MATHEMATICS OF COMPUTATION
Volume 80, Number 276, October 2011, Pages 1979–1995
S 0025-5718(2011)02487-7
Article electronically published on March 9, 2011

# $\mathcal{C}^0$ PENALTY METHODS FOR THE FULLY NONLINEAR MONGE-AMPÈRE EQUATION

SUSANNE C. BRENNER, THIRUPATHI GUDI, MICHAEL NEILAN, AND LI-YENG SUNG

Abstract. In this paper, we develop and analyze $\mathcal{C}^0$ penalty methods for the fully nonlinear Monge-Ampère equation $\det(D^2u) = f$ in two dimensions. The key idea in designing our methods is to build discretizations such that the resulting discrete linearizations are symmetric, stable, and consistent with the continuous linearization. We are then able to show the well-posedness of the penalty method as well as quasi-optimal error estimates using the Banach fixed-point theorem as our main tool. Numerical experiments are presented which support the theoretical results.

## 1. Introduction

Consider the following boundary-value problem for the Monge-Ampère equation: [25, 23, 10]:

$$\det(D^2u) = f \qquad \text{in } \Omega, \tag{1.1a}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{1.1b}$$

Here, $\Omega \subset \mathbf{R}^2$ is a convex domain, $f$ is a strictly positive function on $\Omega$, and

$$\det(D^2u) = \frac{\partial^2 u}{\partial x_1^2}\frac{\partial^2 u}{\partial x_2^2} - \left(\frac{\partial^2 u}{\partial x_1 \partial x_2}\right)^2$$

denotes the determinant of the Hessian matrix $D^2u$.

The goal of this paper is to develop and analyze $\mathcal{C}^0$ penalty methods for classical solutions of the Monge-Ampère equation. In particular, we assume that (1.1) has a strictly convex solution $u \in H^s(\Omega)$ with $s > 3$. We will also assume that $\Omega$ is either a convex polygon or a smooth convex domain. In the case where $\Omega$ is smooth, the smoothness of $u$ follows from the smoothness of $f$ and $g$ by the results in Caffarelli-Nirenberg-Spruck [10]. In fact, in this case there are exactly two solutions: one being convex, the other concave [14]. The case of less smooth (viscosity) solutions as well as more general types of Monge-Ampère equations (where $f$ depends on $u$ and $\nabla u$) and the three-dimensional case will be addressed in future works.

Even for linear problems, the case of curved boundaries require special care in the finite element context. To cope with this difficulty, we use Nitsche's method

Received by the editor May 9, 2010 and, in revised form, September 1, 2010.
2010 *Mathematics Subject Classification.* Primary 65N30, 35J60.
*Key words and phrases.* Monge-Ampère equation, fully nonlinear PDEs, finite element method, convergence analysis.
The work of the first and fourth authors were supported in part by the National Science Foundation under Grants No. DMS-07-13835 and DMS-10-16332. The work of the third author was supported by the National Foundation under Grant No. DMS-09-02683.

[28] to enforce the Dirichlet boundary condition (1.1b). The crux of this method is that a weak form of the essential boundary condition (1.1b) is included into the variational formulation by penalization techniques.

Another hurdle (and the most obvious one) that must be overcome is the strong nonlinearity in the PDE (1.1a). In order to deal with this difficulty, in both the construction and analysis of the numerical scheme, we derive consistent discretizations of (1.1) such that the resulting (discrete) linearization is stable. This simple idea leads to not-so-obvious discretizations of (1.1). For example, the most natural finite element method using Nitsche's technique is to find a finite element function $u_h$ such that

$$\int_\Omega \big(f - \det(D_h^2 u_h)\big) v \, dx + \sigma \sum_{e \in \mathcal{E}_h^b} h_e^{-1} \int_e (u_h - g) v \, ds = 0$$

for all test functions $v$. However this naive method does not work either in practice or analysis, as the linearization of the discrete problem is not consistent with the linearization of the continuous problem. In the next section, we derive some consistent schemes which inherit a stable discrete linearization.

Once we are able to construct such methods, we carry out the numerical analysis using the Banach fixed-point theorem as our main tool. We prove optimal order error estimates in $H^1$-like and $H^2$-like norms. Furthermore, by way of a duality argument we are able to derive (almost) optimal order error estimates in the $L^2$ norm.

Due to their important role in many application areas [11, 31, 9] there has been a growing interest and a surge of papers in recent years towards developing numerical schemes for the Monge-Ampère equation. The first attempt was by Oliker and Prussner [30] who constructed schemes for computing the Aleksandrov measure induced by $D^2u$ and obtained the solution of problem (1.1) as a by-product. Using convexity arguments, they showed that their sequences converge monotonically to the solution, although no rates of convergence were provided. More recently, Oberman [29, 22] constructed a wide stencil difference scheme for nonlinear elliptic PDEs which can be written as functions of eigenvalues of the Hessian matrix (such as the Monge-Ampère equation). It was proved that the finite difference scheme satisfies the convergence criterion (consistency, stability, and monotonicity) established by Barles and Souganidis [2], although no rates of convergence were given. Other relevant finite difference schemes include [3] and [1].

The implementation and the convergence theory of finite element methods for the Monge-Ampère equation is less understood. Dean and Glowinski [15] presented an augmented Lagrange multiplier method and a least squares method for the Monge-Ampère equation by treating the nonlinear equations as a constraint and using a variational principle to select a particular solution. The convergence of their method still remains an open problem. Böhmer [5] introduced a projection method using $\mathcal{C}^1$ finite element functions for a certain class of fully nonlinear second order elliptic PDEs and analyzes the methods using consistency and stability arguments. Finally, Feng and the third author considered fourth order singular perturbations of (1.1) by adding a small multiple of the biharmonic operator to the PDE (1.1a) [18]. Many numerical methods for the regularized problem were proposed in [19, 20, 27].

The finite element method considered here is a projection-type method as is the method proposed in [5]. But we use the standard Lagrange finite elements

which were originally designed for second order linear problems. Advantages of our method in comparison to the aforementioned works include:

- Lagrange elements are simple to use and are available on practically all finite element commercial software.
- The method has a similar number of degrees of freedom compared to non-conforming finite element methods [27], but is easier to implement.
- The method has a small number of degrees of freedom in comparison to $\mathcal{C}^1$ finite element methods [5, 20] and mixed finite element methods [19].
- Curved boundaries can be handled easily in comparison to finite difference methods [29, 30, 3].
- We can handle general Monge-Ampère equations, where the function $f$ depends on both $u$ and its gradient.
- The method and analysis can be extended to the three-dimensional setting [8].

The rest of the paper is organized as follows. In Section 2, we set the notation and give the motivation behind the $\mathcal{C}^0$ penalty method. We end this section with a few standard lemmas that are used frequently throughout the paper. In Section 3, we give a complete analysis of the discrete nonlinear problem. Using the Banach fixed point theorem, we are able to establish the well-posedness of the method as well as derive quasi-optimal error estimates. We end this section by deriving $L^2$ estimates using a duality argument. Finally, in Section 4 we provide some numerical examples which show the efficiency of the method as well as back up the theoretical findings.

## 2. Notation and discretization

Throughout the paper, we use $H^r(\Omega)$ $(r \geq 0)$ to denote the set of all $L^2(\Omega)$ functions whose distributional derivatives up to order $r$ are in $L^2(\Omega)$, and we use $H_0^r(\Omega)$ to denote the set of functions whose traces vanish up to order $r-1$ on $\partial\Omega$. For a normed linear space $X$, we denote by $X'$ its dual and $\langle \cdot, \cdot \rangle$ the pairing between $X$ and $X'$.

We let $\mathcal{T}_h$ be a quasi-uniform, simplicial, and conforming triangulation [12, 7, 4] of the domain $\Omega$ where each triangle on the boundary has at most one curved side. Set $h_T = \operatorname{diam}(T)$ $\forall T \in \mathcal{T}_h$, $h = \max_{T\in\mathcal{T}_h} h_T$, and define the piecewise Sobolev space associated with the mesh as

$$H^r(\Omega; \mathcal{T}_h) = \prod_{T\in\mathcal{T}_h} H^r(T).$$

We denote by $\mathcal{E}_h^i$ the set of interior edges in $\mathcal{T}_h$, $\mathcal{E}_h^b$ the set of boundary edges, $\mathcal{E}_h = \mathcal{E}_h^i \cup \mathcal{E}_h^b$, and $h_e = \operatorname{diam}(e)$ $\forall e \in \mathcal{E}_h$.

