ESAIM: COCV 19 (2013) 780–810
DOI: 10.1051/cocv/2012033

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

A LEAST-SQUARES METHOD FOR THE NUMERICAL SOLUTION OF THE DIRICHLET PROBLEM FOR THE ELLIPTIC MONGE–AMPÈRE EQUATION IN DIMENSION TWO *

Alexandre Caboussat[1,2], Roland Glowinski[3] and Danny C. Sorensen[4]

Abstract. We address in this article the computation of the convex solutions of the Dirichlet problem for the real elliptic Monge–Ampère equation for general convex domains in two dimensions. The method we discuss combines a least-squares formulation with a relaxation method. This approach leads to a sequence of Poisson–Dirichlet problems and another sequence of low dimensional algebraic eigenvalue problems of a new type. Mixed finite element approximations with a smoothing procedure are used for the computer implementation of our least-squares/relaxation methodology. Domains with curved boundaries are easily accommodated. Numerical experiments show the convergence of the computed solutions to their continuous counterparts when such solutions exist. On the other hand, when classical solutions do not exist, our methodology produces solutions in a least-squares sense.

Résumé. Nous étudions, dans cet article, une méthode numérique, pour le calcul des solutions convexes du problème de Dirichlet pour l'équation de Monge–Ampère elliptique, dans des domaines bi-dimensionnel convexes. Une méthode de moindres carrés est couplée à un algorithme de relaxation, conduisant à la résolution d'une suite de problèmes de Poisson–Dirichlet, et d'une suite de problèmes de valeurs propres de petite dimension d'un type nouveau. Une approximation par éléments finis mixtes, couplée à une méthode de régularisation, est utilisée pour implémenter la méthode de moindres-carrés/relaxation ci-dessus, de sorte que les domaines avec frontière courbe sont traités facilement. Des expériences numériques montrent la convergence des solutions calculées vers la solution convexe du problème continu, lorsqu'une telle solution existe. Par ailleurs, si le problème n'a pas de solution classique, notre méthodologie fournit des solutions au sens des moindres carrés.

Mathematics Subject Classification. 65N30, 65K10, 65F30, 49M15, 49K20.

Received January 31, 2011. Revised August 21, 2012
Published online June 3, 2013.

Keywords and phrases. Monge–Ampère equation, least-squares method, biharmonic problem, conjugate gradient method, quadratic constraint minimization, mixed finite element methods.

* *This work was partially supported by the National Science Foundation (Grants NSF DMS-0913982 and DMS-0412267).*

[1] Haute École de Gestion/Geneva School of Business Administration, Genève, Switzerland. alexandre.caboussat@hesge.ch

[2] University of Houston, Department of Mathematics, 4800 Calhoun Rd, Houston, 77204-3008 Texas, USA. caboussat@math.uh.edu

[3] University of Houston, Department of Mathematics, 4800 Calhoun Rd, Houston, 77204-3008 Texas, USA. roland@math.uh.edu

[4] Rice University, Department of Computational and Applied Mathematics, MS 134, Houston, 77251-1892 Texas, USA. sorensen@rice.edu

1. Introduction

If f is *positive*, the canonical *Monge–Ampère equation*

$$\det \mathbf{D}^2\psi = f,$$

is considered by many mathematicians as the prototypical *fully nonlinear elliptic equation*. As such, it has recently received considerable attention from both the analytical and computational standpoints as shown by, *e.g.*, [2,5,6,14,24,37,40,41,43,44,50], with applications in geometry, mechanics and physics.

In particular, *augmented Lagrangian algorithms* and *least-squares techniques* have been used for the numerical solution of the Dirichlet problem for the Monge–Ampère equation in dimension two. These methods are discussed in [13,16–21,32,33]; actually, [32] contains a review of several methods for the solution of the Monge–Ampère equation and related fully nonlinear elliptic equations such as Pucci's.

Let Ω be a bounded, convex domain of $\mathbb{R}^2$; we denote by $\partial\Omega$ the boundary of Ω. Assuming that $f \in L^1(\Omega)$ and $g \in H^{3/2}(\partial\Omega)$, it makes sense (since the operator $\varphi \to \det \mathbf{D}^2\varphi$ is continuous from $H^2(\Omega)$ to $L^1(\Omega)$) to look in $H^2(\Omega)$ for the solutions (convex, in particular) of the Dirichlet problem for the Monge–Ampère equation, that is

$$\det \mathbf{D}^2\psi = f \quad \text{in } \Omega, \qquad \psi = g \quad \text{on } \partial\Omega; \tag{1.1}$$

see [10,11,20,37] for details. Suppose that problem (1.1) has convex (or concave) solutions with the H^2-regularity; if Ω is a square domain, using the augmented Lagrangian and least-squares methods discussed in [13,16–21,32,33], combined with piecewise linear continuous finite element approximations, one has been able to solve problem (1.1) rather accurately. Indeed, the numerical experiments reported in the above references show that the L^2 approximation error is $\mathcal{O}(h^2)$, which is, generically, optimal for second order elliptic problems, using this type of approximations. Using the above methodology, one has been able to compute least-squares solutions of (1.1) when, despite the smoothness of the data f and g, this problem has no classical solutions, as it is the case for example when $\Omega = (0,1)^2$, $f = 1$, and $g = 0$ [11,20,37]. Moreover, our method can be easily generalized to systems, unlike viscosity solutions which are based on maximum principles.

The least-squares methodology discussed in this article was introduced in [17] and further discussed in [21,32,33]. Actually, the most detailed account-published so far-of our least-squares approach can be found in [21] (for a detailed description of the augmented Lagrangian based methodology see [19]). The methodology discussed in [17,21,32,33] relies on the following ingredients:

(i) A well-chosen least-squares formulation in appropriate Hilbert spaces [4].
(ii) Associating with the optimality conditions of the above least-squares problem an initial value problem (flow in the dynamical system terminology).
(iii) The time-discretization of the above initial value problem by an operator-splitting scheme decoupling nonlinearity and differential operators.
(iv) The solution of the nonlinear (resp., linear) problems resulting from the splitting by a Newton's type algorithm (resp., by a preconditioned conjugate gradient algorithm).
(v) A mixed finite element approximation [8] of the Monge–Ampère problem (1.1) based on piecewise linear continuous approximations of ψ and of its three second order derivatives.

Actually, since (a) the speed of convergence of the operator-splitting based iterative method mentioned in (iii) improves as the time discretization step increases, and (b) the above algorithm reduces to a simpler to implement relaxation algorithm *à la* block Gauss–Seidel when the time discretization step converges to $+\infty$, it was decided that relaxation will be the method of choice to go beyond the methodology discussed in [17,21,32,33].

In [21] and related publications, all the test problems considered were posed in $\Omega = (0,1)^2$ and the finite element spaces were associated with uniform triangulations like the one on the left in Figure 2 (see Sect. 10). When applied to problems where Ω has a curved boundary requiring unstructured meshes, or when using uniform meshes like the one on the right in Figure 2, we observed a deterioration of the convergence properties

when $h \to 0$, and even divergence for some test problems. This issue is addressed in this article: an obvious way to overcome this difficulty is to proceed as in, *e.g.*, [24, 25], that is, use mixed finite element approximations of the convex solutions of problem (1.1), and of their second order derivatives, based on continuous, piecewise polynomial functions of degree ≥ 2. This approach has several drawbacks, the main ones being that: (i) unlike piecewise linear approximations, the higher order ones do not preserve the maximum principle when this principle holds. (ii) Compared to piecewise linear approximations, the higher order ones are not easy to implement for domains Ω with curved boundaries. Instead, in order to "rescue" the piecewise linear approximations, we advocate a *Tychonoff-like regularization method* [49] when defining the discrete analogues of the second order derivatives. With this approach we recover convergence of optimal (or nearly optimal) order, as $h \to 0$, even for unstructured meshes, or for pathological structured ones like the triangulation on the right in Figure 2.

To summarize, in this article, we advocate a *relaxation* algorithm for the solution of a well-chosen *least-squares* variant of problem (1.1). With such an algorithm we are able to *decouple* the treatment of the differential operators from the treatment of the nonlinearities. Indeed, the treatment of the differential operators leads to the solution of a sequence of elliptic linear biharmonic problems. The nonlinearity requires the solution of an infinite family of low dimensional constrained minimization problems, one for almost every point of Ω (in practice, one for each interior vertex of the finite element triangulation of Ω).

To solve the above linear biharmonic problems we advocate a conjugate gradient algorithm operating in well-chosen sub-spaces of $H^2(\Omega)$. On the other hand, two quite different methods are considered for the solution of the low dimensional constrained minimization problems: the first one based on the Newton's method combined with an appropriate parametrization of the two-dimensional manifold $\{\mathbf{z} = \{z_i\}_{i=1}^3,\ z_1 > 0,\ z_2 > 0,\ z_1 z_2 - z_3^2 = 1\}$. The second method is based on a novel algorithm for quadratically constrained minimization problems (denoted by $\mathbf{Q_{min}}$ and introduced in [48]). Following [16–22, 35], mixed finite element approximations are used for the discretization of (1.1). A regularization procedure for the approximation of second derivatives on arbitrary meshes allows obtaining optimal (or nearly optimal) convergence properties.

This article is structured as follows: in Section 2, we introduce some fundamental function spaces and sets, and use them to provide a least-squares formulation of problem (1.1). The relaxation algorithm is described in Section 3. In Sections 4 and 5, we discuss the solutions of the local low dimensional constrained minimization problems and of the linear variational bi-harmonic problems. The mixed finite element approximation of problem (1.1) is discussed in Section 6, while Sections 7, 8 and 9 are dedicated to the discrete analogues of the problems discussed in Sections 3, 4 and 5. In Section 10, the methodology discussed in the preceding sections is applied to the solution of test problems, some of them borrowed from [13, 16–21, 32, 33]; these numerical experiments include test cases where Ω has a curved boundary and/or when problem (1.1) has no solution in $H^2(\Omega)$ [11, 20, 37].

The methodology described in this article owes much to *Calculus of Variations* and *Optimal Control*. Indeed the least-squares criterion that we use is nothing but a multi-dimensional integral defined on the subset of a functional space *à la* Sobolev. Moreover *adjoint equation techniques* are used to compute some of the derivatives of the discrete cost functional, resulting in substantial memory and computational time savings.

2. Formulation of the Dirichlet problem for the elliptic Monge–Ampère equation in two dimensions

Let Ω be a bounded convex domain of $\mathbb{R}^2$; we denote by Γ the boundary of Ω. The Dirichlet problem for the canonical Monge–Ampère equation reads as follows:

$$\det \mathbf{D}^2\psi = f \ \text{ in } \Omega, \quad \psi = g \ \text{ on } \Gamma, \tag{2.1}$$

where $\mathbf{D}^2\psi$ is the *Hessian* of the unknown function ψ, that is $\mathbf{D}^2\psi = \left(\dfrac{\partial^2\psi}{\partial x_i \partial x_j}\right)_{1\leq i,j\leq 2}$.

When problem (2.1) has no solution (in $H^2(\Omega)$), we were tempted to call generalized solutions the solutions "captured" by the least-squares/relaxation methodology discussed in the following sections. Actually, in the

context of Monge–Ampère equations, *generalized solution* has a very precise meaning, namely the one introduced by Aleksandrov (see [1,46]) and further discussed in [37], Chapter 1. Following [37], ψ is a generalized solution of problem (2.1) if $M\psi = f$ in Ω and $\psi = g$ on $\partial\Omega$, $M\psi$ being the *Monge–Ampère measure* associated with ψ, a particular Borel measure whose precise definition can be found in [37], Chapter 1 (if ψ is smooth enough and convex, we can identify $M\psi$ with $\det \mathbf{D}^2\psi$).

Concerning the existence and uniqueness of generalized solutions, it is proved in the above reference that if $\Omega \subset \mathbb{R}^d$ is a bounded strictly convex domain, μ is a Borel measure in Ω with $\mu(\Omega) < +\infty$, and $g \in C^0(\partial\Omega)$, then there exists a unique $\psi \in C^0(\overline{\Omega})$ that is a convex solution to the boundary value problem $M\psi = \mu$ in Ω and $\psi = g$ on $\partial\Omega$. It is difficult to be more general, particularly if we compare with the very demanding conditions required from f, g and Ω, so that (2.1) will have a classical solution (see, *e.g.*, [29], Chap. 17 for details). It is also difficult to do better than Aleksandrov generalized solutions as long as problem (2.1) is concerned; unfortunately, the Aleksandrov's notion of weak solution does not generalize easily to other fully nonlinear second order elliptic equations.

For those more general situations, the right concept seems to be the notion of *viscosity solutions* introduced in the early eighties by M. Crandall and P.L. Lions. The basic reference concerning the viscosity solution of second order partial differential equations is [15] (for application to a variety of nonlinear elliptic equations, including Monge–Ampère's, see [10–12, 26, 39, 43] or [37] and the references therein). Actually, it is proved in [37], Chapter 1 that if ψ is a generalized solution to $M\psi = f$ with f continuous, then ψ is a viscosity solution of the Monge–Ampère equation. Conversely, it is also proved that if $f \in C^0(\overline{\Omega})$, $f > 0$, and ψ is a viscosity solution of $\det \mathbf{D}^2\psi = f$, then ψ is a generalized solution of $M\psi = f$. Numerical methods for the solution of problem (1.1), based on the Aleksandrov and viscosity solution approaches, can be found in [43,44].

Among the various methods available for the solution of (2.1), we advocate the following one of the *nonlinear least-squares* type:

$$\begin{cases} \text{Find } (\psi, \mathbf{p}) \in V_g \times \mathbf{Q}_f \text{ such that} \\ J(\psi, \mathbf{p}) \leq J(\varphi, \mathbf{q}), \quad \forall (\varphi, \mathbf{q}) \in V_g \times \mathbf{Q}_f, \end{cases} \tag{2.2}$$

where:

$$J(\varphi, \mathbf{q}) = \frac{1}{2} \int_\Omega \left| \mathbf{D}^2\varphi - \mathbf{q} \right|^2 \mathrm{d}\mathbf{x}, \tag{2.3}$$

$|\cdot|$ being the Fröbenius norm, that is $|\mathbf{T}| = \sqrt{\mathbf{T} : \mathbf{T}}$, with $\mathbf{S} : \mathbf{T} = \sum_{i,j=1}^2 s_{ij} t_{ij}$, for all $\mathbf{S} = (s_{ij})$, $\mathbf{T} = (t_{ij}) \in \mathbb{R}^{2\times 2}$. The functional spaces and sets in (2.2) are defined by:

$$V_g = \left\{ \varphi \in H^2(\Omega)\,,\ \varphi = g \ \text{ on } \Gamma \right\}, \tag{2.4}$$

$$\mathbf{Q}_f = \left\{ \mathbf{q} \in \mathbf{Q}\,,\ \det \mathbf{q} = f,\ q_{11} > 0,\ q_{22} > 0 \right\}, \quad \mathbf{Q} = \left\{ \mathbf{q} \in L^2(\Omega)^{2\times 2},\ \mathbf{q} = \mathbf{q}^t \right\}. \tag{2.5}$$

The space $\mathbf{Q}$ in (2.5) is a Hilbert space for the scalar product $(\mathbf{q}, \mathbf{q}') \to \int_\Omega \mathbf{q} : \mathbf{q}' \mathrm{d}\mathbf{x}$, and the associated norm. In order to have V_g and $\mathbf{Q}_f$ both non-empty, we assume from now on that $f \in L^1(\Omega)$, $f > 0$ and $g \in H^{3/2}(\Gamma)$. The introduction of the set $\mathbf{Q}_f$ allows the decoupling of the differential operators (acting linearly on the unknown function ψ) and of the nonlinearities (acting on the unknown tensor-valued function $\mathbf{p}$). Indeed, the burden of nonlinearity (algebraic here) has been reported on $\mathbf{p}$, explaining the introduction of the nonlinear manifold $\mathbf{Q}_f$.

Note that the existence and uniqueness of a convex solution to the least-squares problem (2.2) is still an open problem; however, the numerical experiments reported in Section 10, will show that our least-squares based method never failed finding the convex solution of the test problems considered there, assuming it exists with the right regularity properties (and sometimes even less, as shown in Sect. 10.12).

Remark 2.1. As shown in, *e.g.*, [19–21], problem (2.2) may have smooth solutions, even if (2.1) has no such solutions as it is the case if $\Omega = (0,1)^2$, $f = 1$ and $g = 0$ [11,37]. Generally speaking, (2.1) admits a smooth solution when $\mathbf{D}^2 V_g \cap \mathbf{Q}_f \neq \emptyset$, as illustrated in Figure 1 (left). On the other hand, when $\mathbf{D}^2 V_g \cap \mathbf{Q}_f = \emptyset$, it makes sense to search for a solution, in the sense of (2.2) (see Fig. 1 (right)).

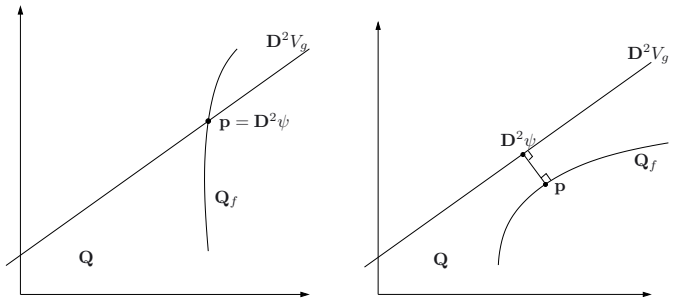

FIGURE 1. The Monge−Ampère problem (2.1) has a solution in V_g (left), or no solution in V_g (right).

