ESAIM: M2AN 46 (2012) 979–1001
DOI: 10.1051/m2an/2011067

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

FINITE ELEMENT APPROXIMATIONS OF THE THREE DIMENSIONAL MONGE-AMPÈRE EQUATION

Susanne Cecelia Brenner[1] and Michael Neilan[2]

Abstract. In this paper, we construct and analyze finite element methods for the three dimensional Monge-Ampère equation. We derive methods using the Lagrange finite element space such that the resulting discrete linearizations are symmetric and stable. With this in hand, we then prove the well-posedness of the method, as well as derive quasi-optimal error estimates. We also present some numerical experiments that back up the theoretical findings.

Mathematics Subject Classification. 65N30, 35J60.

Received July 19, 2010. Revised May 19, 2011.
Published online February 13, 2012.

1. Introduction

Let u be a smooth solution to the Monge-Ampère equation [7, 20, 27]:

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

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

Here, $\Omega \subset \mathbb{R}^3$ is either a strictly convex polygonal domain or a strictly convex domain with smooth boundary, f is a strictly positive function, and $\det(D^2u)$ denotes the determinant of the Hessian matrix D^2u. We assume that $u \in H^s(\Omega)$ for some $s > 7/2$ and is strictly convex. In the case when Ω is smooth, the regularity and strict convexity of u follows from the smoothness of f and g by the results in Caffarelli-Nirenberg-Spruck [8] (also see [28], Chap. 4).

The present article is motivated by the results in [5]. Here, the authors constructed convergent finite element methods for the two dimensional Monge-Ampère equation using the popular and simple Lagrange finite element spaces. In order to build convergent methods, the authors constructed consistent numerical schemes such that the resulting discrete linearizations are stable. As emphasized in [5], this simple idea leads to not-so-obvious discretizations.

Keywords and phrases. Monge-Ampère equation, three dimensions, finite element method, convergence analysis.

[1] Department of Mathematics and Center for Computation & Technology, Louisiana State University, Baton Rouge, 70803 LA, USA. brenner@math.lsu.edu,
Supported in part by the National Science Foundation under Grant Numbers DMS-07-13835 and DMS-10-16332.

[2] Department of Mathematics, University of Pittsburgh, 15260 PA, USA. neilan@pitt.edu,
Supported in part by the National Science Foundation under Grant Numbers DMS-09-02683 and DMS-11-15421.

Article published by EDP Sciences

© EDP Sciences, SMAI 2012

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

In this paper, we expand on these results and study the three dimensional case. Similar to the analysis of the two dimensional counterpart, we use Banach's fixed point theorem as our main tool to prove existence of a solution to the discrete problem as well as derive quasi-optimal order error estimates. Although the general strategy is similar, the fine details of the analysis in the three dimensional case prove to be much more difficult. The underlying reason for the added difficulty is that the mapping $u \to \det(D^2 u)$ is cubic (rather than quadratic in 2D). As a result, the analysis of the 2D case does not carry over, and new techniques must be introduced (*cf.* Rem. 3.1).

There has been a recent flux of papers on numerical methods for the Monge-Ampère equation. However, the three dimensional case is noticeably less prevalent in the literature. We give a brief review in this direction. In [16,17], Froese and Oberman generalized the two dimensional finite difference scheme given in [25] by constructing wide-stencil finite difference methods for the Monge-Ampère equation in dimensions greater than two. Using the framework developed by Barles and Souganidis [1], the authors proved convergence of their method by showing their scheme is monotone, consistent, and stable. In [26] Sorenson and Glowinski considered numerical methods for a σ_2-operator problem, which can be written as a three-dimensional Monge-Ampère-type equation. Extending the previous work in [11], the authors used a least-squares methodology to compute the solution of the nonlinear problem. Böhmer [3] introduced a projection method using C^1 finite element functions and analyzes the method using consistency and stability arguments. However, it is very difficult to construct C^1 finite element spaces in three dimensions and would require the use of piecewise polynomials of degree nine or higher [29]. No numerical experiments were given in [3]. Feng and the second author considered fourth order singular perturbations of (1.1) by adding a small multiple of the biharmonic operator to the PDE [14]. Two different numerical methods for the three dimensional regularized problem were proposed in [15,22]. Finally, we mention the method of Zheligovsky *et al.* [30] who develop numerical methods for the Monge-Ampère equation with periodic boundary conditions based on its Fourier integral form.

In contrast to the C^1 finite element method, the method proposed in this paper is relatively simple to implement and is computationally efficient. Moreover, unlike the scheme in [26], the method is provably convergent for smooth solutions. Furthermore, the method can handle curved boundaries and can naturally be extended to handle more general Monge-Ampère equations such as the equation of prescribed Gauss curvature [18].

The results in this paper are useful for many applications in differential geometry. For other applications (such as optimal transport [28]) it is important to capture weak solutions (*i.e.* viscosity or Aleksandrov solutions). Although the numerical experiments below indicate that the regularity condition $u \in H^s(\Omega),\ s > 7/2$ can be relaxed (*cf.* Sect. 5), it is not clear how to extend the analysis to the case of nonclassical solutions. This is because the analysis is based on the linearization of the numerical scheme which relies on the smoothness of the Hessian matrix $D^2 u$. One option to compute weak solutions is to use the proposed method in conjunction with the vanishing moment method (*cf.* [14,15,22] and Sect. 5). Preliminary numerical experiments of this concept are promising, but the convergence analysis has not been explored yet.

The rest of the paper is organized as follows. In Section 2 we set the notation and state some standard identities and inequalities. In Section 3 we derive the finite element method for the Monge-Ampère equation so that the resulting discrete linearization is stable and symmetric. With this in hand, in Section 4 we use a fixed-point argument to simultaneously prove well-posedness of the method as well as derive quasi-optimal error estimates. We end this section with L^2 error estimates obtained by a duality argument. In Section 5 we present three numerical experiments which back up the theoretical findings. We end the paper with some concluding remarks.

2. Notation and some preliminary results

Let $\mathcal{T}_h$ be a quasi-uniform, simplicial, and conforming triangulation [2,4,9] of the domain Ω such that each simplex has at most one curved face. We denote by $\mathcal{F}_h^i$ the set of interior faces, $\mathcal{F}_h^b$ the set of boundary faces, and $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^b$ the set of all of the faces in $\mathcal{T}_h$. We set $h_T = \operatorname{diam}(T)$ for all $T \in \mathcal{T}_h$, $h_{\mathcal{F}} = \operatorname{diam}(\mathcal{F})$ for all $\mathcal{F} \in \mathcal{F}_h$, and note that by the assumption of quasi-uniformity of the mesh, $h_T \approx h_{\mathcal{F}} \approx h := \max_{T \in \mathcal{T}_h} h_T$.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

For a face $\mathcal{F} \in \mathcal{F}_h^i$, there exist two simplexes $T^+, T^- \in \mathcal{T}_h$ such that $\mathcal{F} = \partial T^+ \cap \partial T^-$. We shall denote the average of a piecewise smooth vector function $\mathbf{w} \in \mathbb{R}^3$ across $\mathcal{F}$ as

$$\{\!\{\boldsymbol{w}\}\!\}\big|_{\mathcal{F}} = \frac{1}{2}\big(\boldsymbol{w}^+\big|_{\mathcal{F}} + \boldsymbol{w}^-\big|_{\mathcal{F}}\big),$$

where we use the notation $\boldsymbol{w}^{\pm} = \boldsymbol{w}\big|_{T^{\pm}}$. In the case that $\mathcal{F}$ is a boundary face with $\mathcal{F} = \partial\Omega \cap \partial T^+$, we set

$$\{\!\{\boldsymbol{w}\}\!\}\big|_{\mathcal{F}} = \boldsymbol{w}^+|_{\mathcal{F}}.$$

Similarly for a matrix $\underline{\boldsymbol{w}} \in \mathbb{R}^{3\times 3}$, we denote the average across $\mathcal{F}$ as

$$\begin{aligned} \{\!\{\underline{\boldsymbol{w}}\}\!\}\big|_{\mathcal{F}} &= \frac{1}{2}\big(\underline{\boldsymbol{w}}^+\big|_{\mathcal{F}} + \underline{\boldsymbol{w}}^-\big|_{\mathcal{F}}\big) && \text{if } \mathcal{F} = \partial T^+ \cap \partial T^- \in \mathcal{F}_h^i, \\ \{\!\{\underline{\boldsymbol{w}}\}\!\}\big|_{\mathcal{F}} &= \underline{\boldsymbol{w}}^+\big|_{\mathcal{F}} && \text{if } \mathcal{F} = \partial\Omega \cap \partial T^+ \in \mathcal{F}_h^b. \end{aligned}$$

Next, we define the jump of a vector function $\boldsymbol{w}$ (which is a scalar) as

$$\begin{aligned} [\![\boldsymbol{w}]\!]\big|_{\mathcal{F}} &= \boldsymbol{w}^+ \cdot \boldsymbol{n}_+\big|_{\mathcal{F}} + \boldsymbol{w}^- \cdot \boldsymbol{n}_-\big|_{\mathcal{F}} && \text{if } \mathcal{F} = \partial T^+ \cap \partial T^- \in \mathcal{F}_h^i, \\ [\![\boldsymbol{w}]\!]\big|_{\mathcal{F}} &= \boldsymbol{w} \cdot \boldsymbol{n}_+\big|_{\mathcal{F}} && \text{if } \mathcal{F} = \partial\Omega \cap \partial T^+ \in \mathcal{F}_h^b, \end{aligned}$$

where $\boldsymbol{n}_{\pm}$ denotes the unit outward normal of $T^{\pm}$.

We use $W^{m,q}(\Omega)$ to denote the set of all $L^q(\Omega)$ functions whose distributional derivatives up to order m are in $L^q(\Omega)$ and set $H^m(\Omega) = W^{m,2}(\Omega)$. We also define the piecewise Sobolev spaces as

$$W^{m,q}(\mathcal{T}_h) = \prod_{T\in\mathcal{T}_h} W^{m,q}(T), \qquad H^m(\mathcal{T}_h) = W^{m,2}(\mathcal{T}_h).$$

For a normed linear space X, we denote by X' its dual and $\langle\cdot,\cdot\rangle$ the pairing between X' and X.

Denote by $\nabla_h v$ the piecewise gradient of v, and by $D_h^2 v$ the piecewise Hessian matrix of v. We also set $\mathrm{cof}(D_h^2 v)$ to be the cofactor matrix of $D_h^2 v$; that is

$$\mathrm{cof}(D_h^2 v)_{ij} = (-1)^{i+j} \det(D_h^2 v\big|_{ij}) \quad i,j = 1,2,3,$$

where $D_h^2 v\big|_{ij}$ denotes the 2×2 matrix resulting from deleting the ith row and jth column of $D_h^2 v$.