We define the jump of a vector function $\mathbf{w}$ on an interior edge $e = \partial T^+ \cap \partial T^-$ as follows:

$$[\![\mathbf{w}]\!]\big|_e = \mathbf{w}^+ \cdot \mathbf{n}_+\big|_e + \mathbf{w}^- \cdot \mathbf{n}_-\big|_e \in \mathbf{R},$$

where $\mathbf{w}^\pm = \mathbf{w}\big|_{T^\pm}$ and $\mathbf{n}_\pm$ is the outward unit normal of $T^\pm$. On a boundary edge $e \in \mathcal{E}_h^b$, we define

$$[\![\mathbf{w}]\!]\big|_e = \mathbf{w} \cdot \mathbf{n}\big|_e \in \mathbf{R}.$$

Next, for a matrix $\underline{\mathbf{w}} \in \mathbf{R}^{2\times 2}$, we define the average of $\underline{\mathbf{w}}$ on $e = \partial T^+ \cap \partial T^-$ by

$$\{\!\{\underline{\mathbf{w}}\}\!\}\big|_e = \frac{1}{2}\left(\underline{\mathbf{w}}^+\big|_e + \underline{\mathbf{w}}^-\big|_e\right) \in \mathbf{R}^{2\times 2}.$$

On the boundary $e \in \mathcal{E}_h^b$, we take

$$\{\!\{\underline{\mathbf{w}}\}\!\}\big|_e = \underline{\mathbf{w}}\big|_e \in \mathbf{R}^{2\times 2}.$$

We denote by $D_h^2 w$ the piecewise Hessian of a function $w \in H^2(\Omega;\mathcal{T}_h)$ and use $\operatorname{cof}(D_h^2 w)$ to denote the piecewise cofactor matrix of $D_h^2 w$; i.e.,

$$\operatorname{cof}(D_h^2 w)\big|_T = \begin{pmatrix} \dfrac{\partial^2 w_T}{\partial x_2^2} & -\dfrac{\partial^2 w_T}{\partial x_1 \partial x_2} \\ -\dfrac{\partial^2 w_T}{\partial x_1 \partial x_2} & \dfrac{\partial^2 w_T}{\partial x_1^2} \end{pmatrix} \qquad \forall T \in \mathcal{T}_h,$$

where $w_T = w\big|_T$.

To construct a consistent scheme which inherits a stable discrete linearization, we take $w \in H^3(\Omega;\mathcal{T}_h) \cap H^1(\Omega)$ and $v \in H^2(\Omega;\mathcal{T}_h) \cap H^1(\Omega)$. By using (1.1a), the divergence-free row property of cofactor matrices (cf. Lemma 2.3), and integration by parts, we have

$$\begin{aligned}
&\int_\Omega \Big(f - \det(D_h^2(u+w))\Big) v \, dx \\
&\quad = -\int_\Omega \big(\det(D_h^2 w)\big) v \, dx - \sum_{T\in\mathcal{T}_h} \int_T \Big(\nabla \cdot \big(\operatorname{cof}(D^2 u)\nabla w\big)\Big) v \, dx \\
&\quad = -\int_\Omega \big(\det(D_h^2 w)\big) v \, dx + \int_\Omega \big(\operatorname{cof}(D^2 u)\nabla w\big)\cdot\nabla v \, dx - \sum_{e\in\mathcal{E}_h} \int_e \big[\!\big[\operatorname{cof}(D^2 u)\nabla w\big]\!\big] v \, ds.
\end{aligned}$$

Therefore, by rearranging and adding terms on both sides of the equation, we get

$$\begin{aligned}
&\int_\Omega \Big(f - \det\big(D_h^2(u+w)\big)\Big) v \, dx + \sum_{e\in\mathcal{E}_h^i} \int_e \big[\!\big[ \{\!\{\operatorname{cof}(D_h^2(u+w))\}\!\} \nabla(u+w)\big]\!\big] v \, ds \\
&\qquad = \int_\Omega \big(\operatorname{cof}(D^2 u)\nabla w\big)\cdot\nabla v \, dx - \sum_{e\in\mathcal{E}_h^b} \int_e \big[\!\big[\operatorname{cof}(D^2 u)\nabla w\big]\!\big] v \, ds \\
&\qquad\quad - \int_\Omega \big(\det(D_h^2 w)\big) v \, dx + \sum_{e\in\mathcal{E}_h^i} \int_e \big[\!\big[ \{\!\{\operatorname{cof}(D_h^2 w)\}\!\} \nabla w\big]\!\big] v \, ds.
\end{aligned} \tag{2.1}$$

The guiding principle here and in the derivation of (2.2) below is that we want to write the left-hand side as a functional in $u + w$ and the right-hand side as the sum of a linear and a quadratic functional in $w$ (cf. (2.3) below).

Observe that the bilinear form

$$(w, v) \to \int_\Omega \big(\operatorname{cof}(D^2 u)\nabla w\big)\cdot\nabla v \, dx - \sum_{e\in\mathcal{E}_h^b} \int_e \big[\!\big[\operatorname{cof}(D^2 u)\nabla w\big]\!\big] v \, ds$$

that appears on the right-hand side of (2.1) can be symmetrized and stabilized to become a consistent and stable bilinear form for the second order differential operator $-\nabla\cdot\big(\operatorname{cof}(D^2 u)\nabla(\cdot)\big)$, as developed by Nitsche in [28]. After symmetrization

and stabilization, equation (2.1) becomes

$$
\begin{aligned}
(2.2)\quad & \int_\Omega \Big(f - \det\big(D_h^2(u+w)\big)\Big) v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \Big[\!\Big[ \big\{\!\!\big\{ \operatorname{cof}(D_h^2(u+w)) \big\}\!\!\big\} \nabla(u+w) \Big]\!\Big] v\,ds \\
& - \sum_{e\in\mathcal{E}_h^b} \Big[\!\Big[ \operatorname{cof}(D_h^2(u+w))\nabla v \Big]\!\Big] (u+w)\,ds + \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D_h^2(u+w))\nabla v \Big]\!\Big] g\,ds \\
& \quad + \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e (u+w) v\,ds - \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e g v\,ds \\
& = \int_\Omega \big(\operatorname{cof}(D^2u)\nabla w\big)\cdot\nabla v\,dx - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D^2u)\nabla w \Big]\!\Big] v\,ds \\
& \quad - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D^2u)\nabla v \Big]\!\Big] w\,ds + \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e w v\,ds \\
& \qquad - \int_\Omega \big(\det(D_h^2 w)\big) v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \Big[\!\Big[ \big\{\!\!\big\{ \operatorname{cof}(D_h^2 w) \big\}\!\!\big\} \nabla w \Big]\!\Big] v\,ds \\
& \qquad\quad - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D_h^2 w)\nabla v \Big]\!\Big] w\,ds,
\end{aligned}
$$

where $\sigma$ is a positive penalty parameter. Equation (2.2) can be written compactly as

$$
F(u+w) = Lw + Rw, \tag{2.3}
$$

where the nonlinear mappings $F, R : H^3(\Omega;\mathcal{T}_h) \to \big[H^2(\Omega;\mathcal{T}_h)\cap H^1(\Omega)\big]'$ and the linear mapping $L : H^2(\Omega;\mathcal{T}_h) \to \big[H^2(\Omega;\mathcal{T}_h)\cap H^1(\Omega)\big]'$ are defined by

$$
\begin{aligned}
(2.4)\quad \big\langle Fw, v\big\rangle &= \int_\Omega \Big(f - \det(D_h^2 w)\Big) v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \Big[\!\Big[ \big\{\!\!\big\{ \operatorname{cof}(D_h^2 w) \big\}\!\!\big\} \nabla w \Big]\!\Big] v\,ds \\
&\quad - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D_h^2 w)\nabla v \Big]\!\Big] (w-g)\,ds + \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e (w-g) v\,ds, \\
(2.5)\quad \big\langle Rw, v\big\rangle &= -\int_\Omega \big(\det(D_h^2 w)\big) v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \Big[\!\Big[ \big\{\!\!\big\{ \operatorname{cof}(D_h^2 w) \big\}\!\!\big\} \nabla w \Big]\!\Big] v\,ds \\
&\quad - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D_h^2 w)\nabla v \Big]\!\Big] w\,ds, \\
(2.6)\quad \big\langle Lw, v\big\rangle &= \int_\Omega \big(\operatorname{cof}(D^2u)\nabla w\big)\cdot\nabla v\,dx - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D^2u)\nabla w \Big]\!\Big] v\,ds \\
&\quad - \sum_{e\in\mathcal{E}_h^b} \int_e \Big[\!\Big[ \operatorname{cof}(D^2u)\nabla v \Big]\!\Big] w\,ds + \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e v w\,ds.
\end{aligned}
$$

Lastly, to define the finite element method, we must define the appropriate finite element spaces. For an integer $k \geq 3$, we define the finite element space $V_h \subset H^1(\Omega)$ as follows:

- If $T \in \mathcal{T}_h$ does not have a curved edge, then $v\big|_T$ is a polynomial of (total) degree $\leq k$ in the rectilinear coordinates for $T$;
- If $T \in \mathcal{T}_h$ has one curved edge, then $v\big|_T$ is a polynomial of degree $\leq k$ in the curvilinear coordinates of $T$ that correspond to the rectilinear coordinates on the reference triangle (Example 2, p. 1216 of [4]).

