_h - E_h v_h|_{H^2(\mathcal{T}_h)} \\ &\leq \|\Delta E_h v_h\|_{L^2(\mathcal{T}_h)} + |v_h - E_h v_h|_{H^2(\mathcal{T}_h)} \\ &\leq \|\Delta(v_h - E_h v_h)\|_{L^2(\mathcal{T}_h)} + \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + |v_h - E_h v_h|_{H^2(\mathcal{T}_h)}. \end{aligned} \tag{2.3.1}$$

We then use the identity $\|\Delta v_h\|_{L^2(\mathcal{T}_h)} \leq \sqrt{d}|v_h|_{H^2(\mathcal{T}_h)}$ and Lemma 3 to get

$$\|D^2 v_h\|_{L^2(\mathcal{T}_h)} \leq \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + (1 + \sqrt{d})|v_h - E_h v_h|_{H^2(\mathcal{T}_h)}$$

$$\leq \|\Delta v_h\|_{L^2(\mathcal{T}_h)} + C(1+\sqrt{d})\Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\|[\![\partial v_h/\partial n_f]\!]\|_{L^2(f)}^2\Big)^{1/2}.$$

The proof is complete with $C_\dagger = C(1+\sqrt{d})$. □

Corollary 1. *There holds, for all $\tau \in (0,1)$ and $v_h \in V_h$,*

$$\|\Delta v_h\|_{L^2(\mathcal{T}_h)}^2 \geq (1-\tau)\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 - \frac{C_\dagger^2}{\tau}\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2.$$

Proof. Applying the Cauchy–Schwarz inequality to Theorem 1 yields

$$\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 \leq \big(1+\rho\big)\|\Delta v_h\|_{L^2(\mathcal{T}_h)}^2 + C_\dagger^2\big(1+\frac{1}{\rho}\big)\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2$$

for any $\rho > 0$. Letting $\tau = \rho/(1+\rho) \in (0,1)$ and rearranging terms, we have

$$\begin{aligned}\|\Delta v_h\|_{L^2(\mathcal{T}_h)}^2 &\geq (1-\tau)\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 - C_\dagger^2\frac{(1-\tau)}{\tau}\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2\\ &\geq (1-\tau)\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 - \frac{C_\dagger^2}{\tau}\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2.\end{aligned}$$

□

3.0 APPLICATIONS TO LINEAR PROBLEMS IN NONDIVERGENCE FORM

In this chapter, we develop a numerical method for the linear PDE problem in non-divergence form. We show that the method is well-defined and give some numerical examples.

3.1 ANALYSIS

In this section, motivated by the discrete Miranda-Talenti estimate, we construct simple convergent finite element methods to approximate strong solutions for elliptic problems in non-divergence form:

$$\mathcal{L}u := A : D^2u + \boldsymbol{b} \cdot \nabla u - cu = g \qquad \text{in } \Omega, \tag{3.1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{3.1.1b}$$

Here, $A : B := \sum_{i,j=1}^{d} A_{i,j}B_{i,j}$ denotes the Frobenius inner product of two matrices. We say that u is a strong solution to (3.1.1) if it has regularity

$$u \in V := H^2(\Omega) \cap H_0^1(\Omega),$$

and satisfies (3.1.1a) almost everywhere in Ω.

To ensure the well-posedness of problem (3.1.1) we assume that $g \in L^2(\Omega)$, the coefficients satisfy $A \in [L^\infty(\Omega)]^{d\times d}$, $\boldsymbol{b} \in [L^\infty(\Omega)]^d$, $c \in L^\infty(\Omega)$ with $c \geq 0$, and that A is uniformly positive definite in Ω, i.e., there exists $\underline{\nu}, \bar{\nu} > 0$ such that

$$\underline{\nu}|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^t A(x)\boldsymbol{\xi} \leq \bar{\nu}|\boldsymbol{\xi}|^2 \quad \text{a.e. } x \in \Omega,$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$. Here, $|\boldsymbol{\xi}|$ denotes the Euclidean distance of $\boldsymbol{\xi}$ from the origin. More importantly, we assume that the coefficients satisfy the *Cordes Condition*.

Definition 3. *The coefficients satisfy the* Cordes Condition *if*

(i) whenever $c \not\equiv 0$ or $\boldsymbol{b} \not\equiv 0$, there exists $\lambda > 0$ and $\epsilon \in (0,1)$ such that

$$\frac{|A|^2 + |\boldsymbol{b}|^2/2\lambda + (c/\lambda)^2}{(\operatorname{tr} A + c/\lambda)^2} \leq \frac{1}{d+\epsilon} \quad \text{a.e. } x \in \Omega. \tag{3.1.2a}$$

Here, $|A| = \sqrt{A : A}$, and $\operatorname{tr} A = \sum_{i=1}^d A_{ii}$ is the trace of A.

(ii) whenever $c \equiv 0$ and $\boldsymbol{b} \equiv 0$, there exists $\epsilon \in (0,1)$ such that

$$\frac{|A|^2}{(\operatorname{tr} A)^2} \leq \frac{1}{d-1+\epsilon} \quad \text{a.e. } x \in \Omega. \tag{3.1.2b}$$

Remark 7. *In two dimensions, and in the case $c \equiv 0$ and $\boldsymbol{b} \equiv 0$, uniform ellipticity of A implies the Cordes condition with $\epsilon = 2\underline{\nu}/(\underline{\nu} + \bar{\nu})$ [31, Example 2].*

Proof. In the case $d = 2$, let the two eigenvalues of A be $\underline{\lambda} \leq \bar{\lambda}$, we have $|A|^2 = \operatorname{tr} AA^T = \underline{\lambda}^2 + \bar{\lambda}^2$ and $\operatorname{tr} A = \underline{\lambda} + \bar{\lambda}$. Thus, we have

$$\frac{|A|^2}{(\operatorname{tr} A)^2} = \frac{\underline{\lambda}^2 + \bar{\lambda}^2}{\underline{\lambda}^2 + 2\underline{\lambda}\bar{\lambda} + \bar{\lambda}^2} = \frac{1}{2 - 1 + \frac{2\underline{\lambda}\bar{\lambda}}{\underline{\lambda}^2 + \bar{\lambda}^2}} = \frac{1}{2 - 1 + \frac{2}{\underline{\lambda}/\bar{\lambda} + \bar{\lambda}/\underline{\lambda}}}$$

Since A is uniform ellipticity, we have $\underline{\nu} \leq \underline{\lambda} \leq \bar{\lambda} \leq \bar{\nu}$. Therefore, we get

$$\frac{|A|^2}{(\operatorname{tr} A)^2} \leq \frac{1}{2 - 1 + \frac{2}{1 + \bar{\nu}/\underline{\nu}}} = \frac{1}{2 - 1 + 2\underline{\nu}/(\underline{\nu} + \bar{\nu})} = \frac{1}{d - 1 + \epsilon}.$$

□

Remark 8. *Uniformly ellipticity of $A \in [L^\infty(\Omega)]^{d\times d}$ is not sufficient to ensure that there exists a unique strong solution to problem (3.1.1), at least in three dimensions. The following classical example from [15] illustrates this feature.*

Example 1. *Let $d = 3$, $\Omega = B_1(0)$ be the unit ball, $c \equiv 0$, $\boldsymbol{b} \equiv 0$, and*

$$A(x) = I_3 + \Big(\frac{1+\alpha}{1-\alpha}\Big)\frac{xx^t}{|x|^2},$$

where $1/2 < \alpha < 1$ and I_3 denotes the 3×3 identity matrix. Clearly A is essentially bounded and uniformly positive definite with $\underline{\nu} = 1$ and $\bar{\nu} = 2/(1-\alpha)$.

The function $u(x) = |x|^\alpha - 1$ satisfies $u \in V$,

$$D^2u(x) = \alpha(\alpha-2)|x|^{\alpha-4}xx^t + \alpha|x|^{\alpha-2}I_3,$$

and since $I_3 : (xx^t) = (xx^t) : (xx^t)/|x|^2 = |x|^2$,

$$A(x) : D^2u(x) = \alpha(\alpha-2)|x|^{\alpha-2}\Big(1+\frac{1+\alpha}{1-\alpha}\Big) + \alpha|x|^{\alpha-2}\Big(3+\frac{1+\alpha}{1-\alpha}\Big) = 0.$$

Therefore both $u = |x|^\alpha - 1$ and the zero function are strong solutions to (3.1.1) with $g \equiv 0$.

Note that

$$|A|^2 = 3 + 2\Big(\frac{1+\alpha}{1-\alpha}\Big) + \Big(\frac{1+\alpha}{1-\alpha}\Big)^2 = \frac{2(\alpha^2-2\alpha+3)}{(1-\alpha)^2},$$
$$(\operatorname{tr} A)^2 = \Big(3+\frac{1+\alpha}{1-\alpha}\Big)^2 = \frac{4(\alpha^2-4\alpha+4)}{(1-\alpha)^2},$$

and therefore

$$|A|^2 - \frac{1}{2}(\operatorname{tr} A)^2 = \frac{2(2\alpha-1)}{(1-\alpha)^2} > 0.$$

Thus, the coefficients do not satisfy the Cordes condition.

Define the function $\gamma \in L^\infty(\Omega)$ by

$$\gamma := \frac{\operatorname{tr} A + c/\lambda}{|A|^2 + |\boldsymbol{b}|^2/2\lambda + (c/\lambda)^2}. \tag{3.1.3}$$

Since A is positive definite and c is non-negative, we clearly see that $\gamma > 0$. In particular, if the Cordes condition (3.1.2a) is satisfied, then

$$\gamma \geq \frac{d+\epsilon}{\operatorname{tr} A + c/\lambda} \geq \frac{d+\epsilon}{d\bar{\nu} + \|c\|_{L^\infty(\Omega)}/\lambda} =: \gamma_0.$$

To state the well-posedness of problem (3.1.1), we define the operator $\mathcal{L}_\lambda : V \to L^2(\Omega)$ by

$$\mathcal{L}_\lambda v := \Delta v - \lambda v. \tag{3.1.4}$$

where in the case that $\boldsymbol{b} \equiv 0$ and $c \equiv 0$, we set $\lambda = 0$ in (3.1.4). Note that, since λ is nonnegative and Ω is convex, the mapping $\mathcal{L}_\lambda : V \to L^2(\Omega)$ is surjective. Moreover, since $\gamma \geq \gamma_0 > 0$ a.e. in Ω. A simple argument will show that $u \in V$ satisfies (3.1.1a) if and only if $\gamma\mathcal{L}u = \gamma g$ a.e. in Ω.

Proof. If $u \in V$ satisfies (3.1.1a), then $\mathcal{L}u - g = 0$ a.e. in Ω. Thus $\gamma(\mathcal{L}u - g) = 0 \Rightarrow \gamma\mathcal{L}u = \gamma g$ a.e. in Ω.
Conversely, if $\gamma\mathcal{L}u = \gamma g$ a.e. in Ω, we have $\gamma(\mathcal{L}u - g) = 0$ a.e. in Ω. Since $\gamma \geq \gamma_0 > 0$, we have $\mathcal{L}u - g = 0/\gamma = 0$ a.e. in Ω, therefore u satisfies (3.1.1a). □

Thus, these two observations show that $u \in V$ is a strong solution to (3.1.1) if and only if

$$B(u,v) := \int_\Omega (\gamma\mathcal{L}u)(\mathcal{L}_\lambda v)\,dx = \int_\Omega \gamma g(\mathcal{L}_\lambda v)\,dx \qquad \forall v \in V. \tag{3.1.5}$$

Lemma 4. *Under the given assumptions, there holds the following inequality a.e. in* Ω*:*

$$|\gamma\mathcal{L}w - \mathcal{L}_\lambda w| \leq \sqrt{1-\epsilon}\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2} \quad \forall w \in V. \tag{3.1.6}$$

Proof. Suppose that $\boldsymbol{b} \not\equiv 0$ or $c \not\equiv 0$.

Applying the definitions of the operators and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |\gamma\mathcal{L}w - \mathcal{L}_\lambda w| &\leq |\gamma A - I_d||D^2 w| + |\gamma||\boldsymbol{b}||\nabla w| + |\lambda - c\gamma||w| \\ &\leq \sqrt{M}\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2}, \end{aligned} \tag{3.1.7}$$

with

$$M := |\gamma A - I_d|^2 + |\gamma|^2 \frac{|\boldsymbol{b}|^2}{2\lambda} + \frac{|\lambda - c\gamma|^2}{\lambda^2}.$$

Expanding this expression out and using the definition of γ and the Cordes condition (3.1.2a), we have

$$\begin{aligned} M &= d + 1 - 2\gamma(\operatorname{tr} A + \frac{c}{\lambda}) + |\gamma|^2(|A|^2 + \frac{|\boldsymbol{b}|^2}{2\lambda} + \frac{|c|^2}{\lambda^2}) \\ &= d + 1 - \gamma\big(\operatorname{tr} A + c/\lambda\big) \\ &= d + 1 - \frac{\big(\operatorname{tr} A + c/\lambda\big)^2}{|A|^2 + |\boldsymbol{b}|^2/(2\lambda) + (c/\lambda)^2} \\ &\leq 1 - \epsilon. \end{aligned}$$

Combining this inequality with (3.1.7) yields (3.1.6).

Likewise, for the special case $\boldsymbol{b} \equiv 0$, $c \equiv 0$ and $\lambda = 0$, we have by (3.1.2b),

$$\begin{aligned}
|\gamma \mathcal{L} w - \mathcal{L}_\lambda w| &\leq |\gamma A - I_d||D^2 w| \\
&= \sqrt{d - 2\gamma \operatorname{tr} A + |\gamma|^2 |A|^2}|D^2 w| \\
&= \sqrt{d - \gamma \operatorname{tr} A}|D^2 w| \\
&\leq \sqrt{1-\epsilon}|D^2 w|.
\end{aligned}$$

□

Lemma 5. *If Ω is convex, then there holds*

$$\|\mathcal{L}_\lambda v\|^2_{L^2(\Omega)} \geq \int_\Omega \left(|D^2 v|^2 + 2\lambda |\nabla v|^2 + \lambda^2 |v|^2\right) dx \quad \forall v \in V.$$

Proof. Integration by parts gives

$$\|\mathcal{L}_\lambda v\|^2_{L^2(\Omega)} = \int_\Omega \left(|\Delta v|^2 + 2\lambda |\nabla v|^2 + \lambda^2 |v|^2\right) dx.$$

An application of the Miranda-Talenti estimate now yields the result. □

Theorem 3. *There exists a unique strong solution to (3.1.1) provided the Cordes condition is satisfied.*

Proof. A proof of this result is given in [23] (also see [30, 31]). However, we give it here for completeness and to motivate the numerical analysis of the method given in the next section.