3. A relaxation algorithm for the solution of problem (2.2)

In order to compute a *convex* solution of problem (2.2) (or at least to force the convexity of the solution) we advocate the following *relaxation algorithm*: solve

$$-\Delta\psi^0 = -2\sqrt{f} \ \text{ in } \Omega, \quad \psi^0 = g \ \text{ on } \Gamma. \tag{3.1}$$

Then, for $n \geq 0$, assuming that ψ^n is known, compute $\mathbf{p}^n, \psi^{n+1/2}$ and ψ^{n+1} as follows:

$$\mathbf{p}^n = \arg\min_{\mathbf{q}\in\mathbf{Q}_f} J(\psi^n, \mathbf{q}), \tag{3.2}$$

$$\psi^{n+1/2} = \arg\min_{\varphi\in V_g} J(\varphi, \mathbf{p}^n), \tag{3.3}$$

$$\psi^{n+1} = \psi^n + \omega(\psi^{n+1/2} - \psi^n), \tag{3.4}$$

with ω, $0 < \omega < \omega_{\max} \leq 2$, a relaxation parameter.

Remark 3.1 (initialization strategy)**.** The rationale behind (3.1) is as follows: denote by λ_1 and λ_2 the eigenvalues of $\mathbf{D}^2\psi$; we have then $\lambda_1\lambda_2 = f$. It follows from $(\lambda_1+\lambda_2)^2 - (\lambda_1-\lambda_2)^2 = 4\lambda_1\lambda_2$, that, if λ_1 and λ_2 are close to each other, then $\Delta\psi = \lambda_1 + \lambda_2 \simeq 2\sqrt{\lambda_1\lambda_2} = 2\sqrt{f}$, justifying thus the initialization (3.1).

The relaxation algorithm (3.1)–(3.4) looks simple but the solution of problems (3.2) and (3.3) leads to technical issues that we will address in the following sections.

4. Numerical solution of the sub-problems (3.2)

4.1. Explicit formulation of problem (3.2)

An explicit formulation of problem (3.2) is given by

$$\mathbf{p}^n = \arg\min_{\mathbf{q}\in\mathbf{Q}_f} \left[\frac{1}{2}\int_\Omega |\mathbf{q}|^2 \,\mathrm{d}\mathbf{x} - \int_\Omega \mathbf{D}^2\psi^n : \mathbf{q}\,\mathrm{d}\mathbf{x}\right]. \tag{4.1}$$

Since neither integrands in (4.1) contains derivatives of $\mathbf{q}$, the minimization problem (4.1) can be solved *pointwise* (in practice at the vertices of a finite element or finite difference grid). This leads us, a.e. in Ω, to the solution of the following finite dimensional minimization problem

$$\mathbf{p}^n(\mathbf{x}) = \arg\min_{\mathbf{q}\in\mathbf{E}_f(\mathbf{x})} \left[\frac{1}{2}|\mathbf{q}|^2 - \mathbf{D}^n(\mathbf{x}) : \mathbf{q}\right], \tag{4.2}$$

where $\mathbf{D}^n(\mathbf{x}) = \mathbf{D}^2\psi^n(\mathbf{x})$ is a symmetric matrix and $\mathbf{E}_f(\mathbf{x}) = \{\mathbf{q} \in \mathbb{R}^{2\times 2}, \mathbf{q} = \mathbf{q}^t, \det\mathbf{q} = f(\mathbf{x}), q_{11} > 0, q_{22} > 0\}$.

4.2. A Newton-type method for the numerical solution of problem (4.2)

Taking advantage of the symmetry of $\mathbf{q}$ and $\mathbf{D}^n(\mathbf{x})$, and using the notation $z_1 = q_{11}, z_2 = q_{22}, z_3 = q_{12} = q_{21}$ and $\mathbf{D}^n(\mathbf{x})_{ij} = d^n_{ij}(\mathbf{x})$, the minimization problem in (4.2) can be rewritten as

$$\min_{\mathbf{z}\in\mathbf{Z}_f(\mathbf{x})} \left[\frac{1}{2}(z_1^2 + z_2^2 + 2z_3^2) - d^n_{11}(\mathbf{x})z_1 - d^n_{22}(\mathbf{x})z_2 - 2d^n_{12}(\mathbf{x})z_3\right], \tag{4.3}$$

with $\mathbf{Z}_f(\mathbf{x}) = \{\mathbf{z}\in\mathbb{R}^3,\, z_1>0,\, z_2>0,\, z_1z_2 - z_3^2 = f(\mathbf{x})\}$. To transform (4.3) into an unconstrained minimization problem in $\mathbb{R}^2$, we perform the change of variables $z_1 = \sqrt{f(\mathbf{x})}e^{\rho}\cosh\theta$, $z_2 = \sqrt{f(\mathbf{x})}e^{-\rho}\cosh\theta$, $z_3 = \sqrt{f(\mathbf{x})}\sinh\theta$, for $(\rho,\theta)\in\mathbb{R}^2$, so that (4.3) becomes

$$\min_{(\rho,\theta)\in\mathbb{R}^2} j(\rho,\theta),$$

with $j(\rho,\theta) = \frac{\sqrt{f(\mathbf{x})}}{2}(\cosh 2\rho\cosh 2\theta + \cosh 2\rho + \cosh 2\theta - 1) - (d^n_{11}(\mathbf{x})e^{\rho} + d^n_{22}(\mathbf{x})e^{-\rho})\cosh\theta - 2d^n_{12}(\mathbf{x})\sinh\theta$. This leads us in turn to the solution of $Dj(\rho,\theta) = \mathbf{0}$, where $Dj(\cdot)$ is the differential of the functional $j(\cdot)$. This 2×2 nonlinear system actually reads as follows:

$$\begin{aligned} Dj(\rho,\theta)_1 &= \sqrt{f(\mathbf{x})}(1+\cosh 2\theta)\sinh 2\rho - (d^n_{11}(\mathbf{x})e^{\rho} - d^n_{22}(\mathbf{x})e^{-\rho})\cosh\theta &&= 0,\\ Dj(\rho,\theta)_2 &= \sqrt{f(\mathbf{x})}(1+\cosh 2\rho)\sinh 2\theta - (d^n_{11}(\mathbf{x})e^{\rho} + d^n_{22}(\mathbf{x})e^{-\rho})\sinh\theta - 2d^n_{12}(\mathbf{x})\cosh\theta &&= 0. \end{aligned}$$

This system can be solved by using a *Newton method*. Let $(\rho^0,\theta^0)\in\mathbb{R}^2$ be given. For $k\geq 0$, we compute $(\rho^{k+1},\theta^{k+1})$ from (ρ^k,θ^k) *via* the solution of

$$D^2j(\rho^k,\theta^k)\begin{pmatrix}\rho^{k+1}-\rho^k\\ \theta^{k+1}-\theta^k\end{pmatrix} = -Dj(\rho^k,\theta^k),$$

where $D^2j(\rho,\theta) = (D^2j(\rho,\theta)_{ij})_{1\leq i,j\leq 2}$ is given by:

$$\begin{aligned} D^2j(\rho,\theta)_{11} &= 2\sqrt{f(\mathbf{x})}\cosh 2\rho(1+\cosh 2\theta) - \left(d^n_{11}(\mathbf{x})e^{\rho} + d^n_{22}(\mathbf{x})e^{-\rho}\right)\cosh\theta,\\ D^2j(\rho,\theta)_{12} &= D^2j(\rho,\theta)_{21} = 2\sqrt{f(\mathbf{x})}\sinh 2\rho\sinh 2\theta - \left(d^n_{11}(\mathbf{x})e^{\rho} - d^n_{22}(\mathbf{x})e^{-\rho}\right)\sinh\theta,\\ D^2j(\rho,\theta)_{22} &= 2\sqrt{f(\mathbf{x})}\cosh 2\theta(1+\cosh 2\rho) - \left(d^n_{11}(\mathbf{x})e^{\rho} + d^n_{22}(\mathbf{x})e^{-\rho}\right)\cosh\theta - 2d_{12}(\mathbf{x})\sinh\theta. \end{aligned}$$

Remark 4.1 (choice of the scalar product)**.** Since we are dealing with symmetric matrices, we can equip $\mathbf{Q}$ with the following scalar product $(\mathbf{q},\mathbf{q}') \to \int_\Omega (q_{11}q'_{11} + q_{22}q'_{22} + q_{12}q'_{12})\mathrm{d}\mathbf{x}$. As shown in [13, 18, 32], this new scalar product has given better results than the one defined by $(\mathbf{q},\mathbf{q}') \to \int_\Omega \mathbf{q} : \mathbf{q}'\mathrm{d}\mathbf{x}$ when applied to the numerical solution of the two-dimensional Dirichlet problem for the *Pucci's equation*, that is $\alpha\lambda^+ + \lambda^- = f$ in Ω, together with $\psi = g$ on Γ, where λ^+ (resp., λ^-) denotes the largest (resp., the smallest) eigenvalue of the Hessian $\mathbf{D}^2\psi$ of the unknown function ψ, and where $\alpha\geq 1$. Using this new scalar product, (4.3) would be replaced by

$$\min_{\mathbf{z}\in\mathbf{Z}_f(\mathbf{x})} \left[\frac{1}{2}\left(z_1^2 + z_2^2 + z_3^2\right) - d^n_{11}(\mathbf{x})z_1 - d^n_{22}(\mathbf{x})z_2 - d^n_{12}(\mathbf{x})z_3\right],$$

with $\mathbf{Z}_f(\mathbf{x})$ defined similarly. The same change of variables and Newton method can be applied to this problem.

4.3. The quadratically constrained minimization method for the numerical solution of problem (4.3)

In [48], a class of quadratically constrained minimization problems has been addressed with a new algorithm denoted by $\mathbf{Q}_{\min}$. This algorithm allows the solution of some specific eigenvalue-constrained matrix optimization problems of dimension N ($\geq$2), its complexity being $\mathcal{O}(N^3)$. The particular case associated with $N=2$ corresponds to (4.2).

This method relies on the following equivalent formulation of problem (4.2):

$$\mathbf{p}^n(\mathbf{x}) = \mathbf{S}^n(\mathbf{x})\boldsymbol{\Lambda}^n(\mathbf{x})\mathbf{S}^n(\mathbf{x})^t, \quad (\boldsymbol{\Lambda}^n(\mathbf{x}), \mathbf{S}^n(\mathbf{x})) = \arg \min_{(\boldsymbol{\Lambda},\mathbf{S})\in\mathcal{E}_f} \left[\frac{1}{2}(\mu_1^2+\mu_2^2) - \text{trace}\left(\mathbf{D}^n(\mathbf{x})\mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^t\right)\right], \tag{4.4}$$

where $\mathcal{E}_f(\mathbf{x}) = \{(\boldsymbol{\Lambda}, \mathbf{S}), \boldsymbol{\Lambda} = \text{diag}(\mu_1, \mu_2), \mu_1\mu_2 = f(\mathbf{x}), \mathbf{S}^t\mathbf{S} = \mathbf{I}\}$. The algorithm developed in [48] applies beautifully to the solution of (4.4). After scaling of $\mathbf{D}^n(\mathbf{x})$ by $\sqrt{f(\mathbf{x})}$, (4.4) is equivalent to

$$\arg \min_{\mathbf{A}\in\mathcal{A}_1} \text{trace}\left[\mathbf{A}\mathbf{A} - 2\mathbf{D}^n\mathbf{A}\right], \tag{4.5}$$

where $\mathcal{A}_1 = \{\mathbf{A} \in \mathbf{R}^{2\times 2}, \mathbf{A} = \mathbf{A}^t, \ell^t\mathbf{M}\ell = 2, \mathbf{M}\ell \geq 0\}$, $\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and $\ell = (\mu_1, \mu_2)^t$, $\{\mu_1, \mu_2\}$ being the spectrum of $\mathbf{A}$. The constraint $\ell^t\mathbf{M}\ell = 2$ corresponds to $\lambda_1 + \lambda_2 = 1$, while the constraint $\mathbf{M}\ell \geq 0$ ensures $\lambda_1, \lambda_2 \geq 0$ to obtain convex solutions. Ultimately, for $N = 2$ the solution is found by solving a simple rational equation of the form

$$\frac{\beta_1^2}{(1+\mu)^2} = 2 + \frac{\beta_2^2}{(1-\mu)^2},$$

where $\beta_1 = (\lambda_1 + \lambda_2)/\sqrt{2}$ and $\beta_2^2 = (\lambda_1^2 + \lambda_2^2)/2 - \lambda_1\lambda_2$, $\{\lambda_1, \lambda_2\}$ being the spectrum of $\mathbf{D}^n(\mathbf{x})/\sqrt{f(\mathbf{x})}$. Remarkably, the same rational equation holds essentially for arbitrary $N \geq 2$. This equation is efficiently solved numerically by first taking reciprocals and then square roots on both sides and applying Newton's method. With a starting guess $\mu_0 = -1$, the method converges typically in 3 to 5 iterations. This occurs because the reciprocal square root transformation yields a problem that is essentially the intersection of two straight lines. For more details, see [48], where this algorithm is developed for arbitrary $N \geq 2$.

5. Conjugate gradient solution of the sub-problems (3.3)

Written in variational form, the Euler−Lagrange equation of the sub-problem (3.3) reads as follows:

$$\text{Find } \psi^{n+1/2} \in V_g \text{ such that } \int_\Omega \mathbf{D}^2\psi^{n+1/2} : \mathbf{D}^2\varphi d\mathbf{x} = \int_\Omega \mathbf{p}^n : \mathbf{D}^2\varphi d\mathbf{x}, \quad \forall\varphi \in V_0, \tag{5.1}$$

where $V_0 = H^2(\Omega) \cap H_0^1(\Omega)$. The linear variational problem (5.1) is well-posed and belongs to the following family of linear variational problems:

$$u \in V_g \ : \ \int_\Omega \mathbf{D}^2 u : \mathbf{D}^2 v d\mathbf{x} = L(v), \quad \forall v \in V_0, \tag{5.2}$$

with the functional $L(\cdot)$ linear and continuous over $H^2(\Omega)$; problem (5.2) is clearly of the biharmonic type. The *conjugate gradient solution* of linear variational problems in Hilbert spaces, such as (5.2), has been addressed in, *e.g.*, [30], Chapter 3. Following the above reference, we are going to solve (5.2) by a conjugate gradient algorithm operating in the spaces V_0 and V_g, both spaces being equipped with the scalar product defined by $(v, w) \to \int_\Omega \Delta v \Delta w d\mathbf{x}$, and the corresponding norm. This conjugate gradient algorithm reads as follows:

Step 1.

$$u^0 \in V_g \text{ given.} \tag{5.3}$$

Step 2. Solve:

$$\text{Find } g^0 \in V_0 \text{ such that } \int_\Omega \Delta g^0 \Delta v d\mathbf{x} = \int_\Omega \mathbf{D}^2 u^0 : \mathbf{D}^2 v d\mathbf{x} - L(v), \quad \forall v \in V_0, \tag{5.4}$$

and set the first descent direction:

$$w^0 = g^0. \tag{5.5}$$

Then, for $k \geq 0$, u^k, g^k, and w^k being known, the last two different from zero, we compute u^{k+1}, g^{k+1} and, if necessary, w^{k+1} as follows.

Step 3. Solve:

$$\text{Find } \bar{g}^k \in V_0 \text{ such that } \int_\Omega \Delta\bar{g}^k \Delta v \mathrm{d}\mathbf{x} = \int_\Omega \mathbf{D}^2 w^k : \mathbf{D}^2 v \mathrm{d}\mathbf{x}, \quad \forall v \in V_0, \tag{5.6}$$

and compute the new iterates as follows:

$$\rho_k = \frac{\int_\Omega |\Delta g^k|^2 \,\mathrm{d}\mathbf{x}}{\int_\Omega \Delta\bar{g}^k \Delta w^k \mathrm{d}\mathbf{x}}, \tag{5.7}$$

$$u^{k+1} = u^k - \rho_k w^k, \tag{5.8}$$

$$g^{k+1} = g^k - \rho_k \bar{g}^k. \tag{5.9}$$

Step 4. Compute

$$\delta_k = \frac{\int_\Omega |\Delta g^{k+1}|^2 \,\mathrm{d}\mathbf{x}}{\int_\Omega |\Delta g^0|^2 \,\mathrm{d}\mathbf{x}}. \tag{5.10}$$

If $\delta_k < \varepsilon$ (meaning that the residual is small enough), take $u = u^{k+1}$; otherwise, compute:

$$\gamma_k = \frac{\int_\Omega |\Delta g^{k+1}|^2 \,\mathrm{d}\mathbf{x}}{\int_\Omega |\Delta g^k|^2 \,\mathrm{d}\mathbf{x}}, \tag{5.11}$$

and update the descent direction *via*

$$w^{k+1} = g^{k+1} + \gamma_k w^k. \tag{5.12}$$

Step 5. Do $k+1 \to k$ and return to **Step 3**.

The numerical experiments reported in Section 10 show that the conjugate gradient algorithm (5.3)–(5.12) enjoys a fast convergence, typically less than 10 iterations for all the meshes and mesh sizes which have been considered. Combined with an appropriate mixed finite element approximation, it requires, at each iteration, the solution of two discrete Poisson problems.

