Calcolo (2015) 52:503–518
DOI 10.1007/s10092-014-0127-7

Quadratic mixed finite element approximations of the Monge–Ampère equation in 2D

Gerard Awanou

Received: 5 August 2014 / Accepted: 29 October 2014 / Published online: 11 November 2014

Abstract We give error estimates for a mixed finite element approximation of the two-dimensional elliptic Monge–Ampère equation with the unknowns approximated by Lagrange finite elements of degree two. The variables in the formulation are the scalar variable and the Hessian matrix.

Keywords Rescaling argument · Monge–Ampère · Discrete Hessian · Mixed methods

Mathematics Subject Classification (2010) 65N30 · 35J25

1 Introduction

Let Ω be a convex polygonal domain of $\mathbb{R}^2$ with boundary $\partial\Omega$. We are interested in a mixed finite element method for the nonlinear elliptic Monge–Ampère equation: find a smooth convex function u such that

$$\begin{aligned} \det(D^2u) &= f \quad \text{in } \Omega \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{1}$$

For $u \in C^2(\Omega)$, $D^2u = ((\partial^2 u)/(\partial x_i \partial x_j))_{i,j=1,\ldots,2}$ denotes the Hessian matrix of u and $\det D^2u$ denotes its determinant. The function f defined on Ω is assumed to satisfy $f \geq c_0 > 0$ for a constant $c_0 > 0$ and we assume that $g \in C(\partial\Omega)$ can be extended to a function $\tilde{g} \in C(\overline{\Omega})$ which is convex in Ω.

G. Awanou (✉)
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

We consider a mixed formulation with unknowns the scalar variable u and the Hessian D^2u. The scalar variable and the components of the Hessian are approximated by Lagrange elements of degree $k \geq 2$. The method considered in this paper was analyzed from different point of views in [11] and [5] for smooth solutions of (1). In both [11] and [5] the convergence of the method for Lagrange elements of degree $k = 1$ and $k = 2$ was left unresolved. In this paper we resolve this issue for quadratic elements.

The ingredients of our approach consist in a fixed point argument, which yields the convergence of a time marching method, a "rescaling argument", i.e. the solution of a rescaled version of the equation, and the continuity of the eigenvalues of a matrix as a function of its entries. This is the same approach we took in the case of the standard finite element discretization of the Monge–Ampère equation [4].

With the mixed methods, as implemented in [10,11], one can apply directly Newton's method to the discrete nonlinear problem and still have numerical evidence of convergence to a larger class of non smooth solutions than what is possible with the standard finite element discretization. We refer to [10,11] for the numerical results. Moreover with the standard finite element discretization [4], convexity must be enforced weakly through appropriate iterative methods. Although the number of unknowns in the mixed methods is higher, in [10,11] the discrete Hessian was eliminated from the discrete equations in the implementation.

However, as observed in [5], this prevents numerical convergence for smooth solutions when linear elements are used to approximate all the unknowns. It is well known that mixed methods for second order linear elliptic equations lead to saddle point problems and that when the (1, 1) block is nonsingular a Schur complement can be used to reduce the size of the linear systems to solve. The mixed methods discussed in this paper were implemented using a Schur complement in [10,11]. It is a distinguished feature of Monge–Ampère type equations that how one solves the discrete nonlinear system equations can lead to dramatically different results for non smooth solutions. Given the degeneracy of saddle point problems, for non smooth solutions, the method analyzed in this paper should be implemented as in [10,11]. Thus we do not reproduce the numerical results in this paper. The reader should note carefully that for the point of view of analysis of the methods, one can use either forms. In this paper, it is the fixed point mapping associated to the reduced equation which will prove useful, c.f. Lemma 6. This approach can be related to the one taken in [11].

As pointed out in [5,10,11] the mixed method discussed in this paper is the formal limit of the vanishing moment methodology in mixed form [9]. In fact, as pointed out in [3], the approach in [9] may be viewed as an iterative method for solving the nonlinear system resulting from the discretization of (1) by a mixed finite element method. As usual in the analysis of nonlinear finite element problems, we use a fixed point argument coupled with a linearization. The argument used in this paper is similar to the one used in [9] but with several key differences. The fixed point argument used in this paper leads to the proof of convergence of an iterative method for quadratic and higher order elements. We use the continuity of the eigenvalues of a matrix as a function of its entries, a tool which allowed us to solve in [3] the open problem on the convexity of the solution obtained with the vanishing moment methodology. The continuity of the eigenvalues coupled with the use of a rescaling argument allow us in this paper to give error estimates for quadratic elements. The lowest order elements allowed in [9] are cubic elements.

Although it is the fixed point mapping associated to the reduced equation which is the key tool used, we also introduce the fixed point mapping associated with the mixed formulation. This is done to highlight the similarity between the analysis in [9], [5] and this paper.

We note that in [11] a stabilized method was proposed which works numerically for non smooth solutions in two dimension. It consists in using piecewise constants for the discrete Hessian and linear elements for the scalar variable. The analysis for smooth solutions of the lowest order methods discussed in [5,11] cannot be done with the approach of this paper. The techniques used in this paper generalize to the three-dimensional problem but only for $k \geq 3$. It should be possible to extend the approach taken in this paper to the formulation where discontinuous elements are used to approximate the unknowns [11]. Numerical results reported in [11] indicate the latter approach could lead to a less accurate approximation of the Hessian. For simplicity, and to focus on the methodology we present, we do not consider such an extension in this paper.

We organize the paper as follows. In the Sect. 2 we introduce some notation and preliminaries. The error analysis of the mixed method is done in Sect. 3.