Let $v \in V$, and write

$$B(v, v) = \int_\Omega |\mathcal{L}_\lambda v|^2 + \int_\Omega \left(\gamma \mathcal{L} v - \mathcal{L}_\lambda v\right)(\mathcal{L}_\lambda v)\, dx.$$

Applying Lemmas 4–5 yields

$$B(v, v) \geq \left(1 - \sqrt{1-\epsilon}\right) \left\|\mathcal{L}_\lambda v\right\|^2_{L^2(\Omega)},$$

and therefore $B(\cdot,\cdot)$ is coercive on V. Since $v \to \int_\Omega \gamma g \mathcal{L}_\lambda v \, dx$ is clearly a bounded linear form on V, with $\left| \int_\Omega \gamma g \mathcal{L}_\lambda v \, dx \right| \leq \|\gamma\|_{L^\infty(\Omega)} \|g\|_{L^2(\Omega)} \|\mathcal{L}_\lambda v\|_{L^2(\Omega)}$, the Lax–Milgram theorem shows that there exists a unique $u \in V$ satisfying $B(u,v) = \int_\Omega \gamma g \mathcal{L}_\lambda v \, dx$ for all $v \in V$. Equivalently, there exists a unique solution $u \in V$ satisfying (3.1.1). □

3.2 FINITE ELEMENT METHOD

Based on the discrete Miranda–Talenti estimate and the arguments given in Theorem 3, we propose the following finite element scheme to approximate the solution to (3.1.1): Find $u_h \in V_h$ such that

$$B_h(u_h, v_h) = \sum_{T \in \mathcal{T}_h} \int_T \gamma g \mathcal{L}_\lambda v_h \, dx \qquad \forall v_h \in V_h, \tag{3.2.1}$$

where the bilinear form $B(\cdot,\cdot)$ is given by

$$B_h(w,v) = \sum_{T \in \mathcal{T}_h} \int_T (\gamma \mathcal{L} w)(\mathcal{L}_\lambda v_h) \, dx + \sigma \sum_{f \in \mathcal{F}_h^I} h_f^{-1} \int_f [\![\partial w/\partial n_f]\!][\![\partial v/\partial n_f]\!] \, ds,$$

$\sigma > 0$ is a positive penalization parameter, and we recall that V_h is the Lagrange finite element space of degree k.

We immediately notice that the scheme (3.2.1) is consistent. Indeed, if $u \in V$ is a strong solution to (3.1.1) then $\gamma \mathcal{L} u = \gamma g$ a.e. in Ω and $[\![\partial u/\partial n_f]\!] = 0$ on $\mathcal{F}_h^I$; thus,

$$B_h(u, v_h) = \sum_{T \in \mathcal{T}_h} \int_T \gamma g \mathcal{L}_\lambda v_h \, dx \qquad \forall v_h \in V_h.$$

To analyze method (3.2.1) and to show that there exists a unique solution, we introduce the following norm on $V + V_h$:

$$\|v\|_h^2 := \|D^2 v\|_{L^2(\mathcal{T}_h)}^2 + 2\lambda\|\nabla v\|_{L^2(\Omega)}^2 + \lambda^2\|v\|_{L^2(\Omega)}^2 + \sum_{f\in\mathcal{F}_h^I} h_f^{-1}\,\|[\![\partial v/\partial n_f]\!]\|_{L^2(f)}^2\,. \quad (3.2.2)$$

Note that if $\|v\|_h = 0$ with $v \in V + V_h$, then the Hessian of v vanishes on each element $T \in \mathcal{T}_h$, and $[\![\partial v/\partial n_f]\!] = 0$ on all $f \in \mathcal{F}_h^I$. This implies that v is a linear polynomial on Ω. Since v vanishes on $\partial\Omega$, then we conclude that $v \equiv 0$. Thus, $\|\cdot\|_h$ is indeed a norm on $V + V_h$ for $\lambda \geq 0$.

The next lemma, a discrete analogue of Lemma 5, relates the discrete norm $\|\cdot\|_h$ with the operator $\mathcal{L}_\lambda$ on V_h.

Lemma 6. *There exists a constant $C_1 > 0$, depending on k and the shape-regularity of the mesh such that, for all $\tau \in (0,1)$,*

$$\|\mathcal{L}_\lambda v_h\|_{L^2(\mathcal{T}_h)}^2 \geq (1-\tau)\|v_h\|_h^2 - C_1\tau^{-1}\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2 \quad \forall v_h \in V_h.$$

Proof. Using the definition of $\mathcal{L}_\lambda$ and integrating by parts, we have

$$\begin{aligned}\|\mathcal{L}_\lambda v_h\|_{L^2(\mathcal{T}_h)}^2 &= \sum_{T\in\mathcal{T}_h}\int_T \Big(|\Delta v_h|^2 + \lambda^2|v_h|^2 - 2\lambda v_h\Delta v_h\Big)\,dx \\ &= \|\Delta v_h\|_{L^2(\mathcal{T}_h)}^2 + 2\lambda\|\nabla v_h\|_{L^2(\Omega)}^2 + \lambda^2\|v_h\|_{L^2(\Omega)}^2 \\ &\quad - 2\lambda\sum_{f\in\mathcal{F}_h^I}\int_f v_h[\![\partial v_h/\partial n_f]\!]\,ds.\end{aligned}$$

Therefore by Corollary 1,

$$\begin{aligned}\|\mathcal{L}_\lambda v_h\|_{L^2(\mathcal{T}_h)}^2 \geq{}& (1-\tau)\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 + 2\lambda\|\nabla v_h\|_{L^2(\Omega)}^2 + \lambda^2\|v_h\|_{L^2(\Omega)}^2 \\ &- \frac{C_\dagger^2}{\tau}\sum_{f\in\mathcal{F}_h^I}\big\|[\![\partial v_h/\partial n_f]\!]\big\|_{L^2(f)}^2 - 2\lambda\sum_{f\in\mathcal{F}_h^I}\int_f v_h[\![\partial v_h/\partial n_f]\!]\,ds \end{aligned} \quad (3.2.3)$$

for all $\tau \in (0,1)$.

By the Cauchy-Schwarz inequality and scaling, we find that

$$2\lambda \sum_{f\in\mathcal{F}_h^I} \int_f v_h [\![\partial v_h/\partial n_f]\!]\, ds \le C\lambda \|v_h\|_{L^2(\Omega)} \Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1} \big\| [\![\partial v_h/\partial n_f]\!] \big\|_{L^2(f)}^2 \Big)^{1/2}$$
$$\le \lambda^2 \tau \|v_h\|_{L^2(\Omega)}^2 + \frac{C^2}{4\tau} \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \big\| [\![\partial v_h/\partial n_f]\!] \big\|_{L^2(f)}^2.$$

Applying this estimate to (3.2.3) and applying the definition of $\|\cdot\|_h$ yields

$$\begin{aligned}
\|\mathcal{L}_\lambda v_h\|_{L^2(\mathcal{T}_h)}^2 &\ge (1-\tau)\|D^2 v_h\|_{L^2(\mathcal{T}_h)}^2 + 2\lambda \|\nabla v_h\|_{L^2(\Omega)}^2 + \lambda^2(1-\tau)\|v_h\|_{L^2(\Omega)}^2 \\
&\quad - \Big(\frac{C_\dagger^2}{\tau} + \frac{C^2}{4\tau}\Big) \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \big\| [\![\partial v_h/\partial n_f]\!] \big\|_{L^2(f)}^2 \\
&\ge (1-\tau)\|v_h\|_h^2 - \Big(1 - \tau + \frac{C_\dagger^2}{\tau} + \frac{C^2}{4\tau}\Big) \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \big\| [\![\partial v_h/\partial n_f]\!] \big\|_{L^2(f)}^2 \\
&\ge (1-\tau)\|v_h\|_h^2 - \Big(\tau^{-1} + \frac{C_\dagger^2}{\tau} + \frac{C^2}{4\tau}\Big) \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \big\| [\![\partial v_h/\partial n_f]\!] \big\|_{L^2(f)}^2.
\end{aligned}$$

Setting $C_1 = 1 + C_\dagger^2 + C^2/4$ yields the result. □

Lemma 7. *For any $\alpha \in (0,1)$, there exists $\sigma_\alpha > 0$, independent of h, such that if $\sigma \ge \sigma_\alpha$, there holds*

$$B_h(v_h, v_h) \ge \alpha\big(1 - \sqrt{1-\epsilon}\big)\|v_h\|_h^2.$$

Consequently, there exists a unique solution $u_h \in V_h$ to (3.2.1) provided σ is sufficiently large.

Proof. We add and subtract $\mathcal{L}_\lambda v_h$ and apply Lemma 4 and the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
B_h(v_h,v_h) =& \sum_{T\in\mathcal{T}_h}\int_T \big(\gamma\mathcal{L}v_h - \mathcal{L}_\lambda v_h\big)(\mathcal{L}_\lambda v_h)\,dx + \|\mathcal{L}_\lambda v_h\|^2_{L^2(\mathcal{T}_h)}\\
&+\sigma\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|^2_{L^2(f)}\\
\geq& \|\mathcal{L}_\lambda v_h\|^2_{L^2(\mathcal{T}_h)} - \sqrt{1-\epsilon}\|v_h\|_h\|\mathcal{L}_\lambda v_h\|_{L^2(\mathcal{T}_h)} + \sigma\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|^2_{L^2(f)}\\
\geq& \big(1-\frac{1}{2}\sqrt{1-\epsilon}\big)\|\mathcal{L}_\lambda v_h\|^2_{L^2(\mathcal{T}_h)} - \frac{1}{2}\sqrt{1-\epsilon}\|v_h\|_h^2\\
&+\sigma\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|^2_{L^2(f)}.
\end{aligned}$$

Using Lemma 6 we find that

$$\begin{aligned}
B_h(v_h,v_h) \geq& \Big((1-\tau)\big(1-\frac{1}{2}\sqrt{1-\epsilon}\big) - \frac{1}{2}\sqrt{1-\epsilon}\Big)\|v_h\|_h^2\\
&+\Big(\sigma - C_1\tau^{-1}\big(1-\frac{1}{2}\sqrt{1-\epsilon}\big)\Big)\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|^2_{L^2(f)}
\end{aligned}$$

for any $\tau\in(0,1)$. For given $\alpha\in(0,1)$, we set $\tau = (1-\alpha)(1-\sqrt{1-\epsilon})/(1-\frac{1}{2}\sqrt{1-\epsilon})$. This yields

$$\begin{aligned}
B_h(v_h,v_h) \geq& \alpha\big(1-\sqrt{1-\epsilon}\big)\|v_h\|_h^2\\
&+\Big(\sigma - C_1\frac{(1-\frac{1}{2}\sqrt{1-\epsilon})^2}{(1-\alpha)(1-\sqrt{1-\epsilon})}\Big)\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v_h/\partial n_f]\!]\big\|^2_{L^2(f)}.
\end{aligned}$$

This inequality provides the desired result provided that

$$\sigma \geq \sigma_\alpha := 1 + \frac{C_1}{(1-\alpha)(1-\sqrt{1-\epsilon})}. \tag{3.2.4}$$

□

Lemma 8. *There holds*

$$|B_h(v, w_h)| \leq C\|v\|_h\|w_h\|_h$$

for all $v \in V + V_h$ *and* $w_h \in V_h$.

Proof. We assume that $\lambda > 0$; the other case is proved in a similar fashion.

Applying the definition of $B_h(\cdot,\cdot)|$ together with the Cauchy–Schwarz inequality yields

$$\begin{aligned}|B_h(v, w_h)| \leq&\|\gamma\mathcal{L}v\|_{L^2(\mathcal{T}_h)}\|\mathcal{L}_\lambda w_h\|_{L^2(\mathcal{T}_h)}\\ &+ \sigma\Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v/\partial n_f]\!]\big\|_{L^2(f)}^2\Big)^{1/2}\Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial w_h/\partial n_f]\!]\big\|_{L^2(f)}^2\Big)^{1/2}.\end{aligned}$$

We easily find that

$$\begin{aligned}\|\gamma\mathcal{L}v\|_{L^2(\mathcal{T}_h)}^2 \leq&2\|\gamma\|_{L^\infty(\Omega)}^2\big(\|A\|_{L^\infty(\Omega)}^2\|D^2v\|_{L^2(\mathcal{T}_h)}^2\\ &+ \|\boldsymbol{b}\|_{L^\infty(\Omega)}^2\|\nabla v\|_{L^2(\Omega)}^2 + \|c\|_{L^\infty(\Omega)}^2\|v\|_{L^2(\Omega)}^2\big)\\ \leq&2\|\gamma\|_{L^\infty(\Omega)}^2\max\{\|A\|_{L^\infty(\Omega)}^2, \|\boldsymbol{b}\|_{L^\infty(\Omega)}^2/(2\lambda), \|c\|_{L^\infty(\Omega)}^2/\lambda^2\}\|v\|_h^2,\end{aligned}$$

and

$$\|\mathcal{L}_\lambda w_h\|_{L^2(\mathcal{T}_h)}^2 \leq 2\big(\|\Delta w_h\|_{L^2(\mathcal{T}_h)}^2 + \lambda^2\|w_h\|_{L^2(\Omega)}^2\big) \leq 2d\|w_h\|_h^2.$$

Thus, we find that

$$\begin{aligned}|B_h(v, w_h)| \leq&C_*\|v\|_h\|w_h\|_h\\ &+ \sigma\Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial v/\partial n_f]\!]\big\|_{L^2(f)}^2\Big)^{1/2}\Big(\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\big\|[\![\partial w_h/\partial n_f]\!]\big\|_{L^2(f)}^2\Big)^{1/2}\\ \leq&\big(\sigma + C_*\big)\|v\|_h\|w_h\|_h,\end{aligned}$$

with

$$C_* = 2\sqrt{d}\|\gamma\|_{L^\infty(\Omega)} \max\{\|A\|_{L^\infty(\Omega)}, \|\boldsymbol{b}\|_{L^\infty(\Omega)}/\sqrt{2\lambda}, \|c\|_{L^\infty(\Omega)}/\lambda\}.$$

□

Theorem 4. *Suppose that the solution to (3.1.1) has regularity $u \in H^s(\Omega)$ for some $2 \le s \le k+1$, and let $u_h \in V_h$ satisfy (3.2.1). Then there holds*

$$\|u - u_h\|_h^2 \le C \inf_{v_h \in V_h} \|u - v_h\|_h^2 \lesssim \sum_{T \in \mathcal{T}_h} h_T^{2s-4} \|u\|_{H^s(T)}^2. \tag{3.2.5}$$

Proof. The first inequality is a result of Lemmas 7–8 and Cea's Lemma. The second inequality follows from standard approximation theory and scaling [5]. □

3.3 NUMERICAL EXPERIMENT FOR LINEAR PROBLEM

In this chapter we perform some numerical experiments and test the accuracy of the finite element methods for linear problems in non-divergence form. The penalty parameter is taken to be $\sigma = 10$ in all experiments.

