IMA Journal of Numerical Analysis Advance Access published May 30, 2014

IMA Journal of Numerical Analysis (2014) Page 1 of 17
doi:10.1093/imanum/dru028

Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: classical solutions

GERARD AWANOU
Department of Mathematics, Statistics, and Computer Science, M/C 249. University of Illinois at Chicago, Chicago, IL 60607-7045, USA
awanou@uic.edu

[Received on 6 November 2013; revised on 1 March 2014]

We propose a new variational formulation of the elliptic Monge–Ampère equation and show how classical Lagrange elements can be used for the numerical resolution of classical solutions of the equation. Error estimates are given for Lagrange elements of degree $d \geqslant n$ in dimensions $n = 2$ and $n = 3$. No jump term is used in the variational formulation. We propose to solve the discrete nonlinear system of equations by a time marching method, and numerical evidence is given which indicates that one approximates in two dimension a larger class of nonsmooth solutions than what is possible if one simply uses Newton's method.

Keywords: elliptic Monge–Ampère; smooth solutions; time marching; Lagrange elements.

1. Introduction

This paper addresses the numerical resolution of the Dirichlet problem for the Monge–Ampère equation

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

A classical solution of (1.1) is a convex function $u \in C(\bar{\Omega}) \cap C^2(\Omega)$ which satisfies (1.1). The domain $\Omega \subset \mathbb{R}^n, n = 2, 3,$ is assumed to be convex with (polygonal) boundary $\partial\Omega$. Here, D^2u denotes the Hessian of u with $(D^2u)_{i,j} = (\partial^2 u)/(\partial x_i \partial x_j), i, j = 1, \ldots, n,$ and f, g are given functions with $f \geqslant 0$ and $g \in C(\partial\Omega)$ with g convex on any line segment in $\partial\Omega$. A smooth solution of (1.1) solves the variational problem: find $u \in W^{2,\infty}(\Omega)$ such that $u = g$ on $\partial\Omega$ and

$$\int_\Omega (\det D^2 u) v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x \quad \text{for all } v \in H_0^1(\Omega).$$

We propose to solve numerically (1.1) with standard Lagrange finite element spaces V_h of degree $d \geqslant n$ by analysing the (nonconforming) variational problem: find $u_h \in V_h \subset H^1(\Omega)$ such that $u_h = g_h$ on $\partial\Omega$ for an interpolant g_h of g and

$$\sum_{K \in \mathscr{T}_h} \int_K (\det D^2 u_h) v_h \, \mathrm{d}x = \int_\Omega f v_h \, \mathrm{d}x \quad \forall v_h \in V_h \cap H_0^1(\Omega). \tag{1.2}$$

Here, $\mathscr{T}_h$ denotes a quasi-uniform, simplicial and conforming triangulation of the domain. Error estimates for smooth solutions are derived. We propose to solve the discrete nonlinear system of equations by a time marching method, cf. Theorem 3.3. Numerical evidence is given, which indicates that one approximates in two dimension a larger class of nonsmooth solutions than what is possible if one simply uses Newton's method.

Closely related to this paper are Brenner & Neilan (2012), Brenner *et al.* (2011) and Neilan (2013). Like the authors of these papers, we also use a fixed-point argument, but our approach is essentially different. No jump term is used in our variational formulation. We are able to give error estimates for Lagrange elements of degree $d \geqslant n$ with no smoothness assumption on the boundary. This is achieved by a rescaling argument and the effective use of the continuity of the eigenvalues of a matrix as a function of its entries. The fixed-point argument we use to establish the well-posedness of (1.2) also yields the theoretical convergence of the time marching iterative method.

The use of the standard Lagrange finite element spaces in connection with the numerical resolution of (1.1) also appears in mixed methods. A least squares formulation was used in Glowinski (2009) and recently a direct mixed formulation was presented in Lakkis & Pryer (2013). The latter is essentially the limiting case of the mixed method for the vanishing moment methodology; cf. Feng & Neilan (2009) and the references therein. The vanishing moment methodology is a singular perturbation approach to the Monge–Ampère equation with the perturbation a multiple of the bilaplacian. The convergence and error estimates for the methods introduced in Glowinski (2009) are still open problems and mixed methods typically lead to a large system of equations.

In view of having numerical results for nonsmooth solutions, it is natural to use a time marching method, and not Newton's method, for solving the discrete nonlinear system of equations. The numerical experiments indicate that the method may be valid for the so-called viscosity solutions. This is a fascinating and challenging issue and its resolution involves additional new ideas different from the techniques for error analysis used in this paper. We wish to address this issue in a separate work (Awanou, 2013d).

Our approach may be viewed as a variant of the method introduced in Brenner *et al.* (2011). As pointed out in Brenner *et al.* (2011), a numerical method based on Lagrange elements and the formulation (1.2) does not work in theory in the sense that it is difficult to use a fixed-point argument with a mapping defined through a second-order elliptic equation in divergence form with coefficient matrix the cofactor matrix of D^2u. The authors in Brenner *et al.* (2011) ingeniously added jump terms to facilitate the above approach. On the other hand, our numerical experiments indicate that (1.1) can be solved numerically if the discrete nonlinear system of equations is solved by a time marching method. An advantage of the time marching method is that the user only needs access to a Poisson solver to implement the scheme. The main advantage, however, is that one has numerical evidence of convergence for nonsmooth solutions in two dimension for a larger class of nonsmooth solutions than what is possible if one simply uses Newton's method. Our numerical experiments are for the case f real valued with $f > 0$ on Ω. Obviously, the time marching method can also be applied to the discretization proposed in Brenner *et al.* (2011), but we believe that, in the context of nonsmooth solutions, the jump terms in the discretization proposed there may not be necessary. We have chosen not to treat curved boundaries for simplicity and to focus on the main ideas. The main motivation to assume that the domain is smooth and strictly convex is to guarantee the existence of a smooth solution for smooth data. One then faces the difficulty of practically imposing Dirichlet boundary conditions, a problem solved in Brenner *et al.* (2011) by the use of the Nitsche method. Here instead, we will make the assumption ubiquitous in the finite element analysis of numerous problems that the solution is smooth on a polygonal domain.

We believe that the fixed-point argument used in this paper and/or the strategy of rescaling the Monge–Ampère equation would prove useful in resolving other outstanding issues about the numerical analysis of Monge–Ampère-type equations; see, for example, Awanou (2013a). For another example, our fixed point-rescaling argument provides an alternative to Neilan (2013) for the proof of the well-posedness of the discretization proposed in Brenner *et al.* (2011) for quadratic finite elements. Essentially, the rescaling argument is appropriate whenever an argument can be made that a result holds for the Monge–Ampère equation, provided the solution is sufficiently small. Thus, instead of describing the whole rescaling argument, one may simply prove results for the case when the exact solution is sufficiently small.

In fact, the results of this paper are similar to the ones announced in the context of C^1 conforming approximations in a technical report by the author Awanou (2013b), but the analysis is more involved. Exploiting that similarity, pseudo-transient continuation methods can be developed for (1.1) by taking appropriate nonconforming discretizations of the iterative methods proposed in Awanou (2013b). We do not pursue this line of investigation in this paper. The properties of the Lagrange finite element spaces used in our analysis, namely an approximation property and inverse estimates, also hold for certain C^1 conforming approximations. Thus, our error estimates hold for these as well. The error estimates hold for the following assumption on the exact solution: $u \in W^{3,\infty}(K)$ on each element K is strictly convex on each element and solves (1.2).