2 Notation and preliminaries

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 $H_0^1(\Omega)$ is the subset of $H^1(\Omega)$ of elements with vanishing trace on $\partial\Omega$. We also 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^2 the space of vector fields with components in X and by $X^{2\times 2}$ the space of matrix fields with each component in X.

The norm in X is denoted by $||.||_X$ and we omit the subscript Ω and superscripts 2 and 2×2 when it is clear from the context. The inner product in $L^2(\Omega)$, $L^2(\Omega)^2$, and $L^2(\Omega)^{2\times 2}$ is denoted by $(,)$ and we use $\langle , \rangle$ for the inner product on $L^2(\partial\Omega)$ and $L^2(\partial\Omega)^2$. For inner products on subsets of Ω, we will simply append the subset notation.

We denote by n the unit outward normal vector to $\partial\Omega$. We recall that for a matrix A, A_{ij} denote its entries and the cofactor matrix of A, denoted cof A, is the matrix with entries $(\text{cof } A)_{ij} = (-1)^{i+j} \det(A)_i^j$ where $\det(A)_i^j$ is the determinant of the matrix obtained from A by deleting its ith row and its jth column. For two matrices $A = (A_{ij})$ and $B = (B_{ij})$, $A : B = \sum_{i,j=1}^2 A_{ij} B_{ij}$ denotes their Frobenius inner product. A quantity which is constant is simply denoted by C.

For a scalar function v we denote by Dv its gradient vector and recall that $D^2 v$ denotes the Hessian matrix of second order derivatives. The divergence of a matrix field is understood as the vector obtained by taking the divergence of each row.

In this section and Sect. 3 we assume that (1) has a solution which is sufficiently smooth. Put $\sigma = D^2u$. Then the unique convex solution $u \in H^3(\Omega)$ of (1) satisfies the following mixed problem: Find $(u, \sigma) \in H^2(\Omega) \times H^1(\Omega)^{2\times 2}$ such that

$$\begin{aligned} &(\sigma, \tau) + (\operatorname{div} \tau, Du) - \langle Du, \tau n\rangle = 0, && \forall \tau \in H^1(\Omega)^{2\times 2} \\ &(\det \sigma, v) = (f, v), && \forall v \in H_0^1(\Omega) \\ &u = g && \text{on } \partial\Omega. \end{aligned} \tag{2}$$

It is proved in [5] that the above variational problem is well defined.

2.1 Discrete variational problem

We denote by $\mathcal{T}_h$ a triangulation of Ω into simplices K and assume that $\mathcal{T}_h$ is quasi-uniform. We denote by V_h the standard Lagrange finite element space of degree $k \geq 2$ and denote by Σ_h the space of symmetric matrix fields with components in the Lagrange finite element space of degree $k \geq 2$. Let I_h denote the standard Lagrange interpolation operator from $H^s(\Omega), s \geq k + 1$ into the space V_h. We use as well the notation I_h for the matrix version of the Lagrange interpolation operator mapping $H^s(\Omega)^{2\times 2}$, for $s \geq k+1$, into Σ_h. We consider the problem: find $(u_h, \sigma_h) \in V_h \times \Sigma_h$ such that

$$\begin{aligned} &(\sigma_h, \tau) + (\operatorname{div} \tau, Du_h) - \langle Du_h, \tau n\rangle &&= 0, \quad \forall \tau \in \Sigma_h \\ &(\det \sigma_h, v) &&= (f, v), \quad \forall v \in V_h \cap H_0^1(\Omega) \\ &u_h &&= g_h \text{ on } \partial\Omega, \end{aligned} \tag{3}$$

where $g_h = I_h \tilde{g}$. It follows from the analysis in [5,11] that (3) is well-posed for $k \geq 3$ and error estimates were given. In Sect. 3 we give an error analysis valid for $k \geq 2$.

For $v_h \in V_h$, we will make the abuse of notation of using $D^2 v_h$ to denote the Hessian of v_h computed element by element. We will need the broken Sobolev norm

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

2.2 Properties of the Lagrange finite element spaces

We recall some properties of the Lagrange finite element space of degree $k \geq 1$ that will be used in this paper. They can be found in [6,8]. We have

Interpolation error estimates.

$$\begin{aligned} &||v - I_h v||_{H^j} \leq Ch^{k+1-j} ||v||_{H^{k+1}}, \quad \forall v \in H^s(\Omega), \ j = 0, 1, \\ &||v - I_h v||_{L^\infty} \leq Ch^k |v|_{H^{k+1}}, \quad \forall v \in H^s(\Omega). \end{aligned} \tag{4}$$

Inverse inequalities

$$||v||_{L^\infty} \leq Ch^{-1}||v||_{L^2}, \quad \forall v \in V_h \tag{5}$$

$$||v||_{H^1} \leq Ch^{-1}||v||_{L^2}, \quad \forall v \in V_h \tag{6}$$

$$||v||_{H^{k+1}(\mathcal{T}_h)} \leq Ch^{-k-1}||v||_{L^2}, \quad \forall v \in V_h. \tag{7}$$

Scaled trace inequality

$$||v||_{L^2(\partial\Omega)} \leq Ch^{-\frac{1}{2}}||v||_{L^2}, \quad \forall v \in V_h. \tag{8}$$

2.3 Algebra with matrix fields

We collect in the following lemma some properties of matrix fields, the proof of which can be found in [1,5].

Lemma 1 *For $K \in \mathcal{T}_h$ and $u, v \in C^2(K)$ we have*

$$\det D^2u - \det D^2v = \operatorname{cof}(tD^2u + (1-t)D^2v) : (D^2u - D^2v), \tag{9}$$

for some $t \in [0, 1]$. It can be shown that $t = 1/2$, [7].