3.3.1 Test 1

In the first experiment we solve the linear problem (3.1.1) in two dimensions on the domain $\Omega = (-\pi, \pi)^2$. The coefficients are taken to be

$$A = 10I_2 + \frac{xx^t}{|x|^2}, \quad \boldsymbol{b} = \boldsymbol{0}, \quad c = 0. \tag{3.3.1}$$

The right-hand side function g is chosen such that the exact solution to (3.1.1)

$$u(x_1, x_2) = \sin(5x_1)\sin(5x_2)/(3x_1^2 + x_2^4 + 2). \tag{3.3.2}$$

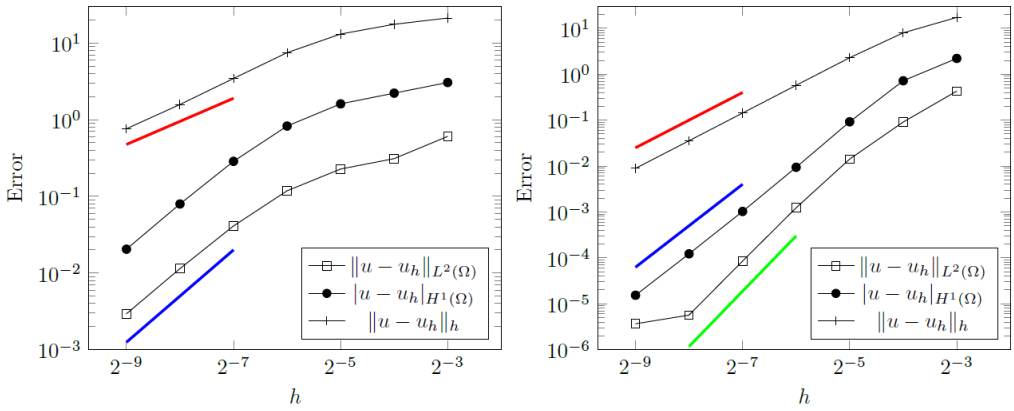

Figure 2: Test 1: Convergence plot of the two-dimensional linear problem with $k = 2$ (left) and $k = 3$ (right). The red reference line has slope $(k - 1)$, the blue reference line has slope k, and the green reference line has slope $(k + 1)$.

It is easy to see that $9|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^t A(x)\boldsymbol{\xi} \leq 11|\boldsymbol{\xi}|^2$ for all $\boldsymbol{\xi} \in \mathbb{R}^2$, and therefore the Cordes condition is satisfied with $\epsilon = 0.9$ (cf. Remark 8).We compute the numerical scheme (3.2.1) for polynomial degrees $k = 2$ and $k = 3$ and report the resulting errors in Figure 2. The figure clearly shows asymptotic $(k - 1)$th order convergence in the H^2-type norm; this agrees with the theoretical results given in Theorem 4. In addition, the experiments indicate that the method converges with optimal kth order convergence in the H^1 norm. The L^2 error converges with (sub-optimal) second order convergence when $k = 2$ and (optimal) fourth order convergence when $k = 3$.

3.3.2 Test 2

We again solve the linear problem (3.1.1) but in three dimensions with $\Omega = (-\pi, \pi)^3$, and with lower order terms:

$$A = 10I_3 + \frac{xx^t}{|x|^2}, \quad \boldsymbol{b} = (1\ 0\ 0)^t, \quad c = 10. \tag{3.3.3}$$

Note that $\operatorname{tr} A = 31$, $|A|^2 = 321$, and therefore

$$\frac{|A|^2 + |\boldsymbol{b}|^2/(2\lambda) + (c/\lambda)^2}{(\operatorname{tr} A + c/\lambda)^2} = \frac{321 + 1/(2\lambda) + (10/\lambda)^2}{(31 + 10/\lambda)^2}.$$

Taking (for example) $\lambda = 1/2$ yields

$$\frac{|A|^2 + |\boldsymbol{b}|^2/2\lambda + (c/\lambda)^2}{(\operatorname{tr} A + c/\lambda)^2} = \frac{722}{2601},$$

and therefore the Cordes condition is satisfied with $\epsilon = 435/722$ (cf. Definition 3). In the numerical experiments, the right-hand side function g is chosen such that the exact solution to (3.1.1) is given by

$$u(x_1, x_2, x_3) = \sin(5x_1)\sin(5x_2)\sin(5x_3)/(3x_1^2 + x_2^4 + 2).$$

The computed errors, listed in Figure 3, show similar behavior as the previous two-dimensional experiments. Namely, we observe asymptotic $(k-1)$th order convergence in the H^2-type norm, and

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h^k), \qquad \|u - u_h\|_{L^2(\Omega)} = \begin{cases} \mathcal{O}(h^2) & k = 2, \\ \mathcal{O}(h^4) & k = 3. \end{cases}$$

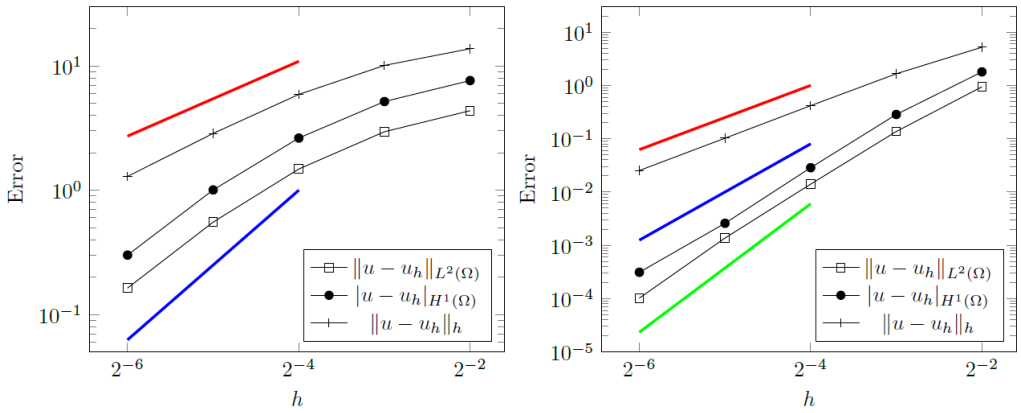

Figure 3: Test 2: Convergence plot of the three-dimensional linear problem with $k = 2$ (left) and $k = 3$ (right). The red reference line has slope $(k-1)$, the blue reference line has slope k, and the green reference line has slope $(k+1)$.

4.0 APPLICATIONS TO NONLINEAR PROBLEMS IN NONDIVERGENCE FORM

In this chapter, we develop a numerical method for the nonlinear Hamilton-Jacobi-Bellmen problem in nondivergence form. We show that the method is well defined and give some numerical examples.

4.1 ANALYSIS

In this chapter, we extend the method and analysis of Section 3, and consider numerical approximations of the Hamilton–Jacobi–Bellman equation:

$$\mathcal{F}[u] := \sup_{\alpha\in\mathcal{A}}[\mathcal{L}^\alpha u - g^\alpha] = 0 \qquad \text{in } \Omega, \tag{4.1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{4.1.1b}$$

where $\mathcal{A}$ is a compact metric space, and $\{\mathcal{L}^\alpha\}_{\alpha\in\mathcal{A}}$ is a family of second–order operators in non–divergence form, namely,

$$\mathcal{L}^\alpha v = A^\alpha : D^2 v + \boldsymbol{b}^\alpha \cdot \nabla v - c^\alpha v. \tag{4.1.2}$$

As in the previous section, $\Omega \subset \mathbb{R}^d$ is a convex domain, but we assume the coefficients satisfy the stronger conditions $\boldsymbol{b}^\alpha \in [C(\bar{\Omega})]^d$, $c^\alpha \in C(\bar{\Omega})$, and $A^\alpha \in [C(\bar{\Omega})]^{d\times d}$ for all $\alpha \in \mathcal{A}$, and that the data is continuous with respect to α, e.g., the function $\alpha \to A^\alpha(x)$ is continuous on $\mathcal{A}$ for fixed $x \in \bar{\Omega}$. In addition, we assume that c^α is nonnegative, and that the family of operators $\{A^\alpha\}_{\alpha\in\mathcal{A}}$ is uniformly positive definite and uniformly satisfies the Cordes condition with respect to α, i.e.,

$$\nu|\boldsymbol{\xi}|^2 \le \sum_{i,j=1}^{d} A_{ij}^\alpha(x)\xi_i\xi_j \le \bar{\nu}|\boldsymbol{\xi}|^2 \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \forall x \in \Omega, \forall \alpha \in \Lambda, \tag{4.1.3}$$

and if $c^\alpha \not\equiv 0$ or $\boldsymbol{b}^\alpha \not\equiv 0$ for some $\alpha \in \mathcal{A}$, there exists $\lambda > 0$ and $\epsilon \in (0,1)$ such that

$$\frac{|A^\alpha|^2 + |\boldsymbol{b}^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2}{(\operatorname{tr} A^\alpha + c^\alpha/\lambda)^2} \le \frac{1}{d+\epsilon} \quad \forall x \in \Omega. \tag{4.1.4a}$$

Otherwise, if $c^\alpha \equiv 0$ and $\boldsymbol{b}^\alpha \equiv 0$ for all $\alpha \in \mathcal{A}$, there exists $\epsilon \in (0,1)$ such that

$$\frac{|A^\alpha|^2}{(\operatorname{tr} A^\alpha)^2} \le \frac{1}{d-1+\epsilon} \quad \forall x \in \Omega. \tag{4.1.4b}$$

Under these conditions, there holds the following result [31, Theorem 3].

Theorem 5. *Under the given conditions, there exists a unique strong solution $u \in V$ to (4.1.1).*

Proof. We refer to [31, Theorem 3] for a complete proof of this result. Here, we just state the main ideas of the proof.

Analogous to (3.1.3), for each $\alpha \in \mathcal{A}$, we define the (positive) function

$$\gamma^\alpha := \frac{\operatorname{tr} A^\alpha + c^\alpha/\lambda}{|A^\alpha|^2 + |\boldsymbol{b}^\alpha|^2/2\lambda + (c^\alpha/\lambda)^2} \quad \text{in } \Omega, \tag{4.1.5}$$

and in the special case where $\boldsymbol{b}^\alpha \equiv 0$ and $c^\alpha \equiv 0$ for all $\alpha \in \mathcal{A}$, we set

$$\gamma^\alpha := \frac{\operatorname{tr} A^\alpha}{|A^\alpha|^2} \quad \text{in } \Omega.$$

The Cordes condition and the uniform ellipticity of A shows that there exists γ_0 such that $\gamma \geq \gamma_0$ for all $\alpha \in \mathcal{A}$.

Define the operator $\mathcal{F}_\gamma : H^2(\Omega) \to L^2(\Omega)$ by

$$\mathcal{F}_\gamma[u] := \sup_{\alpha \in \mathcal{A}} \gamma^\alpha \big(\mathcal{L}^\alpha u - g^\alpha \big).$$

With $\mathcal{L}_\lambda$ defined by (3.1.4), one concludes that $u \in V$ is a strong solution to (4.1.1) if and only if

$$\langle \mathcal{M}[u], v \rangle := \int_\Omega \mathcal{F}_\gamma[u] \mathcal{L}_\lambda v \, dx = 0 \qquad \forall v \in V, \tag{4.1.6}$$

where $\langle \cdot, \cdot \rangle$ is the dual pairing between V^* and V. Continuity of the data and the compactness of $\mathcal{A}$ implies that $\mathcal{M}$ is Lipschitz continuous. Let $u, v, z \in V$ we have

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

And, by using the Cordes condition, one can show that $\mathcal{M}$ is strongly monotone.

Lemma 9. *Let Ω be a bounded convex polygonal domain of $\mathbb{R}^d$. Suppose that (4.1.3) holds, and that the coefficients and continuous and satisfy the Cordes condition (4.1.4). Then there holds the following inequality:*

$$|\mathcal{F}_\gamma[v] - \mathcal{F}_\gamma[z] - \mathcal{L}_\lambda(v - z)| \leq \sqrt{1 - \epsilon} \sqrt{|D^2(v - z)|^2 + 2\lambda |\nabla(v - z)|^2 + \lambda^2 |(v - z)|^2}. \tag{4.1.7}$$

Proof. Using Lemma 4, we have

$$\left|\gamma^\alpha \mathcal{L}^\alpha w - \mathcal{L}_\lambda w\right| \le \sqrt{1-\epsilon}\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2} \qquad \forall \alpha \in \mathcal{A},$$

and therefore

$$\sup_{\alpha\in\mathcal{A}} \left|\gamma^\alpha \mathcal{L}^\alpha w - \mathcal{L}_\lambda w\right| \le \sqrt{1-\epsilon}\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2}.$$

It then follows, with $w = v - z$, that

$$\begin{aligned} |\mathcal{F}_\gamma[v] - \mathcal{F}_\gamma[z] - \mathcal{L}_\lambda w| &= \left| \sup_{\alpha\in\mathcal{A}}(\gamma^\alpha(L^\alpha v - g^\alpha)) - \sup_{\alpha\in\mathcal{A}}(\gamma^\alpha(\mathcal{L}^\alpha z - g^\alpha) - \mathcal{L}_\lambda w\right| \\ &\le \sup_{\alpha\in\mathcal{A}} |\gamma^\alpha \mathcal{L}^\alpha w - \mathcal{L}_\lambda w| \\ &\le \sqrt{1-\epsilon}\sqrt{|D^2 w|^2 + 2\lambda|\nabla w|^2 + \lambda^2|w|^2}. \end{aligned}$$

□

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \langle \mathcal{M}[v], w\rangle - \langle \mathcal{M}[z], w\rangle = \sum_{T\in\mathcal{T}_h} \int_T (\mathcal{F}_\gamma[v_h] - \mathcal{F}_\gamma[z_h] - \mathcal{L}_\lambda w_h)\, \mathcal{L}_\lambda w_h \, dx \\ + \|\mathcal{L}_\lambda w\|^2_{L^2(\mathcal{T}_h)} \end{aligned} \tag{4.1.8}$$

Continuing as in the proof of Lemma 7, we conclude that, for any $\alpha \in (0,1)$, we have

$$\langle \mathcal{M}[v], w\rangle - \langle \mathcal{M}[z], w\rangle \ge \alpha\big(1 - \sqrt{1-\epsilon}\big)\|w\|_h^2,$$

provided that (3.2.4) is satisfied; thus, $\mathcal{M}_h$ is strongly monotone. The Browder–Minty Theorem then gives the existence and uniqueness of $u \in V$ satisfying (4.1.6), and thus (4.1.1). We adopt this framework in the finite element analysis below.

□

4.2 FINITE ELEMENT METHOD

Define the operator $\mathcal{M}_h : V_h + V \to V_h^*$ such that

$$\langle \mathcal{M}_h[w], v\rangle := \sum_{T\in\mathcal{T}_h} \int_T \mathcal{F}_\gamma[w]\mathcal{L}_\lambda v_h \, dx + \sigma \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \int_f [\![\partial w/\partial n_f]\!][\![\partial v_h/\partial n_f]\!] \, ds.$$

We consider the following finite element method for problem (4.1.1): Find $u_h \in V_h$ such that

$$\langle \mathcal{M}_h[u_h], v_h\rangle = 0 \qquad \forall v_h \in V_h. \tag{4.2.1}$$

Note that since the exact (strong) solution to (4.1.1) has regularity $u \in H^2(\Omega)$, and therefore, since u satisfies (4.1.1) almost everywhere in Ω, we conclude that $\langle \mathcal{M}_h[u], v_h\rangle = 0$ for all $v_h \in V_h$, i.e., the method is consistent.