We organize the paper as follows. In Section 2, we give the notation used and recall some facts about determinants and Lagrange finite element spaces. The properties of the finite element spaces needed for our analysis are stated as well as the requirements on the exact solution. We prove the existence and uniqueness of the discrete problem (1.2) with the convergence of the time marching method in Section 3. In Section 4, we give the numerical results. We conclude with some remarks.

2. Notation and preliminaries

Let $\mathbb{P}_d$ denote the space of polynomials of degree $\leqslant d$. We use the usual notation $L^p(\Omega), 2 \leqslant p \leqslant \infty$ for the Lebesgue spaces and $W^{s,p}(\Omega), 1 \leqslant s < \infty$ for the Sobolev spaces of elements of $L^p(\Omega)$ with weak derivatives of order $\leqslant s$ in $L^p(\Omega)$. The norms and seminorms in $W^{s,p}(\Omega)$ are denoted by $\|.\|_{s,p}$ and $|.|_{s,p}$, respectively, and, when $p = 2$, we will simply use $\|.\|_s$ and $|.|_s$. Thus, the L^p-norm is denoted by $\|.\|_0$. We will use the simpler notation $\|.\|_\infty$ for the norm in $L^\infty(\Omega)$.

For a function defined on an element K or more generally on a subdomain S, we will add K or S to the norm and seminorm notation. We will need a broken Sobolev norm

$$\|v\|_{s,p,\mathscr{T}_h} = \left(\sum_{K \in \mathscr{T}_h} \|v\|_{s,p,K}^p \right)^{1/p},$$

with the above conventions for the case when $p = 2$.

For a matrix field A, we define $\|A\|_\infty = \max_{i,j=1,\dots,n} \|A_{ij}\|_\infty$. We denote by n the unit outward normal vector to $\partial\Omega$, and by n_K the unit outward normal vector to ∂K for an element K.

For two matrices $A = (A_{ij})$ and $B = (B_{ij})$, $A : B = \sum_{i,j=1}^n A_{ij} B_{ij}$ denotes their Frobenius inner product. The divergence of a matrix field is understood as the vector obtained by taking the divergence of each row. We use the notation Du to denote the gradient vector and, for a matrix A, $\operatorname{cof} A$ denotes the matrix of cofactors of A.

A quantity that is independent of h, but which may depend on s, p, Ω, etc., is simply denoted by C. Throughout the paper, for a discrete function v_h, the Hessian $D^2 v_h$ is always computed element by element. We will assume that $0 < h \leqslant 1$.

2.1 *Computations with determinants*

LEMMA 2.1 For $u, v \in C^2(K)$, we have

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

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

Proof. The result follows from the mean value theorem and the expression of the derivative of the mapping $F : u \to \det D^2u$. We have $F'(u)(v) = (\operatorname{cof} D^2u) : D^2v$. First, note that $\partial(\det A)/(\partial A_{ij}) = (\operatorname{cof} A)_{ij}$; see, for example, formula (23) p. 440 of Evans (1998). The result then follows from the chain rule. □

LEMMA 2.2 For $n = 2$ and $n = 3$, and two matrix fields η and τ

$$\| \operatorname{cof}(\eta) : \tau \|_0 \leqslant C \|\eta\|_\infty^{n-1} \|\tau\|_0, \tag{2.1}$$

$$\| \operatorname{cof}(\eta) - \operatorname{cof}(\tau) \|_0 \leqslant C(\|t\eta + (1-t)\tau\|_\infty)^{n-2} \|\eta - \tau\|_0. \tag{2.2}$$

Proof. The bound (2.1) is given by a direct computation. For $n = 2$, we have $\operatorname{cof}(\eta) - \operatorname{cof}(\tau) = \operatorname{cof}(\eta - \tau)$ from which the result follows. For $n = 3$, 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

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

for some $t \in [0,1]$. Direct computation then gives (2.2). □

2.2 *Assumptions on the approximation spaces*

For the discretization (1.2), one can use either the Lagrange finite element spaces or certain finite-dimensional spaces of C^1 functions. To make our results applicable to other types of discretizations, we formulate our assumptions on the approximation spaces.

ASSUMPTION 2.3 (Approximation property) The finite-dimensional space $V_h \subset H^1(\Omega)$ contains the Lagrange space of degree d

$$\{ v_h \in C^0(\bar{\Omega}),\ v_h|_K \in \mathbb{P}_d,\ \forall K \in \mathscr{T}_h \},$$

and there exists a linear interpolation operator I_h mapping $C^r(\bar{\Omega})$ for $r = 0$ or $r = 1$ into V_h and a constant C such that if w is in the Sobolev space $W^{l+1,p}(\Omega)$, $1 \leqslant p \leqslant \infty$, $0 \leqslant l \leqslant d$,

$$\|w - I_h w\|_{k,p,\mathscr{T}_h} \leqslant C_{ap} h^{l+1-k} |w|_{l+1,p}, \tag{2.3}$$

for $k = 0, 1, 2$.

The interpolant g_h used in (1.2) is taken as I_h applied to a continuous extension of g.

When V_h is the Lagrange finite element space, the interpolant I_h is taken as the standard interpolation operator defined from the degrees of freedom. It is then known that Assumption 2.3 holds (Brenner & Scott, 2002).

As a consequence of (2.3),

$$\|I_h w\|_{k,p,\mathscr{T}_h} \leqslant (1+C_{ap})\|w\|_{k,p}, \quad w \in W^{k,p}(\Omega), \quad k=0,1,2, \tag{2.4}$$

for all p.

The Monge–Ampère equation (1.1) involves D^2u and the time marching method involves solving Poisson equations. It turns out that, for the fixed-point argument, one needs to relate the H^1-norm of v_h for $v_h \in V_h$ to $\|v_h\|_{2,\infty,\mathscr{T}_h}$. We will need inverse estimates.

Assumption 2.4 Inverse estimates

$$\|w_h\|_{t,p,\mathscr{T}_h} \leqslant C_{\text{inv}} h^{s-t+\min(0,n/p-n/q)} \|w_h\|_{s,q,\mathscr{T}_h}, \tag{2.5}$$

for $0 \leqslant s \leqslant t, 1 \leqslant p,q \leqslant \infty$ and $w_h \in V_h$.

The inverse estimates hold for the Lagrange finite element spaces as a consequence of the quasi-uniformity assumption on the triangulation (Brenner & Scott, 2002).

2.3 *Assumptions on the exact solution*

Let $\lambda_1(A)$ and $\lambda_n(A)$ denote the smallest and largest eigenvalues of a symmetric matrix A. We make the following assumption on the exact solution.