We define the discrete (semi)norms

$$\begin{aligned} \|v\|_{W^{2,3}(\mathcal{T}_h)} = \Bigg(&\sum_{T\in\mathcal{T}_h} \Big(\|D_h^2 v\|_{L^3(T)}^3 + h_T \|D_h^2 v\|_{L^3(\partial T)}^3 \Big) \\ &+ \sum_{\mathcal{F}\in\mathcal{F}_h} \frac{1}{h_{\mathcal{F}}^2} \big\| [\![\nabla_h v]\!] \big\|_{L^3(\mathcal{F})}^3 + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}^5} \|v\|_{L^3(\mathcal{F})}^3 \Bigg)^{\frac{1}{3}}, \end{aligned} \tag{2.1}$$

$$\|v\|_{H^2(\mathcal{T}_h)} = \left(\|D_h^2 v\|_{L^2(\Omega)}^2 + \sum_{\mathcal{F}\in\mathcal{F}_h} \Big(h_{\mathcal{F}} \big\| \{\!\{D_h^2 v\}\!\} \big\|_{L^2(\mathcal{F})}^2 + \frac{1}{h_{\mathcal{F}}} \big\| [\![\nabla_h v]\!] \big\|_{L^2(\mathcal{F})}^2 \Big) + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}^3} \|v\|_{L^2(\mathcal{F})}^2 \right)^{\frac{1}{2}}, \tag{2.2}$$

$$\|v\|_{H^1(\mathcal{T}_h)} = \left(\|\nabla_h v\|_{L^2(\Omega)}^2 + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \Big(h_{\mathcal{F}} \|\nabla_h v\|_{L^2(\mathcal{F})}^2 + \frac{1}{h_{\mathcal{F}}} \|v\|_{L^2(\mathcal{F})}^2 \Big) \right)^{\frac{1}{2}}, \tag{2.3}$$

$$\|v\|_{H^{-1}(\mathcal{T}_h)} = \sup_{w\in V_h} \frac{\langle v, w\rangle}{\|w\|_{H^1(\mathcal{T}_h)}}. \tag{2.4}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Remark 2.1. The norms (2.1)–(2.4) are well-defined for functions in $W^{3,3}(\mathcal{T}_h)$.

Let k be an integer greater than or equal to three and define the finite element space $V_h \subset H^1(\Omega)$ as follows:

- if $T \in \mathcal{T}_h$ does not have a curved face, 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 face, then $v\big|_T$ is a polynomial of degree $\leq k$ in the curvilinear coordinates of T that are associated with the reference simplex (Ex. 2, p. 1216 of [2]).

Remark 2.2. The reason for the requirement $k \geq 3$ as well as the regularity condition $u \in H^s(\Omega)$ for $s > 7/2$ will be made clear in the proof of Theorem 4.12.

We end this section with some preliminary results and identities that are needed in both the derivation of the scheme and the convergence analysis.

Lemma 2.3 (divergence free row property of cofactor matrices, [13])**.** *For any piecewise smooth function v,*

$$\nabla_h \cdot \big(\operatorname{cof}(D_h^2 v)_i\big) = \sum_{j=1}^{3} \frac{\partial}{\partial x_j}\big(\operatorname{cof}(D_h^2 v)_{ij}\big) = 0 \qquad \text{for } i = 1,2,3, \tag{2.5}$$

where $\operatorname{cof}(D_h^2 v)_i$ and $\operatorname{cof}(D_h^2 v)_{ij}$ denote respectively the ith row and the (i,j)-entry of the cofactor matrix $\operatorname{cof}(D_h^2 v)$, and $\nabla_h \cdot$ denotes the piecewise divergence operator.

Lemma 2.4 (determinant and cofactor expansions)**.** *For any piecewise smooth v, w there holds*

$$\det(D_h^2(v+w)) = \det(D_h^2 v) + \nabla_h \cdot \big(\operatorname{cof}(D_h^2 v)\nabla_h w\big) + \nabla_h \cdot (\operatorname{cof}(D_h^2 w)\nabla_h v) + \det(D_h^2 w), \tag{2.6}$$

and

$$\operatorname{cof}(D_h^2(v+w)) = \operatorname{cof}(D_h^2 v) + \operatorname{cof}(D_h^2 w) + A(v,w), \tag{2.7}$$

where $A(v,w) \in \mathbb{R}^{3\times 3}$ is defined such that

$$A(v,w)_{ij} = (-1)^{i+j}\operatorname{cof}(D_h^2 v\big|_{ij}) : D_h^2 w\big|_{ij} \qquad i,j = 1,2,3, \tag{2.8}$$

and $D_h^2 v\big|_{ij}$ denotes the 2×2 matrix resulting from deleting the ith row and jth column of $D_h^2 v$. Here, $B : C$ denotes the Frobenius inner product between two matrices B and C, i.e., $B : C = \sum_{i,j} B_{ij} C_{ij}$.

Proof. For any matrices 3×3 matrices B and C, there holds [21]

$$\begin{aligned}
\det(B+C) &= \sum_{\nu\in S_3} \operatorname{sign}(\nu) \prod_{i=1}^{3} \big[B_{i,\nu(i)} + C_{i,\nu(i)}\big] \\
&= \sum_{\nu\in S_3} \operatorname{sign}(\nu) \prod_{i=1}^{3} B_{i,\nu(i)} + \sum_{\nu\in S_3} \operatorname{sign}(\nu) \sum_{i=1}^{3} C_{i,\nu(i)} \prod_{j\neq i} B_{j,\nu(j)} \\
&\quad + \sum_{\nu\in S_3} \operatorname{sign}(\nu) \sum_{i=1}^{3} B_{i,\nu(i)} \prod_{j\neq i} C_{j,\nu(j)} + \sum_{\nu\in S_3} \operatorname{sign}(\nu) \prod_{i=1}^{3} C_{i,\nu(i)} \\
&= \det(B) + \sum_{i,j=1}^{3} C_{i,j} \Bigg(\sum_{\substack{\nu\in S_3 \\ \nu(i)=j}} \operatorname{sign}(\nu) \prod_{j\neq i} B_{j,\nu(j)} \Bigg) \\
&\quad + \sum_{i,j=1}^{3} B_{i,j} \Bigg(\sum_{\substack{\nu\in S_3 \\ \nu(i)=j}} \operatorname{sign}(\nu) \prod_{j\neq i} C_{j,\nu(j)} \Bigg) + \det(C) \\
&= \det(B) + \operatorname{cof}(B) : C + \operatorname{cof}(C) : B + \det(C),
\end{aligned} \tag{2.9}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

where S_3 consists of all permutations of the set $\{1,2,3\}$. It then follows from (2.9) and Lemma 2.3 that

$$\begin{aligned}\det(D_h^2(v+w)) &= \det(D_h^2 v) + \mathrm{cof}(D_h^2 v) : D_h^2 w + \mathrm{cof}(D_h^2 w) : D_h^2 v + \det(D_h^2 w)\\ &= \det(D_h^2 v) + \nabla\cdot(\mathrm{cof}(D_h^2 v)\nabla_h w) + \nabla\cdot(\mathrm{cof}(D_h^2 w)\nabla_h v) + \det(D_h^2 w).\end{aligned}$$

To prove (2.7), we first use the definition of cofactor matrices.

$$\mathrm{cof}(D_h^2(v+w))_{ij} = (-1)^{i+j}\det\left(D_h^2(v+w)\big|_{ij}\right) \qquad i,j=1,2,3.$$

It can readily be checked that (*cf.* [6])

$$\det\left(D_h^2(v+w)\big|_{ij}\right) = \det\left(D_h^2 v\big|_{ij}\right) + \det\left(D_h^2 w\big|_{ij}\right) + \mathrm{cof}\left(D_h^2 v|\right)_{ij} : D_h^2 w\big|_{ij},$$

and therefore by (2.8),

$$\begin{aligned}\mathrm{cof}(D_h^2(v+w))_{ij} &= (-1)^{i+j}\left(\det(D_h^2 v\big|_{ij}\right) + \det\left(D_h^2 w\big|_{ij}\right) + \mathrm{cof}\left(D_h^2 v\big|_{ij}\right) : D_h^2 w\big|_{ij}\right)\\ &= \mathrm{cof}(D_h^2 v)_{ij} + \mathrm{cof}(D_h^2 w)_{ij} + A(v,w)_{ij}.\end{aligned}$$

□

Remark 2.5. The mapping $(v,w)\to A(v,w)$ is bilinear and symmetric.

Remark 2.6. In order to avoid the proliferation of constants, we shall use the notation $A\lesssim B$ to represent the relation $A\le$ constant $\times B$, where the constant is independent of the mesh parameter h and the penalty parameter σ.

Lemma 2.7 (inverse estimates)**.** *For any $v\in V_h$, there holds*

$$h^{1/2}\|v\|_{L^\infty(\Omega)} + h\|v\|_{H^2(\mathcal{T}_h)} + h^{3/2}\|\nabla v\|_{L^\infty(\Omega)} + h^{3/2}\|v\|_{W^{2,3}(\mathcal{T}_h)} \lesssim \|v\|_{H^1(\mathcal{T}_h)}. \tag{2.10}$$

Proof. By the inverse inequality [4,9] followed by a Sobolev embedding, we have

$$\|v\|_{L^\infty(\Omega)} \lesssim h^{-1/2}\|v\|_{L^6(\Omega)} \lesssim h^{-1/2}\|v\|_{H^1(\Omega)} \le h^{-1/2}\|v\|_{H^1(\mathcal{T}_h)}.$$

The other three inequalities in (2.10) follow from standard scaling arguments. □

Lemma 2.8 (approximation properties of V_h [2])**.** *Let m,ℓ be two integers such that $0\le m\le \ell\le k+1$. Then for any $\chi\in H^\ell(\Omega)$, there exists a $v\in V_h$ such that*

$$\left(\sum_{T\in\mathcal{T}_h}\|\chi-v\|_{H^m(T)}^2\right)^{\frac12} \lesssim h^{\ell-m}\|\chi\|_{H^\ell(\Omega)}.$$

Furthermore if $H^\ell(\Omega)\subset W^{m,3}(\Omega)$, then

$$\left(\sum_{T\in\mathcal{T}_h} h_T^{3/2}\|\chi-v\|_{W^{m,3}(T)}^3\right)^{\frac13} \lesssim h^{\ell-m}\|\chi\|_{H^\ell(\Omega)}.$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

3. Derivation of the finite element method

To derive the finite element method for (1.1), we follow arguments similar to those presented in [5] where the two dimensional case was considered. To motivate the method, we first note that the linearized Monge-Ampère problem reads [6]

$$-\nabla\cdot\big(\mathrm{cof}(D^2u)\nabla w\big)=0 \qquad \text{in } \Omega, \tag{3.1a}$$

$$w=0 \qquad \text{on } \partial\Omega. \tag{3.1b}$$

The finite element discretization of the linearization (3.1) using Nitsche's method [24] is defined by

$$\begin{aligned}(w,v)\to &\int_\Omega\big(\mathrm{cof}(D^2u)\nabla_h w\big)\cdot\nabla_h v\,\mathrm{d}x-\sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\big[\![\mathrm{cof}(D^2u)\nabla_h w]\!\big]\,v\,\mathrm{d}s\\ &-\sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\big[\![\mathrm{cof}(D^2u)\nabla_h v]\!\big]\,w\,\mathrm{d}s+\sigma\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\int_{\mathcal{F}}vw\,\mathrm{d}s.\end{aligned} \tag{3.2}$$

Our goal is to construct a scheme such that the linearization of the discrete Monge-Ampère problem is the discretization of the linearized Monge-Ampère problem (3.1) by Nitsche's method; *i.e.*, that the linearized discrete Monge-Ampère problem is (3.2). With such a scheme in hand, the discrete linearization will be stable (*cf.* Rem. 4.1) which is a key ingredient in the convergence analysis.