Theorem 6. *Let Ω be a bounded convex polygonal domain of $\mathbb{R}^d$, let $\mathcal{T}_h$ be a simplicial, conforming, and shape-regular mesh of Ω without hanging nodes. Suppose that the coefficients are continuous in $\overline{\Omega}$ and satisfy the Cordes condition (4.1.4). Then there exist a unique solution $u_h \in V_h$ satisfying (4.2.1) provided σ is sufficiently large. Moreover, there holds*

$$\|u - u_h\|_h \le C \inf_{v_h\in V_h} \|u - v_h\|_h \le C \sum_{T\in\mathcal{T}_h} h_T^{2s-4} |u|_{H^s(T)} \tag{4.2.2}$$

provided that $u \in H^s(\Omega)$ for some $2 \le s \le k+1$.

Proof. Let $v_h, z_h \in V_h$, and set $w_h = v_h - z_h$, then by the Cauchy-Schwarz inequality, we have

$$\langle \mathcal{M}[v_h], w_h\rangle - \langle \mathcal{M}[z_h], w_h\rangle = \sum_{T\in\mathcal{T}_h} \int_T \left(\mathcal{F}_\gamma[v_h] - \mathcal{F}_\gamma[z_h] - \mathcal{L}_\lambda w_h\right) \mathcal{L}_\lambda w_h \, dx + \|\mathcal{L}_\lambda w\|^2_{L^2(\mathcal{T}_h)} + \sigma \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \int_f [\![\partial w_h/\partial n_f]\!][\![\partial w_h/\partial n_f]\!] \, ds \tag{4.2.3}$$

Continuing as in the proof of Lemma 7, we conclude that, for any $\alpha \in (0,1)$, we have

$$\langle \mathcal{M}[v_h], w_h\rangle - \langle \mathcal{M}[z_h], w_h\rangle \geq \alpha\big(1 - \sqrt{1-\epsilon}\big)\|w_h\|_h^2,$$

provided that (3.2.4) is satisfied; thus, $\mathcal{M}_h$ is strongly monotone. Continuing as in Lemma 8, we also conclude that $\mathcal{M}_h$ is Lipschitz continuous (with respect to $\|\cdot\|_h$). By the Browder-Minty theorem there exist a unique solution $u_h \in V_h$ to (4.2.1). Finally, the error estimate (4.2.2) follows from the consistency of the scheme, the monotonicity and Lipschitz continuity of $\mathcal{M}_h$, and standard interpolation estimates. □

To implement the method, we use the discretization of $\mathcal{A}$ and Howard's algorithm 1 to solve the nonlinear system. By [25, Section 5.3], the Howard's algorithm converges superlinearly to u_h with a good initial guess α_0.

4.3 NUMERICAL EXPERIMENT FOR NONLINEAR PROBLEM

In this section we perform some numerical experiments and test the accuracy of the finite element methods for nonlinear problems in non-divergence form. The penalty parameter is taken to be $\sigma = 10$ in all experiments.

Algorithm 1 Howard's algorithm

1: Initialize $\alpha_0 \in \mathcal{A}$,

2: **while** $i \geq 0$ **do**

3: Find u_h^i such that $\forall v_h \in V_h$,

$$\sum_{T\in\mathcal{T}_h} \int_T \gamma^{\alpha_i}(\mathcal{L}^{\alpha_i}u_h^i - g^{\alpha_i})\mathcal{L}_\lambda v_h \, dx + \sigma \sum_{f\in\mathcal{F}_h^I} h_f^{-1} \int_f [\![\partial w/\partial n_f]\!][\![\partial v_h/\partial n_f]\!] \, ds = 0.$$

4: **if** $i \geq 1$ and $\|u_h^i - u_h^{i-1}\|_h \leq tolerance$ **then** Stop.

5: **end if**

6: $\alpha_{i+1} = argmax_{\alpha\in\mathcal{A}}(\mathcal{L}^\alpha u_h^i - g^\alpha)$,

7: $i = i + 1$.

8: **end while**

4.3.1 Test 3

In these series of experiments, we solve the nonlinear Hamilton-Jacobi-Bellman problem (4.1.1) with $d = 2$, $\Omega = (-\pi, \pi)^2$, $\mathcal{A} = \{1, 2\}$, and

$$A^1 = \begin{pmatrix} 2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix} + \frac{x_1}{|x_1|}\frac{x_2}{|x_2|}\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/2 \end{pmatrix},$$
$$A^2 = \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 2 \end{pmatrix} + \frac{x_1}{|x_1|}\frac{x_2}{|x_2|}\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1 \end{pmatrix},$$
$$\boldsymbol{b}^1 = \boldsymbol{b}^2 = (1\ 0)^t, \quad c^1 = c^2 = 1.$$

The source functions $\{g^\alpha\}_{\alpha\in\mathcal{A}}$ are chosen so that the solution of (4.1.1) is

$$u(x_1, x_2) = \sin(x_1)\sin(x_2). \tag{4.3.1}$$

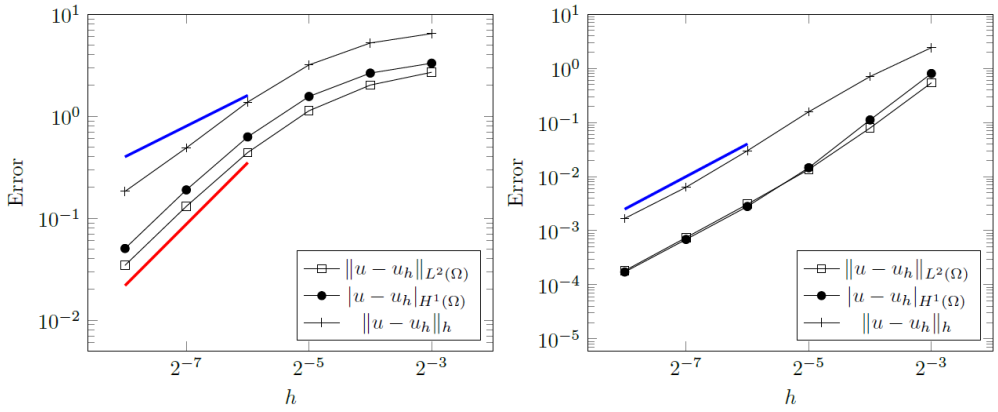

Figure 4: Test 3: Convergence plot of the two-dimensional nonlinear problem with $k = 2$ (left) and $k = 3$ (right). The red reference line has slope k and the blue reference line has slope $(k-1)$.

With this data, we can verify that the Cordes condition (3.1.2) holds with $\lambda = 1$ and $\epsilon = 1/6$. Note that the matrices are discontinuous at the lines $x_1 = 0$ and $x_2 = 0$. Therefore the problem does not satisfy the conditions assumed in Theorems 5 and 6. Nonetheless, the plots of the errors given in Figure 4 show that for $k = 2$, the method converges at sub-optimal rate for L^2 error and optimal rates for H^1 and h norm error. For $k = 3$, all the errors converge at rate $k - 1$. Namely, the numerical experiments indicate that the method converges with at least the following convergence rates:

$$\|u - u_h\|_{L^2(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_h = \mathcal{O}(h^{k-1}).$$

4.3.2 Test 4

In this experiment, we solve the nonlinear Hamilton-Jacobi-Bellman problem (4.1.1) with $d = 2$, $\Omega = (-\pi, \pi)^2$, $\mathcal{A} = [0, 1]$, and

$$A^\alpha = \begin{pmatrix} 2+\alpha & 1 \\ 1 & 1+\alpha \end{pmatrix}, \quad \boldsymbol{b}^1 = \boldsymbol{b}^2 = (0\ 0)^t, \quad c^1 = c^2 = 0.$$

The source functions $\{g^\alpha\}_{\alpha\in\mathcal{A}}$ are chosen so that the solution of (4.1.1) is

$$u(x_1, x_2) = \sin(x_1)\sin(x_2). \tag{4.3.2}$$

With this data, we can verify that the Cordes condition (3.1.2) holds with $\epsilon = 2/7$. The plots of the errors given in Figure 5 show that the method converges as the mesh is refined, at optimal rate for h-norm error when $k = 2, 3$, but at sub-optimal rates for both L^2 norm and H^1 norm when $k = 3$. Namely, the numerical experiments indicate that the method converges with at least the following convergence rates:

$$\|u - u_h\|_{L^2(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_h = \mathcal{O}(h^{k-1}).$$

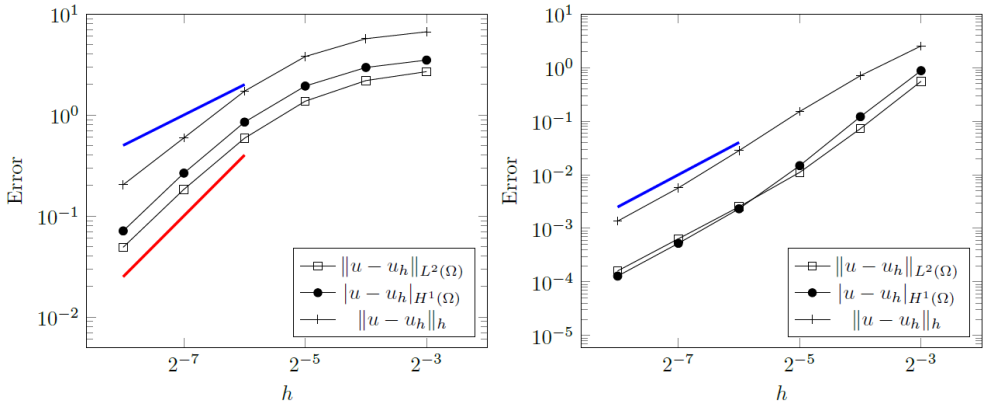

Figure 5: Test 4: Convergence plot of the two-dimensional nonlinear problem with $k = 2$ (left) and $k = 3$ (right). The red reference line has slope k and the blue reference line has slope $(k-1)$.

5.0 MONGE-AMPÈRE PROBLEM

In this chapter, we consider the following Dirichlet boundary value problem of a fully nonlinear elliptic Monge-Ampère equation:

$$det(D^2u) = (\frac{f}{d})^d \quad in\ \Omega, \tag{5.0.1a}$$

$$u(x) = g(x) \quad on\ \partial\Omega, \tag{5.0.1b}$$

where Ω and $\partial\Omega$ denote, respectively, a bounded convex polygonal domain in $\mathbb{R}^d (d \geq 2)$ and its boundary. The Hessian of the function u is denoted D^2u. The function $f : \Omega \to [0, \infty)$ and $g : \partial\Omega \to \mathbb{R}$ are assumed to be bounded and continuous. The Monge-Ampère equation is a well-known example of fully nonlinear second order PDEs. It arises from many fields in science and engineering such as astrophysics, antenna design, differential geometry, geostrophic fluid dynamics, image processing, materials science, mathematical finance, mesh generation, meteorology, and optimal transport [8]. For the classical solution theory, see [14] for existence and uniqueness of the solution for Monge-Ampère equations. For the numerical solution, the fully nonlinear structure of the PDE prevents any Galerkin-type numerical method to be used directly. In this chapter, we approach the Monge-Ampère equations by an equivalent Hamilton-Jacobi-Bellman (HJB) formulation. Then using the Discrete

Miranda Talanti inequality we design a numerical method to solve the HJB equations.

5.1 HAMILTON-JACOBI-BELLMAN FORM OF THE MONGE-AMPÈRE EQUATION

Let $\mathbf{S}$ be the set of $d \times d$ symmetric real matrices, let $\mathbf{S}_+ := \{A \in \mathbf{S} : A \geq 0\}$ be the set of nonnegative symmetric matrices and $\mathbf{S}_1 := \{B \in \mathbf{S}_+ : trB = 1\}$. $\mathbf{S}_1$ is a compact subset of $\mathbf{S}_+$ and $\mathbf{S}_1$ is bounded in the Euclidean norm.
To simplify the notation, we define the Bellman operator

$$H(A, f) := \sup_{B \in \mathbf{S}_1} (-B : A + f \sqrt[d]{det(B)}) \quad \forall A \in \mathbf{S}, f \in [0, \infty), \tag{5.1.1}$$

and the Monge-Ampère operator

$$M(A, f) := \left(\frac{f}{d}\right)^d - det(A) \quad \forall A \in \mathbf{S}, f \in [0, \infty). \tag{5.1.2}$$

Then the Monge-Ampère problem (5.0.1) can be rewritten as

$$M(D^2u(x), f(x)) = 0 \qquad \forall x \ in \ \Omega, \tag{5.1.3a}$$

$$u(x) = g(x) \quad \forall x \ on \ \partial\Omega. \tag{5.1.3b}$$

We can also define the Bellman problem

$$H(D^2u(x), f(x)) = 0 \qquad \forall x \ in \ \Omega, \tag{5.1.4a}$$

$$u(x) = g(x) \quad \forall x \ on \ \partial\Omega. \tag{5.1.4b}$$

The following Lemma is given by [21, Page 49].

Lemma 10. *Let $B \in \mathbf{S}_+$, and $c_i(B)$ denote the determinant of the submatrix resulting from removing the first i rows and columns of B. Then we have*

$$det(B) \leq \left(\prod_{k=1}^{i} B_{kk} \right) c_i(B)$$

for $1 \leq i \leq d$.

Lemma 11. *Let $f \in [0, \infty)$ and $A \in \mathbf{S}$. Then $H(A, f) = 0$ holds if and only if $M(A, f) = 0$ and $A \in \mathbf{S}_+$.*

Proof. For any $A \in \mathbf{S}$, there exists a unitary matrix Q and a diagonal matrix D such that $A = QDQ^T$, where the diagonal entries of D are the eigenvalues of A. Then for any $B \in \mathbf{S}_1$, we have

$$\begin{aligned} -B : A + f\sqrt[d]{det(B)} &= -\mathrm{tr}\,(BA) + f\sqrt[d]{det(B)} \\ &= -\mathrm{tr}\,(BQDQ^T) + f\sqrt[d]{det(B)} \\ &= -\mathrm{tr}\,(Q^TBQD) + f\sqrt[d]{det(Q^T)det(B)det(Q)} \\ &= -(Q^TBQ) : D + f\sqrt[d]{det(Q^TBQ)}. \end{aligned}$$

Since Q is unitary matrix, we have $Q^T\mathbf{S}_1Q = \mathbf{S}_1$. Therefore

$$\begin{aligned} \sup_{B \in \mathbf{S}_1} (-B : A + f\sqrt[d]{det(B)}) &= \sup_{B \in \mathbf{S}_1} (-(Q^TBQ) : D + f\sqrt[d]{det(Q^TBQ)}) \\ &= \sup_{B \in \mathbf{S}_1} (-B : D + f\sqrt[d]{det(B)}). \end{aligned}$$

Thus the supremum with the matrix A is the same as the supremum with the diagonal matrix D, i.e., $H(A, f) = H(D, f)$. And by $det(A) = det(QDQ^T) = det(A)$, we have $M(A, f) = M(D, f)$. Hence we can assume that A is a diagonal matrix without loss of generality.