6. On a mixed finite element approximation

6.1. Generalities

Considering the highly variational flavor of the methodology discussed in the preceding sections, it makes sense to look for finite element based methods for the approximation of (2.1). In order to avoid the complications associated with the construction of finite element sub-spaces of $H^2(\Omega)$ (see, however, [5,25] for such an approach), we employ here a mixed finite element approximation (closely related to those discussed in, *e.g.*, [22, 23, 31, 36, 47] for the solution of linear and nonlinear bi-harmonic problems). Following this approach, it is possible to solve (2.1) employing approximations commonly used for the solution of second order elliptic problems (piecewise linear and globally continuous over a triangulation of Ω for example). The use of low order finite elements is justified in order to have the flexibility to consider computational domains with arbitrary (convex) shapes. However, since low order finite elements may have difficulties at handling some biharmonic problems, an additional regularization may be required, when approximating the second derivatives.

6.2. Mixed finite element approximation

For simplicity, we assume that Ω is a bounded polygonal domain of $\mathbb{R}^2$. Let us denote by $\mathcal{T}_h$ a finite element triangulation of Ω as discussed in, *e.g.*, [31], Appendix 1. From $\mathcal{T}_h$, we approximate the spaces $L^2(\Omega)$, $H^1(\Omega)$ and $H^2(\Omega)$ (respectively, $H^1_0(\Omega)$ and $H^2(\Omega)\cap H^1_0(\Omega)$) by the finite dimensional space V_h (respectively, V_{0h}) defined by:

$$V_h=\left\{v\in C^0\left(\overline{\Omega}\right),\ v|_T\in\mathbb{P}_1,\ \forall T\in\mathcal{T}_h\right\},\quad V_{0h}=V_h\cap H^1_0(\Omega)=\left\{v\in V_h,\ v=0\text{ on }\Gamma\right\},\tag{6.1}$$

with $\mathbb{P}_1$ the space of the two-variables polynomials of degree ≤ 1.

For a function φ being given in $H^2(\Omega)$, we denote $\partial^2\varphi/\partial x_i\partial x_j$ by $D^2_{ij}(\varphi)$. It follows from *Green's formula* that

$$\int_\Omega \frac{\partial^2\varphi}{\partial x_i\partial x_j}v\mathrm{d}\mathbf{x}=-\frac{1}{2}\int_\Omega\left[\frac{\partial\varphi}{\partial x_i}\frac{\partial v}{\partial x_j}+\frac{\partial\varphi}{\partial x_j}\frac{\partial v}{\partial x_i}\right]\mathrm{d}\mathbf{x},\ \forall v\in H^1_0(\Omega),\ \forall i,j=1,2.\tag{6.2}$$

Consider now $\varphi\in V_h$. Taking advantage of the relations (6.2), we define the discrete analogues of the differential operators D^2_{ij} by

$$D^2_{hij}(\varphi)\in V_{0h},\quad \int_\Omega D^2_{hij}(\varphi)v\mathrm{d}\mathbf{x}=-\frac{1}{2}\int_\Omega\left[\frac{\partial\varphi}{\partial x_i}\frac{\partial v}{\partial x_j}+\frac{\partial\varphi}{\partial x_j}\frac{\partial v}{\partial x_i}\right]\mathrm{d}\mathbf{x},\ \forall v\in V_{0h},\ \forall i,j=1,2.\tag{6.3}$$

The functions $D^2_{hij}(\varphi)$ are uniquely defined by the relations (6.3). However, in order to simplify the computation of the above discrete second order partial derivatives, it is tempting to consider using the trapezoidal rule to evaluate the integrals in the left hand sides of (6.3). Owing to their practical importance, let us detail these calculations:

(i) First, we introduce the set Σ_h of the vertices of $\mathcal{T}_h$ and then $\Sigma_{0h}=\{P\in\Sigma_h,\ P\notin\Gamma\}$. Next, we define the integers N_h and N_{0h} by $N_h=\mathrm{Card}(\Sigma_h)$ and $N_{0h}=\mathrm{Card}(\Sigma_{0h})$. We have then $\dim V_h=N_h$ and $\dim V_{0h}=N_{0h}$. We suppose that $\Sigma_{0h}=\{P_j\}_{j=1}^{N_{0h}}$ and $\Sigma_h=\Sigma_{0h}\cup\{P_j\}_{j=N_{0h}+1}^{N_h}$.

(ii) With each $P_k\in\Sigma_h$, we associate the function w^k uniquely defined by

$$w^k\in V_h,\ w^k(P_k)=1,\ w^k(P_l)=0,\ \forall l=1,\ldots,N_h,\ l\neq k.$$

It is well-known (see, *e.g.*, [31], Appendix 1) that the sets $\mathcal{B}_h=\{w^k\}_{k=1}^{N_h}$ and $\mathcal{B}_{0h}=\{w^k\}_{k=1}^{N_{0h}}$ are *vector bases* for V_h and V_{0h}, respectively.

(iii) Let us denote by A_k the area of the polygonal domain which is the union of those triangles of $\mathcal{T}_h$ which have P_k as a common vertex. Applying the trapezoidal rule to the integrals in the left-hand side of the relations (6.3), we obtain

$$D^2_{hij}(\varphi)\in V_{0h},\quad D^2_{hij}(\varphi)(P_k)=-\frac{3}{2A_k}\int_\Omega\left[\frac{\partial\varphi}{\partial x_i}\frac{\partial w^k}{\partial x_j}+\frac{\partial\varphi}{\partial x_j}\frac{\partial w^k}{\partial x_i}\right]\mathrm{d}\mathbf{x},\ \forall k=1,\ldots,N_{0h},\ \forall i,j=1,2.\tag{6.4}$$

Computing the integrals in the right hand side of (6.4) is quite simple since the first order derivatives of φ and w^k are *piecewise constant*. Finally, with $\varphi\in V_h$, we associate $\Delta_h\varphi\in V_{0h}$ uniquely defined by $\Delta_h\varphi(P_k)=D^2_{h11}(\varphi)(P_k)+D^2_{h22}(\varphi)(P_k)$, for $k=1,\ldots,N_{0h}$.

Taking the above relations into account, approximating problem (2.1) is now fairly straightforward. Assuming that the boundary function g is continuous over Γ (which is definitely the case if $g\in H^{3/2}(\Gamma)$), let us denote by g_h the interpolant of g associated with the triangulation $\mathcal{T}_h$. We approximate the affine space V_g by $V_{gh}=\{\varphi\in V_h,\ \varphi(P)=g(P),\forall P\in\Sigma_h\cap\Gamma\}$ and then problem (2.1) by:

$$\text{Find}\ \ \psi_h\in V_{gh}\ \ \text{such that}\ \ D^2_{h11}(\psi_h)(P_k)D^2_{h22}(\psi_h)(P_k)-\left|D^2_{h12}(\psi_h)(P_k)\right|^2=f_h(P_k),\quad k=1,\ldots,N_{0h},\tag{6.5}$$

where f_h is a continuous approximation of f (we can always assume that $f_h \in V_h$). In addition, we define the discrete equivalent of $\mathbf{Q}_f$ as follows:

$$\mathbf{Q}_{fh} = \{\mathbf{q} \in \mathbf{Q}_h\,,\, \det \mathbf{q}(P_k) = f_h(P_k),\ q_{11}(P_k) > 0,\ q_{22}(P_k) > 0,\ k = 1, \ldots, N_{0h}\}\,,$$

with $\mathbf{Q}_h = \{\mathbf{q} \in (V_{0h})^{2\times 2},\ \mathbf{q}(P_k) = \mathbf{q}^t(P_k),\ k = 1, \ldots, N_{0h}\}$. We associate with V_{0h} and $\mathbf{Q}_h$ the following discrete scalar products and corresponding Euclidean norms:

$$(v, w)_{0h} = \frac{1}{3}\sum_{k=1}^{N_h} A_k v(P_k) w(P_k),\ \forall v, w \in V_{0h}, \qquad ||v||_{0h}^2 = (v, v)_{0h},\ \forall v \in V_{0h},$$

$$((\mathbf{S}, \mathbf{T}))_{0h} = \frac{1}{3}\sum_{k=1}^{N_{0h}} A_k \mathbf{S}(P_k) : \mathbf{T}(P_k),\ \forall \mathbf{S}, \mathbf{T} \in \mathbf{Q}_h, \quad |||\mathbf{S}|||_{0h}^2 = ((\mathbf{S}, \mathbf{S}))_{0h},\ \forall \mathbf{S} \in \mathbf{Q}_h.$$

The solution of problem (6.5) will be discussed in the sequel.

Remark 6.1. Suppose that $\Omega = (0, 1)^2$ and that the triangulation $\mathcal{T}_h$ is uniform like the one shown in Figure 2 (left). Suppose that $h = 1/(I + 1)$, I being a positive integer greater than one. In this particular case, the sets Σ_h and Σ_{0h} are given by $\Sigma_h = \{P_{ij} = (ih, jh)\,,\, 0 \le i, j \le I + 1\}$, and $\Sigma_{0h} = \{P_{ij} = (ih, jh)\,,\, 1 \le i, j \le I\}$, implying that $N_h = (I+2)^2$ and $N_{0h} = I^2$. It follows then from the relations (6.4) that (with obvious notation):

$$D_{h11}^2(\varphi)(P_{ij}) = \frac{\varphi_{i+1,j} + \varphi_{i-1,j} - 2\varphi_{ij}}{h^2}, \quad 1 \le i, j \le I,$$

$$D_{h22}^2(\varphi)(P_{ij}) = \frac{\varphi_{i,j+1} + \varphi_{i,j-1} - 2\varphi_{ij}}{h^2}, \quad 1 \le i, j \le I,$$

$$D_{h12}^2(\varphi)(P_{ij}) = \frac{\varphi_{i+1,j+1} + \varphi_{i-1,j-1} + 2\varphi_{ij} - (\varphi_{i+1,j} + \varphi_{i-1,j} + \varphi_{i,j+1} + \varphi_{i,j-1})}{2h^2}, \quad 1 \le i, j \le I.$$

The above discrete second order derivatives of finite difference type have the easily verified yet remarkable property that they are *exact for polynomial functions of degree* ≤ 2.

6.3. A smoothing procedure for the approximation of the second derivatives

As emphasized in [45], when using piecewise linear mixed finite elements, the *a priori* estimates for the error on the second derivatives of the solution ψ are, in general, $\mathcal{O}(1)$ in the L^2-norm. Therefore the convergence properties of the solution method depend strongly on the type of triangulations one employs. Indeed, assuming that the discrete second order derivatives have been computed *via* (6.3) and (6.4), numerical experiments performed by the authors showed the triangulation dependence of the convergence; non-convergence cases (in the L^2-norm) were also observed. Unfortunately, the approximations of $\dfrac{\partial^2 \varphi}{\partial x_i \partial x_j}$ provided by (6.3) and (6.4) converge to the above second derivative, no better than in $H^{-1}(\Omega)$ in general. This allows oscillations and explains the growth of the approximation error in $L^2(\Omega)$ and $H^1(\Omega)$ as $h \to 0$. Such pathological behavior can be observed in the results presented in Section 10. From that point of view a dramatic confirmation of these non-convergence properties is provided by the numerical results associated with the structured symmetric mesh shown on the right of Figure 2 (also called "British flag" mesh or "crisscross" pattern). To cure the non-convergence properties associated with the approximations (6.3) and (6.4) of the second derivatives, we see two options:

(i) Use, as in, *e.g.*, [24, 25], mixed finite elements methods based on piecewise polynomial approximations of degree $\ge$2. This approach has several drawbacks, among them: (a) it is more complicated to implement than the mixed methods described in Section 6.2, particularly if Ω has a curved boundary. (b) These higher order polynomial approximations do not preserve the maximum principle, if this principle takes place for the continuous problem.

(ii) Use a *regularization procedure à la Tychonoff* [49], while keeping a piecewise linear approximation based mixed finite element approach.

Focusing on the second approach, a simple and novel (in this context) way to obtain better convergence properties of the discrete second order derivatives is to use the following regularization procedure: with $C > 0$ and $|K| = \text{meas}(K)$, when computing the discrete second derivatives $D^2_{hij}(\varphi)$ replace (6.3) by:

$$\begin{aligned}&\text{Find } D^2_{hij}(\varphi) \in V_{0h} \text{ such that, } \forall v \in V_{0h},\ i,j = 1,2,\\ &\int_\Omega D^2_{hij}(\varphi) v \mathrm{d}\mathbf{x} + C \sum_{K\in\mathcal{T}_h} |K| \int_K \nabla D^2_{hij}(\varphi)\cdot\nabla v \mathrm{d}\mathbf{x} = -\frac{1}{2}\int_\Omega \left[\frac{\partial\varphi}{\partial x_i}\frac{\partial v}{\partial x_j} + \frac{\partial\varphi}{\partial x_j}\frac{\partial v}{\partial x_i}\right]\mathrm{d}\mathbf{x},\end{aligned} \tag{6.6}$$

and (6.4) by

$$\begin{aligned}&\text{Find } D^2_{hij}(\varphi) \in V_{0h} \text{ such that, } \forall v \in V_{0h},\ i,j = 1,2,\\ &(D^2_{hij}(\varphi), v)_{0h} + C \sum_{K\in\mathcal{T}_h} |K| \int_K \nabla D^2_{hij}(\varphi)\cdot\nabla v \mathrm{d}\mathbf{x} = -\frac{1}{2}\int_\Omega \left[\frac{\partial\varphi}{\partial x_i}\frac{\partial v}{\partial x_j} + \frac{\partial\varphi}{\partial x_j}\frac{\partial v}{\partial x_i}\right]\mathrm{d}\mathbf{x}.\end{aligned} \tag{6.7}$$

The above linear systems can be solved by a sparse Cholesky solver (with the Cholesky factorization made once and for all at the beginning of the algorithm). The overhead in computational time appears to be non significant. Numerical results in Section 10 show that the above regularization procedure generally provides a significant improvement to the orders of convergence of the approximations of the solution ψ of problem (2.1). On the other hand, in the particular case of triangulations like the one on the left of Figure 2, the regularization associated with (6.6) or (6.7), deteriorates significantly the $L^2(\Omega)$-approximation error, while preserving optimal orders of convergence.

Remark 6.2. The regularization method we employed in (6.6) and (6.7) is reminiscent of the stabilization one employed by Hughes *et al.* in [38] to construct convergent approximations of the Stokes problem using, essentially, the same finite element spaces to approximate velocity and pressure (equal-order interpolation), a very popular method indeed.

7. Discrete least-squares formulation and discrete relaxation algorithm

We advocate the following *nonlinear least-squares* method for the solution of problem (6.5):

$$\text{Find } (\psi_h, \mathbf{p}_h) \in V_{gh}\times\mathbf{Q}_{fh} \text{ such that } J_h(\psi_h,\mathbf{p}_h) \le J_h(\varphi,\mathbf{q}),\quad \forall(\varphi,\mathbf{q})\in V_{gh}\times\mathbf{Q}_{fh}, \tag{7.1}$$

where

$$J_h(\varphi,\mathbf{q}) = \frac{1}{2}\left|\left|\left|\mathbf{D}^2_h(\varphi) - \mathbf{q}\right|\right|\right|^2_{0h}.$$

In order to solve the nonlinear least-squares problem (7.1), we suggest the following *relaxation* algorithm:

$$\text{Find } \psi^0_h \in V_{gh} \text{ such that } \int_\Omega \nabla\psi^0_h\cdot\nabla\varphi\mathrm{d}\mathbf{x} = -2(\sqrt{f_h},\varphi)_{0h},\quad \forall\varphi\in V_{0h}. \tag{7.2}$$

For $n \ge 0$, assuming that ψ^n_h is known, compute $\mathbf{p}^n_h, \psi^{n+1/2}_h$ and ψ^{n+1}_h as follows:

$$\mathbf{p}^n_h = \arg\min_{\mathbf{q}\in\mathbf{Q}_{fh}} J_h(\psi^n_h,\mathbf{q}), \tag{7.3}$$

$$\psi^{n+1/2}_h = \arg\min_{\varphi\in V_{gh}} J_h(\varphi,\mathbf{p}^n_h), \tag{7.4}$$

$$\psi^{n+1}_h = \psi^n_h + \omega(\psi^{n+1/2}_h - \psi^n_h), \tag{7.5}$$

with $0 < \omega < \omega_{\max} \le 2$. The solution of the finite dimensional problems (7.3) and (7.4) will be addressed in the following sections.

8. Numerical solution of the discrete sub-problems (7.3)

An explicit formulation of problem (7.3) is given by

$$\mathbf{p}_h^n = \arg \min_{\mathbf{q}\in\mathbf{Q}_{fh}} \left[\frac{1}{2} |||\mathbf{q}|||_{0h}^2 - ((\mathbf{D}_h^2(\psi_h^n), \mathbf{q}))_{0h} \right].$$

This minimization problem can be solved *point-wise*, at each vertex of $\mathcal{T}_h$ belonging to Σ_{0h}, that is:

$$\mathbf{p}_h^n(P_k) = \arg \min_{\mathbf{q}\in\mathbf{E}_{fh}(P_k)} \left[\frac{1}{2} |\mathbf{q}|^2 - \mathbf{D}_h^n(P_k) : \mathbf{q} \right], \ k = 1, \ldots, N_{0h},$$

where $\mathbf{D}_h^n(P_k) = \mathbf{D}_h^2(\psi_h^n)(P_k)$ and $\mathbf{E}_{fh}(P_k) = \left\{ \mathbf{q} \in \mathbb{R}^{2\times 2}, \mathbf{q} = \mathbf{q}^t, \det \mathbf{q} = f_h(P_k), q_{11} > 0, q_{22} > 0 \right\}$. Both the Newton's and the $\mathbf{Q}_{\min}$ methods presented in Section 4 apply here, after replacing $\mathbf{x}$ by P_k, $k = 1, \ldots, N_{0h}$.

