Numerische Mathematik (2021) 148:497–524
https://doi.org/10.1007/s00211-021-01210-x

Numerische Mathematik

A convexity enforcing C^0 interior penalty method for the Monge–Ampère equation on convex polygonal domains

Susanne C. Brenner[1] · Li-yeng Sung[1] · Zhiyu Tan[2] · Hongchao Zhang[1]

Received: 14 December 2020 / Revised: 19 April 2021 / Accepted: 23 May 2021 / Published online: 26 June 2021
© The Author(s) 2021

Abstract
We design and analyze a C^0 interior penalty method for the approximation of classical solutions of the Dirichlet boundary value problem of the Monge–Ampère equation on convex polygonal domains. The method is based on an enhanced cubic Lagrange finite element that enables the enforcement of the convexity of the approximate solutions. Numerical results that corroborate the *a priori* and *a posteriori* error estimates are presented. It is also observed from numerical experiments that this method can capture certain weak solutions.

Mathematics Subject Classification 65N30 · 65K10 · 35G30 · 90C06 · 90C26

1 Introduction

The Monge–Ampère equation is a fully nonlinear partial differential equation that appears in geometric analysis and related applications. Various aspects of this important equation can be found in the monographs [6,20,38,42,44,49,72].

This work of the first two authors was supported in part by the National Science Foundation under Grant No. DMS-19-13035. The work of the fourth author was supported in part by the National Science Foundation under Grant Nos. DMS-18-19161 and DMS-21-10722.

✉ Susanne C. Brenner
brenner@math.lsu.edu

Li-yeng Sung
sung@math.lsu.edu

Zhiyu Tan
ztan@cct.lsu.edu

Hongchao Zhang
hzhang@math.lsu.edu

[1] Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA

[2] Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA

The Dirichlet boundary value problem for the Monge–Ampère equation is given by

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

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

If Ω is a smooth and strictly convex domain, $\psi \in C^3(\bar{\Omega})$ is strictly positive on $\bar{\Omega}$ and $\phi \in C^{4,\delta}(\bar{\Omega})$ for some $\delta \in (0, 1)$, then (1.1) has a unique strictly convex solution $u \in C^{4,\delta}(\bar{\Omega})$ (cf. [21, p. 371, Remark 2]). Our goal is to develop finite element methods that can capture such smooth convex solutions of (1.1).

Remark 1.1 Throughout this paper we will follow the standard notation for differential operators, function spaces and norms that can be found for example in [1,17,23].

As a first step, we consider a finite element method for (1.1) on polygonal domains. Accordingly we assume that $\Omega \subset \mathbb{R}^2$ is a bounded convex polygon,

$$\phi \in H^4(\Omega), \tag{1.2}$$

$$\psi \in H^2(\Omega) \text{ is strictly positive on } \bar{\Omega}, \tag{1.3}$$

and

$$\text{the boundary value problem (1.1) has a strictly convex solution } u \in H^4(\Omega), \tag{1.4}$$

i.e., there exists a positive constant $\alpha_\sharp$ such that the Hessian $D^2 u$ satisfies

$$\xi^t (D^2 u)(x) \xi \geq \alpha_\sharp |\xi|^2 \qquad \forall\, x \in \Omega,\ \xi \in \mathbb{R}^2. \tag{1.5}$$

Remark 1.2 The extension of our method to strictly convex smooth domains, where the regularity (1.4) follows from appropriate regularity of the data, will be carried out in a forthcoming paper.

Remark 1.3 Since a 2×2 symmetric matrix and its cofactor matrix have identical eigenvalues, the estimate (1.5) is equivalent to

$$\xi^t \mathrm{Cof}(D^2 u)(x) \xi \geq \alpha_\sharp |\xi|^2 \qquad \forall\, x \in \Omega,\ \xi \in \mathbb{R}^2. \tag{1.6}$$

Remark 1.4 Note that under assumption (1.3) a sufficiently smooth solution of (1.1) is strictly convex if and only if $\Delta u \geq 0$ on Ω. This is the key motivation for the finite element method in this paper.

There are many numerical approaches to the Dirichlet boundary value problem of the Monge–Ampère equation (and related equations) in 2 and 3 spatial dimensions, with respect to different solution classes (classical solutions, Aleksandrov solutions [2] and viscosity solutions [54]). They include (i) geometric finite difference methods [63,66,68,69], (ii) monotone finite difference methods [7–9,39–41,48,50,67], (iii)

augmented Lagrangian and least-squares finite element methods [19,28–31], (iv) finite element methods based on the vanishing moment approach [3,35–37,57], (v) finite element methods based on L_2 projection [4,5,10,11,13,15,27,51,58–60], (vi) finite element methods based on a reformulation of the Monge–Ampère equation as a Hamilton–Jacobi–Bellman equation [14,34], and (v) two-scale methods [53,64,65]. Comprehensive reviews of the literature can be found in [33,61].

The method in this paper is also based on a nonlinear least-squares approach. It is different from the least-squares method of Dean and Glowinski [19,29,31] in that our least-squares problem is posed only on the finite element spaces and the discrete problems are solved purely as optimization problems.

The key ingredient in our method is an enhanced cubic Lagrange element with exotic degrees of freedom (dofs) that enables us to enforce the convexity of the finite element solutions, which then allows us to develop a simple error analysis based on existing results for second order elliptic problems in non-divergence form.

The rest of the paper is organized as follows. We introduce the enhanced cubic Lagrange element in Sect. 2 together with the discrete nonlinear least-squares problem. We then present *a priori* and *a posteriori* error analyses in Sect. 3 and numerical results in Sect. 4. We end with some concluding remarks in Sect. 5. We also put some of the details in three appendices so that the main flow of the presentation is not distracted. Appendix A contains the derivation of a stability result for elliptic problems in non-divergence form needed for the error analysis in Sect. 3. Details of the optimization algorithm that we use to solve the discrete problems are given in Appendix B. An algorithm that we use to check the elementwise convexity of the approximate solutions is outlined in Appendix C.

Throughout the paper we will use C to denote a generic positive constant independent of the mesh size.

2 The discrete problem

The discrete problem is a nonlinear least-squares problem with box constraints. It is based on an exotic finite element space whose degrees of freedom (dofs) can enforce the convexity of the solutions.

2.1 An enhanced cubic Lagrange finite element

We begin by introducing a new finite element where some of the degrees of freedom (dofs) are associated with nodes outside the element domain. Consequently the construction of the local basis requires information from outside an element. Below we will treat a polynomial on an element as the restriction of a polynomial on $\mathbb{R}^2$ and use the same notation to denote both. In other words, we will identify $P_k(\mathbb{R}^2)$ with the space $P_k(T)$ of polynomials of (total) degree $\leq k$ on a triangle T.

The construction of the finite element is based on the following lemma, where $\hat{T}$ is the reference simplex with vertices (0, 0), (1, 0) and (0, 1), and $\varphi_{\hat{T}} = x_1x_2(1-x_1-x_2)$ is the cubic bubble function that vanishes on the boundary of $\hat{T}$.

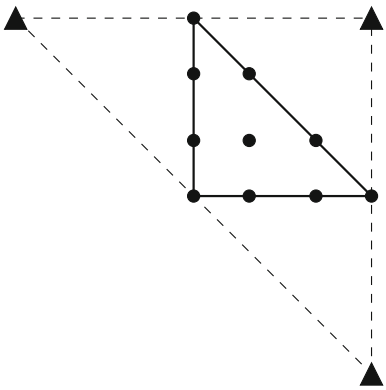

Fig. 1 dofs of the enhanced P_3 Lagrange element

Lemma 2.1 *A function* $v \in P_3(\hat{T}) \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$ *is uniquely determined by the 10 dofs of the standard cubic Lagrange finite element together with the values of* Δv *at the three points* $(1, 1)$, $(-1, 1)$ *and* $(1, -1)$ (*cf. Fig. 1*).

Proof Suppose v vanishes at the 9 vertex and edge nodes, then v belongs to $\langle \varphi_{\hat{T}} \rangle \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$. A direct calculation shows that 0 is the only polynomial in $\langle \varphi_{\hat{T}} \rangle \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$ that vanishes at the center of $\hat{T}$ and whose Laplacian also vanishes at the three points $(1, 1)$, $(-1, 1)$ and $(1, -1)$. □

Remark 2.2 The space $P_3(\hat{T}) \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$ and the 13 dofs in Lemma 2.1 do not define a finite element on $\hat{T}$ in the classical sense of Ciarlet in [23, page 78] because the shape functions are treated as functions defined globally on $\mathbb{R}^2$, and not as functions defined just on the element domain.

Remark 2.3 The vertices of the reference simplex are the midpoints of the edges of the triangle with vertices $(1, 1)$, $(-1, 1)$ and $(1, -1)$. For a general triangle T, the triangle whose midpoints are the vertices of T will be denoted by $T_\dagger$.

On an arbitrary triangle T, the space of shape functions of the enhanced cubic Lagrange element is $P_3(T) \oplus \varphi_T^2 P_1(T)$, where φ_T is the cubic bubble function that vanishes on the boundary of T. The dofs of $v \in P_3(T) \oplus \varphi_T^2 P_1(T)$ are (i) the values of v at the three vertices, (ii) the values of v at the two points that trisect each edge, (iii) the value of v at the center of T, and (iv) the values of $\mathrm{tr}(J_T^t D^2 v J_T)$ at the three vertices of $T_\dagger$, where $J_T \in \mathbb{R}^{2\times 2}$ is the Jacobian matrix of an affine map that maps the reference simplex to T.

Remark 2.4 The enhanced cubic Lagrange element is affine-equivalent (cf. [17,23]) by construction. The exotic dofs at the vertices of $T_\dagger$ are responsible for enforcing the elementwise convexity of the discrete solutions of (1.1).

2.2 The finite element space V_h

Let $\mathcal{T}_h$ be a quasi-uniform simplicial triangulation of Ω. A function v belongs to the finite element space $V_h \subset H^1(\Omega)$ if and only if (i) v belongs to $C(\bar{\Omega})$ and (ii) the restriction of v to $T \in \mathcal{T}_h$ belongs to $P_3(T) \oplus \varphi_T^2 P_1(T)$.

The (global) dofs of $v \in V_h$ are (i) the values of v at the vertices of $\mathcal{T}_h$, (ii) the values of v at the points that trisect the edges of $\mathcal{T}_h$, (iii) the values of v at the centers of the triangles in $\mathcal{T}_h$, and (iv) the values of $\mathrm{tr}(J_T^t D^2 v_T J_T)$ at the three vertices of $T_\dagger$ for each $T \in \mathcal{T}_h$, where v_T is the restriction of v to T and J_T is the Jacobian of an affine map that maps the reference simplex to T.