For two 2×2 matrix fields η and τ

$$|| \operatorname{cof}(\eta) : \tau||_{L^2} \leq C||\eta||_{L^\infty}||\tau||_{L^2}, \tag{10}$$

$$\operatorname{cof}(\eta) - \operatorname{cof}(\tau) = \operatorname{cof}(\eta - \tau). \tag{11}$$

2.4 Continuity of the eigenvalues of a matrix as a function of its entries

Let $\lambda_1(A)$ and $\lambda_2(A)$ denote the smallest and largest eigenvalues of the symmetric matrix A. We have

Lemma 2 ([4, Lemma 3.1]) *There exists constants $m, M > 0$ independent of h and a constant $C_{conv} > 0$ independent of h such that for all $v_h \in V_h$ with $v_h = g_h$ on $\partial\Omega$ and*

$$||v_h - I_h u||_{H^1} < C_{conv}h^2,$$

we have

$$m \leq \lambda_1(\operatorname{cof} D^2v_h(x)) \leq \lambda_2(\operatorname{cof} D^2v_h(x)) \leq M, \quad \forall x \in K, K \in \mathcal{T}_h.$$

The following lemma was used implicitly in [1,2,4].

Lemma 3 *Assume $0 < \alpha < 1$ and $\alpha \leq (m + M)/(2m)$ for constants $m, M > 0$. Let B be a symmetric matrix field such that*

$$0 < m\alpha \leq \lambda_1(B(x)) \leq \lambda_2(B(x)) \leq M\alpha, \quad \forall x \in \Omega.$$

Then for $\nu = (m + M)/2$

$$\gamma \equiv \sup_{\substack{v,w \in V_h \\ |v|_{H^1}=1, |w|_{H^1}=1}} \left| (Dv, Dw) - \frac{1}{\nu}(BDv, Dw) \right|,$$

satisfies $0 < \gamma < 1$.

Proof Since $\lambda_1(B)$ and $\lambda_2(B)$ are the minimum and maximum respectively of the Rayleigh quotient $((Bz) \cdot z)/||z||^2$, where $||z||$ denotes the Euclidean norm of $\mathbb{R}^2$, we have for $x \in \Omega$

$$m\alpha ||z||^2 \leq (B(x)z) \cdot z \leq M\alpha ||z||^2, \quad z \in \mathbb{R}^2.$$

This implies

$$\alpha |w|^2_{H^1} \leq \int_\Omega [B(x)Dw(x)] \cdot Dw(x)\ dx \leq M\alpha |w|^2_{H^1}, \quad w \in V_h.$$

If we assume in addition that $|w|_{H^1} = 1$, we get

$$m\alpha \leq \int_\Omega [B(x)Dw(x)] \cdot Dw(x)\ dx \leq M\alpha, \quad w \in V_h.$$

It follows that

$$\left(1 - \frac{M\alpha}{\nu}\right) \leq \int_\Omega [I - \frac{1}{\nu}B(x)Dw(x)] \cdot Dw(x)\ dx \leq \left(1 - \frac{m\alpha}{\nu}\right), \quad w \in V_h.$$

Since $\nu = (m + M)/2$, we have

$$\begin{aligned} 1 - \frac{\alpha M}{\nu} &= \frac{m + M - 2M\alpha}{m + M} < 1 \\ 1 - \frac{\alpha m}{\nu} &= \frac{m + M - 2m\alpha}{m + M} < 1. \end{aligned}$$

If we define

$$\beta \equiv \sup_{v \in V_h, |v|_{H^1}=1} \left| (Dv, Dv) - \frac{1}{\nu}(BDv, Dv) \right|,$$

by the assumptions on α, we have

$$0 < \beta < 1.$$

We can define a bilinear form on V_h by the formula

$$(p, q) = \int_{\Omega} \left[\left(I - \frac{1}{\nu}(B(x) \right) Dp(x) \right] \cdot Dq(x)\, dx.$$

Then because

$$(p, q) = \frac{1}{4}((p+q, p+q) - (p-q, p-q)),$$

and using the definition of β, we get assuming that $|p|_{H^1} = |q|_{H^1} = 1$,

$$\begin{aligned} |(p, q)| &\leq \frac{\beta}{4}(p+q, p+q) + \frac{\beta}{4}(p-q, p-q) \\ &\leq \frac{\beta}{4}|p+q|^2_{H^1} + \frac{\beta}{4}|p-q|^2_{H^1} = \beta. \end{aligned}$$

This completes the proof. □

3 Error analysis of the mixed method for smooth solutions

We will assume without loss of generality that $h \leq 1$. The goal of this section is to prove the local solvability of (3) for Lagrange elements of degree $k \geq 2$. We define for $\rho > 0$,

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

We are interested in elements $(w_h, \eta_h) \in V_h \times \Sigma_h$ satisfying

$$(\eta_h, \tau) + (\operatorname{div} \tau, Dw_h) - \langle Dw_h, \tau n \rangle = 0, \quad \forall \tau \in \Sigma_h. \tag{12}$$

We define

$$Z_h = \{(w_h, \eta_h) \in V_h \times \Sigma_h, w_h = g_h \text{ on } \partial\Omega, (w_h, \eta_h) \text{ solves (12)}\} \text{ and}$$

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

In [5] the local solvability of (3) was obtained by a fixed point argument which consists in a linearization at the exact solution of (1). To be able to obtain results for quadratic elements we use a time marching method combined with a rescaling argument. This is the point of view we took in [2,4]. We first describe the time marching method at the continuous level.