9. Conjugate gradient solution of the discrete sub-problems (7.4)

9.1. Formulation of (7.4) as a discrete linear variational problem

The *Euler–Lagrange equation* associated with problem (7.4) reads as follows:

$$\text{Find } \psi_h^{n+1/2} \in V_{gh} \text{ such that } ((\mathbf{D}_h^2(\psi_h^{n+1/2}), \mathbf{D}_h^2(\varphi)))_{0h} = ((\mathbf{p}_h^n, \mathbf{D}_h^2(\varphi)))_{0h}, \quad \forall \varphi \in V_{0h}. \tag{9.1}$$

Problem (9.1) is a well-posed linear variational problem in the affine space V_{gh}. Following [30], Chapter 3, the solution of problem (9.1) will be discussed in Section 9.3. However, as written, the linear problem (9.1) leads to excessive computer resource requirements. This is easy to understand: to derive the linear system equivalent to (9.1), we need to compute-via the solution of (6.6) or (6.7)-the matrix-valued functions $\mathbf{D}_h^2(w^j)$, where the functions w^j form a basis of V_{0h}. To avoid this difficulty, we are going to employ an *adjoint equation* approach to derive an equivalent formulation of (9.1), well-suited to solution by a conjugate gradient algorithm.

9.2. An adjoint equation based equivalent formulation of problem (9.1)

Problem (9.1) is equivalent to:

$$\text{Find } \psi_h^{n+1/2} \in V_{gh} \text{ such that } \left\langle \frac{\partial J_h}{\partial \varphi}(\psi_h^{n+1/2}, \mathbf{p}_h^n), \theta \right\rangle = 0, \quad \forall \theta \in V_{0h}, \tag{9.2}$$

where, more generally, $\left\langle \frac{\partial J_h}{\partial \varphi}(\varphi, \mathbf{q}), \theta \right\rangle$ denotes the action of the partial derivative $\frac{\partial J_h}{\partial \varphi}(\varphi, \mathbf{q})$ on the test function θ. Suppose that $\mathbf{D}_h^2(\varphi)$ is obtained from φ *via* relations (6.7); proceeding as in, *e.g.*, [34] one can easily show that, for all $(\varphi, \mathbf{p}) \in V_{gh} \times \mathbf{Q}_h$:

$$\left\langle \frac{\partial J_h}{\partial \varphi}(\varphi, \mathbf{q}), \theta \right\rangle = \int_\Omega \left[\frac{\partial \lambda_{11}}{\partial x_1} \frac{\partial \theta}{\partial x_1} + \frac{\partial \lambda_{22}}{\partial x_2} \frac{\partial \theta}{\partial x_2} + \frac{\partial \lambda_{12}}{\partial x_1} \frac{\partial \theta}{\partial x_2} + \frac{\partial \lambda_{12}}{\partial x_2} \frac{\partial \theta}{\partial x_1} \right] \mathrm{d}\mathbf{x}, \quad \forall \theta \in V_{0h}, \tag{9.3}$$

where $(\lambda_{11}, \lambda_{12}, \lambda_{22})$ is obtained from φ *via* the solution of the following (adjoint) system, for $1 \le i \le j \le 2$:

$$\lambda_{ij} \in V_{0h}, \quad (\lambda_{ij}, \theta)_{0h} + C \sum_{K\in\mathcal{T}_h} |K| \int_K \nabla\lambda_{ij} \cdot \nabla\theta \mathrm{d}\mathbf{x} = (q_{ij} - D_{hij}^2(\varphi), \theta)_{0h}, \quad \forall \theta \in V_{0h}. \tag{9.4}$$

Modifying the adjoint system (9.4), in order to handle (6.6) instead of (6.7) is straightforward. The solvers used to compute the $D_{hij}^2(\varphi)$ (*via* (6.6) or (6.7)) still apply to the solution of the linear problems in (9.4).

9.3. Conjugate gradient solution of problem (9.1)

Assume that $D^2_{hij}(\varphi)$ is obtained from φ *via* (6.7). Then, for the solution of problem (9.1), we can use a conjugate gradient algorithm operating in the spaces V_{0h} and V_{gh} equipped with the scalar product $(v,w) \to (\Delta_h v, \Delta_h w)_{0h}$ and the associated norm. Taking advantage of the results of Section 9.2, this algorithm reads as follows:

Step 1.

$$\psi_h^{n+1/2,0} \in V_{gh} \quad \text{given } (\psi_h^{n+1/2,0} = \psi_h^n \text{ for example}). \tag{9.5}$$

Compute $D^2_{hij}(\psi_h^{n+1/2,0})$ *via* the solution of:
Find $D^2_{hij}(\psi_h^{n+1/2,0}) \in V_{0h}$ such that, for $1 \le i,j \le 2$:

$$(D^2_{hij}(\psi_h^{n+1/2,0}),\theta)_{0h} + C \sum_{K\in\mathcal{T}_h} |K| \int_\Omega \nabla D^2_{hij}(\psi_h^{n+1/2,0}) \cdot \nabla\theta \mathrm{d}\mathbf{x} = - \frac{1}{2} \int_\Omega \left[\frac{\partial \psi_h^{n+1/2,0}}{\partial x_i} \frac{\partial \theta}{\partial x_j} + \frac{\partial \psi_h^{n+1/2,0}}{\partial x_j} \frac{\partial \theta}{\partial x_i} \right] \mathrm{d}\mathbf{x}, \ \forall \theta \in V_{0h}. \tag{9.6}$$

and then $(\lambda_{11}^{n+1/2,0}, \lambda_{12}^{n+1/2,0}, \lambda_{22}^{n+1/2,0}) \in (V_{0h})^3$ *via* the solution of the adjoint system:
Find $\lambda_{ij}^{n+1/2,0} \in V_{0h}$, for $1 \le i \le j \le 2$, such that:

$$(\lambda_{ij}^{n+1/2,0},\theta)_{0h} + C \sum_{K\in\mathcal{T}_h} |K| \int_K \nabla \lambda_{ij}^{n+1/2,0} \cdot \nabla\theta \mathrm{d}\mathbf{x} = (p_{ij}^n - D^2_{hij}(\psi_h^{n+1/2,0}),\theta)_{0h}, \quad \forall \theta \in V_{0h}. \tag{9.7}$$

Step 2. Solve:
Find $g^{n+1/2,0} \in V_{0h}$ such that

$$(\Delta_h g^{n+1/2,0}, \Delta_h \varphi)_{0h} = \int_\Omega \left[\frac{\partial \lambda_{11}^{n+1/2,0}}{\partial x_1} \frac{\partial \varphi}{\partial x_1} + \frac{\partial \lambda_{22}^{n+1/2,0}}{\partial x_2} \frac{\partial \varphi}{\partial x_2} + \frac{\partial \lambda_{12}^{n+1/2,0}}{\partial x_1} \frac{\partial \varphi}{\partial x_2} + \frac{\partial \lambda_{12}^{n+1/2,0}}{\partial x_2} \frac{\partial \varphi}{\partial x_1} \right] \mathrm{d}\mathbf{x}, \qquad \forall \varphi \in V_{0h}, \tag{9.8}$$

and set

$$w^{n+1/2,0} = g^{n+1/2,0}. \tag{9.9}$$

Then, for $k \ge 0$, assuming that $\psi_h^{n+1/2,k}, g^{n+1/2,k}$ and $w^{n+1/2,k}$ are known, the last two different from zero, we compute $\psi_h^{n+1/2,k+1}, g^{n+1/2,k+1}$ and, if necessary, $w^{n+1/2,k+1}$ as follows.

Step 3. Compute $D^2_{hij}(w^{n+1/2,k})$ *via* the solution of:
Find $D^2_{hij}(w^{n+1/2,k}) \in V_{0h}$ such that, for $1 \le i,j \le 2$:

$$(D^2_{hij}(w^{n+1/2,k}),\theta)_{0h} + C \sum_{K\in\mathcal{T}_h} |K| \int_\Omega \nabla D^2_{hij}(w^{n+1/2,k}) \cdot \nabla\theta \mathrm{d}\mathbf{x} = - \frac{1}{2} \int_\Omega \left[\frac{\partial w^{n+1/2,k}}{\partial x_i} \frac{\partial \theta}{\partial x_j} + \frac{\partial w^{n+1/2,k}}{\partial x_j} \frac{\partial \theta}{\partial x_i} \right] \mathrm{d}\mathbf{x}, \ \forall \theta \in V_{0h}, \tag{9.10}$$

and then $(\bar{\lambda}_{11}^{n+1/2,k}, \bar{\lambda}_{12}^{n+1/2,k}, \bar{\lambda}_{22}^{n+1/2,k}) \in (V_{0h})^3$ *via* the solution of the adjoint system:
Find $\bar{\lambda}_{ij}^{n+1/2,k} \in V_{0h}$, for $1 \leq i \leq j \leq 2$, such that:

$$(\bar{\lambda}_{ij}^{n+1/2,k}, \theta)_{0h} + C \sum_{K \in \mathcal{T}_h} |K| \int_K \nabla \bar{\lambda}_{ij}^{n+1/2,k} \cdot \nabla \theta \mathrm{d}\mathbf{x} = -(D_{hij}^2(w^{n+1/2,k}), \theta)_{0h}, \quad \forall \theta \in V_{0h}. \tag{9.11}$$

Solve:
Find $\bar{g}^{n+1/2,k} \in V_{0h}$ such that

$$(\Delta_h \bar{g}^{n+1/2,k}, \Delta_h \varphi)_{0h} =$$

$$\int_\Omega \left[\frac{\partial \bar{\lambda}_{11}^{n+1/2,k}}{\partial x_1} \frac{\partial \varphi}{\partial x_1} + \frac{\partial \bar{\lambda}_{22}^{n+1/2,k}}{\partial x_2} \frac{\partial \varphi}{\partial x_2} + \frac{\partial \bar{\lambda}_{12}^{n+1/2,k}}{\partial x_1} \frac{\partial \varphi}{\partial x_2} + \frac{\partial \bar{\lambda}_{12}^{n+1/2,k}}{\partial x_2} \frac{\partial \varphi}{\partial x_1} \right] \mathrm{d}\mathbf{x}, \qquad \forall \varphi \in V_{0h}, \tag{9.12}$$

and compute the new iterate and residual as follows:

$$\rho_k^{n+1/2} = \frac{\left|\left|\Delta_h g^{n+1/2,k}\right|\right|_{0h}^2}{(\Delta_h \bar{g}^{n+1/2,k}, \Delta_h w^{n+1/2,k})_{0h}}, \tag{9.13}$$

$$\psi_h^{n+1/2,k+1} = \psi_h^{n+1/2,k} - \rho_k^{n+1/2} w^{n+1/2,k}, \tag{9.14}$$

$$g^{n+1/2,k+1} = g^{n+1/2,k} - \rho_k^{n+1/2} \bar{g}^{n+1/2,k}. \tag{9.15}$$

Step 4. Compute

$$\delta_k^{n+1/2} = \frac{\left|\left|\Delta_h g^{n+1/2,k+1}\right|\right|_{0h}^2}{\left|\left|\Delta_h g^{n+1/2,0}\right|\right|_{0h}^2}. \tag{9.16}$$

If $\delta_k^{n+1/2} < \varepsilon$ (meaning that the residual is small enough), take $\psi_h^{n+1/2} = \psi_h^{n+1/2,k+1}$; otherwise, compute:

$$\gamma_k^{n+1/2} = \frac{\left|\left|\Delta_h g^{n+1/2,k+1}\right|\right|_{0h}^2}{\left|\left|\Delta_h g^{n+1/2,k}\right|\right|_{0h}^2}, \tag{9.17}$$

and update the descent direction *via*

$$w^{n+1/2,k+1} = g^{n+1/2,k+1} + \gamma_k^{n+1/2} w^{n+1/2,k}. \tag{9.18}$$

Step 5. Do $k+1 \to k$ and return to **Step 3**.

Remark 9.1. Modifying algorithm (9.5)–(9.18) in order to accommodate the construction of the discrete second order derivatives associated with (6.6) is straightforward. Since the results of numerical experiments (not reported in this article) have shown that the method based on (6.6) is no more accurate than the one based on (6.7) we will focus on the latter, which has also the advantage of being less computer time consuming, everything else being the same.

Remark 9.2. The choice of ε in the stopping criterion of algorithm (9.5)–(9.18) is a delicate issue which has been briefly discussed in [30], Chapter 3 (see also the references therein). As expected other stopping criteria are possible, a rather natural one being

$$\frac{\left(\Delta_h g^{n+1/2,k+1}, \Delta_h g^{n+1/2,k+1}\right)_{0h}}{\max\left\{\left(\Delta_h g^{n+1/2,0}, \Delta_h g^{n+1/2,0}\right)_{0h}, \left(\Delta_h \psi_h^{n+1/2,k+1}, \Delta_h \psi_h^{n+1/2,k+1}\right)_{0h}\right\}} < \varepsilon.$$

Remark 9.3 (solution of the biharmonic problems)**.** Concerning the solution of the discrete bi-harmonic problems in (9.8) and (9.12), let us observe that both problems are of the following type:

$$\text{Find } r_h \in V_{0h} \text{ such that } (\Delta_h r_h, \Delta_h v)_{0h} = \Lambda_h(v), \quad \forall v \in V_{0h}, \tag{9.19}$$

the functional $\Lambda_h(\cdot)$ being linear over V_h. Let us denote $-\Delta_h r_h$ by ω_h. It follows then from (6.4) that problem (9.19) is equivalent to the following system of two coupled discrete Poisson−Dirichlet problems

$$\begin{aligned} &\omega_h \in V_{0h}, \quad \int_\Omega \nabla\omega_h \cdot \nabla v \mathrm{d}\mathbf{x} = \Lambda_h(v), \quad \forall v \in V_{0h}, \\ &r_h \in V_{0h}, \quad \int_\Omega \nabla r_h \cdot \nabla v \mathrm{d}\mathbf{x} = (\omega_h, v)_{0h}, \quad \forall v \in V_{0h}. \end{aligned} \tag{9.20}$$

Both problems are well-posed. Actually, the solution (by direct or iterative methods) of discrete Poisson problems, such as (9.20) has motivated an important literature; some related references can be found in [30], Chapter 5.

10. Numerical experiments

10.1. Generalities

In this section, we shall validate the methodology discussed in Sections 2 to 9. The validation will be achieved *via* the solution of a variety of test problems associated with domains Ω of different shapes, including some with curved boundaries. We will investigate, in particular, the mesh dependence of the computed solutions. The results of our numerical experiments suggest that the methodology based on the regularization procedure associated with relations (6.6) and (6.7) is the only one, so far, able to solve the Monge−Ampère problem (2.1) accurately on domains of arbitrary convex shapes using piecewise linear continuous approximations on unstructured finite element meshes.

The first test problems to be considered concern (not surprisingly) the case where Ω is the unit square $(0,1)^2$. In order to study the mesh dependence of the computed solution, the three types of triangulations visualized in Figure 2 have been used. The structured triangulations (resp., the un-structured one) have been built using Modulef [3] (resp., Gmsh [28]). The uniform triangulation on the left of Figure 2 is called asymmetric despite the fact that it has some (but not many) symmetry properties; this terminology has been used to distinguish it from triangulations, like the one on the right of Figure 2, which have many symmetry properties. Recall (see Remark 6.1) that on uniform asymmetric triangulations, the discrete second order derivatives provided by relation (6.4) are exact for polynomial functions of degree ≤ 2.

10.2. First test problem

In this section, and below, we have denoted by $||\cdot||_{0h}$ the discrete variants of the L^2-errors (obtained by numerical integration). The *first test problem* that we consider is defined by

$$\det \mathbf{D}^2\psi(x_1,x_2) = 1, \ \forall (x_1,x_2) \in \Omega = (0,1)^2, \qquad \psi(x_1,x_2) = \frac{5}{2}x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2, \ \forall (x_1,x_2) \in \Gamma. \tag{10.1}$$

This convex solution of the Monge−Ampère−Dirichlet problem (10.1) is the function ψ given by

$$\psi(x_1,x_2) = \frac{5}{2}x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2, \ \forall (x_1,x_2) \in \Omega. \tag{10.2}$$

Note that, following [11,37] for instance, the convex solution is unique.

Its solution being a convex polynomial of degree 2, problem (10.1) looks rather simple. The condition number of $\mathbf{D}^2\psi$ ($\mathbf{D}^2\psi = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$ here) is $\frac{3+2\sqrt{2}}{3-2\sqrt{2}} \simeq 34$, making ψ fairly *anisotropic*. In general, this implies a strong

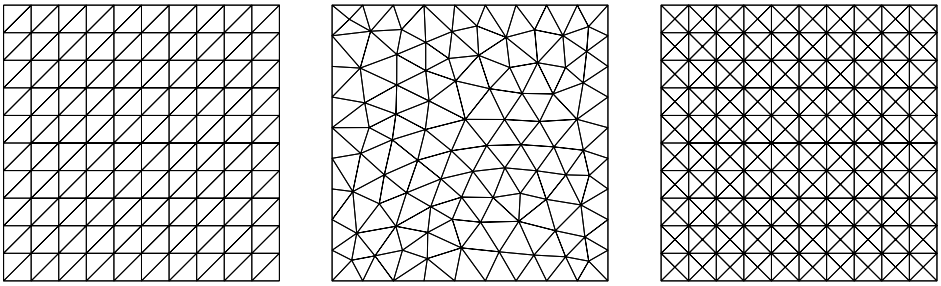

FIGURE 2. Typical triangulations of the unit square $\Omega = (0,1)^2$. Left: structured (asymmetric) mesh; middle: unstructured (isotropic) mesh; right: structured (symmetric) mesh.

mesh dependence of the approximate solution, particularly if one uses the non-smoothed discrete second order derivatives associated with either (6.3) or (6.4).