Let $\mathbf{S}_D := \{B_D \in \mathbf{S}_1 : B_D \textit{ is a diagonal matrix}\}$. Set $a_i = A_{ii}$ and set $b_i = B_{ii}$. Consider the case $f > 0$. By Lemma 10, choose $i = d - 1$, we have that

$$det(B) \leq \left(\prod_{k=1}^{d-1} B_{kk}\right) c_{d-1}(B) = \left(\prod_{i=1}^{d} b_i\right).$$

For any $B \in \mathbf{S}_1 \subset \mathbf{S}_+$, we have $b_i \geq 0$ for all i and $\operatorname{tr} B = \sum_{i=1}^{d} b_i = 1$. Then there exist a matrix $B_D \in \mathbf{S}_D \subset \mathbf{S}_1$ with $(B_D)_{ii} = b_i$ for all i and

$$-B : A + f\sqrt[d]{det(B)} \leq -\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} = -B_D : A + f\sqrt[d]{det(B_D)},$$

so that

$$\sup_{B\in\mathbf{S}_1}(-B : A + f\sqrt[d]{det(B)}) \leq \sup_{B_D\in\mathbf{S}_D}(-B_D : A + f\sqrt[d]{det(B_D)}). \tag{5.1.5}$$

Since $\mathbf{S}_D \subset \mathbf{S}_1$, we also have

$$\sup_{B\in\mathbf{S}_1}(-B : A + f\sqrt[d]{det(B)}) \geq \sup_{B_D\in\mathbf{S}_D}(-B_D : A + f\sqrt[d]{det(B_D)}). \tag{5.1.6}$$

By (5.1.5) and (5.1.6), we have that

$$H(A, f) = \sup_{B\in\mathbf{S}_1}(-B : A + f\sqrt[d]{det(B)}) = \sup_{B_D\in\mathbf{S}_D}(-B_D : A + f\sqrt[d]{det(B_D)}).$$

Then to prove the Lemma, we can instead prove

$$\sup\{-\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} : b_i \geq 0, \sum_{i=1}^{d} b_i = 1\} = 0 \tag{5.1.7}$$

if and only if

$$a_i \geq 0, f^d = d^d \prod_{i=1}^{d} a_i. \tag{5.1.8}$$

First we show that (5.1.7) implies (5.1.8). When $d = 1$, we have $f = a_1$, and the two equation are the same. When $d \geq 2$, if $b_i = 1$ and $b_j = 0$ for all $j \neq i$, by (5.1.7) we

have $a_i \geq 0$ for all i. If $a_1 = 0$, let $b_1 = 1-(d-1)t$, $b_2 = b_3 = \ldots = b_d = t > 0$, then we have

$$-t\sum_{i=1}^{d} a_i + f((1-(d-1)t)t^{d-1})^{\frac{1}{d}} \leq 0.$$

However, for t is small enough, we have

$$-t\sum_{i=1}^{d} a_i + f((1-(d-1)t)t^{d-1})^{\frac{1}{d}} \approx -t\sum_{i=1}^{d} a_i + t^{\frac{d-1}{d}} f > 0.$$

Which contradicts the assumption, so $a_1 > 0$. Thus, by similar arguments, $a_i > 0$ for all i. Hence, for any $B \in \mathbf{S}_1$, we have

$$-\sum_{i=1}^{d} a_i b_i \leq -\min_{1\leq i\leq d}\{a_i\} < 0.$$

Since $\mathbf{S}_1$ is a closed set and $-B : A + f\sqrt[d]{det(B)}$ is a continuous function with respect to B, there exist a maximizer $B' \in \mathbf{S}_1$ that maximize (5.1.7). Then $f(\prod_{i=1}^{d} b_i')^{\frac{1}{d}} > 0$ implies $b_i' > 0$ for all i. Therefore, by the inequality between arithmetic and geometric means, we have that

$$\begin{aligned} 0 =& -\sum_{i=1}^{d} a_i b_i' + f(\prod_{i=1}^{d} b_i')^{\frac{1}{d}} \\ \leq& (\prod_{i=1}^{d} b_i')^{\frac{1}{d}}(f - d(\prod_{i=1}^{d} a_i)^{\frac{1}{d}}). \end{aligned}$$

This implies $f - d(\prod_{i=1}^{d} a_i)^{\frac{1}{d}} \geq 0$. For the other direction, let $b_i = a_i^{-1}(\sum_{i=1}^{d} a_i^{-1})^{-1}$, which satisfy $b_i \geq 0, \sum_{i=1}^{d} b_i = 1$. By (5.1.7), we have

$$\begin{aligned} 0 \geq& -\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} \\ =& -d(\sum_{i=1}^{d} a_i^{-1})^{-1} + f\left(\prod_{i=1}^{d} a_i\right)^{-\frac{1}{d}} (\sum_{i=1}^{d} a_i^{-1})^{-1} \end{aligned}$$

$$= (\sum_{i=1}^{d} a_i^{-1})^{-1} \left(\prod_{i=1}^{d} a_i\right)^{-\frac{1}{d}} (f - d\left(\prod_{i=1}^{d} a_i\right)^{\frac{1}{d}}).$$

This implies

$$f - d\left(\prod_{i=1}^{d} a_i\right)^{\frac{1}{d}} \leq 0.$$

Thus, (5.1.8) holds.

Second to show that (5.1.8) implies (5.1.7). By the inequality between arithmetic and geometric means, we have

$$-\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} \leq (\prod_{i=1}^{d} b_i)^{\frac{1}{d}} (f - d(\prod_{i=1}^{d} a_i)^{\frac{1}{d}}) = 0.$$

So

$$\sup\{-\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} : b_i \geq 0, \sum_{i=1}^{d} b_i = 1\} \leq 0.$$

Let $b_i = a_i^{-1}(\sum_{i=1}^{d} a_i^{-1})^{-1}$, then we have

$$-\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} = (\sum_{i=1}^{d} a_i^{-1})^{-1} \left(\prod_{i=1}^{d} a_i\right)^{-\frac{1}{d}} (f - d\left(\prod_{i=1}^{d} a_i\right)^{\frac{1}{d}}) = 0.$$

Hence $\sup\{-\sum_{i=1}^{d} a_i b_i + f(\prod_{i=1}^{d} b_i)^{\frac{1}{d}} : b_i \geq 0, \sum_{i=1}^{d} b_i = 1\} \geq 0$, implies that (5.1.7) holds.

For the case where $f = 0$, the solution is $a_i = 0$ for some i. Thus both conditions are satisfied. □

Lemma 12. *We have the following inequality:*

$$det(A)^{\frac{1}{d}} det(B)^{\frac{1}{d}} \leq \frac{1}{d} \mathrm{tr}\, AB = \frac{1}{d}(A : B), \quad \forall A, B \in \mathbf{S}_+$$

with equality holds if and only if $B^{1/2}AB^{1/2} = cI$ for some scalar $c \geq 0$, where $B^{1/2}$ is the symmetric nonnegative square root of B.

Proof. By the inequality between arithmetic and geometric means, we have $\frac{1}{d}\mathrm{tr}\, M \geq (det(M))^{\frac{1}{d}}$ for any matrix $M \in \mathbf{S}_+$, and the equality holds if and only if $M = cI$ for some scalar $c \geq 0$. Let $C = B^{1/2}AB^{1/2}$, $C \in \mathbf{S}_+$, then $\frac{1}{d}\mathrm{tr}\, C \geq (det(C))^{\frac{1}{d}}$ with equality holds if and only if $C = cI$ for some $c \geq 0$. By $det(C) = det(B^{1/2}AB^{1/2}) = det(A)det(B)$ and $\mathrm{tr}\, C = \mathrm{tr}\, AB^{1/2}B^{1/2}$=$\mathrm{tr}\, AB = A : B$, we have $det(A)^{\frac{1}{d}}det(B)^{\frac{1}{d}} \leq \frac{1}{d}(A : B)$. □

Lemma 13. *There exists a maximizer $B' \in \mathbf{S}_1$ of the supremum in (5.1.1) which commutes with $A \in \mathbf{S}$. In particular, there is a coordinate transformation, depending on A, which simultaneously diagonalizes A and B'.*

Proof. Let A be the solution for $H(A, f) = 0$. Then by Lemma 11, $f = d(det(A))^{\frac{1}{d}}$. By Lemma 12 we have $-B : A + f(det(B))^{\frac{1}{d}} \leq (det(B))^{\frac{1}{d}}(f - d(det(A))^{\frac{1}{d}}) = 0$, the equality holding if and only if $B^{1/2}AB^{1/2} = cI$ for some scalar $c \geq 0$. If B' is the maximizer in (5.1.1), then $-B' : A + f(det(B'))^{\frac{1}{d}} = H(A, f) = 0$, thus $B'^{1/2}AB'^{1/2} = cI$, B' must be commute with A. □

Lemma 14. *If A is invertiable and $A = QDQ^T$, then the maximizer $B' \in \mathbf{S}_1$ of the supremum in (5.1.1) is $B' = Q\frac{D^{-1}}{\mathrm{tr}\, D^{-1}}Q^T$.*

Proof. By Lemma 13, if $A \in \mathbf{S}_+$, we can compute the supremum by finding the maximizer $B' \in \mathbf{S}_1$. If A is not invertible, then the supremum is 0. If A is invertible, then there exist unitary matrix Q and diagonal matrix $D > 0$ such that $A = QDQ^T$. By B' commutes with A and $B'^{1/2}AB'^{1/2} = cI$, we have $B' = cQD^{-1}Q^T$. Since $B' \in \mathbf{S}_1$, we have $c = \frac{1}{\mathrm{tr}\,(QD^{-1}Q^T)} = \frac{1}{\mathrm{tr}\, D^{-1}}$. Therefore, the maximizer $B' = Q\frac{D^{-1}}{\mathrm{tr}\, D^{-1}}Q^T$ and $H(A, f) = -\frac{d}{\mathrm{tr}\, D^{-1}} + f(det(D^{-1})^{\frac{1}{d}})$. □

5.2 MONGE-AMPÈRE EQUATION WITH PERTURBATION

In this section, we find the relation between the solution to the HJB problem (5.1.4) and the Monge-Ampère problem (5.1.3). To ensure the existence of a strong solution to the HJB problem, we add a small perturbation to it. Then we do the analysis for the HJB problem with perturbation and the corresponding Monge-Ampère problem with perturbation.

Theorem 7. *Any H^2 regular strong solution u to the HJB problem (5.1.4) is also a strong solution to the Monge-Ampère problem (5.1.3).*

Proof. Let $u \in H^2(\Omega)$ be a strong solution to (5.1.4). Then we have that u satisfied (5.1.4) almost everywhere, i.e. $H(D^2u, f) = 0$ a.e. in Ω and $u = g$ at $\partial\Omega$. Thus, by Lemma 11, $M(D^2u, f) = 0$ a.e. in Ω. Which implies u satisfied (5.1.3) almost everywhere. Therefore, u is a strong solution of (5.1.3). □

In the case $d = 2$, we have $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbf{S}_1$ but $\frac{1^2+0^2}{(1+0)^2} = 1 > \frac{1}{1+\epsilon}$ for any $\epsilon > 0$. So the Bellman operator does not satisfy the Cordes condition, and the existence of a strong solution is not guaranteed. To make the Bellman operator satisfy the Cordes condition, we introduce a small perturbation to it.

We define the Bellman operator with perturbation as following.

$$H_\epsilon(A, f) := \sup_{B \in \mathbf{S}_1} (-\epsilon \operatorname{tr} A - B : A + f \sqrt[d]{det(B)}) \quad \forall A \in \mathbf{S}, f \in [0, \infty). \tag{5.2.1}$$

And we define the corresponding Monge-Ampère operator with perturbation

$$M_\epsilon(A, f) := \left(\frac{f}{d}\right)^d - det(A) - (\epsilon^2 + \epsilon)(\operatorname{tr} A)^2 \quad \forall A \in \mathbf{S}, f \in [0, \infty), \tag{5.2.2}$$

and consider the Monge-Ampère problem with perturbation

$$M_\epsilon(D^2u(x), f(x)) = 0 \qquad \forall x \text{ in } \Omega, \tag{5.2.3a}$$

$$u(x) = g(x) \quad \forall x \text{ on } \partial\Omega. \tag{5.2.3b}$$

We also define the Bellman problem with perturbation

$$H_\epsilon(D^2u(x), f(x)) = 0 \qquad \forall x \text{ in } \Omega, \tag{5.2.4a}$$

$$u(x) = g(x) \quad \forall x \text{ on } \partial\Omega. \tag{5.2.4b}$$

Then we have a similar result as Lemma 11.

Lemma 15. *Let $d = 2$ and $f \geq 0$. Then $H_\epsilon(A, f) = 0$ if and only if $M_\epsilon(A, f) = 0$ and $\operatorname{tr} A \geq 0$.*

Proof. Without lost of generality, we can assume that A, B are diagonal matrix by the same reasoning as Lemma 10. Let $A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}$ and $B = \begin{bmatrix} \lambda & 0 \\ 0 & 1-\lambda \end{bmatrix}$. Let

$$h(\lambda) = -\epsilon(a_1 + a_2) - a_1\lambda - a_2(1-\lambda) + f\sqrt{\lambda(1-\lambda)}.$$

Thus, we have that

$$H_\epsilon(A, f) = \sup_{0 \leq \lambda \leq 1} h(\lambda).$$

For the case $f > 0$, by taking the derivative of $h(\lambda)$, we have

$$h'(\lambda) = a_2 - a_1 + \frac{f(1-2\lambda)}{2\sqrt{\lambda(1-\lambda)}},$$

$$h''(\lambda) = -\frac{f}{4(\lambda - \lambda^2)^{3/2}}.$$

So $h''(\lambda) < 0$ for all λ and $h'(\lambda) = 0$ when $\lambda_* = \left(1 - \frac{a_1 - a_2}{\sqrt{(a_1-a_2)^2+f^2}}\right)/2$. Therefore $h(\lambda)$ achieves a maximum at λ_* and

$$H_\epsilon(A, f) = h(\lambda_*) = -(\epsilon + \frac{1}{2})(a_1 + a_2) + \frac{1}{2}\sqrt{(a_1 - a_2)^2 + f^2}.$$

For the case $f = 0$, we have

$$H_\epsilon(A, f) = \max\{h(0), h(1)\} = -(\epsilon + \frac{1}{2})(a_1 + a_2) + \frac{1}{2}|a_1 - a_2|.$$

Which has the same form as the case $f > 0$.