Let $\nu > 0$. We consider the sequence of problems

$$\begin{aligned} -\nu \Delta u^{r+1} &= -\nu \Delta u^r + \det D^2 u^r - f \quad \text{in} \quad \Omega \\ u^{r+1} &= g \quad \text{on} \quad \partial\Omega. \end{aligned}$$

Put $\sigma^{r+1} = D^2 u^{r+1}$. We obtain the equivalent problems

$$\begin{aligned} \sigma^{r+1} &= D^2 u^{r+1} \quad \text{in } \Omega \\ -\nu \operatorname{tr} \sigma^{r+1} &= -\nu \operatorname{tr} \sigma^r + \det \sigma^r - f, \quad \text{in } \Omega \\ u^{r+1} &= g \quad \text{on } \partial\Omega, \end{aligned}$$

where tr A denotes the trace of the matrix A.

We are thus lead to consider the sequence of discrete problems: find $(u_h^{r+1}, \sigma_h^{r+1}) \in V_h \times \Sigma_h$ such that $u_h^{r+1} = g_h$ on $\partial\Omega$ and

$$(\sigma_h^{r+1}, \tau) + (\operatorname{div} \tau, D u_h^{r+1}) - \langle D u_h^{r+1}, \tau n \rangle = 0, \quad \forall \tau \in \Sigma_h \tag{13}$$

$$-\nu (\operatorname{tr} \sigma^{r+1}, v) = -\nu (\operatorname{tr} \sigma^m, v) + (\det \sigma_h^r - f, v), \quad \forall v \in V_h \cap H_0^1(\Omega), \tag{14}$$

given an initial guess (u_h^0, σ_h^0). We prove below the convergence of $(u_h^{r+1}, \sigma_h^{r+1})$ to a local solution (u_h, σ_h) of the discrete problem (3).

Let $\alpha > 0$. We 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

$$\begin{aligned} (\eta_h - T_2(w_h, \eta_h), \tau) &+ (\operatorname{div} \tau, D(w_h - T_1(w_h, \eta_h))) \\ &- \langle D(w_h - T_1(w_h, \eta_h)), \tau n \rangle = (\eta_h, \tau) \\ &+ (\operatorname{div} \tau, D w_h) - \langle D w_h, \tau n \rangle, \quad \forall\, \tau \in \Sigma_h \end{aligned} \tag{15}$$

$$-\nu (\operatorname{tr} T_2(w_h, \eta_h), v) = -\nu (\operatorname{tr} \eta_h, v) + (\det \eta_h - \alpha^2 f, v), \quad \forall\, v \in V_h \cap H_0^1(\Omega) \tag{16}$$

$$T_1(w_h, \eta_h) = w_h \quad \text{on } \partial\Omega. \tag{17}$$

Note that (15) is equivalent to

$$(T_2(w_h, \eta_h), \tau) + (\operatorname{div} \tau, D T_1(w_h, \eta_h)) - \langle D T_1(w_h, \eta_h), \tau n \rangle = 0 \; \forall\, \tau \in \Sigma_h. \tag{18}$$

Let I denote the 2×2 identity matrix. We first make the following important observation.

For $v \in V_h \cap H_0^1(\Omega)$ and $\tau = vI$, we have $\operatorname{div} \tau = Dv$ and since $v = 0$ on $\partial\Omega$, we have in addition $\tau n = 0$ on $\partial\Omega$. Thus using (18) we obtain

$$-\nu(\operatorname{tr} T_2(w_h, \eta_h), v) = -\nu(T_2(w_h, \eta_h), vI) = \nu(DT_1(w_h, \eta_h), Dv). \tag{19}$$

Similarly, we obtain that if (w_h, η_h) solves (12), then

$$(\operatorname{tr} \eta_h, v) = -(Dw_h, Dv), \quad \forall v \in V_h \cap H_0^1(\Omega). \tag{20}$$

Lemma 4 *The mapping T is well defined and if $(\alpha w_h, \alpha\eta_h)$ is a fixed point of* (15)–(17) *with $w_h = g_h$ on $\partial\Omega$, then (w_h, η_h) solves the nonlinear problem* (3).

Proof To prove the first assertion, it is enough to prove that if $(w_h, \eta_h) \in V_h \times \Sigma_h$ is such that $w_h = 0$ on $\partial\Omega$ and

$$\begin{aligned}(\eta_h, \tau) + (\operatorname{div} \tau, Dw_h) - \langle Dw_h, \tau n\rangle = 0, &\quad \forall \tau \in \Sigma_h\\ -\nu(\operatorname{tr} \eta_h, v) = 0, &\quad \forall v \in V_h \cap H_0^1(\Omega),\end{aligned}$$

then $w_h = 0$ and $\eta_h = 0$.

Using (20), we obtain $0 = -(\operatorname{tr} \eta_h, v) = (Dw_h, Dv)$, for all $v \in V_h \cap H_0^1(\Omega)$. Thus $|w_h|_{H^1}^2 = 0$. This proves that $w_h = 0$ by Poincaré's inequality. Using $\tau = \eta_h$ we obtain as well $\eta_h = 0$.