Assumption 2.5 Local piecewise smooth and strict convexity assumption. The solution u of (1.1) is in $W^{3,\infty}(\mathscr{T}_h) \cap H^1(\Omega)$, strictly convex on each element K and for constants $m', M' > 0$, independent of h,

$$m' \leqslant \lambda_1(D^2u(x)) \leqslant \lambda_n(D^2u(x)) \leqslant M' \quad \forall x \in K, \quad K \in \mathscr{T}_h.$$

Moreover, we require the exact solution u to solve the problem: find $u \in W^{2,\infty}(\mathscr{T}_h)$, strictly convex on each element T, such that $u = g$ on $\partial\Omega$ and

$$\sum_{K \in \mathscr{T}_h} \int_K (\det D^2u) v \, dx = \int_\Omega f v \, dx \quad \forall v \in V_h \cap H_0^1(\Omega). \tag{2.6}$$

We note that Assumption 2.5 trivially holds for a strictly convex solution u in $C^3(\bar{\Omega})$. In that case $f \geqslant c_0 > 0$ for a constant c_0.

3. Well-posedness of the discrete problem and error estimates

The proof of all lemmas in this section are given at the end of the section.

We first state a fundamental observation about the behaviour of discrete functions near the interpolant $I_h u$.

Lemma 3.1 There exists $\delta > 0$ such that, for h sufficiently small and for all $v_h \in V_h$ with $\|v_h - I_h u\|_1 < \delta/(2C_{\text{inv}})h^{1+n/2}$, $D^2(v_h|_K)$ is positive definite with

$$\frac{m'}{2} \leqslant \lambda_1 D^2(v_h|_K) \leqslant \lambda_n D^2(v_h|_K) \leqslant \frac{3M'}{2},$$

where m' and M' are the constants of Assumption 2.5. Thus, $\operatorname{cof} D^2(v_h|_K)$ is invertible on each element T.

Let

$$\delta_h = \frac{\delta}{2C_{\text{inv}}} h^{1+n/2}. \tag{3.1}$$

By Lemma 3.1, for $v_h \in V_h$, $\|v_h - I_h u\|_1 \leqslant \delta_h$, v_h is piecewise strictly convex with smallest eigenvalue bounded below by $m'/2$ and above by $3M'/2$. Put

$$X_h = \{v_h \in V_h,\ v_h = g_h \text{ on } \partial\Omega,\ \|v_h - I_h u\|_1 < \delta_h\}.$$

As a consequence of Assumption 2.5, we obtain the following lemma..

LEMMA 3.2 There exists constants $m, M > 0$ independent of h such that, for all $v_h \in X_h$,

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

It follows that

$$m|w|^2_{1,K} \leqslant \int_K [(\operatorname{cof}\, D^2 v_h(x)) Dw(x)] \cdot Dw(x)\, \mathrm{d}x \leqslant M\, |w|^2_{1,K}, \quad w \in H^1(K). \tag{3.2}$$

The main result of this section is the following theorem.

THEOREM 3.3 Let the finite-dimensional spaces $V_h \subset H^1(\Omega)$ contain piecewise polynomials of degree $d \geqslant n$, $n = 2, 3$. Assume that the spaces V_h satisfy Assumption 2.3 of approximation property and Assumption 2.5 of inverse estimates. Assume also that the exact solution $u \in W^{l+1,\infty}(\mathscr{T}_h) \cap H^1(\Omega)$, $n \leqslant l \leqslant d$ satisfies Assumption 2.5 of strict convexity and solves (2.6). Then, the problem (1.2) has a unique local solution u_h in a small neighbourhood of $I_h u$. The solution u_h is strictly convex on each element and we have the error estimates

$$\|u - u_h\|_{2,\mathscr{T}_h} \leqslant C h^{l-1},$$
$$\|u - u_h\|_1 \leqslant C h^l,$$

for h sufficiently small. Moreover, with a sufficiently close initial guess u_h^0, the sequence defined by, $u_h^{k+1} = g_h$ on $\partial\Omega$,

$$\frac{\nu}{\alpha^{n-1}} \int_\Omega D u_h^{k+1} \cdot D v_h\, \mathrm{d}x = \frac{\nu}{\alpha^{n-1}} \int_\Omega D u_h^k \cdot D v_h\, \mathrm{d}x - \int_\Omega f v_h\, \mathrm{d}x + \sum_{K \in \mathscr{T}_h} \int_K (\det D^2 u_h^k) v_h\, \mathrm{d}x, \tag{3.3}$$

$\forall v_h \in V_h \cap H_0^1(\Omega)$, converges linearly to u_h in the H^1-norm for $\nu = (M + m)/2, \alpha = h^5$ and for h sufficiently small.

Before we give the proof of the above theorem, we will state several lemmas whose proofs are given at the end of the section.

We recall that $\alpha > 0$ is a small parameter which may depend on h. For $\rho > 0$, let

$$B_h(\rho) = \{v_h \in V_h,\ v_h = g_h \text{ on } \partial\Omega,\ \|v_h - I_h u\|_1 \leqslant \rho\}.$$

The ball $B_h(\rho)$ is nonempty as it contains $I_h u$. If $v_h \in B_h(\rho)$, $\|\alpha v_h - \alpha I_h u\|_1 \leqslant \alpha\rho$. Note also that if v_h is strictly convex, so is αv_h.

For a given $v_h \in V_h$, $v_h = g_h$ on $\partial\Omega$, define $T(\alpha v_h) \in V_h$ as the solution of

$$\nu \int_\Omega DT(\alpha v_h) \cdot Dw_h \,\mathrm{d}x = \nu \int_\Omega D(\alpha v_h) \cdot Dw_h \,\mathrm{d}x + \alpha^n \sum_{K \in \mathscr{T}_h} \int_K (\det D^2 v_h) w_h \,\mathrm{d}x - \alpha^n \int_\Omega f w_h \,\mathrm{d}x \quad \forall w_h \in V_h \cap H_0^1(\Omega), \tag{3.4}$$

with $\alpha v_h - T(\alpha v_h) = 0$ on $\partial\Omega$ and we recall that $\nu = (M+m)/2$, where M and m are the constants of Lemma 3.2.

We will show that T has a unique fixed point αu_h with u_h in $B_h(\rho)$ for h sufficiently small.

The motivation to introduce the damping parameter α is that it allows one to solve a rescaled version of (1.1). Indeed, $\det D^2 u = f$ is equivalent to $\det \alpha D^2 u = \alpha^n f$. Taking α as a power of h will play a crucial role in proving the well-posedness of (1.2) and obtaining optimal error estimates.

LEMMA 3.4 The mapping T is well defined and if αu_h is a fixed point of T, i.e., $T(\alpha u_h) = \alpha u_h$, then u_h solves (1.2).

The next lemma says that the mapping T does not move the centre $I_h u$ of a ball $B_h(\rho)$ too far.

LEMMA 3.5 We have

$$\|\alpha I_h u - T(\alpha I_h u)\|_1 \leqslant C_1 \alpha^n h^{l-1} \|u\|_{2,\infty}^{n-1} \|u\|_{l+1}. \tag{3.5}$$

The next two lemmas establish the contraction mapping property of T under the assumption that $d \geqslant n$ and $\alpha = h^5$.