If $H_\epsilon(A, f) = 0$, we have

$$\begin{aligned}
-(\epsilon + \frac{1}{2})(a_1 + a_2) + \frac{1}{2}\sqrt{(a_1 - a_2)^2 + f^2} =& 0, \\
\frac{1}{2}\sqrt{(a_1 - a_2)^2 + f^2} =& (\epsilon + \frac{1}{2})(a_1 + a_2), \\
\frac{1}{4}((a_1 - a_2)^2 + f^2) =& (\epsilon + \frac{1}{2})^2(a_1 + a_2)^2, \\
(\frac{f}{2})^2 - a_1 a_2 - (\epsilon^2 + \epsilon)(a_1 + a_2)^2 =& 0, \\
M_\epsilon(A, f) =& 0.
\end{aligned}$$

Furthermore,

$$\operatorname{tr} A = a_1 + a_2 = \frac{1}{2\epsilon + 1}\sqrt{(a_1 - a_2)^2 + f^2} \geq 0.$$

Thus $H_\epsilon(A, f) = 0$ implies $M_\epsilon(A, f) = 0$ and $\operatorname{tr} A \geq 0$.

If $M_\epsilon(A, f) = 0$, then $(\frac{f}{2})^2 - a_1 a_2 - (\epsilon^2 + \epsilon)(a_1 + a_2)^2 = 0$. Also if $\operatorname{tr} A = a_1 + a_2 \geq 0$, we have

$$\begin{aligned}
(\frac{f}{2})^2 - a_1 a_2 - (\epsilon^2 + \epsilon)(a_1 + a_2)^2 =& 0, \\
\frac{1}{4}((a_1 - a_2)^2 + f^2) =& (\epsilon + \frac{1}{2})^2(a_1 + a_2)^2, \\
\frac{1}{2}\sqrt{(a_1 - a_2)^2 + f^2} =& (\epsilon + \frac{1}{2})(a_1 + a_2),
\end{aligned}$$

$$-(\epsilon+\frac{1}{2})(a_1+a_2)+\frac{1}{2}\sqrt{(a_1-a_2)^2+f^2}=0,$$
$$H_\epsilon(A,f)=0.$$

Therefore, we have that $H_\epsilon(A,f)=0$ if and only if $M_\epsilon(A,f)=0$ and $\operatorname{tr} A\geq 0$. □

Lemma 16. *When $d=2$, the operator H_ϵ satisfies the Cordes condition.*

Proof. Let eigenvalue of $B\in\mathbf{S}_1$ to be $\lambda, 1-\lambda$. Then we have

$$\frac{|\epsilon I+B|^2}{(\operatorname{tr}\epsilon I+\operatorname{tr} B)^2}=\frac{2\epsilon^2+2\epsilon+\lambda^2+(1-\lambda)^2}{4\epsilon^2+4\epsilon+1}\leq\frac{2\epsilon^2+2\epsilon+1}{4\epsilon^2+4\epsilon+1}=\frac{1}{1+\frac{2\epsilon^2+2\epsilon}{2\epsilon^2+2\epsilon+1}}.$$

Since $0<\frac{2\epsilon^2+2\epsilon}{2\epsilon^2+2\epsilon+1}<1$ is fixed for any λ, (5.2.1) satisfies the Cordes condition. □

Remark 9. *This methodology does not work for the case $d\geq 3$. For example, consider for some $\frac{1}{3}>\epsilon>0$ and $B=\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}\in\mathbf{S}_1$. Then $\frac{|\epsilon I+B|^2}{(\operatorname{tr}(\epsilon I+B))^2}=\frac{(1+\epsilon)^2+2\epsilon^2}{(1+3\epsilon)^2}>\frac{1}{2}$, and thus the Cordes condition fails.*

Theorem 8. *There exist a unique H^2 strong solution to the regularized HJB problem (5.2.4).*

Proof. Since the Bellman operator with perturbation satisfies the Cordes condition when $d=2$, by Miranda-Talanti inequality, (5.2.4) has a unique strong solution u_ϵ. □

Theorem 9. *The unique H^2 regular strong solution u_ϵ to the regularized HJB problem (5.2.4) is also a strong solution to the regularized Monge-Ampère problem (5.2.3).*

Proof. By Theorem 8, we have a unique H^2 strong solution to (5.2.4). Hence, by Lemma 15 and similar argument of the proof of Theorem 7, u_ϵ is also a strong solution to (5.2.3). □

The converse is also true.

Theorem 10. *There exist a unique subharmonic H^2 strong solution u_ϵ to the regularized Monge-Ampère problem (5.2.3). And the unique subharmonic H^2 regular strong solution u_ϵ to the regularized Monge-Ampère problem (5.2.3) is also a strong solution to the regularized HJB problem (5.2.4).*

Proof. By Theorem 8 and Theorem 9, there exist a H^2 strong solution u_ϵ to the regularized Monge-Ampère problem (5.2.3). By Lemma 15, we have $\Delta u_\epsilon \geq 0$ *a.e.*, so u_ϵ is a subharmonic H^2 strong solution.

Suppose there exist another subharmonic H^2 regular strong solution $u_{\epsilon 2}$ to the regularized Monge-Ampère problem (5.2.3). Then by Lemma 15, $u_{\epsilon 2}$ satisfies (5.2.3) a.e., which implies that $u_{\epsilon 2}$ satisfies (5.2.4) a.e. Therefore $u_{\epsilon 2}$ is also a strong solution to (5.2.4). By Theorem 8, the strong solution to (5.2.4) is unique implies $u_\epsilon = u_{\epsilon 2}$. Thus, u_ϵ is unique. □

5.3 VISCOSITY SOLUTION

Since the existence of classical solutions strongly depends on both the regularity of the data and boundary $\partial\Omega$, the classical solution may not exist without enough regularity of the domain. Thus, the strong solution of Monge-Ampère problem does not exist in general. Instead, we can consider the weak solution theories and look

for weak derivatives that are not defined in pointwise sense. Therefore, we will solve the Monge-Ampère problem for the weak soluion.

Due to no variational formulation for the fully nonlinear second order PDEs, we can not use integration by parts to shift the derivative from solution to test functions. Hence, we have the following non-variational concept of viscosity solution.

Consider the second-order fully nonlinear PDE of the form

$$F(D^2u, Du, u, x) = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega, \tag{5.3.1}$$

where $F : \mathbf{S} \times \mathbb{R}^d \times \mathbb{R} \times \Omega \to \mathbb{R}$.

Definition 4. *We say that the operator F is degenerate elliptic if $F(X, p, r, x) \leq F(Y, p, r, x)$ whenever $Y \leq X$, $X, Y \in \mathbf{S}$.*

Definition 5. *We say a function $f : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$ is upper (resp. lower) semicontinuous at $x_0 \in \Omega$ if and only if either $f(x_0) = \infty$ (resp. $f(x_0) = -\infty$) or for any $\epsilon > 0$, there exist a neighborhood U of x_0 such that $f(x) \leq f(x_0)+\epsilon$ (resp. $f(x) \geq f(x_0) - \epsilon$) for all $x \in U$. We denote the space of upper semicontinuous functions by $USC(\Omega) := \{f : f \text{ is upper semicontinuous at all } x \in \Omega\}$ and the space of lower semicontinuous functions by $LSC(\Omega) := \{f : f \text{ is lower semicontinuous at all } x \in \Omega\}$.*

Then we have the following definition of viscosity solution from [6].

Definition 6. *Let F be degenerate elliptic. A function $u \in USC(\Omega)$ (resp., $u \in LSC(\Omega)$) is called a viscosity subsolution (resp. viscosity supersolution) of (5.3.1) if for all $\varphi \in C^2(\Omega)$ such that $u - \varphi$ has a local maximum (resp., minimum) at $x \in \Omega$ we have*

$$F(D^2\varphi, D\varphi, \varphi, x) \leq 0 \quad (resp., F(D^2\varphi, D\varphi, \varphi, x) \geq 0),$$

and

$$u \leq g \text{ on } \partial\Omega \quad (resp., u \geq g \text{ on } \partial\Omega).$$

The function u is said to be a viscosity solution if it is simultaneously a viscosity subsolution and viscosity supersolution.

Since the Monge-Ampère equation is degenerate elliptic only on the set of convex functions, we need to restrict u and φ to be convex functions. Thus we have the following definition.

Definition 7. *Let F be degenerate elliptic. A convex function $u \in USC(\Omega)$ (resp., $u \in LSC(\Omega)$) is called a viscosity subsolution (resp. viscosity supersolution) of (5.3.1) on the set of convex functions if for all convex function $\varphi \in C^2(\Omega)$ such that $u - \varphi$ has a local maximum (resp., minimum) at $x \in \Omega$ we have*

$$F(D^2\varphi, D\varphi, \varphi, x) \leq 0 \quad (resp., F(D^2\varphi, D\varphi, \varphi, x) \geq 0),$$

and

$$u \leq g \text{ on } \partial\Omega \quad (resp., u \geq g \text{ on } \partial\Omega).$$

The function u is said to be a viscosity solution on the set of convex functions if it is simultaneously a viscosity subsolution and viscosity supersolution on the set of convex functions.

For the Monge-Ampère problem, we have $F(D^2u, Du, u, x) = M(D^2u, f)$. And for Bellman problem, we have $F(D^2u, Du, u, x) = H(D^2u, f)$.

Then by the theorem and proposition from [18], we have the following theorem.

Theorem 11. *Let $\Omega \subset \mathbb{R}^n$ be a bounded convex domain and $f \in C(\bar{\Omega})$ is positive in Ω. A function $u \in C(\Omega)$ is a viscosity solution of the Dirichlet problem $\det D^2u = f$ in Ω, $u = g$ on $\partial\Omega$ if and only if $g \in C(\partial\Omega)$ can be extended to a function $\tilde{g} \in C(\bar{\Omega})$ that is convex in Ω.*

Thus, if $f > 0$ in Ω and g can be extended to a convex function in Ω, there will exist a unique viscosity solution to the Monge-Ampère problem. And by [10], we have this theorem.

Theorem 12. *Let $f \in C(\Omega)$ be nonnegtive, then u is a viscosity solution of HJB problem (5.1.4) if and only if u is a viscosity solution of Monge-Ampère problem (5.1.3) on the set of convex functions.*

Therefore, the existence of a unique viscosity solution to the Monge-Ampère problem (5.1.3) implies that there exists a unique viscosity solution of Bellman problem (5.1.4). To compute this viscosity solution, we will use the viscosity solution u_ϵ of the regularized HJB problem (5.2.4) as an approximation. First we need to show the uniqueness and existence of the viscosity solution for (5.2.4). For uniqueness, we need the operator to satisfy the comparison principle.

Definition 8. *We say that the nonlinear PDE problem (5.3.1) satisfies the comparison principle if u_* is a viscosity subsolution of (5.3.1) and u^* is a viscosity supersolution of (5.3.1), then $u_* \leq u^*$.*

Lemma 17. *If (5.3.1) satisfies the comparison principle, then there exist at most one viscosity solution to (5.3.1).*

Proof. Suppose there exist two viscosity solutions u, v to (5.3.1). Then by Lemma 17, since u is a viscosity subsolution and v is a viscosity supersolution there holds $u \leq v$. And v is viscosity subsolution and u is viscosity supersolution imply that $v \leq u$. Thus, $u = v$ and so viscosity solutions are unique. □

To show H_ϵ that satisfies the comparison principle, we use some sufficient conditions for the existence of comparison principle. The following theorem is from [26, Theorem 2.71].

Theorem 13. *The Dirichlet problem satisfies comparison principle if the operator F is degenerate elliptic, independent of r and p variables and satisfies, for some $\gamma > 0$,*

$$F(x, M) \geq F(x, M + tI) + \gamma t, \ \forall t \geq 0.$$

We apply Theorem 13 to the HJB problem with perturbation. First we show that this problem is degenerate elliptic.

Lemma 18. *H_ϵ is degenerate elliptic operator, i.e., for any $Y \leq X$, we have $H_\epsilon(X, f) \leq H_\epsilon(Y, f)$.*

Proof. For any $0 \leq B \in \mathbf{S}_1$, $\epsilon > 0$ and $Y \leq X$, we have

$$\begin{aligned} -(\epsilon I + B) : X &\leq -(\epsilon I + B) : Y, \\ -\epsilon \mathrm{tr}\, X - B : X + f \sqrt[d]{det(B)} &\leq -\epsilon \mathrm{tr}\, Y - B : Y + f \sqrt[d]{det(B)}, \\ H_\epsilon(X, f) &\leq H_\epsilon(Y, f). \end{aligned}$$

Thus H_ϵ is degenerate elliptic. □

By Theorem 13 and Lemma 18, we have the following result.

Lemma 19. *The HJB problem with perturbation (5.2.4) satisfies the comparison principle.*

Proof. By Lemma 18, H_ϵ is a degenerate elliptic operator. For the HJB problem with perturbation, the nonlinear operator is H_ϵ. For any matrix M, we have

$$\begin{aligned} H_\epsilon(x, M + tI) &= \sup_{B \in \mathbf{S}_1} (-(B + \epsilon I) : (M + tI) + f \sqrt[d]{det(B)}), \\ &= \sup_{B \in \mathbf{S}_1} (-(B + \epsilon I) : M + f \sqrt[d]{det(B)}) - (1 + \epsilon)t, \\ &= H_\epsilon(x, M) - (1 + \epsilon)t. \end{aligned}$$

Thus $H_\epsilon(x, M) \geq H_\epsilon(x, M + tI) + (1 + \epsilon)t$, and by Theorem 13, (5.2.4) satisfies the comparison principle. □

Thus, by Lemma 17, we have at most one viscosity solution of (5.2.4); uniqueness is proved. For the existence of the viscosity solution, first we need an extra condition for the operator.

Definition 9. *The operator F is called uniformly elliptic if for any matrix $P \geq 0$, there exist $0 < \mu \leq \nu$ such that*

$$F(x, r, p, X) - \nu \text{tr}\, P \leq F(x, r, p, X + P) \leq F(x, r, p, X) - \mu \text{tr}\, P.$$

Lemma 20. *The operator H_ϵ is uniformly elliptic.*

Proof. For any $B \in \mathbf{S}_1$, we have $0 \leq B \leq I$. Then $P \geq 0$ implies that $0 \leq B : P \leq \text{tr}\, P$. Thus,

$$\begin{aligned}
H_\epsilon(x, X + P) &= \sup_{B \in \mathbf{S}_1} (-(B + \epsilon I) : (X + P) + f \sqrt[d]{det(B)}), \\
&= \sup_{B \in \mathbf{S}_1} (-B : X - B : P + f \sqrt[d]{det(B)}) - \epsilon \text{tr}\, P, \\
&\Rightarrow \sup_{B \in \mathbf{S}_1} (-B : X + f \sqrt[d]{det(B)}) - (1 + \epsilon) \text{tr}\, P \\
&\leq H_\epsilon(x, X + P) \\
&\leq \sup_{B \in \mathbf{S}_1} (-B : X + f \sqrt[d]{det(B)}) - \epsilon \text{tr}\, P \\
\Rightarrow H_\epsilon(x, X) - (1 + \epsilon) \text{tr}\, P \leq & H_\epsilon(x, X + P) \leq H_\epsilon(x, X) - \epsilon \text{tr}\, P.
\end{aligned}$$

Therefore, with $\nu = 1 + \epsilon, \mu = \epsilon$, H_ϵ is uniformly elliptic. □

We can define the Pucci operator and a more general structure condition.