In Figure 3, we have reported for the three types of meshes shown in Figure 2, the convergence results for the errors $||\psi_h - \psi||_{0h}$ and $||\nabla\psi_h - \nabla\psi||_{0h}$, as functions of the mesh size h, for both the non-regularized (relations (6.3) or (6.4)) and regularized (relations (6.6) and (6.7), with $C = 2$) discrete second order derivatives; both the Newton's method and the $\mathbf{Q}_{\min}$ algorithm have been used to solve the local nonlinear problems (see Sects. 4.2 and 4.3). The lines with slope 1 and 2 in Figure 3 and following denote the lines corresponding to $\mathcal{O}(h)$ and $\mathcal{O}(h^2)$ convergence orders for graphical comparison. These results deserve several comments:

(i) When both algorithms *relaxation/Newton* and *relaxation/*$\mathbf{Q}_{\min}$ converge, they lead essentially to the same solution. However, *relaxation/*$\mathbf{Q}_{\min}$ requires significantly fewer iterations to achieve convergence, and there are situations where it converges while *relaxation/Newton* does not. Typically *relaxation/Newton* requires twice as many iterations as *relaxation/*$\mathbf{Q}_{\min}$ when using the numerical integration (6.3) (between 30 and 200 iterations, *vs.* between 40 and 5000 iterations when the mesh size varies). When using the numerical integration with smoothing (6.4), the number of iterations typically decreases, but the difference remains the same (between 30 and 100 iterations, *vs.* between 25 and 2500 iterations when the mesh size varies). There are several meshes for which convergence can now be achieved also with *relaxation/Newton* when using (6.4) instead of (6.3). Also, $\mathbf{Q}_{\min}$ requires fewer iterations than the Newton algorithm. Moreover, it seems far less sensitive to initialization than Newton's. Actually, the (well-known) sensitivity to initialization of the standard Newton's method has forced us to take $\omega = 0.5$ in some cases, slowing down significantly the convergence of the relaxation method. On the contrary, the greater robustness of $\mathbf{Q}_{\min}$ allowed us to work with $\omega = 1.5$, making the overall algorithm about 20% faster. On the basis of the superior performances of *relaxation/*$\mathbf{Q}_{\min}$, this method has been retained for the solution of the test problems discussed in the following sections.

(ii) To illustrate how the various iterative methods embedded in the relaxation algorithm perform, let us assume that the stopping criterion for the relaxation iterations is the one mentioned above (that is, $|||\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n|||_{0h} < 10^{-4}$); if one takes a 10^{-5} tolerance to stop the conjugate gradient algorithm (9.5)–(9.18) (resp., the Newton's method or the $\mathbf{Q}_{\min}$ algorithm), we observe the following behavior: the Newton's method (resp., $\mathbf{Q}_{\min}$) requires on the average 5–10 (resp., 2–5) iterations to converge, while the number of conjugate gradient iterations varies between 9 and 25 and increases as h decreases (as does the number of relaxation iterations).

(iii) The best convergence results as $h \to 0$, and the fastest convergence of the relaxation method, are obtained by combining the uniform asymmetric triangulations (like the one on the left of Fig. 2) with the non-regularized approximations of the second derivatives (given by relations (6.4)) and $\mathbf{Q}_{\min}$. As expected, in this particular case, the (approximated) L^2-norm of the approximation errors is quite small (of the order of 10^{-7}), since (*cf.* Remark 6.1) for this type of triangulations, the discrete second order derivatives

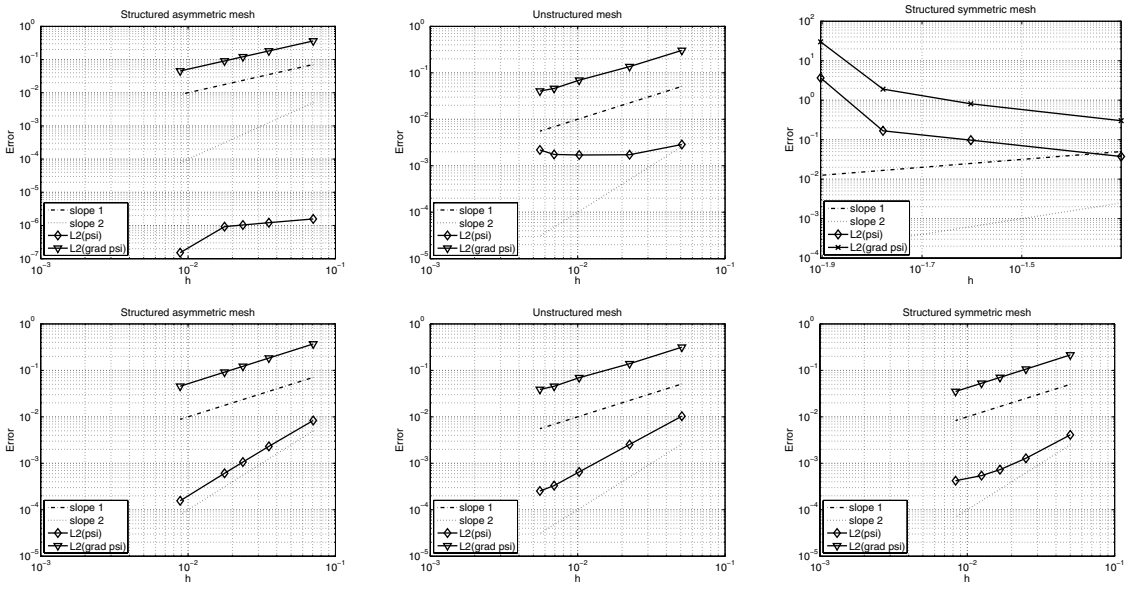

FIGURE 3. First test problem. Convergence (log-log scale) of the errors $||\psi_h - \psi||_{0h}$, $||\nabla(\psi_h - \psi)||_{0h}$; first row: when using non-smoothed approximation of the second derivatives (6.4). Second row: when using smoothed approximation of the second derivatives (6.7). Left: structured asymmetric meshes; middle: unstructured meshes; right: structured symmetric meshes. All results obtained with $\mathbf{Q}_{\min}$.

associated with (6.4) are exact for polynomial functions of degree ≤ 2, as is the convex solution (given by (10.2)) of problem (10.1). Considering the various errors associated with, among others, the solvers involved in our methodology and the mesh generator, we never expected results exact up to machine precision. On the other hand, the uniform asymmetric meshes associated with the non-regularized discrete second order derivatives defined by (6.4) lead to $||\nabla(\psi_h - \psi)||_{0h} = \mathcal{O}(h)$, which is generically optimal when approximating the solution of second-order elliptic equations, using piecewise linear continuous finite element approximations.

(iv) Unlike the uniform asymmetric triangulations, the other types of meshes lead to approximation results ranging from poor (for the unstructured isotropic meshes) to terrible (for the structured symmetric meshes) if one uses the non-regularized discrete second order derivatives defined by (6.4). We observe however that for the unstructured isotropic meshes, although $||\psi_h - \psi||_{0h}$ shows no tendency to converge to 0, we have $||\nabla(\psi_h - \psi)||_{0h} = \mathcal{O}(h)$ for the range of values of h which has been considered. However, there is no contradiction with the Poincaré inequality since, according to [45], we should expect, ultimately, a reduction of the order of convergence for $||\nabla(\psi_h - \psi)||_{0h}$ as $h \to 0$.

(v) For the three types of meshes the regularization of the discrete second order derivatives lead to approximation errors of optimal orders in the range of mesh sizes which has been considered.

10.3. Second test problem

Numerical results for test cases on the unit square $\Omega = (0,1)^2$ introduced, *e.g.*, in [20] are presented. Let us consider the test problem defined by $f(x_1, x_2) = \left(1 + (x_1^2 + x_2^2)\right) e^{(x_1^2+x_2^2)}$, and $g(x_1, x_2) = e^{\frac{1}{2}(x_1^2+x_2^2)}$, whose exact solution is the radial function $\psi(x_1, x_2) = e^{\frac{1}{2}(x_1^2+x_2^2)}$, $(x_1, x_2) \in \Omega$. Figure 4 illustrates the solution ψ_h obtained with various types of triangulations. The method for solving the algebraic problems (4.1) is the $\mathbf{Q}_{\min}$

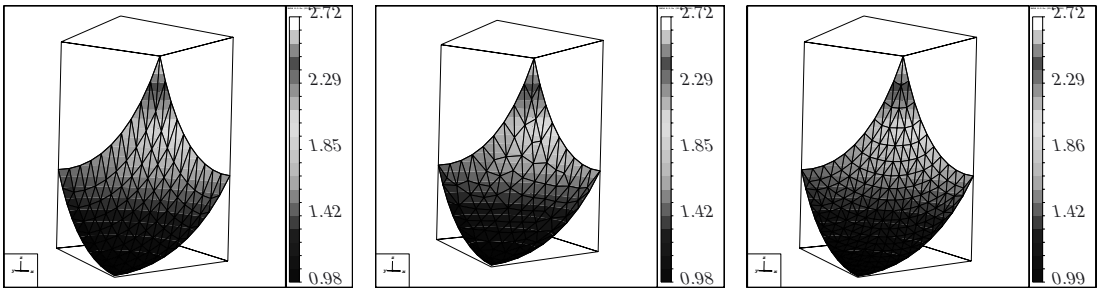

FIGURE 4. Second test problem ($\psi(x_1,x_2)=e^{\frac{1}{2}(x_1^2+x_2^2)}$). Graph of the numerical solution ψ_h. Left: structured asymmetric mesh ($h\simeq 0.0707$); middle: unstructured mesh ($h\simeq 0.0509$); right: structured symmetric mesh ($h=0.05$).

algorithm. The CG algorithm for the solution of the biharmonic problem is stopped when $\delta_k<10^{-5}$, and the tolerance for the $\mathbf{Q}_{\min}$ algorithm is 10^{-5} on successive iterates. The relaxation parameter is $\omega=1.0$.

Remark 10.1. The stopping criterion for the iterative solution method can be any one of the following three: (i) $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n)-\mathbf{p}_h^n\right|\right|\right|_{0h}<10^{-4}$; (ii) $\left|\left|\psi^{n+1}-\psi^n\right|\right|_{0h}<10^{-9}$; or (iii) a maximum of 100 relaxation iterations. Numerical results have shown similar convergence behaviors for all types of stopping criterion, and therefore (i) is used in the whole article (when there is an exact solution). Note that, when using the stopping criterion (ii), numerical results show that the residual $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n)-\mathbf{p}_h^n\right|\right|\right|_{0h}$ varies like h^2 approximately, which agrees with the results in [19, 21].

In Figure 5, we have visualized the order of the approximation errors verified by ψ_h, and its first order derivatives, when they converge to their continuous counterparts, as h converge to zero. One observe that the relaxation algorithm converges faster for the structured asymmetric meshes than for the other types of triangulations. Moreover, the approximations are more accurate since they do not require the use of smoothing techniques. Typically, the CG algorithm converges in 7–10 iterations, while the $\mathbf{Q}_{\min}$ algorithm takes 3–5 iterations. The number of relaxation iterations (with $\omega=1$) is typically less than 20 for all types of triangulation considered and all h considered when using (6.4), and less than 40 when using (6.7). Conclusions are similar to those of the first test problem.

Remark 10.2. The value of the "smoothing parameter" C in (6.7) has been set to $C=2$. However this choice is not critical. Indeed, numerical results results show that the optimal convergence order for the error $||\psi_h-\psi||_{0h}$ is recovered for any value of $C>0$, using all kind of meshes. However, since the accuracy of the computed solutions deteriorates (slowly) as C increases, we have systematically used $C=2$ in the sequel, since it seems to provide a quasi-optimal compromise between regularization (stability) and accuracy.

10.4. Third test problem

Let us consider the test problem, defined, for $R\geq\sqrt{2}$, by $f(x_1,x_2)=\frac{R^2}{\left(R^2-(x_1^2+x_2^2)\right)^2}$, and $g(x_1,x_2)=-\sqrt{R^2-(x_1^2+x_2^2)}$, whose exact solution is the convex function $\psi(x_1,x_2)=-\sqrt{R^2-(x_1^2+x_2^2)}$, $(x_1,x_2)\in\Omega$. When $R>\sqrt{2}$, the exact solution satisfies $\psi\in C^\infty\left(\overline{\Omega}\right)$, while $\psi\in W^{1,p}(\Omega)$, $p\in[1,4)$, when $R=\sqrt{2}$. Therefore it is interesting to see the performance of the algorithm and the quality of the approximation when R tends to $\sqrt{2}$ from above. In order to highlight this effect, we consider two values of R, namely $R=2$ (in that case, ψ is smooth), and $R=\sqrt{2}+0.1$, which is close to the threshold value of $\sqrt{2}$.

Figure 6 shows the graph of ψ_h for $R=2$. Figure 7 illustrates the computational costs and convergence errors for the three types of triangulations. The numerical experiments show consistent second order accuracy of the solution if one smoothes the discrete second order derivatives when employing unstructured isotropic

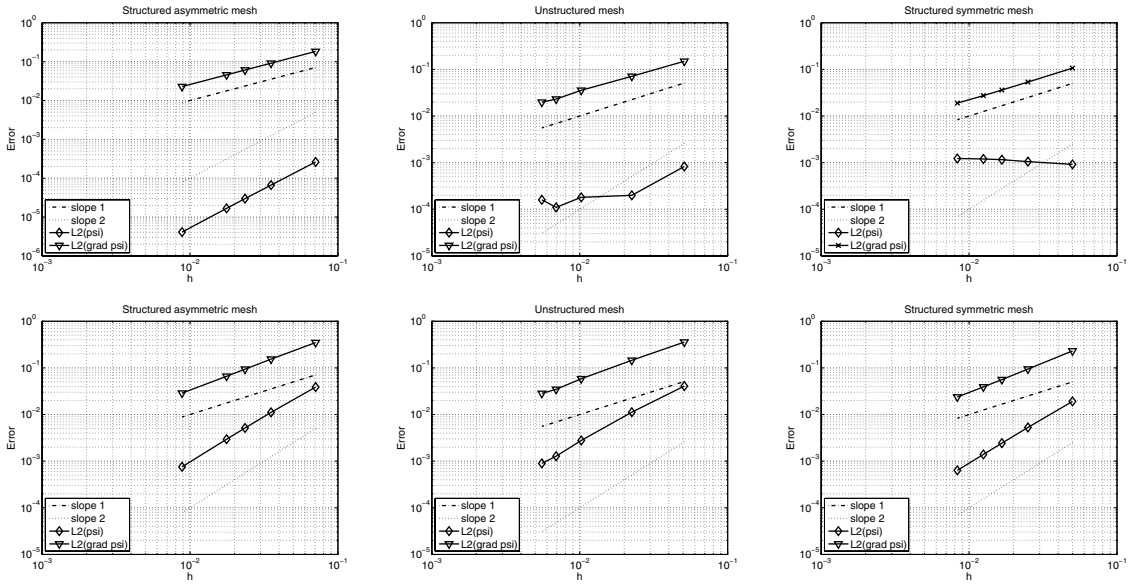

FIGURE 5. Second test problem. Convergence (log-log scale) of the errors $||\psi_h - \psi||_{0h}$, $||\nabla\psi_h - \nabla\psi||_{0h}$; first row: when using non-smoothed approximation of the second derivatives; second row: when using smoothed approximation of the second derivatives. Left: structured asymmetric meshes; middle: unstructured meshes; right: structured symmetric meshes. Stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right|_{0h} < 10^{-4}$.

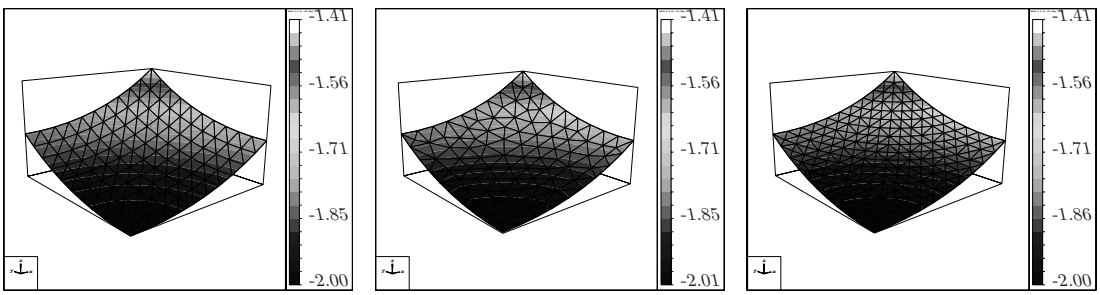

FIGURE 6. Third test problem ($\psi(x_1, x_2) = -\sqrt{R^2 - (x_1^2 + x_2^2)}$, $R = 2$). Graph of the numerical solution ψ_h. Left: structured asymmetric mesh ($h \simeq 0.0707$); middle: unstructured mesh ($h \simeq 0.0509$); right: structured symmetric mesh ($h = 0.05$).

and structured symmetric meshes; they also show that the performances of the method are not altered by the closeness of R to $\sqrt{2}$ (however non-convergence has been observed if $R = \sqrt{2} + 0.01$).

For comparison, Figure 8 shows the graph of ψ_h for $R = \sqrt{2} + 0.1$. Figure 9 illustrates the convergence errors for the three types of triangulations. The numerical experiments still show second order accurate approximation of the solution (with the help of smooth approximations of the second derivatives), showing that the performance of the method is not altered by the lack of regularity of the solution. The number of relaxation iterations increases as R gets closer to $\sqrt{2}$.