*Remark* 2.1. The requirement $k \geq 3$ as well as the regularity condition $u \in H^s(\Omega)$, $s > 3$ will be made obvious in Theorem 3.1.

Let $F_h : V_h \to V_h'$ be the restriction of $F$ to $V_h$. Then the penalty method for (1.1) is to find $u_h \in V_h$ such that

$$F_h u_h = 0; \tag{2.7}$$

that is,

$$\int_\Omega \Big(f - \det(D_h^2 u_h)\Big) v\,dx + \sum_{e\in\mathcal{E}_h^i} \int_e \big[\!\big[ \{\!\{ \operatorname{cof}(D_h^2 u_h) \}\!\} \nabla u_h \big]\!\big] v\,ds$$
$$- \sum_{e\in\mathcal{E}_h^b} \int_e \big[\!\big[ \operatorname{cof}(D_h^2 u_h)\nabla v \big]\!\big] (u_h - g)\,ds + \sigma \sum_{e\in\mathcal{E}_h^b} h_e^{-1} \int_e (u_h - g) v\,ds = 0 \quad \forall v \in V_h.$$

The well-posedness as well as the error estimates of the penalty method (2.7) are established in the following section. For now, we end the current section with a few remarks and technical lemmas which are used throughout the paper.

*Remark* 2.2. Noting $R$ is quadratic in its arguments, we conclude from (2.3) that the operator $L$ is the linearization of $F$ at $u$. The motivation of the finite element method (2.7) is based on the fact that $L$ is the operator associated with Nitsche's method for the second order differential operator $-\nabla \cdot \big(\operatorname{cof}(D^2 u)\nabla(\cdot)\big)$, i.e., the linearization of the nonlinear operator (1.1a).

*Remark* 2.3. In order to avoid the proliferation of constants, we shall use the notation $A \lesssim B$ to represent the relation $A \leq \text{constant} \times B$, where the constant is independent of the mesh parameter $h$ and the penalty parameter $\sigma$.

**Lemma 2.1** (An Algebraic Identity)**.** *For any $v, w \in H^2(\Omega; \mathcal{T}_h)$, there holds*

$$\begin{aligned} \det(D_h^2 v) - \det(D_h^2 w) &= \frac{1}{2}\operatorname{cof}(D_h^2 v) : D_h^2 v - \frac{1}{2}\operatorname{cof}(D_h^2 w) : D_h^2 w \\ &= \frac{1}{2}\big(\operatorname{cof}(D_h^2 v) + \operatorname{cof}(D_h^2 w)\big) : \big(D_h^2 v - D_h^2 w\big). \end{aligned} \tag{2.8}$$

**Lemma 2.2** (A Discrete Sobolev Inequality [6, 7])**.** *For any $v \in V_h$ there holds*

$$\|v\|_{L^\infty(\Omega)} \lesssim (1 + |\ln h|^{\frac{1}{2}}) \|v\|_{H^1(\Omega)}. \tag{2.9}$$

**Lemma 2.3** (Divergence-Free Property of Cofactor Matrices [17])**.** *For any smooth function $v$,*

$$\nabla \cdot \big(\operatorname{cof}(D^2 v)_i\big) = \sum_{j=1}^2 \frac{\partial}{\partial x_j}\big(\operatorname{cof}(D^2 v)_{ij}\big) = 0 \qquad \text{for } i = 1, 2, \tag{2.10}$$

*where $\operatorname{cof}(D^2 v)_i$ and $\operatorname{cof}(D^2 v)_{ij}$ denote respectively the $i$th row and the $(i,j)$-entry of the cofactor matrix $\operatorname{cof}(D^2 v)$.*

**Lemma 2.4** (Trace and Inverse Inequalities [7, 12]). *Let $D \subset \mathbf{R}^2$ be a regular and star-like domain. Then there holds*

$$\|v\|_{L^2(\partial D)} \lesssim \operatorname{diam}(D)^{\frac{1}{2}}\|v\|_{H^1(D)} + \operatorname{diam}(D)^{-\frac{1}{2}}\|v\|_{L^2(D)} \qquad \forall v \in H^1(D),$$
$$\|v\|_{W^{m,q}(D)} \lesssim \operatorname{diam}(D)^{n-m+2\min\{0,1/q-1/p\}}\|v\|_{W^{n,p}(D)} \qquad \forall v \in \mathbb{P}_k(D),$$

*where $\mathbb{P}_k(D)$ denotes the set of polynomials up to degree $k$ restricted to $D$.*

**Lemma 2.5** (Approximation Properties of $V_h$ [4]). *Let $m, \ell$ be two integers such that $0 \le m \le \ell \le k+1$. Then for any $\chi \in H^\ell(\Omega)$, there exists $v \in V_h$ such that*

$$\left(\sum_{T\in\mathcal{T}_h} \|\chi - v\|^2_{H^m(T)}\right)^{\frac{1}{2}} \lesssim h^{\ell-m}\|\chi\|_{H^\ell(\Omega)}.$$

*Remark* 2.4. Due to interpolation of Sobolev spaces (e.g. [7, Theorem 14.3.3]), the parameter $\ell$ appearing in Lemma 2.5 can be taken to be any real number such that $0 \le m \le \ell \le k+1$.

## 3. Convergence analysis

Let $L_h : V_h \to V_h'$ be the restriction of $L$ (defined by (2.6)) to $V_h$, and let $L_h^{-1} : V_h' \to V_h$ denote its inverse (which exists if $\sigma$ is sufficiently large; see Lemma 3.1). Define the mapping $M : H^3(\Omega;\mathcal{T}_h) \to V_h$ as

$$M = L_h^{-1}\big(L - F\big), \tag{3.1}$$

and let $M_h : V_h \to V_h$ be the restriction of $M$ to $V_h$; that is,

$$M_h = Id_h - L_h^{-1}F_h, \tag{3.2}$$

where $Id_h$ is the identity map on $V_h$. The existence of a solution to (2.7) near $u$ will be proven by establishing a fixed point for $M_h$ in a small ball centered at $u_{c,h} \in V_h$, where

$$u_{c,h} = L_h^{-1}Lu. \tag{3.3}$$

Define the discrete $H^1$-like and $H^2$-like norms as

$$\|v\|^2_{1,h} = \|v\|^2_{H^1(\Omega)} + \sum_{e\in\mathcal{E}_h^b}\Big(h_e^{-1}\|v\|^2_{L^2(e)} + h_e\|\nabla v\|^2_{L^2(e)}\Big), \tag{3.4}$$

$$\|v\|^2_{2,h} = \sum_{T\in\mathcal{T}_h}|v|^2_{H^2(T)} + \sum_{e\in\mathcal{E}_h}\Big(h_e^{-1}\big\|\big[\!\big[\nabla v\big]\!\big]\big\|^2_{L^2(e)} + h_e\big\|\big\{\!\big\{D_h^2 v\big\}\!\big\}\big\|^2_{L^2(e)}\Big) + \sum_{e\in\mathcal{E}_h^b} h_e^{-3}\|v\|^2_{L^2(e)}, \tag{3.5}$$

and define the corresponding discrete negative norm as

$$\|q\|_{-1,h} = \sup_{0\neq v\in V_h}\frac{\langle q, v\rangle}{\|v\|_{1,h}}. \tag{3.6}$$

*Remark* 3.1. By Lemma 2.4 and scaling, there holds

$$\|v\|_{2,h} \lesssim h^{-1}\|v\|_{1,h} \qquad \forall v \in V_h. \tag{3.7}$$

*Remark* 3.2. By Lemma 2.5 and scaling, if $u \in H^s(\Omega)$, then there exists $v \in V_h$ such that

$$\|u - v\|_{1,h} + h\|u - v\|_{2,h} \lesssim h^{\ell-1}\|u\|_{H^\ell(\Omega)}, \tag{3.8}$$

where $\ell = \min\{k+1, s\}$ and $k$ denotes the polynomial degree of the finite element space $V_h$.