Definition 10. *We define the Pucci extremal operators as the following:*

$$\mathcal{P}^+(X) = -\mu \mathrm{tr}\,(X^+) + \nu \mathrm{tr}\,(X^-),\ \mathcal{P}^-(X) = -\nu \mathrm{tr}\,(X^+) + \mu \mathrm{tr}\,(X^-),$$

where $X = X^+ - X^-$, $X^+, X^- \geq 0$ are the positive and negative parts of X. And we say the operator F satisfies structure condition if

$$\mathcal{P}^-(X - Y) \leq F(x, X) - F(x, Y) \leq \mathcal{P}^+(X - Y),$$

for any $X, Y \in \mathbf{S}$.

Lemma 21. *The operator H_ϵ satisfies structure condition.*

Proof. By similar argument of Lemma 11, we can assume X, Y are diagonal matrices without loss of generality. Since H_ϵ uniformly elliptic, $X - Y = (X-Y)^+ - (X-Y)^-$ and $(X-Y)^+, (X-Y)^- \geq 0$, there holds

$$\begin{aligned}
-\nu \mathrm{tr}\,(X-Y)^- \leq H_\epsilon(x, Y + (X-Y)^+) - H_\epsilon(x, X) \leq -\mu \mathrm{tr}\,(X-Y)^-,\\
-\nu \mathrm{tr}\,(X-Y)^+ \leq H_\epsilon(x, Y + (X-Y)^+) - H_\epsilon(x, Y) \leq -\mu \mathrm{tr}\,(X-Y)^+.
\end{aligned}$$

Thus, combine the above two inequality, we have

$$\begin{aligned}
-\nu \mathrm{tr}\,(X-Y)^+ + \mu \mathrm{tr}\,(X-Y)^- \leq & H_\epsilon(x, X) - H_\epsilon(x, Y)\\
\leq & -\mu \mathrm{tr}\,(X-Y)^+ + \nu \mathrm{tr}\,(X-Y)^-,\\
\Rightarrow \mathcal{P}^-(X-Y) \leq & H_\epsilon(x, X) - H_\epsilon(x, Y) \leq \mathcal{P}^+(X-Y).
\end{aligned}$$

Therefore, H_ϵ satisfies structure condition. □

By [3, Lemma 2.5], we have the following theorem.

Theorem 14. *Let $F(x, X)$ satisfies the structure condition. If u is a strong solution of $F = f$ on Ω, then u is a viscosity solution of $F = f$ on Ω. The converse is also true.*

Then we can prove the following theorem.

Theorem 15. *There exist unique viscosity solution for the HJB problem with perturbation* (5.2.4).

Proof. By Theorem 8, we have a unique strong solution to (5.2.4). Then by Lemma 21 and Theorem 14, the unique strong solution is a viscosity solution. We have the existence of viscosity solution. Therefore, together with the uniqueness of viscosity solution for (5.2.4), the unique strong solution of HJB problem with perturbation (5.2.4) is also the unique viscosity solution of (5.2.4). □

5.4 A PRIORI ESTIMATES

In this section we are looking for the priori estimate for the viscosity solution u_ϵ to (5.2.4) in the case $d = 2$.

Lemma 22. *Let $g = 0$ and u_ϵ be the unique viscosity solution to* (5.2.4)*, we have* $\|\Delta u_\epsilon\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{8\epsilon}}\|f\|_{L^2(\Omega)}$.

Proof. By u_ϵ is also a strong solution to (5.2.4) and Theorem 9, we have that $M_\epsilon(D^2u_\epsilon, f) = 0$ a.e. in Ω and $u_\epsilon \in H^2(\Omega) \cap H_0^1(\Omega)$. So u_ϵ satisfies Miranda-Talenti estimate, $|u_\epsilon|_{H^2(\Omega)} \leq \|\Delta u_\epsilon\|_{L^2(\Omega)}$. By $d = 2$, we have

$$det(D^2u_\epsilon) = \frac{1}{2}|\Delta u_\epsilon|^2 - \frac{1}{2}|D^2u_\epsilon|^2 \Rightarrow \int_\Omega det(D^2u_\epsilon) = \frac{1}{2}(\|\Delta u_\epsilon\|^2_{L^2(\Omega)} - |u_\epsilon|^2_{H^2(\Omega)}) \geq 0.$$

Thus, by $M_\epsilon(D^2u_\epsilon, f) = 0$ a.e. in Ω, for small ϵ, we have

$$\begin{aligned}
&(\epsilon^2+\epsilon)|\Delta u_\epsilon|^2 + det(D^2u_\epsilon) = \left(\frac{f^2}{4}\right)\\
\Rightarrow &\int_\Omega \left(\frac{f^2}{4}\right) \geq \int_\Omega (\epsilon^2+\epsilon)|\Delta u_\epsilon|^2\\
\Rightarrow &\|\Delta u_\epsilon\|^2_{L^2(\Omega)} \leq \frac{1}{4(\epsilon^2+\epsilon)}\|f\|^2_{L^2(\Omega)} \leq \frac{1}{8\epsilon}\|f\|^2_{L^2(\Omega)}\\
\Rightarrow &\|\Delta u_\epsilon\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{8\epsilon}}\|f\|_{L^2(\Omega)}.
\end{aligned}$$

□

Lemma 23. *Let $g = 0$, u_ϵ be the unique viscosity solution to (5.2.4) and u be the unique convex viscosity solution to (5.1.4), we have $u \leq u_\epsilon$ in Ω.*

Proof. Suppose that for some $\phi \in C^2(\Omega)$, $u - \phi$ has a maximum at some $x_0 \in \Omega$. Since u is a viscosity solution to (5.1.4), we have $H(D^2\phi(x_0), f(x_0)) \leq 0$. And by [17, Remark 1.3.2], u is convex implies $D^2\phi(x_0) \geq 0$. Thus, we have

$$H_\epsilon(D^2\phi(x_0), f(x_0)) = -\epsilon\Delta\phi(x_0) + H(D^2\phi(x_0), f(x_0)) \leq 0.$$

Hence, u is a viscosity subsolution to (5.2.4). By Lemma 19, H_ϵ satisfies comparison principle so we have $u \leq u_\epsilon$. □

Lemma 24. *Let $g = 0$, $u_{\epsilon_1}, u_{\epsilon_2}$ be the unique subharmonic viscosity solutions to (5.2.4) with $\epsilon = \epsilon_1, \epsilon_2$ and $\epsilon_1 < \epsilon_2$, then we have $u_{\epsilon_1} \leq u_{\epsilon_2}$ in Ω.*

Proof. Similar to the proof of Lemma 23, we have $H_{\epsilon_1}(D^2\phi(x_0), f(x_0)) \leq 0$. And u_{ϵ_1} is subharmonic implies $\Delta\phi(x_0) \geq 0$. Thus, we have

$$H_{\epsilon_2}(D^2\phi(x_0), f(x_0)) = -(\epsilon_2 - \epsilon_1)\Delta\phi(x_0) + H_{\epsilon_1}(D^2\phi(x_0), f(x_0)) \leq 0.$$

Hence, u_{ϵ_1} is a viscosity subsolution to (5.2.4) with $\epsilon = \epsilon_2$. By Lemma 19, H_{ϵ_2} satisfies comparison principle so we have $u_{\epsilon_1} \leq u_{\epsilon_2}$. □

Remark 10. *Since u is convex, by maximum principle, we have* $\max_{x\in\Omega} u(x) = \max_{x\in\partial\Omega} u(x) = 0$, $u \leq 0$. *Similar for u_ϵ is subharmonic, $u_\epsilon \leq 0$. Thus $u \leq u_\epsilon \leq 0$ implies $\|u_\epsilon\|_{L^\infty(\Omega)} \leq \|u\|_{L^\infty(\Omega)}$. $\{u_\epsilon\}$ is uniformly bounded.*

A good property of the viscosity solution is its stability. We have the following theorem from [19, Theorem 3.2].

Theorem 16. *Let $\{F_k\}_{k\in\mathbb{N}}$ be a sequence of uniformly degenerate elliptic operators and let $\{u_k\}_{k\in\mathbb{N}} \subset C(\Omega)$ be, for each k, viscosity subsolutions to the equations*

$$F_k(D^2u_k, Du_k, u_k, x) = 0, \quad \text{in } \Omega.$$

If $F_k \to F$ uniformly on compact subsets of $\mathbf{S} \times \mathbb{R}^d \times \mathbb{R} \times \Omega$ and $u_k \to u$ uniformly in compact subsets of Ω as $k \to \infty$, then u is a viscosity subsolution of

$$F\left(D^2u, Du, u, x\right) = 0, \text{in } \Omega.$$

Hence we can use the strong solution of (5.2.4) to approach the viscosity solution of (5.1.3).

Suppose that $\{u_\epsilon\}$ is uniformly bounded and satisfies Hölder continuity of order α, $0 < \alpha \leq 1$ with a fixed constant M.

$$|u_\epsilon(x) - u_\epsilon(y)| \leq M|x - y|^\alpha, x, y \in \Omega.$$

This result will be the topic of future research.

If we have the Hölder continuity for u_ϵ, then there exist a uniformly convergent subsequence of $\{u_\epsilon\}$.

Definition 11. *Let F be a family of functions on Ω. We say that F is equicontinuous if for every $\epsilon > 0$ and $x \in \Omega$, there is $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for all $f \in F$ when $|x - y| < \delta$. We say that F is uniformly equicontinuous if δ does not depends on x.*

Under the given assumptions, $\{u_\epsilon\}$ is uniformly equicontinuous by setting $\delta < (\frac{\epsilon}{M})^{-\alpha}$, thus $|u_\epsilon(x) - u_\epsilon(y)| \leq M|x - y|^\alpha < \epsilon,\ for\ all\ |x - y| < \delta$.

Theorem 17 (Arzelà-Ascoli theorem)**.** *Let F be a family of continuous functions on Ω. If F is uniformly bounded and equicontinuous, then there exists a subsequence $\{f_{n_k}\}_{k\in\mathbb{N}}$ that converges uniformly.*

Proof. Fix an enumeration $\{x_i\}_{i\in\mathbb{N}}$ where x_i are all the $x \in \Omega$ with rational coordinate. Since F is uniformly bounded, the set of points $\{f(x_1)\}_{f\in F}$ are bounded. By BolzanoWeierstrass theorem, there is a sequence $\{f_{n_1}\}$ of distinct functions in F such that $\{f_{n_1}(x_1)\}$ converges. Repeat the same argument for x_2, there is a subsequence $\{f_{n_2}\}$ of $\{f_{n_1}\}$ such that $\{f_{n_2}(x_2)\}$ converges. By induction, we have $\{f_{n_1}\} \supseteq \{f_{n_2}\} \supseteq ...$ and for each $k \in \mathbb{N}$, $\{f_{n_k}\}$ converges at $x_1, x_2, ..., x_k$. Then form a new subsequence $\{f\}$ such that $f_m = f_{n_{mm}}$, the mth term of $\{f\}$ is the mth term of $\{f_{n_m}\}$. Thus, $\{f\}$ converge at all $\{x_i\}$. Therefore, for any $\epsilon > 0$ and x_i, there exist an integer $N_i = N(\epsilon, x_i)$ such that for all $n, m > N_i$ we have $|f_n(x_i) - f_m(x_i)| < \frac{\epsilon}{3}$. Since F is equicontinuous, for every x_i we have a open ball $B_i = B_{x_i}(\delta_i)$ such that $|f(s) - f(t)| < \frac{\epsilon}{3}$ for all $f \in F$ and $s, t \in B_i$. $\cup B_i$ forms an open cover of Ω, Ω is compact implies it has only finite subcover B_{i_n}, $B_{i_1}, B_{i_2}, ..., B_{i_J}$. Thus for any $x \in \Omega$, there exist k such that x, x_k are in the same B_{i_k}. Hence, we have

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f_n(x_k)| + |f_n(x_k) - f_m(x_k)| + |f_m(x) - f_m(x_k)| \leq \epsilon$$

for all $n, m > N = max_{1\leq n\leq J}\{N_{i_n}\}$. Therefore, $\{f\}$ converges uniformly. □

Thus, by Theorem 17 and if $\{u_\epsilon\}$ is uniformly bounded and equicontinuous, there exists a sub sequence of $\{u_\epsilon\}$ converges uniformly to a function u.

Lemma 25. *Suppose that $\{u_\epsilon\}$ is uniformly Holder continuous with respect to ϵ. Then $\{u_\epsilon\}$ converges uniformly to u as $\epsilon \to 0$.*

Proof. $\{u_{\frac{1}{n}}\}$ is a sub sequence of $\{u_\epsilon\}$, by similar argument in the proof of Theorem 17, $\{u_{\frac{1}{n}}\}$ has a uniformly convergent subsequence $\{u_{\frac{1}{n_k}}\}$. Hence for any $\epsilon' > 0$, there exist N such that for any $p > q > N$, we have $|u_{\frac{1}{n_p}} - u_{\frac{1}{n_q}}| < \epsilon'$. Then for any $0 < \epsilon_1 < \epsilon_2 < \frac{1}{n_{N+1}}$, there exist $p > q > N$ such that

$$\frac{1}{n_p} < \epsilon_1 < \epsilon_2 < \frac{1}{n_p}.$$

Hence, by Lemma 24, we have

$$u_{\frac{1}{n_p}} \leq u_{\epsilon_1} \leq u_{\epsilon_2} \leq u_{\frac{1}{n_q}}.$$

Therefore we have $|u_{\epsilon_1} - u_{\epsilon_2}| \leq |u_{\frac{1}{n_p}} - u_{\frac{1}{n_q}}| < \epsilon'$, $\{u_\epsilon\}$ converges uniformly to some function u. And by Theorem 16, u is the solution to (5.0.1). □

5.5 FINITE ELEMENT METHOD

Consider the case for $d = 2$. To solve the Monge-Ampère problem, we can solve the Bellman problem with perturbation instead. We denote the set of all $d \times d$ orthonormal matrices by $O(d)$ and set $\Lambda := [0, 1] \times O(d)$. The set Λ is compact. Let $\alpha := (\lambda, Q) \in \Lambda$ and $D = \begin{bmatrix} \lambda & 0 \\ 0 & 1-\lambda \end{bmatrix}$. Equation (5.2.4) can be rewritten as

$$\inf_{\alpha \in \Lambda} \left(L^\alpha D^2 u(x) - f^\alpha(x) \right) = 0 \qquad \forall x \text{ in } \Omega \tag{5.5.1a}$$

$$u(x) = g(x) \quad \forall x \ in \ \partial\Omega, \tag{5.5.1b}$$

where $L^\alpha A = (Q(\epsilon I_d + D)Q^T) : A$, $f^\alpha = f(x)\sqrt[d]{det(QDQ^T)}$.