To this end, for $w\in W^{3,3}(\mathcal{T}_h)$ and $v\in V_h$, we first state the following identity, which follows from (1.1a) and Lemmas 2.3–2.4:

$$\begin{aligned}\int_\Omega\Big(f-\det(D_h^2(u+w))\Big)v\,\mathrm{d}x&=-\int_\Omega\Big(\det(D_h^2w)+\mathrm{cof}(D_h^2w):D_h^2u\Big)v\,\mathrm{d}x-\int_\Omega\Big(\nabla_h\cdot\big(\mathrm{cof}(D^2u)\nabla_h w\big)\Big)v\,\mathrm{d}x\\&=-\int_\Omega\Big(\det(D_h^2w)+\mathrm{cof}(D_h^2w):D^2u\Big)v\,\mathrm{d}x\\&\quad+\int_\Omega\big(\mathrm{cof}(D^2u)\nabla_h w\big)\cdot\nabla_h v\,\mathrm{d}x-\sum_{\mathcal{F}\in\mathcal{F}_h}\int_{\mathcal{F}}\big[\![\mathrm{cof}(D^2u)\nabla_h w]\!\big]\,v\,\mathrm{d}s.\end{aligned}$$

Therefore, by rearranging the last term in the expression above we have

$$\begin{aligned}&\int_\Omega\Big(f-\det(D_h^2(u+w))\Big)v\,\mathrm{d}x+\sum_{\mathcal{F}\in\mathcal{F}_h^i}\int_{\mathcal{F}}\big[\![\mathrm{cof}(D^2u)\nabla_h w]\!\big]\,v\,\mathrm{d}s\\&=-\int_\Omega\Big(\det(D_h^2w)+\mathrm{cof}(D_h^2w):D^2u\Big)v\,\mathrm{d}x\\&\quad+\int_\Omega\big(\mathrm{cof}(D^2u)\nabla_h w\big)\cdot\nabla_h v\,\mathrm{d}x-\sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\big[\![\mathrm{cof}(D^2u)\nabla_h w]\!\big]\,v\,\mathrm{d}s.\end{aligned}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Adding terms on both sides of the equation and noting $[\![\nabla u]\!]\big|_{\mathcal{F}} = 0\ \forall \mathcal{F} \in \mathcal{F}_h^i$, we have by (2.7) that

$$\begin{aligned}
&\int_\Omega \Big(f - \det\big(D_h^2(u+w)\big)\Big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D_h^2(u+w)) \}\!\} \nabla_h (u+w)]\!] v \,\mathrm{d}s \qquad (3.3)\\
&\quad = \int_\Omega \Big(f - \det\big(D_h^2(u+w)\big)\Big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D^2 u) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s\\
&\qquad + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D_h^2 w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ A(u,w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s\\
&\quad = \int_\Omega \big(\mathrm{cof}(D^2 u)\nabla_h w\big)\cdot\nabla_h v \,\mathrm{d}x - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla_h w]\!] v \,\mathrm{d}s\\
&\qquad - \int_\Omega \big(\mathrm{cof}(D_h^2 w) : D^2 u\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ A(u,w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s\\
&\qquad - \int_\Omega \big(\det(D_h^2 w)\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D_h^2 w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s.
\end{aligned}$$

Note that the bilinear form

$$(w,v) \to \int_\Omega \big(\mathrm{cof}(D^2 u)\nabla_h w\big)\cdot\nabla_h v \,\mathrm{d}x - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla_h w]\!] v \,\mathrm{d}s$$

that appears on the right-hand side of (3.3) can be symmetrized and stabilized to become the consistent and stable bilinear form defined by (3.2). Imposing symmetrization and stabilization, (3.3) becomes

$$\begin{aligned}
&\int_\Omega \Big(f - \det\big(D_h^2(u+w)\big)\Big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D_h^2(u+w)) \}\!\} \nabla_h (u+w)]\!] v \,\mathrm{d}s \qquad (3.4)\\
&\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} [\![\mathrm{cof}(D_h^2(u+w))\nabla_h v]\!] (u+w) \,\mathrm{d}s + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D_h^2(u+w))\nabla_h v]\!] g \,\mathrm{d}s\\
&\quad + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} (u+w) v \,\mathrm{d}s - \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} g v \,\mathrm{d}s\\
&= \int_\Omega \big(\mathrm{cof}(D^2 u)\nabla_h w\big)\cdot\nabla_h v \,\mathrm{d}x - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla_h w]\!] v \,\mathrm{d}s\\
&\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla_h v]\!] w \,\mathrm{d}s + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} w v \,\mathrm{d}s\\
&\quad - \int_\Omega \big(\mathrm{cof}(D_h^2 w) : D^2 u\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ A(u,w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![A(u,w)\nabla_h v]\!] w \,\mathrm{d}s\\
&\quad - \int_\Omega \big(\det(D_h^2 w)\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} [\![\{\!\{ \mathrm{cof}(D_h^2 w) \}\!\} \nabla_h w]\!] v \,\mathrm{d}s\\
&\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} [\![\mathrm{cof}(D_h^2 w)\nabla_h v]\!] w \,\mathrm{d}s,
\end{aligned}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

where σ is a positive penalty parameter independent of h. Equation (3.4) can be written compactly as

$$F(u+w) = Lw + Qw + Rw, \tag{3.5}$$

where the operators $F, R, Q, L : W^{3,3}(\mathcal{T}_h) \to V_h'$ are defined as

$$\begin{aligned}\langle Fw, v\rangle &= \int_\Omega \Big(f - \det(D_h^2 w)\Big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} \big[\!\big[\{\!\{ \mathrm{cof}(D_h^2 w) \}\!\} \nabla_h w \big]\!\big] v \,\mathrm{d}s \\ &\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[\mathrm{cof}(D_h^2 w) \nabla_h v \big]\!\big] (w-g) \,\mathrm{d}s + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} (w-g) v \,\mathrm{d}s,\end{aligned} \tag{3.6}$$

$$\begin{aligned}\langle Rw, v\rangle &= -\int_\Omega \big(\det(D_h^2 w)\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} \big[\!\big[\{\!\{ \mathrm{cof}(D_h^2 w) \}\!\} \nabla_h w \big]\!\big] v \,\mathrm{d}s \\ &\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[\mathrm{cof}(D_h^2 w) \nabla_h v \big]\!\big] w \,\mathrm{d}s,\end{aligned} \tag{3.7}$$

$$\begin{aligned}\langle Qw, v\rangle &= -\int_\Omega \big(\mathrm{cof}(D_h^2 w) : D^2 u\big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} \big[\!\big[\{\!\{ A(u,w) \}\!\} \nabla_h w \big]\!\big] v \,\mathrm{d}s \\ &\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[A(u,w) \nabla_h v \big]\!\big] w \,\mathrm{d}s,\end{aligned} \tag{3.8}$$

$$\begin{aligned}\langle Lw, v\rangle &= \int_\Omega \big(\mathrm{cof}(D^2 u) \nabla_h w\big) \cdot \nabla_h v \,\mathrm{d}x - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[\mathrm{cof}(D^2 u) \nabla_h w \big]\!\big] v \,\mathrm{d}s \\ &\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[\mathrm{cof}(D^2 u) \nabla_h v \big]\!\big] w \,\mathrm{d}s + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} vw \,\mathrm{d}s.\end{aligned} \tag{3.9}$$

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

$$F_h u_h = 0, \tag{3.10}$$

that is,

$$\begin{aligned}&\int_\Omega \Big(f - \det(D_h^2 u_h)\Big) v \,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \int_{\mathcal{F}} \big[\!\big[\{\!\{ \mathrm{cof}(D_h^2 u_h) \}\!\} \nabla u_h \big]\!\big] v \,\mathrm{d}s \\ &\quad - \sum_{\mathcal{F}\in\mathcal{F}_h^b} \int_{\mathcal{F}} \big[\!\big[\mathrm{cof}(D_h^2 u_h) \nabla v \big]\!\big] (u_h - g) \,\mathrm{d}s + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_F (u_h - g) v \,\mathrm{d}s = 0 \qquad \forall v \in V_h.\end{aligned}$$

Remark 3.1. The finite element method (3.10) is the same as the two dimensional method studied in [5]. However, the decomposition (3.5) is not, as the operator Q does not appear in the two dimensional case. This difference as well as the fact that R is cubic and not quadratic in its arguments makes the analysis a bit more involved than the two dimensional counterpart.

4. Convergence analysis

4.1. Strategy and some preliminary estimates

The proofs of both existence as well as error estimates of the finite element method (3.10) proceed by using a relatively simple linearization fixed-point strategy. To this end, we let $L_h : V_h \to V_h'$ be the restriction of L to

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

V_h, that is,

$$\langle L_h v, w\rangle = \langle Lv, w\rangle \qquad \forall v, w \in V_h.$$

We then define $u_{c,h} \in V_h$ such that

$$u_{c,h} = L_h^{-1} L u, \tag{4.1}$$

where $L_h^{-1} : V_h' \to V_h$ denotes the inverse operator of L_h.

Lemma 4.1. *For σ sufficiently large, the operator L_h is invertible, and we have the following estimates:*

$$\|L_h^{-1}\phi\|_{H^1(\mathcal{T}_h)} \lesssim \|\phi\|_{H^{-1}(\mathcal{T}_h)} \qquad \forall \phi \in V_h', \tag{4.2}$$

$$\|Lw\|_{H^{-1}(\mathcal{T}_h)} \lesssim (1+\sigma)\|w\|_{H^1(\mathcal{T}_h)} \qquad \forall w \in H^2(\mathcal{T}_h) \cap H^1(\Omega), \tag{4.3}$$

and

$$\|u - u_{c,h}\|_{H^1(\mathcal{T}_h)} + h\|u - u_{c,h}\|_{H^2(\mathcal{T}_h)} \lesssim (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}, \tag{4.4}$$

where $\ell = \min\{s, k+1\}$.

The proof of Lemma 4.1 can be found, *e.g.*, in [5], Lemma 3.1, and also [24]. For completeness, we provide a proof of Lemma 4.1 in the appendix.

Remark 4.2. For the rest of the paper, we assume that σ is large enough so that (4.2)–(4.4) hold.

Lemma 4.3. *There holds the following estimate:*

$$\|u - u_{c,h}\|_{W^{2,3}(\mathcal{T}_h)} \lesssim (1+\sigma)h^{\ell-5/2}\|u\|_{H^\ell(\Omega)}. \tag{4.5}$$

Proof. By the triangle inequality and the inverse inequality (2.10), we have for any $v \in V_h$

$$\begin{aligned}\|u - u_{c,h}\|_{W^{2,3}(\mathcal{T}_h)} &\lesssim \|u - v\|_{W^{2,3}(\mathcal{T}_h)} + h^{-1/2}\|u_{c,h} - v\|_{H^2(\mathcal{T}_h)} \\ &\le \|u - v\|_{W^{2,3}(\mathcal{T}_h)} + h^{-1/2}\|u - v\|_{H^2(\mathcal{T}_h)} + h^{-1/2}\|u - u_{c,h}\|_{H^2(\mathcal{T}_h)}.\end{aligned}$$

The estimate (4.5) then follows from Lemma 2.8 and (4.4). □

Define the mapping $M : W^{3,3}(\mathcal{T}_h) \to V_h$ as

$$M = L_h^{-1}\big(L - F\big), \tag{4.6}$$

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

$$M_h = Id_h - L_h^{-1}F_h, \tag{4.7}$$

where Id_h denotes the identity operator on V_h. The goal now is to show that M_h has a unique fixed point in a neighborhood of $u_{c,h}$, which will be a solution to the finite element method (3.10).

To achieve this goal, we first note that by (4.6), (3.5), and (4.1)

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

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

and therefore,

$$u_{c,h} - Mw = L_h^{-1}\Big(R(w-u) + Q(w-u)\Big) \qquad \forall w \in W^{3,3}(\mathcal{T}_h), \tag{4.8}$$

and

$$Mw_1 - Mw_2 = L_h^{-1}\Big(R(w_2-u) - R(w_1-u) + Q(w_2-u) - Q(w_1-u)\Big) \qquad \forall w_1, w_2 \in W^{3,3}(\mathcal{T}_h). \tag{4.9}$$

To estimate the right-hand sides of (4.8) and (4.9), we introduce the Gâteaux derivative of Q and the first and second Gâteaux derivatives of R:

$$\begin{aligned} DQ[w](z) &= \lim_{t\to 0} \frac{Q(w+tz) - Q(w)}{t}, \\ DR[w](z) &= \lim_{t\to 0} \frac{R(w+tz) - R(w)}{t}, \\ D^2R[w](z,q) &= \lim_{t\to 0} \frac{DR[w+tq](z) - DR[w](z)}{t}. \end{aligned}$$

Remark 4.4. By (3.7)–(3.8) and the expansions (2.6)–(2.7),

$$\begin{aligned} \big\langle DQ[w](z), v\big\rangle = &-\int_\Omega \big(A(w,z) : D^2u\big)v\,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i}\int_{\mathcal{F}}\Big(\big[\!\big[\{\!\{A(u,z)\}\!\}\nabla_h w\big]\!\big] + \big[\!\big[\{\!\{A(u,w)\}\!\}\nabla_h z\big]\!\big]\Big)v\,\mathrm{d}s \\ &- \sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\Big(\big[\!\big[A(u,w)\nabla_h v\big]\!\big]z + \big[\!\big[A(u,z)\nabla_h v\big]\!\big]w\Big)\mathrm{d}s, \end{aligned} \tag{4.10}$$

$$\begin{aligned} \big\langle DR[w](z), v\big\rangle = &-\int_\Omega \big(\mathrm{cof}(D_h^2w) : D_h^2z\big)v\,\mathrm{d}x + \sum_{\mathcal{F}\in\mathcal{F}_h^i}\int_{\mathcal{F}}\Big(\big[\!\big[\{\!\{\mathrm{cof}(D_h^2w)\}\!\}\nabla_h z\big]\!\big] + \big[\!\big[\{\!\{A(w,z)\}\!\}\nabla_h w\big]\!\big]\Big)v\,\mathrm{d}s \\ &- \sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\Big(\big[\!\big[\mathrm{cof}(D_h^2w)\nabla_h v\big]\!\big]z + \big[\!\big[A(w,z)\nabla_h v\big]\!\big]w\Big)\mathrm{d}s, \end{aligned} \tag{4.11}$$

$$\begin{aligned} \big\langle D^2R[w](z,q), v\big\rangle = &-\int_\Omega \big(A(w,q) : D_h^2z\big)v\,\mathrm{d}x \\ &+ \sum_{\mathcal{F}\in\mathcal{F}_h^i}\int_{\mathcal{F}}\Big(\big[\!\big[\{\!\{A(w,q)\}\!\}\nabla_h z\big]\!\big] + \big[\!\big[\{\!\{A(w,z)\}\!\}\nabla_h q\big]\!\big] + \big[\!\big[\{\!\{A(q,z)\}\!\}\nabla_h w\big]\!\big]\Big)v\,\mathrm{d}s \\ &- \sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}}\Big(\big[\!\big[A(w,q)\nabla_h v\big]\!\big]z + \big[\!\big[A(w,z)\nabla_h v\big]\!\big]q + \big[\!\big[A(q,z)\nabla_h v\big]\!\big]w\Big)\mathrm{d}s. \end{aligned} \tag{4.12}$$

We note that the mapping $(w,z) \to DQ[w](z)$ is bilinear, and the mapping $(w,z,q) \to D^2R[w](z,q)$ is trilinear. Furthermore, we have the following symmetry properties [3]:

$$\begin{aligned} DQ[w](z) &= DQ[z](w), \\ D^2R[w](z,q) &= D^2R[q](z,w) = D^2R[q](w,z). \end{aligned}$$

[3] To see the second symmetry property of D^2R, set $\Phi(s,t) = R(w+sq+tz)$ and note that $D^2R[w](z,q) = \frac{\partial^2\Phi}{\partial s\partial t}(0,0) = \frac{\partial^2\Phi}{\partial t\partial s}(0,0) = D^2R[w](q,z)$.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Since $DQ[\cdot](\cdot)$ is bilinear, we have for any $z_1, z_2 \in W^{3,3}(\mathcal{T}_h)$

$$Q(z_1) - Q(z_2) = \int_0^1 DQ[tz_1 + (1-t)z_2](z_1 - z_2)\,\mathrm{d}t = DQ\left[\frac{1}{2}(z_1 + z_2)\right](z_1 - z_2).$$

In particular,

$$Q(w_1 - u) - Q(w_2 - u) = DQ\left[\frac{1}{2}(w_1 + w_2) - u\right](w_1 - w_2). \tag{4.13}$$

Moreover, using the trilinearity and symmetry of D^2R we have

$$\begin{aligned}
R(z_1) - R(z_2) &= \int_0^1 DR[tz_1 + (1-t)z_2](z_1 - z_2)\,\mathrm{d}t \\
&= \int_0^1 \left[\int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} DR\big[s(tz_1 + (1-t)z_2)\big](z_1 - z_2)\,\mathrm{d}s\right]\mathrm{d}t \\
&= \int_0^1 \left[\int_0^1 sD^2R[tz_1 + (1-t)z_2](z_1 - z_2, tz_1 + (1-t)z_2)\,\mathrm{d}s\right]\mathrm{d}t \\
&= \frac{1}{2}\int_0^1 D^2R[tz_1 + (1-t)z_2](z_1 - z_2, tz_1 + (1-t)z_2)\,\mathrm{d}t \\
&= \frac{1}{2}\int_0^1 D^2R[z_1 - z_2](tz_1 + (1-t)z_2, tz_1 + (1-t)z_2)\,\mathrm{d}t \\
&= \frac{1}{6}\Big(D^2R[z_1 - z_2](z_1, z_1) + D^2R[z_1 - z_2](z_1, z_2) + D^2R[z_1 - z_2](z_2, z_2)\Big).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
R(w_1 - u) - R(w_2 - u) = \frac{1}{6}\Big(&D^2R[w_1 - w_2](w_1 - u, w_1 - u) \\
&+ D^2R[w_1 - w_2](w_1 - u, w_2 - u) \\
&+ D^2R[w_1 - w_2](w_2 - u, w_2 - u)\Big).
\end{aligned} \tag{4.14}$$

In light of the identities (4.8)–(4.9) and (4.13)–(4.14) we must first derive estimates for the operators DQ and D^2R in order to show that M_h is a contraction mapping in a neighborhood of $u_{c,h}$. We establish these bounds in the following lemmas.

Lemma 4.5 (estimate of DQ)**.** *We have for any $w, z \in W^{3,3}(\mathcal{T}_h)$,*

$$\big\|DQ[w](z)\big\|_{H^{-1}(\mathcal{T}_h)} \lesssim h^{-1/2}\|w\|_{H^2(\mathcal{T}_h)}\|z\|_{H^2(\mathcal{T}_h)}. \tag{4.15}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Proof. By (4.10) and (2.10), we have for any $v \in V_h$

$$\begin{aligned}
\big\langle DQ[w](z), v\big\rangle &\leq \Bigg(\big\|A(z,w)\big\|_{L^1(\Omega)} \|D^2 u\|_{L^\infty(\Omega)} + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \Big(\big\|\{\!\{A(u,z)\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h w]\!]\big\|_{L^2(\mathcal{F})} \\
&\quad + \big\|\{\!\{A(u,w)\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h z]\!]\big\|_{L^2(\mathcal{F})} \Big) \Bigg) \|v\|_{L^\infty(\Omega)} \\
&\quad + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \Big(\big\|A(u,w)\big\|_{L^2(\mathcal{F})} \|z\|_{L^2(\mathcal{F})} + \big\|A(u,z)\big\|_{L^2(\mathcal{F})} \|w\|_{L^2(\mathcal{F})} \Big) \|\nabla v\|_{L^\infty(\Omega)} \\
&\lesssim h^{-1/2} \Bigg(\big\|A(z,w)\big\|_{L^1(\Omega)} + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \Big(\big\|\{\!\{A(u,z)\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h w]\!]\big\|_{L^2(\mathcal{F})} \\
&\quad + \big\|\{\!\{A(u,w)\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h z]\!]\big\|_{L^2(\mathcal{F})} \Big) + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \Big(\big\|A(u,w)\big\|_{L^2(\mathcal{F})} \|z\|_{L^2(\mathcal{F})} \\
&\quad + \big\|A(u,z)\big\|_{L^2(\mathcal{F})} \|w\|_{L^2(\mathcal{F})} \Big) \Bigg) \|v\|_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

Therefore by (2.8) and the Cauchy-Schwarz inequality

$$\begin{aligned}
\big\langle DQ[w](z), v\big\rangle &\lesssim h^{-1/2} \Bigg(\|D_h^2 z\|_{L^2(\Omega)} \|D_h^2 w\|_{L^2(\Omega)} \\
&\quad + \sum_{\mathcal{F}\in\mathcal{F}_h^i} \Big(\big\|\{\!\{D_h^2 z\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h w]\!]\big\|_{L^2(\mathcal{F})} + \big\|\{\!\{D_h^2 w\}\!\}\big\|_{L^2(\mathcal{F})} \big\|[\![\nabla_h z]\!]\big\|_{L^2(\mathcal{F})} \Big) \\
&\quad + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \Big(\|D_h^2 w\|_{L^2(\mathcal{F})} \|z\|_{L^2(\mathcal{F})} + \|D_h^2 z\|_{L^2(\mathcal{F})} \|w\|_{L^2(\mathcal{F})} \Big) \Bigg) \|v\|_{H^1(\mathcal{T}_h)} \\
&\lesssim h^{-1/2} \|z\|_{H^2(\mathcal{T}_h)} \|w\|_{H^2(\mathcal{T}_h)} \|v\|_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

The estimate (4.15) then follows from the definition (2.4). □

Lemma 4.6. *For any $w, z, q \in W^{3,3}(\mathcal{T}_h)$, we have*

$$\sum_{\mathcal{F}\in\mathcal{F}_h} \int_{\mathcal{F}} \big\|[\![\{\!\{A(w,z)\}\!\}\nabla_h q]\!]\big\|_{L^1(\mathcal{F})} \lesssim \|w\|_{W^{2,3}(\mathcal{T}_h)} \|z\|_{W^{2,3}(\mathcal{T}_h)} \|q\|_{W^{2,3}(\mathcal{T}_h)}, \tag{4.16}$$

$$\sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \int_{\mathcal{F}} \big\|A(w,z)q\big\|_{L^1(\mathcal{F})} \lesssim \|w\|_{W^{2,3}(\mathcal{T}_h)} \|z\|_{W^{2,3}(\mathcal{T}_h)} \|q\|_{W^{2,3}(\mathcal{T}_h)}. \tag{4.17}$$