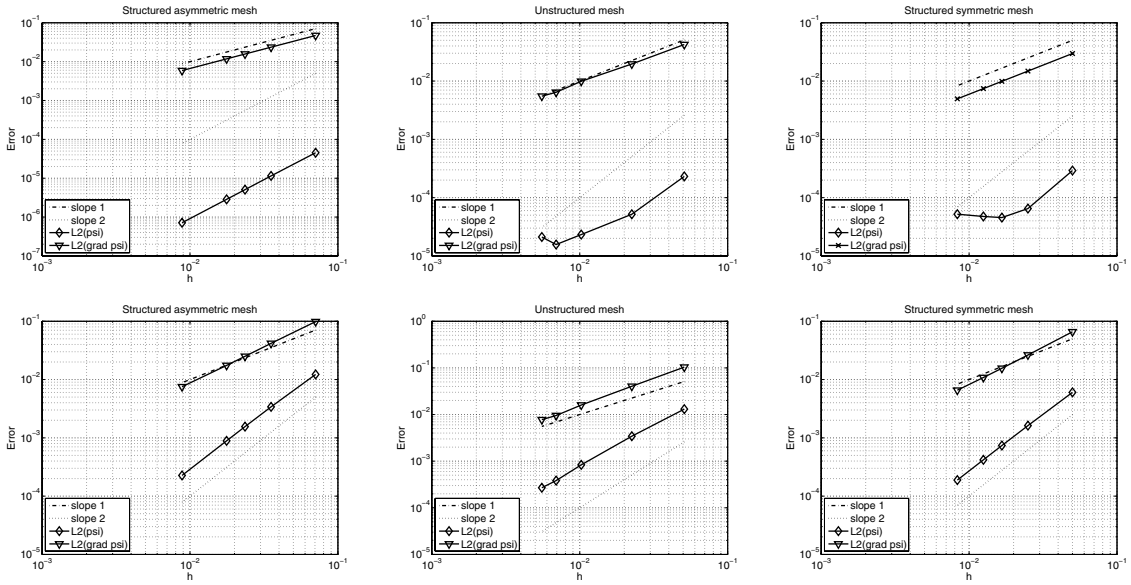

FIGURE 7. Third test problem ($\psi(x_1, x_2) = -\sqrt{R^2 - (x_1^2 + x_2^2)}$, $R = 2$). Convergence (log-log scale) of the errors $||\psi_h - \psi||_{0h}$, $||\nabla\psi_h - \nabla\psi||_{0h}$; first row: when using non-smoothed approximation of the second derivatives; second row: when using smoothed approximation of the second derivatives. Left: structured asymmetric meshes; middle: unstructured meshes; right: structured symmetric meshes. Stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right| < 10^{-4}$.

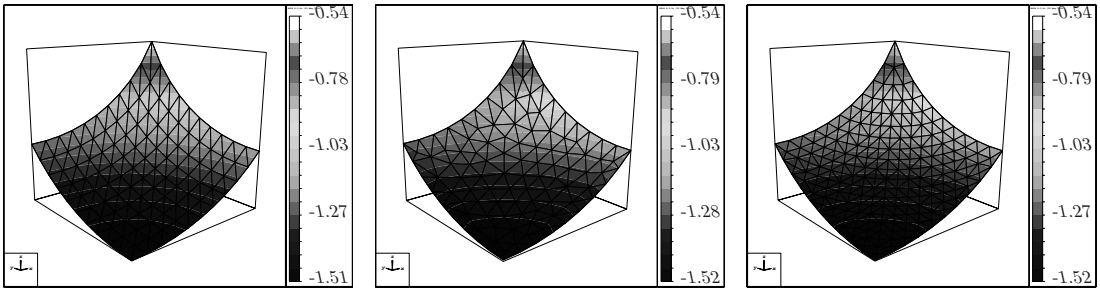

FIGURE 8. Third test problem ($\psi(x_1, x_2) = -\sqrt{R^2 - (x_1^2 + x_2^2)}$, $R = 0.1 + \sqrt{2}$). Graph of the numerical solution ψ_h. Left: structured asymmetric mesh ($h \simeq 0.0707$); middle: unstructured mesh ($h \simeq 0.0509$); right: structured symmetric mesh ($h = 0.05$).

10.5. Fourth test problem

The fourth test problem is defined by $f(x_1, x_2) = \frac{1}{\sqrt{x_1^2+x_2^2}}$, and $g(x_1, x_2) = \frac{\left(2\sqrt{x_1^2+x_2^2}\right)^{3/2}}{3}$; its exact convex solution is given by $\psi(x_1, x_2) = \frac{\left(2\sqrt{x_1^2+x_2^2}\right)^{3/2}}{3}$, $(x_1, x_2) \in \Omega$. Figure 10 illustrates the solution ψ_h. Convergence results are given in Figure 11 for the various types of triangulations. Conclusions are similar as in the previous cases, and the importance of the smoothing procedure for the approximation of the second derivatives is again highlighted. The number of relaxation iterations (with $\omega = 1$) is typically less than 20 for all h considered when using (6.4), and less than 40 when using (6.7).

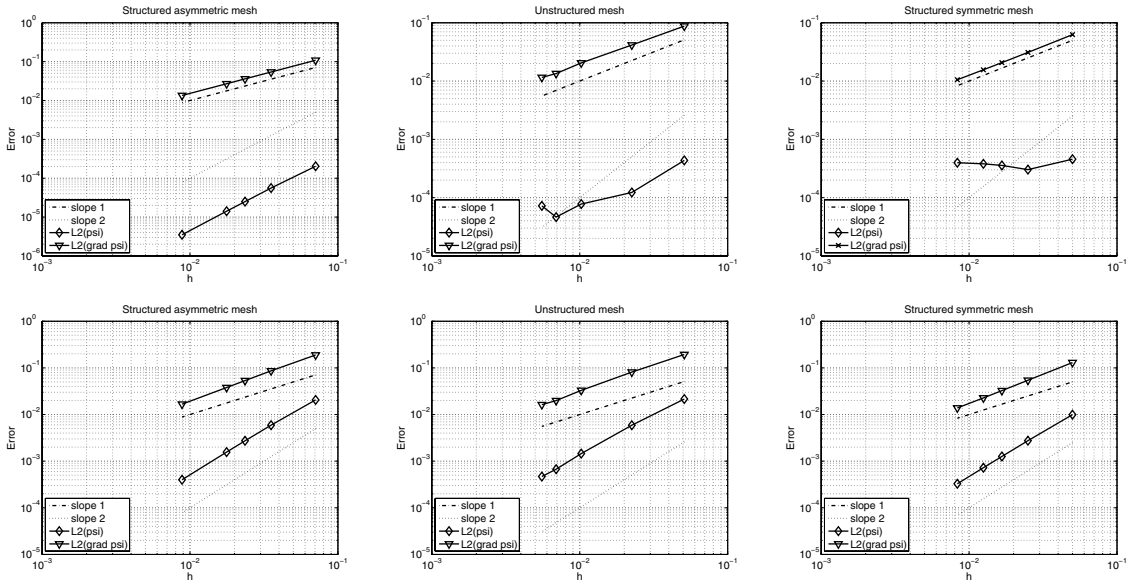

FIGURE 9. Third test problem ($\psi(x_1, x_2) = -\sqrt{R^2 - (x_1^2 + x_2^2)}$, $R = 0.1 + \sqrt{2}$). Convergence (log-log scale) of the errors $||\psi_h - \psi||_{0h}$, $||\nabla\psi_h - \nabla\psi||_{0h}$; first row: when using non-smoothed approximation of the second derivatives. second row: when using smoothed approximation of the second derivatives. Left: structured asymmetric meshes; middle: unstructured meshes; right: structured symmetric meshes. Stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right| < 10^{-4}$.

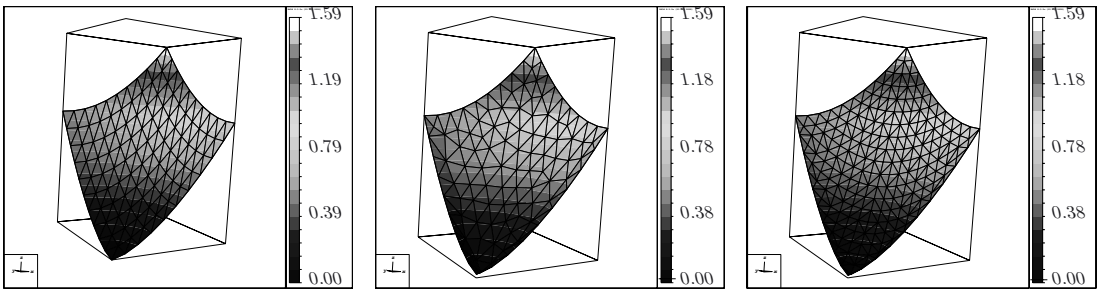

FIGURE 10. Fourth test problem ($\psi(x_1, x_2) = \dfrac{\left(2\sqrt{x_1^2+x_2^2}\right)^{3/2}}{3}$). Graph of the numerical solution ψ_h. Left: structured asymmetric mesh ($h \simeq 0.0707$); middle: unstructured mesh ($h \simeq 0.0509$); right: structured symmetric mesh ($h = 0.05$).

Remark 10.3. The fourth test problem is particularly interesting in the sense that the exact solution $\psi \in H^2(\Omega)$ (in fact $\psi \in W^{2,p}(\Omega)$ for $1 \leq p < 4$) but $\psi \notin C^2\left(\overline{\Omega}\right)$. However, our methodology (which has been constructed to capture solutions with the H^2-regularity) provides optimal order error estimates (without regularization if one uses the uniform asymmetric mesh in Figure 2 (left), and with regularization for the other meshes).

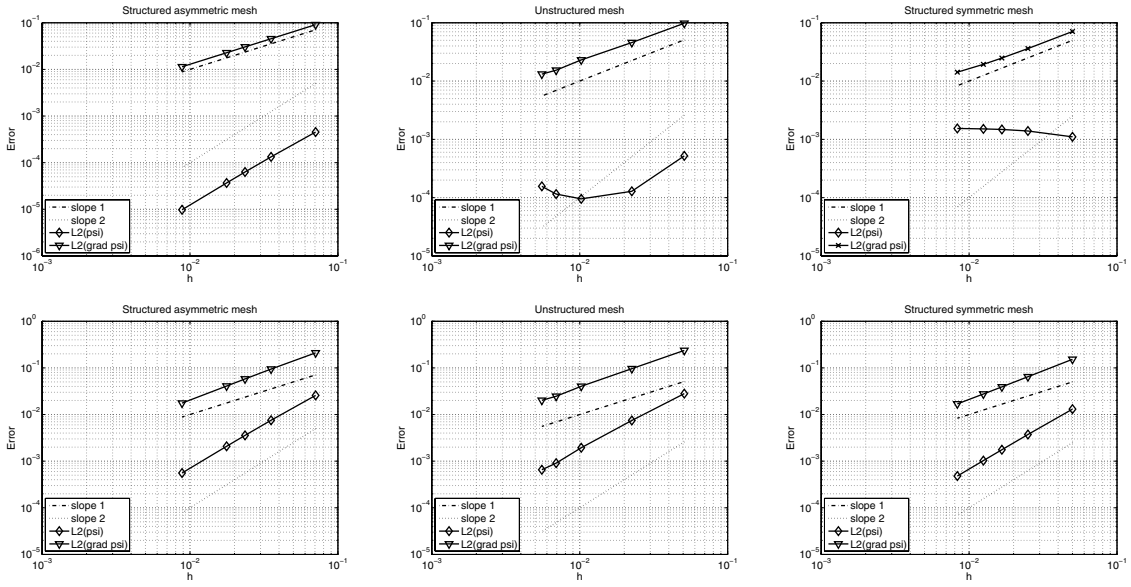

FIGURE 11. Fourth test problem ($\psi(x_1,x_2)=\frac{\left(2\sqrt{x_1^2+x_2^2}\right)^{3/2}}{3}$). Convergence (log-log scale) of the errors $||\psi_h-\psi||_{0h}$, $||\nabla\psi_h-\nabla\psi||_{0h}$; first row: when using non-smoothed approximation of the second derivatives; second row: when using smoothed approximation of the second derivatives. Left: structured asymmetric meshes; middle: unstructured meshes; right: structured symmetric meshes. Stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n)-\mathbf{p}_h^n\right|\right|\right|<10^{-4}$.

10.6. Fifth test problem

The last test problem on the unit square is defined, as in the introduction, by $f(x_1,x_2)=1$, and $g(x_1,x_2)=0$. In that case, the Monge–Ampère equation does not have solutions belonging to $H^2(\Omega)$ (it has however viscosity solutions), despite the smoothness of the data. The problem stems from the non-strict convexity of Ω (see [11, 20, 37] for details). Therefore, the solution obtained can only be compared with computational results from the literature, *e.g.*, in [17, 21, 25, 43]. We use the $\mathbf{Q}_{\min}$ algorithm in the following discussion, smoothed approximations of the second derivatives (6.7), and $\omega=1$. The stopping criterion is $\left|\left|\psi_h^n-\psi_h^{n+1}\right|\right|_{0h}<10^{-7}$.

Figure 12 illustrates the solution of the Monge–Ampère equation obtained with all types of triangulations. Figure 13 illustrates the determinant of its computed Hessian. Figure 14 shows a cut of the solution (corresponding respectively to the solutions in Fig. 12) for $x_2=1/2$ and $x_1=x_2$ respectively for several mesh sizes. The solution, in particular the solution magnitude, appropriately matches the solution presented in [19–21, 25, 32]. Table 1 shows the values of the residual and the number of iterations for various values of the mesh size h and types of triangulations. A close inspection of the numerical results shows that the curvature of the graph of ψ_h is slightly negative close to the corners of Ω, implying that the Monge–Ampère equation is violated here (indeed the curvature is given by $\det\mathbf{D}^2\psi/(1+|\nabla\psi|^2)^2$). The equation is also violated along the boundary (as emphasized in [32], p. 176); however, the Monge–Ampère equation $\det\mathbf{D}^2\psi=1$ is verified with a very high precision sufficiently far away from Γ. For more information on the solutions of $\det\mathbf{D}^2\psi=1$, see [37], Chapter 4 and references therein.

Remark 10.4. An inspection of Figure 14 shows that the minimum values of the discrete solutions converge to a limit very close to -0.18, for the three types of meshes which have been considered. There is a remarkable agreement between these results and the ones reported in [21, 25]. The agreement with [21] is not surprising since the numerical results presented there, and in this paragraph, were obtained using the same least-squares

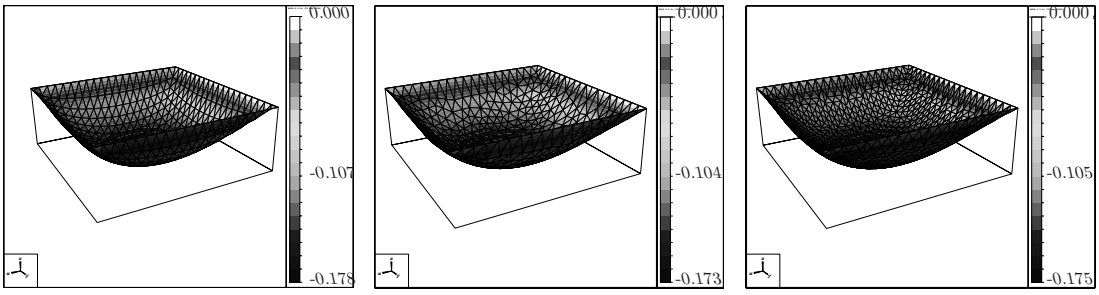

FIGURE 12. Fifth test problem ($f = 1$, $g = 0$). Graph of the numerical solution ψ_h. Left: structured asymmetric mesh ($h \simeq 0.0353$, after 140 iterations). Middle: unstructured mesh ($h \simeq 0.0225$, after 105 iterations). Right: structured symmetric mesh ($h = 0.025$, after 146 iterations).

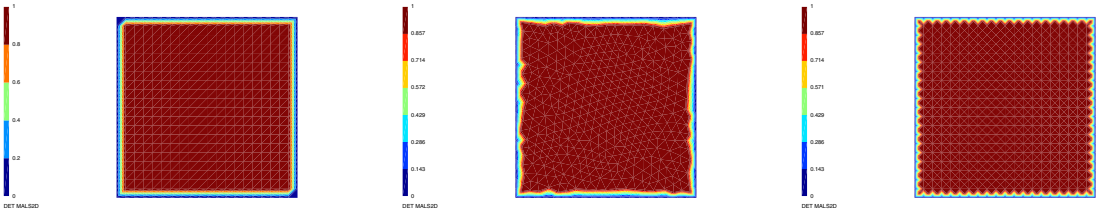

FIGURE 13. Fifth test problem ($f = 1$, $g = 0$). Contours of the determinant of the discrete Hessian $\mathbf{D}_h^2(\psi_)h$. Left: structured asymmetric mesh ($h \simeq 0.0353$, after 140 iterations). Middle: unstructured mesh ($h \simeq 0.0225$, after 105 iterations). Right: structured symmetric mesh ($h = 0.025$, after 146 iterations).

TABLE 1. Fifth test problem ($f = 1$, $g = 0$). Convergence results and computational costs when the stopping criterion is $\left|\left|\psi_h^n - \psi_h^{n+1}\right|\right|_{0h} < 10^{-7}$ ($\omega = 1$).

