Journal of Scientific Computing (2019) 79:1–47
https://doi.org/10.1007/s10915-018-0839-y

A Finite Element/Operator-Splitting Method for the Numerical Solution of the Two Dimensional Elliptic Monge–Ampère Equation

Roland Glowinski[1,2] · Hao Liu[3] · Shingyu Leung[3] · Jianliang Qian[4]

Received: 13 November 2017 / Revised: 17 September 2018 / Accepted: 20 September 2018 /
Published online: 27 September 2018

Abstract
We discuss in this article a novel method for the numerical solution of the two-dimensional elliptic Monge–Ampère equation. Our methodology relies on the combination of a time-discretization by operator-splitting with a mixed finite element based space approximation where one employs the same finite-dimensional spaces to approximate the unknown function and its three second order derivatives. A key ingredient of our approach is the reformulation of the Monge–Ampère equation as a nonlinear elliptic equation in divergence form, involving the cofactor matrix of the Hessian of the unknown function. With the above elliptic equation we associate an initial value problem that we discretize by operator-splitting. To enforce the pointwise positivity of the approximate Hessian we employ a hard thresholding based projection method. As shown by our numerical experiments, the resulting methodology is robust and can handle a large variety of triangulations ranging from uniform on rectangles to unstructured on domains with curved boundaries. For those cases where the solution is smooth and isotropic enough, we suggest also a two-stage method to improve the computational efficiency, the second stage being reminiscent of a Newton-like method. The methodology discussed in this article is able to handle domains with curved boundaries and unstructured meshes, using piecewise affine continuous approximations, while preserving optimal, or nearly optimal, convergence orders for the approximation error.

Keywords Fully nonlinear elliptic partial differential equations · Monge–Ampère equations · Operator-splitting method · Finite element approximations · Mixed finite element methods · Tychonoff regularization · Variational crimes

1 Introduction

Let Ω be a bounded domain of $\mathbb{R}^2$. The Dirichlet problem for the canonical Monge–Ampère equation reads as

$$\begin{cases} \det \mathbf{D}^2 u = f \text{ in } \Omega, \\ u = g \text{ on } \partial\Omega, \end{cases} \tag{1}$$

✉ Hao Liu
hliuas@ust.hk

Extended author information available on the last page of the article

$\mathbf{D}^2u$ being the Hessian matrix of u, that is $\mathbf{D}^2u = \begin{pmatrix} \frac{\partial^2 u}{\partial x_1^2} & \frac{\partial^2 u}{\partial x_1 \partial x_2} \\ \frac{\partial^2 u}{\partial x_1 \partial x_2} & \frac{\partial^2 u}{\partial x_2^2} \end{pmatrix}$, the functions f and g being given. If $f > 0$, problem (1) is a prototypical fully nonlinear elliptic boundary value problem. The existence and regularity properties of the solutions to fully nonlinear elliptic problems have been discussed in [2,10,25], a particular attention being given to the canonical Monge–Ampère equation in [31]. As shown in, e.g., [3,9,18,24,35], the Monge–Ampère equation has a wide range of applications, differential geometry, optimal transportation, physics and mechanics among them.

Starting with [39] various numerical methods have been developed for the numerical solution of fully nonlinear elliptic boundary problems, problem (1) being the most investigated by far. The fast multiplication of these methods during the last decade has made keeping track of all of them an almost impossible task. Several of them have been reported in [18], but a visit to Google Scholar has become a must to have a more complete view. Focusing on those approaches with which we have some familiarity, we will classify them roughly into two families. The methods of the first family treat a finite difference or finite element approximation of the equation under consideration (possibly coupled to a regularization procedure as done in [19,20]; see also [18]); such methods, and the iterative solution of the resulting discrete problems, are discussed in, e.g., [1,4–6,22,23,34,43]. Another approach is to reformulate the nonlinear elliptic problem as an optimization one; this can be done via least-squares or via the introduction of a well-chosen augmented Lagrangian algorithm. Such optimization based methods are discussed in e.g. [8,11–16,28–30,35].

The method discussed in this article concerns problem (1) specifically. It relies on: (i) An equivalent divergence formulation of (1). (ii) An initial value problem associated with a mixed variant of the above divergence formulation. (iii) The time discretization by operator-splitting of the above initial value problem. (iv) An eigenvalue projection algorithm to enforce the pointwise positive semi-definiteness of the approximate Hessian at each time step. (v) A mixed finite element implementation of the above methodology. Also for those problems where the solution of (1) is sufficiently smooth and isotropic, we propose a two-stage strategy to speed up the convergence: during the first stage, the dynamical system (flow) we consider is associated with $\partial u/\partial t$, while during the second stage we use $\partial(Su)/\partial t$, S being a well-chosen elliptic operator, giving to this second stage a Newton-like flavor.

As reported in, e.g., various chapters of [30] (see also the references therein), operator-splitting methods have a long history for providing efficient solution methods for a large variety of problems modeled by partial differential equations and inequalities. Fully nonlinear elliptic equations are no exceptions: To the best of our knowledge, the first publication making use of an operator-splitting method for the solution of a fully nonlinear elliptic problem is the celebrated article [3] by J.D. Benamou and Y. Brenier on the solution of the Monge–Kantorovich optimal mass transfer problem by the alternating direction method of multipliers (ADMM), a particular operator-splitting method. The solution of (1) by another ADMM algorithm is discussed in [11,13,14,16,29,30]. The numerical solution of the Monge–Ampère equation for domains with a curved boundary is one of the objectives of this article. Actually, one discussed in [6,21,32,33] efficient methods to achieve that goal. In particular, the methods in [21,32] are (among other things) mesh-free variants of the wide stencil finite difference methods developed by A. Oberman and his collaborators (see, e.g., [22,38] and the references therein). Solution methods for the Monge–Ampère equation on two and three-dimensional domains with curved boundaries are discussed also in [7,8]. Albeit relying on continuous piecewise affine approximations, like the methods in this article, the methods

discussed in [7,8] are more complicated to implement, and less robust and flexible than the quite simple and modular ones described in the following sections.

This article is organized as follows: In Sect. 2, we reformulate (1) as an elliptic problem in divergence form with which we associate an initial value problem whose time-discretization, by operator-splitting, is discussed in Sect. 3. The finite element implementation of the above operator-splitting method is discussed in Sect. 4. The two-stage strategy we mentioned above is discussed in Sect. 5. In Sect. 6 we present the results of numerical experiments. These results validate the methodology discussed in the preceding sections, including the two-stage strategy described in Sect. 5; they also show that problem (1) is solvable on domains with curved boundaries, using piecewise affine approximations associated with unstructured triangulations.

The main goal of this article was to access the possibility of solving a large variety of test problems, some of them quite singular, using continuous piecewise affine finite element approximations, associated with possibly unstructured meshes on domains with a curved boundary, while preserving good accuracy properties. The results of the many numerical experiments we have performed are promising and suggest further investigations and applications: among them, accelerating the convergence of our algorithm being an important objective, the two-stage method introduced in Sect. 5 (a Newton-like method) being just a first step in that direction. Actually, preliminary promising results suggest that the methodology introduced in the present article can be generalized to three dimensional problems, obstacle problems for the Monge–Ampère operator (like those discussed in [42]) and to the Gaussian curvature equation $\det \mathbf{D}^2 u = K(1+|\nabla u|^2)^{1+d/2}$, with $d = 2$ or 3, K (the given curvature) being a positive function.

2 A Divergence Formulation of Problem (1) and an Associated Initial Value Problem

Let us denote by $\mathrm{cof}\,(\mathbf{D}^2 u)$ the matrix-valued function $\begin{pmatrix} \frac{\partial^2 u}{\partial x_2^2} & -\frac{\partial^2 u}{\partial x_1 \partial x_2} \\ -\frac{\partial^2 u}{\partial x_1 \partial x_2} & \frac{\partial^2 u}{\partial x_1^2} \end{pmatrix}$. One can easily show that problem (1) is equivalent to

$$\begin{cases} -\nabla \cdot \left(\mathrm{cof}\left(\mathbf{D}^2 u\right) \nabla u\right) + 2f = 0 \text{ in } \Omega, \\ u = g \text{ on } \partial\Omega. \end{cases} \tag{2}$$

Similarly, one can also easily show that (2) characterizes formally u as being either a minimizer or a maximizer of the functional I over the space V_g where

$$I(v) = \int_\Omega \left(\mathrm{cof}\left(\mathbf{D}^2 v\right)\right) \nabla v \cdot \nabla v \, dx + 6 \int_\Omega f v dx \text{ and } V_g = \{v | v \text{ smooth}, v = g \text{ on } \partial\Omega\}.$$

Assume that u is solution of (1), (2). Since the symmetric matrix-valued functions $\mathbf{D}^2 u$ and $\mathrm{cof}\,(\mathbf{D}^2 u)$ are either point-wise positive definite or negative definite in the neighborhood of u, one can easily show that functional I is either convex or concave in the neighborhood of u, justifying using a well-initialized descent type method to compute the solutions of (1), (2); this can be achieved via the time integration of an initial value problem associated with (2). To partly overcome the nonlinear coupling between u and its second order derivatives we introduce a matrix-valued function $\mathbf{p}$ verifying the linear relation $\mathbf{p} = \mathbf{D}^2 u$. Problem (2)

is clearly equivalent to the following system of partial differential equations:

$$\begin{cases} \begin{cases} -\nabla \cdot (\mathrm{cof}(\mathbf{p})\nabla u) + 2f = 0 \text{ in } \Omega, \\ u = g \text{ on } \partial\Omega, \end{cases} \\ \mathbf{p} - \mathbf{D}^2 u = \mathbf{0}. \end{cases} \tag{3}$$

To handle those situations where $\inf_{x\in\Omega} f(x) = 0$, or for the solution of obstacle problems for the Monge–Ampère operator (as those considered in [42]), we found that one is on the safe side if one considers the following variant of (3), obtained by regularization:

$$\begin{cases} \begin{cases} -\nabla \cdot \left[(\varepsilon \mathbf{I} + \mathrm{cof}(\mathbf{p}))\, \nabla u \right] + 2f = 0 \text{ in } \Omega, \\ u = g \text{ on } \partial\Omega, \end{cases} \\ \mathbf{p} - \mathbf{D}^2 u = \mathbf{0}, \end{cases} \tag{4}$$

ε being a small positive number (of the order of h^2 in practice, h being the space discretization step). In order to solve (4), we associate with it the following initial value problem (*flow* in the dynamical system terminology):

$$\begin{cases} \begin{cases} \frac{\partial u}{\partial t} - \nabla \cdot \left[(\varepsilon \mathbf{I} + \mathrm{cof}(\mathbf{p}))\, \nabla u \right] + 2f = 0 \text{ in } \Omega \times (0, +\infty), \\ u = g \text{ on } \partial\Omega \times (0, +\infty), \end{cases} \\ \frac{\partial \mathbf{p}}{\partial t} + \gamma \left(\mathbf{p} - \mathbf{D}^2 u \right) = \mathbf{0} \text{ in } \Omega \times (0, +\infty), \\ u(0) = u_0, \ \mathbf{p}(0) = \mathbf{p}_0, \end{cases} \tag{5}$$

with γ a positive constant [above and below, $\phi(t)$ denotes the function $x \to \phi(x, t)$].

Before discussing (in Sect. 3) the time discretization of problem (5), we will address two important issues, namely: (i) The choice of γ, and (ii) the choice of u_0 and $\mathbf{p}_0$. Concerning γ, the idea is to pick a value so that $\mathbf{p}(t)$ evolves in time roughly like $u(t)$. Taking advantage of the fact that $\mathbf{p}$ and $\mathrm{cof}(\mathbf{p})$ have the same eigenvalues, we suggest taking

$$\gamma = \beta \lambda_0 \left(\varepsilon + \sqrt{\alpha} \right),$$

where λ_0 is the smallest eigenvalue of operator $-\nabla^2$ in $H_0^1(\Omega)$, α is the lower bound of function f, and β is a constant of the order of 1. Assuming that we are looking for the convex solutions of (1), several possibilities (not exclusive of each other) do exist in order to force this convexity property. The simplest one is a proper choice of u_0 and $\mathbf{p}_0$ in (5). Following the discussion in [28,29], we suggest taking for u_0 the solution of the following Poisson problem

$$\nabla^2 u_0 = 2\lambda\sqrt{f} \text{ in } \Omega, \ u_0 = g \text{ on } \partial\Omega, \tag{6}$$

with $\lambda(> 0)$ of the order of 1. Concerning $\mathbf{p}_0$, an obvious choice is $\mathbf{p}_0 = \mathbf{D}^2 u_0$, a simpler alternative being $\mathbf{p}_0 = \lambda\sqrt{f}\,\mathbf{I}$.

Remark 1 A natural variant of (5) is obtained by replacing $\frac{\partial u}{\partial t}$ by $-\frac{\partial}{\partial t}\nabla^2 u$. We did not pursue in that direction for two main reasons: (i) $\frac{\partial u}{\partial t}$ is much better suited than $-\frac{\partial}{\partial t}\nabla^2 u$ for the numerical solution of obstacle problems for the Monge–Ampère operator (like those discussed in [42]), and (ii) the numerical experiments we performed showed that the *the two-stage method* discussed in Sect. 5 is as fast and, in general, more robust than the methodology based on $-\frac{\partial}{\partial t}\nabla^2 u$.

3 On the Time Discretization of System (5) by Operator-Splitting

The structure of system (5) suggests using operator-splitting for its time-discretization. Among the many possible operator-splitting schemes (see, e.g., [30] for further information on operator-splitting methods) we advocate the particular Lie scheme described below, where $\Delta t (> 0)$ is a time-discretization step and $t^n = n\Delta t$:

$$u^0 = u_0\,,\ \mathbf{p}^0 = \mathbf{p}_0. \tag{7}$$

For $n \geq 0$, $\{u^n, \mathbf{p}^n\} \to \left\{u^{n+1/2}, \mathbf{p}^{n+1/2}\right\} \to \left\{u^{n+1}, \mathbf{p}^{n+1}\right\}$ as follows
Fractional Step 1: Solve

$$\begin{cases} \frac{\partial u}{\partial t} - \nabla \cdot \left[\left(\varepsilon \mathbf{I} + \operatorname{cof}\left(\mathbf{p}^n\right) \right) \nabla u \right] + 2f = 0 \text{ in } \Omega \times \left(t^n, t^{n+1}\right), \\ u = g \text{ on } \partial\Omega \times \left(t^n, t^{n+1}\right), \\ \frac{\partial \mathbf{p}}{\partial t} = \mathbf{0} \text{ in } \Omega \times \left(t^n, t^{n+1}\right), \\ u\left(t^n\right) = u^n,\ \ \mathbf{p}\left(t^n\right) = \mathbf{p}^n, \end{cases} \tag{8}$$

and set $u^{n+1/2} = u(t^{n+1})\,,\ \mathbf{p}^{n+1/2} = \mathbf{p}^n$.
Fractional Step 2: Solve

$$\begin{cases} \frac{\partial u}{\partial t} = 0 \text{ in } \Omega \times \left(t^n, t^{n+1}\right), \\ \frac{\partial \mathbf{p}}{\partial t} + \gamma \mathbf{p} = \gamma \mathbf{D}^2 u^{n+1/2} \text{ in } \Omega \times \left(t^n, t^{n+1}\right), \\ u\left(t^n\right) = u^{n+1/2},\ \ \mathbf{p}\left(t^n\right) = \mathbf{p}^{n+1/2}, \end{cases} \tag{9}$$

and set

$$u^{n+1} = u^{n+1/2}, \mathbf{p}^{n+1} = P_+\left[\mathbf{p}\left(t^{n+1}\right)\right], \tag{10}$$

where in (10), P_+ denotes a projection operator (to be defined in Sect. 4.5) on the convex cone of the symmetric positive semi-definite 2×2 matrices, the projection being done pointwise. Scheme (7)–(10) is first-order accurate at most, and semi-constructive since we still have to solve the sub-initial value problems (8) and (9). There is no difficulty with (9) since it has a closed form solution. To time-discretize (8), we advocate just taking one step of the backward Euler scheme. The resulting scheme (of the *Markchuk–Yanenko* type) reads as follows (using a more compact notation):

$$u^0 = u_0,\ \ \mathbf{p}^0 = \mathbf{p}_0. \tag{11}$$

For $n \geq 0$, $\{u^n, \mathbf{p}^n\} \to \left\{u^{n+1}, \mathbf{p}^{n+1}\right\}$ as follows

$$\begin{cases} \frac{u^{n+1}-u^n}{\Delta t} - \nabla \cdot \left[\left(\varepsilon \mathbf{I} + \operatorname{cof}\left(\mathbf{p}^n\right) \right) \nabla u^{n+1} \right] + 2f = 0 \text{ in } \Omega, \\ u^{n+1} = g \text{ on } \partial\Omega, \end{cases} \tag{12}$$

$$\mathbf{p}^{n+1} = P_+\left[e^{-\gamma \Delta t}\mathbf{p}^n + \left(1 - e^{-\gamma \Delta t}\right)\mathbf{D}^2 u^{n+1}\right]. \tag{13}$$

If the matrix-valued function $\mathbf{p}^n$ is positive semi-definite, then (12) is a formally well-posed elliptic boundary value problem.

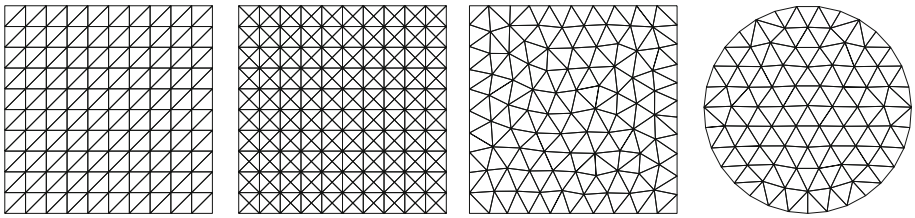

Fig. 1 Four meshes for two different domains used in the numerical experiments. **a** A regular mesh on the unit square. **b** A (highly) symmetric mesh on the unit square. **c** An isotropic unstructured mesh on the unit square. **d** An isotropic unstructured mesh on a half-unit disk

4 On the Finite Element Implementation of the Operator-Splitting Scheme

4.1 Synopsis

The equivalent divergence formulations (2) and (3) of problem (1) strongly suggest employing space approximations based on variational principles. To achieve such a goal, we are going to use finite element spaces consisting of functions which are globally continuous and piecewise affine on triangulations of Ω. As in, e.g., [8,29] (see also the references therein), we are going to use a mixed finite element method, relying basically on the same finite dimensional spaces to approximate u, its three second order derivatives, and the entries of the matrix-valued function $\mathbf{p}$.

4.2 The Basic Finite Element Spaces

We follow the presentations in [8,14,26,29]: Assuming that Ω is a polygonal domain of $\mathbb{R}^2$ (or has been approximated by such a domain), we introduce a family $(\mathcal{T}_h)_h$ of triangulations of Ω, like the ones in Fig. 1; usually, one denotes by h the length of the largest edge(s) of $\mathcal{T}_h$.

The first finite element space we introduce is the finite dimensional space V_h defined by

$$V_h = \left\{ v | v \in C^0\left(\bar{\Omega}\right), v|_T \in P_1, \forall T \in \mathcal{T}_h \right\}, \tag{14}$$

where P_1 is the space of the polynomials of two variables of degree ≤ 1. Let us denote by Σ_h the set of the vertices of the triangles of $\mathcal{T}_h$; we have then $\Sigma_h = \left\{ Q_j \right\}_{j=1}^{N_h}$. Next, we associate with each vertex Q_j the (shape) function w_j, uniquely defined by:

$$w_j \in V_h, w_j(Q_j) = 1, w_j(Q_k) = 0, \forall k = 1, \ldots, N_h, k \neq j.$$

The set $\mathcal{B}_h = \left\{ w_j \right\}_{j=1}^{N_h}$ is a vector basis of the space V_h; it verifies $v = \sum_{j=1}^{N_h} v(Q_j) w_j, \forall v \in V_h$, implying that $\dim V_h = N_h$. We observe that the support of the basis function w_j is the union of those triangles of $\mathcal{T}_h$ which have Q_j as a common vertex.

Assuming that $g \in C^0(\partial\Omega)$, we define V_{gh}, an affine subspace of V_h, by $V_{gh} = \left\{ v | v \in V_h, v(Q_j) = g(Q_j), \forall Q_j \in \Sigma_h \cap \partial\Omega \right\}$. Note that if $g = 0$, then $V_{gh} = V_{0h}$, where $V_{0h} = \{ v | v \in V_h, v = 0 \text{ on } \partial\Omega \} \left(= V_h \cap H_0^1(\Omega) \right)$.

In Sect. 4.3 below, we are going to address the approximation of $\partial^2 u / \partial x_1^2$, $\partial^2 u / \partial x_2^2$ and $\partial^2 u / \partial x_1 \partial x_2$, an important issue indeed.