LEMMA 3.6 For h sufficiently small, and $0 < \rho < \delta_h$, T is a strict contraction mapping in the ball $\alpha B_h(\rho)$, i.e., for $v_h, w_h \in B_h(\rho)$,

$$\|T(\alpha v_h) - T(\alpha w_h)\|_1 \leqslant a \|\alpha v_h - \alpha w_h\|_1, \quad 0 < a < 1.$$

LEMMA 3.7 For h sufficiently small and $\rho = \delta/(4C_{\mathrm{inv}})h^l, n \leqslant l \leqslant d$, where δ and C_{inv} are the constants in (3.1), T is a strict contraction in $\alpha B_h(\rho)$ and maps $\alpha B_h(\rho)$ into itself.

The previous lemmas will readily allow us to conclude the solvability of (1.2) and derive error estimates in the H^1-norm by using the explicit expression of the radius ρ of the above lemma. We can now give the proof of Theorem 3.3.

Proof of Theorem 3.3. Since the mapping T is a strict contraction which maps $\alpha B_h(\rho)$ into itself, the existence of a fixed point αu_h with $u_h \in B_h(\rho)$ follows from the Banach fixed-point theorem. By Lemma 3.4, u_h solves (1.2).

From the expression of ρ given in Lemma 3.7, we get, using the value of $\alpha = h^5$,

$$\|u - u_h\|_1 \leqslant \|u - I_h u\|_1 + \|I_h u - u_h\|_1 \leqslant Ch^l |u|_{l+1} + Ch^l \leqslant Ch^l,$$

which proves the H^1 error estimate. By (2.3) and (2.5),

$$\begin{aligned}\|u-u_h\|_{2,\mathscr{T}_h} &\leqslant \|u-I_hu\|_{2,\mathscr{T}_h}+\|I_hu-u_h\|_{2,\mathscr{T}_h}\\ &\leqslant \|u-I_hu\|_{2,\mathscr{T}_h}+h^{-1}\|I_hu-u_h\|_1\\ &\leqslant Ch^{l-1}|u|_{l+1}+Ch^{l-1},\end{aligned}$$

which proves that

$$\|u-u_h\|_{2,\mathscr{T}_h}\leqslant Ch^{l-1}.$$

Finally, we prove the convergence of the time marching method (3.3). Since T is a strict contraction in $\alpha B_h(\rho)$, the sequence defined by $\alpha u_h^{k+1}=T(\alpha u_h^k)$, $u_h^{k+1}=u_h^k$ on $\partial\Omega$ converges linearly to αu_h. Simplifying by α^n, we get the convergence of (3.3). □

We conclude this section with the proofs of Lemmas 3.1–3.7.

Proof of Lemma 3.1. Recall that the eigenvalues of a (symmetric) matrix are continuous functions of its entries, as roots of the characteristic equation, Ostrowski (1960) Appendix K, or Harris & Martin (1987). Thus, for all $\epsilon>0$, there exists $\delta>0$ such that for $v\in W^{2,\infty}(\Omega)$, $|v-u|_{2,\infty}\leqslant\delta$ implies $|\lambda_1(D^2v(x))-\lambda_1(D^2u(x))|<\epsilon$ a.e. in Ω.

By Assumption 2.5, $\lambda_1(D^2u(x))\geqslant m'$, a.e. in Ω, and with $\epsilon=m'/2$, we obtain $\lambda_1(D^2v(x))>m'/2$, a.e. in Ω. We conclude that, for $|v-u|_{2,\infty}\leqslant\delta$, $\lambda_1(D^2v(x))>m'/2$ a.e. in Ω.

Now, by (2.3), $|u-I_hu|_{2,\infty}\leqslant C_{ap}h^{d-1}|u|_{d+1,\infty}$. So, for h sufficiently small, $|u-I_hu|_{2,\infty}\leqslant\delta/2$. Moreover, by (2.5) and the assumption of the lemma

$$|v_h-I_hu|_{2,\infty}\leqslant C_{\mathrm{inv}}h^{-1-n/2}\|v_h-I_hu\|_1\leqslant\frac{\delta}{2}.$$

It follows that $\lambda_1(D^2v_h(x))>m'/2$ a.e. in Ω, as claimed.

If necessary, by taking δ smaller, we have $|\lambda_n(D^2v_h(x))-\lambda_n(D^2u(x))|<M'/2$ a.e. in Ω. Thus, $\lambda_n(D^2v_h(x))\leqslant\lambda_n(D^2u(x))+M'/2\leqslant 3M'/2$. This concludes the proof. □

Proof of Lemma 3.2. We first note that, by Lemma 3.1, there exists constants $m,M>0$ such that $m\leqslant\lambda_1(\operatorname{cof}D^2v_h(x))\leqslant\lambda_n(\operatorname{cof}D^2v_h(x))\leqslant M$ a.e. in Ω for $v_h\in X_h$. To prove this, recall that, for an invertible matrix A, $\operatorname{cof}A=(\det A)(A^{-1})^{\mathrm{T}}$. Since a matrix and its transpose have the same set of eigenvalues, the eigenvalues of $\operatorname{cof}A$ are of the form $\det A/\lambda_i$, where $\lambda_i, i=1,\ldots,n$ is an eigenvalue of A. Applying this observation to $A=D^2u(x)$ and using Lemma 3.1, we obtain that the eigenvalues of $\operatorname{cof}D^2v_h(x)$ are a.e. uniformly bounded below by $m=(m')^n/M'$ and above by $M=(M')^n/m'$.

Since $\lambda_1(D^2v_h(x))$ and $\lambda_n(D^2v_h(x))$ are the minimum and maximum, respectively, of the Rayleigh quotient $[(\operatorname{cof}\,D^2v_h(x))z]\cdot z/\|z\|^2$, where $\|z\|$ denotes the standard Euclidean norm in $\mathbb{R}^n$, we have

$$m'\|z\|^2\leqslant[(\operatorname{cof}\,D^2v_h(x))z]\cdot z\leqslant M'\|z\|^2,\quad z\in\mathbb{R}^n.$$

This implies

$$m|w|_{1,K}^2\leqslant\int_K[(\operatorname{cof}\,D^2v_h(x))Dw(x)]\cdot Dw(x)\,\mathrm{d}x\leqslant M|w|_{1,K}^2,\quad w\in H^1(K).$$

□

Proof of Lemma 3.4. The existence of $T(\alpha v_h)$ solving (3.4) is an immediate consequence of the Lax–Milgram lemma.