The analysis of the discrete problem (2.7) is based on the following lemmas.

**Lemma 3.1.** *We have*

$$\|Lv\|_{-1,h} \lesssim (1+\sigma)\|v\|_{1,h} \qquad \forall v \in H^2(\Omega; \mathcal{T}_h) \cap H^1(\Omega). \tag{3.9}$$

*Moreover, there exists a $\sigma_0 > 0$ that depends on $u$ and the minimum angle of $\mathcal{T}_h$ such that the map $L_h : V_h \to V_h'$ is invertible and*

$$\|L_h^{-1} q\|_{1,h} \lesssim \|q\|_{-1,h}, \tag{3.10}$$

*provided $\sigma \geq \sigma_0$.*

*Proof.* Since $u$ is strictly convex in $\Omega$, the matrix $\operatorname{cof}(D^2 u)$ is positive definite. Furthermore by a Sobolev embedding, if $u \in H^s(\Omega)$ for some $s > 3$, then $u \in W^{2,\infty}(\Omega)$. Therefore, there exist constants $\lambda, \Lambda > 0$ such that

$$\lambda\|w\|_{H^1(\Omega)}^2 \leq \int_\Omega \big(\operatorname{cof}(D^2 u)\nabla w\big) \cdot \nabla w \, dx \leq \Lambda \|w\|_{H^1(\Omega)}^2 \qquad \forall w \in H^1(\Omega). \tag{3.11}$$

The conclusion of the lemma then follows from (3.11) and the results of Nitsche [28]. □

*Remark* 3.3. By Lemma 3.1, both $u_{c,h}$ and the operators $M, M_h$ are well defined if $\sigma \geq \sigma_0$, which we assume for the rest of the paper.

**Lemma 3.2.** *Suppose that $u \in H^s(\Omega)$ and let $u_{c,h} \in V_h$ be defined by* (3.3). *Then there holds*

$$\|u - u_{c,h}\|_{1,h} + h\|u - u_{c,h}\|_{2,h} \lesssim (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}, \tag{3.12}$$

*where $\ell = \min\{k+1, s\}$ and $k$ denotes the polynomial degree of the finite space $V_h$.*

*Proof.* By (3.10), (3.3), and (3.9), we have for any $v \in V_h$,

$$\begin{aligned} \|u - u_{c,h}\|_{1,h} &\leq \|u - v\|_{1,h} + \|L_h^{-1} L_h(v - u_{c,h})\|_{1,h} \\ &\lesssim \|u - v\|_{1,h} + \|L(v - u)\|_{-1,h} \\ &\lesssim (1+\sigma)\|u - v\|_{1,h}. \end{aligned}$$

As $v$ was arbitrary, the first estimate in (3.12) follows from (3.8).

To obtain the second error estimate in (3.12), we use the inverse inequality (3.7) to conclude that

$$\begin{aligned} \|u - u_{c,h}\|_{2,h} &\lesssim \|u - v\|_{2,h} + h^{-1}\|u_{c,h} - v\|_{1,h} \\ &\lesssim \|u - v\|_{2,h} + h^{-1}\|u - v\|_{1,h} + h^{-1}\|u - u_{c,h}\|_{1,h}. \end{aligned}$$

Again, choosing $v$ so that (3.8) holds, we obtain (3.12). □

With the preliminary analysis completed, we turn our attention to the nonlinear problem (2.7) and the mapping $M$ defined in (3.1). Note that by (3.1), (2.3), and (3.3), for all $w \in H^3(\Omega;\mathcal{T}_h)$,

$$\begin{aligned} Mw &= L_h^{-1}\big(Lw - Fw\big) \\ &= L_h^{-1}\big(Lw - L(w-u) - R(w-u)\big) \\ &= L_h^{-1}\big(Lu - R(w-u)\big) \\ &= u_{c,h} - L_h^{-1}R(w-u), \end{aligned} \tag{3.13}$$

and hence

$$Mw_1 - Mw_2 = L_h^{-1}\Big(R(w_2-u) - R(w_1-u)\Big) \qquad \forall w_1, w_2 \in H^3(\Omega;\mathcal{T}_h). \tag{3.14}$$

From (3.14) and (3.13) it is clear that the crucial ingredient for applying the Banach fixed point theorem to the map $M_h$ in a small ball around $u_{c,h}$ is the contraction estimate of $R$ established in the next lemma.

**Lemma 3.3** (Contraction Estimate of $R$)**.** *For any $w_1, w_2 \in H^3(\Omega;\mathcal{T}_h)$, there holds*

$$\big\|Rw_1 - Rw_2\big\|_{-1,h} \lesssim (1 + |\ln h|^{\frac{1}{2}})\big(\|w_1\|_{2,h} + \|w_2\|_{2,h}\big)\|w_1 - w_2\|_{2,h}. \tag{3.15}$$

*Proof.* By (2.5) and (2.8) we have for any $v \in V_h$,

$$\begin{aligned} \big\langle Rw_1 - Rw_2, v\big\rangle &= \int_\Omega \Big(\det(D_h^2 w_2) - \det(D_h^2 w_1)\Big) v\,dx \\ &\quad + \sum_{e\in\mathcal{E}_h^i}\int_e \Big(\big[\!\big[\big\{\!\!\big\{\mathrm{cof}(D_h^2 w_1)\big\}\!\!\big\}\nabla w_1\big]\!\big] - \big[\!\big[\big\{\!\!\big\{\mathrm{cof}(D_h^2 w_2)\big\}\!\!\big\}\nabla w_2\big]\!\big]\Big) v\,ds \\ &\quad - \sum_{e\in\mathcal{E}_h^b}\int_e \Big(\big[\!\big[\mathrm{cof}(D_h^2 w_1)\nabla v\big]\!\big] w_1 - \big[\!\big[\mathrm{cof}(D_h^2 w_2)\nabla v\big]\!\big] w_2\Big) ds \\ &\quad = \frac{1}{2}\int_\Omega \Big(\mathrm{cof}(D_h^2(w_1+w_2)) : D_h^2(w_2-w_1)\Big) v\,dx \\ &\quad + \sum_{e\in\mathcal{E}_h^i}\int_e \Big(\big[\!\big[\big\{\!\!\big\{\mathrm{cof}(D_h^2(w_1-w_2))\big\}\!\!\big\}\nabla w_1\big]\!\big] - \big[\!\big[\big\{\!\!\big\{\mathrm{cof}(D_h^2 w_2)\big\}\!\!\big\}\nabla(w_2-w_1)\big]\!\big]\Big) v\,ds \\ &\quad - \sum_{e\in\mathcal{E}_h^b}\int_e \Big(\big[\!\big[\mathrm{cof}(D_h^2 w_1)\nabla v\big]\!\big](w_1-w_2) - \big[\!\big[\mathrm{cof}(D_h^2(w_2-w_1))\nabla v\big]\!\big] w_2\Big) ds. \end{aligned}$$