4.3 Finite Element Approximation of the Three Second Order Derivatives

Unlike the collocation methods discussed in [7,8,29], the values taken on $\partial\Omega$ by the discrete second order derivatives affect the approximate solutions. From that point of view, enforcing (as done in [7,8,29]) these discrete derivatives to vanish on $\partial\Omega$ leads to large approximation errors on the solutions of problem (12), a consequence of the poor approximation of $\mathrm{cof}(\mathbf{p}^n)$ in the neighborhood of $\partial\Omega$. To overcome this difficulty, several approaches are available. We will focus on two of them. The starting point of the *first* one is to observe that the Divergence Theorem implies:

$$\begin{cases} \forall i,\, j = 1, 2, \forall v \in H^2(\Omega), \\ \int_\Omega \frac{\partial^2 v}{\partial x_i \partial x_j} w dx = -\frac{1}{2} \int_\Omega \left[\frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i} \right] dx \\ \forall w \in H_0^1(\Omega). \end{cases} \tag{15}$$

This suggests approximating $\frac{\partial^2 v}{\partial x_i \partial x_j}$ by $D^2_{ijh}(v)$ verifying

$$\begin{cases} \forall i,\, j = 1, 2, \forall v \in V_h,\, D^2_{ijh}(v) \in V_h \\ \int_\Omega D^2_{ijh}(v) w dx = -\frac{1}{2} \int_\Omega \left[\frac{\partial v}{\partial x_i} \frac{\partial w}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_i} \right] dx \\ \forall w \in V_{0h}. \end{cases} \tag{16}$$

The finite dimensional problem (16) being undetermined, additional relations are required in order to force solution uniqueness. We suggest imposing (approximately)

$$\frac{\partial}{\partial \mathbf{n}} D^2_{ijh}(v) = 0 \text{ on } \partial\Omega. \tag{17}$$

Among the various options available to impose (17), the one we selected reads as

$$\int_{\omega_k} \nabla D^2_{ijh}(v) \cdot \nabla w_k dx = 0, \forall k = N_{0h} + 1, \ldots, N_h, \tag{18}$$

where in (18): (i) $N_{0h} = \dim V_{0h}$. (ii) $\{Q_k\}_{k=N_{0h}+1}^{N_h} = \Sigma_h \cap \partial\Omega$, $\Sigma_h = \{Q_k\}_{k=1}^{N_h}$ being the set of the vertices of $\mathcal{T}_h$. (iii) w_k is the basis (shape) function of V_h associated with Q_k. (iv) $\omega_k =$ support of w_k is the union of those triangles of $\mathcal{T}_h$ which have Q_k as a common vertex; $|\omega_k|$ will denote the area of ω_k.

The rationale behind (18) is easy to understand: we have (Green's formula)

$$\int_{\partial\omega_k \cap \partial\Omega} \frac{\partial \psi}{\partial \mathbf{n}} w_k d(\partial\Omega) = \int_{\omega_k} \nabla^2 \psi w_k dx + \int_{\omega_k} \nabla\psi \cdot \nabla w_k dx, \forall \psi \in H^2(\Omega). \tag{19}$$

Suppose now that ψ is *harmonic* (that is $\nabla^2\psi = 0$), then relation (19) reduces to

$$\int_{\partial\omega_k \cap \partial\Omega} \frac{\partial \psi}{\partial \mathbf{n}} w_k d(\partial\Omega) = \int_{\omega_k} \nabla\psi \cdot \nabla w_k dx. \tag{20}$$

Albeit the function $D^2_{ijh}(v)$ is *piecewise harmonic* only (being piecewise affine), we used (20) to impose (17), explaining where (18) comes from. It is clear that replacing the above homogeneous Neumann boundary condition by (18) is a typical example of *variational crime* (in the sense of [44]). The numerical experiment results reported in Sect. 6 will validate a regularized variant of the approach we just described, based essentially on relations (16) and (18). Indeed, numerical experiments performed with (16), (18) show that the quality of the approximation deteriorates as $h \to 0$, unless Ω is a rectangle and that one uses regular

triangulations like the one shown in Fig. 1a. To overcome the above difficulty, we advocate (inspired by [7,8,29]) the simple (Tychonoff) regularization procedure where one replaces system (16), (18) by:

$$\begin{cases} \forall i,j=1,2, \forall v \in V_h, D^2_{ijh}(v) \in V_h, \\ C \sum_{T \in \mathcal{T}_h^k} |T| \int_T \nabla D^2_{ijh}(v) \cdot \nabla w_k dx + \int_{\omega_k} D^2_{ijh}(v) w_k dx \\ \quad = -\frac{1}{2} \int_{\omega_k} \left[\frac{\partial v}{\partial x_i} \frac{\partial w_k}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w_k}{\partial x_i} \right] dx, \forall k = 1, \ldots, N_{0h}, \\ \int_{\omega_k} \nabla D^2_{ijh}(v) \cdot \nabla w_k dx = 0, \forall k = N_{0h}+1, \ldots, N_h, \end{cases} \tag{21}$$

where in (21), C is a positive constant of the order of 1, $\mathcal{T}_h^k$ being the set of those triangles of $\mathcal{T}_h$ which have Q_k as a common vertex.

Remark 2 Suppose that one uses the *trapezoidal rule* to approximate the $L^2(\Omega)$-inner products in (21), then the above system reduces to

$$\begin{cases} \forall i,j=1,2, \forall v \in V_h, D^2_{ijh}(v) \in V_h, \\ C \sum_{T \in \mathcal{T}_h^k} |T| \int_T \nabla D^2_{ijh}(v) \cdot \nabla w_k dx + \frac{|\omega_k|}{3} D^2_{ijh}(Q_k) \\ \quad = -\frac{1}{2} \int_{\omega_k} \left[\frac{\partial v}{\partial x_i} \frac{\partial w_k}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w_k}{\partial x_i} \right] dx, \forall k = 1, \ldots, N_{0h}, \\ \int_{\omega_k} \nabla D^2_{ijh}(v) \cdot \nabla w_k dx = 0, \forall k = N_{0h}+1, \ldots, N_h. \end{cases} \tag{22}$$

System (22) being simpler than system (21) from a matrix point of view, it is the one we have used for those computations relying on (18). The practical use of (22) to define discrete second order derivatives requires this finite dimensional variational problem to be well-posed. We will prove it is true in the particular case where $\mathcal{T}_h$ being *isotropic* one replaces (22) by

$$\begin{cases} \forall i,j=1,2, \forall v \in V_h, D^2_{ijh}(v) \in V_h, \\ Ch^2 \int_{\omega_k} \nabla D^2_{ijh}(v) \cdot \nabla w_k dx + \frac{|\omega_k|}{3} D^2_{ijh}(v)(Q_k) \\ \quad = -\frac{1}{2} \int_{\omega_k} \left[\frac{\partial v}{\partial x_i} \frac{\partial w_k}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w_k}{\partial x_i} \right] dx, \forall k = 1, \ldots, N_{0h}, \\ \int_{\omega_k} \nabla D^2_{ijh}(v) \cdot \nabla w_k dx = 0, \forall k = N_{0h}+1, \ldots, N_h. \end{cases} \tag{23}$$

Theorem 1 *Suppose that in* (23)*, the function* v *is given in* V_h*. Then the associated linear system providing* $D^2_{ijh}(v)$ *has a unique solution.*

Proof The proof is very simple once we introduce the space $M_h (\subset V_h)$ defined by

$$M_h = \left\{ \mu \in V_h, \mu = \sum_{k=N_{0h}+1}^{N_h} \mu_k w_k, (\mu_k)_{k=N_{0h}+1}^{N_h} \in \mathbb{R}^{N_h - N_{0h}} \right\}. \tag{24}$$

We clearly have

$$V_h = V_{0h} \oplus M_h \text{ and } \mu \in M_h \Leftrightarrow \mu|_T = 0, \forall T \in \mathcal{T}_h \text{ such that } \partial T \cap \partial \Omega = \varnothing. \tag{25}$$

Next, we denote by $(\cdot,\cdot)_{0h}$ the inner-product over V_h defined by

$$(v, w)_{0h} = \frac{1}{3}\sum_{k=1}^{N_h} |\omega_k| v(Q_k) w(Q_k), \tag{26}$$

where $(Q_k)_{k=1}^{N_h} = \Sigma_h$. We have then $V_{0h}^{\perp} = M_h$ for the inner product defined by (26).

In order to prove that (23) is well-posed, it is sufficient to show that

$$\begin{cases} p \in V_h, \\ Ch^2 \int_{\omega_k} \nabla p \cdot \nabla w_k dx + \frac{|\omega_k|}{3} p(Q_k) = 0, \forall k = 1, \ldots, N_{0h}, \\ \int_{\omega_k} \nabla p \cdot \nabla w_k dx = 0, \forall k = N_{0h}+1, \ldots, N_h, \end{cases} \quad \Rightarrow p = 0. \tag{27}$$

The variational system in the left part of (27) is equivalent to

$$\begin{cases} p \in V_h, \\ Ch^2 \int_{\Omega} \nabla p \cdot \nabla v dx + (p, v)_{0h} = 0, \forall v \in V_{0h}, \\ \int_{\Omega} \nabla p \cdot \nabla \mu dx = 0, \forall \mu \in M_h. \end{cases} \tag{28}$$

It follows from (25) that $p = p_0 + p_1$ with $p_0 \in V_{0h}$ and $p_1 \in M_h$, the above decomposition being unique. We have $(p_0, p_1)_{0h} = 0$ and [from (28)] $\int_{\Omega} \nabla p \cdot \nabla p_1 dx = 0$. Taking $v = p_0$ in (28), the above two relations imply

$$Ch^2 \int_{\Omega} |\nabla p|^2 dx + (p_0, p_0)_{0h} = 0. \tag{29}$$

It follows from (29) that $p_0 = 0$ and $\nabla p = \nabla(p_0 + p_1) = \nabla p_1 = 0$. Function p_1 is thus a constant, but since $p_1(Q_k) = 0, \forall k = 1, \ldots, N_{0h}$, this constant is 0, implying $p = p_0 + p_1 = 0$. □

Remark 3 We recall that h is the length of the largest edge(s) of $\mathcal{T}_h$. Denote by $h_{\min}$ the length of the smallest edge(s) of $\mathcal{T}_h$. If the ratio $h/h_{\min} \approx 1$ (a situation we already considered in Remark 2), then one can advantageously replace the term $C \sum_{T \in \mathcal{T}_h^k} |T| \int_T \nabla D_{ijh}^2(v) \cdot \nabla w_k dx$ in (22) by the following simpler one $\varepsilon_1 \int_{\omega_k} \nabla D_{ijh}^2(v) \cdot \nabla w_k dx$, with ε_1 a positive parameter of the order of h^2.

Remark 4 Suppose that $\Omega = (0, 1)^2$ and that the triangulation $\mathcal{T}_h$ is of the same type as the one in Fig. 1a. Suppose also that $h = \frac{1}{I+1}$, I being an integer greater than 1. In this particular case, we have $\Sigma_h = \{Q_{ij} | Q_{ij} = (ih, jh), 0 \le i, j \le I+1\}$ and $\sum_{0h} = \{Q_{ij} | Q_{ij} = (ih, jh), 1 \le i, j \le I\}$, implying that $N_h = (I+2)^2$ and $N_{0h} = I^2$. Suppose that $C = 0$ in (22) or (23), then if $1 \le i, j \le I$, the relations giving $D_{klh}^2(v)(Q_{ij})(1 \le k, l \le 2)$ reduce to simple finite difference formulas, exact for the polynomials of degree ≤ 2. These formulas can be found in [29] pages 390 and 391. □

Although proving good convergence results, particularly if problem (1) has a smooth solution, the *variational crime* at the basis of relations (22) and (23) was leaving us uncomfortable. When using models (4), (5) to solve the Monge–Ampère equation (1), one can expect increased robustness (possibly at the expense of accuracy) if one uses smooth approximation of $\mathbf{p}$. Suppose that $\psi \in H^2(\Omega)$ and consider (with $\varepsilon > 0$) the following well-posed elliptic linear variational problem

$$\begin{cases} p_{ij}^{\varepsilon} \in H_0^1(\Omega), \\ \varepsilon \int_\Omega \nabla p_{ij}^{\varepsilon} \cdot \nabla \phi dx + \int_\Omega p_{ij}^{\varepsilon} \phi dx = -\frac{1}{2} \int_\Omega \left[\frac{\partial \psi}{\partial x_i} \frac{\partial \phi}{\partial x_j} + \frac{\partial \psi}{\partial x_j} \frac{\partial \phi}{\partial x_i} \right] dx, \forall \phi \in H_0^1(\Omega). \end{cases} \tag{30}$$

Function p_{ij}^{ε} verifies

$$\lim_{\varepsilon \to 0} p_{ij}^{\varepsilon} = \frac{\partial^2 \psi}{\partial x_i \partial x_j} \text{ in } L^2(\Omega), \tag{31}$$

and

$$\begin{cases} -\varepsilon \nabla^2 p_{ij}^{\varepsilon} + p_{ij}^{\varepsilon} = \frac{\partial^2 \psi}{\partial x_i \partial x_j} \text{ in } \Omega, \\ p_{ij}^{\varepsilon} = 0 \text{ on } \partial \Omega \end{cases} \tag{32}$$

(in the sense of distributions). From the *convexity* of Ω, it follows from (32) that $p_{ij}^{\varepsilon} \in H_0^1(\Omega) \cap H^2(\Omega) \left(\subset C^0(\bar{\Omega}) \right)$. Discrete variants of the above regularization method were applied in [7,8] to the solution of the Monge–Ampère equation in dimensions two and three. The numerical results reported in [7,8] show that the least-squares collocation method discussed there has no problem to accommodate the boundary condition $p_{ij}^{\varepsilon} = 0$ on $\partial \Omega$, unlike the divergence formulation (2) of the Monge–Ampère equation we employ in this article. In order to overcome this difficulty, we suggest introducing a correcting step, namely

$$\begin{cases} -\varepsilon \nabla^2 \tilde{p}_{ij}^{\varepsilon} + \tilde{p}_{ij}^{\varepsilon} = p_{ij}^{\varepsilon} \text{ in } \Omega, \\ \frac{\partial \tilde{p}_{ij}^{\varepsilon}}{\partial \mathbf{n}} = 0 \text{ on } \partial \Omega, \end{cases} \tag{33}$$

whose variational formulation reads as

$$\begin{cases} \tilde{p}_{ij}^{\varepsilon} \in H^1(\Omega), \\ \varepsilon \int_\Omega \nabla \tilde{p}_{ij}^{\varepsilon} \cdot \nabla \phi dx + \int_\Omega \tilde{p}_{ij}^{\varepsilon} \phi dx = \int_\Omega p_{ij}^{\varepsilon} \phi dx, \forall \phi \in H^1(\Omega). \end{cases} \tag{34}$$

Function $\tilde{p}_{ij}^{\varepsilon}$ verifies $\lim_{\varepsilon \to 0} \tilde{p}_{ij}^{\varepsilon} = \frac{\partial^2 \psi}{\partial x_i \partial x_j}$ in $L^2(\Omega)$, and $\tilde{p}_{ij}^{\varepsilon} \in H^4(\Omega)$. These properties are at the foundation of our *second approach* concerning the approximation of the second order derivatives. Suppose that $v \in V_h$; one proceeds as follows to compute the discrete analogue $D_{ijh}^2(v)$ of $\frac{\partial^2 v}{\partial x_i \partial x_j} (1 \leq i, j \leq 2)$:
Solve:

$$\begin{cases} p_{ij} \in V_{0h}, \\ C \sum_{T \in \mathcal{T}_h^k} |T| \int_T \nabla p_{ij} \cdot \nabla w_k dx + \frac{|\omega_k|}{3} p_{ij}(Q_k) = -\frac{1}{2} \int_{\omega_k} \left[\frac{\partial v}{\partial x_i} \frac{\partial w_k}{\partial x_j} + \frac{\partial v}{\partial x_j} \frac{\partial w_k}{\partial x_i} \right] dx, \\ \forall k = 1, \ldots, N_{0h}, \end{cases} \tag{35}$$

and then

$$\begin{cases} D_{ijh}^2(v) \in V_h, \\ C \sum_{T \in \mathcal{T}_h^k} |T| \int_T \nabla D_{ijh}^2(v) \cdot \nabla w_k dx + \frac{|\omega_k|}{3} D_{ijh}^2(v)(Q_k) = \frac{|\omega_k|}{3} p_{ij}(Q_k), \\ \forall k = 1, \ldots, N_h, \end{cases} \tag{36}$$

C being of the order of 1. Anticipating on the numerical results reported in Sect. 6, we would like to mention the following facts concerning the two approaches we presented concerning the approximation of the second order derivatives: Although resting on more solid mathematical foundations the *second approach* is less accurate than the *first one* if the Monge–Ampère problem under consideration has smooth enough solutions. However, the

second approach is *more robust* in the sense it can handle non-smooth situations more easily than the *first one* [typical examples in that direction being provided by problems (1) where f is not a function, but a positive measure (a Dirac's one for example)].

Remark 3 still applies for the approximation of the second order derivatives defined by relations (35), (36).

Remark 5 When implementing the double regularization approach associated with relations (32), (33), it makes sense (and is even tempting) to replace ε in (32) [resp. (33)] by η_1 of the order of $h^{2+\gamma}$ (resp., η_2 of the order of $h^{2-\gamma}$) with $\gamma(>0)$ not too large. Numerical experiments done with $\gamma = 1/4, 1/3$ and $1/2$ showed that for some test problems the introduction of a positive γ may improve convergence. However since we found examples where it has the opposite effect, we decided to stay with $\gamma = 0$.

4.4 Finite Element Implementation of the Operator-Splitting Scheme (11)–(13)

Let us recall that the set Σ_{0h} of the vertices of $\mathcal{T}_h$ interior to Ω has been ordered so that $\Sigma_{0h} = \{Q_k\}_{k=1}^{N_{0h}}$ and $\Sigma_h = \Sigma_{0h} \cup \{Q_k\}_{k=N_{0h}+1}^{N_h}$. In the sequel, we will denote by $\mathbf{Q}_h$ the space $\{\mathbf{q} | \mathbf{q} \in (V_h)^{2\times 2}, \mathbf{q} = \mathbf{q}^T\}$.

4.4.1 Implementation of Scheme (11)–(13)

A fully discrete analogue of scheme (11)–(13) reads as

$$u^0 = u_{0h} \left(\in V_{gh}\right), \mathbf{p}^0 = \mathbf{p}_{0h} \left(\in \mathbf{Q}_h\right). \tag{37}$$

For $n \geq 0$, $\{u^n, \mathbf{p}^n\} \rightarrow \left\{u^{n+1}, \mathbf{p}^{n+1}\right\}$ as follows:

Solve

$$\begin{cases} u^{n+1} \in V_{gh}, \\ \int_\Omega u^{n+1} v dx + \Delta t \int_\Omega \left(\varepsilon \mathbf{I} + \text{cof}\left(\mathbf{p}^n\right)\right) \nabla u^{n+1} \cdot \nabla v dx \\ \quad = \int_\Omega u^n v dx - 2\Delta t \int_\Omega f_h v dx, \forall v \in V_{0h}, \end{cases} \tag{38}$$

and compute $\mathbf{p}^{n+1}$ via

$$\begin{cases} \forall k = 1, \ldots, N_h, \text{ one has } \mathbf{p}^{n+\frac{1}{2}}(Q_k) = e^{-\gamma \Delta t} \mathbf{p}^n(Q_k) \\ \quad + \left(1 - e^{-\gamma \Delta t}\right) \begin{pmatrix} D^2_{11h}\left(u^{n+1}\right)(Q_k) & D^2_{12h}\left(u^{n+1}\right)(Q_k) \\ D^2_{12h}\left(u^{n+1}\right)(Q_k) & D^2_{22h}\left(u^{n+1}\right)(Q_k) \end{pmatrix}, \\ \mathbf{p}^{n+1}(Q_k) = P_+\left[\mathbf{p}^{n+1/2}(Q_k)\right], \end{cases} \tag{39}$$

where in (38), $f_h(>0)$ is a continuous approximation of f, and where in (39), the discrete second order derivatives of u^{n+1} are computed using the methods discussed in Sect. 4.3, operator P_+ being defined in Sect. 4.5. The initialization of scheme (37)–(39) and the solution of the discrete variational problem (38) will be discussed in Sects. 4.4.2 and 4.4.3, respectively.

4.4.2 Initialization of Scheme (37)–(39)

Our starting point will be problem (6) defining u_0. The most natural discrete analogue of problem (6) is given by

$$u_{0h} \in V_{gh}, \int_\Omega \nabla u_{0h} \cdot \nabla v dx = -2\lambda \int_\Omega \sqrt{f_h} v dx, \forall v \in V_{0h}. \tag{40}$$

From a practical point of view, we advocate using the trapezoidal rule to compute (approximately) the integral $\int_\Omega \sqrt{f_h} v dx$. We obtain then (using notation from previous sections):

$$\int_\Omega \sqrt{f_h} v dx \approx \frac{1}{3} \Sigma_{j=1}^{N_{0h}} |\omega_j| \sqrt{f_h(Q_j)} v(Q_j). \tag{41}$$

If Ω is a rectangle and the triangulation we employ is of the Fig. 1a type, we advocate using finite differences and a fast Poisson solver to compute u_{0h} from (6). The solution of (40) for more general domains and/or triangulations $\mathcal{T}_h$ will be (briefly) discussed in Sect. 4.4.3. Once u_{0h} has been computed, we use the methods discussed in Sect. 4.3 to define $\mathbf{p}_{0h}$ by

$$\mathbf{p}_{0h} = \sum_{k=1}^{N_h} P_+ \left[\begin{pmatrix} D_{11h}^2(u_{0h}) & D_{12h}^2(u_{0h}) \\ D_{12h}^2(u_{0h}) & D_{22h}^2(u_{0h}) \end{pmatrix} (Q_k) \right] w_k. \tag{42}$$

An alternative to (42) is to define $\mathbf{p}_{0h}$ by $\mathbf{p}_{0h} = \lambda \sum_{k=1}^{N_h} \sqrt{f_h(Q_k)} \mathbf{I}$, a discrete analogue of $\mathbf{p}_0 = \lambda \sqrt{f} \mathbf{I}$.

4.4.3 On the Solution of Problems (38) and Related Linear Variational Problems

What follows is fairly classical (see, for example Appendix 1 of [26] and Chapter 5 of [27], and the references therein). Problems (38) and (40) are particular cases of the following finite dimensional linear variational problem (of the Dirichlet type):

$$\psi \in V_{gh}, a \int_\Omega \psi \phi dx + \int_\Omega \mathbf{M} \nabla \psi \cdot \nabla \phi dx = \int_\Omega f^* \phi dx, \forall \phi \in V_{0h}, \tag{43}$$

where in (43), a is a non-negative constant, $\mathbf{M}$ is a piecewise affine uniformly positive definite symmetric matrix-valued function, f^* being a given continuous function. From the positivity of $\mathbf{M}$, problem (43) has a unique solution. We will return on the matrix $\mathbf{M}$ positivity issue in Sect. 4.5.

