INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 11, Number 4, Pages 745–761

ERROR ANALYSIS OF A MIXED FINITE ELEMENT METHOD FOR THE MONGE-AMPÈRE EQUATION

GERARD AWANOU AND HENGGUANG LI

(Communicated by Ragnar Winther)

Abstract. We analyze the convergence of a mixed finite element method for the elliptic Monge-Ampère equation in dimensions 2 and 3. The unknowns in the formulation, the scalar variable and a discrete Hessian, are approximated by Lagrange finite element spaces. The method originally proposed by Lakkis and Pryer can be viewed as the formal limit of a Hermann-Miyoshi mixed method proposed by Feng and Neilan in the context of the vanishing moment methodology. Error estimates are derived under the assumption that the continuous problem has a smooth solution.

Key words. Monge-Ampère, mixed finite elements, Lagrange elements, fixed point.

1. Introduction

We are interested in the numerical approximation of convex solutions of the nonlinear elliptic Monge-Ampère equation

$$\begin{aligned} \det D^2 u &= f \,\text{in}\, \Omega \\ u &= g \,\text{on}\, \partial\Omega. \end{aligned} \tag{1.1}$$

Here Ω is a convex polygonal domain of $\mathbb{R}^d$ and $f \in C(\Omega), g \in C(\partial\Omega)$ with $f \geq c_0 > 0$ for a constant $c_0 > 0$. We give an analysis of a mixed finite element approximation of (1.1) for dimensions $d = 2$ and $d = 3$. The unknowns in the formulation are the scalar variable and a discrete Hessian and both are approximated by Lagrange finite element spaces of degree $k \geq 1$.

The numerical study of Monge-Ampère type equations is a recent active research area where it appears that techniques to prove convergence to the so-called viscosity solutions of (1.1) are inherently different from the ones needed to derive error estimates for smooth solutions. It has been documented in [7, 8] for the two-dimensional problem that the method of Lakkis and Pryer with Lagrange elements of degree $k \geq 2$ captures viscosity solutions of the Monge-Ampère equation. Some numerical methods proposed for the Monge-Ampère equation, e.g. [3], do not perform well for non smooth solutions when the discrete problem is solved by Newton's method. On the other hand, with the mixed method one can use Newton's method and still have numerical convergence for non smooth solutions. This offers the possibility of numerical solvers faster than the iterative methods proposed in [1]. In this paper we assume that (1.1) has a smooth solution.

To guarantee the existence of a smooth solution, one has to assume that the domain is smooth and strictly convex and the data f and g are also smooth [9]. The convex polygonal domain may be assumed to be an approximation of a smooth and strictly convex domain. Another approach would be to consider elements with curved faces and enforce Dirichlet boundary conditions by a penalty method as in [3].

Received by the editors August 26, 2013 and, in revised form, January 25, 2014.
2000 *Mathematics Subject Classification.* 65N30, 35J25.

The method of Lakkis and Pryer has been recently generalized in [8] where a discontinuous finite element space is used to approximate the discrete Hessian. This results in a more efficient numerical method and an analysis of both types of methods were given in [8] for the two dimensional problem. The connection of the method of Lakkis and Pryer with a Herman-Miyoshi mixed finite element was also noted in [8]. But the idea to analyze the method from the point of view of mixed methods, or to view it as the formal limit of the mixed method proposed in the context of the vanishing moment methodology in [6], was not considered. One possible reason is that Herman-Miyoshi type mixed methods were originally studied for equations involving the biharmonic operator. Several technical arguments have to be made as the linearized Monge-Ampère equation is a second order elliptic equation. The contributions of this paper are:

(1) An analysis valid in both dimensions 2 and 3 and different from the one given in [8] for the two dimensional problem.
(2) Error estimates for Lagrange elements of degree $k \geq 3$ in dimensions 2 and 3.
(3) Numerical experiments for smooth solutions and Lagrange elements of degree $k = 1$. Previous authors in their implementation eliminated the discrete Hessian, which does not necessarily converge for $k = 1$, and concluded the divergence of the method for linear elements.

The approach taken in this paper could help in the investigation of the method for low order elements, i.e. for $k = 1, 2$.

The paper is organized as follows. In the second section we introduce some notations, recall classical finite element results, present the mixed method for the Monge-Ampère equation and useful facts about computations with determinants. Our variational formulation is well posed for dimensions $d = 2$ and $d = 3$ but other general statements are valid for arbitrary dimension d. In section 3 we give the error analysis. The last section is devoted to the numerical results.

2. Preliminaries

2.1. Notation and assumptions. Let Ω be an open convex bounded subset of $\mathbb{R}^d$ with boundary $\partial\Omega$ and let $\mathcal{T}_h$ denote a triangulation of Ω into simplices K. We denote by h_K the diameter of the element K and $h = \max_{K \in \mathcal{T}_h} h_K$. We make the assumption that the triangulation is conforming and satisfies the usual shape regularity condition, i.e. there exists a constant $\sigma > 0$ such that $h_K/\rho_K \leq \sigma$, for all $K \in \mathcal{T}_h$ where ρ_K denotes the radius of the largest ball inside K. To be able to use global inverse estimates, c.f. (2.2) and (2.3) below, we require the triangulation to be also quasi-uniform, i.e. there is a constant $C > 0$ such that $h \leq C h_K$ for all $K \in \mathcal{T}_h$.

We use the usual notation $L^p(\Omega), 2 \leq p \leq \infty$ for the Lebesgue spaces and $H^s(\Omega), 1 \leq s < \infty$ for the Sobolev spaces of elements of $L^2(\Omega)$ with weak derivatives of order less than or equal to s in $L^2(\Omega)$. We recall that $W^{s,\infty}(\Omega)$ is the Sobolev space of functions with weak derivatives of order less than or equal to s in $L^\infty(\Omega)$. For a given normed space X, we denote by X^d the space of vector fields with components in X and by $X^{d \times d}$ the space of matrix fields with each component in X. The norm in X is denoted by $|| \, ||_X$ and we omit the subscripts Ω, d, and $d \times d$ when it is clear from the context. We will use the standard notation $| \, |_{H^s}$ for the semi norm on $H^s(\Omega), H^s(\Omega)^d$ and $H^s(\Omega)^{d \times d}$. The inner product in $L^2(\Omega), L^2(\Omega)^d$, and $L^2(\Omega)^{d \times d}$ is denoted by $(,)$ and we use $\langle , \rangle$ for the inner product on $L^2(\partial\Omega)$

and $L^2(\partial\Omega)^d$. For inner products on subsets of Ω, we will simply append the subset notation. We denote by n the unit outward normal vector to $\partial\Omega$.

For a scalar function v we denote by Dv the gradient vector and by D^2v the Hessian matrix of second order derivatives. For two matrices $A = (A_{ij})$ and $B = (B_{ij})$, $A : B = \sum_{i,j=1}^d A_{ij}B_{ij}$ denotes their Frobenius inner product. The divergence of a matrix field is understood as the vector obtained by taking the divergence of each row. A quantity which is constant is simply denoted by C.

2.2. Lagrange finite element spaces. Let V_h denote the standard Lagrange finite element space of degree $k \geq 1$ and $\Sigma_h = V_h^{d\times d}$. Thus elements of Σ_h are not necessarily symmetric matrix fields. We recall that $H_0^1(\Omega)$ is the subset of $H^1(\Omega)$ of elements with vanishing trace on $\partial\Omega$. Let I_h denote the standard Lagrangian interpolation operator from $H^s(\Omega), s \geq k+1$ into the space V_h. We have the following approximation property

$$
\begin{aligned}
&||v - I_h v||_{H^j} \leq Ch^{k+1-j}||v||_{H^{k+1}}, \forall v \in H^s(\Omega), j = 0, 1, \\
&||v - I_h v||_{L^\infty} \leq Ch^{k+1-\frac{d}{2}}|v|_{H^{k+1}}, \forall v \in H^s(\Omega).
\end{aligned} \tag{2.1}
$$

We use the notation I_h for the matrix version of the interpolation operator into $V_h^{d\times d}$. For a continuous function g defined on $\partial\Omega$, we let g_h denote its piecewise Lagrange interpolant on $\partial\Omega$. Finally we denote by I the $d \times d$ identity matrix.