Using the inverse inequality (3.7), (2.9), (3.5), and the Cauchy-Schwarz inequality we obtain

$$\begin{aligned} \big\langle Rw_1 - Rw_2, v\big\rangle \le \Bigg(&\frac{1}{2}\sum_{T\in\mathcal{T}_h} |w_1+w_2|_{H^2(T)}|w_1-w_2|_{H^2(T)} \\ &+ \sum_{e\in\mathcal{E}_h^i}\Big(\big\|\big\{\!\!\big\{D_h^2(w_1-w_2)\big\}\!\!\big\}\big\|_{L^2(e)}\big\|\big[\!\big[\nabla w_1\big]\!\big]\big\|_{L^2(e)} \\ &+ \big\|\big\{\!\!\big\{D_h^2 w_2\big\}\!\!\big\}\big\|_{L^2(e)}\big\|\big[\!\big[\nabla(w_1-w_2)\big]\!\big]\big\|_{L^2(e)}\Big)\Bigg)\|v\|_{L^\infty(\Omega)} \\ &+ \sum_{e\in\mathcal{E}_h^b}\Big(\|D_h^2 w_1\|_{L^2(e)}\|w_1-w_2\|_{L^2(e)} \end{aligned}$$

$$
\begin{aligned}
&\qquad + \|D_h^2(w_1-w_2)\|_{L^2(e)}\|w_2\|_{L^2(e)}\Big)\|\nabla v\|_{L^\infty(\Omega)}\\
&\lesssim (1+|\ln h|^{\frac12})\Bigg(\big(\|w_1\|_{2,h}+\|w_2\|_{2,h}\big)\|w_1-w_2\|_{2,h}\\
&\qquad + \sum_{e\in\mathcal{E}_h^i}\Big(h_e^{\frac12}\big\|\{\!\{D_h^2(w_1-w_2)\}\!\}\big\|_{L^2(e)}h^{-\frac12}\big\|[\![\nabla w_1]\!]\big\|_{L^2(e)}\\
&\qquad\quad + h^{\frac12}\big\|\{\!\{D_h^2 w_2\}\!\}\big\|_{L^2(e)}h^{-\frac12}\big\|[\![\nabla(w_1-w_2)]\!]\big\|_{L^2(e)}\Big)\\
&\qquad\quad + \sum_{e\in\mathcal{E}_h^b}h_e^{-1}\Big(h^{\frac12}\big\|D_h^2 w_1\big\|_{L^2(e)}h^{-\frac12}\|w_1-w_2\|_{L^2(e)}\\
&\qquad\quad + h^{\frac12}\big\|D_h^2(w_1-w_2)\big\|_{L^2(e)}h^{-\frac12}\|w_2\|_{L^2(e)}\Big)\Bigg)\|v\|_{1,h}\\
&\lesssim (1+|\ln h|^{\frac12})\big(\|w_1\|_{2,h}+\|w_2\|_{2,h}\big)\|w_1-w_2\|_{2,h}\|v\|_{1,h}.
\end{aligned}
$$

The estimate (3.15) then follows from (3.6). □

In light of (3.13)–(3.14), Lemma 3.3 immediately gives us the following three results.

**Lemma 3.4** (Contraction Property of $M$ on $H^3(\Omega;\mathcal{T}_h)$)**.** *For any $w_1, w_2 \in H^3(\Omega;\mathcal{T}_h)$ there holds*

$$
\begin{aligned}
&\|Mw_1-Mw_2\|_{1,h}\\
&\qquad\lesssim (1+|\ln h|^{\frac12})\big(\|u-w_1\|_{2,h}+\|u-w_2\|_{2,h}\big)\|w_1-w_2\|_{2,h}.
\end{aligned}
\tag{3.16}
$$

*Proof.* By (3.14), (3.10), and (3.15), we have

$$
\begin{aligned}
\|Mw_1-Mw_2\|_{1,h} &= \big\|L_h^{-1}\big(R(w_2-u)-R(w_1-u)\big)\big\|_{1,h}\\
&\lesssim \|R(w_2-u)-R(w_1-u)\|_{-1,h}\\
&\lesssim (1+|\ln h|^{\frac12})\big(\|u-w_1\|_{2,h}+\|u-w_2\|_{2,h}\big)\|w_1-w_2\|_{2,h}.
\end{aligned}
$$
□

**Lemma 3.5** (Contraction Property of $M_h$ on $V_h$)**.** *Define the discrete closed ball with center $u_{c,h}$ and radius $\rho$ as*

$$
\mathbb{B}_\rho(u_{c,h}) = \{v\in V_h;\ \|u_{c,h}-v\|_{1,h}\le\rho\}. \tag{3.17}
$$

*Then there exists a constant $C_1>0$ such that for any $v_1, v_2 \in \mathbb{B}_\rho(u_{c,h})$ there holds*

$$
\begin{aligned}
&\|M_hv_1-M_hv_2\|_{1,h}\\
&\qquad\le C_1h^{-2}(1+|\ln h|^{\frac12})\big(\rho+(1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}\big)\|v_1-v_2\|_{1,h}.
\end{aligned}
\tag{3.18}
$$

*Proof.* By (3.16), (3.17), the inverse inequality (3.7) and (3.12), we have

$$
\begin{aligned}
\|M_hv_1-M_hv_2\|_{1,h} &\lesssim (1+|\ln h|^{\frac12})\big(\|u-v_1\|_{2,h}+\|u-v_2\|_{2,h}\big)\|v_1-v_2\|_{2,h}\\
&\lesssim h^{-1}(1+|\ln h|^{\frac12})\big(\|u-u_{c,h}\|_{2,h}+h^{-1}\rho\big)\|v_1-v_2\|_{1,h}\\
&\lesssim h^{-2}(1+|\ln h|^{\frac12})\big(\rho+(1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}\big)\|v_1-v_2\|_{1,h}.
\end{aligned}
$$
□

**Lemma 3.6** (Mapping Property of $M_h$)**.** *There holds for any $v \in \mathbb{B}_\rho(u_{c,h})$,*

$$\|u_{c,h} - M_h v\|_{1,h} \le C_2 h^{-2}(1 + |\ln h|^{\frac{1}{2}})\big(\rho^2 + (1+\sigma)^2 h^{2(\ell-1)}\|u\|^2_{H^\ell(\Omega)}\big). \tag{3.19}$$

*Proof.* For any $w \in H^3(\Omega; \mathcal{T}_h)$, we have by (3.13), (3.10), and (3.15) that

$$\begin{aligned}\|u_{c,h} - Mw\|_{1,h} &= \big\|L_h^{-1}R(w-u)\big\|_{1,h}\\ &\lesssim \|R(w-u)\|_{-1,h}\\ &\lesssim (1 + |\ln h|^{\frac{1}{2}})\|u-w\|^2_{2,h}.\end{aligned}$$

Therefore by the inverse inequality (3.7), (3.17) and (3.12), we have for $v \in \mathbb{B}_\rho(u_{c,h})$,

$$\begin{aligned}\|u_{c,h} - M_h v\|_{1,h} &\lesssim (1 + |\ln h|^{\frac{1}{2}})\|u-v\|^2_{2,h}\\ &\lesssim (1 + |\ln h|^{\frac{1}{2}})\big(\|u-u_{c,h}\|^2_{2,h} + \|u_{c,h}-v\|^2_{2,h}\big)\\ &\lesssim h^{-2}(1 + |\ln h|^{\frac{1}{2}})\big((1+\sigma)^2 h^{2(\ell-1)}\|u\|^2_{H^\ell(\Omega)} + \rho^2\big).\end{aligned}$$

□

With the preceding results established, we are now in position to prove the first main result.

**Theorem 3.1** (Main Theorem I)**.** *There exists an $h_0(\sigma) > 0$ such that for $h \le h_0(\sigma)$ there exists a solution $u_h$ to the penalty method* (2.7)*. Moreover,*

$$\|u - u_h\|_{1,h} \lesssim (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}, \tag{3.20}$$

$$\|u - u_h\|_{2,h} \lesssim (1+\sigma)h^{\ell-2}\|u\|_{H^\ell(\Omega)}. \tag{3.21}$$

*Proof.* Since $\ell = \min\{k+1, s\} > 3$, we can choose $h_0(\sigma) > 0$ such that $h \le h_0(\sigma)$ implies

$$\delta = 2\max\{C_1, C_2\}h^{-2}(1 + |\ln h|^{\frac{1}{2}})(1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)} < 1. \tag{3.22}$$