The proof of the second assertion is immediate. □

We recall from [5, Remark 3.6], see also [10,11], that for $v_h \in V_h$, there exists a unique $\eta_h \in \Sigma_h$ denoted $H(v_h)$, such that

$$(H(v_h), \tau) + (\operatorname{div} \tau, Dv_h) - \langle Dv_h, \tau n\rangle = 0, \quad \forall \tau \in \Sigma_h, \tag{21}$$

holds. To see this consider the problem: find $\eta_h \in \Sigma_h$ such that

$$(\eta_h, \tau) = -(\operatorname{div} \tau, Dv_h) + \langle Dv_h, \tau n\rangle, \quad \forall \tau \in \Sigma_h. \tag{22}$$

For $\tau \in \Sigma_h$, we define $F(\tau) = -(\operatorname{div} \tau, Dv_h) + \langle Dv_h, \tau n\rangle$. Clearly F is linear. By the Schwarz inequality, (6) and (8)

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

Thus a unique solution $\eta_h = H(v_h)$ exists by the Lax-Milgram Lemma.

Remark 1 From the definition of $H(v_h)$ (21) and (22), we have for $v_h \in V_h$,

$$H(\alpha v_h) = \alpha H(v_h).$$

Lemma 5 *Let $v_h \in V_h$ such that $||v_h - I_h u||_{H^1} \leq \mu$. Then*

$$||H(v_h) - I_h \sigma||_{L^2} \leq Ch^{-1}\mu + Ch^{k-1}.$$

Proof For $\tau \in \Sigma_h$, by (2) and (21) we have

$$\begin{aligned}(H(v_h) - I_h\sigma, \tau) &= (H(v_h) - \sigma, \tau) + (\sigma - I_h\sigma, \tau)\\ &= (\sigma - I_h\sigma, \tau) - (\operatorname{div}\tau, D(v_h - u)) + \langle D(v_h - u), \tau n\rangle\\ &= (\sigma - I_h\sigma, \tau) - (\operatorname{div}\tau, D(v_h - I_h u)) + \langle D(v_h - I_h u), \tau n\rangle\\ &\quad - (\operatorname{div}\tau, D(I_h u - u)) + \langle D(I_h u - u), \tau n\rangle.\end{aligned}$$

Let $\tau = H(v_h) - I_h\sigma$. By the Schwarz inequality, (6) and (8)

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

Therefore

$$\begin{aligned}\|\tau\|_{L^2} &\leq Ch^{k+1} + Ch^{-1}\mu + Ch^{k-1} + Ch^{k-\frac{1}{2}}\\ &\leq Ch^{-1}\mu + Ch^{k-1}.\end{aligned}$$

This proves the result. □

It follows from Lemma 5, with $\mu = 0$, that $(I_h u, H(I_h u)) \in B_h(\rho)$, i.e. the ball $B_h(\rho) \neq \emptyset$ for $\rho = C_0 h^k$ for a constant $C_0 > 0$. See also [5, Lemma 3.5]. As a consequence, see also [11],

$$||H(I_h u) - I_h\sigma||_{L^2} \leq C_0 h^{k-1}. \tag{23}$$

Let

$$\tilde{B}_h(\rho) = \{v_h \in V_h, v_h = g_h \text{ on } \partial\Omega,\ ||v_h - I_h u||_{H^1} \leq \rho\},$$

and consider the mapping

$$\tilde{T}_1 : V_h \to V_h,\ \text{defined by } \tilde{T}_1(v_h) = T_1(v_h, H(v_h)).$$

The motivation to introduce a discrete Hessian $H(v_h)$ in this paper, as opposed to the approach in [5], is given by Lemma 6 below.

Lemma 6 *If w_h is a fixed point of $\tilde{T}_1$, then $(w_h, H(w_h))$ is a fixed point of T and equivalently, if (w_h, η_h) is a fixed point of T, then w_h is a fixed point of $\tilde{T}_1$.*

Proof The result was given as [5, Remark 3.6]. Let w_h be a fixed point of $\tilde{T}_1$. We have $T_1(w_h, H(w_h)) = w_h$ and by (18) and (21), $T_2(w_h, H(w_h)) = H(T_1(w_h, H(w_h))) = H(w_h)$. This proves that $(w_h, H(w_h))$ is a fixed point of T.

Conversely if (w_h, η_h) is a fixed point of T, then $\tilde{T}_1(w_h) = T_1(w_h, H(w_h)) = T_1(w_h, \eta_h) = w_h$. This completes the proof. □

Lemma 7 *We have for* $0 \leq \alpha \leq 1$

$$||\alpha I_h u - T_1(\alpha I_h u, H(\alpha I_h u))||_{H^1} \leq \frac{C_1}{\nu}\alpha^2 h^{k-1}, \tag{24}$$

for a positive constant C_1.

Proof Since $T_1(\alpha I_h u, H(\alpha I_h u)) - \alpha I_h u = 0$ on $\partial\Omega$, by (19) and (16) we have using $w_h = \alpha I_h u$, $\eta_h = H(\alpha I_h u)$ and $v = T_1(w_h, \eta_h) - w_h$

$$\nu(DT_1(w_h, \eta_h), Dv) = -\nu(\operatorname{tr} T_2(w_h, \eta_h), v) = -\nu(\operatorname{tr} \eta_h, v) + (\det \eta_h - \alpha^2 f, v).$$

It follows that

$$\nu|Dv|^2_{L^2} = -\nu(Dw_h, Dv) - \nu(\operatorname{tr} \eta_h, v) + (\det \eta_h - \alpha^2 f, v).$$