We will need the inverse estimates, c.f. Theorem 4.5.11 of [4],

$$
||v||_{L^\infty} \leq Ch^{-\frac{d}{2}}||v||_{L^2}, \forall v \in V_h \tag{2.2}
$$

$$
||v||_{H^1} \leq Ch^{-1}||v||_{L^2}, \forall v \in V_h, \tag{2.3}
$$

and the trace inequality

$$
||v||_{L^2(\partial\Omega)} \leq C||v||_{H^1(\Omega)}, \forall v \in H^1(\Omega), \tag{2.4}
$$

which gives the scaled trace inequality by standard scaling arguments

$$
||v||^2_{L^2(\partial\Omega)} \leq C(h^{-1}||v||^2_{L^2} + h||\nabla v||^2_{L^2}), \ \forall v \in V_h. \tag{2.5}
$$

We note that (2.5) holds for all $v \in H^1(\Omega)$. The scaled trace inequality and the inverse estimate imply

$$
||v||_{L^2(\partial\Omega)} \leq Ch^{-\frac{1}{2}}||v||_{L^2}, \ \forall v \in V_h. \tag{2.6}
$$

The discrete Sobolev inequalities give estimates sharper than the inverse inequality (2.2)

$$
||v||_{L^\infty} \leq C(1 + |\ln h|^{\frac{1}{2}})||v||_{H^1}, \forall v \in V_h \text{ and } d = 2 \tag{2.7}
$$

$$
||v||_{L^\infty} \leq Ch^{-\frac{1}{2}}||v||_{H^1}, \forall v \in V_h \text{ and } d = 3. \tag{2.8}
$$

The first one can be found in [2] and the second follows from an inverse estimate and the embedding of $H^1(\Omega)$ in $L^6(\Omega)$.

2.3. Variational formulations. We make the assumption that (1.1) has a unique strictly convex solution $u \in H^s(\Omega), s > 3$ for $d = 2$ and $s > 4$ for $d = 3$. Additional assumptions about the regularity of u will be made for the error analysis. By Sobolev embedding $u \in C^2(\overline{\Omega})$. Moreover the unique convex solution of (1.1) satisfies the following mixed problem: Find $(u,\sigma) \in H^2(\Omega) \times H^1(\Omega)^{d\times d}$ such that

$$
\begin{aligned}
(\sigma,\tau) + (\operatorname{div}\tau, Du) - \langle Du, \tau n\rangle &= 0, \forall \tau \in H^1(\Omega)^{d\times d}\\
(\det \sigma, v) &= (f, v), \forall v \in H_0^1(\Omega) \\
u &= g \text{ on } \partial\Omega.
\end{aligned}
\tag{2.9}
$$

To see that the quantity $(\det\sigma, v)$ is finite for $v \in L^2(\Omega)$ and $\sigma \in H^1(\Omega)^{d\times d}$, note that for $d = 2$, $\det\sigma$ is a quadratic function of the entries of σ. For $\sigma_1, \sigma_2 \in H^1(\Omega)$, by Hölder's inequality and the embedding of $H^1(\Omega)$ into $L^q(\Omega), 1 \leq q < \infty$ for $d = 2$

$$\int_\Omega \sigma_1\sigma_2 v dx \leq ||\sigma_1||_{L^4}||\sigma_2||_{L^4}||v||_{L^2} \leq ||\sigma_1||_{H^1}||\sigma_2||_{H^1}||v||_{L^2}.$$

For $d = 3$, $\sigma_1, \sigma_2, \sigma_3 \in H^1(\Omega)$, by Hölder's inequality and the embedding of $H^1(\Omega)$ into $L^q(\Omega), 1 \leq q \leq 6$ for $d = 3$

$$\int_\Omega \sigma_1\sigma_2\sigma_3 v dx \leq ||\sigma_1||_{L^6}||\sigma_2||_{L^6}||\sigma_3||_{L^6}||v||_{L^2} \leq ||\sigma_1||_{H^1}||\sigma_2||_{H^1}||\sigma_3||_{H^1}||v||_{L^2}.$$

A mixed formulation of (2.9) consists in finding $(u_h, \sigma_h) \in V_h \times \Sigma_h$ such that

$$
\begin{aligned}
(\sigma_h,\tau) + (\operatorname{div}\tau, Du_h) - \langle Du_h, \tau n\rangle &= 0, \forall \tau \in \Sigma_h\\
(\det \sigma_h, v) &= (f, v),\ \forall v \in V_h \cap H_0^1(\Omega) \\
u_h &= g_h \text{ on } \partial\Omega.
\end{aligned}
\tag{2.10}
$$

The condition $\tau \in H^1(\Omega)^{d\times d}$ in the formulation (2.9) may be replaced by $\tau \in L^2(\Omega)^{d\times d}$ with $\operatorname{div}\tau \in L^2(\Omega)^d$. Also, we need $v \in H_0^1(\Omega)$ only to be able to take traces on $\partial\Omega$.

Remark 2.1. *The mixed method* (2.10) *is a nonconforming mixed method as we require* $u \in H^2(\Omega)$ *for the term* $\langle Du, \tau n\rangle$ *to be well defined.*

2.4. Computation with determinants. For a matrix A, we denote by A_{ij} its entries and by $\operatorname{cof} A$ its cofactor matrix, i.e. $(\operatorname{cof} A)_{ij} = (-1)^{i+j}\det(A)_i^j$ where $\det(A)_i^j$ is the determinant of the matrix obtained from A by deleting the ith row and the jth column.

Lemma 2.2. *For a* $d\times d$ *matrix* A, $\det A = d^{-1}(\operatorname{cof} A) : A$ *and for* $u \in C^3(\Omega)$, $\det D^2u = d^{-1}\operatorname{div}((\operatorname{cof} D^2u)Du)$.

Proof. The first statement follows from the row expansion definition of the determinant, expanding $\det A$ in d different ways using each row.

For a vector field $v = (v_i)$, let Dv be the matrix such that $(Dv)_{ij} = (\partial v_i)/(\partial x_j)$. We claim that $\operatorname{div}(Av) = (\operatorname{div} A^T)\cdot v + A : (Dv)^T$. Indeed

$$
\begin{aligned}
\operatorname{div}(Av) &= \sum_{i=1}^d \frac{\partial}{\partial x_i}(Av)_i = \sum_{i=1}^d \frac{\partial}{\partial x_i}\left(\sum_{j=1}^d A_{ij}v_j\right)\\
&= \sum_{i=1}^d\sum_{j=1}^d \left(\frac{\partial A_{ij}}{\partial x_i}v_j + A_{ij}\frac{\partial v_j}{\partial x_i}\right) = \sum_{j=1}^d\left(\sum_{i=1}^d \frac{\partial A_{ij}}{\partial x_i}\right)v_j + \sum_{i=1}^d\sum_{j=1}^d A_{ij}\frac{\partial v_j}{\partial x_i}\\
&= (\operatorname{div} A^T)\cdot v + A : (Dv)^T.
\end{aligned}
$$

We take $v = Du$ and note that $\operatorname{div}\operatorname{cof} Dv = \operatorname{div}\operatorname{cof} D^2u = 0$ by the divergence-free row property of the cofactor matrix, p. 440 of [5]. Since $\operatorname{cof} D^2u$ and D^2u are symmetric matrices, the result then follows. □

Lemma 2.3. *Fréchet derivative of the determinant. For* $F(u) = \det D^2 u$*, we have* $F'(u)(v) = (\operatorname{cof} D^2 u) : D^2 v$.

Proof. We have $\partial(\det A)/(\partial A_{ij}) = (\operatorname{cof} A)_{ij}$. See for example formula (23) p. 440 of [5]. The result then follows from the chain rule. □

Lemma 2.4. *Mean value theorem for the determinant. For* $K \in \mathcal{T}_h$ *and* $u, v \in C^2(K)$ *we have on* K

$$\det D^2 u - \det D^2 v = \operatorname{cof}(tD^2 u + (1-t)D^2 v) : (D^2 u - D^2 v),$$

for some $t \in [0, 1]$.

Proof. The result follows immediately from Lemma 2.3 and the mean value theorem. □

Lemma 2.5. *For* $d = 2$ *and* $d = 3$*, and two matrix fields* η *and* τ

$$|| \operatorname{cof}(\eta) : \tau||_{L^2} \leq C||\eta||_{L^\infty}^{d-1} ||\tau||_{L^2}.$$

Proof. The proof follows from direct computation. □

Lemma 2.6. *For* $d = 2$ *and* $d = 3$*, and two matrix fields* η *and* τ

$$|| \operatorname{cof}(\eta) - \operatorname{cof}(\tau)||_{L^2(K)} \leq C(||t\eta + (1-t)\tau||_{L^\infty(K)})^{d-2} ||\eta - \tau||_{L^2(K)},$$

for some $t \in [0, 1]$.

Proof. For $d = 2$, we have $\operatorname{cof}(\eta) - \operatorname{cof}(\tau) = \operatorname{cof}(\eta - \tau)$ from which the result follows. For $d = 3$ we use the mean value theorem. It is enough to estimate the first entry of $\operatorname{cof}(\eta) - \operatorname{cof}(\tau)$ which is equal to

$$\begin{aligned} \det \begin{pmatrix} \eta_{22} & \eta_{23} \\ \eta_{32} & \eta_{33} \end{pmatrix} - \det \begin{pmatrix} \tau_{22} & \tau_{23} \\ \tau_{32} & \tau_{33} \end{pmatrix} &= \operatorname{cof}\left(t \begin{pmatrix} \eta_{22} & \eta_{23} \\ \eta_{32} & \eta_{33} \end{pmatrix} + (1-t) \begin{pmatrix} \tau_{22} & \tau_{23} \\ \tau_{32} & \tau_{33} \end{pmatrix} \right) : \\ &\qquad \begin{pmatrix} \eta_{22} - \tau_{22} & \eta_{23} - \tau_{23} \\ \eta_{32} - \tau_{32} & \eta_{33} - \tau_{33} \end{pmatrix} \\ &= \operatorname{cof} \begin{pmatrix} t\eta_{22} + (1-t)\tau_{22} & t\eta_{23} + (1-t)\tau_{23} \\ t\eta_{32} + (1-t)\tau_{32} & t\eta_{33} + (1-t)\tau_{33} \end{pmatrix} : \\ &\qquad \begin{pmatrix} \eta_{22} - \tau_{22} & \eta_{23} - \tau_{23} \\ \eta_{32} - \tau_{32} & \eta_{33} - \tau_{33} \end{pmatrix}, \end{aligned}$$

for some $t \in [0, 1]$. The result then follows from Lemma 2.5. □

3. Error analysis of the Monge-Ampère equation

We use a fixed point argument which consists of linearizing the nonlinear equation at the exact solution and use the stability of the linearized problem. This technique has recently been used for the vanishing moment methodology approach to the Monge-Ampère equation in [6] . We first derive and study the linearized problem. Then we prove the existence and uniqueness of a solution to the nonlinear discrete equations.

3.1. The linearized Monge-Ampère equation. Put $F(u) = f - \det D^2 u$ and recall that the Fréchet derivative of F is given by $F'(u)(v) = -\operatorname{div}\big((\operatorname{cof} D^2 u)Dv\big)$. See Lemma 2.3. We are thus led to consider the linearized problem: find $w \in H^1(\Omega)$

$$
\begin{aligned}
-\operatorname{div}\big((\operatorname{cof} D^2 u)Dw\big) &= m \text{ in } \Omega \\
w &= l \text{ on } \partial\Omega,
\end{aligned}
\tag{3.1}
$$

for given $m \in L^2(\Omega)$ and $l \in C(\partial\Omega)$.

By the assumptions on the right hand side f of (1.1), its solution u is strictly convex and thus $A = \operatorname{cof}(D^2 u)$ is uniformly positive definite. Problem (3.1) is therefore well posed.

A mixed formulation of (3.1) consists in finding $(w, \eta) \in H^1(\Omega) \times L^2(\Omega)^{d\times d}$ such that

$$
\begin{aligned}
\eta &= D^2 w \text{ in } \Omega \\
-\operatorname{div}\operatorname{cof}(D^2 u)Dw &= m \text{ in } \Omega \\
w &= l \text{ on } \partial\Omega.
\end{aligned}
$$

A weak formulation of the above problem is given by: Find $(w, \eta) \in H^1(\Omega) \times L^2(\Omega)^{d\times d}$

$$
\begin{aligned}
(\eta, \tau) + (\operatorname{div}\tau, Dw) - \langle Dw, \tau n\rangle &= 0, \forall \tau \in H^1(\Omega)^{d\times d}, \\
((\operatorname{cof}(D^2 u)Dw, Dv) &= (m, v),\ \forall v \in H_0^1(\Omega), \\
w &= l \text{ on } \partial\Omega,
\end{aligned}
$$

The discrete problem consists in finding $(w_h, \eta_h) \in V_h \times \Sigma_h$

$$
\begin{aligned}
(\eta_h, \tau) + (\operatorname{div}\tau, Dw_h) - \langle Dw_h, \tau n\rangle &= 0, \forall \tau \in \Sigma_h, \\
((\operatorname{cof}(D^2 u)Dw_h, Dv) &= (m, v),\ \forall v \in V_h \cap H_0^1(\Omega), \\
w_h &= l_h \text{ on } \partial\Omega.
\end{aligned}
\tag{3.2}
$$

Theorem 3.1. *Problem* (3.2) *has a solution which is unique.*

Proof. To prove existence and uniqueness of the problem (3.2), we assume $m = 0, l_h = 0$ and show that $w_h = 0$ and $\eta_h = 0$. Taking $v = w_h$ and $\tau = \eta_h$ in (3.2), by the strict convexity of u, we obtain $|w_h|_{H^1} = 0$ which gives $w_h = 0$. It then follows that $\eta_h = 0$ as well. ☐

Remark 3.2. *We make the observation that the last two equations of* (3.2)*, which solve a linear diffusion equation, completely decouple from the first equation. Thus, for the linearized problem, we view η_h as a projection of w_h.*

3.2. Error analysis of the nonlinear problem. Without loss of generality, we will assume that $h \le 1$. Define a mapping $T : V_h \times \Sigma_h \to V_h \times \Sigma_h$ by

$$
T(w_h, \eta_h) = (T_1(w_h, \eta_h), T_2(w_h, \eta_h)),
$$

where $T_1(w_h, \eta_h)$ and $T_2(w_h, \eta_h)$ satisfy

$$
\begin{aligned}
(\eta_h - T_2(w_h, \eta_h), \tau) &+ (\operatorname{div}\tau, D(w_h - T_1(w_h, \eta_h))) \\
&- \langle D(w_h - T_1(w_h, \eta_h)), \tau n\rangle = (\eta_h, \tau) \\
&+ (\operatorname{div}\tau, Dw_h) - \langle Dw_h, \tau n\rangle, \quad \forall\, \tau \in \Sigma_h
\end{aligned}
\tag{3.3}
$$

$$
((\operatorname{cof} D^2 u)D(w_h - T_1(w_h, \eta_h)), Dv) = (f, v) - (\det \eta_h, v),\ \forall\ v \in V_h \cap H_0^1(\Omega)
\tag{3.4}
$$

$$
w_h - T_1(w_h, \eta_h) = 0 \quad \text{on} \quad \partial\Omega.
\tag{3.5}
$$

The motivation of the definition of the mapping T is given by Lemma 3.3 and 3.4 below.

Lemma 3.3. *T is well defined by the well-posedness of the linearized problem, i.e. Theorem 3.1 applied to* (3.2)*. The proof is immediate.*

Lemma 3.4. *A fixed point of* (3.3)–(3.5) *with $w_h = g_h$ on $\partial\Omega$ solves the nonlinear problem* (2.10)*. The proof is immediate.*

We denote by $\nu > 0$ a lower bound of the smallest eigenvalue of $\operatorname{cof} D^2 u$. Let $(u, \sigma) \in H^{k+3}(\Omega) \times H^{k+1}(\Omega)^{d\times d}$ denote the unique convex solution of (2.9) with $k \geq 1$. Note that by Sobolev embedding we then have $\sigma \in L^\infty(\Omega)^{d\times d}$. For $\rho > 0$, define

$$\bar{B}_h(\rho) = \{(w_h, \eta_h) \in V_h \times \Sigma_h,\ \|w_h - I_h u\|_{H^1} \leq \rho,\ \|\eta_h - I_h\sigma\|_{L^2} \leq h^{-1}\rho\}$$

$$\begin{aligned} Z_h = \{\, (w_h, \eta_h) \in V_h \times \Sigma_h, w_h = g_h \text{ on } \partial\Omega, \\ (\eta_h, \tau) + (\operatorname{div}\tau, Dw_h) - \langle Dw_h, \tau n\rangle = 0, \forall \tau \in \Sigma_h \,\} \text{ and} \end{aligned} \tag{3.6}$$

$$B_h(\rho) = \bar{B}_h(\rho) \cap Z_h. \tag{3.7}$$

Lemma 3.5. *$B_h(\rho) \neq \emptyset$ for h sufficiently small and $\rho = C_0 h^k$, for a positive constant $C_0 > 0$.*

Proof. We show that there exists $\eta_h \in \Sigma_h$ such that $(I_h u, \eta_h) \in Z_h$ for h sufficiently small. By (3.6) the problem: find $\eta_h \in \Sigma_h$ such that

$$(\eta_h, \tau) = -(\operatorname{div}\tau, DI_h u) + \langle DI_h u, \tau n\rangle, \quad \forall \tau \in \Sigma_h,$$

has a unique solution η_h by the Lax-Milgram Lemma. Clearly the right hand side defines a linear functional of $\tau \in \Sigma_h$. To see that it is a continuous functional, note that by the Schwarz inequality, (2.3) and (2.6)

$$\begin{aligned} |-(\operatorname{div}\tau, DI_h u) + \langle DI_h u, \tau\cdot n\rangle| &\leq C||\tau||_{H^1}||I_h u||_{H^1} + C||I_h u||_{H^1(\partial\Omega)}||\tau||_{L^2(\partial\Omega)} \\ &\leq C(h^{-1}||I_h u||_{H^1} + h^{-\frac{1}{2}}||I_h u||_{H^1(\partial\Omega)})||\tau||_{L^2}. \end{aligned}$$

Next, recall from (2.9)

$$(\sigma, \tau) = -(\operatorname{div}\tau, Du) + \langle Du, \tau n\rangle.$$

Therefore

$$(\eta_h - \sigma, \tau) = -(\operatorname{div}\tau, D(I_h u - u)) + \langle D(I_h u - u), \tau n\rangle.$$

Thus,

$$(\eta_h - I_h\sigma, \tau) = (\sigma - I_h\sigma, \tau) - (\operatorname{div}\tau, D(I_h u - u)) + \langle D(I_h u - u), \tau n\rangle.$$

Let $\tau = \eta_h - I_h\sigma$. Then by the Schwarz inequality, (2.3) and (2.6)

$$\begin{aligned} &\|\eta_h - I_h\sigma\|_{L^2}^2 \\ \leq\ &\|\sigma - I_h\sigma\|_{L^2}\|\eta_h - I_h\sigma\|_{L^2} + C\|\eta_h - I_h\sigma\|_{H^1}\|D(I_h u - u)\|_{L^2} \\ &+C\|D(I_h u - u)\|_{L^2(\partial\Omega)}\|\eta_h - I_h\sigma\|_{L^2(\partial\Omega)} \\ \leq\ &\|\sigma - I_h\sigma\|_{L^2}\|\eta_h - I_h\sigma\|_{L^2} + Ch^{-1}\|\eta_h - I_h\sigma\|_{L^2}\|D(I_h u - u)\|_{L^2} \\ &+Ch^{-\frac{1}{2}}\|D(I_h u - u)\|_{L^2(\partial\Omega)}\|\eta_h - I_h\sigma\|_{L^2}. \end{aligned}$$

Therefore

$$\begin{aligned} \|\eta_h - I_h\sigma\|_{L^2} &\leq \|\sigma - I_h\sigma\|_{L^2} + Ch^{-1}\|D(I_h u - u)\|_{L^2} + Ch^{-1/2}\|D(I_h u - u)\|_{L^2(\partial\Omega)} \\ &\leq Ch^{k+1} + Ch^{k-1} + Ch^{k-\frac{1}{2}} = h^{k-1}(Ch^2 + C + h^{\frac{1}{2}}) \\ &\leq h^{-1}(C_0 h^k) = h^{-1}\rho. \end{aligned}$$

This proves the result.

□

Remark 3.6. *For the solvability of* (3.3)–(3.5) *it is enough to study a certain mapping* $\tilde{T}_1 : V_h \to V_h$ *defined as follows. Given* $w_h \in V_h$ *there exists a unique* $\eta_h \in \Sigma_h$ *which satisfies the condition in* (3.6)*. The proof is analogue to the proof of the previous lemma. We define* $\tilde{T}_1(w_h) = T_1(w_h, \eta_h)$*. Next, note that with* (w_h, η_h) *satisfying* (3.6), $T_2(w_h, \eta_h)$ *is uniquely determined by* w_h, η_h *and* $T_1(w_h, \eta_h)$*. It then follows that if* u_h *is a fixed point of* $\tilde{T}_1$*, i.e.* $T_1(u_h, \sigma_h) = u_h$*, then* $T_2(u_h, \sigma_h) = \sigma_h$*. Thus* (u_h, σ_h) *is a fixed point of* T*. It is possible to describe the approach in* [8] *in terms of the mapping* $\tilde{T}_1$ *by referring to* η_h *as a discrete Hessian.*

The next lemma characterizes pairs $(w_h, \eta_h) \in V_h \times \Sigma_h$ which are in Z_h defined by (3.6).

Lemma 3.7. *Let* $(w_h, \eta_h) \in Z_h$*. Then*

$$|((\operatorname{cof} D^2 u) : \eta_h, v) + ((\operatorname{cof} D^2 u) D w_h, D v)| \leq C h ||v||_{H^1} ||w_h||_{H^1},$$

for all $v \in V_h \cap H_0^1(\Omega)$*.*

Proof. Recall that elements of Σ_h are continuous across inter-elements. We denote by $\mathcal{E}_h^i$ the set of interior faces. For a vector field, we denote by $[[w]] = w_{K^+} - w_{K^-}$ its jump across the intersection of the elements K^+ and K^-. We use n to denote the unit outward normal to the face e. Let h_e measure the size of the face e and denote by P_{Σ_h} the L^2 projection into the space Σ_h. With $A = \operatorname{cof} D^2 u$ we have for $v \in V_h \cap H_0^1(\Omega)$ and using (3.6)

$$\begin{aligned}
(A : \eta_h, v) &= (\eta_h, vA) = (\eta_h, P_{\Sigma_h}(vA)) \\
&= -(\operatorname{div} P_{\Sigma_h}(vA), D w_h) + \langle D w_h, (P_{\Sigma_h}(vA)) n \rangle_{\partial\Omega} \\
&= -\sum_{K \in \mathcal{T}_h} (\operatorname{div} P_{\Sigma_h}(vA), D w_h)_K + \langle D w_h, (P_{\Sigma_h}(vA)) n \rangle_{\partial\Omega} \\
&= \sum_{K \in \mathcal{T}_h} (P_{\Sigma_h}(vA), D^2 w_h)_K - \sum_{K \in \mathcal{T}_h} \langle D w_h, (P_{\Sigma_h}(vA)) n \rangle_{\partial K} \\
&\qquad + \langle D w_h, (P_{\Sigma_h}(vA)) n \rangle_{\partial\Omega} \\
&= \sum_{K \in \mathcal{T}_h} (P_{\Sigma_h}(vA), D^2 w_h)_K - \sum_{e \in \mathcal{E}_h^i} \langle [[D w_h]], (P_{\Sigma_h}(vA)) n \rangle_e \\
&= \sum_{K \in \mathcal{T}_h} (P_{\Sigma_h}(vA) - vA, D^2 w_h)_K + \sum_{K \in \mathcal{T}_h} (vA, D^2 w_h)_K \\
&\qquad - \sum_{e \in \mathcal{E}_h^i} \langle [[D w_h]], (P_{\Sigma_h}(vA)) n \rangle_e \\
&= \sum_{K \in \mathcal{T}_h} (P_{\Sigma_h}(vA) - vA, D^2 w_h)_K - \sum_{K \in \mathcal{T}_h} (\operatorname{div}(vA), D w_h)_K \\
&\qquad + \sum_{K \in \mathcal{T}_h} \langle D w_h, (vA) n \rangle_{\partial K} - \sum_{e \in \mathcal{E}_h^i} \langle [[D w_h]], (P_{\Sigma_h}(vA)) n \rangle_e \\
&= \sum_{K \in \mathcal{T}_h} (P_{\Sigma_h}(vA) - vA, D^2 w_h)_K - \sum_{K \in \mathcal{T}_h} (A D v, D w_h)_K \\
&\qquad + \sum_{e \in \mathcal{E}_h^i} \langle [[D w_h]], (vA) n \rangle_e - \sum_{e \in \mathcal{E}_h^i} \langle [[D w_h]], (P_{\Sigma_h}(vA)) n \rangle_e
\end{aligned}$$

$$= -(ADw_h, Dv) + \sum_{K\in\mathcal{T}_h} (P_{\Sigma_h}(vA) - vA, D^2 w_h)_K - \sum_{e\in\mathcal{E}_h^i} \langle [[Dw_h]], (P_{\Sigma_h}(vA))n - (vA)n\rangle_e,$$

where we used $v = 0$ on $\partial\Omega$, the divergence-free row property of $\operatorname{cof} D^2u$, i.e. $\operatorname{div} A = 0$ and the symmetry of A. Using again integration by parts, we obtain

$$\begin{aligned}
&(A:\eta_h, v) + (ADw_h, Dv)\\
&= -\sum_{K\in\mathcal{T}_h} (\operatorname{div}(P_{\Sigma_h}(vA) - vA), Dw_h)_K + \sum_{K\in\mathcal{T}_h} \langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_{\partial K}\\
&\quad - \sum_{e\in\mathcal{E}_h^i} \langle [[Dw_h]], (P_{\Sigma_h}(vA))n - (vA)n\rangle_e\\
&= -\sum_{K\in\mathcal{T}_h} (\operatorname{div}(P_{\Sigma_h}(vA) - vA), Dw_h)_K + \langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_{\partial\Omega}.
\end{aligned}$$