Fix $h \le h_0(\sigma)$ and set

$$\rho_0 = (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}. \tag{3.23}$$

Then for any $v \in \mathbb{B}_{\rho_0}(u_{c,h})$, we have by (3.19), (3.22) and (3.23) that

$$\begin{aligned}\|u_{c,h} - M_h v\|_{1,h} &\le C_2 h^{-2}(1 + |\ln h|^{\frac{1}{2}})\big(\rho_0^2 + (1+\sigma)^2 h^{2(\ell-1)}\|u\|^2_{H^\ell(\Omega)}\big)\\ &= 2\Big(C_2 h^{-2}(1 + |\ln h|^{\frac{1}{2}})(1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}\Big)\rho_0\\ &\le \rho_0,\end{aligned}$$

and so $M_h$ maps $\mathbb{B}_{\rho_0}(u_{c,h})$ into $\mathbb{B}_{\rho_0}(u_{c,h})$. Moreover, by (3.18) and (3.22) for $v_1, v_2 \in \mathbb{B}_{\rho_0}(u_{c,h})$,

$$\begin{aligned}\|M_h v_1 - M_h v_2\|_{1,h} &\le C_1 h^{-1}(1 + |\ln h|^{\frac{1}{2}})\big(\rho_0 + (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}\big)\|v_1 - v_2\|_{1,h}\\ &= 2C_1 h^{-1}(1 + |\ln h|^{\frac{1}{2}})(1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}\|v_1 - v_2\|_{1,h}\\ &\le \delta\|v_1 - v_2\|_{1,h}.\end{aligned}$$

Hence $M_h$ has a unique fixed point $u_h \in \mathbb{B}_\rho(u_{c,h})$, which is a solution of (2.7). Moreover, by (3.23) and (3.12),

$$\|u - u_h\|_{1,h} \le \|u - u_{c,h}\|_{1,h} + \rho_0 \lesssim (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}.$$

Finally, by the inverse inequality, we have

$$\|u - u_h\|_{2,h} \le \|u - u_{c,h}\|_{2,h} + h^{-1}\rho_0 \lesssim (1+\sigma)h^{\ell-2}\|u\|_{H^\ell(\Omega)}.$$

□

*Remark* 3.4. Theorem 3.1 shows that $\|u_h - u_{c,h}\|_{1,h} \leq \rho_0 = (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}$, but this estimate is not sharp. Indeed, by (3.13) we have

$$u_h = M_h u_h = u_{c,h} - L_h^{-1} R(u_h - u), \tag{3.24}$$

and therefore by (3.1), (3.15), and (3.21),

$$\begin{aligned} \|u_h - u_{c,h}\|_{1,h} &= \|L_h^{-1} R(u_h - u)\|_{1,h} \\ &\lesssim \|R(u_h - u)\|_{-1,h} \\ &\lesssim (1 + |\ln h|^{\frac{1}{2}})\|u - u_h\|_{2,h}^2 \\ &\lesssim (1 + |\ln h|^{\frac{1}{2}})(1+\sigma)^2 h^{2(\ell-2)}\|u\|_{H^\ell(\Omega)}^2. \end{aligned} \tag{3.25}$$

**Theorem 3.2** (Main Theorem II)**.** *In addition to the hypotheses of Theorem* 3.1*, assume that* $u \in W^{3,\infty}(\Omega)$*. Then there holds*

$$\|u - u_h\|_{L^2(\Omega)} \lesssim (1+\sigma)^2\big(h^\ell \|u\|_{H^\ell(\Omega)} + (1 + |\ln h|^{\frac{1}{2}})h^{2(\ell-2)}\|u\|_{H^\ell(\Omega)}^2\big). \tag{3.26}$$

*Proof.* Let $\psi \in H_0^1(\Omega)$ be the solution to the following auxiliary problem:

$$-\nabla \cdot \big(\mathrm{cof}(D^2 u)\nabla\psi\big) = u - u_h \qquad \text{in } \Omega, \tag{3.27a}$$

$$\psi = 0 \qquad \text{on } \partial\Omega. \tag{3.27b}$$

Since $u \in W^{3,\infty}(\Omega)$, there holds $\mathrm{cof}(D^2 u) \in \big[W^{1,\infty}(\Omega)\big]^{2\times 2}$. Therefore by elliptic regularity theory [23, 17, 24], we have

$$\|\psi\|_{H^2(\Omega)} \lesssim \|u - u_h\|_{L^2(\Omega)}. \tag{3.28}$$

Let $\psi_h \in V_h$ be chosen so that

$$\|\psi - \psi_h\|_{1,h} \lesssim h\|\psi\|_{H^2(\Omega)} \lesssim h\|u - u_h\|_{L^2(\Omega)}. \tag{3.29}$$

Note that by the trace inequality,

$$\Big(\sum_{e\in\mathcal{E}_h^b} h_e\|\nabla\psi\|_{L^2(e)}^2\Big)^{\frac{1}{2}} \lesssim \|\psi\|_{H^2(\Omega)},$$

and therefore by (3.4) and (3.27b),

$$\|\psi\|_{1,h} \lesssim \|\psi\|_{H^2(\Omega)}.$$

Hence, by (3.29) we have

$$\|\psi_h\|_{1,h} \leq \|\psi_h - \psi\|_{1,h} + \|\psi\|_{1,h} \lesssim \|\psi\|_{H^2(\Omega)} \lesssim \|u - u_h\|_{L^2(\Omega)}. \tag{3.30}$$

Using (3.27), we can write

$$\begin{aligned} \|u - u_h\|_{L^2(\Omega)}^2 &= \big\langle L(u - u_h), \psi\big\rangle \\ &= \big\langle L(u - u_h), \psi - \psi_h\big\rangle + \big\langle L(u - u_h), \psi_h\big\rangle. \end{aligned} \tag{3.31}$$

From (3.9), (3.29) and (3.28), we have

$$\begin{aligned} \big\langle L(u - u_h), \psi - \psi_h\big\rangle &\leq \|L(u - u_h)\|_{-1,h}\|\psi - \psi_h\|_{1,h} \\ &\lesssim (1+\sigma)\|u - u_h\|_{1,h}\|\psi - \psi_h\|_{1,h} \\ &\lesssim h(1+\sigma)\|u - u_h\|_{1,h}\|u - u_h\|_{L^2(\Omega)}. \end{aligned} \tag{3.32}$$

On the other hand, (3.24), (3.15), (3.30) and (3.28) imply

$$\begin{aligned}\big\langle L(u_h-u),\psi_h\big\rangle &= \big\langle L_h(u_h-u_{c,h}),\psi_h\big\rangle \\ &= \big\langle R(u-u_h),\psi_h\big\rangle \\ &\le \|R(u-u_h)\|_{-1,h}\|\psi_h\|_{1,h} \\ &\lesssim (1+|\ln h|^{\frac12})\|u-u_h\|_{2,h}^2\|\psi_h\|_{1,h} \\ &\lesssim (1+|\ln h|^{\frac12})\|u-u_h\|_{2,h}^2\|u-u_h\|_{L^2(\Omega)}.\end{aligned} \tag{3.33}$$

Combining (3.31)–(3.33), dividing by $\|u-u_h\|_{L^2(\Omega)}$, and using (3.20)–(3.21), we obtain

$$\|u-u_h\|_{L^2(\Omega)} \lesssim (1+\sigma)^2\big(h^\ell\|u\|_{H^\ell(\Omega)} + (1+|\ln h|^{\frac12})h^{2(\ell-2)}\|u\|_{H^\ell(\Omega)}^2\big). \qquad \square$$

*Remark* 3.5. Since $\ell > 3$, the error estimate (3.26) is of higher order than (3.20). Moreover, the $L^2$ estimate is (almost) of the optimal order $k+1$ provided $s \ge 4$.

## 4. Numerical experiments

In this section, we perform some numerical tests that back up the theoretical results proved in the previous section, as well as show the effectiveness and efficiency of the method. We solve the finite element method (2.7) using the COMSOL Multiphysics software package [13], and solve the resulting nonlinear algebraic system using Newton's method. Given $u_0 \in V_h$, the Newton approximations to $u_h$ form a sequence $\{u_k\}_{k=0}^\infty \subset V_h$ satisfying

$$DF_h[u_k](u_{k+1}-u_k) = -F_h u_k, \tag{4.1}$$

where $DF_h : V_h \to \mathcal{L}(V_h; V_h')$ denotes the Gâteaux derivative of $F_h$. Here, $\mathcal{L}(V_h; V_h')$ denotes the space of linear maps from $V_h$ to $V_h'$.