Therefore, using (20), we get

$$\nu|Dv|^2_{L^2} = (\det \eta_h - \alpha^2 f, v). \tag{25}$$

On the other hand since $f = \det D^2 u = \det \sigma$, by (9) and Remark 1, on each element K

$$\begin{aligned}\det \eta_h - \alpha^2 f &= \det H(\alpha I_h u) - \alpha^2 \det \sigma = \det \alpha H(I_h u) - \alpha^2 \det \sigma \\ &= \alpha^2(\det H(I_h u) - \det \sigma) \\ &= \alpha^2(\operatorname{cof}(tH(I_h u) + (1-t)\sigma) : (H(I_h u) - \sigma)),\end{aligned} \tag{26}$$

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

By (4) we have $\|I_h \sigma\|_{L^\infty} \leq C\|\sigma\|_{L^\infty}$. Thus by (23) and (5)

$$\begin{aligned}||H(I_h u)||_{L^\infty} &\leq ||H(I_h u) - I_h\sigma||_{L^\infty} + \|I_h\sigma\|_{L^\infty} \leq Ch^{-1}||H(I_h u) - I_h\sigma||_{L^2} \\ &\quad + \|I_h\sigma\|_{L^\infty} \leq Ch^{k-2} + C\|\sigma\|_{L^\infty} \leq C, \quad \text{since } k \geq 2.\end{aligned}$$

Thus by (10) and (23)

$$\begin{aligned}\| \det(H(I_h u)) - \det\sigma \|_{L^2(K)} &\leq C\|tH(I_h u) + (1-t)\sigma\|_{L^\infty(K)} \|H(I_h u) - \sigma\|_{L^2(K)} \\ &\leq C\|H(I_h u) - \sigma\|_{L^2(K)} \\ &\leq C\|H(I_h u) - I_h\sigma\|_{L^2(K)} + C\|I_h\sigma - \sigma\|_{L^2(K)} \\ &\leq Ch^{k-1}.\end{aligned}$$

Therefore by (4) and (26)

$$\| \det \eta_h - \alpha^2 f\|_{L^2} = \alpha^2 \| \det(H(I_h u)) - \det\sigma\|_{L^2} \leq C\alpha^2 h^{k-1}. \tag{27}$$

And so combining (25)–(27), (23), Cauchy–Schwarz inequality, the interpolation error estimate (4) and Poincare's inequality, we get

$$|v|^2_{H^1} \leq \frac{C}{\nu}\alpha^2 h^{k-1} ||v||_{L^2} \leq \frac{C}{\nu}\alpha^2 h^{k-1}||v||_{H^1},$$

from which (24) follows. □

We will need the following lemma

Lemma 8 *Let $(w_h, \eta_h) \in Z_h$. Then for a piecewise smooth symmetric matrix field P*

$$((\text{cof } P) : \eta_h, v) + ((\text{cof } P) D w_h, Dv) \leq Ch||v||_{H^1}||w_h||_{H^1}, \tag{28}$$

for all $v \in V_h \cap H^1_0(\Omega)$ and for a constant C which depends on $|| \text{cof } P||_{H^{k+1}(\mathcal{T}_h)}$.

Proof The proof is the same as the proof of [5, Lemma 3.7]. There the proof was given for $P = D^2 u$, but it carries over to the general case of this lemma line by line. The dependence of the constant C on $|| \text{cof } P||_{H^{k+1}(\mathcal{T}_h)}$ arises from the use in the proof of the approximation property $||P_{\Sigma_h}(v \text{ cof } P) - v \text{ cof } P||_{H^m(\mathcal{T}_h)} \leq Ch^{k+1-m}||v \text{ cof } P||_{H^{k+1}(\mathcal{T}_h)}$. Here P_{Σ_h} denotes the L^2 projection operator into Σ_h. □

Lemma 9 *For $(w_h, \eta_h) \in B_h(\rho)$, $\rho = C_0 h^k$, we have*

$$||\eta_h - D^2 w_h||_{L^\infty} \leq Ch^{k-2}.$$

Proof Recall that for $(w_h, \eta_h) \in B_h(\rho)$, we have $\eta_h = H(w_h)$. We have by (5), (23)

$$\begin{aligned}||\eta_h - D^2 w_h||_{L^\infty} &\leq ||H(w_h) - D^2 w_h||_{L^\infty} \\ &\leq ||H(w_h) - I_h\sigma||_{L^\infty} + ||I_h\sigma - D^2 w_h||_{L^\infty} \\ &\leq Ch^{-1}||H(w_h) - I_h\sigma||_{L^2} + ||I_h\sigma - D^2 u||_{L^\infty} + ||D^2 u - D^2 w_h||_{L^\infty} \\ &\leq Ch^{k-2} + Ch^{k+1} + ||D^2 u - D^2 I_h u||_{L^\infty} + ||D^2 I_h u - D^2 w_h||_{L^\infty} \\ &\leq Ch^{k-2} + Ch^{-1}||I_h u - w_h||_{H^1} \\ &\leq Ch^{k-2}.\end{aligned}$$

□

The next lemma states a crucial contraction property of the mapping T_1 in $\alpha B_h(\rho)$.

Lemma 10 *Let* $(w_1, \eta_1), (w_2, \eta_2) \in B_h(\rho)$ *with* $\rho \leq \min(C_0, C_{conv})h^k$. *We have*

$$|T_1(\alpha w_1, \alpha\eta_1) - T_1(\alpha w_2, \alpha\eta_2)|_{H^1} \leq a|\alpha w_1 - \alpha w_2|_{H^1}, \tag{29}$$

for $0 < a < 1$, *h sufficiently small,* $\alpha = h^{k+2}$ *and* $\nu = (m + M)/2$.