We define

$$\Gamma = -\sum_{K\in\mathcal{T}_h} (\operatorname{div}(P_{\Sigma_h}(vA) - vA), Dw_h)_K + \langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_{\partial\Omega}.$$

Therefore, $(A : \eta_h, v) + (ADw_h, Dv) = \Gamma$. To estimate Γ, we proceed with an approach similar to the one taken in [8]. By Lemma 4.4 and 4.5 of [8], one obtains

$$||P_{\Sigma_h}(vA) - vA||_{H^1} \leq Ch||v||_{H^1} \tag{3.8}$$

$$\Big(\sum_{e\in\partial\Omega} h_e^{-1} ||P_{\Sigma_h}(vA) - vA||^2_{L^2(e)}\Big)^{1/2} \leq Ch||v||_{H^1}. \tag{3.9}$$

The results in [8] are stated in terms of A_h the L^2 projection of A into Σ_h. But the analysis there easily holds. Put

$$||v||_{H^k(\mathcal{T}_h)} = \Bigg(\sum_{K\in\mathcal{T}_h} ||v||^2_{H^k(K)}\Bigg)^{\frac{1}{2}}.$$

For example, to prove (3.8), note that v is a piecewise polynomial of degree k and hence $||v||_{H^{k+1}(\mathcal{T}_h)} = ||v||_{H^k(\mathcal{T}_h)}$. Thus using an inverse estimate and the approximation properties of P_{Σ_h}, we have for $m = 0, 1$

$$\begin{aligned}
||P_{\Sigma_h}(vA) - vA||_{H^m(\mathcal{T}_h)} &\leq Ch^{k+1-m}||v||_{H^{k+1}(\mathcal{T}_h)} = Ch^{k+1-m}||v||_{H^k(\mathcal{T}_h)}\\
&\leq Ch^{k+1-m}h^{1-k}||v||_{H^1(\mathcal{T}_h)} \leq Ch^{2-m}||v||_{H^1(\mathcal{T}_h)}.
\end{aligned}$$

Similarly, one proves (3.9) using the above result, the trace inequality and inverse estimates. Thus

$$\Bigg| -\sum_{K\in\mathcal{T}_h} (\operatorname{div}(P_{\Sigma_h}(vA) - vA), Dw_h)_K \Bigg| \leq Ch||v||_{H^1}||w||_{H^1}. \tag{3.10}$$

Moreover

$$\begin{aligned}
&\left|\langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_{\partial\Omega}\right| \\
&= \left|\sum_{e\in\partial\Omega} \langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_e\right| \\
&= \left|\sum_{e\in\partial\Omega} \langle h_e^{-1/2}(P_{\Sigma_h}(vA) - vA)n, h_e^{1/2}Dw_h\rangle_e\right| \\
&\leq \left(\sum_{e\in\partial\Omega} h_e^{-1}||P_{\Sigma_h}(vA) - vA||^2_{L^2(e)}\right)^{1/2}\left(\sum_{e\in\partial\Omega} h_e||Dw_h||^2_{L^2(e)}\right)^{1/2}.
\end{aligned}$$

By (3.9) and the trace-inverse inequality (2.6), we get

$$|\langle (P_{\Sigma_h}(vA) - vA)n, Dw_h\rangle_{\partial\Omega}| \leq Ch||v||_{H^1}||w||_{H^1}. \tag{3.11}$$

By (3.10) and (3.11), we obtain $|\Gamma| \leq Ch||v||_{H^1}||w||_{H^1}$. This completes the proof. □

Lemma 3.8. *The mapping T does not move the center $(I_hu, I_h\sigma)$ of the ball $\bar{B}_h(\rho)$ too far, i.e.*

$$||I_hu - T_1(I_hu, I_h\sigma)||_{H^1} \leq Ch^{k+1}\|\sigma\|^{d-1}_{L^\infty}\|\sigma\|_{H^{k+1}} \equiv C_1h^{k+1} \tag{3.12}$$

$$\begin{aligned}
\|I_h\sigma - T_2(I_hu, I_h\sigma)\|_{L^2} &\leq Ch^{k-1}(\|\sigma\|^{d-1}_{L^\infty}\|\sigma\|_{H^{k+1}} + \|\sigma\|_{H^{k+1}} + \|u\|_{H^{k+1}}) \\
&\equiv C_2h^{k-1}.
\end{aligned} \tag{3.13}$$

Proof. Since $T_1(I_hu, I_h\sigma) = I_hu$ on $\partial\Omega$ by (3.5), we have $v = I_hu - T_1(I_hu, I_h\sigma) \in V_h \cap H^1_0(\Omega)$. Using it in (3.4), $\det D^2u = \det\sigma = f$ and using the strict convexity of D^2u and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\nu|I_hu - T_1(I_hu, I_h\sigma)|^2_{H^1} &\leq ||\det I_h\sigma - \det\sigma||_{L^2}||I_hu - T_1(I_hu, I_h\sigma)||_{L^2} \\
&\leq ||\det I_h\sigma - \det\sigma||_{L^2}||I_hu - T_1(I_hu, I_h\sigma)||_{H^1}.
\end{aligned}$$

By Poincaré's inequality

$$||I_hu - T_1(I_hu, I_h\sigma)||_{H^1} \leq C||\det I_h\sigma - \det\sigma||_{L^2}.$$

By Lemma 2.4, on each element K

$$\det I_h\sigma - \det\sigma = \mathrm{cof}(tI_h\sigma + (1-t)\sigma) : (I_h\sigma - \sigma),$$

for some $t \in [0,1]$. By (2.1) we have $\|I_h\sigma\|_{L^\infty} \leq C\|\sigma\|_{L^\infty}$. Thus by Lemma 2.5

$$\begin{aligned}
\|\det(I_h\sigma) - \det\sigma\|_{L^2(K)} &\leq C\|tI_h\sigma + (1-t)\sigma\|^{d-1}_{L^\infty(K)}\|I_h\sigma - \sigma\|_{L^2(K)} \\
&\leq C\|\sigma\|^{d-1}_{L^\infty}\|I_h\sigma - \sigma\|_{L^2(K)}.
\end{aligned}$$

Therefore

$$\|\det(I_h\sigma) - \det\sigma\|_{L^2} \leq C\|\sigma\|^{d-1}_{L^\infty}\|I_h\sigma - \sigma\|_{L^2}.$$

And so by (2.1)

$$||I_hu - T_1(I_hu, I_h\sigma)||_{H^1} \leq C\|\sigma\|^{d-1}_{L^\infty}\|I_h\sigma - \sigma\|_{L^2} \leq Ch^{k+1}\|\sigma\|^{d-1}_{L^\infty}\|\sigma\|_{H^{k+1}},$$

which proves (3.12).

Next, use $w_h = I_hu$, $\eta_h = I_h\sigma$ and $\tau = I_h\sigma - T_2(I_hu, I_h\sigma)$ in (3.3) to obtain

$$\begin{aligned}
||\tau||^2_{L^2} = -(\mathrm{div}\,\tau, D(w_h - T_1(w_h,\eta_h))) + \langle D(w_h - T_1(w_h,\eta_h)), \tau n\rangle + (\eta_h, \tau) \\
+ (\mathrm{div}\,\tau, Dw_h) - \langle Dw_h, \tau n\rangle.
\end{aligned}$$

Note that

$$(\sigma,\tau)+(\operatorname{div}\tau, Du)-\langle Du,\tau n\rangle=0,\quad \forall\, \tau\in H^1(\Omega),$$

and thus

$$\begin{aligned}||\tau||_{L^2}^2 &= -(\operatorname{div}\tau, D(I_hu-T_1(I_hu,I_h\sigma)))+\langle D(I_hu-T_1(I_hu,I_h\sigma)),\tau n\rangle\\ &\quad +(I_h\sigma-\sigma,\tau)+(\operatorname{div}\tau, D(I_hu-u))-\langle D(I_hu-u),\tau n\rangle.\end{aligned}$$

By Cauchy-Schwarz and Poincaré's inequalities, the inverse estimate (2.3), (3.12), the trace-inverse inequality (2.6) and the interpolation estimates (2.1), we have

$$\begin{aligned}&||\tau||_{L^2}^2\\ &\leq ||\tau||_{H^1}||I_hu-T_1(I_hu,I_h\sigma)||_{H^1}+C||I_hu-T_1(I_hu,I_h\sigma)||_{H^1(\partial\Omega)}||\tau||_{L^2(\partial\Omega)}\\ &\quad +||I_h\sigma-\sigma||_{L^2}||\tau||_{L^2}+||\tau||_{H^1}||I_hu-u||_{H^1}+C||I_hu-u||_{H^1(\partial\Omega)}||\tau||_{L^2(\partial\Omega)}\\ &\leq Ch^k\|\sigma\|_{L^\infty}^{d-1}\|\sigma\|_{H^{k+1}}||\tau||_{L^2}+Ch^{-1}||I_hu-T_1(I_hu,I_h\sigma)||_{H^1}||\tau||_{L^2}\\ &\quad +Ch^{k+1}\|\sigma\|_{H^{k+1}}||\tau||_{L^2}+Ch^{k-1}\|u\|_{H^{k+1}}||\tau||_{L^2}+Ch^{k-1}|u|_{H^{k+1/2}(\partial\Omega)}||\tau||_{L^2}\\ &\leq Ch^k\|\sigma\|_{L^\infty}^{d-1}\|\sigma\|_{H^{k+1}}||\tau||_{L^2}+Ch^{k+1}\|\sigma\|_{H^{k+1}}||\tau||_{L^2}+Ch^{k-1}\|u\|_{H^{k+1}}||\tau||_{L^2}.\end{aligned}$$

We conclude that $||\tau||_{L^2}\leq Ch^{k-1}(\|\sigma\|_{L^\infty}^{d-1}\|\sigma\|_{H^{k+1}}+\|\sigma\|_{H^{k+1}}+\|u\|_{H^{k+1}})$ which is (3.13). □

Lemma 3.9. *Let $\rho>0$ and (w_1,η_1) and (w_2,η_2) in $B_h(\rho)$. We have*

$$||T_2(w_1,\eta_1)-T_2(w_2,\eta_2)||_{L^2}\leq C_4h^{-1}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1},\tag{3.14}$$

and C_4 can be chosen such that $C_4\geq 1$.

Proof. For (w_1,η_1) and (w_2,η_2) in $B_h(\rho)$. We have using (3.3)

$$\begin{aligned}((T_2(w_1,\eta_1)-T_2(w_2,\eta_2),\tau) &= -(\operatorname{div}\tau, D((T_1(w_1,\eta_1)-T_1(w_2,\eta_2)))\\ &\quad +\langle D((T_1(w_1,\eta_1)-T_1(w_2,\eta_2)),\tau n\rangle.\end{aligned}$$

Choosing $\tau=T_2(w_1,\eta_1)-T_2(w_2,\eta_2)$ and using Cauchy-Schwarz and Poincaré's inequalities, the inverse estimate (2.3) and the trace-inverse inequality (2.6), we obtain

$$\begin{aligned}&||\tau||_{L^2}^2\\ &\leq ||\tau||_{H^1}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1}+Ch^{-1}||\tau||_{L^2}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1}\\ &\leq Ch^{-1}||\tau||_{L^2}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1}.\end{aligned}$$

We conclude that (3.14) holds. □

Lemma 3.10. *Let $\rho>0$ and (w_1,η_1) and (w_2,η_2) in $B_h(\rho)$. We have*

$$\begin{aligned}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1}^2 &\leq C(h^{k+1-d/2}||u||_{H^{k+3}}+h^{-\frac{d}{2}-1}\rho+||u||_{W^{2,\infty}})^{d-2}\\ &\quad (h^{k+1}||u||_{H^{k+3}}+h^{-1}\rho)||\eta_1-\eta_2||_{L^2}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{L^\infty}\\ &\quad +Ch||w_1-w_2||_{H^1}||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1}.\end{aligned}\tag{3.15}$$

Proof. Using (3.4) we have

$$\begin{aligned}&((\operatorname{cof} D^2u)D(T_1(w_1,\eta_1)-T_1(w_2,\eta_2)),Dv)\\ &=((\operatorname{cof} D^2u)D(w_1-w_2),Dv)+(\det\eta_1-\det\eta_2,v)+((\operatorname{cof} D^2u):(\eta_1-\eta_2),v)\\ &\quad -((\operatorname{cof} D^2u):(\eta_1-\eta_2),v),\end{aligned}$$

for all $v \in V_h$. By the definition of $B_h(\rho)$, (3.7) and Lemma 3.7, we have

$$\begin{aligned}(3.16)\qquad &((\operatorname{cof} D^2u)D(T_1(w_1,\eta_1)-T_1(w_2,\eta_2)),Dv)\\ &=-((\operatorname{cof} D^2u):(\eta_1-\eta_2),v)+(\det\eta_1-\det\eta_2,v)+\Gamma,\end{aligned}$$

for all $v \in V_h$ with

$$(3.17)\qquad |\Gamma| \leq Ch||w_1-w_2||_{H^1}||v||_{H^1}.$$

By Lemma 2.4, on each element K, for some $t \in [0,1]$ we have

$$\det\eta_1-\det\eta_2=\operatorname{cof}(t\eta_1+(1-t)\eta_2):(\eta_1-\eta_2),$$

where for simplicity we do not explicitly indicate the dependence of t on K. Therefore on each element K

$$\begin{aligned}(3.18)\qquad (\operatorname{cof} D^2u):(\eta_1-\eta_2)-(\det\eta_1-\det\eta_2)=\big(&(\operatorname{cof} D^2u)\\ &-\operatorname{cof}(t\eta_1+(1-t)\eta_2)\big):(\eta_1-\eta_2).\end{aligned}$$

We have $T_1(w_1,\eta_1)-T_1(w_2,\eta_2)=w_1-w_2=0$ on $\partial\Omega$ by (3.5). We can then use $v=T_1(w_1,\eta_1)-T_1(w_2,\eta_2)$ in (3.16). By (3.18), and with $\sigma=D^2u$, we get

$$(3.19)\qquad \nu|v|^2_{H^1} \leq \Big|\sum_{K\in\mathcal{T}_h}\big(\big((\operatorname{cof}\sigma)-\operatorname{cof}(t\eta_1+(1-t)\eta_2)\big):(\eta_1-\eta_2),v\big)_K\Big|+|\Gamma|.$$

Let us define

$$A_K=\big(\big((\operatorname{cof}\sigma)-\operatorname{cof}(t\eta_1+(1-t)\eta_2)\big):(\eta_1-\eta_2),v\big)_K.$$

By Hölder's inequality, Lemma 2.6, the interpolation estimate (2.1) and the definition of $B_h(\rho)$ (3.7), we have for some $s \in [0,1]$ which depends on K

$$\begin{aligned}A_K &\leq C||s\sigma+(1-s)(t\eta_1+(1-t)\eta_2)||^{d-2}_{L^\infty(K)}||\sigma-(t\eta_1+(1-t)\eta_2)||_{L^2(K)}\\ &\qquad\qquad ||\eta_1-\eta_2||_{L^2(K)}||v||_{L^\infty}\\ &\leq C||s(\sigma-I_h\sigma)+(1-s)(t(\eta_1-I_h\sigma)+(1-t)(\eta_2-I_h\sigma))+I_h\sigma||^{d-2}_{L^\infty(K)}\\ &\qquad ||\sigma-(t\eta_1+(1-t)\eta_2||_{L^2(K)}||\eta_1-\eta_2||_{L^2(K)}||v||_{L^\infty}\\ &\leq C(||\sigma-I_h\sigma||_{L^\infty}+t||\eta_1-I_h\sigma||_{L^\infty}+(1-t)||\eta_2-I_h\sigma||_{L^\infty}\\ &\qquad +\|I_h\sigma||_{L^\infty})^{d-2}||\sigma-(t\eta_1+(1-t)\eta_2)||_{L^2(K)}||\eta_1-\eta_2||_{L^2(K)}||v||_{L^\infty}\\ &\leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}}+th^{-\frac{d}{2}}||\eta_1-I_h\sigma||_{L^2}+(1-t)h^{-\frac{d}{2}}||\eta_2-I_h\sigma||_{L^2}\\ &\qquad +\|I_h\sigma||_{L^\infty})^{d-2}||\sigma-(t\eta_1+(1-t)\eta_2)||_{L^2(K)}||\eta_1-\eta_2||_{L^2(K)}||v||_{L^\infty}\\ &\leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}}+h^{-\frac{d}{2}-1}\rho+||\sigma||_{L^\infty})^{d-2}\\ &\qquad ||\sigma-(t\eta_1+(1-t)\eta_2)||_{L^2(K)}||\eta_1-\eta_2||_{L^2(K)}||v||_{L^\infty}.\end{aligned}$$

Moreover

$$\begin{aligned}\sigma-(t\eta_1+(1-t)\eta_2)&=\sigma-I_h\sigma+tI_h\sigma+(1-t)I_h\sigma-(t\eta_1+(1-t)\eta_2)\\ &=\sigma-I_h\sigma+t(I_h\sigma-\eta_1)+(1-t)(I_h\sigma-\eta_2).\end{aligned}$$