Remark 2.5 The dofs in (iii) and (iv) define the bubble functions in $\langle \varphi_T \rangle \oplus \varphi_T^2 P_1(T)$.

It follows from the extension theorems for Sobolev spaces (cf. [1, Chapter 5]) that the solution $u \in H^4(\Omega)$ of (1.1) can be extended to a strictly convex function in $H^4(\tilde{\Omega})$ where $\tilde{\Omega}$ is an open set that contains $\bar{\Omega}$ in its interior. We will denote this extension again by u. We assume that h is sufficiently small so that

$$T_\dagger \subset \tilde{\Omega} \qquad \forall\, T \in \mathcal{T}_h. \tag{2.1}$$

The nodal interpolant $\Pi_h u \in V_h$ is then defined by the condition that u and $\Pi_h u$ share the same global dofs mentioned above.

We will denote the piecewise Hessian operator by D_h^2, the set of the interior edges of $\mathcal{T}_h$ by $\mathcal{E}_h^i$, the length of an edge e by $|e|$, and the jump of the normal derivative of v across an (interior) edge by $[\![\partial v/\partial n]\!]$.

Lemma 2.6 *The following estimates are valid for* $\Pi_h u$*:*

$$\begin{aligned}
&\|u - \Pi_h u\|_{L_2(\Omega)} + h|u - \Pi_h u|_{H^1(\Omega)} + h\|u - \Pi_h u\|_{L_\infty(\Omega)} \\
&\quad + h^2 \|D_h^2(u - \Pi_h u)\|_{L_2(\Omega)} + h^4 \Big(\sum_{T \in \mathcal{T}_h} |D_h^2(\Pi_h u)|_{H^2(T)}^2 \Big)^{\frac{1}{2}} \le C h^4 |u|_{H^4(\tilde{\Omega})},
\end{aligned} \tag{2.2}$$

$$|u - \Pi_h u|_{W^{2,\infty}(T)} \le C h |u|_{H^4(\tilde{\Omega})} \qquad \forall\, T \in \mathcal{T}_h, \tag{2.3}$$

$$\begin{aligned}
&\sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial(\Pi_h u)/\partial n]\!]\|_{L_2(e)}^2 = \sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial(u - \Pi_h u)/\partial n]\!]\|_{L_2(e)}^2 \\
&\quad \le C h^4 |u|_{H^4(\tilde{\Omega})}^2,
\end{aligned} \tag{2.4}$$

$$\|D_h^2(\Pi_h u)\|_{L_2(\Omega)}^2 + \sum_{T \in \mathcal{T}_h} |D_h^2(\Pi_h u)|_{H^2(T)}^2 + \max_{T \in \mathcal{T}_h} |\Pi_h u|_{W^{2,\infty}(T)}^2 \le C \|u\|_{H^4(\tilde{\Omega})}^2. \tag{2.5}$$

Proof The estimates (2.2) and (2.3) follow from the invariance of cubic polynomials under the local nodal interpolation operator, the Bramble-Hilbert lemma [12] and scaling. The estimate (2.2) then implies the estimate (2.4) through the trace theorem

with scaling, and the estimate (2.5) follows from (2.2) and (2.3) through the triangle inequality and the Sobolev Embedding Theorem [1, Theorem 4.12]. □

In view of the estimate for $\|D_h^2(u - \Pi_h u)\|_{L_2(\Omega)}$ in (2.2) and the bound for $\max_{T\in\mathcal{T}_h} |\Pi_h u|_{W^{2,\infty}(T)}$ in (2.5), we immediately arrive at

$$\|\det D_h^2(\Pi_h u) - \det D^2 u\|_{L_2(\Omega)} \leq Ch^2, \tag{2.6}$$

where the positive constant C is independent of h.

Remark 2.7 All the estimates for $\Pi_h u$ are also valid for $\Pi_h\phi$. In particular we have

$$\|D_h^2(\phi - \Pi_h\phi)\|_{L_2(\Omega)} + \Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial(\Pi_h\phi)/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}} \leq Ch^2.$$

2.3 A nonlinear least-squares problem with box constraints

Let ϕ_h be the one dimensional cubic Lagrange interpolant of ϕ along $\partial\Omega$ and the (convex) subset L_h of V_h be defined by

$$\begin{aligned} L_h = \{v \in V_h :\ & v = \phi_h \text{ on } \partial\Omega \text{ and } \mathrm{tr}(J_T^t D^2 v_T J_T) \geq 0 \text{ at the vertices of } T\\ & \text{for every } T \in \mathcal{T}_h\}. \end{aligned} \tag{2.7}$$

Remark 2.8 The inequality constraints in the definition of L_h are motivated by the observation in Remark 1.4 and they are box constraints for the dofs of V_h introduced at the beginning of Sect. 2.2.

Remark 2.9 Note that $\phi_h = \Pi_h\phi = \Pi_h u$ on $\partial\Omega$ and hence $v = \Pi_h u$ on $\partial\Omega$ for all $v \in L_h$.

The discrete problem is to find

$$u_h = \operatorname*{argmin}_{v\in L_h} \mathcal{J}_h(v), \tag{2.8}$$

where the cost function $\mathcal{J}_h$ is defined by

$$\begin{aligned} \mathcal{J}_h(v) = {} & \frac{h^4}{2}\|D_h^2 v\|_{L_2(\Omega)}^2 + \frac{1}{2}\sum_{T\in\mathcal{T}_h} |\mathrm{tr}(J_T^t D^2 v_T J_T)|_{H^2(T)}^2\\ & + \frac{1}{2}\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2\\ & + \frac{1}{2}\|\det D_h^2 v - \psi\|_{L_2(\Omega)}^2, \end{aligned} \tag{2.9}$$

and v_T is the restriction of v to T. Note that the Frobenius norm of J_T satisfies

$$|J_T| \approx h. \tag{2.10}$$

We will use $\|\cdot\|_h$ to denote the mesh-dependent norm defined by

$$\|v\|_h^2 = \|D_h^2 v\|_{L_2(\Omega)}^2 + \sum_{e\in\mathcal{E}_h^i} |e|^{-1} \|[[\partial v/\partial n]]\|_{L_2(e)}^2. \tag{2.11}$$

Remark 2.10 The first two terms in the definition of $\mathcal{J}_h$ are regularization terms that are crucial for the well-posedness of the discrete problem and for enforcing the elementwise convexity of the discrete solutions. The third term is a penalty term that compensates for the fact that $V_h \not\subset H^2(\Omega)$. The last term is the least-squares term for (1.1a).

The solvability of (2.8) is justified by the following result.

Lemma 2.11 *The cost function $\mathcal{J}_h : L_h \longrightarrow [0, \infty)$ has a global minimizer.*

Proof According to the Poincaré-Friedrichs inequality for piecewise H^2 functions in [18], we have

$$\|v\|_{L_2(\Omega)} \le C\|v\|_h \qquad \forall\, v \in V_h \cap H_0^1(\Omega),$$

which implies (cf. Remark 2.9)

$$\|v - \Pi_h u\|_{L_2(\Omega)} \le C\|v - \Pi_h u\|_h \qquad \forall\, v \in L_h. \tag{2.12}$$

It follows from (2.2), (2.4), (2.5), (2.11) and (2.12) that $\|v\|_h$ (and hence $\mathcal{J}_h(v)$) approaches ∞ if v belongs to L_h and $\|v\|_{L_2(\Omega)}$ approaches ∞. □

3 Error analysis

We will show that any u_h satisfying (2.8) will converge to the solution u of (1.1) as $h \downarrow 0$ and the order of convergence is 2. Since our optimization algorithm does not guarantee that a global minimizer of $\mathcal{J}_h$ can be found, it is also useful to have an *a posteriori* error estimate that can demonstrate the convergence of our method numerically.

3.1 Some *a priori* bounds for u_h

Since $\Pi_h u$ belongs to L_h, it follows from (1.1a), (2.4)–(2.6) and (2.10) that

$$\begin{aligned}
&h^4\|D_h^2 u_h\|_{L_2(\Omega)}^2 + \sum_{T\in\mathcal{T}_h} |\mathrm{tr}(J_T^t D^2(u_h)_T J_T)|_{H^2(T)}^2 \\
&\quad + \sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial u_h/\partial n]\!]\|_{L_2(e)}^2 + \|\det D_h^2 u_h - \psi\|_{L_2(\Omega)}^2 \\
&= 2J_h(u_h) \\
&\le 2J_h(\Pi_h u) \\
&= h^4\|D_h^2(\Pi_h u)\|_{L_2(\Omega)}^2 + \sum_{T\in\mathcal{T}_h} |\mathrm{tr}(J_T^t D^2(\Pi_h u)_T J_T)|_{H^2(T)}^2 \\
&\quad + \sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial(\Pi_h u)/\partial n]\!]\|_{L_2(e)}^2 + \|\det D_h^2(\Pi_h u) - \det D^2 u\|_{L_2(\Omega)}^2 \\
&\le Ch^4. \qquad (3.1)
\end{aligned}$$

Consequently we have

$$\|D_h^2 u_h\|_{L_2(\Omega)} \le C, \qquad (3.2)$$

$$\Big(\sum_{T\in\mathcal{T}_h} |\mathrm{tr}(J_T^t D^2(u_h)_T J_T)|_{H^2(T)}^2\Big)^{\frac{1}{2}} \le Ch^2, \qquad (3.3)$$

$$\Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial u_h/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}} \le Ch^2, \qquad (3.4)$$

$$\|\det D_h^2 u_h - \psi\|_{L_2(\Omega)} \le Ch^2. \qquad (3.5)$$

Let $T \in \mathcal{T}_h$ be arbitrary and ψ_T be the P_1 Lagrange interpolant of ψ on T. We have, by a standard inverse estimate (cf. [17,23]),

$$\begin{aligned}
\|\det D_h^2 u_h - \psi\|_{L_\infty(T)} &\le \|\det D_h^2 u_h - \psi_T\|_{L_\infty(T)} + \|\psi - \psi_T\|_{L_\infty(T)} \\
&\le Ch^{-1}\|\det D_h^2 u_h - \psi_T\|_{L_2(T)} + \|\psi - \psi_T\|_{L_\infty(T)} \\
&\le Ch^{-1}\big(\|\det D_h^2 u_h - \psi\|_{L_2(T)} + \|\psi - \psi_T\|_{L_2(T)}\big) \\
&\quad + \|\psi - \psi_T\|_{L_\infty(T)},
\end{aligned}$$

which together with (3.5) and the assumption that $\psi \in H^2(\Omega)$ implies

$$\|\det D_h^2 u_h - \psi\|_{L_\infty(T)} \le Ch \qquad \forall\, T \in \mathcal{T}_h. \qquad (3.6)$$