Proof Put $v = T_1(\alpha w_1, \alpha\eta_1) - T_1(\alpha w_2, \alpha\eta_2)$. By assumption $v \in V_h \cap H_0^1(\Omega)$. Using (19) and (16) we obtain

$$\begin{aligned}
&\nu(DT_1(\alpha w_1, \alpha\eta_1) - DT_1(\alpha w_2, \alpha\eta_2), Dv)\\
&\quad = -\nu(\operatorname{tr} T_2(\alpha w_1, \alpha\eta_1) - \operatorname{tr} T_2(\alpha w_2, \alpha\eta_2), v)\\
&\quad = -\nu(\operatorname{tr} \alpha\eta_1 - \operatorname{tr} \alpha\eta_2, v) + (\det \alpha\eta_1 - \det \alpha\eta_2, v).
\end{aligned}$$

Therefore, using (9), we have for some $t \in [0, 1]$ and with the notation

$$Q = t\eta_1 + (1 - t)\eta_2 \text{ and } \overline{Q} = tD^2 w_1 + (1 - t)D^2 w_2,$$

$$\begin{aligned}
|v|^2_{H^1} &= -(\operatorname{tr} \alpha\eta_1 - \operatorname{tr} \alpha\eta_2, v)\\
&\qquad + \tfrac{1}{\nu}((\operatorname{cof} \alpha(t\eta_1 + (1 - t)\eta_2)) : \alpha(\eta_1 - \eta_2), v)\\
&= \left(\left(-I + \tfrac{1}{\nu} \operatorname{cof} \alpha Q\right) : \alpha(\eta_1 - \eta_2), v\right)\\
&= -(I : \alpha(\eta_1 - \eta_2), v) - (D\alpha(w_1 - w_2), Dv)\\
&\quad + \tfrac{1}{\nu}((\operatorname{cof} \alpha Q) : \alpha(\eta_1 - \eta_2), v) + \tfrac{1}{\nu}((\operatorname{cof} \alpha Q)D\alpha(w_1 - w_2), Dv)\\
&\quad + (D\alpha(w_1 - w_2), Dv) - \tfrac{1}{\nu}((\operatorname{cof} \alpha \overline{Q})D\alpha(w_1 - w_2), Dv)\\
&\quad + \tfrac{1}{\nu}((\operatorname{cof} \alpha \overline{Q})D\alpha(w_1 - w_2), Dv) - \tfrac{1}{\nu}((\operatorname{cof} \alpha Q)D\alpha(w_1 - w_2), Dv).
\end{aligned} \tag{30}$$

For $(w_1, \eta_1), (w_2, \eta_2) \in B_h(\rho)$, $t(w_1, \eta_1) + (1 - t)(w_2, \eta_2) \in B_h(\rho)$ and thus for h sufficiently small, by Lemmas 2 and 3 we get

$$\left|(D(w_1 - w_2), Dv) - \frac{1}{\nu}((\operatorname{cof} \alpha \overline{Q})D(w_1 - w_2), Dv)\right| \leq \gamma |w_1 - w_2|_{H^1}|v|_{H^1}, \tag{31}$$

for $0 < \gamma < 1$.

On the other hand, by Lemma 8, with $P = I$, we have

$$|-(I : (\eta_1 - \eta_2), v) - (D(w_1 - w_2), Dv)| \leq Ch|w_1 - w_2|_{H^1}|v|_{H^1}. \tag{32}$$

Applying Lemma 8, with $P = Q$, we get

$$\begin{aligned}
|((\operatorname{cof} Q) : (\eta_1 - \eta_2), v) + ((\operatorname{cof} Q)D(w_1 - w_2), Dv)| &\leq Ch||\operatorname{cof} Q||_{H^{k+1}(\mathcal{T}_h)}\\
&\qquad |w_1 - w_2|_{H^1}|v|_{H^1}.
\end{aligned} \tag{33}$$

Finally, since by (11)

$$\operatorname{cof} Q - \operatorname{cof} \overline{Q} = \operatorname{cof}(Q - \overline{Q}) = \operatorname{cof}(t(\eta_1 - D^2 w_1) + (1-t)(\eta_2 - D^2 w_2)),$$

we get using Lemma 9

$$|| \operatorname{cof} Q - \operatorname{cof} \overline{Q}||_{L^\infty} \leq C h^{k-2} \leq C, \quad \text{since } k \geq 2.$$

Thus

$$\left| \tfrac{1}{\nu}((\operatorname{cof} \overline{Q}) D(w_1 - w_2), Dv) - \tfrac{1}{\nu}((\operatorname{cof} Q) D(w_1 - w_2), Dv) \right| \leq C|w_1 - w_2|_{H^1} |v|_{H^1}. \tag{34}$$

We conclude from (30)–(34) that

$$|v|_{H^1} \leq (\gamma + Ch + C\alpha h || \operatorname{cof} Q||_{H^{k+1}(\mathcal{T}_h)} + C\alpha)|\alpha w_1 - \alpha w_2|_{H^1}. \tag{35}$$

Using the inverse estimate (7) and noting that $\rho \leq h^2$

$$\begin{aligned}
|| \operatorname{cof} Q||_{H^{k+1}(\mathcal{T}_h)} &\leq C h^{-k-1} || \operatorname{cof} Q||_{L^2} \leq C h^{-k-1} ||Q||_{L^2} \\
&\leq C h^{-k-1} ||t\eta_1 + (1-t)\eta_2||_{L^2} \\
&\leq C h^{-k-1} (||\eta_1||_{L^2} + ||\eta_2||_{L^2}) \\
&\leq C h^{-k-1} (||\eta_1 - I_h\sigma||_{L^2} + ||\eta_2 - I_h\sigma||_{L^2} + 2||I_h\sigma||_{L^2}) \\
&\leq C h^{-k-1} (h^{-1}\rho + ||\sigma||_{L^2}) \leq C h^{-k-1} (Ch + ||\sigma||_{L^2}) \leq C h^{-k-1}.
\end{aligned}$$