Thus, the method we developed from the Discrete Miranda-Talanti inequality can be used to solve this problem. Define the operator $F(u) := \inf_{\alpha\in\Lambda'}(L^\alpha D^2 u(x) - f^\alpha(x))$, the function $\gamma^\alpha := \frac{1}{\lambda^2+(1-\lambda)^2}$, and the operator $F_\gamma(u) := \inf_{\alpha\in\Lambda'} \gamma^\alpha\big(\mathcal{L}^\alpha u - f^\alpha\big)$. Let $\mathcal{T}_h$ be a shape-regular mesh of Ω, $\mathcal{F}_h^I$ be the set of all interior faces, set the form

$$B(u,v) := \sum_{T\in\mathcal{T}_h}\int_T F_\gamma(u)\Delta v\,dx + \sigma\sum_{f\in\mathcal{F}_h^I} h_f^{-1}\int_f [\![\partial u/\partial n_f]\!][\![\partial v/\partial n_f]\!]\,ds.$$

Let V_h be the kth degree Lagrange finite element space, then the finite element method is to find $u_h \in V_h$ such that $B(u_h, v_h) = 0$ for all $v_h \in V_h$ and let $\epsilon \to 0$. We will use the following algorithm 2 to find the solution.

By Lemma 16, L^α satisfies the Cordes condition, thus there exist a unique strong solution $u_\epsilon \in H^2(\Omega)$ to (5.5.1a). Therefore, since u_ϵ satisfies (5.5.1a) almost everywhere in Ω, we conclude that $B(u_\epsilon, v_h) = 0$ for all $v_h \in V_h$. When $\epsilon \to 0$, the result of $u_\epsilon \to u$ is not clear. By Theorem 6, there exist a unique solution u_h and $\|u_\epsilon - u_h\|_h \le C\sum_{T\in\mathcal{T}_h} h_T^{2s-4}|u_\epsilon|_{H^s(T)}$ provided that $u_\epsilon \in H^s(\Omega)$ for some $2 \le s \le k+1$.

5.6 NUMERICAL EXPERIMENT

5.6.1 Test 5

In this experiment, we solve the Monge-Ampère problem (5.0.1) with $d = 2$, $\Omega = [0,1]^2$. The source functions f are chosen so that the solution of (5.0.1) is

$$u(x_1, x_2) = \frac{4}{7}(x_1^2 + x_2^2)^{\frac{7}{4}}. \tag{5.6.1}$$

Algorithm 2 Solve the Bellman problem with perturbation

k=0
Choose a initial $\alpha_0 \in \Lambda$
for $k \leq Maxiteration$ **do**
    Find u_k solves $L^{\alpha_0} u_k - f^{\alpha_0} = 0$
    Find $Q(x) \in O(2)$ such that Q diagonalize $D^2 u_k$
    Find $\lambda(x) = \arg\min_{\lambda \in [0,1]} (Q(\epsilon I_d + D)Q^T) : D^2 u_k(x) - f(x) \sqrt[d]{det(QDQ^T)}$
    $\alpha_{k+1} = (\lambda, Q)$
    if $k \geq 1$ **then**
        if $\|u_k - u_{k-1}\|_{L^2(\Omega)} \leq tol$ **then**
            Break
        end if
    end if
    $k = k + 1$
end for

And we assume that $\epsilon = h^2$. The plots of the errors given in Figure 6 show that the method converges as the mesh is refined, at optimal rate for H^1 and h-norm error when $k = 2$ and at sub-optimal rates for L^2 norm. When $k = 3$, only h-norm error reaches optimal convergence rate. Namely, the numerical experiments indicate that the method converges with at least the following convergence rates:

$$\|u - u_h\|_{L^2(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h^{k-1}), \quad \|u - u_h\|_h = \mathcal{O}(h^{k-1}).$$

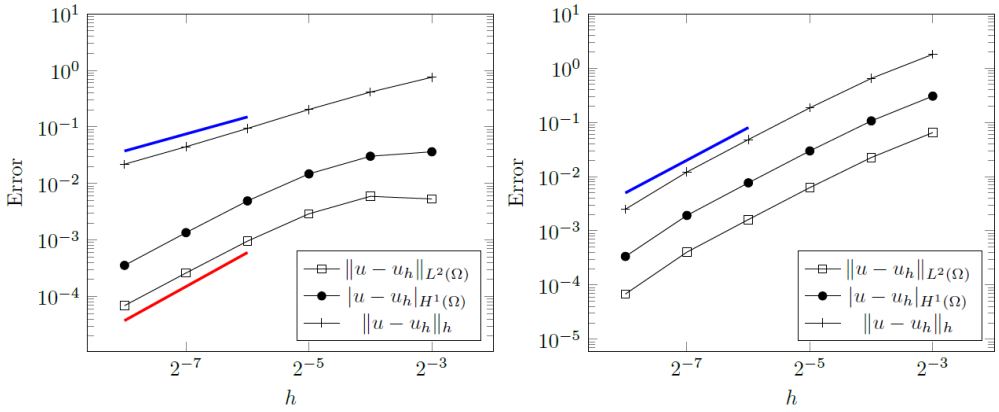

Figure 6: Test 5: Convergence plot of the two-dimensional nonlinear problem with $k = 2$ (left) and $k = 3$ (right). The red reference line has slope k and the blue reference line has slope $(k-1)$.

5.6.2 Test 6

In this experiment, we solve the Monge-Ampère problem (5.0.1) with $d = 2$, $\Omega = [0,1]^2$. The source functions f are chosen so that the solution of (5.0.1) is

$$u(x_1, x_2) = \frac{4}{3}(x_1^2 + x_2^2)^{\frac{3}{4}}. \tag{5.6.2}$$

And we assume that $\epsilon = h^2$. The plots of the errors given in Figure 7 show that the method converges as the mesh is refined, at rate 0.5 for h-norm error for both $k = 2, 3$.

$$\|u - u_h\|_h = \mathcal{O}(h^{0.5}).$$

This result is understandable since the true solution is $u \in H^{2.5}(\Omega)$, thus the best approximation for h-norm is at rate 0.5.

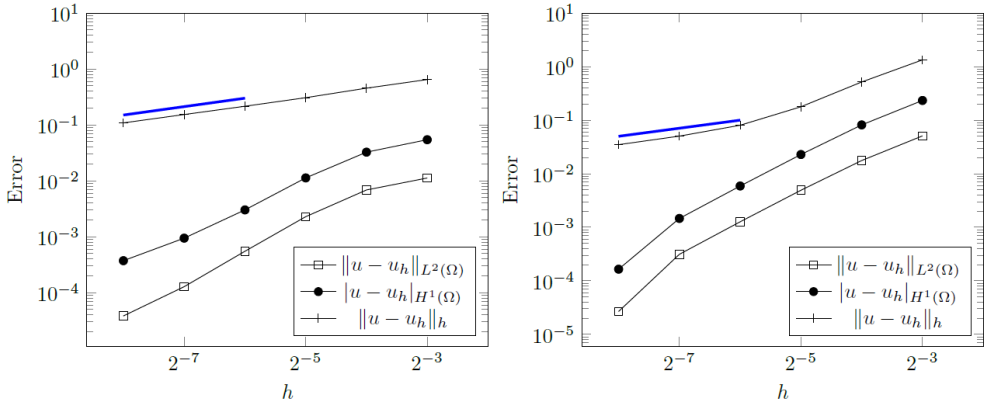

Figure 7: Test 6: Convergence plot of the two-dimensional nonlinear problem with $k = 2$ (left) and $k = 3$ (right). The blue reference line has slope 0.5.

6.0 CONCLUSION AND FUTURE RESEARCH

In my thesis, we introduce a new inequality, Discrete Miranda-Talanti inequality, which is the extension of Miranda-Talanti inequality. The proof is done by introducing an enriching operator from Lagrange element space to the Clough-Tocher finite element space. By applying the Discrete Miranda-Talanti inequality, we develop finite element methods to solve linear and nonlinear PDEs in non-divergence form. Our method has relatively less extra terms than previous method and is easy to implement with numerical software. We test our method with several numerical test for both linear and nonlinear case, the result matches our expectation.

We further study the Monge-Ampère problem, a special nonlinear PDE form. We find the Hamilton-Jacobi-Bellman representation of the Monge-Ampère problem and then introduce a perturbation term ϵ to ensure the Cordes condition is satisfied. A generalized type of solution, viscosity solution, is used for solving the Monge-Ampère problem, and we prove the equivalence of viscosity solution between Monge-Ampère problem and Hamilton-Jacobi-Bellman representation with or without the perturbation. For a priori estimate, we assume that $\{u_\epsilon\}$ satisfied Hölder continuity. The proof of this Hölder continuity is to be done in future research. And with this assumption, the strong solutions of Hamilton-Jacobi-Bellman representation with perturbation will converge uniformly to the viscosity solution of the Monge-Ampère

problem. The error estimate is also for the future research. Finally, we develop a new finite element method to solve the Monge-Ampère problem and do a few test for it. The result reaches optimal convergence rate for h-norm error.

My future research is about proving the Hölder continuity and develop error estimate with ϵ and h.

BIBLIOGRAPHY

[1] Susanne C. Brenner, Thirupathi Gudi, and Li-yeng Sung. An a posteriori error estimator for a quadratic C^0-interior penalty method for the biharmonic problem. *IMA J. Numer. Anal.*, 30(3):777–798, 2010.

[2] Susanne C. Brenner and Li-yeng Sung. Virtual enriching operators. 2019.

[3] L. Caffarelli, M. G. Crandall, M. Kocan, and A. wich. On viscosity solutions of fully nonlinear equations with measurable ingredients. *Communications on Pure and Applied Mathematics*, 49(4):365–398, 1996.

[4] Filippo Chiarenza, Michele Frasca, and Placido Longo. Interior $W^{2,p}$ estimates for nondivergence elliptic equations with discontinuous coefficients. *Ricerche Mat.*, 40(1):149–168, 1991.

[5] Philippe G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

[6] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions. User's guide to viscosity solutions of second order partial differential equations. 1992.

[7] Jim Douglas, Jr., Todd Dupont, Peter Percell, and Ridgway Scott. A family of C^1 finite elements with optimal approximation properties for various Galerkin methods for 2nd and 4th order problems. *RAIRO Anal. Numér.*, 13(3):227–255, 1979.

[8] Xiaobing Feng, Roland Glowinski, and Michael Neilan. Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Review*, 55(2):205–267, 2013.

[9] Xiaobing Feng, Lauren Hennings, and Michael Neilan. Finite element methods for second order linear elliptic partial differential equations in non-divergence form. *Math. Comp.*, 86(307):2025–2051, 2017.

[10] Xiaobing Feng and Max Jensen. Convergent semi-lagrangian methods for the monge-ampère equation on unstructured grids. 2016.

[11] Xiaobing Feng, Michael Neilan, and Stefan Schnake. Interior penalty discontinuous galerkin methods for second order linear non-divergence form elliptic pdes. *Journal of Scientific Computing*, 74(3):1651–1676, 2018.

[12] Dietmar Gallistl. Variational formulation and numerical analysis of linear elliptic equations in nondivergence form with Cordes coefficients. *SIAM J. Numer. Anal.*, 55(2):737–757, 2017.

[13] Emmanuil H. Georgoulis, Paul Houston, and Juha Virtanen. An *a posteriori* error indicator for discontinuous Galerkin approximations of fourth-order elliptic problems. *IMA J. Numer. Anal.*, 31(1):281–298, 2011.

[14] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, New York;Berlin;, reprint of the 2nd, rev. 3rd printing. edition, 1998.

[15] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[16] Pierre Grisvard. *Elliptic problems in nonsmooth domains*, volume 69 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Reprint of the 1985 original [MR0775683], With a foreword by Susanne C. Brenner.

[17] Cristian E. Gutirrez and SpringerLink (Online service). *The Monge-Ampre Equation*, volume 89. Springer International Publishing, Cham, 2nd 2016. edition, 2016.

[18] David Hartenstine. The dirichlet problem for the monge-ampere equation in convex (but not strictly convex) domains. *Electronic Journal of Differential Equations*, 2006(138):1–9, 2006.

[19] Nikos Katzourakis and SpringerLink (Online service). *An Introduction To Viscosity Solutions for Fully Nonlinear PDE with Applications to Calculus of Variations in L*. Springer International Publishing, Cham, 2015 edition, 2015;2014;.

[20] Ellya L. Kawecki. A dgfem for nondivergence form elliptic equations with cordes coefficients on curved domains. *Numerical Methods for Partial Differential Equations*, 2019.

[21] N.V. Krylov. *Nonlinear Elliptic and Parabolic Equations of the Second Order*. Mathematics and its Applications. Springer Netherlands, 1987.

[22] Antonino Maugeri, Dian K. Palagachev, and Lubomira G. Softova. *Elliptic and parabolic equations with discontinuous coefficients*, volume 109 of *Mathematical Research*. Wiley-VCH Verlag Berlin GmbH, Berlin, 2000.

[23] Antonino Maugeri, Dian K. Palagachev, and Lubomira G. Softova. *Elliptic and parabolic equations with discontinuous coefficients*, volume 109 of *Mathematical Research*. Wiley-VCH Verlag Berlin GmbH, Berlin, 2000.

[24] Lin Mu and Xiu Ye. A simple finite element method for non-divergence form elliptic equations. *Int. J. Numer. Anal. Model.*, 14(2):306–311, 2017.

[25] Michael Neilan. Convergence analysis of a finite element method for second order non-variational elliptic problems. *J. Numer. Math.*, 2017. to appear.

[26] Michael Neilan, Abner J. Salgado, and Wujun Zhang. Numerical analysis of strongly nonlinear pdes. 2016.

[27] Ricardo H. Nochetto and Wujun Zhang. Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form. *Found. Comput. Math.*, 2017. to appear.

[28] Abner J. Salgado and Wujun Zhang. Finite element approximation of the isaacs equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 53(2):351–374, 2019.

[29] Iain Smears. Nonoverlapping domain decomposition preconditioners for discontinuous Galerkin approximations of Hamilton-Jacobi-Bellman equations. *J. Sci. Comput.*, 74(1):145–174, 2018.

[30] Iain Smears and Endre Süli. Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients. *SIAM J. Numer. Anal.*, 51(4):2088–2106, 2013.

[31] Iain Smears and Endre Süli. Discontinuous Galerkin finite element approximation of Hamilton-Jacobi-Bellman equations with Cordes coefficients. *SIAM J. Numer. Anal.*, 52(2):993–1016, 2014.

[32] Iain Smears and Endre Süli. Discontinuous Galerkin finite element methods for time-dependent Hamilton-Jacobi-Bellman equations with Cordes coefficients. *Numer. Math.*, 133(1):141–176, 2016.

[33] Tatyana Sorokina and Hal Schenck. Subdivision and spline spaces. *Constr. Approx.*, 2017. to appear.

[34] Chunmei Wang and Junping Wang. A primal-dual weak Galerkin finite element method for second order elliptic equations in non-divergence form. *Math. Comp.*, 2017. to appear.

[35] A. J. Worsey and G. Farin. An n-dimensional Clough-Tocher interpolant. *Constr. Approx.*, 3(2):99–110, 1987.