Similar to other numerical methods for the Monge-Ampère equation [22, 26], the Newton iteration requires an accurate starting value to ensure convergence. In order to obtain a good initial guess, we apply the vanishing moment methodology to the Monge-Ampère equation [18]. The crux of the vanishing moment method is to approximate fully nonlinear PDEs by higher order semi-linear PDEs, in particular, fourth order PDEs. For the case of the Monge-Ampère equation (1.1) the vanishing moment approximation is defined to be the solution to the following fourth order problem:

$$-\varepsilon\Delta^2 u^\varepsilon + \det(D^2u^\varepsilon) = f \qquad 0<\varepsilon\ll 1, \tag{4.2}$$

along with appropriate boundary conditions. We refer the reader to [18, 21] for more details and motivation of the vanishing moment method. Besides being a valuable tool to obtain close approximations of the Monge-Ampère equation, another interesting feature of the vanishing moment methodology is the ability to choose the concave solution (recall there is a convex solution and concave solution to the Monge-Ampère equation) by substituting $\varepsilon$ by $-\varepsilon$ in (4.2). However, we do not pursue this direction in the numerical experiments below.

The $\mathcal{C}^0$ interior penalty method for (4.2) is defined as seeking $u_h^\varepsilon \in V_h$ such that

$$\varepsilon A_h u_h^\varepsilon + F_h u_h^\varepsilon = 0, \tag{4.3}$$

TABLE 1. Example 1. Errors of computed solution and rates of convergence with respect to $h$ with $\sigma = 100$.

| | $h$ | $\|u-u_h\|_{L^2(\Omega)}$ | rate | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{H^2(\mathcal{T}_h)}$ | rate |
|---|---|---|---|---|---|---|---|
| $k=2$ | 1/8 | 3.06E-03 | | 7.69E-02 | | 7.88E+00 | |
| | 1/16 | 1.62E-03 | 0.92 | 4.74E-02 | 0.70 | 6.33E+00 | 0.31 |
| | 1/32 | 1.96E-04 | 3.05 | 1.13E-02 | 2.06 | 3.18E+00 | 0.99 |
| | 1/64 | 2.79E-05 | 2.81 | 2.77E-03 | 2.03 | 1.59E+00 | 1.00 |
| | 1/128 | 7.07E-06 | 1.98 | 6.88E-04 | 2.01 | 7.97E-01 | 1.00 |
| | 1/256 | 2.03E-06 | 1.80 | 1.72E-04 | 2.00 | 3.99E-01 | 1.00 |
| $k=3$ | 1/8 | 1.47E-04 | | 1.87E-03 | | 4.27E-01 | |
| | 1/16 | 5.62E-05 | 1.38 | 9.45E-04 | 0.98 | 2.72E-01 | 0.65 |
| | 1/32 | 3.89E-06 | 3.85 | 9.51E-05 | 3.31 | 6.82E-02 | 2.00 |
| | 1/64 | 2.55E-07 | 3.93 | 1.00E-05 | 3.25 | 1.71E-02 | 2.00 |
| | 1/128 | 1.64E-08 | 3.96 | 1.13E-06 | 3.14 | 4.27E-03 | 2.00 |
| | 1/256 | 1.08E-09 | 3.92 | 1.35E-07 | 3.07 | 1.07E-03 | 2.00 |
| $k=4$ | 1/8 | 5.21E-06 | | 7.99E-05 | | 1.23E-02 | |
| | 1/16 | 1.68E-06 | 1.63 | 3.09E-05 | 1.37 | 6.22E-03 | 0.99 |
| | 1/32 | 6.06E-08 | 4.79 | 1.79E-06 | 4.11 | 7.80E-04 | 3.00 |
| | 1/64 | 2.04E-09 | 4.89 | 1.06E-07 | 4.07 | 9.72E-05 | 3.01 |
| | 1/128 | 6.47E-11 | 4.98 | 6.52E-09 | 4.03 | 1.22E-05 | 3.00 |

where

$$\langle A_h v, w\rangle = \int_\Omega D_h^2 v : D_h^2 w\,dx - \sum_{e\in\mathcal{E}_h^i}\int_e \Big( \{\!\{\partial_{nn}^2 v\}\!\}\,[\![\nabla w]\!] + [\![\nabla v]\!]\,\{\!\{\partial_{nn}^2 w\}\!\} - \sigma h_e^{-1}\,[\![\nabla v]\!]\,[\![\nabla w]\!]\Big) ds \qquad \forall v, w \in V_h,$$

and

$$\{\!\{\partial_{nn}^2 w\}\!\}\big|_e = \frac{1}{2}\big(D_h^2 w^+ \mathbf{n}_+\cdot\mathbf{n}_+\big|_e + D_h^2 w^- \mathbf{n}_-\cdot\mathbf{n}_-\big|_e\big) \qquad e = \partial T^+ \cap \partial T^- \in \mathcal{E}_h^i$$

denotes the average of the second order normal derivative of $w$. To solve (4.3), we again use Newton's method, which creates a sequence $\{u_k^\varepsilon\}_{k=0}^\infty \subset V_h$ satisfying

$$\varepsilon A_h u_{k+1}^\varepsilon + DF_h[u_k^\varepsilon](u_{k+1}^\varepsilon - u_k^\varepsilon) = -F_h u_k^\varepsilon. \tag{4.4}$$

We plan to address the convergence of the Newton iterations (4.1) and (4.4) in the near future. For now, we provide some numerical experiments illustrating the robustness of this technique.

4.1. **Example 1.** In this test, we solve (2.7) for varying values of $h$ and $k$, and choose our data such that the exact solution to the Monge-Ampère equation (1.1) is $u = 20e^{x_1^6/6+x_2}$. We take $\Omega = (0,1)^2$, the unit square and set $\sigma = 100$.

In order to obtain some good initial guesses, we first solve (4.3) with $\varepsilon$-values 1E-2, 1E-4, 1E-6, and 0, using each previous solution as our initial guess in the Newton iteration (4.4) (we take $u_0^\varepsilon = x_1^2 + x_2^2$ as our initial guess for the first iteration with $\varepsilon =$ 1E-2). After computing the solution of (2.7) we calculate the

$L^2$, $H^1$, and broken $H^2$ error and record the errors in Table 1[1]. As predicted by the theoretical results in Theorem 3.1, for $k \geq 3$ we observe $(k+1)$, $(k)$, and $(k-1)$-order convergence in the $L^2$, $H^1$ and $H^2$-norms. Furthermore, the numerical tests also indicate that the method is convergent using quadratic polynomials, although a theoretical proof of such a result has yet to be shown.

4.2. **Example 2.** In this test, we solve the Monge-Ampère equation on the unit disc $\Omega = B(0,1)$ with the same test problem as Example 1. Our finite element space is constructed such that on curved elements, we use polynomial functions of degree $\leq k$ in the curvilinear coordinates for $T$, which in this case, are isoparametric/subparametric finite elements.

We solve (2.7) for varying $h$ and $k$ values and record the errors in Table 2. In order to obtain good initial guesses for the Newton iteration, we use the same strategy as in Example 1, first solving the regularized problem (4.3). As expected, the convergence rates are exactly the same as the previous test with convergence rates of $O(h^4)$, $O(h^3)$, $O(h^2)$ in the $L^2$, $H^1$, and $H^2$-norms using cubic polynomials. Similar to the previous tests, we also observe that the method is also convergent using quadratic polynomials.

TABLE 2. Example 2. Errors of computed solution and rates of convergence with respect to $h$ with $\sigma = 100$ on curved domain.

| | $h$ | $\|u-u_h\|_{L^2(\Omega)}$ | rate | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{H^2(\mathcal{T}_h)}$ | rate |
|---|---|---|---|---|---|---|---|
| $k=2$ | 1/8 | 3.30E-03 | | 1.35E-01 | | 9.73E+00 | |
| | 1/16 | 1.03E-03 | 1.68 | 3.50E-02 | 1.95 | 4.90E+00 | 0.99 |
| | 1/32 | 1.66E-04 | 2.64 | 1.01E-02 | 1.79 | 2.41E+00 | 1.02 |
| | 1/64 | 3.06E-05 | 2.44 | 1.89E-03 | 2.42 | 1.12E+00 | 1.10 |
| | 1/128 | 1.11E-05 | 1.46 | 4.66E-04 | 2.02 | 5.54E-01 | 1.02 |
| $k=3$ | 1/8 | 9.40E-04 | | 4.91E-03 | | 8.73E-01 | |
| | 1/16 | 7.85E-05 | 3.58 | 5.82E-04 | 3.08 | 2.24E-01 | 1.96 |
| | 1/32 | 4.63E-06 | 4.08 | 5.66E-05 | 3.36 | 4.89E-02 | 2.20 |
| | 1/64 | 2.99E-07 | 3.95 | 6.51E-06 | 3.12 | 1.16E-02 | 2.07 |
| | 1/128 | 1.89E-08 | 3.98 | 7.91E-07 | 3.04 | 2.84E-03 | 2.04 |

4.3. **Example 3.** For the last test we solve (2.7) on the unit square $\Omega = (0,1)^2$ using quadratic polynomials and choose our data such that the exact solution is $u = \frac{(4(x_1^2+x_2^2))^{\frac{3}{4}}}{3}$. Unlike the previous two tests, the exact solution is not smooth, and one can readily check that $u \in W^{2,p}(\Omega)$ for all $p \in [1,4)$ but $u \notin \mathcal{C}^2(\overline{\Omega})$ [15].