Since $\gamma < 1$, and $\alpha = h^{k+2}$, for h sufficiently small, $Ch + C\alpha h || \operatorname{cof} Q||_{H^{k+1}(\mathcal{T}_h)} + C\alpha < 1 - \gamma$. We conclude from (35) that (29) holds. □

Lemma 11 *For $\rho = \min(C_0, C_{conv}) h^k$, the mapping $\tilde{T}_1$ has a unique fixed point in $\alpha \tilde{B}_h(\rho)$ for $\alpha = h^{k+2}$.*

Proof Note that by (29), $\tilde{T}_1$ is a strict contraction in $\alpha \tilde{B}_h(\rho)$ for $\rho \leq \min(C_0, C_{conv}) h^k$. We now show that $\tilde{T}_1$ maps $\alpha \tilde{B}_h(\rho)$ into itself. Let $v_h \in \tilde{B}_h(\rho)$. We have by (29) and (24)

$$\begin{aligned}
||\tilde{T}_1(\alpha v_h) - \alpha I_h u||_{H^1} &\leq ||\tilde{T}_1(\alpha v_h) - \tilde{T}_1(\alpha I_h u)||_{H^1} + ||\tilde{T}_1(\alpha I_h u) - \alpha I_h u||_{H^1} \\
&\leq a||\alpha v_h - \alpha I_h u||_{H^1} + C_1 \alpha^2 h^{k-1} \\
&\leq a\alpha\rho + C_1 \alpha h^{2k+1} = a\alpha\rho + C_1 h^{k+1} \alpha h^k.
\end{aligned}$$

Therefore for h sufficiently small, $C_1 h^{k+1} \leq \min(C_0, C_{conv})(1-a)$ and so

$$||\tilde{T}_1(\alpha v_h) - \alpha I_h u||_{H^1} \leq a\alpha\rho + (1-a)\alpha\rho.$$

The result then follows from the Banach fixed point theorem. □

We can now state the main result of this paper

Theorem 1 *Let* $(u,\sigma) \in H^{k+3}(\Omega) \times H^{k+1}(\Omega)^{2\times 2}$ *denote the unique convex solution of* (2)*. Problem* (3) *has a unique local solution* (u_h, σ_h) *for* $k \geq 2$ *and* h *sufficiently small. We have*

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

Proof Recall that for $(u_h, \sigma_h) \in B_h(\rho)$, we have $\sigma_h = H(u_h)$. The result follows from Lemmas 6, 11 and 4, the definition of $B_h(\rho)$ and (23).

The local solution u_h given by Lemma 11 satisfies $||u_h - I_h u||_{H^1} \leq Ch^k$. Since by Lemma 6, $(u_h, H(u_h))$ is a fixed point of T, by Lemma 4, $(u_h, H(u_h))$ solves (3). By the definition of $B_h(\rho)$ $\sigma_h = H(u_h)$ and by (23), we have $||\sigma_h - I_h \sigma||_{H^1} \leq Ch^{k-1}$. □

Remark 2 As with [4], the analysis of this paper extends to the three dimensional case when one assumes $k \geq 3$. In the quadratic case, in three dimension, one obtains that the solution u_h is much closer to the interpolant than what can be expected from the approximation properties of the finite element space. However, upon a rescaling of the equation, this difficulty disappears. The same argument applies to the standard finite element discretization [4].

Remark 3 If it is known that (3) has a solution (u_h, σ_h) with u_h piecewise strictly convex, the fixed point argument of this paper can be repeated with $\alpha = \nu h^{k+2}$ to prove that the solution is locally unique.

Acknowledgments The author was partially supported by a Division of Mathematical Sciences of the US National Science Foundation grant No. 1319640.

References

1. Awanou, G.: Pseudo transient continuation and time marching methods for Monge–Ampère type equations (2013). http://arxiv.org/abs/1301.5891
2. Awanou, G.: On standard finite difference discretizations of the elliptic Monge–Ampère equation (2014). http://arxiv.org/pdf/1311.2812v5.pdf
3. Awanou, G.: Spline element method for the Monge–Ampère equations (2014) (to appear in B.I.T. Numerical Mathematics)
4. Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: classical solutions (2014) (to appear in IMA J. Numer. Anal.)
5. Awanou, G., Li, H.: Error analysis of a mixed finite element method for the Monge–Ampère equation. Int. J. Numer. Anal. Model. **11**, 745–761 (2014)
6. Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. I. Math. Comput. **47**(175), 103–134 (1986)
7. Brenner, S.C., Gudi, T., Neilan, M., Sung, L.Y.: C^0 penalty methods for the fully nonlinear Monge–Ampère equation. Math. Comput. **80**(276), 1979–1995 (2011)
8. Brenner, S.C., Scott, L.R.: The mathematical theory of finite element methods. In: Texts in Applied Mathematics, vol. 15, 2nd edn. Springer-Verlag, New York (2002)
9. Feng, X., Neilan, M.: 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**(2), 1226–1250 (2009)

10. Lakkis, O., Pryer, T.: A finite element method for nonlinear elliptic problems. SIAM J. Sci. Comput. **35**(4), A2025–A2045 (2013)
11. Neilan, M.: Finite element methods for fully nonlinear second order PDEs based on a discrete Hessian with applications to the Monge–Ampère equation. J. Comput. Appl. Math. **263**, 351–369 (2014)