Structured asymmetric mesh			Unstructured isotropic mesh			Structured symmetric mesh		
h	$\left\|\left\|\left\|\mathbf{D}^2(\psi_h^n) - \mathbf{p}_h^n\right\|\right\|\right\|_{0h}$	# iter.	h	$\left\|\left\|\left\|\mathbf{D}^2(\psi_h^n) - \mathbf{p}_h^n\right\|\right\|\right\|_{0h}$	# iter.	h	$\left\|\left\|\left\|\mathbf{D}^2(\psi_h^n) - \mathbf{p}_h^n\right\|\right\|\right\|_{0h}$	# iter.
0.07071	0.10003E-04	43	0.05091	0.52768E-05	48	0.05000	0.10559E-04	48
0.03535	0.78018E-04	140	0.02249	0.42235E-04	105	0.02500	0.64290E-04	146
0.02357	0.19351E-03	337	0.01023	0.22740E-03	203	0.01666	0.15442E-03	261
0.01767	0.29037E-03	763	0.00692	0.47883E-03	259	0.01250	0.28545E-03	377
0.00883	0.80873E-03	2000	0.00554	0.67947E-03	268	0.00833	0.11296E-02	884

formulation and mixed finite element approximation (the difference being in the algorithms used to solve the approximate problems). The agreement with [25] is more "interesting" and significant. The numerical results reported in [25] have been obtained using a mixed finite element implementation of the so-called vanishing moment method, relying on spaces of continuous piecewise quadratic functions and on a biharmonic regularization of the Monge−Ampère equation (see [25] for details). Actually, it has been shown in [24, 25] that when the regularization parameter converges to zero, the solutions produced by the vanishing moment method converge to a viscosity solution. This good matching between the solutions in [25] and the ones in this article, concerning the fifth test problem (a problem without classical solutions) suggests that our least-squares approach has, possibly, some viscosity method features.

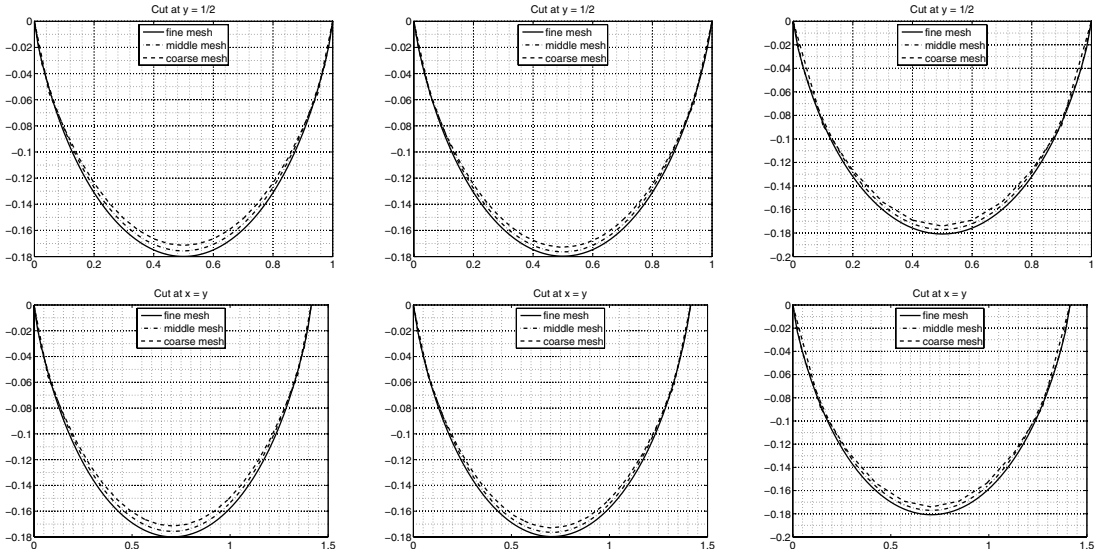

FIGURE 14. Fifth test problem ($f = 1$, $g = 0$). Cut of the graph of the numerical solution ψ_h along the lines $x_2 = 1/2$ (top row) and $x_1 = x_2$ (bottom row). Left: structured asymmetric mesh ($h \simeq 0.0353, 0.0176, 0.0088$). Middle: unstructured mesh ($h \simeq 0.0225, 0.0102, 0.0055$). Right: structured symmetric mesh ($h = 0.05, 0.0166, 0.0083$).

10.7. Summary and comments

Numerical experiments on the unit square have illustrated the properties of the proposed method. The $\mathbf{Q}_{\min}$ algorithm [48] proves to be much more efficient than a more classical Newton method. Not surprisingly, low order finite elements are not accurate for the approximation of the second derivatives and therefore require an additional regularization *à la Tychonoff*. With this regularization procedure, optimal convergence orders are obtained on **all** types of triangulations considered. As the method without regularization fails on some types of triangulations (due to a poor approximation of the second order derivatives), this shows a remarkable robustness improvement with the method proposed in this work. Furthermore, the method has been able to capture (in a least-squares sense) solutions when no classical solution exists. Based on these comments, we will present in the following numerical experiments for domains different from the unit square, typically with curved boundaries, to further emphasize the efficiency of the proposed methodology.

10.8. A test problem on the unit disk

The results presented in the above sections have shown that, if coupled with an appropriate regularization procedure, the piecewise linear approximation methods discussed in [21], have the ability to compute accurate approximate solutions for arbitrary types of triangulations, including pathological ones. These results agree with the results obtained in the literature, as in, *e.g.*, [20, 21, 25]. In this section, we show that the proposed method based on mixed finite elements applies also to domains with *curved boundaries*. Note that, in the case of curved boundaries, the use of mixed piecewise linear finite elements provides a substantial simplification compared to using high order finite elements (as in [24, 42] for instance), or finite differences [26, 27, 43].

We consider the unit disk $\mathcal{S}_1$, with isotropic triangulations built with GMSH [28]. Figure 15 visualizes the solution for $f = 1$ and $g = 0$ on $\mathcal{S}_1$. The exact convex solution is $\psi(x_1, x_2) = 1/2\left[(x_1^2 + x_2^2) - 1\right]$, which is clearly in $C^\infty\left(\overline{\mathcal{S}_1}\right)$. Figure 16 illustrates the convergence results when using the smoothed approximation of the derivatives (6.7); $||\psi_h - \psi||_{0h}$ exhibits second order accuracy.

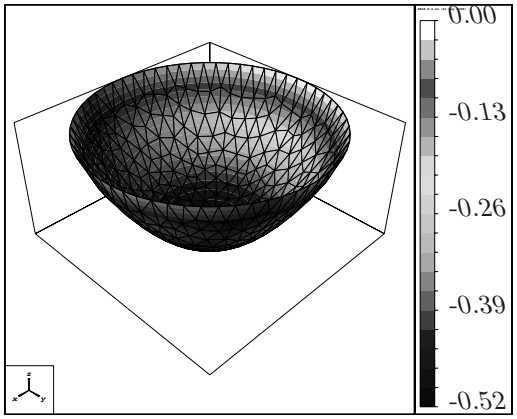

FIGURE 15. Test problem on the unit disk $\mathcal{S}_1$. Graph of the numerical solution ψ_h (for $f = 1$ and $g = 0$) on the unit disk $\mathcal{S}_1$ ($h \simeq 0.04392$, 19 relaxation iterations with $\omega = 1$).

h	$\|\|\psi_h - \psi\|\|_{0h}$	$\|\|\nabla(\psi_h - \psi)\|\|_{0h}$	# iter.
0.04392	0.35182E-01	0.21176E+00	19
0.02788	0.15247E-01	0.12036E+00	23
0.02083	0.89428E-02	0.84006E-01	19
0.01508	0.58031E-02	0.63335E-01	19
0.01349	0.39951E-02	0.49891E-01	18
0.01028	0.22307E-02	0.35496E-01	21

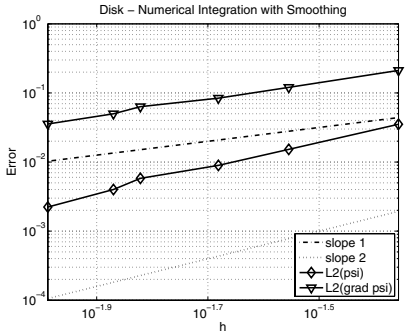

FIGURE 16. Test problem on the unit disk $\mathcal{S}_1$. Left and middle: convergence of the approximation errors $||\psi_h - \psi||_{0h}$ and $||\nabla\psi_h - \nabla\psi||_{0h}$ on $\mathcal{S}_1$. Right column: number of relaxation iterations necessary to achieve convergence (stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right|_{0h} < 10^{-4}$, $\omega = 1$). The algebraic solver is $\mathbf{Q}_{\min}$ and the second order derivatives are approximated using the smoothing procedure (6.6). Left: numerical values; right: plot of the errors (log-log scale).

10.9. A test problem on an ellipse

We consider the elliptical domain $\mathcal{E}_{a,b} = \left\{(x_1, x_2) \in \mathbb{R}^2, x_1^2/a^2 + x_2^2/b^2 < 1\right\}$, with corresponding isotropic triangulations built with GMSH [28]. In particular, let us work with the elliptical domain $\mathcal{E}_{1,2}$, and $f = 1/4$, and $g = 0$. In this case, the exact solution to (2.1) is given by $\psi(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2/4 - 1)$. Figure 15 visualizes the solution ψ_h, while Figure 18 illustrates the convergence results when using the smoothed approximation of the derivatives (6.7); $||\psi_h - \psi||_{0h}$ exhibits again second order accuracy.

Remark 10.5. Note that, for both the unit disk and the elliptical domain, convergence properties are lost when using non-smoothed approximations of the second derivatives (6.4).

10.10. A test problem on the half-disk

Finally let us consider now the half-disk domain $\mathcal{S}_{1,-} := \mathcal{S}_1 \cap \{y < 0\}$, and return to the example presented in Section 10.2, namely $f(x_1, x_2) = 1$ and $g(x_1, x_2) = \frac{5}{2}x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2$. Figure 19 visualizes the contours of the

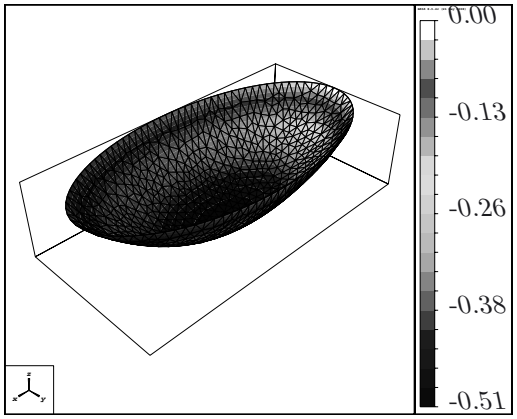

FIGURE 17. Test problem on the elliptical domain $\mathcal{E}_{1,2}$. Graph of the numerical solution ψ_h (for $f = 1/4$ and $g = 0$) on $\mathcal{E}_{1,2}$ ($h \simeq 0.04249$, 72 relaxation iterations with $\omega = 1$).

h	$\|\|\psi_h - \psi\|\|_{0h}$	$\|\|\nabla(\psi_h - \psi)\|\|_{0h}$	# iter.
0.04249	0.31593E-01	0.18896E+00	72
0.01986	0.80691E-02	0.79596E-01	66
0.01633	0.52340E-02	0.61007E-01	74
0.01377	0.36987E-02	0.48918E-01	68

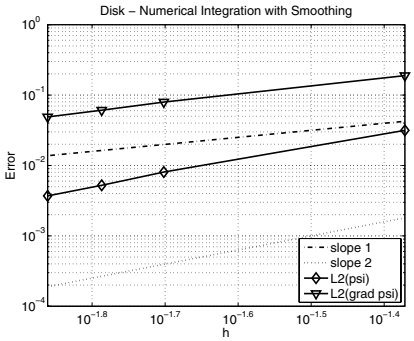

FIGURE 18. Test problem on the elliptical domain $\mathcal{E}_{1,2}$. Left and middle columns: convergence of the approximation errors $||\psi_h - \psi||_{0h}$ and $||\nabla\psi_h - \nabla\psi||_{0h}$. Right column: number of relaxation iterations necessary to achieve convergence (stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right|_{0h} < 10^{-4}$, $\omega = 1$). The algebraic solver is the $\mathbf{Q}_{\min}$ algorithm and the second derivatives are approximated using the smoothing procedure (6.7). Left: numerical values; right: plot of the errors (log-log scale).

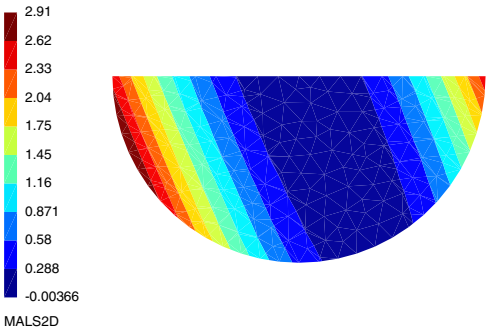

FIGURE 19. Test problem on the half-disk. Contours of the numerical solution ψ_h on $\mathcal{S}_{1,-}$ (for $f(x_1, x_2) = 1$ and $g(x_1, x_2) = \frac{5}{2}x_1^2 + 2x_1x_2 + \frac{1}{2}x_2^2$) ($h \simeq 0.04519$, 140 relaxation iterations with $\omega = 1$).

h	$\|\|\psi_h - \psi\|\|_{0h}$	$\|\|\nabla(\psi_h - \psi)\|\|_{0h}$	# iter.
0.04519	0.95925E-02	0.37198E+00	140
0.02226	0.26230E-02	0.17649E+00	224
0.01009	0.72159E-03	0.84842E-01	281
0.00674	0.34671E-03	0.56310E-01	304

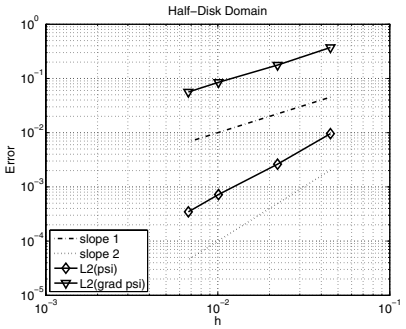

FIGURE 20. Test problem on the half-disk. Left and middle columns: convergence of the approximation errors $||\psi_h - \psi||_{0h}$ and $||\nabla\psi_h - \nabla\psi||_{0h}$. Right column: number of relaxation iterations necessary to achieve convergence (stopping criterion $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right|_{0h} < 10^{-4}$, $\omega = 1$). The algebraic solver is the $\mathbf{Q}_{\min}$ algorithm and the second derivatives are approximated using the smoothing procedure (6.7). Left: numerical values; right: plot of the errors (log-log scale).

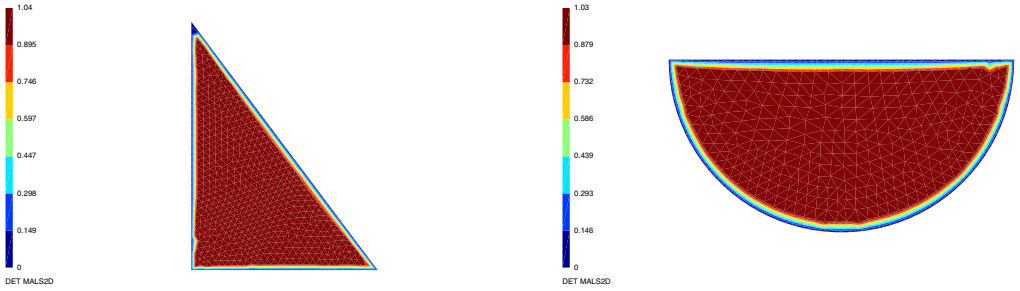

FIGURE 21. Non-smooth example on non strictly convex domains. Contours of the determinant of the discrete Hessian $\mathbf{D}_h^2(\psi_h)$ ($f = 1$ and $g = 0$). Left: triangular domain Ω_T ($h \simeq 0.0457$, 500 relaxation iterations with $\omega = 1$). Right: half disk $\mathcal{S}_{1,-}$ ($h \simeq 0.0284$, 500 relaxation iterations with $\omega = 1$).

solution on $\mathcal{S}_{1,-}$, while Figure 20 illustrates the convergence results when using the smoothed approximation of the derivatives (6.7); $||\psi_h - \psi||_{0h}$ exhibits the expected second order accuracy. The non-strict convexity of the domain increases significantly the number of relaxation iterations, compared to the two previous test cases. Note that, when using the numerical integration method described in (6.4), convergence is not guaranteed. On the other hand, if one uses the smooth variant (6.4) (6.7), the number of iterations decreases.

10.11. Further numerical results

Finally let us focus on some non-smooth cases with $f = 1$ and $g = 0$, and consider the triangular domain Ω_T defined by $\Omega_T = \left\{(x_1, x_2) \in \mathbb{R}^2 \,:\, x_1, x_2 > 0,\ 4x_1 + 3x_2 < 12\right\}$, and the half-disk $\mathcal{S}_{1,-}$.

Figure 21 visualizes the approximation of the determinant of the Hessian $\mathbf{D}_h^2(\psi_h^n)$ for these situations, and shows a loss of convexity of the solution in the neighborhood of the corners (and of the parts of the boundary that are not strictly convex) that is similar to the effects observed on the unit square.

10.12. Non-smooth test problems involving the Dirac measure

To conclude these numerical experiments, we are going to consider two non-smooth cases which are in principle beyond the scope of our approach since for both cases, the solution of the associated problem (2.1) is the function

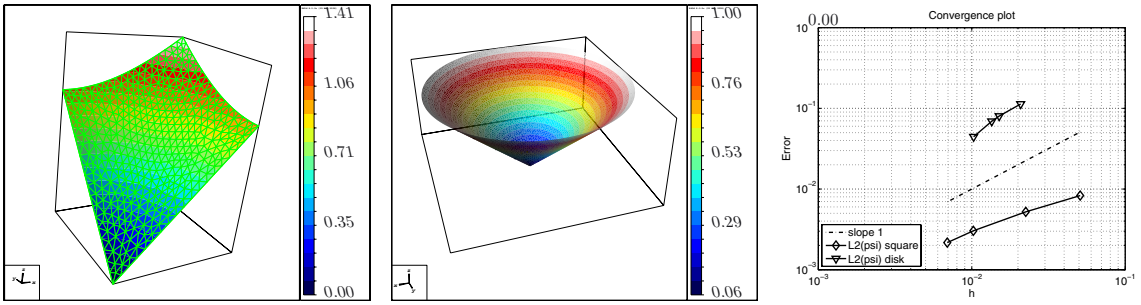

FIGURE 22. Non-smooth test problem with regularization. Left: graph of the numerical solution ψ_h on the unit square ($h \simeq 0.0509, \varepsilon = 10^{-6}$); middle: graph of the numerical solution ψ_h on the unit disk ($h \simeq 0.0103, \varepsilon = 10^{-3}$); right: convergence of ψ_h toward the exact solution $\psi(x_1, x_2) = \sqrt{x_1^2 + x_2^2}$ when the mesh size h tends to zero on the unit square and the unit disk.

ψ defined by $\psi(\mathbf{x}) = |\mathbf{x}|\,(= \sqrt{x_1^2 + x_2^2})$, a (convex) function which does not have the $H^2(\Omega)$-regularity if $(0,0)$ belongs to Ω, and which verifies $M\psi = \pi\delta_{(0,0)}$, M being the Monge–Ampère measure (see, *e.g.*, [37], Chap. 1 for the definition and properties of M) and $\delta_{(0,0)}$ the Dirac measure at $(0,0)$. The first test problem that we consider is: find $\psi : \Omega \to \mathbb{R}$ such that

$$\det \mathbf{D}^2\psi(x_1, x_2) = 0 \quad \text{in} \quad \Omega, \qquad \psi(x_1, x_2) = \sqrt{x_1^2 + x_2^2} \quad \text{on} \quad \partial\Omega, \tag{10.3}$$

with $\Omega = (0,1)^2$. The unique convex solution of the Monge–Ampère problem (10.3) is the function $\mathbf{x} \to |\mathbf{x}|$. Our methodology being suited for strictly positive right-hand sides f only, we approximate problem (10.3) by

$$\det \mathbf{D}^2\psi(x_1, x_2) = \varepsilon \quad \text{in} \quad \Omega, \qquad \psi(x_1, x_2) = \sqrt{x_1^2 + x_2^2} \quad \text{on} \quad \partial\Omega, \tag{10.4}$$

where $\varepsilon > 0$ a (small) positive number. The boundary data does not have the $H^{3/2}(\Gamma)$-regularity (as required -in principle- by our least-squares approach), but only the $H^s(\Gamma)$-regularity with $s < 3/2$. However, the discrete version of our least-squares/relaxation methodology handles problem (10.4) fairly easily. Figure 22 (left) illustrates the solution ψ_h obtained with an unstructured isotropic mesh. The method for solving the algebraic problems (4.1) is the $\mathbf{Q}_{\min}$ algorithm. The CG algorithm for the solution of the biharmonic problem is stopped when $\delta_k < 10^{-5}$, and the tolerance for the $\mathbf{Q}_{\min}$ algorithm is 10^{-5} on successive iterates. The relaxation parameter is $\omega = 1.0$. The stopping criterion is $\left|\left|\left|\mathbf{D}_h^2(\psi_h^n) - \mathbf{p}_h^n\right|\right|\right|_{0h} < 10^{-4}$. We observed that the solution obtained by our least-squares/relaxation method is essentially independent of the value of ε, for ε in the range 10^{-1}–10^{-9}.

Next, we consider a related test problem on the unit open disk $\Omega = \mathcal{S}_1$, namely: find $\psi : \Omega \to \mathbb{R}$ such that

$$\det \mathbf{D}^2\psi(x_1, x_2) = \pi\delta_{(0,0)} \quad \text{in} \quad \Omega, \qquad \psi(x_1, x_2) = 1 \quad \text{on} \quad \partial\Omega, \tag{10.5}$$

The unique convex (generalized) solution of problem (10.5) is, again, the function ψ defined by $\psi(\mathbf{x}) = |\mathbf{x}|$. From a numerical point of view, the Dirac measure in (10.5) has to be regularized to make it compatible with our methodology. Thus, we consider the following regularized approximation of problem (10.5): find $\psi : \Omega \to \mathbb{R}$ such that

$$\det \mathbf{D}^2\psi(x_1, x_2) = \frac{\varepsilon^2}{(\varepsilon^2 + x_1^2 + x_2^2)^2} \quad \text{in} \quad \Omega, \qquad \psi(x_1, x_2) = 1 \quad \text{on} \quad \partial\Omega, \tag{10.6}$$

with $\varepsilon > 0$. This type of regularization is reminiscent of the one used for instance in [9].

Remark 10.6. The rationale behind the regularized problem (10.6) is the fact that an explicit calculation shows that the relation $-\Delta \ln(1/\sqrt{x_1^2+x_2^2}) = 2\pi\delta_{(\mathbf{0})}$ (in the sense of distributions) can be approximated by $-\Delta \ln(1/\sqrt{\varepsilon^2+x_1^2+x_2^2}) = \frac{2\varepsilon^2}{(\varepsilon^2+x_1^2+x_2^2)^2}$. The fact that the regularized right-hand side in (10.6) is strictly positive facilitates also the applicability of our methodology.

Figure 22 (middle) illustrates the solution ψ_h obtained with an unstructured isotropic mesh, showing the ability of the least-squares approach in handling the regularized version of the Monge−Ampère equation. Figure 22 (right) illustrates the convergence of the approximation ψ_h to the exact solution ψ, measured as the $L^2(\Omega)$-norm of the difference $\psi_h - \psi$. Good convergence properties are obtained by taking $\varepsilon = 10^{-6}$ in (10.4) and $\varepsilon = h$ in (10.6), h being the mesh size.

11. Conclusions and perspectives

A numerical method for the approximation of the Dirichlet problem for the real elliptic Monge−Ampère equation for arbitrary domains in two dimensions has been presented.

This least-squares method allows to obtain approximations of the convex solution that satisfy exactly the boundary condition, while satisfying the equation in a weak sense. The relaxation algorithm allows to decouple the differential operators from the nonlinearities. It includes a novel (with respect to [21]) algorithm for the solution of local nonlinear eigenvalue problems.

Mixed piecewise linear finite elements, together with a Tychonoff regularization, allow to find approximations of the solution of the Monge−Ampère equation in arbitrary domains and with arbitrary types of triangulations. In particular, our methodology can handle quite easily and accurately (convex) domains with curved boundaries.

Perspectives include the extension of this methodology to other fully nonlinear elliptic equations in two and three dimensions of space in arbitrary domains [7, 26].

Acknowledgements. The authors acknowledge the partial support of the National Science Foundation Grants NSF DMS-0412267 and NSF DMS-0913982. The authors thank Prof. L. Caffarelli (Univ. of Texas at Austin), Prof. E. Dean (Univ. of Houston), Prof. X. Feng (Univ. of Tennessee at Knoxville), Prof. M. Picasso (EPFL), and the two anonymous reviewers for helpful comments and discussions. The first author gratefully acknowledges the partial support of the company Ycoor Systems SA, Switzerland, and the Chair and Analysis and Numerical Simulation, Ecole Polytechnique Fédérale de Lausanne, Switzerland.

References

[1] A.D. Aleksandrov, Uniqueness conditions and estimates for the solution of the Dirichlet problem. *Amer. Math. Soc. Trans.* **68** (1968) 89–119.

[2] J.D. Benamou, B.D. Froese and A.M. Oberman, Two numerical methods for the elliptic Monge−Ampère equation. *ESAIM: M2AN* **44** (2010) 737–758.

[3] M. Bernadou, P.L. George, A. Hassim, P. Joly, P. Laug, A. Perronet, E. Saltel, D. Steer, G. Vanderborck and M. Vidrascu, Modulef, a modular library of finite elements. Technical report, INRIA (1988).

[4] P.B. Bochev and M.D. Gunzburger, *Least-Squares Finite Element Methods*. Springer-Verlag, New York (2009).

[5] K. Boehmer, On finite element methods for fully nonlinear elliptic equations of second order. *SIAM J. Numer. Anal.* **46** (2008) 1212–1249.

[6] S.C. Brenner, T. Gudi, M. Neilan and L.-Y. Sung, c^0 penalty methods for the fully nonlinear Monge−Ampere equation. *Math. Comput.* **80** (2011) 1979–1995.

[7] S.C. Brenner and M. Neilan, Finite element approximations of the three dimensional Monge−Ampère equation. *ESAIM: M2AN* **46** (2012) 979–1001.

[8] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, NewYork (1991).

[9] A. Caboussat and R. Glowinski, Regularization methods for the numerical solution of the divergence equation $\nabla \cdot \mathbf{u} = f$. *J. Comput. Math.* **30** (2012) 354–380.

[10] X. Cabré, Topics in regularity and qualitative properties of solutions of nonlinear elliptic equations. *Discrete Contin. Dyn. Systems* **8** (2002) 289–302.

[11] L.A. Caffarelli, Nonlinear elliptic theory and the Monge−Ampère equation, in *Proc. of the International Congress of Mathematicians*. Higher Education Press, Beijing (2002) 179–187.

[12] L.A. Caffarelli and X. Cabré, *Fully Nonlinear Elliptic Equations*. American Mathematical Society, Providence, RI (1995).

[13] L.A. Caffarelli and R. Glowinski, Numerical solution of the Dirichlet problem for a Pucci equation in dimension two. Application to homogenization. *J. Numer. Math.* **16** (2008) 185–216.

[14] L.A. Caffarelli, S.A. Kochenkgin and V.I. Olicker, On the numerical solution of reflector design with given far field scattering data, in *Monge−Ampère Equation: Application to Geometry and Optimization*, American Mathematical Society, Providence, RI (1999) 13–32.

[15] M.G. Crandall, H. Ishii and P.-L. Lions, User's guide to viscosity solutions of second order partial differential equations. *Bull. Am. Math. Soc.* **27** (1992) 1–67.

[16] E.J. Dean and R. Glowinski, Numerical solution of the two-dimensional elliptic Monge−Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach. *C. R. Acad. Sci. Paris, Ser. I* **336** (2003) 779–784.

[17] E.J. Dean and R. Glowinski, Numerical solution of the two-dimensional elliptic Monge−Ampère equation with Dirichlet boundary conditions: a least-squares approach. *C. R. Acad. Sci. Paris, Ser. I* **339** (2004) 887–892.

[18] E.J. Dean and R. Glowinski, Numerical solution of a two-dimensional elliptic Pucci's equation with Dirichlet boundary conditions: a least-squares approach. *C. R. Acad. Sci. Paris, Ser. I* **341** (2005) 374–380.

[19] E.J. Dean and R. Glowinski, An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge−Ampère equation in two dimensions. *Electronic Transactions in Numerical Analysis* **22** (2006) 71–96.

[20] E.J. Dean and R. Glowinski, Numerical methods for fully nonlinear elliptic equations of the Monge−Ampère type. *Comput. Meth. Appl. Mech. Engrg.* **195** (2006) 1344–1386.

[21] E.J. Dean and R. Glowinski, On the numerical solution of the elliptic Monge−Ampère equation in dimension two: A least-squares approach, in *Partial Differential Equations: Modeling and Numerical Simulation*, vol. 16 of *Comput. Methods Appl. Sci.,* edited by R. Glowinski and P. Neittaanmäki. Springer (2008) 43–63.

[22] E.J. Dean, R. Glowinski and T.W. Pan, Operator-splitting methods and applications to the direct numerical simulation of particulate flow and to the solution of the elliptic Monge−Ampère equation. in *Control and Boundary Analysis*, edited by J.P. Zolésio J. Cagnol, CRC Boca Raton, FLA (2005) 1–27.

[23] E.J. Dean, R. Glowinski and D. Trevas, An approximate factorization/least squares solution method for a mixed finite element approximation of the Cahn-Hilliard equation. *Jpn J. Ind. Appl. Math.* **13** (1996) 495–517.

[24] X. Feng and M. Neilan, Mixed finite element methods for the fully nonlinear Monge−Ampère equation based on the vanishing moment method. *SIAM J. Numer. Anal.* **47** (2009) 1226–1250.

[25] X. Feng and M. Neilan, Vanishing moment method and moment solutions of second order fully nonlinear partial differential equations. *J. Sci. Comput.* **38** (2009) 74–98.

[26] B.D. Froese and A.M. Oberman, Convergent finite difference solvers for viscosity solutions of the elliptic Monge−Ampère equation in dimensions two and higher. *SIAM J. Numer. Anal.* **49** (2011) 1692–1715.

[27] B.D. Froese and A.M. Oberman, Fast finite difference solvers for singular solutions of the elliptic Monge−Ampère equation. *J. Comput. Phys.* **230** (2011) 818–834.

[28] C. Geuzaine and J.-F. Remacle, Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *Int. J. Numer. Meth. Eng.* **79** (2009) 1309–1331.

[29] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin (2001).

[30] R. Glowinski, *Finite Element Methods For Incompressible Viscous Flow*, *Handbook of Numerical Analysis*, edited by P.G. Ciarlet, J.L. Lions. Elsevier, Amsterdam **IX** (2003) 3–1176.

[31] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*. 2nd edition, Springer-Verlag, New York, NY (2008).

[32] R. Glowinski, Numerical methods for fully nonlinear elliptic equations. in *Invited Lectures, 6th Int. Congress on Industrial and Applied Mathematics, Zürich, Switzerland, 16-20 July 2007*. EMS (2009) 155–192.

[33] R. Glowinski, E.J. Dean, G. Guidoboni, H.L. Juarez and T.W. Pan, Applications of operator-splitting methods to the direct numerical simulation of particulate and free surface flows and to the numerical solution of the two-dimensional Monge−Ampère equation. *Jpn J. Ind. Appl. Math.* **25** (2008) 1–63.

[34] R. Glowinski, J.-L. Lions and J.W. He, *Exact and Approximate Controllability for Distributed Parameter Systems: A Numerical Approach*. Encyclopedia of Mathematics and its Applications. Cambridge University Press (2008).

[35] R. Glowinski, D. Marini and M. Vidrascu, Finite-element approximations and iterative solutions of a fourth-order elliptic variational inequality. *IMA J. Numer. Anal.* **4** (1984) 127–167.

[36] R. Glowinski and O. Pironneau, Numerical methods for the first bi-harmonic equation and for the two-dimensional Stokes problem. *SIAM Rev.* **17** (1979) 167–212.

[37] C.E. Gutiérrez, *The Monge−Ampère Equation*. Birkhaüser, Boston (2001).

[38] T.J.R. Hughes, L. Franca and M. Balestra, A new finite element formulation for computational fluid dynamics: V. circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolation. *Comput. Methods Appl. Mech. Engrg.* **59** (1986) 85–100.

[39] H. Ishii and P.-L. Lions, Viscosity solutions of fully nonlinear second-order elliptic partial differential equations. *J. Differ. Eq.* **83** (1990) 26–78.

[40] G. Loeper and F. Rapetti, Numerical solution of the Monge−Ampère equation by a Newton's algorithm. *C. R. Math. Acad. Sci. Paris* **340** (2005) 319–324.

[41] B. Mohammadi, Optimal transport, shape optimization and global minimization. *C. R. Acad Sci Paris, Ser. I* **351** (2007) 591–596.

[42] M. Neilan, A nonconforming Morley finite element method for the fully nonlinear Monge−Ampère equation. *Numer. Math.* **115** (2010) 371–394.

[43] A. Oberman, Wide stencil finite difference schemes for the elliptic Monge−Ampère equations and functions of the eigenvalues of the Hessian. *Discr. Contin. Dyn. Syst. B* **10** (2008) 221–238.

[44] V.I. Oliker and L.D. Prussner, On the numerical solution of the equation $z_{xx}z_{yy} - z_{xy}^2 = f$ and its discretization, I. *Numer. Math.* **54** (1988) 271–293.

[45] M. Picasso, F. Alauzet, H. Borouchaki and P.-L. George, A numerical study of some Hessian recovery techniques on isotropic and anisotropic meshes. *SIAM J. Sci. Comput.* **33** (2011) 1058–1076.

[46] A.V. Pogorelov, *Monge−Ampère Equations of Elliptic Type*. P. Noordhooff, Ltd, Groningen, Netherlands (1964).

[47] L. Reinhart, On the numerical analysis of the Von Kármán equation: mixed finite element approximation and continuation techniques. *Numer. Math.* **39** (1982) 371–404.

[48] D.C. Sorensen and R. Glowinski, A quadratically constrained minimization problem arising from PDE of Monge−Ampère type. *Numer. Algor.* **53** (2010) 53–66.

[49] A.N. Tychonoff, The regularization of incorrectly posed problems. *Doklady Akad. Nauk. SSSR* **153** (1963) 42–52.

[50] V. Zheligovsky, O. Podvigina and U. Frisch, The Monge−Ampère equation: Various forms and numerical solution. *J. Comput. Phys.* **229** (2010) 5043–5061.