3.2 The elementwise convexity of u_h

Let q_T be a polynomial defined on $T \in \mathcal{T}_h$. Recall that q_T is the restriction of a polynomial defined on $\mathbb{R}^2$ which is also denoted by q_T. We define $I_T q_T$ to be the restriction of $I_{T_\dagger} q_T$ on T, where $I_{T_\dagger}$ is the P_1 nodal interpolation operator associated with $T_\dagger$. Note that any linear polynomial on T is invariant under I_T, and according to (2.7),

$$I_T \mathrm{tr}(J_T^t D^2 v_T J_T) \geq 0 \quad \text{on } T \text{ for all } v \in L_h. \tag{3.7}$$

We have, by (3.3), a standard inverse estimate (cf. [17,23]) and the Bramble-Hilbert lemma,

$$\begin{aligned}
&\|\mathrm{tr}(J_T^t D^2 (u_h)_T J_T) - I_T \mathrm{tr}(J_T^t D^2 (u_h)_T J_T)\|_{L_\infty(T)} \\
&\quad \leq Ch^{-1} \|\mathrm{tr}(J_T^t D^2 (u_h)_T J_T) - I_T \mathrm{tr}(J_T^t D^2 (u_h)_T J_T)\|_{L_2(T)} \\
&\quad \leq Ch |\mathrm{tr}(J_T^t D^2 (u_h)_T J_T)|_{H^2(T)} \\
&\quad \leq Ch^3.
\end{aligned} \tag{3.8}$$

Lemma 3.1 *There exists a positive constant $\alpha_\flat$ independent of h such that, for h sufficiently small, we have*

$$\xi^t D_h^2 u_h \xi \geq \alpha_\flat |\xi|^2 \qquad \text{on all } T \in \mathcal{T}_h \text{ and for all } \xi \in \mathbb{R}^2, \tag{3.9}$$

or equivalently the minimum eigenvalue of $D_h^2 u_h$ is bounded below by a positive constant independent of h.

Proof Let $T \in \mathcal{T}_h$ be arbitrary. From (3.6), we have

$$\det D^2 (u_h)_T \geq \frac{1}{2} \min_{x \in \bar{\Omega}} \psi(x) > 0 \qquad \text{on } T$$

if h is sufficiently small. Consequently, in view of (2.10), we also have

$$h^{-4} \det(J_T^t D^2 (u_h)_T J_T) \geq \frac{\delta^2}{2} \min_{x \in \bar{\Omega}} \psi(x) \qquad \forall\, T \in \mathcal{T}_h, \tag{3.10}$$

where the positive constant $\delta \leq \min\{h^{-2} |\det J_T| : T \in \mathcal{T}_h\}$ is independent of h.

On the other hand, on each $T \in \mathcal{T}_h$, we have

$$\begin{aligned}
h^{-2} \mathrm{tr}(J_T^t D^2 (u_h)_T J_T) &\geq h^{-2} \big[\mathrm{tr}(J_T^t D^2 (u_h)_T J_T) - I_T \mathrm{tr}(J_T^t D^2 (u_h)_T J_T)\big] \\
&\geq -h^{-2} \|\mathrm{tr}(J_T^t D^2 (u_h)_T J_T) - I_T \mathrm{tr}(J_T^t D^2 (u_h)_T J_T)\|_{L_\infty(T)} \\
&\geq -Ch
\end{aligned} \tag{3.11}$$

by (3.7) and (3.8), where the positive constant C is independent of h.

We conclude from (3.10) and (3.11) that, for h sufficiently small, the minimum eigenvalue of $h^{-2}J_T^t D^2(u_h)_T J_T$ on the triangle T is bounded below by a positive constant independent of T and h, which implies that the same is true for $D_h^2 u_h$ because $\mathcal{T}_h$ is a quasi-uniform triangulation. □

Therefore, for h sufficiently small, u_h is a strictly convex polynomial on each $T \in \mathcal{T}_h$. Note that (3.9) is equivalent to

$$\xi^t \mathrm{Cof}(D_h^2 u_h)\xi \geq \alpha_\flat |\xi|^2 \qquad \text{on all } T \in \mathcal{T}_h \text{ and for all } \xi \in \mathbb{R}^2. \tag{3.12}$$

3.3 *A priori* error estimates

We have, by the fundamental theorem of calculus (cf. [38, Lemma A.1]),

$$\det D^2 u - \det D_h^2 u_h = \Big[\int_0^1 \mathrm{Cof}\, D_h^2(tu + (1-t)u_h)dt\Big] : D_h^2(u - u_h), \tag{3.13}$$

which is valid for all points in Ω except those on the edges of $\mathcal{T}_h$. Here and below we use the colon to denote the Frobenius inner product between matrices.

Let $A \in [L_\infty(\Omega)]^{2\times 2}$ be defined by the integral on the right-hand side of (3.13). For h sufficiently small, we have, by (1.6), (3.6) and (3.12),

$$\alpha|\xi|^2 \leq \xi^t A(x)\xi \leq \beta|\xi|^2 \qquad \forall\, \xi \in \mathbb{R}^2 \quad \text{and almost all } x \in \Omega, \tag{3.14}$$

where $0 < \alpha \leq \beta$ are constants independent of h.

The proof of the following lemma is given in Appendix A.

Lemma 3.2 *Under the condition* (3.14) *we have*

$$\begin{aligned}\|D_h^2(\zeta - v)\|_{L_2(\Omega)} \leq \frac{1}{1-\delta}\Big[&\alpha^{-1}\|A : D_h^2(\zeta - v)\|_{L_2(\Omega)} \\ &+ 2C_\dagger\Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}\Big]\end{aligned} \tag{3.15}$$

for all $\zeta \in H^2(\Omega) \cap H_0^1(\Omega)$ *and* $v \in V_h \cap H_0^1(\Omega)$, *where*

$$\delta = \frac{\beta - \alpha}{(\alpha^2 + \beta^2)^{\frac{1}{2}}} \tag{3.16}$$

and the positive constant $C_\dagger$ *only depends on the shape regularity of* $\mathcal{T}_h$.

We can now establish an *a priori* error estimate for u_h.

Theorem 3.3 *There exists a positive constant* C *independent of* h *such that*

$$\|u - u_h\|_h \leq Ch^2. \tag{3.17}$$

Proof We can assume h is sufficiently small so that (3.14) is satisfied. We begin with the triangle inequality

$$\|D_h^2(u-u_h)\|_{L_2(\Omega)} \leq \|D_h^2\big((u-\phi)-(u_h-\Pi_h\phi)\big)\|_{L_2(\Omega)} + \|D_h^2(\phi-\Pi_h\phi)\|_{L_2(\Omega)}. \tag{3.18}$$

Note that $u-\phi \in H^2(\Omega)\cap H_0^1(\Omega)$ by (1.1b) and $u_h-\Pi_h\phi \in V_h \cap H_0^1(\Omega)$ by (2.7) and Remark 2.9. Hence it follows from (3.15) that

$$\begin{aligned}
&\|D_h^2\big((u-\phi)-(u_h-\Pi_h\phi)\big)\|_{L_2(\Omega)}\\
&\quad\leq C\Big[\|A : D_h^2\big((u-\phi)-(u_h-\Pi_h\phi)\big)\|_{L_2(\Omega)}\\
&\qquad+\Big(\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial(u_h-\Pi_h\phi)/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}\Big]\\
&\quad\leq C\Big[\|A : D_h^2(u-u_h)\|_{L_2(\Omega)}+\Big(\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial u_h/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}\\
&\qquad+\|D_h^2(\phi-\Pi_h\phi)\|_{L_2(\Omega)}+\Big(\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial\Pi_h\phi/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}\Big].
\end{aligned}\tag{3.19}$$

Putting (1.1a), (3.13), (3.18) and (3.19) together, we have

$$\|D_h^2(u-u_h)\|_{L_2(\Omega)} \leq C\Big[\|\psi-\det D_h^2u_h\|_{L_2(\Omega)}+\Big(\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial u_h/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}} + \mathrm{Osc}(\phi)\Big], \tag{3.20}$$

where

$$\mathrm{Osc}(\phi) = \|D_h^2(\phi-\Pi_h\phi)\|_{L_2(\Omega)}+\Big(\sum_{e\in\mathcal{E}_h^i}|e|^{-1}\|[\![\partial\Pi_h\phi/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}. \tag{3.21}$$

It follows from Remark 2.7, (3.4), (3.5), (3.20) and (3.21) that

$$\|D_h^2(u-u_h)\|_{L_2(\Omega)} \leq Ch^2, \tag{3.22}$$

and then the estimate (3.17) follows from (2.11), (3.4) and (3.22). □

Remark 3.4 A careful tracking of the constant C in (3.17) shows that it takes the form of $C_*(\beta/\alpha)[(1/\alpha)+1]$, where α and β are the constants in (3.14) and the positive constant C_* depends only on u, ϕ and the shape regularity of $\mathcal{T}_h$.

According to the Poincaré-Friedrichs and Sobolev inequalities for piecewise H^2 functions (cf. [16,18]), we have

$$\|\zeta\|_{L_2(\Omega)} + |\zeta|_{H^1(\Omega)} + \|\zeta\|_{L_\infty(\Omega)} \leq C\|\zeta\|_h \tag{3.23}$$

for all $\zeta \in H_0^1(\Omega)$ that is piecewise H^2 with respect to $\mathcal{T}_h$. It follows from (2.2), (2.4), Remark 2.9, (2.11), (3.17), (3.23) and the triangle inequality that

$$\begin{aligned}
&\|u - u_h\|_{L_2(\Omega)} + |u - u_h|_{H^1(\Omega)} + \|u - u_h\|_{L_\infty(\Omega)} \\
&\quad \leq \|u - \Pi_h u\|_{L_2(\Omega)} + |u - \Pi_h u_h|_{H^1(\Omega)} + \|u - \Pi_h u_h\|_{L_\infty(\Omega)} \\
&\qquad + C(\|u - \Pi_h u\|_h + \|u - u_h\|_h) \\
&\quad \leq Ch^2.
\end{aligned} \tag{3.24}$$

Remark 3.5 Numerical results in Example 4.1 indicate that the convergence in $\|\cdot\|_{L_2(\Omega)}$, $|\cdot|_{H^1(\Omega)}$ and $\|\cdot\|_{L_\infty(\Omega)}$ is better than $O(h^2)$.