Using the same strategy as the previous two tests, we first solve the regularized problem (4.3) with $\sigma = 100$ using quadratic polynomials to obtain some good initial guesses. We then solve (2.7) and record the errors in Table 3. Although a theoretical proof of such a result has not been proven, the numerical experiments indicate that the method is convergent with rates of $O(h^2)$, $O(h^{\frac{3}{2}})$, and $O(h^{\frac{1}{2}})$ in the $L^2$, $H^1$, and $H^2$-norms. The suboptimal rates are most likely due to the low regularity of the solution.

[1]We define the piecewise $H^2$-seminorm as $|u-u_h|_{H^2(\mathcal{T}_h)} = \left(\sum_{T\in\mathcal{T}_h} |u-u_h|^2_{H^2(T)}\right)^{\frac{1}{2}}$.

TABLE 3. Example 3. Errors of computed solution and rates of convergence with respect to $h$ with $\sigma = 100$ for nonsmooth solution.

| | $h$ | $\|u-u_h\|_{L^2(\Omega)}$ | rate | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{H^2(\mathcal{T}_h)}$ | rate |
|---|---|---|---|---|---|---|---|
| $k=2$ | 1/8 | 2.96E-05 | | 2.61E-03 | | 2.52E-01 | |
| | 1/16 | 6.33E-06 | 2.22 | 9.46E-04 | 1.46 | 1.79E-01 | 0.49 |
| | 1/32 | 1.48E-06 | 2.09 | 3.44E-04 | 1.46 | 1.27E-01 | 0.49 |
| | 1/64 | 3.73E-07 | 1.99 | 1.26E-04 | 1.45 | 9.03E-02 | 0.50 |
| | 1/128 | 9.74E-08 | 1.94 | 4.66E-05 | 1.43 | 6.41E-02 | 0.50 |

## 5. Concluding remarks

In this paper, we have developed and analyzed $\mathcal{C}^0$ penalty methods for the fully nonlinear Monge-Ampère equation. To build convergent numerical schemes, we constructed discretizations such that the resulting discrete linearization is stable, symmetric, and consistent with the continuous linearization. With this in hand, we proved existence of the numerical solution as well as derived quasi-optimal error estimates using a simple fixed-point technique.

The methodology developed in this paper has been applied to the three dimensional Monge-Ampère equation [8]. Due to the simplicity and flexibility of the method, it is also relatively straightforward to formulate discretizations for more general Monge-Ampère equations in which the function $f$ depends on $\nabla u$ and $u$. We plan to address the convergence properties of such problems in the near future.

## References


[1] F. E. Baginski and N. Whitaker, *Numerical solutions of boundary value problems for K-surfaces in $R^3$*, Numer. Methods for PDEs, 12(4):525–546, 1996. MR1396470 (97h:65152)

[2] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4(3):271–283, 1991. MR1115933 (92d:35137)

[3] J. D. Benamou, B. D. Froese, and A. M. Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, M2AN Math. Model. Numer. Anal., DOI 10.1051/m2an/2010017, 2010.

[4] C. Bernardi, *Optimal finite element interpolation on curved domains*, SIAM J. Numer. Anal., 26(5):1212–1240, 1989. MR1014883 (91a:65228)

[5] K. Böhmer, *On finite element methods for fully nonlinear elliptic equations of second order*, SIAM J. Numer. Anal., 46(3):1212–1249, 2008. MR2390991 (2009b:35092)

[6] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring. I.*, Math. Comp. 47(175):103–134, 1986. MR842125 (87m:65174)

[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, third edition, Springer, 2008. MR2373954 (2008m:65001)

[8] S. C. Brenner and M. Neilan, *Finite element approximations of the three dimensional Monge-Ampère equation*, submitted.

[9] C. J. Budd, W. Huang, and R. D. Russell, *Adaptivity with moving grids*, Acta Numerica 18:111–241, 2009. MR2506041 (2010c:65165)

[10] L. A. Caffarelli, L. Nirenberg, and J. Spruck, *The Dirichlet problem for nonlinear second-order elliptic equations I. Monge-Ampère equation*, Comm. Pure Appl. Math. 37(3):369–402, 1984. MR739925 (87f:35096)

[11] L. A. Caffarelli and M. Milman, *Monge-Ampère Equation: Applications to Geometry and Optimization*, American Mathematical Society, Providence, RI, 1999. MR1660738 (99f:00018)

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978. MR0520174 (58:25001)

[13] COMSOL Multiphysics Software Package, http://www.comsol.com.

[14] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, John Wiley & Sons, Inc., New York, 1989. MR1013360 (90k:35001)

[15] E. J. Dean and R. Glowinski, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Comput. Methods Appl. Mech. Engrg., 195(13-16):1344–1386, 2006. MR2203972 (2006i:65191)

[16] G. L. Delzanno, L. Chacón, J. M. Finn, Y. Chung, and G. Lapenta, *An optimal robust equidistribution method for two-dimensional grid adaptation based on Monge-Kantorovich optimization*, J. Comput. Phys. 227(23):9841–9864, 2008. MR2469037 (2010b:65286)

[17] L. C. Evans, *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI, 1998. MR1625845 (99e:35001)

[18] X. Feng and M. Neilan, *Vanishing moment method and moment solutions for second order fully nonlinear partial differential equations*, J. Scient. Comp., 38(1):74–98, 2009. MR2472219 (2010a:65234)

[19] X. Feng and M. Neilan, *Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method*, SIAM J. Numer. Anal. 47(2):1226–1250, 2009. MR2485451 (2010a:65235)

[20] X. Feng and M. Neilan, *Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation*, J. Sci. Comput., DOI: 10.107/S10915-010-9439-1.

[21] X. Feng and M. Neilan, *The vanishing moment method for fully nonlinear second order partial differential equations: formulation, theory, and numerical analysis*, submitted.

[22] B. D. Froese and A. M. Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher*, arXiv:1007.0765, 2010.

[23] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001. MR1814364 (2001k:35004)

[24] P. Grisvard, *Elliptic Problems on Nonsmooth Domains*, Pitman Publishing Inc., 1985. MR775683 (86m:35044)

[25] C. E. Gutiérrez, *The Monge-Ampère Equation*, volume 44 of *Progress in Nonlinear Differential Equations and Their Applications*, Birkhauser, Boston, MA, 2001. MR1829162 (2002e:35075)

[26] G. Loeper and F. Rapetti, *Numerical solution of the Monge-Ampère equation by a Newton's algorithm*, C. R. Math. Acad. Sci. Paris 340(4):319–324, 2005. MR2121899

[27] M. Neilan, *A nonconforming Morley finite element method for the fully nonlinear Monge-Ampère equation*, Numer. Math., 115(3):371–394, 2010.

[28] J. A. Nitsche, *Über ein Variationspirinzip zur Lösung Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unteworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36:9–15, 1971. MR0341903 (49:6649)

[29] A. M. Oberman, *Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. Ser. B, 10(1):221–238, 2008. MR2399429 (2009f:35101)

[30] V. I. Oliker and L. D. Prussner, *On the numerical solution of the equation* $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - (\frac{\partial^2 z}{\partial x \partial y})^2 = f$ *and its discretizations*, Numer. Math., 54:271–293, 1988. MR971703 (90h:65164)

[31] C. Villani, *Topics in Optimal Transportation*, volume 58 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2003. MR1964483 (2004e:90003)

Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803
*E-mail address*: brenner@math.lsu.edu

Department of Mathematics, Indian Institute of Science, Bangalore, 560012
*E-mail address*: gudi@math.iisc.ernet.in

Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803
*E-mail address*: neilan@math.lsu.edu

Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803
*E-mail address*: sung@math.lsu.edu