If $T(\alpha u_h) = \alpha u_h$, then

$$\alpha^n \sum_{K\in\mathscr{T}_h} \int_K (\det D^2 u_h) v_h \, \mathrm{d}x = \alpha^n \int_\Omega f v_h \, \mathrm{d}x \quad \forall v_h \in V_h \cap H_0^1(\Omega),$$

and thus u_h solves (1.2). Conversely, if u_h solves (1.2), αu_h is a fixed point of T. □

Proof of Lemma 3.5. From (2.3) and (2.4), we obtain

$$\|u - I_h u\|_{2,K} \leqslant C h^{l-1} |u|_{l+1}, \tag{3.6}$$

$$\|I_h u\|_{2,K} \leqslant C \|u\|_2. \tag{3.7}$$

Put $w_h = \alpha I_h u - T(\alpha I_h u)$ and note that $w_h \in H_0^1(\Omega)$. Since the exact solution solves (2.6), we have

$$\int_\Omega f w_h \, \mathrm{d}x = \int_\Omega (\det D^2 u) w_h \, \mathrm{d}x.$$

With $v_h = I_h u$ in (3.4), we obtain

$$\nu \int_\Omega D[T(\alpha v_h) - \alpha v_h] \cdot D w_h \, \mathrm{d}x = \alpha^n \left(\sum_{K\in\mathscr{T}_h} \int_K (\det D^2 I_h u - \det D^2 u) w_h \, \mathrm{d}x \right). \tag{3.8}$$

Put

$$z_h = \det D^2 I_h u - \det D^2 u.$$

We have, by Lemma 2.1,

$$z_h = (\mathrm{cof}(t D^2 I_h u + (1-t) D^2 u)) : (D^2 I_h u - D^2 u),$$

for some $t \in [0,1]$. Thus, by Lemma 2.2, (3.7) and (3.6),

$$\begin{aligned} \|z_h\|_{0,K} &\leqslant C \|t D^2 I_h u + (1-t) D^2 u)\|_\infty^{n-1} \|D^2 I_h u - D^2 u\|_{0,K} \\ &\leqslant C(\|I_h u\|_{2,\infty} + \|u\|_{2,\infty})^{n-1} \|I_h u - u\|_{2,K} \\ &\leqslant C \|u\|_{2,\infty}^{n-1} h^{l-1} \|u\|_{l+1,K} \leqslant C h^{l-1} \|u\|_{2,\infty}^{n-1} \|u\|_{l+1,K}. \end{aligned}$$

By Lemma 3.4 and (3.8), we obtain

$$\begin{aligned} \nu |w_h|_1^2 &\leqslant C\alpha^n \sum_{K\in\mathscr{T}_h} \|z_h\|_{0,K} \|w_h\|_{0,K} \\ &\leqslant C\alpha^n h^{l-1} \|u\|_{2,\infty}^{n-1} \|u\|_{l+1} \|w_h\|_0 \leqslant C\alpha^n h^{l-1} \|u\|_{2,\infty}^{n-1} \|u\|_{l+1} \|w_h\|_1. \end{aligned}$$

The result then follows by Poincaré's inequality. □

Proof of Lemma 3.6. We define

$$V_K = \{v_h|_K,\ K \in \mathscr{T}_h,\ v_h \in X_h\},$$

and denote by V_K' the space of linear continuous functionals on V_K. For $F \in V_K'$, $\|F\|$ will denote the operator norm of F. We define a mapping $T_K : \alpha B_h(\rho) \to V_K'$ defined by

$$\langle T_K(\alpha v_h), z_h\rangle = \alpha \int_K Dv_h \cdot Dz_h \,\mathrm{d}x + \frac{\alpha^n}{\nu}\int_K (\det D^2 v_h) z_h \,\mathrm{d}x - \frac{\alpha^n}{\nu}\int_K f z_h \,\mathrm{d}x.$$

Note that the restriction of elements of $\alpha B_h(\rho)$ to K are in V_K.

Step 1: We claim that, for $v_h \in B_h(\rho)$ and $w_h \in V_K$, $\|T_K'(\alpha v_h)(\alpha w_h)\| \leqslant a\|w_h\|_{1,K}$ for a constant a such that $0 < a < 1$ and h sufficiently small.

$$\begin{aligned}\langle T_K'(\alpha v_h)(\alpha w_h), z_h\rangle &= \alpha \int_K Dw_h \cdot Dz_h \,\mathrm{d}x + \frac{\alpha^n}{\nu}\int_K [\operatorname{div}(\operatorname{cof} D^2 v_h) Dw_h] z_h \,\mathrm{d}x \\ &= \alpha \int_K Dw_h \cdot Dz_h \,\mathrm{d}x - \frac{\alpha^n}{\nu}\int_K [(\operatorname{cof} D^2 v_h) Dw_h]\cdot Dz_h \,\mathrm{d}x \\ &\quad + \frac{\alpha^n}{\nu}\int_{\partial K} z_h[(\operatorname{cof} D^2 v_h) Dw_h]\cdot n_K \,\mathrm{d}s,\end{aligned}$$

and we used the expression of the derivative of the mapping $u \to \det D^2 u$ also used in the proof of Lemma 2.1. Therefore,

$$\begin{aligned}\langle T_K'(\alpha v_h)(\alpha w_h), z_h\rangle = \alpha \int_K \left[\left(I - \frac{1}{\nu}\operatorname{cof} D^2 \alpha v_h\right) Dw_h\right]\cdot Dz_h \,\mathrm{d}x \\ + \frac{\alpha^n}{\nu}\int_{\partial K} z_h[(\operatorname{cof} D^2 v_h) Dw_h]\cdot n_K \,\mathrm{d}s,\end{aligned} \tag{3.9}$$

where I is the $n \times n$ identity matrix. We define

$$\beta = \sup_{w_h \in V_K, |w_h|_{1,K}=1} \left| \int_K \left[\left(I - \frac{1}{\nu}\operatorname{cof} D^2 \alpha v_h)\right) Dw_h\right]\cdot Dw_h \,\mathrm{d}x\right|.$$

By assumption $\rho < \delta_h$. Thus, by (3.2) we obtain

$$\left(1 - \frac{M\alpha^{n-1}}{\nu}\right)|w|_{1,K}^2 \leqslant \int_K \left[\left(I - \frac{1}{\nu}(\operatorname{cof} D^2 \alpha v_h)\right) Dw\right]\cdot Dw \,\mathrm{d}x \leqslant \left(1 - \frac{m\alpha^{n-1}}{\nu}\right)|w|_{1,K}^2.$$

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

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

Thus, since for h sufficiently small

$$\alpha^{n-1} < \frac{m+M}{2M} \leqslant \frac{m+M}{2m},$$

we have

$$0 \leqslant \beta < 1.$$

Define $p_h = w_h/|w_h|_{1,K}$ and $q_h = z_h/|z_h|_{1,K}$ for $w_h \neq 0$ and $z_h \neq 0$. Then,

$$\frac{|\int_K [(I-(1/\nu)\operatorname{cof} D^2\alpha v_h)Dw_h]\cdot Dz_h\,\mathrm{d}x|}{|w_h|_{1,K}|z_h|_{1,K}} = \left|\int_K \left[\left(I - \frac{1}{\nu}\operatorname{cof} D^2\alpha v_h\right) Dp_h\right]\cdot Dq_h\,\mathrm{d}x\right|. \tag{3.10}$$

We can define a bilinear form on V_K by the formula

$$(p,q) = \int_\Omega \left[\left(I - \frac{1}{\nu}(\operatorname{cof} D^2\alpha v_h)\right) Dp\right]\cdot Dq\,\mathrm{d}x.$$

Then, because

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

and using the definition of β, we obtain

$$|(p_h,q_h)| \leqslant \frac{\beta}{4}|p_h+q_h|_{1,K}^2 + \frac{\beta}{4}|p_h-q_h|_{1,K}^2 = \beta,$$

since p_h and q_h are unit vectors in the $|.|_{1,K}$ seminorm. It follows from (3.10) that

$$\frac{|\int_K [(I-(1/\nu)\operatorname{cof} D^2\alpha v_h)Dw_h]\cdot Dz_h\,\mathrm{d}x|}{|w_h|_{1,K}|z_h|_{1,K}} \leqslant \beta.$$