Using the trapezoidal rule to approximate the first and third integrals in (43), the above problem takes the following formulation:

$$\begin{cases} \psi \in V_{gh}, \forall \phi \in V_{0h}, \frac{a}{3} \sum_{l=1}^{N_{0h}} |\omega_l| \, \psi(Q_l) \phi(Q_l) \\ \quad + \int_\Omega \mathbf{M} \nabla \psi \cdot \nabla \phi dx = \frac{1}{3} \sum_{l=1}^{N_{0h}} |\omega_l| \, f^*(Q_l) \phi(Q_l). \end{cases} \tag{44}$$

Since the set $\{w_k\}_{k=1}^{N_{0h}}$ is a vector basis of V_{0h}, and $\psi = \sum_{l=1}^{N_{0h}} \psi(Q_l) + \sum_{l=N_{0h}+1}^{N_h} g(Q_l)$, it follows from (44) that the vector $\{\psi(Q_k)\}_{k=1}^{N_{0h}}$ is clearly the solution of the following [equivalent to (44)] linear system [where ψ_k denotes $\psi(Q_k)$]:

$$\begin{cases} \frac{a}{3} |\omega_k| \, \psi_k + \sum_{l=1}^{N_{0h}} \left(\int_{\omega_k \cap \omega_l} \mathbf{M} \nabla w_k \cdot \nabla w_l dx \right) \psi_l \\ \quad = \frac{1}{3} |\omega_k| \, f^*(Q_k) - \sum_{l=N_{0h}+1}^{N_h} \left(\int_{\omega_k \cap \omega_l} \mathbf{M} \nabla w_k \cdot \nabla w_l dx \right) g(Q_l), \\ k = 1, \ldots, N_{0h}. \end{cases} \tag{45}$$

Since, in practice, h is small compared to the diameter of Ω, the matrix associated with the linear system (45) is sparse. Moreover, this matrix is symmetric positive definite, these properties following from the uniform positivity of the symmetric matrix-valued function $\mathbf{M}$. One can solve thus the linear system (45) by a sparse Cholesky solver, or a diagonally preconditioned conjugate gradient algorithm. An user friendly alternative is to use one of those MATLAB (or other library) programs which decides by itself which linear solver

is the more appropriate to the linear system under consideration. Matrix $\mathbf{M}$ (resp., vector ∇w_k) being piecewise affine (resp., piecewise constant) the various integrals in (45) can be computed exactly by the trapezoidal rule.

The above comments concerning problems (38) and (40) apply also to the linear problems (22), (23) and (35), (36) we used in Sect. 4.3 to compute the discrete second order partial derivatives.

4.5 Enforcing the Local Positive Semi-definiteness of p by Eigenvalue Projection

In relations (10), (13), (39) and (42) we made use of P_+ a (kind of) projection operator mapping the space of the 2×2 symmetric real matrices onto the convex cone of the 2×2 positive semi-definite symmetric real matrices. Suppose that $\mathbf{A}$ is a 2×2 symmetric real matrix. From the symmetry of $\mathbf{A}$ there exists a 2×2 orthogonal matrix $\mathbf{S}$ such that $\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}$ with $\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, λ_1 and λ_2 being the two eigenvalues of matrix $\mathbf{A}$. Operator P_+ is defined by $P_+(\mathbf{A}) = \mathbf{S} \begin{pmatrix} \lambda_1^+ & 0 \\ 0 & \lambda_2^+ \end{pmatrix} \mathbf{S}^{-1}$, with $\lambda_i^+ = \max(0, \lambda_i)$, $\forall i = 1, 2$.

5 A Two-Stage Convergence Acceleration Strategy

In this section, we propose a simple strategy to speed up the convergence to steady states of scheme (11)–(13). Instead of (5), we consider the following initial value problem

$$\begin{cases} \begin{cases} -\frac{\partial(\nabla\cdot[(\varepsilon\mathbf{I}+\mathbf{M})\nabla u])}{\partial\tau} - \nabla\cdot\left[(\varepsilon\mathbf{I} + \mathrm{cof}(\mathbf{p}))\,\nabla u\right] + 2f = 0 \text{ in } \Omega \times (0, +\infty), \\ u = g \text{ on } \partial\Omega \times (0, +\infty), \end{cases} \\ \frac{\partial \mathbf{p}}{\partial \tau} + \gamma\left(\mathbf{p} - \mathbf{D}^2 u\right) = \mathbf{0} \text{ in } \Omega \times (0, +\infty), \\ u(0) = u_1,\ \mathbf{p}(0) = \mathbf{p}_1, \end{cases} \tag{46}$$

where τ is (as is t) an artificial time. In (46), $\mathbf{M}$ is a time independent matrix-valued function, which is supposed to be reasonably close to $\mathrm{cof}(\mathbf{D}^2 u)$, u being the convex solution of problem (1) (assuming that such a solution does exist). Scheme (11)–(13) can be easily modified to accommodate (46): we just have to replace $\frac{u^{n+1}-u^n}{\Delta t}$ in (12) by

$$-\nabla\cdot\left[(\varepsilon\mathbf{I} + \mathbf{M})\,\nabla\left(\frac{u^{n+1} - u^n}{\Delta\tau}\right)\right].$$

Compared to (12), the above preconditioning allows the use of a larger time step $\Delta\tau$, typically of the order of 1 if $\gamma \approx 1$, and u_1 and $\mathbf{p}_1$ are well-chosen. If convergence takes place, we expect that $\lim_{n\to+\infty} u^n = u$. To guarantee convergence, a proper choice of $\{u_1, \mathbf{p}_1\}$ is in order: a simple way to achieve that goal is to proceed as follows:

(i) Start iterating with scheme (11)–(13) for a small value of Δt, then stop time-stepping after a sufficiently (but not too) large number of time steps, denoting by $\{u_1, \mathbf{p}_1\}$ the pair produced by scheme (11)–(13) (further details will be given in Sect. 6).
(ii) Take for $\mathbf{M}$ the matrix-valued function $\mathrm{cof}(\mathbf{D}^2 u_1)$, and using $\{u_1, \mathbf{p}_1\}$ as initializer, switch until convergence to the variant of scheme (11)–(13) associated with the initial value-problem (46).

The reader will notice the *Newton-like* flavor of the above two-stage strategy. Its convergence is discussed in the PhD dissertation of the second author. It is shown in particular that

if $\mathbf{M}$ is close to $\mathrm{cof}(\mathbf{D}^2 u)$, where u is solution to problem (1), large values of $\Delta\tau$ can be used, leading to fast convergence properties.

6 Numerical Experiments

6.1 Generalities

In this section we will apply the computational methods discussed in Sects. 3–5, to the numerical solution of test problems, some of them without smooth solutions or no solution at all. The associated domains Ω will be the unit square $(0, 1)^2$ (as expected), and the disk of radius $1/2$, centered at $(1/2, 1/2)$. In Fig. 1a, b, c, d, we visualized finite element triangulations of these two domains. The mesh in Fig. 1a will be called a *regular* mesh, while the mesh on Fig. 1b will be called a *symmetric* mesh (it has five symmetries, while the mesh in Fig. 1a has three symmetries 'only'). We will say that the meshes in Fig. 1c, d are *unstructured*, although they are quite isotropic.

We use *DistMesh* [40] to generate the meshes shown in Fig. 1c, d. As expected (from the experiments reported in [8]), of all the triangulations we tested, those as the one in Fig. 1a are the only ones not requiring a regularization of the discrete second order derivatives in relation (22) or (23) to produce accurate results, when applied to the solution of test problems with smooth solutions. However, for poorly smooth or non-smooth problems, regularization improves significantly the performances of our methods, even for meshes of Fig. 1a type.

One of our goals with the first two test problems we are going to consider was to test the performances of the original one-stage algorithm (11)–(13) (actually of a finite element variant of it), and compare them with those of the two-stage algorithm discussed in Sect. 5. In all of our tests, without specification, we use $C = 1$ in (23), (35) and (36). We took $\varepsilon = h^2, \beta = 1/4$ and $\Delta t = 2h^2$ in the discrete analogue of algorithm (11)–(13). In the first stage of the two-stage algorithm we used $\|\mathbf{D}^2 u^n - \mathbf{p}^n\| < 1$ as the criterion to switch to stage 2 (above, we defined the matrix-function norm $\|\cdot\|$ by

$$\|\mathbf{S}\|^2 = \frac{1}{3}\sum_{k=1}^{N_h} |\omega_k| \|\mathbf{S}(Q_k)\|^2, \ \forall \mathbf{S} \in \mathbf{Q}_h,$$

the matrix norm of $\mathbf{S}(Q_k)$ being the Fröbenius one). In stage 2, we used $\Delta\tau = 8h$ and took, typically, $\|u^{n+1} - u^n\|_2 < 10^{-7}$, as stopping criterion ($\|\cdot\|_2$ being a trapezoidal rule based approximation of the canonical L_2-norm). When using only the discrete analog of scheme (11)–(13) we used a stopping criterion like the one used for stage 2, but with a smaller tolerance (h^3 seems to be a reasonable choice), in order to compensate that $\Delta t (= 2h^2)$ is quite small.

Remark 6 A smaller stopping criterion can be used for stage 2, but this is not necessary since (as visible on Fig. 2) the L_2 and L_∞ distances to the exact solution do not decrease anymore past some value of n; this appears clearly on Fig. 2b, c, d.

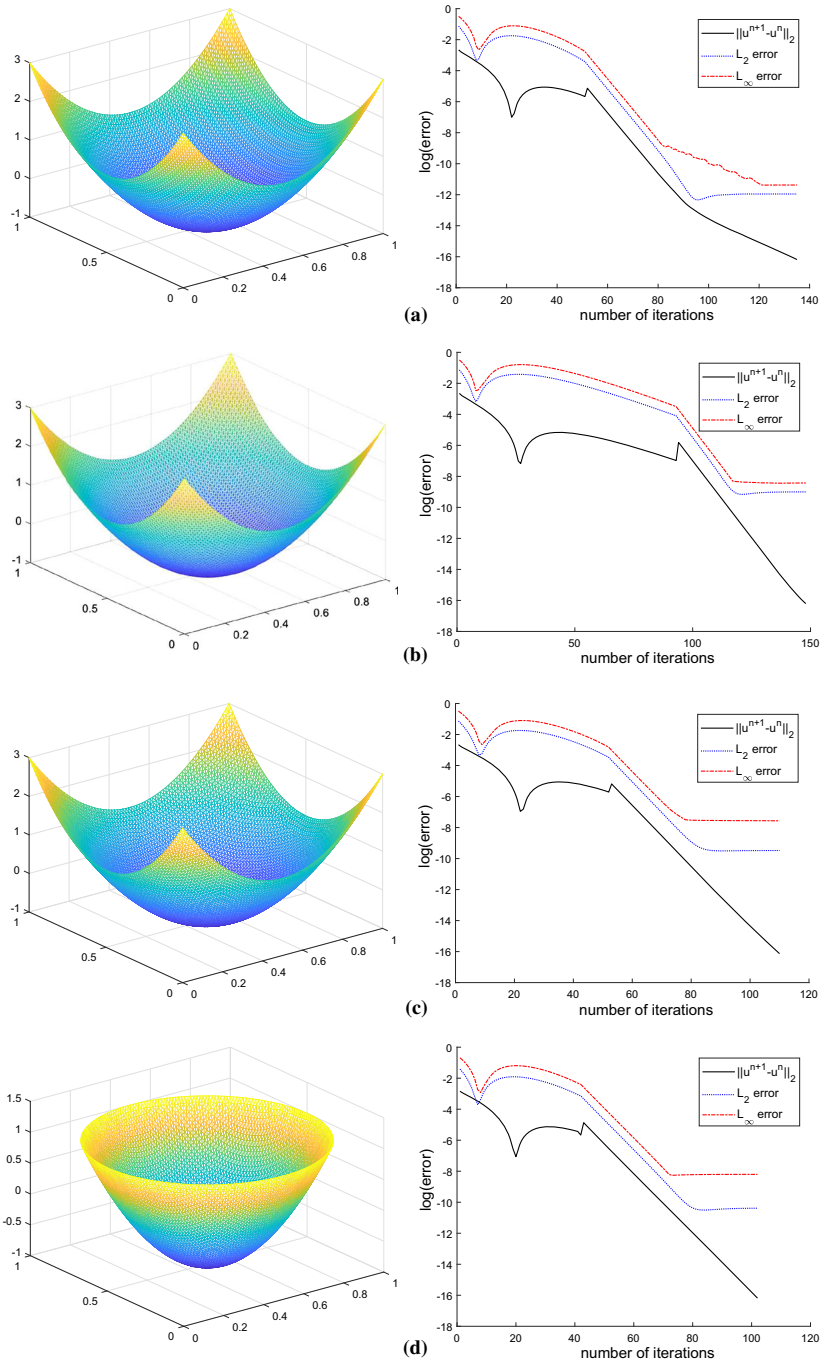

Fig. 2 (Test problem (47), $\alpha = 1$): graphs of the computed solutions obtained via the two-stage strategy with relations (23) and related convergence behavior for: **a** a unit square regular mesh ($h = 1/80$), **b** a unit square symmetric mesh ($h = 1/80$), **c** a unit square unstructured mesh ($h = 1/80$), **d** a half unit disk triangulation ($h = 1/80$)

6.2 A Polynomial Example

The *first test problem* we consider reads as

$$\begin{cases} \det \mathbf{D}^2 u = 256 \text{ in } \Omega, \\ u(x_1, x_2) = 8\left[\alpha\left(x_1 - \frac{1}{2}\right)^2 + \frac{1}{\alpha}\left(x_2 - \frac{1}{2}\right)^2\right] - 1, \forall (x_1, x_2) \in \partial\Omega, \end{cases} \tag{47}$$

with $\alpha \geq 1$, Ω being either the unit square $(0, 1)^2$ or the disk of radius 1/2 centered at $(1/2, 1/2)$. Clearly, the exact convex solution of problem (47) is given by

$$u(x_1, x_2) = 8\left[\alpha\left(x_1 - \frac{1}{2}\right)^2 + \frac{1}{\alpha}\left(x_2 - \frac{1}{2}\right)^2\right] - 1, \forall (x_1, x_2) \in \partial\Omega.$$

The first problem (47) we considered was the one associated with $\alpha = 1$, corresponding clearly to an isotropic solution. On Table 1, we have reported some of the results we obtained when applying the two-stage algorithm to the solution of problem (47), the second order derivatives being approximated by (23). For various values of h we have shown: (i) The number of iterations (time steps) necessary to achieve convergence. (ii) The value of $\|u^{n+1} - u^n\|_2$ at convergence. (iii) The L_2 and L_∞ approximation errors and the associated convergence rates (roughly of the order of two, which is generically optimal for the continuous piecewise affine approximations of the solution of a second order elliptic problem). In addition, we have reported on the 8th column of Table 1, a discrete L_2-norm of the gradient of the function $u^{n_c} - u$ where n_c denotes the number of time steps necessary to achieve convergence (n_c can be found in the second column of Table 1). For all tests with $h = 1/160$, we set $\varepsilon = h$ and the stopping criterion $\|\mathbf{D}^2 u^n - \mathbf{p}^n\| < 10$ for stage-one and stopping criterion $\|u^{n+1} - u^n\|_2 < 10^{-8}$ for stage-two.

Actually, further explanations are in order: indeed, we defined $\|\nabla(u^{n_c} - u)\|_2$ by

$$\left\|\nabla\left(u^{n_c} - u\right)\right\|_2 = \sqrt{\frac{1}{3}\sum_{k=1}^{N_h} |\omega_k| |\nabla_h u^{n_c}(Q_k) - \nabla u(Q_k)|^2}$$

with

$$\nabla_h u^{n_c}(Q_k) = \frac{1}{|\omega_k|}\sum_{T \in \mathcal{T}_h} |T| \left(\nabla u^{n_c}\right)|_T,$$

$\mathcal{T}_h^k$ being, as in Sect. 4.3, the set of those triangles of $\mathcal{T}_h$ which have Q_k as a common vertex. The (classical) *averaging* procedure associated with the definition of $\nabla_h u^{n_c}(Q_k)$ acts as a low pass filter, explaining the higher than one convergence rates reported in the 9^{th} column of Table 1. Let us define by $A_{Tq}, q = 1, 2$ and 3, the three vertices of triangle T; defining $\|\nabla(u^{n_c} - u)\|_2$ by

$$\left\|\nabla\left(u^{n_c} - u\right)\right\|_2 = \sqrt{\frac{1}{3}\sum_{T \in \mathcal{T}_h} |T| \sum_{q=1}^{3} \left|\nabla u^{n_c}|_T - \nabla u\left(A_{Tq}\right)\right|^2}$$

would lead to rates of convergence equal to one for the gradient approximation.

Table 2 is the variant of Table 1 associated with the second order derivative approximations defined by relations (35) and (36). The results reported in Table 2 suggest faster convergence of the iterative method, but lower orders of convergence for the approximation errors as $h \to 0$. Since our method converges faster and the second order derivatives are smoother

Table 1 (Test problem (47), $\alpha = 1$): (i) number of iterations needed for the convergence of the two-stage algorithm. (ii) Approximation errors and convergence rates for the solution and its gradient. (a) Regular meshes of the unit square. (b) Symmetric meshes of the unit square. (c) Unstructured isotropic meshes of the unit square. (d) Unstructured isotropic meshes of the half-unit disk. Relations (23) have been used to approximate the second order derivatives. (In (a), $h = \Delta x_1 = \Delta x_2$)

h	Iterations	$\|u^{n+1} - u^n\|_2$	L_2 error	Rate	L_∞ error	Rate	$\|\nabla(u^{n_c} - u)\|_2$	Rate
(a)								
1/10	20	4.37×10^{-8}	3.89×10^{-4}		7.06×10^{-4}		3.76×10^{-1}	
1/20	22	5.69×10^{-8}	1.00×10^{-4}	1.96	1.80×10^{-4}	1.97	1.33×10^{-1}	1.50
1/40	48	8.61×10^{-8}	2.55×10^{-5}	1.97	4.56×10^{-5}	1.98	4.71×10^{-2}	1.50
1/80	135	9.39×10^{-8}	6.43×10^{-6}	1.99	1.15×10^{-5}	1.99	1.67×10^{-2}	1.50
1/160	295	9.86×10^{-9}	1.61×10^{-6}	2.00	2.87×10^{-6}	2.00	5.89×10^{-3}	1.50
(b)								
1/10	17	4.49×10^{-8}	7.89×10^{-3}		1.40×10^{-2}		1.95×10^{-1}	
1/20	25	9.38×10^{-8}	1.98×10^{-3}	1.99	3.51×10^{-3}	2.00	6.90×10^{-2}	1.50
1/40	53	7.67×10^{-8}	4.94×10^{-4}	2.00	8.79×10^{-4}	2.00	2.44×10^{-2}	1.50
1/80	148	9.18×10^{-8}	1.23×10^{-4}	2.00	2.19×10^{-4}	2.00	8.61×10^{-3}	1.50
1/160	265	9.64×10^{-9}	3.09×10^{-5}	1.99	5.47×10^{-5}	2.00	3.04×10^{-3}	1.50
(c)								
1/10	21	6.16×10^{-8}	4.33×10^{-3}		1.50×10^{-2}		3.59×10^{-1}	
1/20	21	7.06×10^{-8}	9.68×10^{-4}	2.18	4.64×10^{-3}	1.69	1.17×10^{-1}	1.61
1/40	46	8.26×10^{-8}	2.26×10^{-4}	2.10	1.30×10^{-3}	1.84	4.29×10^{-2}	1.45
1/80	110	9.86×10^{-8}	7.61×10^{-5}	1.57	5.18×10^{-4}	1.32	1.68×10^{-2}	1.35
1/160	212	9.94×10^{-9}	5.20×10^{-5}	0.55	3.18×10^{-4}	0.70	1.26×10^{-2}	0.42
(d)								
1/10	20	5.15×10^{-8}	1.93×10^{-3}		1.23×10^{-2}		3.26×10^{-1}	
1/20	20	7.49×10^{-8}	6.47×10^{-4}	1.58	2.95×10^{-3}	2.06	9.44×10^{-2}	1.79
1/40	44	9.41×10^{-8}	1.16×10^{-4}	2.48	8.32×10^{-4}	1.83	3.32×10^{-2}	1.51
1/80	102	9.35×10^{-8}	3.09×10^{-5}	1.91	2.74×10^{-4}	1.60	1.10×10^{-2}	1.59
1/160	143	9.94×10^{-9}	3.12×10^{-5}	–	2.19×10^{-4}	0.32	1.03×10^{-2}	0.09

with relations (35) and (36), a very small stopping criterion is not necessary. Here we use $\|\mathbf{D}^2 u^n - \mathbf{p}^n\| < 10$ for stage-one and $\|u^{n+1} - u^n\|_2 < 10^{-4}$ for stage-two.

On Fig. 2, we have visualized the approximate solutions produced by the two-stage algorithm, the discrete second order derivatives being defined by relations (23). We clearly see that the L_2 and L_∞ approximation errors reach a plateau for n large enough, implying that the actual approximation errors have been reached.

In Table 3, we have reported the CPU time necessary for the two-stage algorithm to solve problem (47), with $\alpha = 1$, for the same types of meshes and space discretization steps that in Table 1, relations (23) being used to approximate the second order derivatives.