Proof. To prove (4.16), we first use Hölder's inequality and (2.8) to obtain

$$\begin{aligned}
\sum_{\mathcal{F}\in\mathcal{F}_h} \int_{\mathcal{F}} \big\|[\![\{\!\{A(w,z)\}\!\}\nabla_h q]\!]\big\|_{L^1(\mathcal{F})} &\leq \sum_{\mathcal{F}\in\mathcal{F}_h} \big\|\{\!\{A(w,z)\}\!\}\big\|_{L^{\frac{3}{2}}(\mathcal{F})} \big\|[\![\nabla_h q]\!]\big\|_{L^3(\mathcal{F})} \\
&\lesssim \sum_{\mathcal{F}\in\mathcal{F}_h^i} \Big(\big\|D_h^2 w^+\big\|_{L^3(\mathcal{F})} \|D_h^2 z^+\|_{L^3(\mathcal{F})} + \|D_h^2 w^-\|_{L^3(\mathcal{F})} \|D_h^2 z^-\|_{L^3(\mathcal{F})} \Big) \big\|[\![\nabla_h q]\!]\big\|_{L^3(\mathcal{F})} \\
&\quad + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \big\|D_h^2 w\big\|_{L^3(\mathcal{F})} \|D_h^2 z\|_{L^3(\mathcal{F})} \big\|[\![\nabla_h q]\!]\big\|_{L^3(\mathcal{F})}.
\end{aligned}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Therefore by Hölder's nequality and (2.1), we have

$$\begin{aligned}
&\sum_{\mathcal{F}\in\mathcal{F}_h}\int_{\mathcal{F}}\left\|\left[\!\left[\left\{\!\!\left\{A(w,z)\right\}\!\!\right\}\nabla_h q\right]\!\right]\right\|_{L^1(\mathcal{F})}\\
&\lesssim\left(\sum_{T\in\mathcal{T}_h}h_T\|D_h w\|^3_{L^3(\partial T)}\right)^{\frac{1}{3}}\left(\sum_{T\in\mathcal{T}_h}h_T\|D_h z\|^3_{L^3(\partial T)}\right)^{\frac{1}{3}}\left(\sum_{\mathcal{F}\in\mathcal{F}_h}\frac{1}{h_{\mathcal{F}}^2}\left\|\left[\!\left[\nabla_h q\right]\!\right]\right\|^3_{L^3(\mathcal{F})}\right)^{\frac{1}{3}}\\
&\le\|w\|_{W^{2,3}(\mathcal{T}_h)}\|z\|_{W^{2,3}(\mathcal{T}_h)}\|q\|_{W^{2,3}(\mathcal{T}_h)}.
\end{aligned}$$

We prove (4.17) using similar techniques. First we have

$$\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\int_{\mathcal{F}}\left\|A(w,z)q\right\|_{L^1(\mathcal{F})}\le\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\|D_h^2 w\|_{L^3(\mathcal{F})}\|D_h^2 z\|_{L^3(\mathcal{F})}\|q\|_{L^3(\mathcal{F})},$$

and therefore

$$\begin{aligned}
&\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\int_{\mathcal{F}}\left\|A(w,z)q\right\|_{L^1(\mathcal{F})}\\
&\lesssim\left(\sum_{T\in\mathcal{T}_h}h_T\|D_h w\|^3_{L^3(\partial T)}\right)^{\frac{1}{3}}\left(\sum_{T\in\mathcal{T}_h}h_T\|D_h z\|^3_{L^3(\partial T)}\right)^{\frac{1}{3}}\left(\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}^5}\|q\|^3_{L^3(\mathcal{F})}\right)^{\frac{1}{3}}\\
&\le\|w\|_{W^{2,3}(\mathcal{T}_h)}\|z\|_{W^{2,3}(\mathcal{T}_h)}\|q\|_{W^{2,3}(\mathcal{T}_h)}.
\end{aligned}$$

□

Lemma 4.7 (estimate of D^2R)**.** *For any $w,z,q\in W^{3,3}(\mathcal{T}_h)$, we have*

$$\left\|D^2R[w](z,q)\right\|_{H^{-1}(\mathcal{T}_h)}\lesssim h^{-1/2}\|w\|_{W^{2,3}(\mathcal{T}_h)}\|z\|_{W^{2,3}(\mathcal{T}_h)}\|q\|_{W^{2,3}(\mathcal{T}_h)}.\tag{4.18}$$

Proof. By (4.11), (2.10), and (2.8), we have for any $v\in V_h$

$$\begin{aligned}
\left\langle D^2R[w](z,q),v\right\rangle\le&\left(\sum_{T\in\mathcal{T}_h}\|D_h^2 w\|_{L^3(T)}\|D_h^2 q\|_{L^3(T)}\|D_h^2 z\|_{L^3(T)}+\sum_{\mathcal{F}\in\mathcal{F}_h^i}\Big(\left\|\left[\!\left[\left\{\!\!\left\{A(w,q)\right\}\!\!\right\}\nabla_h z\right]\!\right]\right\|_{L^1(\mathcal{F})}\right.\\
&\left.+\left\|\left[\!\left[\left\{\!\!\left\{A(w,z)\right\}\!\!\right\}\nabla_h q\right]\!\right]\right\|_{L^1(\mathcal{F})}+\left\|\left[\!\left[\left\{\!\!\left\{A(q,z)\right\}\!\!\right\}\nabla_h w\right]\!\right]\right\|_{L^1(\mathcal{F})}\Big)\right)\|v\|_{L^\infty(\Omega)}\\
&+\sum_{\mathcal{F}\in\mathcal{F}_h^b}\Big(\left\|A(w,q)z\right\|_{L^1(\mathcal{F})}+\left\|A(w,z)q\right\|_{L^1(\mathcal{F})}+\left\|A(q,z)w\right\|_{L^1(\mathcal{F})}\Big)\|\nabla v\|_{L^\infty(\Omega)}\\
\lesssim h^{-1/2}&\left(\|D_h^2 w\|_{L^3(\Omega)}\|D_h^2 q\|_{L^3(\Omega)}\|D_h^2 z\|_{L^3(\Omega)}+\sum_{\mathcal{F}\in\mathcal{F}_h^i}\Big(\left\|\left[\!\left[\left\{\!\!\left\{A(w,q)\right\}\!\!\right\}\nabla_h z\right]\!\right]\right\|_{L^1(\mathcal{F})}\right.\\
&\left.+\left\|\left[\!\left[\left\{\!\!\left\{A(w,z)\right\}\!\!\right\}\nabla_h q\right]\!\right]\right\|_{L^1(\mathcal{F})}+\left\|\left[\!\left[\left\{\!\!\left\{A(q,z)\right\}\!\!\right\}\nabla_h w\right]\!\right]\right\|_{L^1(\mathcal{F})}\Big)\right.\\
&\left.+\sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\Big(\left\|A(w,q)z\right\|_{L^1(\mathcal{F})}+\left\|A(w,z)q\right\|_{L^1(\mathcal{F})}+\left\|A(q,z)w\right\|_{L^1(\mathcal{F})}\Big)\right)\|v\|_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

Therefore by (2.1) and (4.16)–(4.17)

$$\left\langle D^2R[w](z,q),v\right\rangle\lesssim h^{-1/2}\|w\|_{W^{2,3}(\mathcal{T}_h)}\|z\|_{W^{2,3}(\mathcal{T}_h)}\|q\|_{W^{2,3}(\mathcal{T}_h)}\|v\|_{H^1(\mathcal{T}_h)},$$

from which (4.18) immediately follows. □

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

4.2. Main results

With Lemmas 4.5 and 4.7 established, we proceed with the analysis of the nonlinear problem (3.10). First, combining (4.13)–(4.14) with the estimates (4.15) and (4.18), we immediately obtain the next two results.

Lemma 4.8 (contraction estimate of Q)**.** *For any $w_1, w_2 \in W^{3,3}(\mathcal{T}_h)$, we have*

$$\big\|Q(w_1-u)-Q(w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)} \lesssim h^{-1/2}\Big(\|u-w_1\|_{H^2(\mathcal{T}_h)}+\|u-w_2\|_{H^2(\mathcal{T}_h)}\Big)\|w_1-w_2\|_{H^2(\mathcal{T}_h)}. \tag{4.19}$$

Proof. By (4.13) and (4.15), we have

$$\begin{aligned}\big\|Q(w_1-u)-Q(w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)} &= \Big\|DQ\Big[\frac{1}{2}(w_1+w_2)-u\Big](w_1-w_2)\Big\|_{H^{-1}(\mathcal{T}_h)}\\ &\lesssim h^{-1/2}\big(\|u-w_1\|_{H^2(\mathcal{T}_h)}+\|u-w_2\|_{H^2(\mathcal{T}_h)}\big)\|w_1-w_2\|_{H^2(\mathcal{T}_h)}.\end{aligned}$$

□

Lemma 4.9 (contraction estimate of R)**.** *For any $w_1, w_2 \in W^{3,3}(\mathcal{T}_h)$, we have*

$$\big\|R(w_1-u)-R(w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)} \lesssim h^{-1/2}\Big(\|u-w_1\|^2_{W^{2,3}(\mathcal{T}_h)}+\|u-w_2\|^2_{W^{2,3}(\mathcal{T}_h)}\Big)\|w_1-w_2\|_{W^{2,3}(\mathcal{T}_h)}. \tag{4.20}$$

Proof. By (4.14) and (4.18),

$$\begin{aligned}\big\|R(w_1-u)-R(w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)} \le \frac{1}{6}\Big(&\big\|D^2R[w_1-w_2](w_1-u,w_1-u)\big\|_{H^{-1}(\mathcal{T}_h)}\\ &+\big\|D^2R[w_1-w_2](w_1-u,w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)}\\ &+\big\|D^2R[w_1-w_2](w_2-u,w_2-u)\big\|_{H^{-1}(\mathcal{T}_h)}\Big)\\ \lesssim h^{-1/2}&\big(\|u-w_1\|^2_{W^{2,3}(\mathcal{T}_h)}+\|u-w_2\|^2_{W^{2,3}(\mathcal{T}_h)}\big)\|w_1-w_2\|_{W^{2,3}(\mathcal{T}_h)}.\end{aligned}$$

□

Next, using the contraction estimates of Q and R, the inverse inequality, and the identity (4.9), we can derive contraction estimates of M_h in the following lemma.

Lemma 4.10 (contraction property of M_h)**.** *Define the discrete (closed) ball with center $u_{c,h}$ as*

$$\mathbb{B}_\rho(u_{c,h}) = \{v \in V_h;\ \|u_{c,h}-v\|_{H^1(\mathcal{T}_h)} \le \rho\}. \tag{4.21}$$

Then there exists a constant C_1, independent of h, σ, and ρ such that for any $v_1, v_2 \in \mathbb{B}_\rho(u_{c,h})$ there holds

$$\begin{aligned}\|M_hv_1-M_hv_2\|_{H^1(\mathcal{T}_h)} \le C_1\Big(&\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2+\big(h^{-5/2}\rho\big)^2\\ &+(1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}+h^{-5/2}\rho\Big)\|v_1-v_2\|_{H^1(\mathcal{T}_h)},\end{aligned} \tag{4.22}$$

where $\ell = \min\{s, k+1\}$.