We then have, for $v_h, w_h \in V_K$,

$$\left|\int_K \left[\left(I - \frac{1}{\nu}\operatorname{cof} D^2\alpha v_h\right) Dw_h\right]\cdot Dz_h\,\mathrm{d}x\right| \leqslant \beta|w_h|_{1,K}|z_h|_{1,K} \leqslant \beta\|w_h\|_{1,K}\|z_h\|_{1,K}.$$

In other words,

$$\frac{|\int_K [(I-(1/\nu)\operatorname{cof} D^2\alpha v_h)Dw_h]\cdot Dz_h\,\mathrm{d}x|}{\|w_h\|_{1,K}\|z_h\|_{1,K}} \leqslant \beta. \tag{3.11}$$

Next, we bound the second term on the right of (3.9). We need the scaled trace inequality

$$\|v\|_{L^2(\partial K)} \leqslant Ch_K^{-1/2}\|v\|_{L^2(K)} \quad \forall v \in V_h. \tag{3.12}$$

We have, by Schwarz inequality, (3.12) and (2.5),

$$\begin{aligned}\int_{\partial K} z_h[(\operatorname{cof} D^2 v_h)Dw_h]\cdot n_K\,\mathrm{d}s &\leqslant Ch^{-1}\|(\operatorname{cof} D^2 v_h)Dw_h\|_{0,K}\|z_h\|_{0,K}\\ &\leqslant Ch^{-1}|v_h|_{2,\infty,K}^{n-1}\|w_h\|_{1,K}\|z_h\|_{1,K}\\ &\leqslant Ch^{-(1+n/2)(n-1)-1}\|w_h\|_{1,K}\|v_h\|_{1,K}^{n-1}\|z_h\|_{1,K}.\end{aligned} \tag{3.13}$$

By (3.11) and (3.13),

$$\frac{|\langle T_K'(\alpha v_h)(\alpha w_h), z_h\rangle|}{\|w_h\|_{1,K}\|z_h\|_{1,K}} \leqslant \alpha(\beta + C\alpha^{n-1}h^{-(1+n/2)(n-1)-1}\|v_h\|_{1,K}^{n-1}).$$

We conclude, using the expression of $\alpha = h^5$ and assuming $\rho \leqslant 1$, that

$$\begin{aligned}
\|T_K'(\alpha v_h)(\alpha w_h)\| &= \sup_{z_h \neq 0} \frac{|\langle T_K'(\alpha v_h)(\alpha w_h), z_h\rangle|}{\|z_h\|_{1,K}} \\
&\leqslant \alpha(\beta + Ch^{(4-n/2)(n-1)-1}\|v_h\|_{1,K}^{n-1})\|w_h\|_{1,K} \\
&\leqslant \alpha(\beta + Ch^{(4-n/2)(n-1)-1}(\|v_h - I_h u\|_{1,K} + \|I_h u\|_{1,K})^{n-1})\|w_h\|_{1,K} \\
&\leqslant \alpha(\beta + Ch^{(4-n/2)(n-1)-1}(\rho + \|u\|_1)^{n-1})\|w_h\|_{1,K} \\
&\leqslant (\beta + Ch^{1/2}(1 + \|u\|_1)^{n-1})\|\alpha w_h\|_{1,K},
\end{aligned}$$

and we recall that $n = 2, 3$, allowing us to treat the two cases in a unified fashion.

Since $\beta < 1$, for h sufficiently small $a = \beta + Ch^{1/2}(1 + \|u\|_1)^{n-1} < 1$. This proves the result.

Step 2: The mapping T_K is a strict contraction, i.e., for $v_h, w_h \in B_h(\rho)$, $\|T_K(\alpha v_h) - T_K(\alpha w_h)\| \leqslant a\|\alpha v_h - \alpha w_h\|_{1,K}$, $0 < a < 1$.

Using the mean value theorem,

$$\begin{aligned}
\|T_K(\alpha v_h) - T_K(\alpha w_h)\| &= \|\int_0^1 T_K'(\alpha v_h + t(\alpha w_h - \alpha v_h))(\alpha w_h - \alpha v_h)\,\mathrm{d}t\| \\
&\leqslant \int_0^1 \|T_K'(\alpha v_h + t(\alpha w_h - \alpha v_h))(\alpha w_h - \alpha v_h)\|\,\mathrm{d}t.
\end{aligned}$$

Since $B_h(\rho)$ is convex, $v_h + t(w_h - v_h) \in B_h(\rho), t \in [0, 1]$, and by the result established in Step 1,

$$\|T_K(\alpha v_h) - T_K(\alpha w_h)\| \leqslant \int_0^1 a\|\alpha w_h - \alpha v_h\|_{1,K}\,\mathrm{d}t = a\|\alpha w_h - \alpha v_h\|_{1,K}.$$

Step 3: The mapping T is a strict contraction in $\alpha B_h(\rho)$.

$$\begin{aligned}
\int_\Omega D(T(\alpha v_h) - T(\alpha w_h)) \cdot D\psi_h\,\mathrm{d}x &= \alpha \int_\Omega D(v_h - w_h) \cdot D\psi_h\,\mathrm{d}x \\
&\quad + \frac{\alpha^n}{\nu} \sum_{K \in \mathscr{T}_h} \int_K (\det D^2 v_h - \det D^2 w_h)\psi_h\,\mathrm{d}x \\
&= \sum_{K \in \mathscr{T}_h} \langle T_K(\alpha v_h) - T_K(\alpha w_h), \psi_h\rangle.
\end{aligned}$$

With $\psi_h = T(\alpha v_h) - T(\alpha w_h)$, we obtain, using the result from Step 2 and the observation that $\psi_h = 0$ on $\partial\Omega$,

$$
\begin{aligned}
|T(\alpha v_h) - T(\alpha w_h)|_1^2 &\leqslant \sum_{K\in\mathscr{T}_h} \|T_K(\alpha v_h) - T_K(\alpha w_h)\| \, \|\psi_h\|_{1,K} \\
&\leqslant a \sum_{K\in\mathscr{T}_h} \|\alpha v_h - \alpha w_h\|_{1,K} \|\psi_h\|_{1,K} \\
&\leqslant a \left(\sum_{K\in\mathscr{T}_h} \|\alpha v_h - \alpha w_h\|_{1,K}^2 \right)^{1/2} \left(\sum_{K\in\mathscr{T}_h} \|\psi_h\|_{1,K}^2 \right)^{1/2} \\
&= a\|\alpha v_h - \alpha w_h\|_1 \|\psi_h\|_1 \\
&\leqslant aC_p \|\alpha v_h - \alpha w_h\|_1 |\psi_h|_1,
\end{aligned}
$$

where C_p is the constant in the Poincare's inequality. It follows that $\|T(\alpha v_h) - T(\alpha w_h)\|_1 \leqslant a\|\alpha v_h - \alpha w_h\|_1$. $\square$

Proof of Lemma 3.7. Since $\rho = \delta/(4C_{\text{inv}})h^l$ and $l \geqslant n$ for $n = 2, 3$, we obtain $l \geqslant 1 + n/2$ and thus $\rho < \delta_h$.