3.4 An *a posteriori* error estimate

Since finding a global minimizer of a nonconvex function is in general NP-hard, an optimization algorithm usually only produces an approximate stationary point $\tilde{u}_h$ of the cost function $\mathcal{J}_h$. Therefore we need more than the *a priori* error estimate (3.17) to ensure the convergence of the approximate solutions to the solution u of (1.1).

In the following discussion, it suffices to assume that u belonging to $W^{2,\infty}(\Omega)$ is strictly convex, i.e., (1.5) is satisfied almost everywhere in Ω.

Let $\tilde{u}_h \in L_h$ be elementwise strictly convex. Then the relation (3.13) is valid with u_h replaced by $\tilde{u}_h$, and we have, by Lemma 3.2 and the arguments in the proof of Theorem 3.3,

$$\begin{aligned}
\|D_h^2(u - \tilde{u}_h)\|_{L_2(\Omega)} \leq C\Big[&\|\det D_h^2\tilde{u}_h - \psi\|_{L_2(\Omega)} + \Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial\tilde{u}_h/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}} \\
&+ \mathrm{Osc}(\phi)\Big],
\end{aligned} \tag{3.25}$$

where the positive constant C depends only on the minimum and maximum eigenvalues of D^2u and $D_h^2\tilde{u}_h$ over Ω and the shape regularity of $\mathcal{T}_h$.

Therefore, after verifying the elementwise strict convexity of an approximate solution $\tilde{u}_h$ for (2.8), we can monitor the convergence of $\tilde{u}_h$ by evaluating the residual-based error estimator

$$\eta_h(\tilde{u}_h) = \|\det D_h^2\tilde{u}_h - \psi\|_{L_2(\Omega)} + \Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1}\|[\![\partial\tilde{u}_h/\partial n]\!]\|_{L_2(e)}^2\Big)^{\frac{1}{2}}. \tag{3.26}$$

According to (2.11) and (3.25), the estimator η_h is reliable for the error measured by the norm $\|\cdot\|_h$:

$$\|u - \tilde{u}_h\|_h \leq C\big(\eta_h(\tilde{u}_h) + \mathrm{Osc}(\phi)\big). \tag{3.27}$$

Moreover $\mathrm{Osc}(\phi)$ is $O(h^2)$ (cf. Remark 2.7).

In the other direction the obvious relations

$$|e|^{-1}\|[\![\partial\tilde{u}_h/\partial n]\!]\|^2_{L_2(e)} = |e|^{-1}\|[\![\partial(u-\tilde{u}_h)/\partial n]\!]\|^2_{L_2(e)} \tag{3.28}$$

and

$$\|\det D_h^2\tilde{u}_h - \psi\|_{L_2(T)} = \|\det D_h^2\tilde{u}_h - \det D^2 u\|_{L_2(T)} \tag{3.29}$$

imply that $\eta_h(\tilde{u}_h)$ is also locally efficient. Therefore we can use η_h to generate adaptive meshes when the solution of (1.1) is less smooth.

4 Numerical results

We have tested our method on three examples with known solutions. For each example we solve (2.7)–(2.9) by an active set algorithm (cf. Appendix B) that produces an approximate stationary point of the cost function in (2.9). The elementwise convexity of the approximate solutions are checked numerically by Algorithm C.1 in Appendix C.

For Example 4.2 and Example 4.3, where the known solutions do not belong to $H^4(\Omega)$, we have also solved (2.7)–(2.9) on adaptive meshes generated by the error estimator in (3.26) and a Dörfler marking strategy [32].

The relative errors of the approximate solution $\tilde{u}_h$ in various norms are defined by

$$e^r_{2,h} = \frac{\|u-\tilde{u}_h\|_h}{\|u\|_{H^2(\Omega)}}, \quad e^r_{1,h} = \frac{\|u-\tilde{u}_h\|_{H^1(\Omega)}}{\|u\|_{H^1(\Omega)}}, \quad e^r_{0,h} = \frac{\|u-\tilde{u}_h\|_{L^2(\Omega)}}{\|u\|_{L^2(\Omega)}}$$

and

$$e^r_{\infty,h} = \frac{\max\limits_{p\in\mathcal{V}_h}|u(p)-\tilde{u}_h(p)|}{\|u\|_{L^\infty(\Omega)}},$$

where $\mathcal{V}_h$ is the set of the vertices of the triangulation $\mathcal{T}_h$.

All the numerical experiments were carried out on a MacBook Pro laptop computer with a 2.8GHz Quad-Core Intel Core i7 processor and with 16GB 2133 MHz LPDDR3 memory. We use MATLAB (R2018b v.9.5.0) in our computations.

Example 4.1 This example is from [28], where Ω is the unit square $(0,1)^2$,

$$\psi(x) = (1+|x|^2)e^{|x|^2/2} \quad \text{and} \quad \phi(x) = e^{x^2/2}.$$

Table 1 Relative errors versus mesh size h and orders of convergence (Example 4.1)

h	$e^r_{2,h}$	Order	$e^r_{1,h}$	Order	$e^r_{0,h}$	Order	$e^r_{\infty,h}$	Order
2^0	1.2094e-1	–	3.1133e-2	–	6.7122e-3	–	1.1843e-2	–
2^{-1}	2.1221e-2	2.51	2.9131e-3	3.42	5.8949e-4	3.51	7.4311e-4	3.99
2^{-2}	5.5763e-3	1.93	3.2631e-4	3.16	3.0492e-5	4.27	3.8506e-5	4.27
2^{-3}	1.4039e-3	1.99	3.9515e-5	3.05	2.7445e-6	3.47	3.7144e-6	3.37
2^{-4}	3.5033e-4	2.00	4.8710e-6	3.02	2.9277e-7	3.23	3.7768e-7	3.30
2^{-5}	8.7449e-5	2.00	6.5347e-7	2.90	8.1151e-8	1.85	9.2308e-8	2.03

Table 2 Residual, Cost and CPU Time (Example 4.1)

h	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}
$\eta_h(\tilde{u}_h)$	8.4426e-1	1.9585e-1	4.7765e-2	1.1638e-2	2.7580e-3	6.5290e-4
Order	–	2.11	2.03	2.04	2.08	2.08
$\mathcal{J}_h(\tilde{u}_h)$	4.1353e-1	2.4864e-2	1.5275e-3	9.3814e-5	6.2186e-6	6.4513e-6
Order	–	4.01	4.02	4.03	3.92	−0.05
CPU Time (s)	2.8892e0	2.3075e0	2.9910e0	5.2907e0	1.4320e1	7.1900e1

The exact solution is $u = e^{x^2/2}$. The assumptions (1.2)–(1.4) are satisfied.

The errors of the approximate solutions $\tilde{u}_h$ obtained by the optimization algorithm on uniform meshes are reported in Table 1. The order of convergence for $e^r_{2,h}$ is 2, which agrees with the estimate in Theorem 3.3 for the solutions u_h of (2.8). The orders of convergence for $e^r_{1,h}$, $e^r_{0,h}$ and $e^r_{\infty,h}$ are higher.

The residual $\eta_h(\tilde{u}_h)$ and the cost $\mathcal{J}_h(\tilde{u}_h)$ are reported in Table 2, their behaviors agree with the estimates (3.1), (3.4) and (3.5) for the minimizer u_h. It is observed from the CPU times in Table 2 that a good approximate solution at $h = 2^{-2}$ was computed in 3 seconds.

We have verified that all the approximate solutions are elementwise strictly convex, and the reliability of the error estimator η_h can be observed by comparing $e^r_{2,h}$ in Table 1 and $\eta_h(\tilde{u}_h)$ in Table 2.

We have also solved the same problem on four other regular polygons (cf. Fig. 2), where the diameters of these polygons are 2.3660 (triangle), 1.6420 (pentagon), 1.5774 (hexagon) and 1.5307 (octagon).

It is observed from the convergence histories of $e^r_{2,h}$ and $e^r_{\infty,h}$ in Fig. 3 that the performance of our method is similar for all five polygons.

Example 4.2 This example is from [63], where $\Omega = (-1, 1)^2$,

$$\psi(x) = \begin{cases} 16 & \text{in } |x| \le 1/2 \\ 64 - 16|x|^{-1} & \text{in } 1/2 \le |x| \end{cases},$$

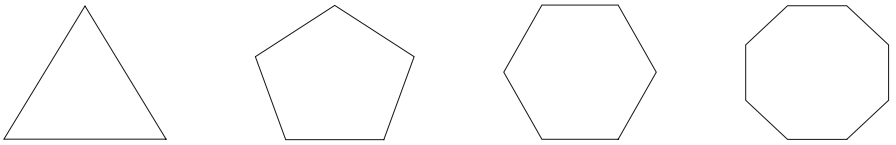

Fig. 2 Regular triangle, pentagon, hexagon and octagon

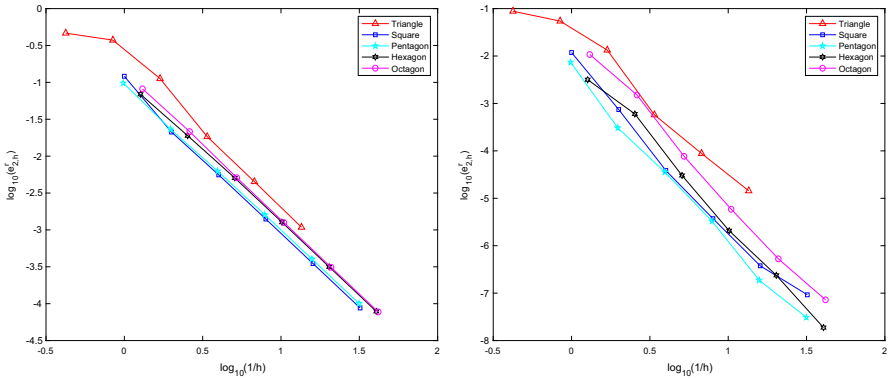

Fig. 3 Convergence histories of $e^r_{2,h}$ (left) and $e^r_{\infty,h}$ (right) for five regular polygons (Example 4.1)

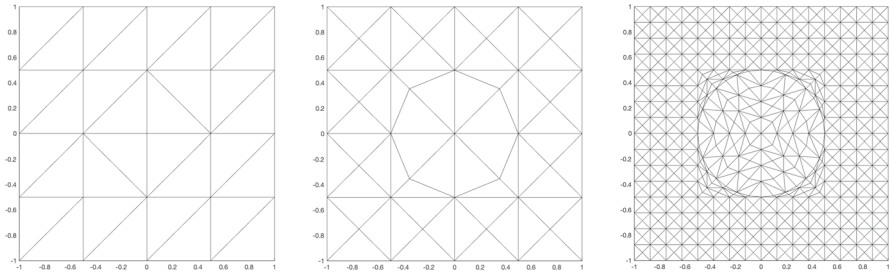

Fig. 4 Computational meshes (Example 4.2)

the exact solution is

$$u(x) = \begin{cases} 2|x|^2 & \text{in } |x| \le 1/2 \\ 2(|x| - 1/2)^2 + 2|x|^2 & \text{in } 1/2 \le |x| \end{cases},$$

and $\phi \in H^4(\Omega)$ equals u in a neighborhood of $\partial\Omega$. For this example, the function ψ is piecewise smooth and discontinuous along the circle defined by $|x| = 1/2$, and the Aleksandrov solution u is a piecewise smooth C^1 function.

The computational meshes are generated by a bisection procedure to fit the circle where ψ is discontinuous. The first two meshes and the final mesh are presented in Fig. 4.

Table 3 Relative errors versus mesh size h and orders of convergence (Example 4.2)

# of dofs	$e^r_{2,h}$	Order	$e^r_{1,h}$	Order	$e^r_{0,h}$	Order	$e^r_{\infty,h}$	Order
265	1.1610e-1	–	5.4486e-2	–	4.8750e-2	–	5.1245e-2	–
505	8.2598e-2	0.53	1.7304e-2	1.78	9.7645e-3	2.49	1.6850e-2	1.72
1009	4.3527e-2	0.93	9.2201e-3	0.91	6.9752e-3	0.49	7.7534e-3	1.12
1969	2.7050e-2	0.71	2.9208e-3	1.72	1.7075e-3	2.10	1.2552e-3	2.72
3937	1.5237e-2	0.83	8.5781e-4	1.77	4.5996e-4	1.89	5.6883e-4	1.14
7777	1.1519e-2	0.41	4.7382e-4	0.87	3.6609e-4	0.34	2.4460e-4	1.24

Table 4 Residual, Cost and CPU Time (Example 4.2)

# of dofs	265	505	1009	1969	3937	7777
$\eta_h(\tilde{u}_h)$	5.0972e0	3.1707e0	2.1286e0	1.7010e0	1.1319e0	9.3084e-1
Order	–	0.74	0.58	0.34	0.59	0.29
$\mathcal{J}_h(\tilde{u}_h)$	1.8770e1	4.5374e0	2.3356e0	1.2051e0	5.5931e-1	3.8182e-1
Order	–	2.20	0.96	0.99	1.11	0.56
CPU time (s)	1.7195e1	6.8276e0	2.1865e0	7.7616e0	6.3374e0	1.5600e2

We have verified that all the approximate solutions are elementwise strictly convex. The relative errors are reported in Table 3. The convergence of $e^r_{2,h}$ is of a reduced order ≈ 0.5, and the orders of convergence are higher for the lower order norms.

The residual $\eta_h(\tilde{u}_h)$, the cost $\mathcal{J}_h(\tilde{u}_h)$ and the CPU time are provided in Table 4. It is observed that a satisfactory approximate solution with 1009 dofs was computed in less than 3 seconds. The reliability estimate (3.27) is confirmed by comparing the values of $e^r_{2,h}$ and $\eta_h(\tilde{u}_h)$.

We also tested the performance of the *a posteriori* error estimator in (3.26) for this example. The convergence histories of $e^r_{2,h}$ and $e^r_{\infty,h}$ on bisection and adaptive meshes are shown in Fig. 5. The advantages of the adaptive meshes can be observed until round-off errors interfere at finer meshes.

The discrete solution on the final bisection mesh and the adaptive mesh with 3385 dofs are displayed in Fig. 6.

Example 4.3 This example is from [64], which is a modification of an example in [48]. The domain Ω is the unit square $= (0, 1)^2$,

$$\psi(x) = \max\left(1 - \frac{0.2}{|x - (1/2, 1/2)|}, 0\right),$$

the exact solution of this example is

$$u(x) = \frac{1}{2}(\max(|x - (1/2, 1/2)| - 0.2), 0)^2,$$

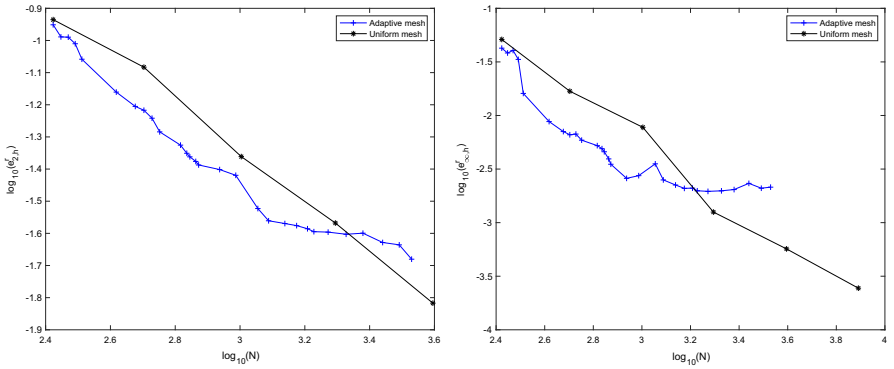

Fig. 5 Convergence histories of $e^r_{2,h}$ (left) and $e^r_{\infty,h}$ (right) on bisection and adaptive meshes (Example 4.2)

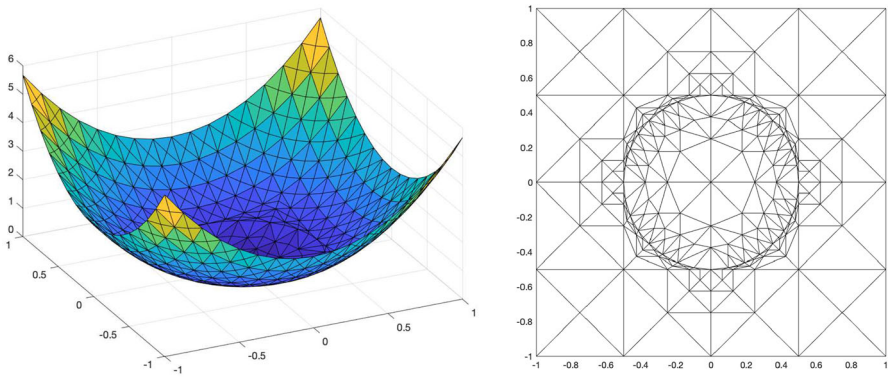

Fig. 6 (Left) graph of the computed solution on the final bisection mesh and (Right) adaptive mesh with 3385 dofs (Example 4.2)

and $\phi \in H^4(\Omega)$ equals u in a neighborhood of $\partial\Omega$. For this example the function ψ vanishes on the disc defined by $|x - (1/2, 1/2)| \leq 0.2$ and the Aleksandrov solution u is a piecewise smooth C^1 function.

We solved this problem by the nonlinear least-squares method on uniform meshes. The approximate solutions are elementwise strictly convex outside the disc where $\psi = 0$. The relative errors of $\tilde{u}_h$ are reported in Table 5. The order of convergence for $e^r_{2,h}$ is roughly 0.5 and the orders of convergence for $e^r_{1,h}$, $e^r_{0,h}$ and $e^r_{\infty,h}$ are better than 1. The discrete solution at $h = 2^{-5}$ can be found in Fig. 8.

The residual $\eta_h(\tilde{u}_h)$, the cost $\mathcal{J}_h(\tilde{u}_h)$ and the CPU time are provided in Table 6. Comparing $e^r_{2,h}$ in Table 5 and $\eta_h(\tilde{u}_h)$ in Table 6, one can see that the reliability estimate (3.27) is no longer valid because of the lack of strict convexity for both the discrete solutions and the continuous solution inside the disc where $\psi = 0$. It can also be seen that a satisfactory approximate solution at $h = 2^{-3}$ was computed in less than 8 seconds.

Table 5 Relative errors versus mesh size h and orders of convergence (Example 4.3)

h	$e^r_{2,h}$	Order	$e^r_{1,h}$	Order	$e^r_{0,h}$	Order	$e^r_{\infty,h}$	Order
2^0	3.4668e-1	–	1.7188e-1	–	1.4617e-1	–	1.5992e-1	–
2^{-1}	2.2904e-1	0.60	8.5079e-2	1.01	7.5088e-2	0.96	7.5322e-2	1.09
2^{-2}	1.4413e-1	0.67	3.0248e-2	1.49	2.2785e-2	1.72	2.0525e-2	1.88
2^{-3}	9.2617e-2	0.64	8.9737e-3	1.75	5.7416e-3	1.99	4.8416e-3	2.08
2^{-4}	7.9606e-2	0.22	5.3381e-3	0.75	3.0459e-3	0.91	2.5211e-3	0.94
2^{-5}	5.4643e-2	0.54	2.0729e-3	1.36	9.1258e-4	1.74	7.4776e-4	1.75

Table 6 Residual, Cost and CPU Time (Example 4.3)

h	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}
$\eta_h(\tilde{u}_h)$	1.3729e-1	7.5432e-2	2.9578e-2	1.1201e-2	4.3008e-3	1.5727e-3
Order	–	0.86	1.35	1.40	1.38	1.45
$\mathcal{J}_h(\tilde{u}_h)$	1.9072e-2	2.6887e-3	3.1878e-4	3.8834e-5	5.7440e-6	6.9137e-7
Order	–	2.83	3.08	3.04	2.76	3.05
CPU Time (s)	1.1282e0	1.9737e0	3.5095e0	7.7817e0	2.9395e1	2.4705e2

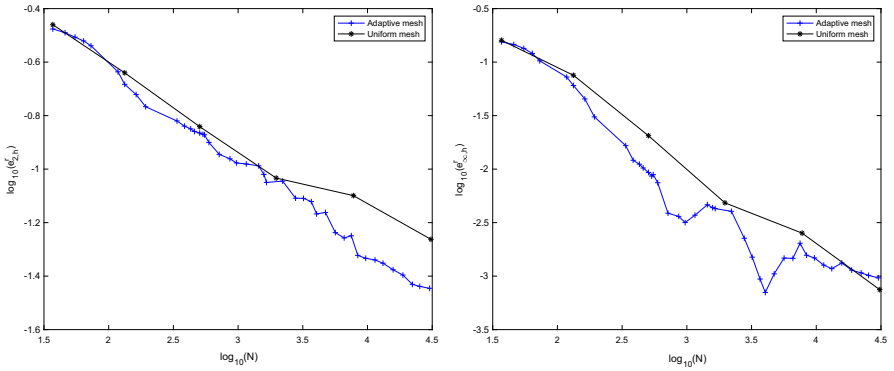

Fig. 7 Convergence histories of $e^r_{2,h}$ (left) $e^r_{\infty,h}$ (right) on uniform and adaptive meshes (Example 4.3)

We also tested the performance of the *a posteriori* residual error estimator in (3.26) for this example. The convergence histories of $e^r_{2,h}$ and $e^r_{\infty,h}$ on uniform and adaptive meshes are shown in Fig. 7. The advantage of adaptive meshes is observed. The adaptive mesh with 30187 dofs in Fig. 8 clearly captures the singularities of the exact solution along the circle defined by $|x - (1/2, 1/2)| = 0.2$.

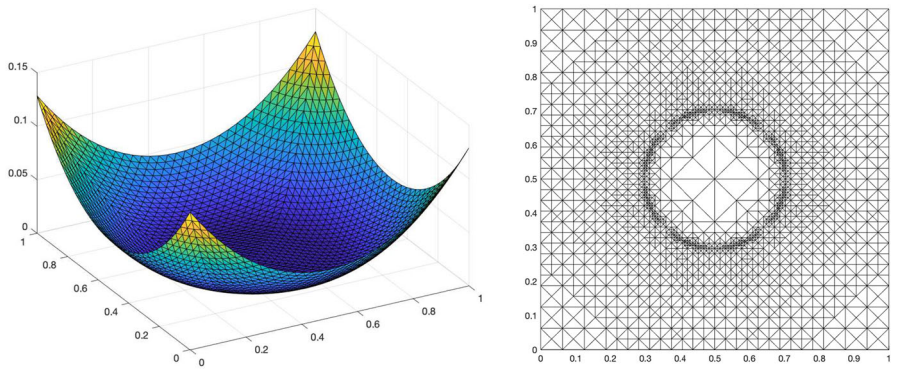

Fig. 8 (Left) graph of the computed solution on uniform mesh with $h = 2^{-5}$ and (Right) adaptive mesh with 30187 dofs (Example 4.3)

5 Concluding remarks

By going beyond the classical definition of a finite element, we are able to construct a C^0 interior penalty method for the Dirichlet boundary value problem of the Monge–Ampère equation, where the elementwise convexity of the approximate solutions can be enforced. This in turn enables us to use existing results for second order elliptic equations in nondivergence form to obtain both *a priori* and *a posteriori* error estimates. The *a posteriori* error estimate is a significant part of our method since it allows us to access the convergence of the approximate solutions generated by optimization algorithms that are not necessarily global minimizers.

The approach in this paper can be extended to smooth domains. We also note that convexity enforcing is useful for the problem of prescribed Gaussian curvature (cf. [38,42]) and the nonlinear least-squares approach can be applied to the Pucci equations (cf. [20,42]). These are some of our ongoing projects.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix A. The Proof of Lemma 3.2

We note that stability estimates similar to 3.2 and in more general settings can be found for example in [62,70]. We include a proof here for self-containedness.

The first ingredient is the Miranda–Talenti estimate

$$\|D^2\zeta\|_{L_2(\Omega)} \leq \|\Delta\zeta\|_{L_2(\Omega)} \qquad \forall\, \zeta \in H^2(\Omega) \cap H_0^1(\Omega) \tag{A.1}$$

that is valid for convex domains (cf. [56,71]). In the case of a polygon, the two sides of (A.1) are actually equal (cf. [43, Sect. 4.3]).

The second ingredient is the existence of an operator $E_h : V_h \cap H_0^1(\Omega) \longrightarrow H^2(\Omega) \cap H_0^1(\Omega)$ that satisfies

$$\|D_h^2(v - E_h v)\|_{L_2(\Omega)} \leq C_\dagger \Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1} \|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \Big)^{\frac{1}{2}} \qquad \forall\, v \in V_h \cap H_0^1(\Omega), \tag{A.2}$$

where $C_\dagger$ only depends on the shape regularity of $\mathcal{T}_h$ (cf. [62, Lemma 3]).

Remark A.1 The operator E_h in [62] is for the standard cubic Lagrange finite element space. The extension of E_h and (A.2) to the enhanced cubic Lagrange finite element space in Sect. 2.1 is straightforward since the additional bubble functions are already in $H^2(\Omega) \cap H_0^1(\Omega)$.

It follows from (A.1) and (A.2) that

$$\begin{aligned}
\|D_h^2(\zeta - v)\|_{L_2(\Omega)} &\leq \|D_h^2(\zeta - E_h v)\|_{L_2(\Omega)} + \|D_h^2(v - E_h v)\|_{L_2(\Omega)} \\
&\leq \|\Delta(\zeta - E_h v)\|_{L_2(\Omega)} + \|D_h^2(v - E_h v)\|_{L_2(\Omega)} \\
&\leq \|\Delta_h(\zeta - v)\|_{L_2(\Omega)} + 2\|D_h^2(v - E_h v)\|_{L_2(\Omega)} \\
&\leq \|\Delta_h(\zeta - v)\|_{L_2(\Omega)} + 2C_\dagger \Big(\sum_{e\in\mathcal{E}_h^i} |e|^{-1} \|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \Big)^{\frac{1}{2}}
\end{aligned} \tag{A.3}$$

for all $\zeta \in H^2(\Omega) \cap H_0^1(\Omega)$ and $v \in V_h \cap H_0^1(\Omega)$.

Following the treatment of second order linear elliptic equations in nondivergence form in [22,55,70], we introduce the function

$$\gamma(x) = \frac{A(x) : I}{A(x) : A(x)}$$

where I is the 2×2 identity matrix.

Note that $A(x) : I$ is the sum of the eigenvalues of $A(x)$ and $A(x) : A(x)$ is the sum of the squares of the eigenvalues of $A(x)$. Therefore we have, by (3.14), the following upper bound of $\gamma(x)$:

$$\gamma(x) \leq \max_{\alpha\leq\lambda_1,\lambda_2\leq\beta} \frac{\lambda_1 + \lambda_2}{\lambda_1^2 + \lambda_2^2} = \frac{1}{\alpha}, \tag{A.4}$$

and also the following Cordes condition (cf. [24]):

$$
\begin{aligned}
|\gamma(x)A(x) - I|^2 &= \gamma(x)^2 A(x) : A(x) - 2\gamma(x)(A(x) : I) + 2 \\
&= 2 - \frac{(A(x) : I)^2}{A(x) : A(x)} \\
&\le \max_{\alpha \le \lambda_1, \lambda_2 \le \beta} \frac{(\lambda_1 - \lambda_2)^2}{\lambda_1^2 + \lambda_2^2} = \frac{(\beta - \alpha)^2}{\alpha^2 + \beta^2} = \delta^2,
\end{aligned} \tag{A.5}
$$

where $\delta\ (< 1)$ is given by (3.16).

It follows from (A.3) and (A.5) that

$$
\begin{aligned}
&\int_\Omega [\gamma A : D_h^2(\zeta - v)]\Delta_h(\zeta - v)\, dx \\
&= \|\Delta_h(\zeta - v)\|_{L_2(\Omega)}^2 + \int_\Omega [\gamma A : D_h^2(\zeta - v) - \Delta_h(\zeta - v)]\Delta_h(\zeta - v) dx \\
&= \|\Delta_h(\zeta - v)\|_{L_2(\Omega)}^2 + \int_\Omega [(\gamma A - I) : D_h^2(\zeta - v)]\Delta_h(\zeta - v) dx \\
&\ge \|\Delta_h(\zeta - v)\|_{L_2(\Omega)}^2 - \delta \|D_h^2(\zeta - v)\|_{L_2(\Omega)} \|\Delta_h(\zeta - v)\|_{L_2(\Omega)} \\
&\ge (1 - \delta)\|\Delta_h(\zeta - v)\|_{L_2(\Omega)}^2 \\
&\quad - 2C_\dagger \delta \|\Delta_h(\zeta - v)\|_{L_2(\Omega)} \Big(\sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \Big)^{\frac{1}{2}},
\end{aligned}
$$

which together with (A.4) implies

$$
\begin{aligned}
\|\Delta_h(\zeta - v)\|_{L_2(\Omega)} \le \frac{\alpha^{-1}}{1 - \delta} &\|A : D_h^2(\zeta - v)\|_{L_2(\Omega)} \\
&+ 2C_\dagger \frac{\delta}{1 - \delta} \Big(\sum_{e \in \mathcal{E}_h^i} |e|^{-1} \|[\![\partial v/\partial n]\!]\|_{L_2(e)}^2 \Big)^{\frac{1}{2}}.
\end{aligned} \tag{A.6}
$$

Finally we arrive at (3.15) through (A.3) and (A.6).

Appendix B. An optimization algorithm

An active set algorithm is implemented to solve the bound constrained optimization

$$
\min \{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{B} \}, \tag{B.1}
$$

where $f : R^n \to R$ is twice continuously differentiable on the the set $\mathcal{B} = \{\mathbf{x} \in R^n : \mathbf{l} \le \mathbf{x} \le \mathbf{u}\}$. Our algorithm is based on the active set approach proposed in [45] for solving nonlinear optimization with bound constraints, which was further developed in [47] for handling more general polyhedral constrained optimization.

Here, we just very briefly explain the structure and convergence results of our algorithm. For more details on the theory of the algorithm, one may refer to references [45,47]. Our active set algorithm consists of two phases: a nonmonotone gradient projection phase and an unconstrained optimization phase, and a set of rules for switching between these two phases for achieving both global and fast local convergence. In particular, a projected cyclic Barzilai-Borwein (PCBB) algorithm is used in the gradient projection phase, where the line search direction at iteration $\mathbf{x}_k$ is generated by

$$\mathbf{d}_k = \mathrm{P}_{\mathcal{B}}(\mathbf{x}_k - \alpha_k \mathbf{g}_k) - \mathbf{x}_k. \tag{B.2}$$

Here, $\mathrm{P}_{\mathcal{B}}(\cdot)$ is the projection on the feasible region $\mathcal{B}$, α_k is the cyclic Barzilai-Borwein stepsize [25] and $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$. Along the search direction $\mathbf{d}_k$, an adaptive nonmonotone line search proposed in [26] is used to ensure global convergence. This PCBB algorithm of phase one is not only robust in the sense that it converges to a stationary point under mild assumptions, but also very effective for identifying the optimal active constraints where the components of the solution are on the boundary of $\mathcal{B}$.

However, the convergence rate of PCBB is often at best linear. Hence, to accelerate the convergence, a more efficient unconstrained optimization algorithm is used in phase two to optimize the objective function by fixing some components of variable $\mathbf{x}$ on the boundary of $\mathcal{B}$, that is

$$\min \{f(\mathbf{x}) : \mathbf{x}_{\mathcal{A}} = \mathbf{b}_{\mathcal{A}}\}. \tag{B.3}$$

Here, $\mathcal{A}$ is the active index set given by phase one and $\mathbf{b}_{\mathcal{A}}$ indicates the partial boundary of $\mathcal{B}$ where the components of $\mathbf{x}$ with index $\mathcal{A}$ are fixed. When one iteration of the phase two algorithm lies out of the feasible region $\mathcal{B}$, the set rules developed in the algorithm determine whether the algorithm will switch to the gradient projection phase or restart the unconstrained optimization phase by projecting the iterate back to $\mathcal{B}$. A limited memory nonlinear conjugate gradient method (L-CG_DESCENT) [46] is used to solve the subspace optimization (B.3) in phase two.

L-CG_DESCENT is a very efficient first-order method which has much more rapid convergence than most gradient descent methods, and maintains cheap iterations since only up to first-order information is used. However, when the optimization problem gets very ill-conditioned, which is often the case for a discrete optimization problem resulting from a finite difference method or a finite element method (such as the C^0 interior penalty method studied in this paper), slow convergence is often detected near the solution. Under this situation, in phase two we would switch L-CG_DESCENT to the second-order Newton's method, which is generally more expensive but often quickly leads to more accurate solutions. The convergence theories developed in [45,47] guarantee our active set algorithm converges at least to a stationary point of problem (B.1). Furthermore, the active set algorithm would asymptotically reduce the bound constrained optimization (B.1) to an unconstrained optimization (B.3) even when the problem is degenerate. Hence, fast local convergence would be expected by combining the more rapid convergence algorithms such as L-CG_DESCENT and Newton's method in the phase two optimization.

Appendix C. Elementwise convexity

Since the enhanced cubic Lagrange finite element is affine-equivalent, we can focus on the reference simplex. It is convenient to first consider the convexity of a tensor product polynomial on the unit square $(0, 1) \times (0, 1)$, for which we will need some explicit inverse estimates.

C.1 Explicit inverse estimates

Let $\mathbb{Q}_{m,n}$ be the space of tensor product polynomials spanned by $x_1^j x_2^k$ for $0 \le j \le m$ and $0 \le k \le n$. Given any $q \in \mathbb{Q}_{m,n}$, we can write

$$q(x) = \sum_{j=0}^{m} \sum_{k=0}^{n} a_{j,k}\, p_j(x_1)\, p_k(x_2)$$

where $p_0, p_1, \ldots$ are the Legendre polynomials.

Let $I = (-1, 1)$. It follows from the properties of the Legendre polynomials [52, (4.4.2), (4.5.1) and (4.5.2)] that

$$\begin{aligned}
\|q\|_{L_\infty(I\times I)} &\le \sum_{j=0}^{m} \sum_{k=0}^{n} |a_{j,k}| \\
&\le \Big(\sum_{j=0}^{m} \sum_{k=0}^{n} (j+\frac{1}{2})(k+\frac{1}{2}) \Big)^{\frac{1}{2}} \Big(\sum_{j=0}^{m} \sum_{k=0}^{n} a_{j,k}^2 (j+\frac{1}{2})^{-1} (k+\frac{1}{2})^{-1} \Big)^{\frac{1}{2}} \\
&= \frac{(m+1)(n+1)}{2} \|q\|_{L_2(I\times I)},
\end{aligned}$$

which, through scaling, implies

$$\|q\|_{L_\infty((0,1)\times(0,1))} \le (m+1)(n+1) \|q\|_{L_2((0,1)\times(0,1))} \qquad \forall\, q \in \mathbb{Q}_{m,n}. \tag{C.1}$$

C.2 Convexity on the unit square

Let $\hat{K}$ be the (closed) unit square $[0, 1] \times [0, 1]$ with vertices $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$, $\hat{p}_4$, and $Q_{\hat{K}}$ be the nodal interpolation operator for the Q_1 finite element on $\hat{K}$.

For any $q \in \mathbb{Q}_{m,n}$ and $x \in \hat{K}$, we have $Q_{\hat{K}} \Delta q - \Delta q \in \mathbb{Q}_{m,n}$ and therefore

$$\begin{aligned}
(Q_{\hat{K}} \Delta q)(x) &= [(Q_{\hat{K}} \Delta q)(x) - \Delta q(x)] + \Delta q(x) \\
&\le (m+1)(n+1) \| Q_{\hat{K}} \Delta q - \Delta q \|_{L_2(\hat{K})} + \Delta q(x)
\end{aligned}$$

by (C.1). It follows that

$$\begin{aligned}\Delta q(x) &\geq \min_{x\in\hat{K}}(Q_{\hat{K}}\Delta q)(x) - (m+1)(n+1)\|\Delta q - Q_{\hat{K}}\Delta q\|_{L_2(\hat{K})} \\ &= \min_{1\leq i\leq 4}\Delta q(\hat{p}_i) - (m+1)(n+1)\|\Delta q - Q_{\hat{K}}(\Delta q)\|_{L_2(\hat{K})} \qquad \forall\, x\in\hat{K}.\end{aligned} \tag{C.2}$$

Similarly, since $\det D^2 q - Q_{\hat{K}}(\det D^2 q) \in \mathbb{Q}_{2m-2,2n-2}$, we have

$$\begin{aligned}(\det D^2 q)(x) &\geq \min_{x\in\hat{K}}[Q_{\hat{K}}(\det D^2 q)](x) \\ &\quad - (2m-1)(2n-1)\|\det D^2 q - Q_{\hat{K}}(\det D^2 q)\|_{L_2(\hat{K})} \\ &= \min_{1\leq i\leq 4}(\det D^2 q)(\hat{p}_i) \\ &\quad - (2m-1)(2n-1)\|\det D^2 q - Q_{\hat{K}}(\det D^2 q)\|_{L_2(\hat{K})} \qquad \forall\, x\in\hat{K}.\end{aligned} \tag{C.3}$$

According to (C.2) and (C.3), if

$$\min_{1\leq i\leq 4}(\Delta q)(\hat{p}_i) - (m+1)(n+1)\|\Delta q - Q_{\hat{K}}(\Delta q)\|_{L_2(\hat{K})} > 0$$

and

$$\min_{1\leq i\leq 4}(\det D^2 q)(\hat{p}_i) - (2m-1)(2n-1)\|\det D^2 q - Q_{\hat{K}}(\det D^2 p)\|_{L_2(\hat{K})} > 0,$$

then

$$\Delta p > 0 \quad \text{and} \quad \det D^2 p > 0 \qquad \forall\, x\in\hat{K},$$

which implies that q is strictly convex on $\hat{K}$.

C.3 Convexity on the reference simplex

Let $q \in P_3(\hat{T}) \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$. Then $q \in \mathbb{Q}_{5,5}$ and we begin by computing (cf. (C.2) and (C.3))

$$L_{\Delta,\hat{K}} = \min_{1\leq i\leq 4}(\Delta q)(\hat{p}_i) - 36\|\Delta q - Q_{\hat{K}}(\Delta q)\|_{L_2(\hat{K})} \tag{C.4}$$

and

$$L_{\det,\hat{K}} = \min_{1\leq i\leq 4}(\det D^2 q)(\hat{p}_i) - 81\|\det D^2 q - Q_{\hat{K}}(\det D^2 q)\|_{L_2(\hat{K})}. \tag{C.5}$$

If $L_{\Delta,\hat{K}} > 0$ and $L_{\det,\hat{K}} > 0$, then q is strictly convex on $\hat{K}$ (and therefore also on $\hat{T}$). If this is not the case, then we divide $\hat{K}$ into four sub-squares and use scaled versions of (C.4) and (C.5) to check the convexity of q on the sub-squares whose intersection with $\hat{T}$ has a positive area. By repeating this procedure we arrive at the following algorithm for checking the convexity of $q \in P_3(\hat{T}) \oplus \varphi_{\hat{T}}^2 P_1(\hat{T})$ on $\hat{T}$.

Algorithm C.1 Let $\mathcal{R}_0 = \{\hat{K}\}$ and L be the maximum refinement level.

1. If $\mathcal{R}_l \neq \emptyset$ and $l \leq L$, compute for $K \in \mathcal{R}_l$ the quantities

$$L_{\Delta,K} = \min_{1\leq i\leq 4} (\Delta q)(p_{K,i}) - \frac{36}{h_l} \|\Delta q - Q_K(\Delta q)\|_{L_2(K)}$$

and

$$L_{\det,K} = \min_{1\leq i\leq 4} (\det D^2 q)(p_{K,i}) - \frac{81}{h_l} \|\det D^2 q - Q_K(\det D^2 q)\|_{L_2(K)},$$

where $p_{K,i}$ $(i = 1, 2, 3, 4)$ are the vertices of K, h_l is the width/height of the squares in $\mathcal{R}_l$, and Q_K is the nodal interpolation operator for the Q_1 element associated with K. Set

$$\mathcal{R}_l^{nc} = \{K \in \mathcal{R}_l : \ L_{\Delta,K} \leq 0 \text{ or } L_{\det,K} \leq 0\}.$$

Stop if $\mathcal{R}_l^{nc} = \emptyset$. The polynomial is strictly convex on $\hat{T}$.

2. If $\mathcal{R}_l^{nc} \neq \emptyset$, divide each $K \in \mathcal{R}_l^{nc}$ into four sub-squares K_j $(j = 1, 2, 3, 4)$ and define

$$\mathcal{R}_l^{nc,d} = \{K_j : \ K \in \mathcal{R}_l^{nc}, \ j = 1, 2, 3, 4\}.$$

Set

$$\mathcal{R}_{l+1} = \{K \in \mathcal{R}_l^{nc,d} : \ |K \cap \hat{T}| > 0\},$$

$h_l = \frac{1}{2} h_l$, $l = l + 1$ and go to 1.

Remark C.2 In our numerical experiments we were able to verify the elementwise strict convexity by observing that the Algorithm C.1 terminated before the refinement reached level 6.