We conclude using again (2.1) that

$$\begin{aligned}||\sigma-(t\eta_1+(1-t)\eta_2)||_{L^2(K)} &\leq ||\sigma-I_h\sigma||_{L^2(K)}+t||I_h\sigma-\eta_1||_{L^2(K)}\\ &\qquad +(1-t)||I_h\sigma-\eta_2||_{L^2(K)}\\ &\leq ||\sigma-I_h\sigma||_{L^2(K)}+||I_h\sigma-\eta_1||_{L^2(K)}\\ &\qquad +||I_h\sigma-\eta_2||_{L^2(K)}.\end{aligned}$$

It follows that

$$\sum_{K\in\mathcal{T}_h} |A_K| \leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}} + h^{-\frac{d}{2}-1}\rho + ||\sigma||_{L^\infty})^{d-2}||v||_{L^\infty}$$
$$\sum_{K\in\mathcal{T}_h} \big(||\sigma - I_h\sigma||_{L^2(K)} + ||I_h\sigma - \eta_1||_{L^2(K)} + ||I_h\sigma - \eta_2||_{L^2(K)}\big)||\eta_1 - \eta_2||_{L^2(K)}$$
$$\leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}} + h^{-\frac{d}{2}-1}\rho + ||\sigma||_{L^\infty})^{d-2}||v||_{L^\infty}$$
$$\big(||\sigma - I_h\sigma||_{L^2} + ||I_h\sigma - \eta_1||_{L^2} + ||I_h\sigma - \eta_2||_{L^2}\big)||\eta_1 - \eta_2||_{L^2}$$
$$\leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}} + h^{-\frac{d}{2}-1}\rho + ||\sigma||_{L^\infty})^{d-2}||v||_{L^\infty}$$
$$(Ch^{k+1}||\sigma||_{H^{k+1}} + Ch^{-1}\rho)||\eta_1 - \eta_2||_{L^2},$$

where we again used the interpolation estimate (2.1) and the definition of $B_h(\rho)$.

Combined with (3.19) we obtain

$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||^2_{H^1} \leq C(Ch^{k+1-d/2}||\sigma||_{H^{k+1}} + h^{-\frac{d}{2}-1}\rho + \|I_h\sigma\|_{L^\infty})^{d-2}$$
$$(Ch^{k+1}||\sigma||_{H^{k+1}} + Ch^{-1}\rho)||\eta_1 - \eta_2||_{L^2}$$
$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{L^\infty} + |\Gamma|$$
$$\leq C(h^{k+1-d/2}||u||_{H^{k+3}} + h^{-\frac{d}{2}-1}\rho + \|\sigma\|_{L^\infty})^{d-2}$$
$$(h^{k+1}||u||_{H^{k+3}} + h^{-1}\rho)||\eta_1 - \eta_2||_{L^2}$$
$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{L^\infty} + |\Gamma|.$$

In view of (3.17), this completes the proof. □

Lemma 3.11. *Let* $\rho(h) = 2C_3h^k$ *where* $C_3 = \max(C_0, C_1, C_2)$ *with* C_0 *the constant in Lemma 3.5 and* C_1, C_2 *the constants from Lemma 3.8. Then the mapping* T_1 *has a strict contraction property in* $B_h(\rho)$ *for* h *sufficiently small. That is*

$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{H^1} \leq \frac{h}{4C_4}||\eta_1 - \eta_2||_{L^2} + \frac{1}{4C_4}||w_1 - w_2||_{H^1}. \tag{3.20}$$

for $(w_1,\eta_1),(w_2,\eta_2) \in B_h(\rho)$.

Proof. The proofs in dimensions 2 and 3 are different.

Case $d = 2$. Using the discrete Sobolev inequality (2.7) and (3.15), we have

$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{H^1} \leq C(h^{k+1}||u||_{H^{k+3}} + h^{-1}\rho)(1 + |\ln h|^{\frac{1}{2}})||\eta_1 - \eta_2||_{L^2}$$
$$+ Ch||w_1 - w_2||_{H^1}$$
$$\leq C(h^k + h^{-2}\rho)(1 + |\ln h|^{\frac{1}{2}})h||\eta_1 - \eta_2||_{L^2}$$
$$+ Ch||w_1 - w_2||_{H^1}$$
$$\leq C(h^k + h^{k-2})(1 + |\ln h|^{\frac{1}{2}})h||\eta_1 - \eta_2||_{L^2}$$
$$+ Ch||w_1 - w_2||_{H^1},$$

where we also used the expression of ρ given in the lemma to be proved.

For $k \geq 3$ and h sufficiently small we have $C(h^k + h^{k-2})(1 + |\ln h|^{\frac{1}{2}}) \leq 1/(4C_4)$ and $Ch \leq 1/(4C_4)$. Thus (3.20) holds.

Case $d=3$. Using the discrete Sobolev inequality (2.8) and (3.15), we have

$$\begin{aligned}
||T_1(w_1,\eta_1)-T_1(w_2,\eta_2)||_{H^1} &\leq Ch^{-\frac{1}{2}}(h^{k-1/2}||u||_{H^{k+3}}+h^{-\frac{5}{2}}\rho+||u||_{W^{2,\infty}})\\
&\qquad (h^{k+1}||u||_{H^{k+3}}+h^{-1}\rho)||\eta_1-\eta_2||_{L^2}\\
&\qquad\quad +Ch||w_1-w_2||_{H^1}\\
&\leq C(h^{k-1}+h^{k-3}+h^{-\frac{1}{2}})(h^{k+1}+h^{k-1})||\eta_1-\eta_2||_{L^2}\\
&\qquad\quad +Ch||w_1-w_2||_{H^1}\\
&\leq C(h^{k-1}+h^{k-3}+h^{-\frac{1}{2}})(h^{k}+h^{k-2})h||\eta_1-\eta_2||_{L^2}\\
&\qquad\quad +Ch||w_1-w_2||_{H^1},
\end{aligned}$$

where we also used the expression of ρ given in the lemma to be proved.

For $k\geq 3$ and h sufficiently small we have $C(h^{k-1}+h^{k-3}+h^{-\frac{1}{2}})(h^k+h^{k-2})\leq 1/(4C_4)$ and $Ch\leq 1/(4C_4)$. Thus (3.20) holds as well.

□

Lemma 3.12. *T maps $B_h(\rho)$ into itself for h sufficiently small and with $\rho(h)$ given in Lemma 3.11.*

Proof. Let $(w_h,\eta_h)\in B_h(\rho)$. By definition, $||w_h-I_hu||_{H^1}\leq\rho$ and $||\eta_h-I_h\sigma||\leq h^{-1}\rho$. By (3.20), (3.12), and using $1/C_4\leq 1$

$$\begin{aligned}
||T_1(w_h,\eta_h)-I_hu||_{H^1} &\leq ||T_1(w_h,\eta_h)-T_1(I_hu,I_h\sigma)||_{H^1}+||T_1(I_hu,I_h\sigma)-I_hu||_{H^1}\\
&\leq \frac{h}{4}||\eta_h-I_h\sigma||_{L^2}+\frac{1}{4}||u_h-I_hu||_{H^1}+C_1h^{k+1}\\
&\leq \frac{\rho}{2}+C_3h^k=\frac{\rho}{2}+\frac{\rho}{2}\\
&\leq\rho,
\end{aligned}$$

for h sufficiently small. In addition, by (3.20), (3.14) and (3.13) and a similar argument we get

$$\begin{aligned}
||T_2(w_h,\eta_h)-I_h\sigma||_{L^2} &\leq ||T_2(w_h,\eta_h)-T_2(I_hu,I_h\sigma)||_{L^2}+||T_2(I_hu,I_h\sigma)-I_h\sigma||_{L^2}\\
&\leq C_4h^{-1}||T_1(w_h,\eta_h)-T_1(I_hu,I_h\sigma)||_{H^1}+||T_2(I_hu,I_h\sigma)-I_h\sigma||_{L^2}\\
&\leq \frac{1}{4}||\eta_h-I_h\sigma||_{L^2}+\frac{h^{-1}}{4}||u_h-I_hu||_{H^1}+C_2h^{k-1}\\
&\leq \frac{h^{-1}\rho}{2}+C_3h^{k-1}=\frac{h^{-1}\rho}{2}+\frac{h^{-1}\rho}{2}\\
&\leq h^{-1}\rho.
\end{aligned}$$

By (3.3) $(T_1(w_h,\eta_h),T_2(w_h,\eta_h))$ is in the space Z_h defined by (3.6). This concludes the proof. □

We can now claim

Theorem 3.13. *Let $(u,\sigma)\in H^{k+3}(\Omega)\times H^{k+1}(\Omega)^{d\times d}$ denotes the unique convex solution of (2.9) with $k\geq 3$. Then the problem (2.10) has a unique solution in $B_h(\rho)\subset V_h\times\Sigma_h$ for h sufficiently small and with $\rho(h)$ given in Lemma 3.11.*

Proof. The proof follows from the Brouwer fixed point theorem. For h sufficiently small and for $(w_1, \eta_1), (w_2, \eta_2) \in B_h(\rho)$, by (3.20) and (3.14)

$$||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{H^1} + ||T_2(w_1,\eta_1) - T_2(w_2,\eta_2)||_{L^2} \leq (1 + Ch^{-1})||T_1(w_1,\eta_1) - T_1(w_2,\eta_2)||_{H^1} \leq (1 + Ch^{-1})||w_1 - w_2||_{H^1} + C||\eta_1 - \eta_2||_{L^2}.$$

Hence the mapping T is continuous in $B_h(\rho)$. Since for h sufficiently small and the choice of $\rho(h)$, the continuous mapping T maps the closed ball $B_h(\rho)$ into itself, there exists a fixed point (u_h, σ_h) in $B_h(\rho)$.