Let $v_h \in B_h(\rho)$. Then, using Lemma 3.5 and the observation that $h^{l+1} \leqslant (1-a)\rho$ for h sufficiently small

$$
\begin{aligned}
\|T(\alpha v_h) - \alpha I_h u\|_1 &\leqslant \|T(\alpha v_h) - T(\alpha I_h u)\|_1 + \|T(\alpha I_h u) - \alpha I_h u\|_1 \\
&\leqslant a\|\alpha v_h - \alpha I_h u\|_1 + C_1 \alpha^n h^{l-1} \|u\|_{l+1,\infty}^n \\
&\leqslant a\|\alpha v_h - \alpha I_h u\|_1 + C\alpha h^{l+5n-6} \\
&\leqslant a\|\alpha v_h - \alpha I_h u\|_1 + C\alpha h^{l+1} \\
&\leqslant a\|\alpha v_h - \alpha I_h u\|_1 + (1-a)\alpha\rho \\
&\leqslant a\alpha\rho + (1-a)\alpha\rho \\
&\leqslant \alpha\rho,
\end{aligned}
$$

and we conclude that

$$\|T(\alpha v_h) - \alpha I_h u\|_1 \leqslant \alpha\rho.$$

This proves the result. $\square$

REMARK 3.8 Let us assume that (1.2) has a strictly convex solution u_h (independently of the smoothness of u). If, in addition, its eigenvalues are bounded below and above by constants independent of h, then using again the continuity of the eigenvalues of a matrix as a function of its entries, we obtain the existence of $\delta' > 0$ such that, for v_h in

$$Y^h = \{v_h \in V_h,\ v_h = g_h \text{ on } \partial\Omega,\ \|v_h - u_h\|_1 < C\delta' h^{1+n/2}\},$$

TABLE 1 *Test 1* $d = 2$, $\bar{\nu} = 50$

h	Iterations	$\|u - u_h\|_{L^2}$	Rate	$\|u - u_h\|_{H^1}$	Rate
$\frac{1}{4}$	71	4.38×10^{-2}		2.05×10^{-1}	
$\frac{1}{8}$	54	2.18×10^{-2}	1.00	1.04×10^{-1}	0.98
$\frac{1}{16}$	461	9.00×10^{-3}	1.28	4.19×10^{-2}	1.31
$\frac{1}{32}$	493	2.76×10^{-3}	1.70	1.28×10^{-2}	1.71
$\frac{1}{64}$	459	7.35×10^{-4}	1.91	3.40×10^{-3}	1.91
$\frac{1}{128}$	448	1.86×10^{-4}	1.98	8.65×10^{-4}	1.97

v_h is convex. It is not difficult to see that the mapping T is also a strict contraction in Y^h for h sufficiently small. One obtains the linear convergence of the iterative method (3.3) to u_h as follows:

$$\|\alpha u_h^{k+1} - \alpha u_h\|_1 = \|T(\alpha u_h^k) - T(\alpha u_h)\|_1 \leqslant a\|\alpha u_h^k - \alpha u_h\|_1, \quad 0 < a < 1.$$

Simplifying by α proves the claim.

4. Numerical results

The implementation is done in Matlab. The computational domain is the unit square $[0, 1]^2$ which is first divided into squares of side length h. Then, each square is divided into two triangles by the diagonal with positive slope. We use standard test functions for numerical convergence to viscosity solutions of nondegenerate Monge–Ampère equations, i.e., for $f > 0$ in Ω.

Test 1: $u(x, y) = e^{(x^2+y^2)/2}$ with corresponding f and g. This solution is infinitely differentiable.

Test 2: $u(x, y) = -\sqrt{2 - x^2 - y^2}$ with corresponding f and g. This solution is not in $H^2(\Omega)$.

Test 3: $g(x, y) = 0$ and $f(x, y) = 1$. No exact solution is known in this case.

The initial guess was taken as the finite element approximation of the solution of the Poisson equation

$$\Delta u^0 = 2\sqrt{f} \text{ in } \Omega, \quad u^0 = g \text{ on } \partial\Omega.$$

We define

$$\bar{\nu} = \frac{\nu}{\alpha^{n-1}},$$

where ν and α are the parameters in (3.3). It is not easy to estimate the lower and upper values m and M. Thus, trial and error are used in the selection of $\bar{\nu}$. If the numerical error is deemed not accurate, one increases $\bar{\nu}$. In Awanou (2013d), it is shown that if one regularizes the data, the sequence of problems associated with (1.2) have solutions which are piecewise strictly convex. In that case, trial and error should also be used for the choice of $\bar{\nu}$. In Tables 1 and 2, we used the value of $\bar{\nu}$ that gives a sufficiently accurate solution at the finest level of refinement displayed. On coarser meshes, a smaller value of $\bar{\nu}$ may be used, and hence the solution could be obtained with fewer iterations.

For the test function in Test 1, which is a smooth function and the one in Test 3 (Fig. 1), we used the iterative method of Theorem 3.3. For the nonsmooth solution of Test 2, we found the following truncated version more efficient. For $m = 1, 2, \ldots$, we consider truncating functions $\chi_m(x)$ defined by $\chi_m(x) = -m$

TABLE 2 *Test 2* $d = 2$, $\bar{\nu} = 150$, $m = 250$

h		$\|u - u_h\|_{L^2}$	Rate	$\|u - u_h\|_{H^1}$	Rate
$\frac{1}{16}$	106	1.79×10^{-1}		1.1718	
$\frac{1}{32}$	249	6.54×10^{-2}	1.45	5.47×10^{-1}	1.10
$\frac{1}{64}$	554	1.24×10^{-2}	2.40	1.52×10^{-1}	1.85
$\frac{1}{128}$	886	2.10×10^{-3}	2.56	6.00×10^{-2}	1.34
$\frac{1}{256}$	993	4.91×10^{-4}	2.09	4.27×10^{-2}	0.49

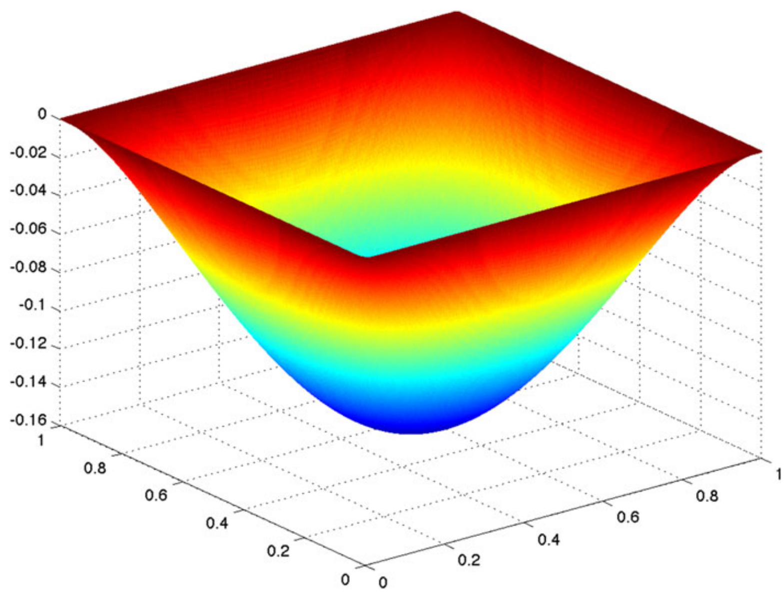

FIG. 1. Test 3 $d = 2$, $h = \frac{1}{2}^8$, $\bar{\nu} = 50$, with three iterations.