Remark 7 We would like to emphasize that the main goals of this article were: (i) Investigate the possibility of solving the Dirichlet problem for the Mong–Ampère equation using continuous piecewise affine finite element approximations, well-suited to domain with curved

Table 2 (Test problem (47), $\alpha = 1$): (i) number of iterations needed for the convergence of the two-stage algorithm. (ii) Approximation errors and convergence rates for the solution and its gradient. (a) Regular meshes of the unit square. (b) Symmetric meshes of the unit square. (c) Unstructured isotropic meshes of the unit square. (d) Unstructured isotropic meshes of the half-unit disk. Relations (35) and (36) have been used to approximate the second order derivatives. (In (a), $h = \Delta x_1 = \Delta x_2$)

h	Iterations	$\|u^{n+1} - u^n\|_2$	L_2 error	Rate	L_∞ error	Rate	$\|\nabla(u^{n_c} - u)\|_2$	Rate
(a)								
1/20	11	7.89×10^{-5}	2.51×10^{-1}		2.94×10^{-1}		1.65×10^{0}	
1/40	27	8.48×10^{-5}	1.28×10^{-1}	0.97	1.50×10^{-1}	0.97	1.16×10^{0}	0.51
1/80	62	9.90×10^{-5}	6.53×10^{-2}	0.97	7.67×10^{-2}	0.97	8.11×10^{-1}	0.52
1/160	182	9.32×10^{-5}	3.36×10^{-2}	0.96	3.90×10^{-2}	0.98	5.69×10^{-1}	0.51
(b)								
1/20	17	5.36×10^{-5}	2.24×10^{-1}		2.60×10^{-1}		1.55×10^{0}	
1/40	31	8.19×10^{-5}	1.15×10^{-1}	0.96	1.29×10^{-1}	1.01	1.06×10^{0}	0.55
1/80	81	8.99×10^{-5}	5.82×10^{-2}	0.98	6.43×10^{-2}	1.00	7.27×10^{-1}	0.54
1/160	265	9.17×10^{-5}	2.98×10^{-2}	0.97	3.22×10^{-2}	1.00	5.02×10^{-1}	0.53
(c)								
1/20	14	8.48×10^{-5}	2.48×10^{-1}		2.94×10^{-1}		1.65×10^{0}	
1/40	27	8.72×10^{-5}	1.26×10^{-1}	0.98	1.49×10^{-1}	0.98	1.14×10^{0}	0.53
1/80	63	8.45×10^{-5}	6.37×10^{-2}	0.98	7.55×10^{-2}	0.98	7.99×10^{-1}	0.51
1/160	196	9.13×10^{-5}	3.24×10^{-2}	0.98	3.62×10^{-2}	1.06	5.56×10^{-1}	0.52
(d)								
1/20	10	5.52×10^{-5}	2.50×10^{-1}		3.20×10^{-1}		1.69×10^{0}	
1/40	27	8.44×10^{-5}	1.40×10^{-1}	0.84	1.70×10^{-1}	0.91	1.27×10^{0}	0.41
1/80	60	8.57×10^{-5}	7.33×10^{-2}	0.93	8.66×10^{-2}	0.97	9.03×10^{-1}	0.49
1/160	172	9.73×10^{-5}	3.71×10^{-2}	0.98	4.38×10^{-2}	0.98	6.20×10^{-1}	0.54

Table 3 (Test problem (47), $\alpha = 1$): CPU times for the two-stage algorithm with regularization (23) and the four types of meshes shown in Fig. 1, the values of h being those in Table 1. (a) Unit square regular meshes. (b) Unit square symmetric meshes. (c) Unit square isotropic unstructured meshes. (d) Half-unit disk isotropic unstructured meshes

Mesh	CPU time (s)			
	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$
(a)	0.92	1.38	5.81	69.00
(b)	1.01	2.12	11.98	249.89
(c)	0.93	1.39	6.35	58.92
(d)	0.88	1.27	5.35	47.20

boundaries. (ii) Use a simple operator-splitting based method, to solve the discrete Monge–Ampère problems. Privileging simplicity versus sophistication, our goal was not at this stage to develop fast solution methods, although the two-stage algorithm (a Newton-like method) is a step in that direction. There is no doubt that CPU times can be improved by using dedicated

Table 4 (Test problem (47), $\alpha = 1$): comparison between the discrete analogue of the one-stage algorithm (11)–(13) and its two-stage variant. These comparisons have been done for Fig. 1a (regular) and b (symmetric) types of meshes: (i) one-stage algorithm and regular mesh. (ii) two-stage algorithm and regular mesh. (iii) one-stage algorithm and symmetric mesh. (iv) two-stage algorithm and symmetric mesh. Relations (23) have been used to approximate the second order derivatives. In (i) and (ii), $h = \Delta x_1 = \Delta x_2$

	h	Iterations	L_2 error	L_∞ error
(i)	1/80	254	6.46×10^{-6}	1.16×10^{-5}
(ii)	1/80	135	6.43×10^{-6}	1.15×10^{-5}
(iii)	1/80	211	1.18×10^{-4}	2.54×10^{-4}
(iv)	1/80	148	1.24×10^{-4}	2.19×10^{-4}

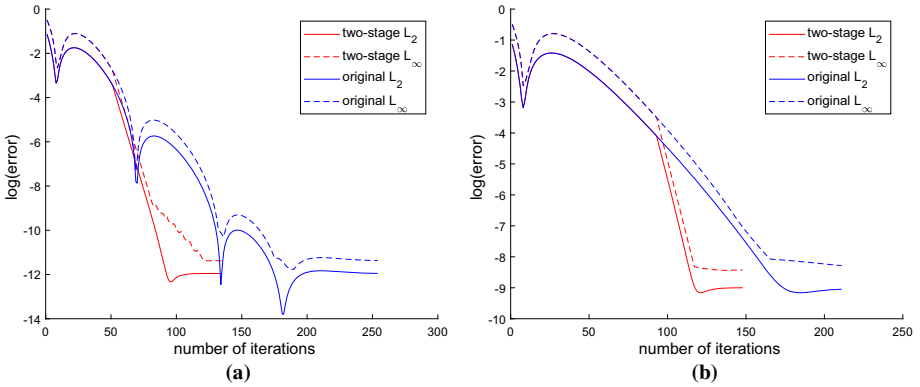

Fig. 3 (Test problem (47), $\alpha = 1$): convergence histories of the L_2 and L_∞ approximation errors: **a** regular mesh for $h(= \Delta x_1 = \Delta x_2) = 1/80$. **b** Symmetric mesh for $h = 1/80$. Relations (23) have been used to approximate the second order derivatives. We observe that the various errors reach a plateau for n sufficiently large

linear solvers more efficient than the more general ones picked by MATLAB, with the user out the loop. □

In Table 4 we have compared the results obtained by the discrete analogue of algorithm (11)–(13) and by its two-stage variant, the meshes being like those in Fig. 1(a) (regular) and (b) (symmetric). These results show the significantly faster convergence of the two-stage algorithm, and the higher accuracy obtained with the regular mesh. We have visualized on Fig. 3 the results of Table 4; the plateaus reached by the L_2 and L_∞ approximation errors as n increases appear clearly on this figure.

Suppose that one takes $\varepsilon = 0$ in the discrete analogue of (12), $C = 0$ in (23), and uses a mesh of Fig. 1a type to solve the discrete Monge–Ampère problem. The solution of problem (47) being a polynomial function of degree 2, it follows from Remark 4 that the relations giving the discrete second order derivatives are exact at the vertices of $\mathcal{T}_h$ interior to Ω. Since $\frac{\partial}{\partial \mathbf{n}} \mathbf{D}^2 u = 0$ on $\partial\Omega$, one expects to reach machine precision for a sufficiently large number of time steps. The results visualized in Fig. 4 show that this prediction is verified. These results have been obtained using the one-stage algorithm with $\Delta t = h^2$, h being equal to 1/40; they show indeed that both the L_2 and L_∞ approximation errors converge to 0 with machine precision.

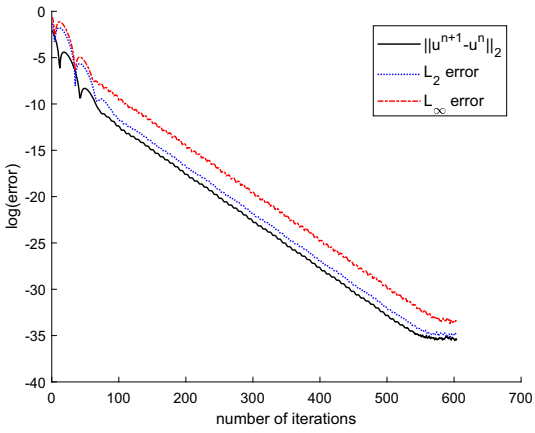

Fig. 4 (Test problem (47), $\alpha = 1$): histories of the L_2 and L_∞ approximation errors for the one-stage algorithm with $\varepsilon = 0$ in (12), $C = 0$ in (23), and $\Delta t = h^2$ (regular mesh on the unit square, $h = 1/40$)

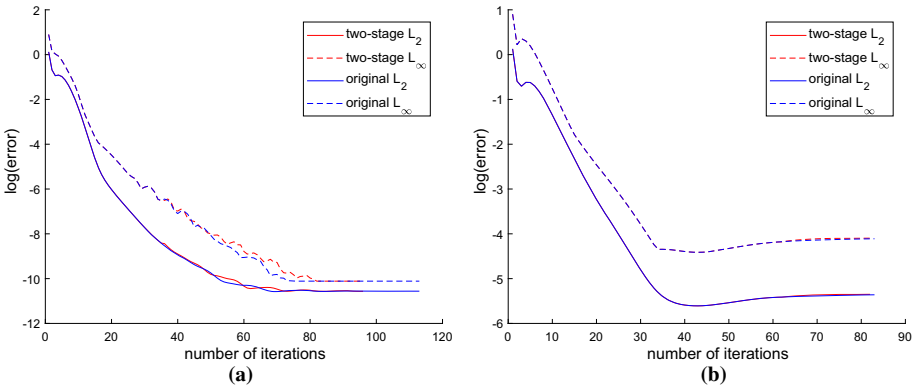

Fig. 5 (Test problem (47), $\alpha = 5$): histories of the L_2 and L_∞ approximation errors for the one-stage and two-stage algorithms. **a** Regular mesh on the unit square with $h = 1/40$. **b** Symmetric mesh with $h = 1/40$. Relations (23) have been used to approximate the second order derivatives. For the two-stage algorithm we used $\gamma = 1$ and $\Delta\tau = 8h$ in the discrete analogue of (46)

Increasing α in (47) increases, as expected, the anisotropy of the solution to the problem. A consequence of this phenomenon is that beyond $\alpha = 5$ there is no gain at using the two-stage algorithm, as shown by the numerical comparisons we performed. This appears clearly on Fig. 5 where we have visualized the convergence histories of the one-stage and two-stage algorithms, when applied to the solution of problem (47) with $\alpha = 5$.

6.3 A Smooth Exponential Example

In this section, we consider the following Monge–Ampère problem

$$\begin{cases} \det \mathbf{D}^2 u = 64\left(1 + 2r^2\right) e^{2r^2} \text{ in } \Omega, \\ u = 4e^{r^2} - \frac{9}{2}, \text{ on } \partial\Omega, \end{cases} \tag{48}$$

with $r = \sqrt{(x_1 - 1/2)^2 + (x_2 - 1/2)^2}$, Ω being either the unit square $(0, 1)^2$ or the open disk of radius 1/2 centered at $(1/2, 1/2)$. We recall that if function ϕ is *radial* with respect to

Table 5 [Test problem (48)]: (i) number of iterations necessary for the convergence of the two-stage algorithm. (ii) Approximation errors and convergence rates for the solution and its gradient. (a) Regular meshes of the unit square. (b) Symmetric meshes of the unit square. (c) Unstructured isotropic meshes of the unit square. (d) Unstructured isotropic meshes of the half-unit disk. Relations (23) have been used to approximate the second order derivatives. (In (a), $h = \Delta x_1 = \Delta x_2$)

h	Iterations	$\|u^{n+1} - u^n\|_2$	L_2 error	Rate	L_∞ error	Rate	$\|\nabla(u^{n_c} - u)\|_2$	Rate
(a)								
1/10	11	5.30×10^{-8}	1.64×10^{-2}		3.75×10^{-2}		3.59×10^{-1}	
1/20	25	9.48×10^{-8}	4.95×10^{-3}	1.73	1.07×10^{-2}	1.81	1.23×10^{-1}	1.55
1/40	55	8.29×10^{-8}	1.28×10^{-3}	1.95	2.67×10^{-3}	2.00	4.23×10^{-2}	1.54
1/80	149	9.28×10^{-8}	3.25×10^{-4}	1.98	6.59×10^{-4}	2.02	1.46×10^{-2}	1.53
1/160	208	9.66×10^{-9}	8.22×10^{-5}	1.98	1.63×10^{-4}	2.09	5.12×10^{-3}	1.51
(b)								
1/10	17	8.07×10^{-8}	2.43×10^{-2}		5.04×10^{-2}		2.24×10^{-1}	
1/20	30	7.06×10^{-8}	8.33×10^{-3}	1.54	1.60×10^{-2}	1.06	8.14×10^{-2}	1.46
1/40	68	7.76×10^{-8}	2.70×10^{-3}	1.63	4.76×10^{-3}	1.75	3.02×10^{-2}	1.43
1/80	219	9.04×10^{-8}	8.31×10^{-4}	1.70	1.37×10^{-3}	1.80	1.13×10^{-2}	1.42
1/160	277	9.93×10^{-9}	2.36×10^{-4}	1.82	3.74×10^{-4}	1.87	4.09×10^{-3}	1.47
(c)								
1/10	13	6.65×10^{-8}	1.70×10^{-2}		4.07×10^{-2}		3.45×10^{-1}	
1/20	25	7.43×10^{-8}	5.50×10^{-3}	1.63	1.14×10^{-2}	1.84	1.06×10^{-1}	1.70
1/40	54	9.86×10^{-8}	1.39×10^{-3}	1.98	3.13×10^{-3}	1.86	3.71×10^{-2}	1.51
1/80	146	8.58×10^{-8}	3.57×10^{-4}	1.96	8.04×10^{-4}	1.96	1.40×10^{-2}	1.41
1/160	225	9.91×10^{-9}	9.30×10^{-5}	1.94	2.66×10^{-4}	1.60	8.96×10^{-3}	0.64
(d)								
1/10	10	4.15×10^{-8}	1.14×10^{-2}		2.93×10^{-2}		2.81×10^{-1}	
1/20	24	6.32×10^{-8}	3.25×10^{-3}	1.81	7.54×10^{-3}	1.96	7.34×10^{-2}	1.94
1/40	51	8.48×10^{-8}	8.85×10^{-4}	1.88	2.09×10^{-3}	1.85	2.44×10^{-2}	1.59
1/80	130	9.35×10^{-8}	2.46×10^{-4}	1.85	6.00×10^{-4}	1.80	7.97×10^{-3}	1.61
1/160	127	9.68×10^{-9}	7.30×10^{-5}	1.75	2.30×10^{-4}	1.38	6.35×10^{-3}	0.33

$(1/2, 1/2)$, then $\det \mathbf{D}^2\phi = \frac{\phi'\phi''}{r}$, implying that the function u defined by $u = \left(4e^{r^2} - \frac{9}{2}\right)\Big|_{\bar{\Omega}}$ is an exact convex solution to problem (48). We have reported on Table 5: (i) The number of iterations needed to have convergence of the two-stage algorithm, using $\|u^{n+1} - u^n\|_2 \leq 10^{-7}$ as stopping criterion, and (ii) various approximation errors. These results were obtained with $\varepsilon = h^2$, the discrete second order derivatives being defined by (23). The orders of convergence of the approximation errors are not optimal but significantly higher than 1, the symmetric meshes having-as expected-the lowest orders of convergence. It is worth noticing that the unstructured isotropic meshes used on the disk give orders of convergence almost as good as those produced by the symmetric meshes on the unit square. For all tests with $h = 1/160$, we set $\varepsilon = h$ and the stopping criterion $\|\mathbf{D}^2 u^n - \mathbf{p}^n\| < 10$ for stage-one and stopping criterion $\|u^{n+1} - u^n\|_2 < 10^{-8}$ for stage-two.

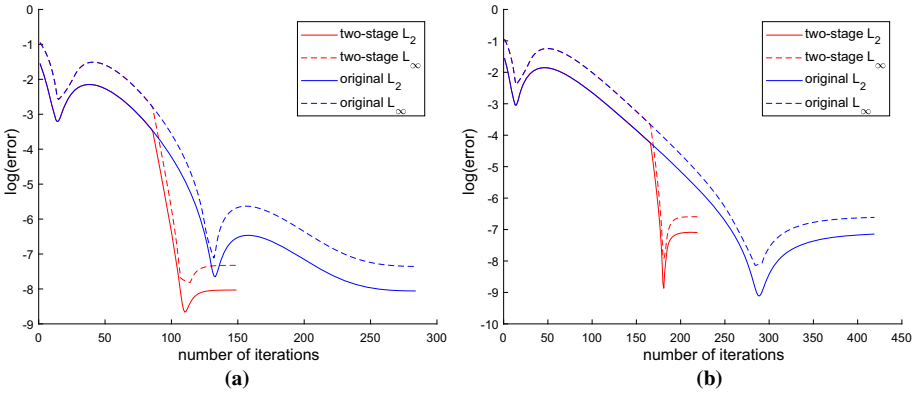

Fig. 6 [Test problem (48)]: convergence histories of the L_2 and L_∞ approximation errors: **a** regular mesh for $h = 1/40$. **b** Symmetric mesh for $h = 1/40$. We observe that the various errors reach a plateau for n sufficiently large. Relations (23) have been used to approximate the second order derivatives. (In **a** $h = \Delta x_1 = \Delta x_2$)

Table 6 [Test problem (48)]: CPU times for the two-stage algorithm and the four types of meshes shown in Fig. 1, the values of h being those in Table 5. (a) Unit square regular meshes. (b) Unit square symmetric meshes. (c) Unit square unstructured meshes. (d) Half-unit disk unstructured meshes

Mesh	CPU time (s)			
	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$
(a)	0.82	1.46	6.57	66.33
(b)	1.01	2.38	15.04	363.98
(c)	0.86	1.51	7.33	77.24
(d)	0.77	1.37	6.03	59.39

In Fig. 6 we have compared the convergence histories of the one-stage and two-stage algorithms, when applied to the solution of problem (48). The comments we did, in Sect. 6.2, about Fig. 3 (concerning the solution of problem (47) with $\alpha = 1$) still apply here.

In Table 6, we have reported the CPU time necessary for the two-stage algorithm to solve problem (48), for the same types of meshes and space discretization steps than in Table 5, relation (23) being used to approximate the second order derivatives.

In Table 7, we have compared the results obtained by the discrete analogue of algorithm (11)–(13) and by its two-stage variant, the meshes being like those in Fig. 1a, b with relations (23). These results show the significantly faster convergence of the two-stage algorithm. The above results, combined with those reported in Sect. 6.2, strongly suggest that for problems with smooth isotropic solutions the two-stage algorithm is clearly faster than the one-stage one.

Finally, we have reported in Table 8, various convergence results obtained when combining the two-stage algorithm with the approximation of the second order derivatives defined by (35), (36). We use stopping criterion $\|\mathbf{D}^2 u^n - \mathbf{p}^n\| < 10$ for stage-one and $\|u^{n+1} - u^n\|_2 < 10^{-4}$ for stage-two, the same criterions as those for Table 2. Comparing with the results in Table 5 the following conclusions can be drawn: (i) The algorithms based on approximation (35), (36) of the second order derivatives are (roughly) twice faster than those based on approximation (23). (ii) Approximation (23) leads to higher orders of accuracy for

Table 7 [Test problem (48)]: comparison between discrete analogue of the one-stage algorithm (11)–(13) and its two-stage variant. These comparisons have been done for meshes as in Fig. 1a (regular) and b (symmetric): (i) one-stage algorithm and regular mesh. (ii) two-stage algorithm and regular mesh. (iii) one-stage algorithm and symmetric mesh. (iv) two-stage algorithm and symmetric mesh. Relations (23) have been used to approximate the second order derivatives. In (i) and (ii), $h = \Delta x_1 = \Delta x_2$

	h	Iterations	L_2 error	L_∞ error
(i)	1/80	284	3.17×10^{-4}	6.39×10^{-4}
(ii)	1/80	149	3.25×10^{-4}	6.59×10^{-4}
(iii)	1/80	419	7.90×10^{-4}	1.34×10^{-3}
(iv)	1/80	219	8.31×10^{-4}	1.37×10^{-3}

Table 8 [Test problem (48)]: (i) number of iterations necessary for the convergence of the two-stage algorithm. (ii) Approximation errors and convergence rates for the solution and its gradient. (a) Regular meshes of the unit square. (b) Symmetric meshes of the unit square. (c) Unstructured isotropic meshes of the unit square. (d) Unstructured isotropic meshes of the half-unit disk. Relations (35) and (36) have been used to approximate the second order derivatives. (In (a), $h = \Delta x_1 = \Delta x_2$)

h	Iterations	$\|u^{n+1} - u^n\|_2$	L_2 error	Rate	L_∞ error	Rate	$\|\nabla(u^{n_c} - u)\|_2$	Rate
(a)								
1/20	17	6.77×10^{-5}	1.58×10^{-1}		1.89×10^{-1}		1.07×10^{0}	
1/40	28	8.26×10^{-5}	8.08×10^{-2}	0.97	9.45×10^{-2}	1.00	7.50×10^{-1}	0.51
1/80	75	9.30×10^{-5}	4.11×10^{-2}	0.98	4.67×10^{-2}	1.02	5.16×10^{-1}	0.54
1/160	248	9.36×10^{-5}	2.13×10^{-2}	0.95	2.34×10^{-2}	1.00	3.57×10^{-1}	0.53
(b)								
1/20	17	9.72×10^{-5}	1.40×10^{-1}		1.67×10^{-1}		1.03×10^{0}	
1/40	36	8.81×10^{-5}	7.19×10^{-2}	0.96	8.09×10^{-2}	1.05	6.90×10^{-1}	0.58
1/80	105	9.50×10^{-5}	3.65×10^{-2}	0.98	3.86×10^{-2}	1.07	4.63×10^{-1}	0.58
1/160	379	9.19×10^{-5}	1.88×10^{-2}	0.96	2.03×10^{-2}	0.93	3.15×10^{-1}	0.56
(c)								
1/20	17	6.34×10^{-5}	1.57×10^{-1}		1.90×10^{-1}		1.09×10^{0}	
1/40	28	8.40×10^{-5}	7.92×10^{-2}	0.99	9.35×10^{-2}	1.02	7.43×10^{-1}	0.55
1/80	76	9.33×10^{-5}	4.01×10^{-2}	0.98	4.57×10^{-2}	1.03	5.08×10^{-1}	0.55
1/160	272	9.33×10^{-5}	2.05×10^{-2}	1.00	2.22×10^{-2}	1.04	3.49×10^{-1}	0.54
(d)								
1/20	16	7.82×10^{-5}	1.57×10^{-1}		2.03×10^{-1}		1.10×10^{0}	
1/40	30	7.67×10^{-5}	8.98×10^{-2}	0.81	1.10×10^{-1}	0.88	8.29×10^{-1}	0.41
1/80	70	8.69×10^{-5}	4.71×10^{-2}	0.93	5.61×10^{-2}	0.97	5.87×10^{-1}	0.50
1/160	228	9.95×10^{-5}	2.42×10^{-2}	0.96	2.89×10^{-2}	0.96	4.01×10^{-1}	0.55

the approximation errors. Further numerical experiments will show that for most non-smooth problems, the algorithms relying on (35), (36) are faster *and* more accurate than those based on (23).