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, 2nd edn. Academic Press, Amsterdam (2003)
2. Aleksandrov, A.D.: Dirichlet's problem for the equation Det $||z_{ij}|| = \varphi(z_1, \cdots, z_n, z, x_1, \cdots, x_n)$. I. Vestnik Leningrad. Univ. Ser. Mat. Meh. Astr. **13**, 5–24 (1958)
3. Awanou, G.: Spline element method for Monge-Ampère equations. BIT **55**, 625–646 (2015)

4. Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equations: classical solutions. IMA J. Numer. Anal. **35**, 1150–1166 (2015)
5. Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge-Ampère equation: Aleksandrov solutions. ESAIM Math. Model. Numer. Anal. **51**, 707–725 (2017)
6. Bakelman, I.J.: Convex Analysis and Nonlinear Geometric Elliptic Equations. Springer, Berlin (1994)
7. Barles, G., Souganidis, P.E.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptotic Anal. **4**, 271–283 (1991)
8. Benamou, J.-D., Collino, F., Mirebeau, J.-M.: Monotone and consistent discretization of the Monge-Ampère operator. Math. Comp. **85**, 2743–2775 (2016)
9. Benamou, J.-D., Froese, B.D., Oberman, A.M.: Two numerical methods for the elliptic Monge-Ampère equation. M2AN Math. Model. Numer. Anal. **44**, 737–758 (2010)
10. Böhmer, K.: On finite element methods for fully nonlinear elliptic equations of second order. SIAM J. Numer. Anal. **46**, 1212–1249 (2008)
11. Böhmer, K., Schaback, R.: A meshfree method for solving the Monge-Ampère equation. Numer. Algorithms, pp 539–551, (2019)
12. Bramble, J.H., Hilbert, S.R.: Estimation of linear functionals on Sobolev spaces with applications to Fourier transforms and spline interpolation. SIAM J. Numer. Anal. **7**, 113–124 (1970)
13. Brenner, S.C., Gudi, T., Neilan, M., Sung, L.-Y.: C^0 penalty methods for the fully nonlinear Monge-Ampère equation. Math. Comp. **80**, 1979–1995 (2011)
14. Brenner, S.C., Kawecki, E.L.: Adaptive C^0 interior penalty methods for Hamilton-Jacobi-Bellman equations with Cordes coefficients. J. Comp. Appl. Math. **1**, 1 (2020). https://doi.org/10.1016/j.cam.2020.11324
15. Brenner, S.C., Neilan, M.: Finite element approximations of the three dimensional Monge-Ampère equation. ESAIM Math. Model. Numer. Anal. **46**, 979–1001 (2012)
16. Brenner, S.C., Neilan, M., Reiser, A., Sung, L.-Y.: A C^0 interior penalty method for a von Kármán plate. Numer. Math. **135**, 803–832 (2017)
17. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods (Third Edition). Springer, New York (2008)
18. Brenner, S.C., Wang, K., Zhao, J.: Poincaré-Friedrichs inequalities for piecewise H^2 functions. Numer. Funct. Anal. Optim. **25**, 463–478 (2004)
19. Caboussat, A., Glowinski, R., Sorensen, D.C.: A least-squares method for the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in dimension two. ESAIM Control Optim. Calc. Var. **19**, 780–810 (2013)
20. Caffarelli, L.A., Cabré, X.: Fully Nonlinear Elliptic Equations. American Mathematical Society, Providence (1995)
21. Caffarelli, L.A., Nirenberg, L., Spruck, J.: The Dirichlet problem for nonlinear second-order elliptic equations. I. Monge-Ampère equation. Comm. Pure Appl. Math. **37**, 369–402 (1984)
22. Campanato, S.: A Cordes type condition for nonlinear nonvariational systems. Rend. Accad. Naz. Sci. XL Mem. Mat. (5) **13**, 307–321 (1989)
23. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
24. Cordes, H.O.: Über die erste Randwertaufgabe bei quasilinearen Differentialgleichungen zweiter Ordnung in mehr als zwei Variablen. Math. Ann. **131**, 278–312 (1956)
25. Dai, Y.-H., Hager, W.W., Schittkowski, K., Zhang, H.: The cyclic Barzilai-Borwein method for unconstrained optimization. IMA J. Numer. Anal. **26**, 604–627 (2006)
26. Dai, Y.-H., Zhang, H.: Adaptive two-point stepsize gradient algorithm. Numer. Algorithms **27**, 377–385 (2001)
27. Davydov, O., Saeed, A.: Numerical solution of fully nonlinear elliptic equations by Böhmer's method. J. Comput. Appl. Math. **254**, 43–54 (2013)
28. Dean, E.J., Glowinski, R.: Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach. C. R. Math. Acad. Sci. Paris **336**, 779–784 (2003)
29. Dean, E.J., Glowinski, R.: Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: a least-squares approach. C. R. Math. Acad. Sci. Paris **339**, 887–892 (2004)
30. Dean, E.J., Glowinski, R.: An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions. Electron. Trans. Numer. Anal. **22**, 71–96 (2006). ((**electronic**))

31. Dean, E.J., Glowinski, R.: Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. Comput. Methods Appl. Mech. Engrg. **195**, 1344–1386 (2006)
32. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
33. Feng, X., Glowinski, R., Neilan, M.: Recent developments in numerical methods for fully nonlinear second order partial differential equations. SIAM Rev. **55**, 205–267 (2013)
34. Feng, X., Jensen, M.: Convergent semi-Lagrangian methods for the Monge-Ampère equation on unstructured grids. SIAM J. Numer. Anal. **55**, 691–712 (2017)
35. Feng, X., Neilan, M.: Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method. SIAM J. Numer. Anal. **47**, 1226–1250 (2009)
36. Feng, X., Neilan, M.: Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. J. Sci. Comput. **38**, 74–98 (2009)
37. Feng, X., Neilan, M.: Analysis of Galerkin methods for the fully nonlinear Monge-Ampère equation. J. Sci. Comput. **47**, 303–327 (2011)
38. Figalli, A.: The Monge-Ampère equation and its applications. European Mathematical Society (EMS), Zürich (2017)
39. Froese, B.D., Oberman, A.M.: Convergent finite difference solvers for viscosity solutions of the elliptic Monge-Ampère equation in dimensions two and higher. SIAM J. Numer. Anal. **49**, 1692–1714 (2011)
40. Froese, B.D., Oberman, A.M.: Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. J. Comput. Phys. **230**, 818–834 (2011)
41. Froese, B.D., Oberman, A.M.: Convergent filtered schemes for the Monge-Ampère partial differential equation. SIAM J. Numer. Anal. **51**, 423–444 (2013)
42. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Classics in Mathematics. Springer, Berlin (2001)
43. Grisvard, P.: Elliptic Problems in Non Smooth Domains. Pitman, Boston (1985)
44. Gutiérrez, C.E.: The Monge-Ampère Equation. Birkhäuser Boston Inc., Boston (2001)
45. Hager, W.W., Zhang, H.: A new active set algorithm for box constrained optimization. SIAM J. Optim. **17**, 526–557 (2006)
46. Hager, W.W., Zhang, H.: The limited memory conjugate gradient method. SIAM J. Optim. **23**, 2150–2168 (2013)
47. Hager, W.W., Zhang, H.: An active set algorithm for nonlinear optimization with polyhedral constraints. Sci. China Math. **59**, 1525–1542 (2016)
48. Hamfeldt, B.F., Salvador, T.: Higher-order adaptive finite difference methods for fully nonlinear elliptic equations. J. Sci. Comput. **75**, 1282–1306 (2018)
49. Krylov, N.V.: Nonlinear Elliptic and Parabolic Equations of the Second Order. D. Reidel Publishing Co., Dordrecht (1987)
50. Kuo, H.J., Trudinger, N.S.: Discrete methods for fully nonlinear elliptic equations. SIAM J. Numer. Anal. **29**, 123–135 (1992)
51. Lakkis, O., Pryer, T.: A finite element method for nonlinear elliptic problems. SIAM J. Sci. Comput. **35**, A2025–A2045 (2013)
52. Lebedev, N.N.: Special Functions and Their Applications. Dover, New York (1972)
53. Li, W., Nochetto, R.H.: Optimal pointwise error estimates for two-scale methods for the Monge-Ampère equation. SIAM J. Numer. Anal. **56**, 1915–1941 (2018)
54. Lions, P.-L.: Fully nonlinear elliptic equations and applications. In *Nonlinear analysis, function spaces and applications, Vol. 2 (Písek, 1982)*, vol. 49, Teubner-Texte zur Math. pp 126–149. Teubner, Leipzig, (1982)
55. Maugeri, A., Palagachev, D.K., Softova, L.G.: Elliptic and Parabolic Equations with Discontinuous Coefficients. Wiley-VCH Verlag Berlin GmbH, Berlin (2000)
56. Miranda, C.: Su di una particolare equazione ellittica del secondo ordine a coefficienti discontinui. An. Şti. Univ. "Al. I. Cuza" Iaşi Secţ. I a Mat. (N.S.) **11B**, 209–215 (1965)
57. Neilan, M.: A nonconforming Morley finite element method for the fully nonlinear Monge-Ampère equation. Numer. Math. **115**, 371–394 (2010)
58. Neilan, M.: Quadratic finite element approximations of the Monge-Ampère equation. J. Sci. Comput. **54**, 200–226 (2013)
59. 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)

60. Neilan, M.: A unified analysis of three finite element methods for the Monge-Ampère equation. Electron. Trans. Numer. Anal. **41**, 262–288 (2014)
61. Neilan, M., Salgado, A.J., Zhang, W.: Numerical analysis of strongly nonlinear PDEs. Acta Numer. **26**, 137–303 (2017)
62. Neilan, M., Wu, M.: Discrete Miranda-Talenti estimates and applications to linear and nonlinear PDEs. J. Comput. Appl. Math. **356**, 358–376 (2019)
63. Neilan, M., Zhang, W.: Rates of convergence in W_p^2-norm for the Monge-Ampère equation. SIAM J. Numer. Anal. **56**, 3099–3120 (2018)
64. Nochetto, R.H., Ntogkas, D., Zhang, W.: Two-scale method for the Monge-Ampère equation: convergence to the viscosity solution. Math. Comp. **88**, 637–664 (2019)
65. Nochetto, R.H., Ntogkas, D., Zhang, W.: Two-scale method for the Monge-Ampère equation: pointwise error estimates. IMA J. Numer. Anal. **39**, 1085–1109 (2019)
66. Nochetto, R.H., Zhang, W.: Pointwise rates of convergence for the Oliker-Prussner method for the Monge-Ampère equation. Numer. Math. **141**, 253–288 (2019)
67. Oberman, A.M.: Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. Dis. Contin. Dyn. Syst. Ser. B **10**, 221–238 (2008)
68. Oliker, V.I., Prussner, L.D.: On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ and its discretizations. I. Numer. Math. **54**, 271–293 (1988)
69. Qiu, W., Tang, L.: A note on the Monge-Ampère type equations with general source terms. Math. Comp. **89**, 2675–2706 (2020)
70. Smears, I., Süli, E.: Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients. SIAM J. Numer. Anal. **51**, 2088–2106 (2013)
71. Talenti, G.: Sopra una classe di equazioni ellittiche a coefficienti misurabili. Ann. Mat. Pura Appl. **4**(69), 285–304 (1965)
72. Villani, C.: Topics in Optimal Transportation. American Mathematical Society, Providence (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.