Assume that (w_h^1, η_h^1) and (w_h^2, η_h^2) are two fixed points of T. Then $T_1(w_h^1, \eta_h^1) = w_h^1$ and $T_1(w_h^2, \eta_h^2) = w_h^2$. By (3.20) and using $1/C_4 \leq 1$, we have

$$||w_h^1 - w_h^2|| \leq \frac{h}{4}||\eta_h^1 - \eta_h^2||_{L^2} + \frac{1}{4}||w_h^1 - w_h^2||_{H^1},$$

and so

$$||w_h^1 - w_h^2|| \leq \frac{h}{3}||\eta_h^1 - \eta_h^2||_{L^2}.$$

We also have $T_2(w_h^1, \eta_h^1) = \eta_h^1$ and $T_2(w_h^2, \eta_h^2) = \eta_h^2$. By (3.14)

$$||\eta_h^1 - \eta_h^2||_{L^2} \leq h^{-1}||w_h^1 - w_h^2|| \leq \frac{1}{3}||\eta_h^1 - \eta_h^2||_{L^2}.$$

This implies $\eta_h^1 = \eta_h^2$ and so $w_h^1 = w_h^2$. This proves uniqueness. □

The following error estimates hold

Theorem 3.14. *Under the assumptions of Theorem 3.13, the solution (u_h, σ_h) of (3.3)–(3.5) satisfies*

$$||u - u_h||_{H^1} \leq Ch^k \tag{3.21}$$

$$||\sigma - \sigma_h||_{L^2} \leq Ch^{k-1}. \tag{3.22}$$

Proof. By the definition of the ball $B_h(\rho)$ (3.7), the existence of the solution (u_h, σ_h) in $B_h(\rho)$ with $\rho = O(h^k)$ given in Lemma 3.11, we have

$$||I_h u - u_h||_{H^1} \leq Ch^k$$
$$||I_h \sigma - \sigma_h||_{L^2} \leq Ch^{k-1}.$$

The estimates (3.21) and (3.22) then follow from triangular inequalities and standard interpolation inequalities. □

Remark 3.15. *For computational efficiency, one may impose that elements of Σ_h are symmetric matrix fields. The analysis of this paper also holds in that case.*

Remark 3.16. *It is not necessary to use the same polynomial degrees for V_h and Σ_h. However for V_h the Lagrange space of degree k_1 and Σ_h a finite element space of matrix fields with each component in a Lagrange space of degree k_2, we need $k_2 \geq k_1 \geq 3$ for the analysis of the paper to hold. Lemma 3.7 breaks down for $k_2 < k_1$. The analogue of (3.8) for v a piecewise polynomial of degree k_1 gives*

$$||P_{\Sigma_h}(vA) - vA||_{H^m(\mathcal{T}_h)} \leq Ch^{k_2+1-m}||v||_{H^{k_2+1}(\mathcal{T}_h)} = Ch^{k_2+1-m}||v||_{H^{k_2}(\mathcal{T}_h)} \leq Ch^{k_2+1-m}h^{1-k_2}||v||_{H^1(\mathcal{T}_h)} \leq Ch^{2-m}||v||_{H^1(\mathcal{T}_h)},$$

only if $k_2 \geq k_1$.

TABLE 1. Linear Lagrange elements for a smooth solution $u(x,y) = e^{(x^2+y^2)/2}$

h	$\|\|u-u_h\|\|_{L^2}$	rate	$\|u-u_h\|_{H^1}$	rate	$\|\|\sigma-\sigma_h\|\|_{L^2}$	rate
1/2	$1.05\ 10^{-1}$		$5.41\ 10^{-1}$		4.14	
1/4	$2.53\ 10^{-2}$	2.05	$2.80\ 10^{-1}$	0.95	3.13	0.40
1/8	$5.95\ 10^{-3}$	2.09	$1.41\ 10^{-1}$	0.99	2.35	0.41
1/16	$1.46\ 10^{-3}$	2.02	$7.08\ 10^{-2}$	0.99	1.71	0.45
1/32	$3.70\ 10^{-4}$	1.98	$3.54\ 10^{-2}$	1	1.22	0.49
1/64	$9.41\ 10^{-5}$	1.97	$1.77\ 10^{-2}$	1	0.87	0.49
1/128	$2.37\ 10^{-5}$	1.99	$8.85\ 10^{-3}$	1	0.61	0.51

4. Numerical Results

We give numerical results for linear finite elements and a smooth solution $u(x,y) = e^{(x^2+y^2)/2}$ on the unit square $[0,1]^2$, c.f. Table 1. The method was implemented with the software freefem++ on a uniform mesh obtained by dividing the domain into squares, then each square is divided into two triangles by taking the diagonal with positive slope. The numerical results indicate a superconvergence result for $||\sigma - \sigma_h||_{L^2}$.

Numerical results for the method analyzed in this paper were reported in [7, 8] for the two dimensional problem and high order elements, i.e. $k \geq 2$. Therefore, we do not repeat these tests here. The authors of [7, 8] reported the divergence of the method for linear finite elements. This is probably the case if the method is implemented in a primal form with the discrete Hessian, which does not necessarily converge for linear elements, eliminated from the equations. It has been reported in [8] that penalizing the jumps of the first derivatives make the method suitable for linear finite elements and non smooth solutions. Our numerical results indicate that for smooth solutions, there is an advantage in considering the method in mixed form using linear elements for all the variables.

The reader interested in discontinuous elements for Σ_h may refer to [8] or prove a version of Lemma 3.7 without using the continuity across inter elements of elements of Σ_h.

Acknowledgments

The authors would like to thank the referee for a careful reading of the manuscript and suggestions to improve the paper. G. Awanou was supported in part by a 2009-2013 Sloan Foundation Fellowship and NSF-DMS grant 1319640. H. Li was partially supported by NSF-DMS grant 1158839.

References

[1] G. Awanou, Pseudo transient continuation and time marching methods for Monge-Ampère type equations, http://arxiv.org/pdf/1301.5891.pdf, 2013.

[2] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, The construction of preconditioners for elliptic problems by substructuring. I, Math. Comp., 47 (1986), no. 175, 103–134.

[3] S. C. Brenner, T. Gudi, M. Neilan, and L.Y. Sung, C^0 penalty methods for the fully nonlinear Monge-Ampère equation, Math. Comp., 80 (2011), no. 276, 1979–1995.

[4] S. C. Brenner and S. L. Ridgway, The Mathematical Theory of Finite Element Methods, second ed., Texts in Applied Mathematics, vol. 15, Springer-Verlag, New York, 2002.

[5] L. C. Evans, Partial Differential Equations, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998.

[6] X. Feng and M. Neilan, Error analysis for mixed finite element approximations of the fully nonlinear Monge-Ampère equation based on the vanishing moment method, SIAM J. Numer. Anal., 47 (2009), no. 2, 1226–1250.

[7] O. Lakkis and T. Pryer, A finite element method for nonlinear elliptic problems, SIAM J. Sci. Comput., 35 (2013), no. 4, A2025–A2045.

[8] M. Neilan, Finite element methods for fully nonlinear second order PDEs based on the discrete Hessian, J. Comput. Appl. Math., 263 (2014), 351–369.

[9] N. S. Trudinger and X. J. Wang, Boundary regularity for the Monge-Ampère and affine maximal surface equations, Ann. of Math., (2) 167 (2008), no. 3, 993–1028.

Department of Mathematics, Statistics, and Computer Science, M/C 249. University of Illinois at Chicago, Chicago, IL 60607-7045, USA
E-mail: `awanou@uic.edu`
URL: `http://www.math.uic.edu/~awanou/`

Department of Mathematics, Wayne State University, 656 W. Kirby, Detroit, MI 48202, USA
E-mail: `hli@math.wayne.edu`
URL: `http://www.math.wayne.edu/~hli/`