6.4 A Popular Monge–Ampère Problem Without Classical Solution

The test problems we considered in Sects. 6.2 and 6.3 had exact, strictly convex classical solutions belonging to $C^\infty(\bar{\Omega})$. Our goal in this section is to investigate the ability of our methodology at handling a 'pathological' (nevertheless very popular) Monge–Ampère problem, namely

$$\begin{cases} \det \mathbf{D}^2 u = 1 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{49}$$

with $\Omega = (0, 1)^2$. Problem (49) (introduced in [11], to the best of our knowledge) is pathological since, despite the simplicity of its data (making it the Monge–Ampère analogue of the Dirichlet problem $-\nabla^2 u = 1$ in Ω, $u = 0$ on $\partial\Omega$, Ω still being the unit square), this problem *does not have smooth classical solutions*. The non-existence of smooth solutions to problem (49) is very simple to prove (see, e.g., [14] for details), the difficulty stemming, from the non-strict convexity of Ω (and from the corners of Ω). The non-existence of classical solutions does not prevent problem (49) to have generalized solutions, *viscosity solutions* for example. Actually, the approximation of the viscosity solutions to problem (49), and their computation, has been thoroughly investigated in, e.g., [4,38]; this is fortunate since it will allow qualitative and quantitative comparisons. The computational method we used to solve problem (49) relies on the one-stage algorithm combined with the approximation (35), (36) of the second order derivatives, the other combinations being slower and less accurate. Using $\|u^{n+1} - u^n\|_2 \leq 10^{-8}$ as stopping criterion, we obtained the results reported and visualized on Table 9 and Figs. 7 and 8 [in Table 9(a) $h = \Delta x_1 = \Delta x_2$, while in Table (b) and (c), h denotes the length of the mesh largest edge(s)].

As $h \to 0$, the results from Table 9 and graphs of Figs. 7 and 8 indicate the convergence of the approximate computed solutions to a convex function whose minimal value is close to -0.183. The three types of meshes we employed share these convergence properties. The graphs in Fig. 7 are qualitatively close to graphs reported in the literature (in references [4,38], for example). Making quantitative comparisons is more difficult. Indeed the minimal values reached by the solutions computed in [4] Section 5.3 on a 141×141 grid (the finest one used in [4]) range from 0.2621 to 0.3024. Since the test problem considered in [4] was

$$\det \mathbf{D}^2 u = 1 \text{ in } \tilde{\Omega}, u = 1 \text{ on } \partial\tilde{\Omega},$$

with $\tilde{\Omega} = (-1, 1)^2$, corrections are needed (subtract 1 and divide by 4), leading to the range $[-0.184475, -0.17315]$ which contains definitely *all* the values reported in the 6th column of Table 9 (which is gratifying in itself). Actually, there is more: Indeed, one can assume reasonably that in Table 8 of [4] the most accurate approximation of the actual minimal value is 0.2695. Two reasons for that statement are that: (i) The value 0.2695 has been obtained with the finer mesh (141×141) and wider stencil (33 points) used in [4], and (ii) it has been proved (see, e.g., [22]) that the associated method guarantees convergence to a viscosity solution. After correction of the value 0.2695, one obtains -0.182625, which is pretty close to the value -0.1831 reported in Table 9(a) for $h = 1/120$.

Remark 8 The above comparisons with the results in [4] suggest that our methodology produces approximate solutions converging to viscosity solutions if one uses (35), (36) to approximate the second order derivatives. Albeit we do not know how to prove this convergence result, it is not completely surprising since our method solves the regularized Mong–Ampère equation

Table 9 [Test problem (49)]: (i) second column—number of iterations of the one-stage algorithm necessary to achieve convergence (approximation (35), (36) of second order derivatives). (ii) Fourth and fifth columns: Discrete L_2-norms of the consistency gap $\mathbf{D}_h^2(u_h) - \mathbf{p}_h$. (iii) Minimal values of the computed solutions. (a) Regular meshes of the unit square. (b) Symmetric meshes of the unit square. (c) Isotropic unstructured meshes of the unit square. (In (a), $h = \Delta x_1 = \Delta x_2$)

h	Iterations	$\|u^{n+1} - u^n\|_2$	$\|\mathbf{D}^2 u^n - \mathbf{p}^n\|$	$\frac{\|\mathbf{D}^2 u^n - \mathbf{p}^n\|}{\|\mathbf{p}^n\|}$	Min value
(a)					
1/20	222	7.73×10^{-9}	3.96×10^{-5}	1.72×10^{-5}	-0.1801
1/40	622	9.04×10^{-9}	2.83×10^{-3}	7.37×10^{-4}	-0.1817
1/80	2468	9.82×10^{-9}	6.42×10^{-3}	1.28×10^{-3}	-0.1834
1/120	3247	9.90×10^{-9}	6.60×10^{-2}	1.07×10^{-2}	-0.1831
(b)					
1/20	324	9.87×10^{-9}	7.82×10^{-4}	3.43×10^{-4}	-0.1793
1/40	1168	9.99×10^{-9}	5.87×10^{-3}	1.75×10^{-3}	-0.1812
1/80	3994	9.99×10^{-9}	3.40×10^{-2}	7.05×10^{-3}	-0.1831
1/120	8211	9.99×10^{-9}	8.05×10^{-2}	1.36×10^{-2}	-0.1839
(c)					
1/20	217	9.43×10^{-9}	3.49×10^{-5}	1.53×10^{-5}	-0.1797
1/40	625	8.75×10^{-9}	2.40×10^{-3}	7.06×10^{-4}	-0.1815
1/80	2492	9.99×10^{-9}	6.17×10^{-3}	1.28×10^{-3}	-0.1831
1/120	15,885	9.99×10^{-9}	1.20×10^{-2}	2.14×10^{-3}	-0.1837

$$-\frac{\varepsilon}{2}\nabla^2 u - \det \mathbf{D}^2 u = -1 \text{ in } \Omega, u = 0 \text{ on } \partial\Omega,$$

with $\varepsilon \approx h^2$ after space discretization.

We mentioned above that the *non-strict convexity* of $(0, 1)^2$ was explaining the non-existence of smooth solutions to problem (49). In order to investigate the influence of boundary corners we considered the following variant of problem (49)

$$\begin{cases} \det \mathbf{D}^2 u = 1 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{50}$$

where Ω is the (eye-shape) *strictly convex* domain defined by

$$\Omega = \{\{x_1, x_2\} | -x_1(1 - x_1) < x_2 < x_1(1 - x_1), 0 < x_1 < 1\},$$

and visualized in Fig. 9 (where we have also visualized one of the triangulations we employed). We have reported on Table 10 and Figs. 10 and 11 numerical results obtained by the one-stage algorithm, using approximations (35), (36) of the second order derivatives, the stopping criterion still being $\|u^{n+1} - u^n\|_2 < 10^{-8}$. Comparing Tables 9 and 10 shows several qualitative similarities, namely: (i) The number of iterations necessary to achieve convergence of the one-stage algorithm is a fast increasing function of $1/h$. (ii) The discrete L^2-norm of the consistency gap $\mathbf{D}_h^2(u_h) - \mathbf{p}_h$ increases with 1/h. These results suggest the non-existence of a smooth convex solution to problem (50). We could have anticipated this result. Suppose indeed that u is a convex solution to (50) belonging to $C^1(\bar{\Omega})$: it follows then from the boundary conditions that

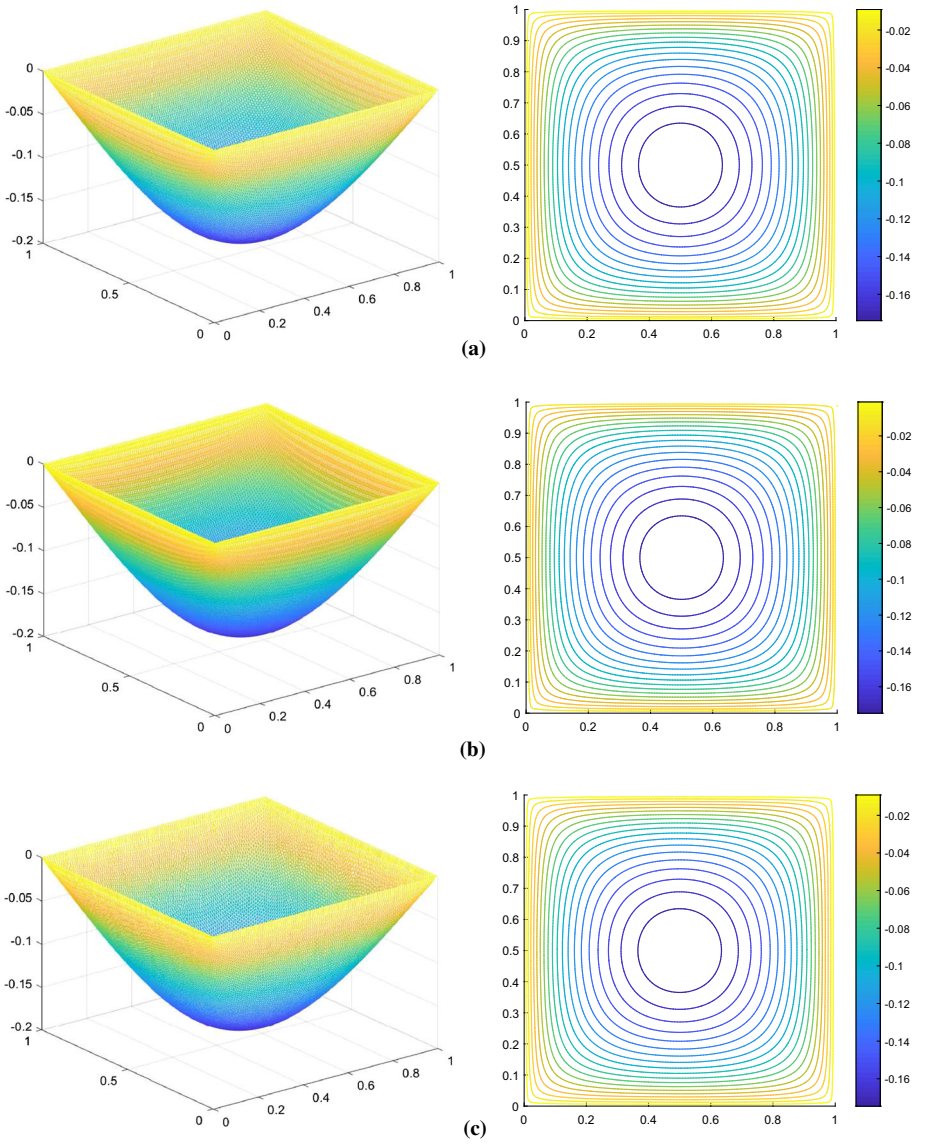

Fig. 7 [Test problem (49)]: graphs and contours of the computed solutions, using the one-stage algorithm and relations (35), (36) to approximate the second order derivatives. **a** Regular mesh with $h = 1/120$. **b** Symmetric mesh with $h = 1/120$. **c** Unstructured isotropic mesh with $h = 1/120$. (In **a** $h = \Delta x_1 = \Delta x_2$)

$$\nabla u(0,0) \cdot \tau_+ = 0, \nabla u(0,0) \cdot \tau_- = 0, \tag{51}$$

where the unit vectors $\tau_+ (= (1/\sqrt{2}, 1/\sqrt{2}))$ and $\tau_- (= (1/\sqrt{2}, -1/\sqrt{2}))$ are tangent at $(0,0)$ at the upper and lower parts of $\partial\Omega$, respectively. Relations (51) imply $\nabla u(0,0) = \mathbf{0}$. Similarly, one can show that $\nabla u(1,0) = \mathbf{0}$. The function u being convex and differentiable

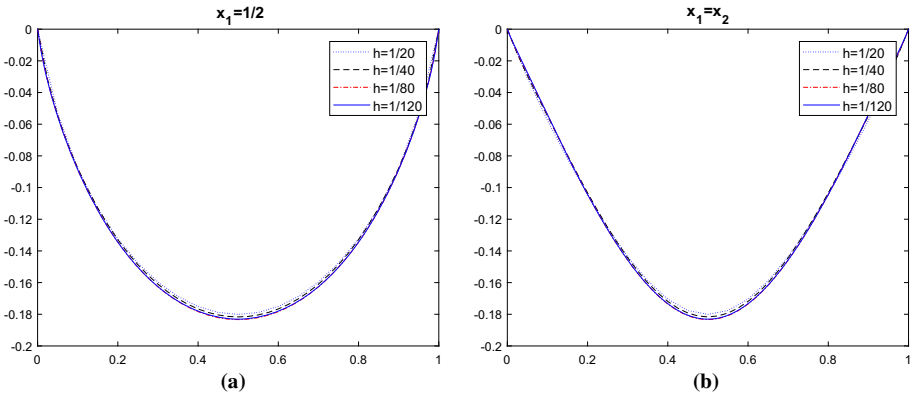

Fig. 8 [Test problem (49)]: graphs of the restrictions of the computed solutions to: **a** the line $x_1 = 1/2$ and **b** the line $x_1 = x_2$, for $h = 1/20, 1/40, 1/80$, and $1/120$ (one-stage algorithm using regular meshes, and approximations (35), (36) of the second order derivatives). **a**, **b** Suggest the uniform convergence of the approximate solutions to a strictly convex generalized solution of problem (49)

Fig. 9 [Test problem (50)]: a triangulation of the eye-shape domain ($h = 1/40$)

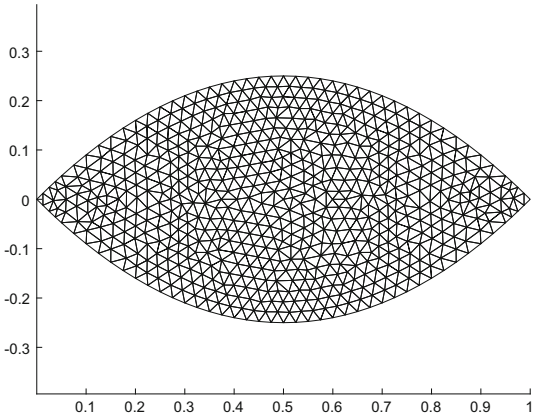

Table 10 [Test problem (50)]: (i) second column—number of iterations of the one-stage algorithm necessary to achieve convergence (approximations (35), (36) of the second order derivatives). (ii) Fourth and fifth columns: Discrete L_2-norms of the consistency gap $\mathbf{D}_h^2(u_h) - \mathbf{p}_h$. (iii) Minimal values of the computed solutions

h	Iterations	$\|u^{n+1} - u^n\|_2$	$\|\mathbf{D}^2 u^n - \mathbf{p}^n\|$	$\frac{\|\mathbf{D}^2 u^n - \mathbf{p}^n\|}{\|\mathbf{p}^n\|}$	Min value
1/20	84	8.73×10^{-9}	1.05×10^{-5}	9.27×10^{-6}	-0.0644
1/40	220	9.73×10^{-9}	6.34×10^{-4}	4.33×10^{-4}	-0.0598
1/80	829	9.53×10^{-9}	1.30×10^{-3}	7.34×10^{-4}	-0.0573
1/120	1766	9.93×10^{-9}	2.38×10^{-3}	1.17×10^{-3}	-0.0578

on the convex set $\bar{\Omega}$, and vanishing on $\partial\Omega$, the relation $\nabla u(0,0) = \mathbf{0}$ implies that $u = 0$ on $\bar{\Omega}$, which implies in turn that $\det \mathbf{D}^2 u = 0$ in Ω, contradicting (50).

Figures 10 and 11 suggest uniform convergence to a strictly convex function as $h \to 0$, the gradient of this solution being infinite on $\partial\Omega \setminus \{\{0,0\},\{1,0\}\}$.

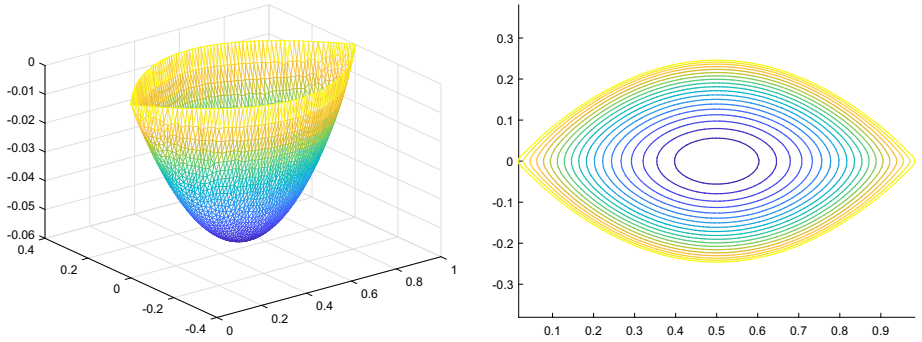

Fig. 10 [Test problem (50)]: graph and contours of the computed approximate solution ($h = 1/80$)

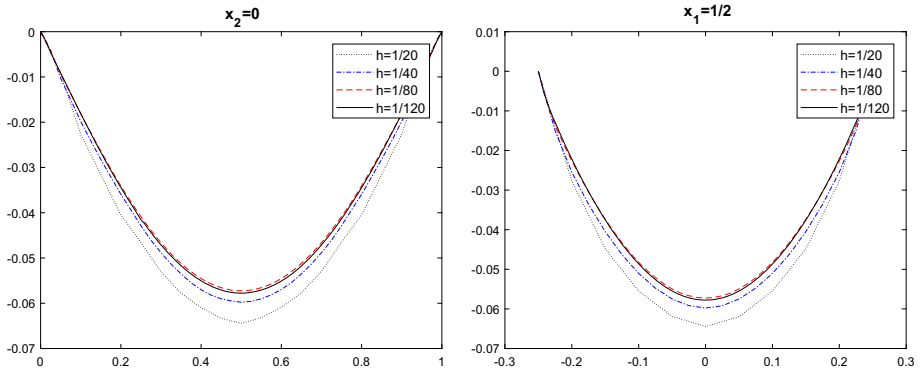

Fig. 11 [Test problem (50)]: graphs of the computed approximate solutions restricted to the lines $x_2 = 0$ (left), and $x_1 = 1/2$ (right) ($h = 1/20, 1/40, 1/80, 1/120$)

6.5 Test Problems with Singular Functions *f*

This subsection is dedicated to Monge–Ampère-Dirichlet problems, where function f in (1) is *singular* (in the sense that $\sup_{x\in\Omega} f(x) = +\infty$). In that direction, the *first problem* we consider is defined by

$$\begin{cases} \det \mathbf{D}^2 u = \frac{1}{r} \text{ in } \Omega, \\ u = \frac{2\sqrt{2}}{3} r^{3/2} \text{ on } \partial\Omega, \end{cases} \tag{52}$$

where $\Omega = (-1, 1)^2$ and $r = \sqrt{(x_1 - 1)^2 + (x_2 - 1)^2}$, implying that function f blows up at the corner $(1, 1)$. The strictly convex function u defined by

$$u(x_1, x_2) = \frac{2\sqrt{2}}{3} r^{3/2}, \forall (x_1, x_2) \in \bar{\Omega},$$

is an exact solution to problem (52). Very clearly, $u \notin C^2(\bar{\Omega})$. On the other hand, one can easily verify that $u \in C^1(\bar{\Omega}) \cap W^{2,s}(\Omega), \forall s \in [1, 4)$, making problem (52) an 'almost' smooth one. To solve problem (52), we used the finite element analogue of the one-stage algorithm (11)–(13) with $\varepsilon = 2h^2$, the triangulations we employed being regular (as in Fig. 1a), or symmetric (as in Fig. 1b). We used relations (23) to approximate the second order derivatives with $C = 2$. We took $\|u^{n+1} - u^n\|_2 < 10^{-5}$ as stopping criterion. For this test problem, h denotes the space discretization step $\Delta x_1 (= \Delta x_2)$.