for $x < -m$, $\chi_m(x) = x$ for $-m \leqslant x \leqslant m$ and $\chi_m(x) = m$ for $x > m$ and the sequence of problems

$$\bar{\nu} \int_{\Omega} Du_h^{k+1,m} \cdot Dv_h \, \mathrm{d}x = \bar{\nu} \int_{\Omega} Du_h^{k,m} \cdot Dv_h \, \mathrm{d}x + \sum_{K \in \mathcal{T}_h} \int_K \chi_m(\det D^2 u_h^{k,m} - f) v_h \, \mathrm{d}x,$$

with $u_h^{k+1,m} = g_h$ on $\partial\Omega$.

Compared with C^1 conforming approximations or mixed methods, the standard finite element method is less able to capture convex solutions. However, we note the unusual high-order convergence rate in the L^2-norm for the nonsmooth solution of Test 2. The optimal convergence rate of Theorem 3.3 is an asymptotic convergence rate. For higher-order elements, better numerical convergence rates are obtained with the iterative methods discussed in Awanou (2013c) or the finite element version of the ones discussed in Benamou *et al.* (2010). It could just be that the time marching method is not accurate enough to approximate the solution for high-order elements. In summary, the discrete problem (1.2) has a unique local solution, and the time marching method proposed in this paper is efficient for nonsmooth solutions, with $f > 0$, and quadratic elements. In particular, the time marching method is not robust enough to handle the situation, where the right-hand side of (1.1) is a measure, i.e., a Dirac distribution. For these situations, in the context of standard discretizations, one may have to use mixed methods as in Lakkis & Pryer (2013) or Neilan (2014). In addition, Newton's method can then be applied directly to the resolution of the nonlinear system resulting from the discretization of (1.1). It seems that the time marching method is enough to handle easily strictly convex viscosity solutions for the discretization discussed in this paper. The extension of the analysis in Awanou (2013d) to the case of mixed methods will be discussed in Awanou (2014).

5. Concluding remarks

REMARK 5.1 The motivation to choose the test functions in (1.2) to be in $H_0^1(\Omega)$ stems from the use of Lagrange elements in (3.3).

REMARK 5.2 Numerical evidence of convergence of standard discretizations of the Monge–Ampère equation to nonsmooth solutions has been discussed for a long time in the finite element context (Feng & Neilan, 2009; Glowinski, 2009; Lakkis & Pryer, 2013) and the references therein. In Benamou *et al.* (2010), for the two-dimensional Monge–Ampère equation, it was proposed to discretize, with the standard finite difference method, the following iterative method:

$$\Delta u_{k+1} = ((\Delta u_k)^2 + 2(f - \det D^2 u_k))^{1/2} \text{ in } \Omega, \quad u_{k+1} = g \text{ on } \partial\Omega.$$

As for the time marching method (3.3), the above approach requires solving only Poisson equations, and can be used for numerical evidence of convergence to nonsmooth solutions even with a finite element discretization. However, even for smooth solutions, its convergence properties are not understood.

REMARK 5.3 With regard to general fully nonlinear equations, the time marching method has been applied to the Pucci equation and the Gauss curvature equation in a previous version of Awanou (2013b). Since the proof of convergence for the Monge–Ampère equation exploits the divergence form of the equation, it is not clear whether the techniques used here can be extended to other fully nonlinear equations. It should also be noted that the Monge–Ampère equation has a geometric structure, as evidenced by the version of the Aleksandrov theory, which consists in approximation by smooth functions. This makes it easier to understand convergence of standard discretizations to nonsmooth solutions; see Awanou (2013d).

Acknowledgements

The author would like to thank the referees for a careful reading of the paper and their suggestions which have led to a better presentation of the paper.

Funding

This work began when the author was supported in part by a 2009–2013 Sloan Foundation Fellowship, and continued while the author was in residence at the Mathematical Sciences Research Institute (MSRI) in Berkeley, California, Fall 2013. The MSRI receives major funding from the National

Science Foundation under Grant No. 0932078000. The author was partially supported by NSF DMS grant No. 1319640.

References

Awanou, G. (2013a) On standard finite difference discretizations of the elliptic Monge–Ampère equation. Submitted.

Awanou, G. (2013b) Pseudo transient continuation and time marching methods for Monge–Ampère type equations. http://arxiv.org/pdf/1301.5891.pdf Accessed 13 May 2014.

Awanou, G. (2013c) Spline element method for the Monge–Ampère equation. http://arxiv.org/abs/1012.1775 Accessed 13 May 2014.

Awanou, G. (2013d) Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: Aleksandrov solutions. http://arxiv.org/pdf/1310.4568v1.pdf Accessed 13 May 2014.

Awanou, G. (2014) Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: mixed methods. In preparation.

Benamou, J.-D., Froese, B. D. & Oberman, A. M. (2010) Two numerical methods for the elliptic Monge–Ampère equation. *M2AN Math. Model. Numer. Anal.*, **44**, 737–758.

Brenner, S. C., Gudi, T., Neilan, M. & Sung, L.-Y. (2011) C^0 penalty methods for the fully nonlinear Monge–Ampère equation. *Math. Comp.*, **80**, 1979–1995.

Brenner, S. C. & Neilan, M. (2012) Finite element approximations of the three dimensional Monge–Ampère equation. *ESAIM Math. Model. Numer. Anal.*, **46**, 979–1001.

Brenner, S. C. & Scott, L. R. (2002) *The Mathematical Theory of Finite Element Methods*, 2nd edn. Texts in Applied Mathematics, vol. 15. New York: Springer, pp. xvi + 361.

Evans, L. C. (1998) *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19. Providence, RI: American Mathematical Society, pp. xviii + 662.

Feng, X. & Neilan, M. (2009) 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**, 1226–1250.

Glowinski, R. (2009) Numerical methods for fully nonlinear elliptic equations. *ICIAM 07—6th International Congress on Industrial and Applied Mathematics*. Zurich: European Mathematical Society, pp. 155–192.

Harris, G. & Martin, C. (1987) The roots of a polynomial vary continuously as a function of the coefficients. *Proc. Amer. Math. Soc.*, **100**, 390–392.

Lakkis, O. & Pryer, T. (2013) A finite element method for nonlinear elliptic problems. *SIAM J. Sci. Comput.*, **35**, A2025–A2045.

Neilan, M. (2013) Quadratic finite element approximations of the Monge–Ampère equation. *J. Sci. Comput.*, **54**, 200–226.

Neilan, M. (2014) 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.

Ostrowski, A. M. (1960) *Solution of Equations and Systems of Equations*. Pure and Applied Mathematics, vol. IX. New York–London: Academic Press, pp. ix + 202.