Proof. It follows from (4.9), (4.2), and (4.19)–(4.20) that

$$\begin{aligned}\|M_hv_1-M_hv_2\|_{H^1(\mathcal{T}_h)} &= \Big\|L_h^{-1}\Big(R(v_1-u)-R(v_2-u)+Q(v_1-u)-Q(v_2-u)\Big)\Big\|_{H^1(\mathcal{T}_h)}\\ &\lesssim \|R(v_1-u)-R(v_2-u)\|_{H^{-1}(\mathcal{T}_h)}+\|Q(v_1-u)-Q(v_2-u)\|_{H^{-1}(\mathcal{T}_h)}\\ &\lesssim h^{-1/2}\Bigg(\Big(\|u-v_1\|_{H^2(\mathcal{T}_h)}+\|u-v_2\|_{H^2(\mathcal{T}_h)}\Big)\|v_1-v_2\|_{H^2(\mathcal{T}_h)}\\ &\quad+\Big(\|u-v_1\|^2_{W^{2,3}(\mathcal{T}_h)}+\|u-v_2\|^2_{W^{2,3}(\mathcal{T}_h)}\Big)\|v_1-v_2\|_{W^{2,3}(\mathcal{T}_h)}\Bigg).\end{aligned}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

Hence by the inverse inequality (2.10), (4.4)–(4.5) and (4.21), we have

$$\begin{aligned}\|M_hv_1 - M_hv_2\|_{H^1(\mathcal{T}_h)} &\lesssim h^{-1/2}\bigg(h^{-1}\Big(\|u-v_1\|_{H^2(\mathcal{T}_h)} + \|u-v_2\|_{H^2(\mathcal{T}_h)}\Big)\\ &\quad + h^{-3/2}\Big(\|u-v_1\|^2_{W^{2,3}(\mathcal{T}_h)} + \|u-v_2\|^2_{W^{2,3}(\mathcal{T}_h)}\Big)\bigg)\|v_1-v_2\|_{H^1(\mathcal{T}_h)}\\ &\lesssim h^{-1/2}\bigg(h^{-1}\Big(\|u-u_{c,h}\|_{H^2(\mathcal{T}_h)} + h^{-1}\rho\Big)\\ &\quad + h^{-3/2}\Big(\|u-u_{c,h}\|^2_{W^{2,3}(\mathcal{T}_h)} + h^{-3}\rho^2\Big)\bigg)\|v_1-v_2\|_{H^1(\mathcal{T}_h)}\\ &\lesssim \Big(\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2 + \big(h^{-5/2}\rho\big)^2\\ &\quad + (1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)} + h^{-5/2}\rho\Big)\|v_1-v_2\|_{H^1(\mathcal{T}_h)}.\end{aligned}$$

□

Lemma 4.11 (mapping property of M_h)**.** *For any $v \in \mathbb{B}_\rho(u_{c,h})$, there holds*

$$\|u_{c,h} - M_hv\|_{H^1(\mathcal{T}_h)} \le C_2\Big((1+\sigma)^3h^{3\ell-8}\|u\|^3_{H^\ell(\Omega)} + h^{-5}\rho^3 + (1+\sigma)^2h^{2\ell-9/2}\|u\|^2_{H^\ell(\Omega)} + h^{-5/2}\rho^2\Big), \tag{4.23}$$

where $C_2 > 0$ is independent of h, σ, and ρ.

Proof. By (4.8), (4.2), and (4.19)–(4.20), we have

$$\begin{aligned}\|u_{c,h} - M_hv\|_{H^1(\mathcal{T}_h)} &\le \big\|L_h^{-1}\big(R(v-u)\big)\big\|_{H^1(\mathcal{T}_h)} + \big\|L_h^{-1}\big(Q(v-u)\big)\big\|_{H^1(\mathcal{T}_h)}\\ &\lesssim \|R(v-u)\|_{H^{-1}(\mathcal{T}_h)} + \|Q(v-u)\|_{H^{-1}(\mathcal{T}_h)}\\ &\lesssim h^{-1/2}\Big(\|u-v\|^3_{W^{2,3}(\mathcal{T}_h)} + \|u-v\|^2_{H^2(\mathcal{T}_h)}\Big).\end{aligned}$$

Therefore, we obtain by (4.21), the inverse inequality (2.10), and (4.4)–(4.5)

$$\begin{aligned}\|u_{c,h} - M_hw\|_{H^1(\mathcal{T}_h)} &\lesssim h^{-1/2}\Big(\|u-u_{c,h}\|^3_{W^{2,3}(\mathcal{T}_h)} + h^{-\frac{9}{2}}\rho^3 + \|u-u_{c,h}\|^2_{H^2(\mathcal{T}_h)} + h^{-2}\rho^2\Big)\\ &\lesssim \Big((1+\sigma)^3h^{3\ell-8}\|u\|^3_{H^\ell(\Omega)} + h^{-5}\rho^3\\ &\quad + (1+\sigma)^2h^{2\ell-9/2}\|u\|^2_{H^\ell(\Omega)} + h^{-5/2}\rho^2\Big).\end{aligned}$$

□

Theorem 4.12 (main result I)**.** *There exists an $h_0(\sigma) > 0$ such that for $h \le h_0(\sigma)$, equation (3.10) has a solution u_h satisfying the estimate*

$$\|u-u_h\|_{H^1(\mathcal{T}_h)} + h\|u-u_h\|_{H^2(\mathcal{T}_h)} + h^{3/2}\|u-u_h\|_{W^{2,3}(\mathcal{T}_h)} \lesssim (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}. \tag{4.24}$$

Proof. Since $s > 7/2$ and $k \ge 3$, we may choose $h_0(\sigma) > 0$ such that $h \le h_0(\sigma)$ implies

$$\delta := 2\max\{C_1, C_2\}\Big(\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2 + (1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\Big) < 1. \tag{4.25}$$

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

$$\rho_0 = (1+\sigma)h^{\ell-1}\|u\|_{H^\ell(\Omega)}, \tag{4.26}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

and let $v_1, v_2 \in \mathbb{B}_{\rho_0}(u_{c,h})$. We then have by (4.22) and (4.25)–(4.26)

$$\begin{aligned}\|M_h v_1 - M_h v_2\|_{H^1(\mathcal{T}_h)} &\le C_1\Big(\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2 + \big(h^{-5/2}\rho_0\big)^2 \\ &\quad + (1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)} + h^{-5/2}\rho_0\Big)\|v_1 - v_2\|_{H^1(\mathcal{T}_h)} \\ &= 2C_1\Big(\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2 + (1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\Big)\|v_1 - v_2\|_{H^1(\mathcal{T}_h)} \\ &\le \delta\|v_1 - v_2\|_{H^1(\mathcal{T}_h)}.\end{aligned} \tag{4.27}$$

Moreover by (4.23) and (4.25)–(4.26), for any $v \in \mathbb{B}_{\rho_0}(u_{c,h})$, we have

$$\begin{aligned}\|u_{c,h} - M_h v\|_{H^1(\mathcal{T}_h)} &\le C_2\Big((1+\sigma)^3 h^{3\ell-8}\|u\|^3_{H^\ell(\Omega)} + h^{-5}\rho_0^3 + (1+\sigma)^2 h^{2\ell-9/2}\|u\|^2_{H^\ell(\Omega)} + h^{-5/2}\rho_0^2\Big) \\ &= 2C_2\Big(\big((1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\big)^2 + (1+\sigma)h^{\ell-7/2}\|u\|_{H^\ell(\Omega)}\Big)\rho_0 \le \rho_0.\end{aligned} \tag{4.28}$$

It then follows from (4.27) and (4.28) that M_h has a unique fixed point u_h in $\mathbb{B}_{\rho_0}(u_{c,h})$ which is a solution to (3.10). To obtain the estimates (4.24), we use the triangle inequality, (4.4)–(4.5), and (4.26) to conclude

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

and

$$\begin{aligned}h^{1/2}\|u - u_h\|_{W^{2,3}(\mathcal{T}_h)} + \|u - u_h\|_{H^2(\mathcal{T}_h)} &\lesssim h^{1/2}\|u - u_{c,h}\|_{W^{2,3}(\mathcal{T}_h)} + \|u - u_{c,h}\|_{H^2(\mathcal{T}_h)} + h^{-1}\rho_0 \\ &\lesssim (1+\sigma)h^{\ell-2}\|u\|_{H^\ell(\Omega)}.\end{aligned}$$

□

Theorem 4.13 (main result II)**.** *In addition to the hypotheses of Theorem 4.12, suppose $u \in W^{3,\infty}(\Omega)$. Then there holds*

$$\|u - u_h\|_{L^2(\Omega)} \lesssim (1+\sigma)^2\Big(h^\ell\|u\|_{H^\ell(\Omega)} + h^{2\ell-9/2}\|u\|^2_{H^\ell(\Omega)} + (1+\sigma)h^{3\ell-8}\|u\|^3_{H^\ell(\Omega)}\Big). \tag{4.29}$$

Proof. Let $\zeta \in H^1(\Omega)$ be the solution to the following problem

$$-\nabla\cdot\big(\mathrm{cof}(D^2 u)\nabla\zeta\big) = u - u_h \qquad \text{in } \Omega, \tag{4.30a}$$

$$\zeta = 0 \qquad \text{on } \partial\Omega. \tag{4.30b}$$

Since $u \in W^{3,\infty}(\Omega)$ we have $\mathrm{cof}(D^2 u) \in [W^{1,\infty}(\Omega)]^{3\times 3}$. Thus by elliptic regularity theory [13, 19], we have

$$\|\zeta\|_{H^2(\Omega)} \lesssim \|u - u_h\|_{L^2(\Omega)}. \tag{4.31}$$

Let $\zeta_h \in V_h$ be chosen such that (*cf.* [5], Thm. 3.2)

$$\|\zeta - \zeta_h\|_{H^1(\mathcal{T}_h)} \lesssim h\|\zeta\|_{H^2(\Omega)}, \qquad \|\zeta_h\|_{H^1(\mathcal{T}_h)} \lesssim \|\zeta\|_{H^2(\Omega)}. \tag{4.32}$$

We then have by (4.30)

$$\|u - u_h\|^2_{L^2(\Omega)} = \big\langle L(u - u_h), \zeta - \zeta_h\big\rangle + \big\langle L(u - u_h), \zeta_h\big\rangle. \tag{4.33}$$

For the first term, we use (4.3), (4.31)–(4.32), and (4.24) to obtain the bound

$$\begin{aligned}\big\langle L(u - u_h), \zeta - \zeta_h\big\rangle &\lesssim (1+\sigma)\|u - u_h\|_{H^1(\mathcal{T}_h)}\|\zeta - \zeta_h\|_{H^1(\mathcal{T}_h)} \\ &\lesssim (1+\sigma)^2 h^\ell\|u\|_{H^\ell(\Omega)}\|u - u_h\|_{L^2(\Omega)}.\end{aligned} \tag{4.34}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

To bound the second term in (4.33), we first note that by (4.1) and (4.8)

$$\big\langle L(u-u_h),\zeta_h\big\rangle = \big\langle L_h(u_{c,h}-u_h),\zeta_h\big\rangle = \big\langle L_h(u_{c,h}-M_hu_h),\zeta_h\big\rangle = \big\langle R(u_h-u)+Q(u_h-u),\zeta_h\big\rangle.$$

Therefore we have by (4.19)–(4.20), (4.31)–(4.32), and (4.24),

$$\begin{aligned}\big\langle L(u-u_h),\zeta_h\big\rangle &\le \big(\|R(u_h-u)\|_{H^{-1}(\mathcal{T}_h)}+\|Q(u_h-u)\|_{H^{-1}(\mathcal{T}_h)}\big)\|\zeta_h\|_{H^1(\mathcal{T}_h)}\\ &\lesssim h^{-1/2}\Big(\|u-u_h\|^3_{W^{2,3}(\mathcal{T}_h)}+\|u-u_h\|^2_{H^2(\mathcal{T}_h)}\Big)\|\zeta\|_{H^2(\Omega)}\\ &\lesssim \Big((1+\sigma)^3h^{3\ell-8}\|u\|^3_{H^\ell(\Omega)}+(1+\sigma)^2h^{2\ell-9/2}\|u\|^2_{H^\ell(\Omega)}\Big)\|u-u_h\|_{L^2(\Omega)}.\end{aligned} \tag{4.35}$$

Applying the estimates (4.34)–(4.35) to (4.33) and dividing by $\|u-u_h\|_{L^2(\Omega)}$, we obtain (4.29). □

Remark 4.14. Since $\ell > 7/2$, the error estimate (4.29) is of higher order than (4.24). Moreover, the estimate (4.29) is of the optimal order $k+1$ provided $s \ge k+1$ and $k \ge 4$.