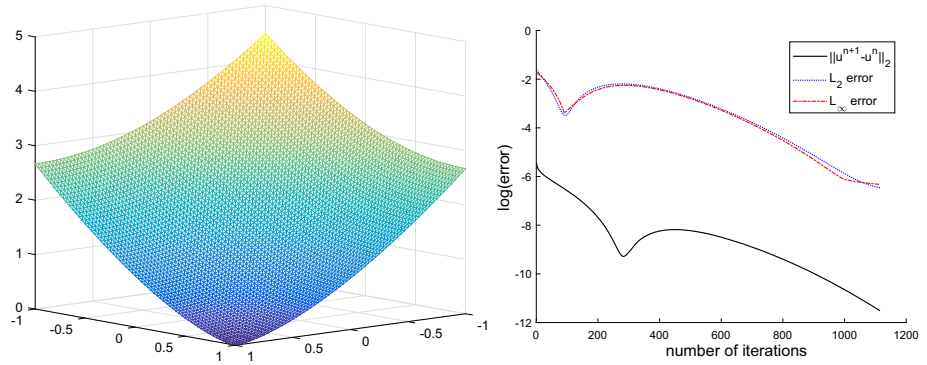

Fig. 12 [Test problem (52)]: graphs of the computed approximate solution and convergence histories (regular triangulation with $h = 1/64$)

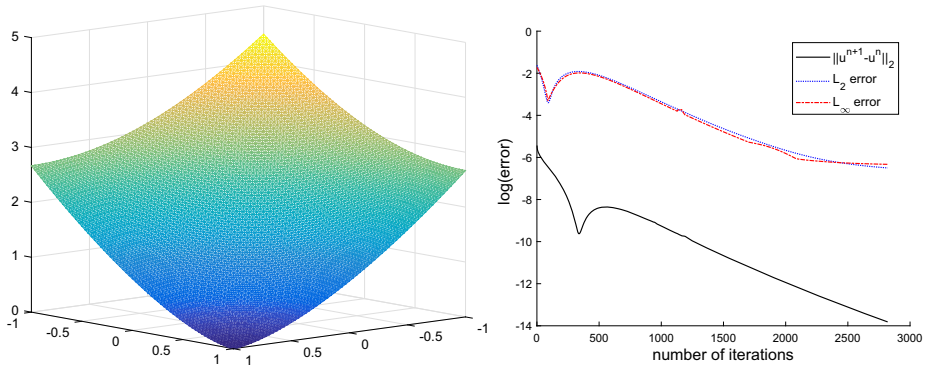

Fig. 13 [Test problem (52)]: graphs of the computed solution and convergence histories (symmetric triangulation with $h = 1/64$)

We have visualized on Fig. 12 (resp., Fig. 13) the graph of the computed approximate solution, and the convergence histories of the residual $\|u^{n+1} - u^n\|_2$ and of the L_2 and L_∞ norms of the approximation error, Fig. 12 (resp., Fig. 13) being associated with the $h = 1/64$ regular (resp., $h = 1/64$ symmetric) triangulation.

In Table 11(a), we have reported results obtained on regular meshes by the method detailed above, namely number of iterations necessary to achieve convergence and related approximation errors for various values of h. For comparison purpose, we have reported in Table 11(b) the related results obtained in [43] on identical meshes by two solution methods, both relying on those wide-stencil finite difference methods discussed in [22]. Comparing Table 11(a) and (b) shows that (for this test problem at least) the method we employed to solve problem (52) is significantly more accurate than the two methods discussed in [43] (however, for this test problem and for the same meshes, they are less accurate and much slower than the least-square/relaxation method discussed in [8]).

Remark 9 The methods discussed in [43] rely on accelerated gradient type algorithms *à la Nesteroy* ([36,45]) whose convergence rate is $O(1/n^2)$, n denoting the number of iterations. Actually, the two methods discussed in [43] are also *cascadic* (in the sense that a first approximate solution is computed on a coarse mesh, and then interpolated on a mesh twice finer, in order to initialize an iterative method). Motivated by [36,43] (and by suggestions of our

Table 11 [Test problem (52)]: number of iterations necessary to achieve convergence and associated approximation errors on regular meshes for various mesh sizes. (a) Results obtained by the methodology discussed in the current article. (b) Results obtained by the two methods discussed in [43] (M1 and M2 are the numbers of iterations needed by the two methods discussed in [43] to achieve convergence)

h	Iterations	$\|u^{n+1}-u^n\|_2$	L_2 error	Rate	L_∞ error	Rate
(a)						
1/16	106	9.69×10^{-6}	9.23×10^{-3}		8.75×10^{-3}	
1/32	311	9.83×10^{-6}	3.71×10^{-3}	1.31	3.59×10^{-3}	1.29
1/64	1115	9.96×10^{-6}	1.56×10^{-3}	1.25	1.79×10^{-3}	1.00

Grid size	L_∞ error	Rate	M1	M2
(b)				
16×16	2.50×10^{-2}		564	601
32×32	1.60×10^{-2}	0.64	585	651
64×64	1.10×10^{-2}	0.54	976	1037

colleague X.C. Tai) we applied the so-called Nesterov method (introduced by Polyak in [41]) to speed up the convergence of the algorithms considered in the present article. The improvement it brings to these algorithms are marginal (if any). A possible explanation is that our algorithms rely on implicit/explicit schemes, unlike the algorithms in [43], all related to fully explicit schemes. □

The *second problem* we consider in this subsection is another classical test problem. It is defined by

$$\begin{cases} \det \mathbf{D}^2 u = \frac{R^2}{(R^2-r^2)^2} \text{ in } \Omega, \\ u = -\sqrt{R^2-r^2} \text{ on } \partial\Omega. \end{cases} \tag{53}$$

where $\Omega=(0,1)^2$, $R\geq 1/\sqrt{2}$, and $r=\sqrt{(x_1-1/2)^2+(x_2-1/2)^2}$. The function u defined by

$$u(x_1,x_2)=-\sqrt{R^2-r^2}, \forall (x_1,x_2)\in\bar{\Omega},$$

is a strictly convex solution to problem (53), its graph being a part of the sphere of radius R centered at (1/2,1/2,0). The above function u belongs to $C^\infty(\bar{\Omega})$ if $R>1/\sqrt{2}$. On the other hand, $u\in C^0(\bar{\Omega})\cap W^{1,s}(\Omega)$, $s\in[1,4)$, if $R=1/\sqrt{2}$, the vector-valued function ∇u being singular at the four corners of $\bar{\Omega}$. Function u reaches its minimal value $(-R)$ at (1/2,1/2). We tested our methodology on the two particular cases of problem (53) associated with $R=\sqrt{2}$ (a smooth case) and $R=1/\sqrt{2}$ (a non-smooth case). For both values of R we used: (i) Regular triangulations (like the one in Fig. 1a, with $h(=\Delta x_1=\Delta x_2)$ taking the values 1/20, 1/40 and 1/80, and (ii) the fully discrete analogue of the one-stage algorithm (11)–(13) with $\varepsilon=h^2$. The results in Table 12(a) [resp., Table 12(b)] have been obtained using approximation (23) [resp., (35), (36)] of the second derivatives, the stopping criterion being $\|u^{n+1}-u^n\|_2<10^{-10}$ (resp., $<10^{-8}$). These results show that the method relying on (23) is second order accurate, while the other one is first order accurate. We observe also that in Table 12(a) [resp., Table 12(b)] the minimal value of the computed solutions decreases (resp., increases) with h, the convergence to the exact value ($-\sqrt{2}(=-1.41421\ldots)$, here)

Table 12 (Test problem (53) with $R = \sqrt{2}$): number of iterations needed for the convergence of the one-stage algorithm, minimal value of the computed solutions, and associated L_2 and L_∞ norms of the approximation errors. (a) Approximations (23) of the second order derivatives. (b) Approximations (35), (36) of the second order derivatives. We used regular triangulations (with $h = \Delta x_1 = \Delta x_2$), as the one in Fig. 1a, to obtain these results. We recall that the minimal value of the exact solution is $-\sqrt{2}(= -0.141421\ldots)$

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	358	9.68×10^{-11}	-1.4138	1.99×10^{-4}		4.08×10^{-4}	
1/40	1161	9.55×10^{-11}	-1.4141	5.06×10^{-5}	1.98	1.03×10^{-4}	1.99
1/80	4505	9.97×10^{-11}	-1.4142	1.26×10^{-5}	2.01	2.54×10^{-5}	2.02
(b)							
1/20	290	9.01×10^{-9}	-1.4266	1.16×10^{-2}		1.38×10^{-2}	
1/40	1050	9.91×10^{-9}	-1.4204	5.99×10^{-3}	0.95	7.03×10^{-3}	0.97
1/80	3024	9.99×10^{-9}	-1.4173	3.04×10^{-3}	0.99	3.56×10^{-3}	0.98

Table 13 (Test problem (53) with $R = 1/\sqrt{2}$): number of iterations needed for the convergence of the one-stage algorithm, minimal value of the computed solutions, and associated L_2 and L_∞ norms of the approximation errors. (a) Approximations (23) of the second derivatives. (b) Approximations (35), (36) of the second derivatives. We used regular triangulations (with $h = \Delta x_1 = \Delta x_2$), as the one in Fig. 1a, to obtain these results. We recall that the minimal value of the exact solution is $-1/\sqrt{2}(= -0.707106\ldots)$

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	142	9.80×10^{-9}	-0.7052	3.90×10^{-3}		2.58×10^{-2}	
1/40	489	9.92×10^{-9}	-0.7065	1.70×10^{-3}	1.20	1.95×10^{-2}	0.40
1/80	1548	9.90×10^{-9}	-0.7069	6.54×10^{-4}	1.38	1.43×10^{-2}	0.45
(b)							
1/20	145	8.29×10^{-9}	-0.7382	3.24×10^{-2}		4.64×10^{-2}	
1/40	539	9.93×10^{-9}	-0.7225	1.58×10^{-2}	1.04	2.67×10^{-2}	0.80
1/80	1846	9.95×10^{-9}	-0.7146	7.72×10^{-3}	1.03	1.56×10^{-2}	0.78

being much faster for the first method. From the smoothness of the $R = \sqrt{2}$ solution, the results reported just above were expected, including the clear superiority of the relation (23) based method (Table 12).

Taking $R = 1/\sqrt{2}$ in (53) makes this problem non-smooth (and therefore more challenging) since ∇u is singular at the four corners of $\bar{\Omega}$. Taking $\|u^{n+1} - u^n\|_2 < 10^{-8}$ as stopping criterion, we obtained the results reported in Table 13(a) and (b).

Comparing Table 13(a) and (b) suggest the following: (i) The two one-stage methods we employed, relying on either (23) or (35), (36) behave similarly as long as iterative properties are concerned. (ii) The first method produces better approximation errors for the values of h considered here, particularly for the discrete L_2 approximation error. Actually, the convergence rates reported in the last column of Table 13(a) are close to those reported in Section 5.4 of [4], for a closely related problem. (iii) We observe that in Table 13(a) [resp., Table 13(b)] the minimal value of the computed solutions decreases (resp., increases) with h . We observed a similar behavior in Table 12. As we already mentioned, the numerical solution of problem (53) with $R = 1/\sqrt{2}$ is discussed in [4]; it is discussed also in [22,37].

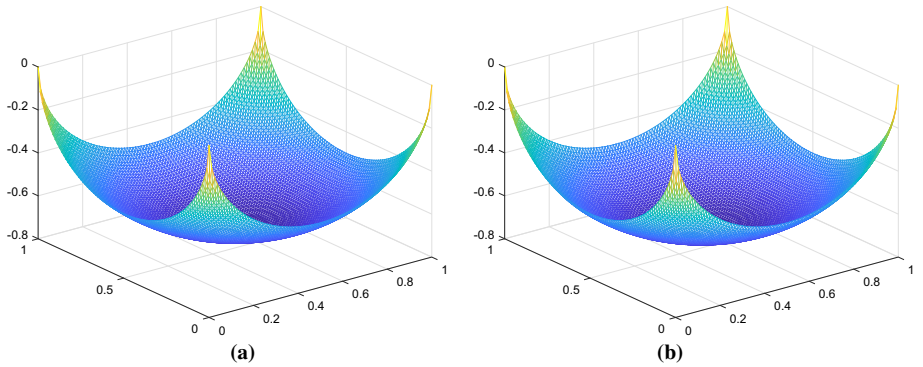

Fig. 14 (Test problem (53) with $R = 1/\sqrt{2}$): graphs of the approximate solutions computed on a regular mesh with $h = 1/80$. **a** Approximations (23) of the second order derivatives. **b** Approximations (35), (36) of the second order derivatives

Table 14 [Test problem (54)]: (i) number of iterations needed for the convergence of the one-stage algorithm. (ii) Approximations errors and orders of convergence. (a) Approximation (23) of the second order derivatives. (b) Approximations (35), (36) of the second order derivatives

h	Iterations	$\|u^{n+1} - u^n\|$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	46	9.21×10^{-6}	-0.4077	6.86×10^{-2}		9.21×10^{-2}	
1/40	138	9.55×10^{-6}	-0.4394	4.85×10^{-2}	0.50	6.05×10^{-2}	0.61
1/80	485	9.85×10^{-6}	-0.4597	3.36×10^{-2}	0.53	4.03×10^{-2}	0.59
1/160	1600	9.99×10^{-6}	-0.4699	2.36×10^{-2}	0.51	3.25×10^{-2}	0.31
(b)							
1/20	33	2.30×10^{-6}	-0.5272	3.66×10^{-2}		5.88×10^{-2}	
1/40	143	9.91×10^{-6}	-0.5344	3.48×10^{-2}	0.07	5.30×10^{-2}	0.15
1/80	500	9.84×10^{-6}	-0.5296	2.83×10^{-2}	0.30	4.31×10^{-2}	0.30
1/160	1652	9.99×10^{-6}	-0.5186	1.93×10^{-2}	0.55	3.31×10^{-2}	0.38

The two graphs in Fig. 14 are quite similar and show very clearly the singular behavior of ∇u close to the four corners of $\bar{\Omega}$.

The next problem with a spherical solution we consider is the following variant of problem (53):

$$\begin{cases} \det \mathbf{D}^2 u = \frac{4}{(1-4r^2)^2} \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{54}$$

where $\Omega = \{\{x_1, x_2\}, (x_1 - 1/2)^2 + (x_2 - 1/2)^2 < 1/4\}$ and $r = \sqrt{(x_1 - 1/2)^2 + (x_2 - 1/2)^2}$. The function u defined by

$$u(x_1, x_2) = -\frac{1}{2}\sqrt{1 - 4r^2} \text{ on } \bar{\Omega}$$

is a strictly convex solution to problem (54), taking its minimal value $-1/2$ at (1/2,1/2). We can easily show that $u \in C^0(\bar{\Omega}) \cap W^{1,s}(\Omega), \forall s \in [1, 2)$.

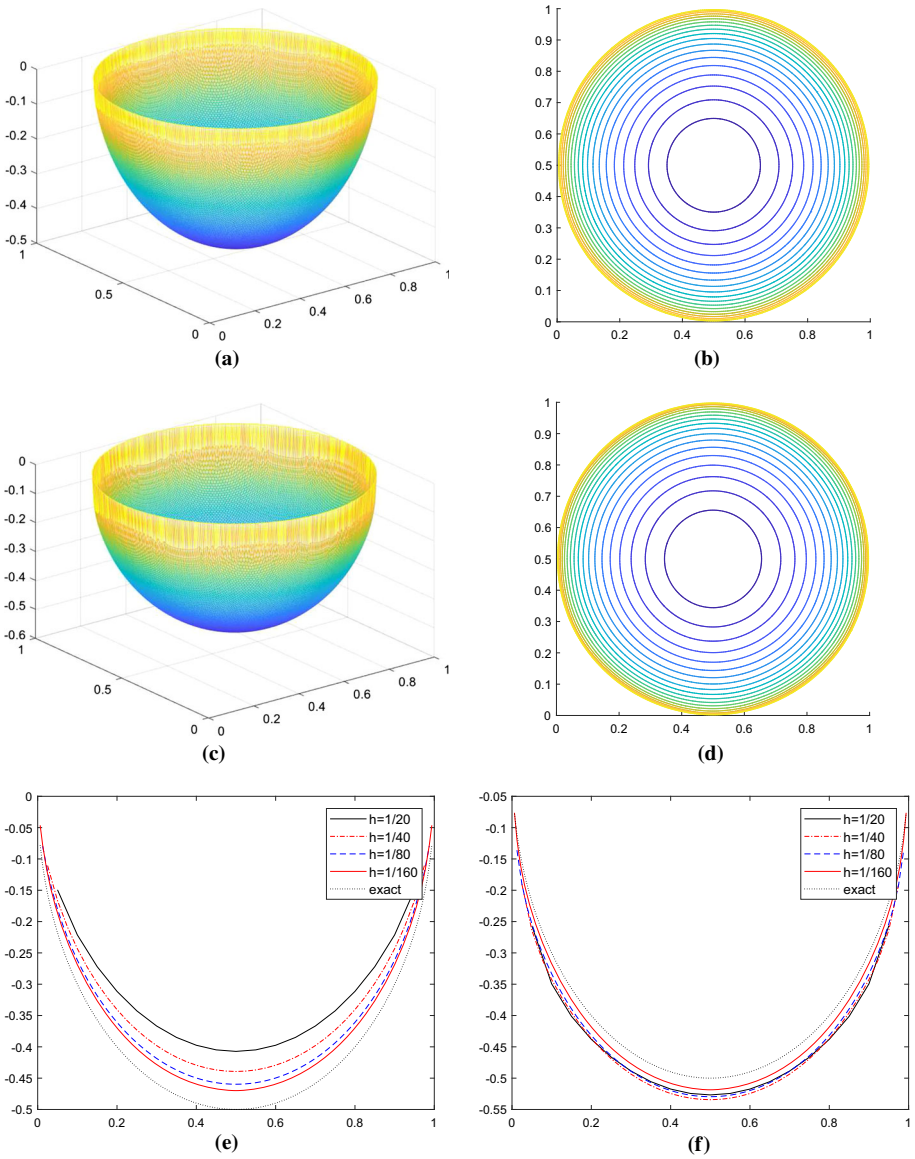

Fig. 15 [Test problem (54)]: **a**, **b** graphs and contours of the approximate solution computed using approximations (23) of the second order derivatives ($h = 1/160$). **c**, **d** Graphs and contours of the approximate solution computed using approximations (35), (36) of the second order derivatives ($h = 1/160$). **e**, **f** Graphs of the restrictions of the computed approximate solutions to the diameter $x_1 = 1/2$, for $h = 1/20, 1/40, 1/80$ and $1/160$ [**e** approximations (23), **f** approximations (35), (36)]

Problem (54) is significantly 'more non-smooth' than problem (53) since $|\nabla u|$ is infinite on the *whole* boundary of $\bar{\Omega}$, making (54) a good test problem to investigate the robustness of our methodology. In order to solve problem (54) we used: (i) Isotropic unstructured finite element meshes like the one in Fig. 1d. (ii) The one-stage algorithm with $\varepsilon = h^2$.

Table 15 [Test problem (55)]: performances of the two-stage algorithm on a family of isotropic unstructured meshes. (a) Approximations (23) of the second order derivatives. (b) Approximations (35), (36) of the second order derivatives. The minimal value of the exact solution is $-\sqrt{2}(=-1.41421\ldots)$

h	Iterations	$\|u^{n+1}-u^n\|$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	27	9.10×10^{-8}	-1.4138	1.43×10^{-4}		3.20×10^{-4}	
1/40	36	9.09×10^{-8}	-1.4141	3.62×10^{-5}	1.98	8.03×10^{-5}	1.99
1/80	40	9.62×10^{-8}	-1.4142	1.11×10^{-5}	1.71	2.35×10^{-5}	1.77
(b)							
1/20	32	7.33×10^{-8}	-1.4283	1.16×10^{-2}		1.49×10^{-2}	
1/40	38	8.32×10^{-8}	-1.422	6.59×10^{-3}	0.82	8.01×10^{-3}	0.90
1/80	43	9.87×10^{-8}	-1.4182	3.44×10^{-3}	0.94	4.09×10^{-3}	0.97

(iii) The approximations of the second order derivatives defined by (23) and (35), (36). (iv) $\|u^{n+1}-u^n\|_2 < 10^{-5}$ as stopping criterion. The results reported in Table 14(a) and (b) show that the two approaches we are considering are very close to each other as long as the number of iterations is concerned. On the other hand, these results show that the second method is more accurate than the first one. We obtain a clear-cut confirmation of the superiority of the second method by comparing Fig. 15e , f where we visualized, for various values of h, the graphs of the restrictions of the computed approximate solutions to the diameter $x_1 = 1/2$. Figure 15a–d (obtained with $h = 1/160$) provide additional details.

Remark 10 When applied to the solution of problems (53) and (54), the two-stage algorithm (described in Sect. 5) does not decrease the number of iterations needed for convergence, when compared to the one-stage algorithm (11)–(13). This disappointing (but not surprising) property follows from the non-smoothness of the solution to both problems. To check again that solution smoothness enhances the speed of convergence of the two-stage algorithm we are going to consider another test problem with a spherical solution, namely

$$\begin{cases} \det \mathbf{D}^2 u = \frac{2}{(2-r^2)^2} \text{ in } \Omega, \\ u = -\sqrt{2-r^2} \text{ on } \partial\Omega, \end{cases} \tag{55}$$

where $\Omega = \{\{x_1, x_2\}, (x_1-1/2)^2+(x_2-1/2)^2 < 1/4\}$ and $r = \sqrt{(x_1-1/2)^2+(x_2-1/2)^2}$. The function u defined by

$$u(x_1, x_2) = -\sqrt{2-r^2} \text{ on } \bar{\Omega}$$

is a strictly convex solution to problem (55), verifying $u \in C^\infty(\bar{\Omega})$. Applying the two-stage algorithm combined with unstructured isotropic triangulations, like those in Fig. 1d, leads to the results reported in Table 15 and visualized in Fig. 16 (we used $\|u^{n+1}-u^n\|_2 < 10^{-7}$ as topping criterion).

Table 15 shows that the two-stage approach we tested behave quite similarly as long as the number of iterations is concerned. On the other hand Table 15 shows that for this smooth test problem, the method based on (23) provides approximation errors 'almost' two orders of magnitude smaller than the one based on (35), (36).

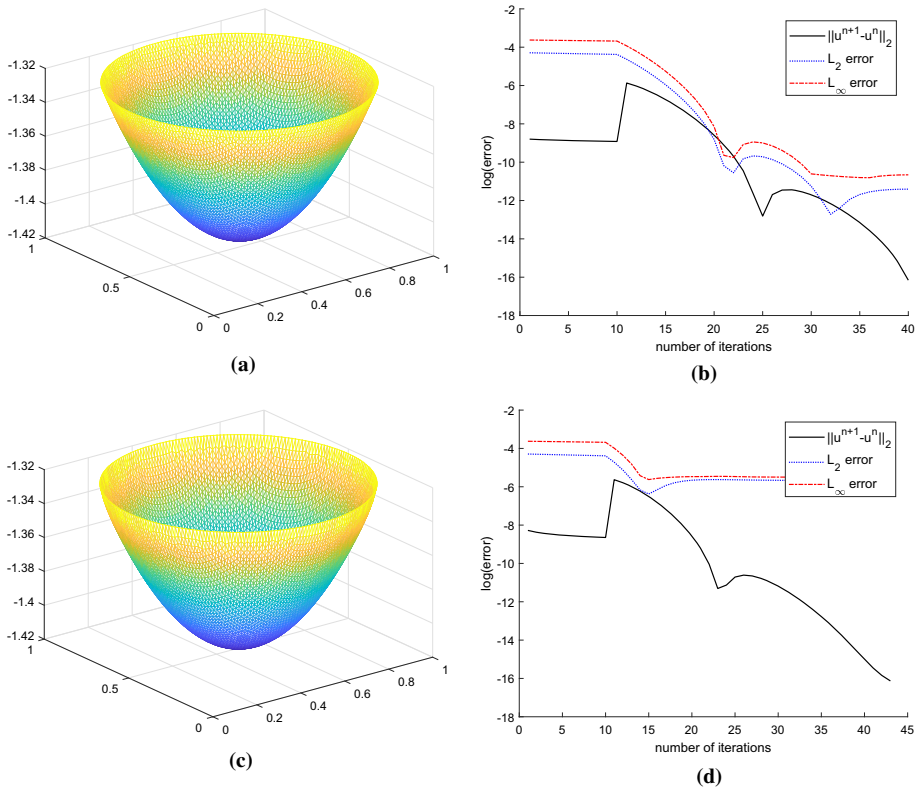

Fig. 16 [Test problem (55)]: graphs of the computed approximate solutions and variations with n of the residual and of the L_2 and L_∞ norms of the approximation errors (two-stage algorithm and isotropic unstructured triangulation of the disk $\bar{\Omega}$ with $h = 1/80$). **a**, **b** Approximations (23) of the second order derivatives. **c**, **d** Approximations (35), (36) of the second order derivatives

Figure 16 shows that, in practice, the final values of the approximation error norms are reached much more quickly, when using relations (35), (36) to approximate the second order derivatives, than when using (23). □

To conclude this sub-section, dedicated to test problems with singular functions f, we are going to consider a test problem, which seems to be new in this context, namely:

$$\begin{cases} \det \mathbf{D}^2 u = f \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{56}$$

with $\Omega = (0, 1)^2$ and f defined by

$$f(x_1, x_2) = \frac{1 - (1 - 2x_1)^2(1 - 2x_2)^2}{16x_1(1 - x_1)x_2(1 - x_2)}, \forall (x_1, x_2) \in \Omega.$$

The strictly convex function u defined by

$$u(x_1, x_2) = -\sqrt{x_1(1 - x_1)x_2(1 - x_2)}, \forall (x_1, x_2) \in \bar{\Omega},$$

Table 16 [Test problem (56)]: performance of the one-stage algorithm combined with regular triangulations: (a) Approximations (23) of the second order derivatives. (b) Approximations (35), (36) of the second order derivatives. The minimal value of the exact solution is -0.25, showing the clear superiority of the second approach

h	Iterations	$\|u^{n+1}-u^n\|$	Min	L_2 error	Rate	L_∞ error	Rate	L_1 error	Rate
(a)									
1/20	132	9.15×10^{-7}	-0.2014	4.86×10^{-2}		4.86×10^{-2}		3.99×10^{-2}	
1/40	314	9.75×10^{-7}	-0.2202	2.73×10^{-2}	0.83	3.05×10^{-2}	0.67	2.65×10^{-2}	0.59
1/80	1209	9.81×10^{-7}	-0.2322	1.68×10^{-2}	0.70	2.03×10^{-2}	0.59	1.65×10^{-2}	0.68
1/160	4973	9.92×10^{-8}	-0.2391	1.02×10^{-2}	0.72	1.40×10^{-2}	0.54	1.01×10^{-2}	0.71
(b)									
1/20	138	9.77×10^{-7}	-0.2313	1.29×10^{-2}		1.87×10^{-2}		1.16×10^{-2}	
1/40	312	9.84×10^{-7}	-0.2376	9.41×10^{-3}	0.46	1.24×10^{-2}	0.59	8.79×10^{-3}	0.40
1/80	1207	9.99×10^{-7}	-0.2427	5.82×10^{-3}	0.69	7.32×10^{-3}	0.76	5.55×10^{-3}	0.66
1/160	4936	9.94×10^{-8}	-0.2455	3.48×10^{-3}	0.74	4.50×10^{-3}	0.70	3.35×10^{-2}	0.73

is solution to problem (56). We clearly have $u \in C^0(\bar{\Omega}) \cap W^{1,s}(\Omega), \forall s \in [1,2)$, with ∇u singular on the whole boundary of Ω. Combining the one-stage algorithm with regular triangulations, we obtained the results reported in Table 16 and visualized in Fig. 17. These results show the clear superiority, in terms of accuracy, of the approach based on approximations (35), (36) of the second order derivatives. Actually, comparing Fig. 17f, h shows that the method based on (35), (36) has better convexity conservation properties than the method based on (23). We used $\|u^{n+1}-u^n\|_2 < 10^{-6}$ (resp. 10^{-7}) as stopping criterion for $h = 1/20$, $1/40$ and $1/80$ (resp., $1/160$).

Problem (56) deserves becoming a classical test problem for Monge–Ampère–Dirichlet solvers: Future will tell.

6.6 Test Problems with a Positive Measure as Right-Hand Side

Test problems where, in (1), f is a Dirac measure (possibly multiplied by some positive constant) are classical nowadays. Our goal in this section is to return to such test problems, and then to go one step further by considering situations where measure f is of the form

$$v \to \int_\gamma v d\gamma,$$

γ being a curve contained in Ω. We consider first the test problem defined by

$$\begin{cases} \det \mathbf{D}^2 u = \pi \delta_{(1/2,1/2)} \text{ in } \Omega, \\ u = r \text{ on } \partial\Omega \end{cases} \tag{57}$$

where in (57): Ω is either the square $(0,1)^2$ or the open disk of radius 1/2 centered at (1/2,1/2), $r = \sqrt{(x_1-1/2)^2+(x_2-1/2)^2}$, and $\delta_{(1/2,1/2)}$ is the Dirac measure at (1/2,1/2). The function u defined by $u = r|_{\bar{\Omega}}$ is an exact convex solution to problem (57), belonging to $W^{1,\infty}(\Omega)$.

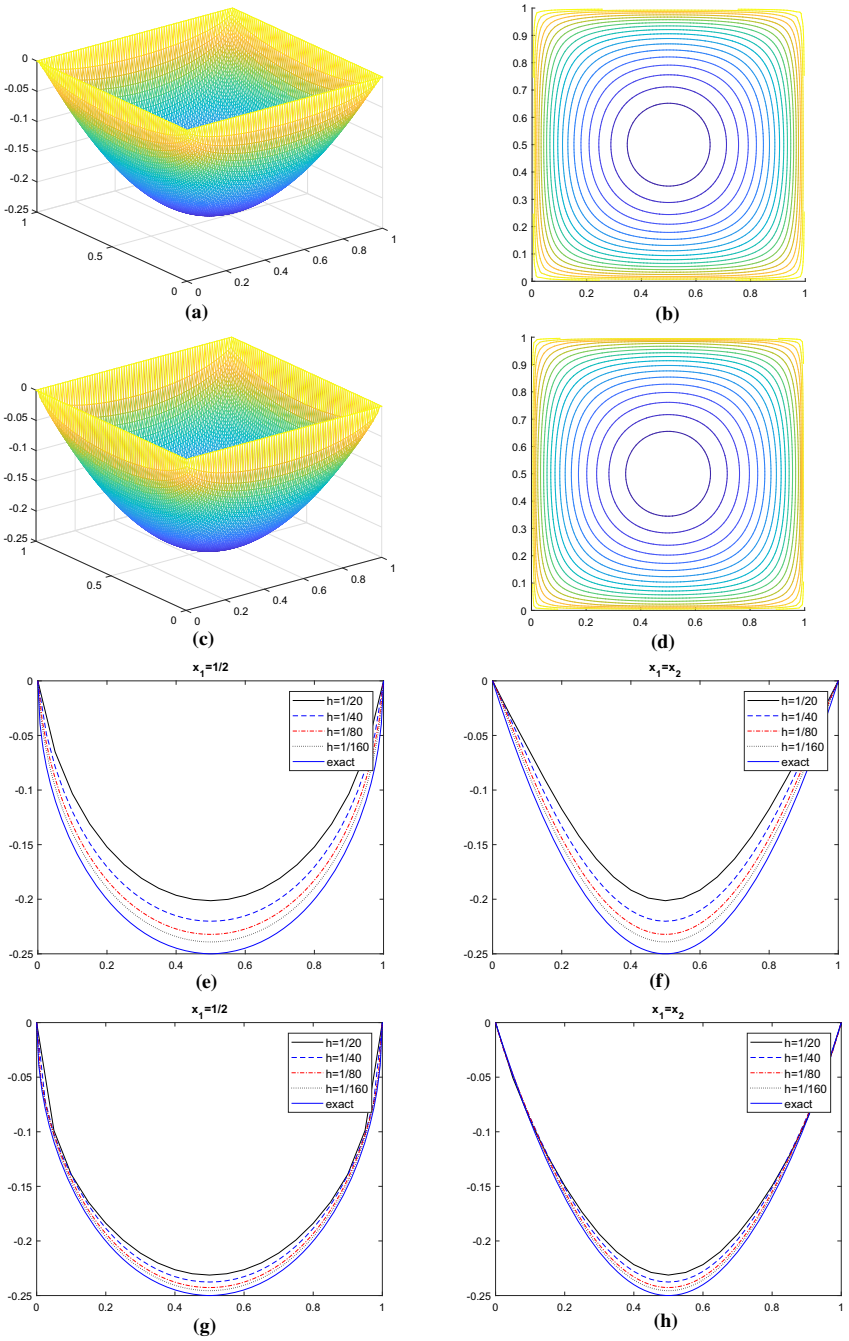

Fig. 17 [Test problem (56)]: **a**, **b**, **c**, **d** graphs and contours of the approximate solution obtained by the one-stage algorithm on a regular mesh with $h(= \Delta x_1 = \Delta x_2) = 1/80$. **e**, **g** Graphs of the restrictions of the computed approximate solutions to the line $x_1 = 1/2$. **f**, **h** Graphs of the restrictions of the computed approximate solutions to the line $x_1 = x_2$. The results visualized on **a**, **b** and **e**, **f** (resp., **c**, **d** and **g**, **h**) have been obtained using relations (23) [resp., (35), (36)] to approximate the second order derivatives

Table 17 (Test problem (57) for the unit square). For $h = 1/20, 1/40$ and $1/80$: (i) number of iterations needed for convergence of the one-stage algorithm and associated residuals. (ii) Minimal value of the computed approximate solutions. (iii) Discrete L_2 and L_∞ norms of the approximation errors and associated rates of convergence. Table (a) [resp., (b)] reports results associated with the approximation (58) [resp., (59)] of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$ (resp., $\rho = 1/16$). The minimal value of the exact solution is 0

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	2586	9.99×10^{-9}	-0.2186	6.27×10^{-2}		2.19×10^{-1}	
1/40	8323	9.99×10^{-9}	-0.0244	1.65×10^{-2}	1.93	2.95×10^{-2}	2.893
1/80	27,987	9.99×10^{-9}	0.0121	7.87×10^{-3}	1.07	1.21×10^{-2}	1.29
(b)							
1/20	887	9.99×10^{-9}	-0.1679	4.67×10^{-2}		1.68×10^{-1}	
1/40	3898	9.99×10^{-9}	-0.0386	2.14×10^{-2}	1.13	4.92×10^{-2}	1.77
1/80	17,327	9.99×10^{-9}	0.0057	8.40×10^{-3}	1.35	1.22×10^{-2}	2.01

In order to apply our methodology to the numerical solution of problem (57) we approximated (as in [8]) the measure $\pi\delta_{(1/2,1/2)}$ by the C^∞ strictly positive function f^1_η defined by

$$f^1_\eta(x_1, x_2) = \frac{\eta^2}{\left[\eta^2 + (x_1 - 1/2)^2 + (x_2 - 1/2)^2\right]^2}, \tag{58}$$

with η a small positive number having the dimension of a distance. In [8], good results were obtained using $\eta = h$. On the other hand, a 'good' function $h \to \eta(h)$ seems more difficult to identify for the methodology we use in the present article. Assuming that (with obvious notation) the relation $\lim_{\eta\to 0} \lim_{h\to 0} u^\eta_h = u$ holds, we fixed η and computed the associated approximate solution for various values of h, leading to the results reported and visualized in Tables 17, 18 and Figs. 18, 19. These results have been obtained using: (i) The one-stage algorithm with $\|u^{n+1} - u^n\|_2 < 10^{-8}$ as stopping criterion. (ii) Regular (resp., unstructured isotropic) triangulations of the unit square (resp., half unit disk). (iii) Approximations (35), (36) of the second order derivatives. Actually, we have also reported and visualized on those tables and figures, results obtained using the function f^2_ρ defined, with $\rho > 0$, by

$$f^2_\rho(x_1, x_2) = \frac{3}{\rho^3} \max\left(0, \rho - \sqrt{(x_1 - 1/2)^2 + (x_2 - 1/2)^2}\right) + h, \tag{59}$$

as approximation of the measure $\pi\delta_{(1/2,1/2)}$.

Remark 11 If Ω is a square, the solution of problem (57) (or closely related ones) has been addressed in various publications ([4,8,22,23] among others), by a variety of computational methods, some of them particularly fast like the ones discussed in [23]. On the other hand, solving (57) on a disk using piecewise affine approximations is a more challenging issue; it has been addressed in particular in [8], using finite element approximations closely related (but not identical) to those discussed in Sect. 4. □

We define as follows the last problem involving a Dirac measure we consider in this article:

$$\begin{cases} \det \mathbf{D}^2 u = 2\delta_{(1/2,1/2)} \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{60}$$

Table 18 (Test problem (57) for the half unit disk). For $h = 1/20, 1/40$ and $1/80$: (i) number of iterations needed for convergence of the one-stage algorithm and associated residuals. (ii) Minimal value of the computed approximate solutions. (iii) Discrete L_2 and L_∞ norms of the approximation errors and associated rates of convergence. Table (a) [resp., (b)] reports results associated with the approximation (58) [resp., (59)] of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$ (resp., $\rho = 1/16$). The minimal value of the exact solution is 0

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	1166	9.91×10^{-9}	-0.148	3.82×10^{-2}		1.54×10^{-1}	
1/40	4232	9.98×10^{-9}	-0.0177	1.67×10^{-2}	1.19	3.07×10^{-2}	2.33
1/80	14,883	9.99×10^{-9}	0.0106	8.35×10^{-3}	1	1.05×10^{-2}	1.55
(b)							
1/20	685	9.90×10^{-9}	-0.1231	3.44×10^{-2}		1.37×10^{-1}	
1/40	3475	9.99×10^{-9}	-0.0254	1.78×10^{-2}	0.95	4.13×10^{-2}	1.73
1/80	16,715	9.99×10^{-9}	0.0028	9.48×10^{-3}	0.91	1.55×10^{-2}	1.41

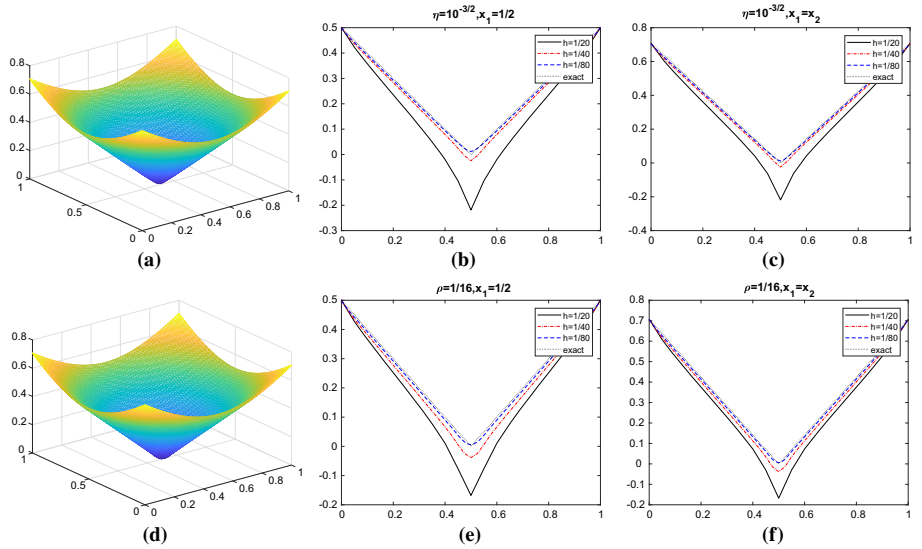

Fig. 18 (Test problem (57) for the unit square): **a**, **b**, **c** (resp., **d**, **e**, **f**) are associated with the approximation (58) with $\eta = 10^{-3/2}$ (resp., (59) with $\rho = 1/16$) of the measure $\pi\delta_{(1/2,1/2)}$. **a**, **d** Graphs of the computed solutions for $h = 1/80$. **b**, **e** Graphs of the restrictions of the computed solution to the line $x_1 = 1/2$ for $h = 1/20, 1/40$ and $1/80$. **c**, **f** Graphs of the restrictions of the computed solution to the line $x_1 = x_2$ for $h = 1/20, 1/40$ and $1/80$. The minimal value of the exact solution is 0

where in (60), Ω is the square $(0, 1)^2$ and $\delta_{(1/2,1/2)}$ is the Dirac measure at (1/2,1/2). It follows from [17] that the function u defined by

$$u(x_1, x_2) = -\min(x_1, (1 - x_1), x_2, (1 - x_2)), \forall (x_1, x_2) \in \bar{\Omega}$$

is an exact convex solution to problem (60), belonging to $W^{1,\infty}(\Omega)$. Its minimal value is clearly -0.5. In order to solve problem (60), we proceeded as follows: (i) For $h = 1/20, 1/40$ and 1/80, we used regular triangulations like the one in Fig. 1a. (ii) Employed the one-stage

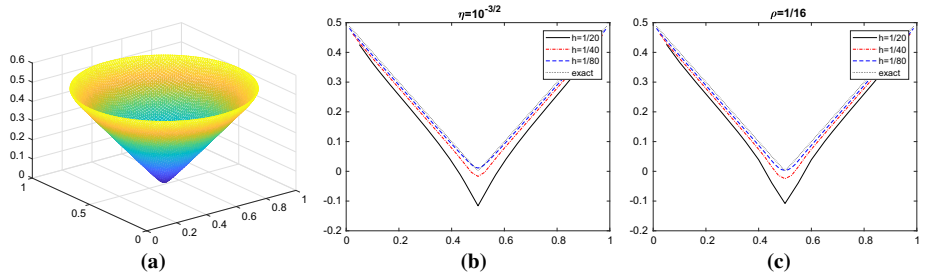

Fig. 19 (Test problem (57) for the half-unit disk): **a** graph of the computed approximate solution obtained for $h = 1/80$ by the one-stage algorithm, using the approximation (58) of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$. **b** Graphs of the restrictions of the computed approximate solutions to the diameter $x_1 = 1/2$, for $h = 1/20$, $1/40$ and $1/80$ (approximation (58) of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$). **c** Graphs of the restrictions of the computed approximate solutions to the diameter $x_1 = 1/2$, for $h = 1/20$, $1/40$ and $1/80$ (approximation (59) of the measure $\pi\delta_{(1/2,1/2)}$ with $\rho = 1/16$). The minimal value of the exact solution is 0

Table 19 [Test problem (60)]. For $h = 1/20$, $1/40$ and $1/80$: (i) number of iterations needed for convergence of the one-stage algorithm and associated residuals. (ii) Minimal value of the computed approximate solutions. (iii) Discrete L_2 and L_∞ norms of the approximation errors and associated rates of convergence. Table (a) [resp., (b)] reports results associated with the approximation (58) [resp., (59)] of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$ (resp., $\rho = 1/16$). The minimal value of the exact solution is -0.5

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Rate	L_∞ error	Rate
(a)							
1/20	4099	8.75×10^{-7}	-0.6356	3.67×10^{-2}		1.36×10^{-1}	
1/40	10,356	9.99×10^{-7}	-0.4937	8.56×10^{-3}	2.10	1.94×10^{-2}	2.81
1/80	14,970	9.99×10^{-9}	-0.4848	1.42×10^{-2}	–	2.28×10^{-2}	–
(b)							
1/20	2116	9.99×10^{-7}	-0.5949	2.45×10^{-2}		9.49×10^{-2}	
1/40	4847	1.06×10^{-6}	-0.5099	1.41×10^{-2}	0.80	2.43×10^{-2}	1.97
1/80	10,939	9.99×10^{-7}	-0.4860	1.05×10^{-2}	0.43	1.69×10^{-2}	0.52

algorithm, using $\|u^{n+1} - u^n\|_2 < 10^{-6}$ or 10^{-8} as stopping criterion. (iii) Approximated the second order derivatives using relations (35), (36). (iv) Approximated the measure $2\delta_{(1/2,1/2)}$ by $\frac{2}{\pi} f_\eta^1$ (resp. $\frac{2}{\pi} f_\rho^2$), with $\eta = 10^{-3/2}$ in (58) [resp., $\rho = 1/16$ in (59)]. The results we obtained have been reported and visualized in Table 19 and Fig. 20.

Table 19 and Fig. 20 support the assumption $\lim_{\eta\to 0}\lim_{h\to 0} u_h^\eta = \lim_{\eta\to 0}\lim_{h\to 0} u_h^\rho = u$, but they show also that a critical issue needing to be addressed is identifying how to pick η and ρ as functions of h, in order to improve the convergence of the approximate solutions when $h \to 0$; we intend to investigate this issue.