5. Numerical experiments

In this section, we perform some numerical tests that back up the theoretical results proved in the previous section. Following similar ideas to those presented in [5], we apply the vanishing moment methodology [15,22] in order to obtain good initial guesses for the Newton solver in our computations. The crux of the vanishing moment method is to approximate fully nonlinear PDEs by higher order quasi-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:

$$-\varepsilon\Delta^2u^\varepsilon + \det(D^2u^\varepsilon) = f \qquad 0<\varepsilon\ll 1, \tag{5.1}$$

along with appropriate boundary conditions.

The finite element method for (5.1) is defined as seeking $u_h^\varepsilon \in V_h$ such that

$$\varepsilon A_hu_h^\varepsilon + F_hu_h^\varepsilon = 0, \tag{5.2}$$

where

$$\begin{aligned}\big\langle A_hv,w\big\rangle = \int_\Omega D_h^2v : D_h^2w\,\mathrm{d}x - \sum_{\mathcal{F}\in\mathcal{F}_h^i}\int_{\mathcal{F}}\Big(\big\{\!\big\{\partial^2_{nn}v\big\}\!\big\}\big[\!\big[\nabla w\big]\!\big] + \big[\!\big[\nabla v\big]\!\big]\big\{\!\big\{\partial^2_{nn}w\big\}\!\big\}\\ -\sigma\frac{1}{h_{\mathcal{F}}}\big[\!\big[\nabla v\big]\!\big]\big[\!\big[\nabla w\big]\!\big]\Big)\mathrm{d}s \qquad \forall v,w\in V_h,\end{aligned}$$

and

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

denotes the average of the second order normal derivative of w.

5.1. Example 1

In this test, we solve (3.10) for varying values of h and k, and choose our data and parameters such that the exact solution is given by

$$u = e^{(x^2+y^2+z^2)/2}, \qquad \Omega=(0,1)^3, \qquad \sigma = 150.$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

TABLE 1. Example 1. Numerical errors and rates of convergence for a smooth solution on the unit cube.

	h	$\|u-u_h\|_{L^2(\Omega)}$	Rate	$\|u-u_h\|_{H^1(\Omega)}$	Rate	$\|D_h^2(u-u_h)\|_{L^2(\Omega)}$	Rate
$k=2$	1/4	3.60E-04		1.32E-02		4.56E-01	
	1/8	4.05E-05	3.15	3.23E-03	2.02	2.24E-01	1.02
	1/12	1.06E-05	3.30	1.29E-03	2.26	1.43E-01	1.11
	1/16	4.88E-06	2.70	7.26E-04	2.01	1.07E-01	1.01
	1/20	2.67E-06	2.71	4.50E-04	2.14	8.44E-02	1.06
$k=3$	1/4	1.63E-05		3.85E-04		3.23E-02	
	1/8	9.86E-07	4.05	4.89E-05	2.98	7.58E-03	2.09
	1/12	1.53E-07	4.60	1.15E-05	3.57	3.00E-03	2.29
	1/16	4.71E-08	4.09	4.88E-06	2.97	1.67E-03	2.04

TABLE 2. Example 1. #DOFs, Max errors, and CPU time for a smooth solution on the unit cube.

	h	# Elements	#DOFs	$\|u-u_h\|_{L^\infty(\Omega)}$	CPU time (s)
$k=2$	1/4	943	1581	9.12E-03	3.89
	1/8	8434	12611	2.29E-03	38.55
	1/12	29 761	42 798	4.49E-04	140.89
	1/16	70 418	99 436	2.73E-04	355.78
	1/20	139 588	195 110	1.21E-04	803.47
$k=3$	1/4	943	4952	4.14E-04	28.78
	1/8	8434	40 985	5.71E-05	140.52
	1/12	29 761	140 861	7.52E-06	874.57
	1/16	70 418	329 244	6.36E-06	2758.98

In order to obtain some good initial guess, we solve (5.2) with ε-values $10^{-1}, 10^{-3}, 10^{-5}$, 10^{-7}, and 0, using each previous solution as our initial guess (we take $u = x_1^2 + x_2^2 + x_3^2$ as our initial guess for the first iteration with $\varepsilon = 10^{-1}$ in all of our numerical tests). After computing the solution of (3.10) we calculate the L^2, H^1, and piecewise H^2 error and record the errors in Table 1. As predicted by the theoretical results in Theorem 4.12, we observe third and second order convergence in the H^1 and H^2-norms, respectively, using cubic polynomials. Furthermore, the numerical tests also indicate that the method convergences optimally in the L^2 norm when using cubic polynomials and converges using quadratic polynomials.

In Table 2 we list the CPU time to compute the solution as well as the numerical error in the L^∞ norm. For comparison, we list the data taken from [16], where a wide-stencil finite difference scheme was used for the same test problem. Comparing the CPU times against the number of degrees of freedom in Tables 2, 3, we observe that the finite difference scheme is faster than the proposed finite element method. This is likely due to (a) the vanishing moment method creates additional overhead and (b) the assembly of the stiffness matrix is more expensive for the finite element method. However, comparing the CPU time against the L^∞ error, the finite element method is superior than that of the finite difference scheme. This behavior is likely due to the fact that the finite element method is of higher order, and therefore more efficient for smooth solutions.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

TABLE 3. #DOFs, Max errors, and CPU time of the wide-stencil difference scheme reported in [16].

N	#DOFs	$\|u - u_N\|_{L^\infty(\Omega)}$	CPU time (s)
7	343	1.51E-02	0.1
11	1331	1.40E-02	0.1
15	3375	1.32E-02	0.5
21	9261	1.27E-02	3.6
31	29 791	1.25E-02	34.7

TABLE 4. Example 2. Numerical errors and rates of convergence for a smooth solution on an ellipsoid.

	h	$\|u - u_h\|_{L^2(\Omega)}$	Rate	$\|u - u_h\|_{H^1(\Omega)}$	Rate	$\|D_h^2(u - u_h)\|_{L^2(\Omega)}$	Rate
$k = 2$	1/2	6.34E-02		3.60E-01		1.71E+00	
	1/4	5.17E-02	0.29	2.91E-01	0.31	1.43E+00	0.25
	1/8	2.79E-03	4.21	2.89E-02	3.33	5.44E-01	1.40
	1/12	8.46E-04	2.94	1.26E-02	2.05	3.38E-01	1.17
	1/16	3.18E-04	3.40	5.70E-03	2.75	2.12E-01	1.62
$k = 3$	1/2	1.72E-03		2.44E-02		4.76E-01	
	1/4	3.62E-04	2.25	7.30E-03	1.74	2.17E-01	1.13
	1/8	2.80E-05	3.69	1.14E-03	2.68	7.14E-02	1.60
	1/12	1.05E-05	2.43	5.69E-04	1.72	4.27E-02	1.27
	1/14	4.88E-06	4.94	3.01E-04	4.12	2.98E-02	2.33

5.2. Example 2

For our second test, we compute (3.10) on an ellipsoid. Namely, we choose our data, domain, and parameters as

$$u = \mathrm{e}^{x^6/6+(y^2+z^2)/2}, \qquad \Omega = \{(x, y, z);\ x^2 + 4y^2 + 4z^2 = 1\}, \qquad \sigma = 150.$$

We note that 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 finite elements.

After solving (3.10), we list the errors in Table 4. In accordance to Theorem 4.12, we observe optimal rates of convergence in the H^2 and H^1 norms when using cubic polynomials. Similar to the previous test, the numerical tests also indicate that the method converges optimally in the L^2 norm and is convergent when using quadratic polynomials.

5.3. Example 3

For the last test we solve (3.10) using quadratic polynomials with data

$$u = \frac{(x^2 + y^2 + z^2)^{\frac{3}{4}}}{3}, \qquad \Omega = (0, 1)^3, \qquad \sigma = 150.$$

Unlike the previous two tests, the solution to this problem is not smooth as it has a singularity at the origin. Nevertheless, we are still able to compute the solution and the numerical errors listed in Table 5 indicate the method converges with order $O(h^2)$, $O(h^2)$, and $O(h)$ in the L^2-norm, H^1-norm and H^2-norm respectively.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

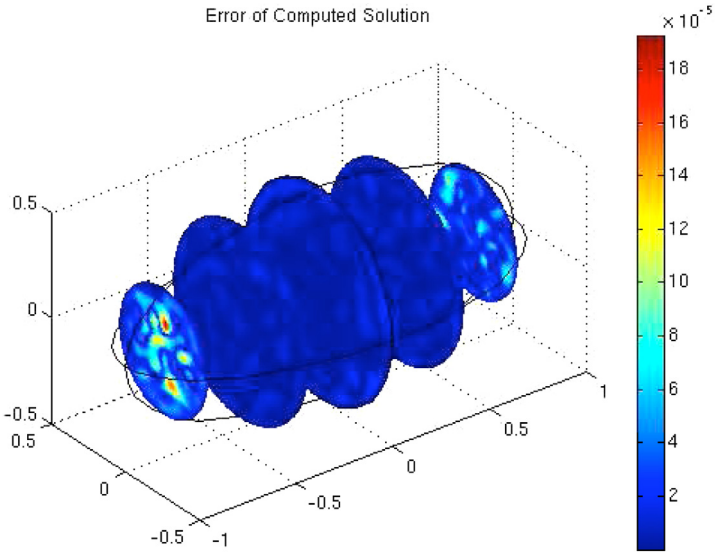


FIGURE 1. Example 2. Cross section plot of the error using cubic polynomials with $h = 1/8$.

TABLE 5. Example 3. Numerical errors and rates of convergence for a non-smooth solution on the unit cube.

	h	$\|u-u_h\|_{L^2(\Omega)}$	Rate	$\|u-u_h\|_{H^1(\Omega)}$	Rate	$\|D_h^2(u-u_h)\|_{L^2(\Omega)}$	Rate
$k=2$	1/2	1.73E-04		3.55E-03		8.82E-02	
	1/4	4.68E-05	1.89	1.59E-03	1.16	6.01E-02	0.55
	1/8	5.66E-06	3.05	3.80E-04	2.07	3.04E-02	0.98
	1/12	2.32E-06	2.19	1.84E-04	1.79	2.21E-02	0.78
	1/16	1.15E-06	2.46	1.04E-04	1.98	1.67E-02	0.96
	1/20	7.12E-07	2.14	6.63E-05	2.02	1.32E-02	1.06

6. CONCLUDING REMARKS

In this paper, we developed and studied finite element approximations of the three dimensional Monge-Ampère equation. Using a fixed point argument, we established existence of a solution to the scheme as well as derived quasi-optimal error estimates provided that the solution is sufficiently smooth. The numerical tests confirm the theory and also suggest that the method works for some non-classical solutions as well.

We end this paper by remarking that the methodology and analysis presented here can naturally be extended to other finite element methods. For example, the corresponding discontinuous Galerkin (DG) method to (3.10) is to find a function $u_h \in V_h^{DG}$ such that

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}\int_T\Big(f-\det(D^2u_h)\Big)v\,\mathrm{d}x+\sum_{e\in\mathcal{F}_h^i}\int_e\Big(\frac{\sigma}{h_{\mathcal{F}}}[\![u_h]\!]\cdot[\![v_h]\!]\\ +[\![\{\!\{\mathrm{cof}(D^2u_h)\}\!\}\nabla u_h]\!]\{\!\{v\}\!\}-\gamma\{\!\{\mathrm{cof}(D^2u_h)\nabla v_h\}\!\}\cdot[\![u_h]\!]\Big)\mathrm{d}s\\ +\sum_{e\in\mathcal{F}_h^b}\int_e\Big(\frac{\sigma}{h_{\mathcal{F}}}(u_h-g)v-\gamma[\![\mathrm{cof}(D^2u_h)\nabla v]\!](u_h-g)\Big)\mathrm{d}s=0\qquad\forall v\in V_h^{DG},\end{aligned}\tag{6.1}$$

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

where the jump of a scalar function is defined as

$$[\![v]\!]\big|_{\mathcal{F}} = v^+ \boldsymbol{n}_+ + v^- \boldsymbol{n}_- \qquad \mathcal{F} = \partial T^+ \cap \partial T^- \in \mathcal{F}_h^i,$$

V_h^{DG} is the space of totally discontinuous piecewise polynomails, and γ is a parameter that can take the values $\{1, -1, 0\}$, which correspond to the symmetric interior penalty method ($\gamma = 1$), non-symmetric interior penalty method ($\gamma = -1$), and incomplete interior penalty method ($\gamma = 0$). We refer the interested reader to the reference [23] for the derivation of the method (6.1). Using similar ideas to those presented here, we expect results similar to those stated in Theorems 4.12 and 4.13 also hold for the DG method (6.1).

Appendix A. Proof of Lemma 4.1

To prove that L_h is invertible as well as the stability estimate (4.2), it suffices to show that L_h is coercive over V_h with respect to the norm $\|\cdot\|_{H^1(\mathcal{T}_h)}$.

By (3.9) we have for $v \in V_h$,

$$\langle L_h v, v\rangle = \int_\Omega \big(\mathrm{cof}(D^2 u)\nabla v\big)\cdot\nabla v \,\mathrm{d}x - 2\sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla v]\!]\, v \,\mathrm{d}s + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b}\frac{1}{h_{\mathcal{F}}}\|v\|^2_{L^2(\mathcal{F})}. \tag{A.1}$$

Since u is strictly convex, the matrix $\mathrm{cof}(D^2 u)$ is positive definite. Thus, there exists a constant $\lambda > 0$ such that

$$\int_\Omega \big(\mathrm{cof}(D^2 u)\nabla v\big)\cdot\nabla v \,\mathrm{d}x \geq \lambda \|\nabla v\|^2_{L^2(\Omega)}. \tag{A.2}$$

Next since $u \in H^s(\Omega)$ with $s > 7/2$ we have $u \in W^{2,\infty}(\Omega)$ by a Sobolev embedding. Therefore by the trace, inverse and Cauchy-Schwarz inequalities, we have

$$\begin{aligned} 2\sum_{\mathcal{F}\in\mathcal{F}_h^b}\int_{\mathcal{F}} [\![\mathrm{cof}(D^2 u)\nabla_h v]\!]\, v \,\mathrm{d}s &\leq 2\|u\|^2_{W^{2,\infty}(\Omega)} \sum_{\mathcal{F}\in\mathcal{F}_h^b} \|\nabla v\|_{L^2(\mathcal{F})}\|v\|_{L^2(\mathcal{F})} \\ &\leq C\left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}\|\nabla v\|^2_{L^2(\mathcal{F})}\right)^{1/2}\left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1}\|v\|^2_{L^2(\mathcal{F})}\right)^{1/2} \\ &\leq C\|\nabla v\|_{L^2(\Omega)}\left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1}\|v\|^2_{L^2(\mathcal{F})}\right)^{1/2} \\ &\leq \frac{\lambda}{2}\|\nabla v\|^2_{L^2(\Omega)} + \frac{C}{\lambda}\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1}\|v\|^2_{L^2(\mathcal{F})}. \end{aligned} \tag{A.3}$$

It then follows from (A.1)–(A.3) that

$$\langle L_h v, v\rangle \geq \frac{\lambda}{2}\|\nabla v\|^2_{L^2(\Omega)} + \left(\sigma - \frac{C}{\lambda}\right)\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1}\|v\|^2_{L^2(\mathcal{F})}.$$

Thus by taking σ_0 sufficiently large, we have for $\sigma \geq \sigma_0$ and a scaling argument,

$$\|v\|^2_{H^1(\mathcal{T}_h)} \lesssim \langle L_h v, v\rangle. \tag{A.4}$$

The invertibility of L_h as well as the stability estimate (4.2) immediately follow from (A.4) and (2.4).

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

To show the bound (4.3), we have by (3.9), the Cauchy-Schwarz inequality and (2.3), for $w \in H^2(\mathcal{T}_h) \cap H^1(\Omega)$ and $v \in V_h$ that

$$\begin{aligned}
\langle Lw, v\rangle &\leq \|u\|^2_{W^{2,\infty}(\Omega)} \Bigg[\|\nabla w\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \|\nabla w\|_{L^2(\mathcal{F})} \|v\|_{L^2(\mathcal{F})} \\
&\quad + \sum_{\mathcal{F}\in\mathcal{F}_h^b} \|w\|_{L^2(\mathcal{F})} \|\nabla v\|_{L^2(\mathcal{F})} + \sigma \sum_{\mathcal{F}\in\mathcal{F}_h^b} \frac{1}{h_{\mathcal{F}}} \|w\|_{L^2(\mathcal{F})} \|v\|_{L^2(\mathcal{F})} \Bigg] \\
&\lesssim \Bigg[\|\nabla w\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}} \|\nabla w\|^2_{L^2(\mathcal{F})} \right)^{1/2} \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1} \|v\|^2_{L^2(\mathcal{F})} \right)^{1/2} \\
&\quad + \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1} \|w\|^2_{L^2(\mathcal{F})} \right)^{1/2} \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_F^{-1} \|\nabla v\|^2_{L^2(\mathcal{F})} \right)^{1/2} \\
&\quad + \sigma \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1} \|w\|^2_{L^2(\mathcal{F})} \right)^{1/2} \left(\sum_{\mathcal{F}\in\mathcal{F}_h^b} h_{\mathcal{F}}^{-1} \|v\|^2_{L^2(\mathcal{F})} \right)^{1/2} \Bigg] \\
&\lesssim (1+\sigma) \|w\|_{H^1(\mathcal{T}_h)} \|v\|_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

The bound (4.3) then follows from (2.4).

We now show the error estimates (4.4). First by (4.2), (4.3) and (4.1), we have for any $v \in V_h$,

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

It then follows from Lemma 2.8 and scaling that

$$\|u - u_{c,h}\|_{H^1(\mathcal{T}_h)} \lesssim (1+\sigma) h^{\ell-1} \|u\|_{H^\ell(\Omega)}. \tag{A.5}$$

Next by the triangle inequality and Lemma 2.7, we have for any $v \in V_h$

$$\begin{aligned}
\|u - u_{c,h}\|_{H^2(\mathcal{T}_h)} &\lesssim \|u - v\|_{H^2(\mathcal{T}_h)} + h^{-1} \|u_{c,h} - v\|_{H^1(\mathcal{T}_h)} \\
&\lesssim \|u - v\|_{H^2(\mathcal{T}_h)} + h^{-1} \|u - v\|_{H^1(\mathcal{T}_h)} + h^{-1} \|u - u_{c,h}\|_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

Thus by Lemma 2.8 and (A.5) we obtain

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

References

[1] G. Barles and P.E. Souganidis, Convergence of approximation schemes for fully nonlinear second order equtions. *Asymptotic Anal.* **4** (1991) 271–283.

[2] C. Bernardi, Optimal finite element interpolation on curved domains. *SIAM J. Numer. Anal.* **26** (1989) 1212–1240.

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

[4] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, 3th edition. Springer (2008).

[5] S.C. Brenner, T. Gudi, M. Neilan and L.-Y. Sung, $\mathcal{C}^0$ penalty methods for the fully nonlinear Monge-Ampère equation. *Math. Comput.* **80** (2011) 1979–1995.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press

[6] L.A. Caffarelli and C.E. Gutiérrez, Properties of the solutions of the linearized Monge-Ampère equation. *Amer. J. Math.* **119** (1997) 423–465.
[7] L.A. Caffarelli and M. Milman, Monge-Ampère Equation: Applications to Geometry and Optimization. *Amer. Math. Soc.* Providence, RI (1999).
[8] 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** (1984) 369–402.
[9] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam (1978).
[10] M.G. Crandall, H. Ishii and P.-L. Lions, User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc.* **27** (1992) 1–67.
[11] E.J. Dean and R. Glowinski, Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. *Comput. Methods Appl. Mech. Engrg.* **195** (2006) 1344–1386.
[12] 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** (2008) 9841–9864.
[13] L.C. Evans, Partial Differential Equations, Graduate Studies in Mathematics. Providence, RI. *Amer. Math. Soc.* **19** (1998).
[14] X. Feng and M. Neilan, Vanishing moment method and moment solutions for second order fully nonlinear partial differential equations. *J. Sci. Comput.* **38** (2009) 74–98.
[15] 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** (2009) 1226–1250.
[16] B.D. Froese and A.M. Oberman, Convergent finite difference solvers for viscosity solutions of the ellptic Monge-Ampère equation in dimensions two and higher. *SIAM J. Numer. Anal.* **49** (2011) 1692–1714.
[17] B.D. Froese and A.M. Oberman, Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. *J. Comput. Phys.* **230** (2011) 818–834.
[18] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, Berlin (2001).
[19] P. Grisvard, *Elliptic Problems on Nonsmooth Domains*. Pitman Publishing Inc. (1985).
[20] C.E. Gutiérrez, The Monge-Ampère Equation, *Progress in Nonlinear Differential Equations and Their Applications* **44**. Birkhauser, Boston, MA (2001).
[21] T. Muir, *A Treatise on the Theory of Determinants*. Dover Publications Inc., New York (1960).
[22] M. Neilan, A nonconforming Morley finite element method for the fully nonlinear Monge-Ampère equation. *Numer. Math.* **115** (2010) 371–394.
[23] M. Neilan, A unified analysis of some finite element methods for the Monge-Ampère equation. Submitted.
[24] 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** (1971) 9–15.
[25] 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** (2008) 221–238.
[26] D.C. Sorensen and R. Glowinski, A quadratically constrained minimization problem arising from PDE of Monge-Ampère type. *Numer. Algorithm* **53** (2010) 53–66.
[27] N.S. Trudinger and X.-J. Wang, The Monge-Ampère equation and its geometric applications, *Handbook of Geometric Analysis* I. International Press (2008) 467–524.
[28] C. Villani, Topics in Optimal Transportation, Graduate Studies in Mathematics. Providence, RI. *Amer. Math. Soc.* **58** (2003).
[29] A. Ženíšek, *Polynomial approximation on tetrahedrons in the finite element method. J. Approx. Theory* **7** (1973) 334–351.
[30] V. Zheligovsky, O. Podvigina and U. Frisch, The Monge-Ampère equation: various forms and numerical solutions. *J. Comput. Phys.* **229** (2010) 5043–5061.

https://doi.org/10.1051/m2an/2011067 Published online by Cambridge University Press