To conclude this section dedicated to the numerical solution of problem (1), when f is a *positive measure*, we consider the following Monge–Ampère problem:

$$\begin{cases} \det \mathbf{D}^2 u = f \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{61}$$

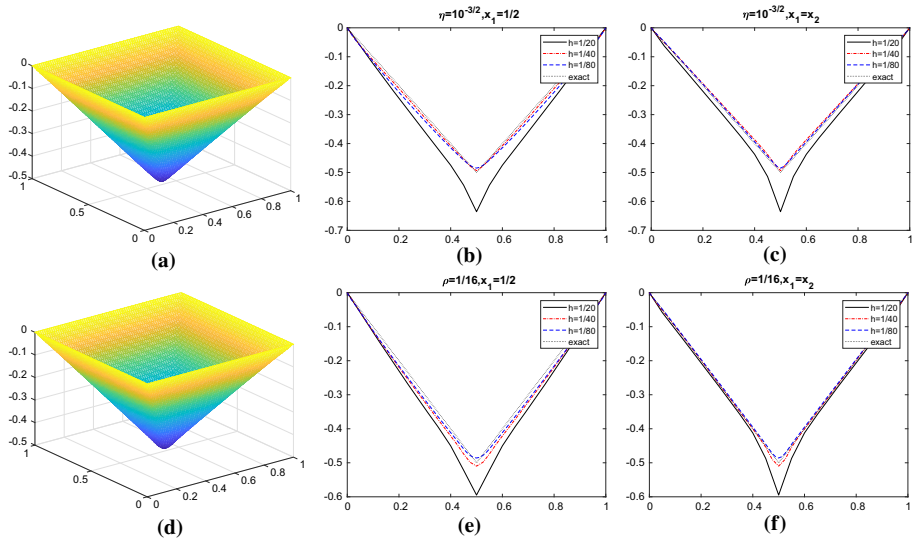

Fig. 20 [Test problem (60)] **a**, **d** graphs of the computed approximate solution obtained for $h = 1/80$ by the one-stage algorithm. **b**, **e** Graphs of the restrictions of the computed approximate solutions to the line $x_1 = 1/2$, for $h = 1/20$, $1/40$ and $1/80$. **c**, **f** Graphs of the restrictions of the computed approximate solutions to the line $x_1 = x_2$, for $h = 1/20$, $1/40$ and $1/80$. **a**, **b**, **c** (resp., **d**, **e**, **f**) correspond to the approximation (58) [resp., (59)] of the measure $\pi\delta_{(1/2,1/2)}$ with $\eta = 10^{-3/2}$ (reps., $\rho = 1/16$). The minimal value of the exact solution is -0.5

where in (62): (i) $\Omega = (0, 1)^2$, and (ii) f is the positive *measure* defined by

$$\langle f, \phi \rangle = \int_\gamma \phi d\gamma, \forall \phi \in H_0^1(\Omega), \tag{62}$$

the 'curve' γ being the cross defined by

$$\gamma = \{(x_1, x_2), 0 < x_1 < 1, x_2 = 1/2\} \cup \{(x_1, x_2), x_1 = 1/2, 0 < x_2 < 1\}.$$

In order to apply to the solution of problem (61) the methodology discussed in Sects. 2–5, we approximate the above measure f by f_η, a strictly positive C^∞ function defined by

$$f_\eta(x_1, x_2) = \eta^2 \left[\frac{1}{\left(\eta^2 + |x_1 - 1/2|^2\right)^{3/2}} + \frac{1}{\left(\eta^2 + |x_2 - 1/2|^2\right)^{3/2}} \right], \forall (x_1, x_2) \in \bar{\Omega}, \tag{63}$$

with $\eta > 0$. We can easily prove that $\lim_{\eta \to 0} f_\eta = f$ in the sense of distributions. The parameter η having the dimension of a distance, we used $\eta = h$ when applying the methodology discussed in Sects. 2–5 to the solution of

$$\begin{cases} \det \mathbf{D}^2 u_\eta = f_\eta \text{ in } \Omega, \\ u_\eta = 0 \text{ on } \partial\Omega. \end{cases} \tag{64}$$

The results reported and visualized in Table 20 and Fig. 21 have been obtain using: (i) Regular triangulations of Ω for $h = 1/20$, $1/40$ and $1/80$ (here, $h = \Delta x_1 = \Delta x_2$). (ii) The one-stage algorithm with $\|u^{n+1} - u^n\| < 10^{-8}$ as stopping criterion. (iii) The approximations (35), (36) of the second order derivatives.

Table 20 [Test problem (61)]. For $h = 1/20, 1/40$ and $1/80$: (i) Number of iterations needed for the convergence of the one-stage algorithm, and related residuals. (ii) Minimum values of the computed approximate solutions. We used $\eta = h$ in (63) when approximating the measure f by the function f_η

h	Iterations	$\|u^{n+1} - u^n\|$	Min
1/20	955	9.99×10^{-9}	-0.4917
1/40	5642	9.99×10^{-9}	-0.4905
1/80	39952	9.99×10^{-9}	-0.4923

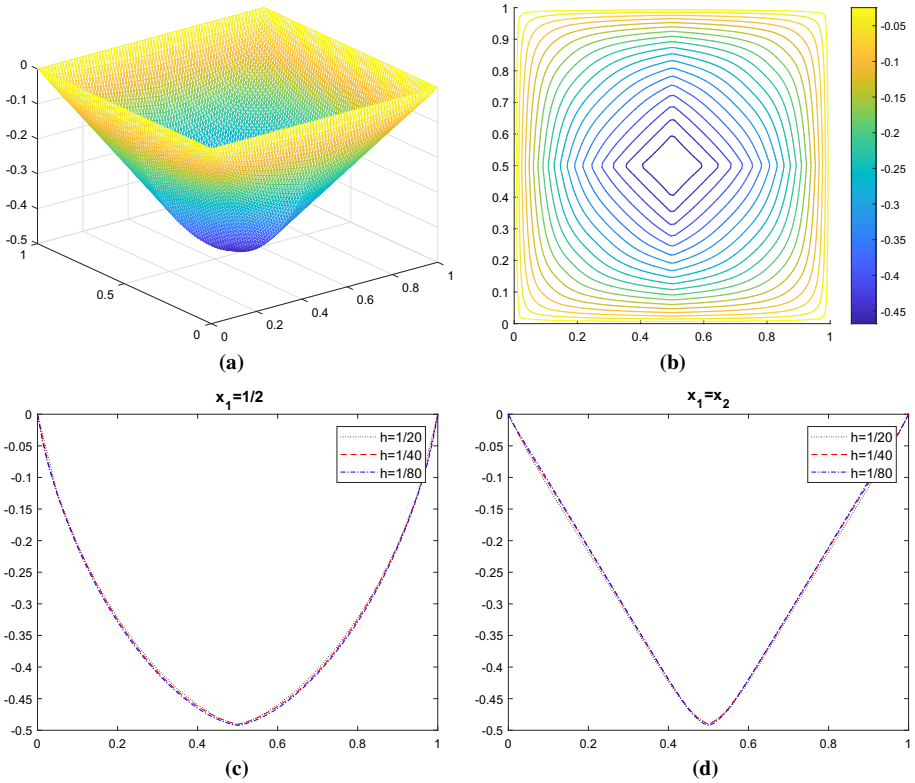

Fig. 21 [Test problem (61)]. **a**, **b** Graph and contours of the computed approximate solution for $h = 1/80$. **c** Graph of the restrictions of the computed approximate solutions to the line $x_1 = 1/2$, for $h = 1/20, 1/40$ and $1/80$. **d** Graphs of the restrictions of the computed approximate solutions to the line $x_1 = x_2$, for $h = 1/20, 1/40$ and $1/80$

From Fig. 21, we observe the convexity of the computed approximate solutions. Figure 21 suggests also a fast uniform convergence to a convex solution of problem (61) the choice $\eta = h$ works 'beautifully'.

Table 21 [Test problem (65)]. For $h = 1/20, 1/40, 1/80$ and $1/160$: (i) Number of iterations needed for the convergence of the one-stage algorithm, and associated residuals. (ii) Minimum values of the computed approximate solutions. (iii) L_2 and L_∞ norms of the approximation errors and related convergence rates. (a) $\eta = 0$. (b) $\eta = h$. (c) $\eta = h^2$. The minimal value of the exact solution is 0

h	Iterations	$\|u^{n+1} - u^n\|_2$	Min	L_2 error	Ratio	L_∞ error	Ratio
(a)							
1/20	66	9.51×10^{-8}	-3.44×10^{-2}	2.39×10^{-2}		3.45×10^{-2}	
1/40	345	9.92×10^{-8}	-3.55×10^{-2}	2.31×10^{-2}		3.56×10^{-2}	
1/80	924	9.98×10^{-8}	-3.66×10^{-2}	2.30×10^{-2}		3.67×10^{-2}	
(b)							
1/20	59	9.68×10^{-8}	-6.21×10^{-3}	1.65×10^{-3}		6.21×10^{-3}	
1/40	146	9.49×10^{-8}	-4.54×10^{-3}	1.18×10^{-3}	0.48	4.54×10^{-3}	0.45
1/80	382	9.89×10^{-8}	-3.10×10^{-3}	7.75×10^{-4}	0.61	3.10×10^{-3}	0.55
1/160	2019	9.95×10^{-9}	-2.08×10^{-3}	5.00×10^{-4}	0.63	2.08×10^{-3}	0.58
(c)							
1/20	231	9.89×10^{-8}	-4.10×10^{-4}	4.13×10^{-4}		1.09×10^{-3}	
1/40	803	9.99×10^{-8}	-1.54×10^{-4}	9.69×10^{-5}	2.09	2.95×10^{-4}	1.89
1/80	3618	9.99×10^{-9}	-1.85×10^{-4}	3.97×10^{-5}	1.29	1.85×10^{-4}	0.67
1/160	29,817	9.99×10^{-11}	-1.17×10^{-4}	2.38×10^{-5}	0.74	1.17×10^{-4}	0.66

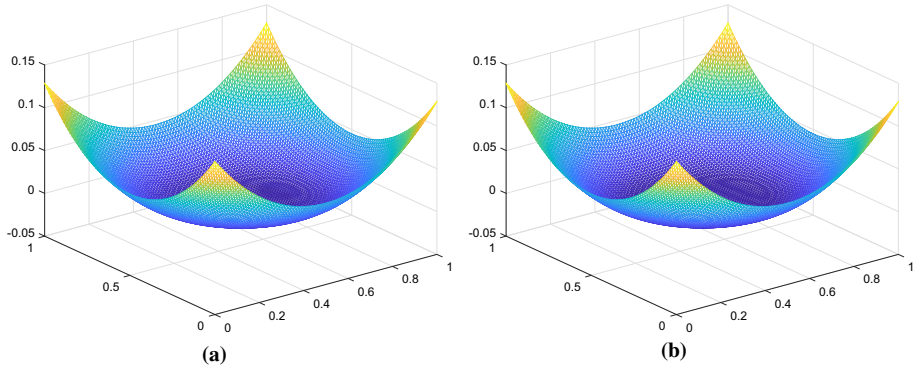

Fig. 22 [Test problem (65)]. Graphs of the computed approximate solutions for $h = 1/80$: **a** $\eta = h$. **b** $\eta = h^2$. At first glance, the two graphs are identical

6.7 A Kind of Obstacle Problem

The last test problem we will consider in this article is defined by

$$\begin{cases} \det \mathbf{D}^2 u = \max\left(1 - \frac{0.2}{r}, 0\right) \text{ in } \Omega, \\ u = \frac{1}{2}(\max(r - 0.2, 0))^2 \text{ on } \partial\Omega, \end{cases} \tag{65}$$

where in (65), $\Omega = (0, 1)^2$ and $r = \sqrt{(x - 1/2)^2 + (_2 - 1/2)^2}$. Problem (65) was suggested to us by one of the anonymous referees of this article; its solution has been addressed in, e.g., [22,37]. The function u defined by

$$u(x_1, x_2) = \frac{1}{2}\left(\max\left(\sqrt{(x_1 - 1/2)^2 + (x_2 - 1/2)^2} - 0.2, 0\right)\right)^2, \forall (x_1, x_2) \in \bar{\Omega},$$

is the exact convex solution to problem (65); we clearly have $u \in C^1(\bar{\Omega})$. Since the function u vanishes in the open disk of radius 0.2 centered at (1/2,1/2), $\det \mathbf{D}^2 u$ shares the same property, making the elliptic problem (65) degenerated. In order to facilitate the solution of problem (65), we approximated the function $\max(1 - \frac{0.2}{r}, 0)$ by $\max(1 - \frac{0.2}{r}, \eta)$, η being a small non-negative parameter converging to 0 with h. We have reported in Table 21 and Fig. 22, numerical results associated with $\eta = 0$, $\eta = h$ and $\eta = h^2$. These results have been obtained using: (i) Regular triangulations with $h(= \Delta x_1 = \Delta x_2) = 1/20$, 1/40, 1/80 and 1/160. (ii) The one-stage algorithm, the stopping criterion being $\|u^{n+1} - u^n\|_2 < tol$ with $10^{-10} \leq tol \leq 10^{-7}$. (iii) Approximations (23) of the second order derivatives.

From an accuracy point of view, the results obtained for $\eta = h$ [Table 21(b)] match those reported in [22,37], obtained using more sophisticated approximations. Moreover, Table 21(c) shows that taking $\eta = h^2$ increases accuracy by (roughly) one order of magnitude, the price to pay for this improvement being a much larger number of iterations in order to achieve the convergence of the one-stage algorithm.

Remark 12 We used the word *obstacle* in the title of this section. To justify this terminology, observe that, in (65), one can replace the equation

$$\det \mathbf{D}^2 u = \max\left(1 - \frac{0.2}{r}, 0\right)$$

by

$$\det \mathbf{D}^2 u = \left(1 - \frac{0.2}{r}\right) \chi_{(u>0)}, \tag{66}$$

where $\chi_{(u>0)}$ is the characteristic function of the set $\{x \in \Omega, u(x) > 0\}$. One encounters equations similar to (66) in [42], a publication dedicated to *obstacle problems* for the Monge–Ampère operator $\phi \to \det \mathbf{D}^2\phi$. The obstacle associated with problem (65) is clearly the plane $x_3 = 0$.

7 Conclusion

In this article, we have developed a relatively easy to implement finite element and operator-splitting based methodology for the numerical solution of the Monge–Ampère equation. The related methods have performed well for various types of triangulations (structured and unstructured) and can handle curved boundaries quite easily since they rely on continuous piecewise affine approximations. We introduced also a Newton-like two-stage variant of our methodology, which accelerates significantly the convergence if the problem under consideration has a smooth convex solution. Our future investigations (some of them quite advanced) will include looking at techniques to further accelerate the convergence of our iterative procedures, and the solution of three-dimensional and obstacle problems for the Monge–Ampère operator $\phi \to \det \mathbf{D}^2\phi$.

Acknowledgements The authors would like to thank the anonymous reviewers of this article for most helpful comments and suggestions. The work of S. Leung is partially supported by the Hong Kong RGC Grants 16303114 and 16309316. The work of J. Qian is partially supported by NSF Grants 1522249 and 1614566.

References

1. Awanou, G.: Standard finite elements for the numerical resolution of the elliptic Monge–Ampère equation: classical solutions. IMA J. Numer. Anal. **35**(3), 1150–1166 (2014)
2. Bakelman, I.J.: Convex Analysis and Nonlinear Geometric Elliptic Equations. Springer, Berlin (1994)
3. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. Numer. Math. **84**(3), 375–393 (2000)
4. Benamou, J.D., Froese, B.D., Oberman, A.M.: Two numerical methods for the elliptic Monge–Ampère equation. ESAIM Math. Model. Numer. Anal. **44**(4), 737–758 (2010)
5. Brenner, S., Gudi, T., Neilan, M., Sung, L.: C^0 penalty methods for the fully nonlinear Monge–Ampère equation. Math. Comput. **80**(276), 1979–1995 (2011)
6. Brenner, S.C., Neilan, M.: Finite element approximations of the three dimensional Monge–Ampère equation. ESAIM Math. Model. Numer. Anal. **46**(5), 979–1001 (2012)
7. Caboussat, A., Glowinski, R., Gourzoulidis, D.: A least-squares/relaxation method for the numerical solution of the three-dimensional elliptic Monge-Ampère equation. J. Sci. Comput. **77**, 1–26 (2018)
8. 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**(3), 780–810 (2013)
9. Caffarelli, L.A., Milman, M.: Monge–Ampère Equation: Applications to Geometry and Optimization: NSF-CBMS Conference on the Monge–Ampère Equation, Applications to Geometry and Optimization, July 9–13, 1997, Florida Atlantic University, Volume 226. American Mathematical Society, Providence (1999)
10. Caffarelli, L.A., Cabre, X.: Fully Nonlinear Elliptic Equations. American Mathematical Society, Providence (1995)

11. 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**(9), 779–784 (2003)
12. 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**(12), 887–892 (2004)
13. 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)
14. Dean, E.J., Glowinski, R.: Numerical methods for fully nonlinear elliptic equations of the Monge–Ampère type. Comput. Methods Appl. Mech. Eng. **195**(13), 1344–1386 (2006)
15. Dean, E.J., Glowinski, R.: On the numerical solution of the elliptic Monge–Ampère equation in dimension two: a least-squares approach. In: Glowinski, R., Neittaanmaki, P. (eds.) Partial Differential Equations, pp. 43–63. Springer, Dordrecht (2008)
16. Dean, E.J., Glowinski, R., Pan, T.W.: Operator-splitting methods and applications to the direct numerical simulation of particulate flow and to the solution of the elliptic Monge–Ampère equation. In: Cagnol, J., Zoésio, J.P. (eds.) Control Boundary Analysis, pp. 1–27. CRC, Boca Raton (2005)
17. D'Onofrio, L., Giannetti, F., Greco, L.: On weak Hessian determinants. In: Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti Lincei. Matematica e Applicazioni, vol. 16(3), pp. 159–169 (2005)
18. Feng, X., Glowinski, R., Neilan, M.: Recent developments in numerical methods for fully nonlinear second order partial differential equations. SIAM Rev. **55**(2), 205–267 (2013)
19. 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**(2), 1226–1250 (2009)
20. Feng, X., Neilan, M.: Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations. J. Sci. Comput. **38**(1), 74–98 (2009)
21. Froese, B.D.: Meshfree finite difference approximations for functions of the eigenvalues of the Hessian. Numer. Math. **138**(1), 75–99 (2018)
22. 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**(4), 1692–1714 (2011)
23. Froese, B.D., Oberman, A.M.: Fast finite difference solvers for singular solutions of the elliptic Monge–Ampère equation. J. Comput. Phys. **230**(3), 818–834 (2011)
24. Froese, B.D.: A numerical method for the elliptic Monge–Ampère equation with transport boundary conditions. SIAM J. Sci. Comput. **34**(3), A1432–A1459 (2012)
25. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (1983)
26. Glowinski, R.: Numerical Nethods for Nonlinear Variational Problems. Springer, New York (1984). (2nd printing: 2008)
27. Glowinski, R.: Finite element methods for incompressible viscous flow. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. IX, pp. 3–1176. North-Holland, Amsterdam (2003)
28. Glowinski, R.: Numerical methods for fully nonlinear elliptic equations. In: 6th International congress on industrial and applied mathermatics, ICIAM, vol. 7, pp. 155–192 (2009)
29. Glowinski, R.: Variational Methods for the Numerical Solution of Nonlinear Elliptic Problems. SIAM, Philadelphia (2015)
30. Glowinski, R., Osher, S., Yin, W. (eds.): Splitting Methods in Communication, Imaging, Science, and Engineering. Springer, Berlin (2016)
31. Gutiérrez, C.E.: The Monge–Ampère Equation. Birkhaüser, Basel (2001)
32. Hamfeldt, B.F., Salvador, T.: Higher-order adaptive finite difference methods for fully nonlinear elliptic equations. J. Sci. Comput. **75**(3), 1282–1306 (2018)
33. Lindsey, M., Rubinstein, Y.A.: Optimal transport via a Monge–Ampère optimization problem. SIAM J. Math. Anal. **49**(4), 3073–3124 (2017)
34. Loeper, G., Rapetti, F.: Numerical solution of the Monge–Ampère equation by a Newton algorithm. C. R. Acad. Sci. Paris Ser. I **340**, 319–324 (2005)
35. Mohammadi, B.: Optimal transport, shape optimization and global minimization. C. R. Math. **344**(9), 591–596 (2007)
36. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**(2), 372–376 (1983)
37. Nochetto, R., Ntogkas, D., Zhang, W.: Two-scale method for the Monge–Ampère equation: Convergence to the viscosity solution. In: Mathematics of Computation (2018). https://doi.org/10.1093/imanum/dry026

38. Oberman, A.M.: Wide stencil finite difference schemes for the elliptic Monge–Ampère equation and functions of the eigenvalues of the Hessian. Discrete Contin. Dyn. Syst. Ser. B **10**(1), 221–238 (2008)
39. Oliker, V.I., Prussner, L.D.: On the numerical solution of the equation $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ and its discretizations, i. Numer. Math. **54**(3), 271–293 (1989)
40. Persson, P.O., Strang, G.: A simple mesh generator in MATLAB. SIAM Rev. **46**(2), 329–345 (2004)
41. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
42. Savin, O.: The obstacle problem for Monge-Ampère equation. Calc. Var. Partial Differ. Equ. **22**(3), 303–320 (2005)
43. Schaeffer, H., Hou, T.Y.: An accelerated method for nonlinear elliptic PDE. J. Sci. Comput. **69**(2), 556–580 (2016)
44. Strang, G., Fix, G.J.: An Analysis of The Finite Element Method, vol. 212. Prentice-Hall, Englewood Cliffs (1973)
45. Su, W., Boyd, S., Candes, E.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**, 1–43 (2016)

Affiliations

Roland Glowinski[1,2] · Hao Liu[3] · Shingyu Leung[3] · Jianliang Qian[4]

Roland Glowinski
roland@math.uh.edu

Shingyu Leung
masyleung@ust.hk

Jianliang Qian
qian@math.msu.edu

1 Department of Mathematics, University of Houston, 4800 Calhoun Road, Houston, TX 77204, USA

2 Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong

3 Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

4 Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA