Numer. Math. (2001) 88: 299–318
Digital Object Identifier (DOI) 10.1007/s002110000235

Numerische Mathematik

A numerical approach to variational problems subject to convexity constraint

G. Carlier[1], **T. Lachand-Robert**[2], **B. Maury**[2]

[1] Université Paris IX Dauphine, Ceremade, France; e-mail: carlier@ceremade.dauphine.fr
[2] Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique, 75252 Paris Cedex 05, France; e-mail: {lachand,maury}@ann.jussieu.fr, http://www.ann.jussieu.fr

Received November 24, 1999 / Published online October 16, 2000 /

Summary. We describe an algorithm to approximate the minimizer of an elliptic functional in the form $\int_\Omega j(x,u,\nabla u)$ on the set $\mathcal{C}$ of *convex* functions u in an appropriate functional space X. Such problems arise for instance in mathematical economics [4]. A special case gives the convex envelope u_0^{**} of a given function u_0. Let (T_n) be any quasiuniform sequence of meshes whose diameter goes to zero, and I_n the corresponding affine interpolation operators. We prove that the minimizer over $\mathcal{C}$ is the limit of the sequence (u_n), where u_n minimizes the functional over $I_n(\mathcal{C})$. We give an implementable characterization of $I_n(\mathcal{C})$. Then the finite dimensional problem turns out to be a minimization problem with linear constraints.

Mathematics Subject Classification (1991): 65K10

1 Introduction

Let Ω be some bounded open convex subset of $\mathbb{R}^2$, and

$$\mathcal{C} := \{u : \Omega \to \mathbb{R}\,;\, u \text{ is convex in } \Omega\}\,.$$

We consider the variational problem subject to convexity constraint:

$$\inf_{u\in\mathcal{C}\cap K} J(u) \quad \text{with} \quad J(u) = \int_\Omega j(x,u(x),\nabla u(x))\,dx \tag{1}$$

where K is a closed convex subset of a given space $X = H^1(\Omega)$ or $X = L^2(\Omega)$, and j is a quadratic function of u and ∇u. We assume that $\mathcal{C}\cap K$ is

Correspondence to: T. Lachand-Robert

non-empty. If J is lower semicontinuous coercive and strictly convex on K, then existence and uniqueness of a minimizer of (1) directly follows from standard arguments. Throughout this paper, we shall always assume that J is such a functional.

A first example of such a functional is

$$J_f(u) = \int_{\Omega} \left[\frac{1}{2} |\nabla u(x)|^2 + f(x)u(x) \right] dx \tag{2}$$

with $f \in L^2(\Omega)$ given. Consider $J = J_f$, and $K = X = H_0^1(\Omega)$, then (1) is a classical projection problem in H_0^1. Indeed it is equivalent to find the projection of $u_f = \Delta^{(-1)} f$ onto $\mathcal{C}$ (the resolvent of the Laplace operator is intended in H_0^1, that is, with homogeneous Dirichlet boundary conditions), as one can see from the identity

$$2J_f(u) = \|\nabla(u - u_f)\|^2_{L^2(\Omega)} - \|\nabla u_f\|^2_{L^2(\Omega)} .$$

We can generalize to $f \in H^{-1}(\Omega)$, writing $\langle f, u \rangle$ instead of $\int f(x)u(x)\, dx$.

Another example is $X = L^2(\Omega)$,

$$J_{u_0}(u) = \|u - u_0\|^2_{L^2(\Omega)} \tag{3}$$

where $u_0 \in L^2(\Omega)$ is given, and

$$K_{u_0} = \{ u \in L^2(\Omega) \; ; \; u \leq u_0 \text{ a.e.} \} \tag{4}$$

Then the solution of (1) with $J = J_{u_0}$ and $K = K_{u_0}$ is the convex envelope u_0^{**} of u_0, that is the largest convex function in K_{u_0} [3], [5].

We will prove in Sect. 4 that the solutions of these two problems with $u_0 \in H_0^1(\Omega)$ and $f = \Delta u_0$ are, in general, different.

The aim of this paper is to give a numerical scheme to approximate solutions of (1).

1.1 Notations

In all the following we will use classical notations and assumptions from numerical analysis. Let $(T_n)_{n\in\mathbb{N}}$ be a sequence of quasiuniform regular triangulations of the domain Ω, $M_i^n = (x_i^n, y_i^n) \in \Omega$, $i = 1, \ldots, k_n$, are the nodes of T_n, and h_n is the largest edge length of all triangles in T_n. We assume that $h_n \to 0$ as $n \to \infty$.

We will note $N_n := \{1, \ldots, k_n\}$ and

$$\partial N_n := \{ i \in N_n : M_i^n \in \partial\Omega \}, \qquad \mathring{N}_n := N_n \setminus \partial N_n.$$

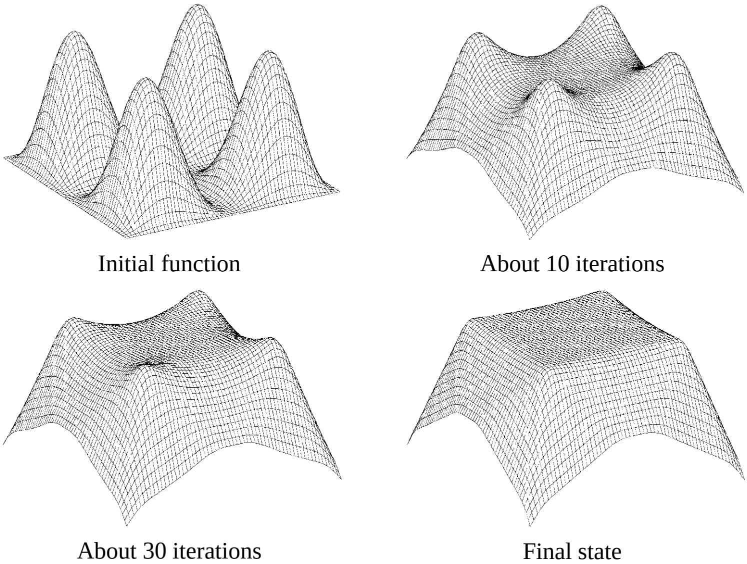

Fig. 1. Iterations of the algorithm. $H^1([0,1]^2)$ projection on $\mathcal{C}$ of $-x(1-x)(2x-1)^2\,y(1-y)(2y-1)^2$

Then we define

$$E_n := \{u \in C^0(\overline{\Omega}) : u \text{ is affine on each triangle of } T_n\} = I_n(C^0(\overline{\Omega})),$$

and

$$\mathcal{C}_n := I_n(\mathcal{C}) \quad \text{and} \quad K_n := I_n(K),$$

where I_n is the affine Lagrange interpolate operator from $C^0(\overline{\Omega})$ to E_n. One can easily show that $\mathcal{C}_n$ is a finite dimensional closed convex cone with nonempty interior.

1.2 Approach

Our basic idea is to approximate problem (1) by:

$$\min_{u \in \mathcal{C}_n \cap K_n} J(u) \tag{5}$$

and let n go to infinity.

This scheme is therefore based on external approximations of the cone of convex functions. As noted by P. Choné [2], there is very little hope that methods where $\mathcal{C}$ is internally approximated by $\mathcal{C} \cap E_n$ should converge to the solution of (1). Indeed the affine Lagrange interpolate of a convex function need not be convex and the cone $\mathcal{C}_n$ is in some sense much bigger than $\mathcal{C} \cap E_n$. This somehow surprising fact is enlighted by the following example: consider $\Omega = (0,1)^2$, with a mesh consisting of two triangles having their common edge in $\{x_1 = x_2\}$. The function $u(x_1, x_2) = (x_1 + x_2)^2$ is convex whereas its interpolate is a *concave* function.

More precisely, we have the proposition:

Proposition 1 *Assume that there are 2 directions h and k such that:*

$$(\nu \cdot h).(\nu \cdot k) \geq 0$$

for every vector ν which is normal to an edge of every triangle of the triangulation T_n, for all n. Then if u is the limit in L^∞_{loc} of a sequence (u_n), with $u_n \in \mathcal{C} \cap E_n$ for all n, we have:

$$\frac{\partial^2 u}{\partial h\, \partial k} \geq 0 \tag{6}$$

in the sense of Radon measures.

For instance, in a structured mesh of the form , the normal vectors are $v_1 = (1,0)$, $\nu_2 = (0,1)$ and $\nu_3 = (1,-1)$. Hence, we can choose $h = (1,0)$, $k = (0,-1)$, and get

$$u = \lim u_n \text{ in } L^\infty_{\text{loc}}, \text{ with } u_n \in \mathcal{C} \cap E_n \quad \Longrightarrow \quad \frac{\partial^2 u}{\partial x\, \partial y} \leq 0$$

in the sense of measures. This inequality obviously does not hold for all convex functions.

As a consequence, it appears that convex functions that do not satisfy the constraints (6) cannot be approximated (even in the sense of distributions) by convex functions of E_n. This is the very reason for which we chose an external approximation scheme.

Proof. This result is proved in [2] but we recall it here for sake of completeness. Let u be in E_n, then $u \in \mathcal{C}$ if and only if, for every pair of adjacent triangles 1 and 2 of T_n, we have:

$$(q_2 - q_1) \cdot \nu_{12} \geq 0,$$

where q_i is the value of ∇u in triangle $i = 1, 2$, and ν_{12} is the normal unit vector pointing from 1 to 2.

Assume now that h and k satisfy the assumption of the previous proposition and let φ be some nonnegative smooth function with compact support in Ω. Summing up Green's Formula in every triangle of T_n yields:

$$\frac{\partial^2 u}{\partial h \partial k}, \varphi > = - \sum_{T \in T_n} \left\langle \frac{\partial u}{\partial h}, \frac{\partial \varphi}{\partial k} \right\rangle$$
$$= \sum_e ((q_2 - q_1) \cdot \nu_{12})(\nu_{12} \cdot h)(\nu_{12} \cdot k) \int_e \varphi(s) ds \geq 0$$

where the last summation is taken over all interior edges e of T_n. □

A similar proposition can be given as a pointwise property:

Proposition 2 *Let $u \in C^2(\overline{\Omega})$ be a convex function which is limit, in $C^0(\Omega)$, of a sequence $(u_n) \subset \mathcal{C} \cap E_n$.*

Let $(M^n_{i_n})$ be a convergent sequence of nodes of the triangulations, with limit $M \in \Omega$. Let, for all n, $M^n_{j_n}$ be a node adjacent to $(M^n_{i_n})$, and ν a cluster point of the sequence $\frac{M^n_{i_n} - M^n_{j_n}}{|M^n_{i_n} - M^n_{j_n}|}$. Then

$$\frac{\partial^2 u}{\partial \nu \, \partial \nu^\perp}(M) \geq 0$$

where $\nu^\perp$ is normal to ν and $(\nu, \nu^\perp)$ is direct.

Proof. Since u_n converges to u uniformly on any compact subset of Ω, and all are convex functions, ∇u_n converges to ∇u a.e. in Ω.

The jump of ∇u_n on the edge $[M^n_{i_n}, M^n_{j_n}]$ is nonnegative by convexity; passing to the limit for a subsequence, the property follows immediately. □

2 Convergence

Let u_n (respectively $\overline{u}$) denote the solution of (5) (respectively (1)). The following convergence result holds:

Theorem 1 *The sequence (u_n) converges to $\overline{u}$ strongly in X and uniformly on all compact subsets of Ω.*

We will prove this theorem only in the case of the projection problem that is, $J = J_f$ and $K = X = H^1_0(\Omega)$. Other cases, with J strictly convex and coercive, are similar. (In particular, the proof is even simpler for the gradient independent case since it does not require an estimate on the gradient.)

In order to prove this property, we first need the technical result:

Lemma 1 *There exists $C > 0$ such that, for all $v \in W^{1,\infty}(\Omega) \cap \mathcal{C}$:*

$$(7) \qquad \forall n \in \mathbb{N}, \qquad \|\nabla I_n(v)\|_{L^\infty(\Omega)} \leq C \|\nabla v\|_{L^\infty(\Omega)} .$$

and $\nabla I_n(v) \to \nabla v$ a.e. in Ω.

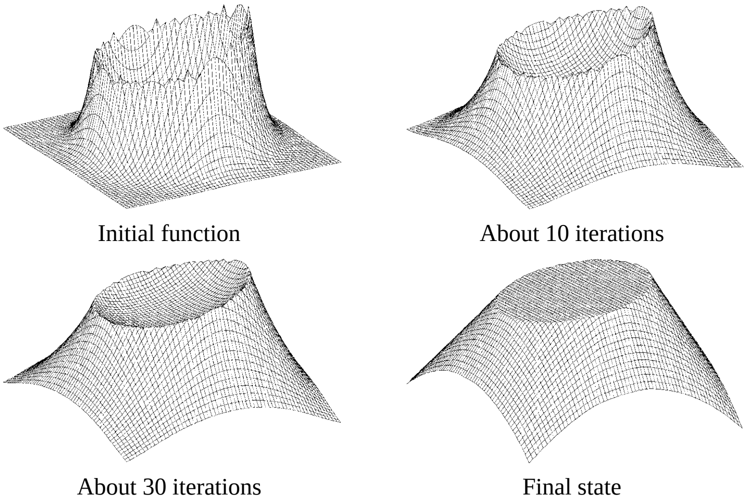

Fig. 2. Iterations of the algorithm. $H^1([0,1]^2)$ projection on $\mathcal{C}$ of $-(4+5xy^2)$ $\mathrm{e}^{-30[(x-\frac{1}{2})^2+(y-\frac{1}{2})^2]}$

Proof.

Step 1. Let T be an element of T_n, with vertices A, B, C. Writing $e_1 = (B-A)/|B-A|$, we have

$$\nabla I_n(v)\cdot e_1 = \frac{v(B)-v(A)}{|B-A|} \le \|\nabla v\|_{L^\infty}.$$

A similar relation holds for $e_2 := (C-A)/|C-A|$. Since the triangulations are quasiuniform there exists $C>0$ such that (7) is satisfied.

Step 2. Let D be the set of differentiability points of v which do not belong to any edge of the triangulations; D is clearly of full Lebesgue measure in Ω. For any $M \in D$, let $([A_n,B_n,C_n])_{n\in\mathbb{N}}$ be the sequence of triangles of T_n containing M and whose vertices A_n, B_n, C_n converge to M. Define also the unit vectors

$$e_1^n := \frac{B_n-A_n}{|B_n-A_n|} \qquad \text{and} \qquad e_2^n := \frac{C_n-A_n}{|C_n-A_n|}.$$

Let p_n, q_n, r_n be some subgradients of v respectively at A_n, B_n, C_n. By monotonicity we have, for all n:

$$p_n\cdot e_1^n \le \frac{v(B_n)-v(A_n)}{|B_n-A_n|} = \nabla[I_n(v)](M)\cdot e_1^n \le q_n\cdot e_1^n$$
$$p_n\cdot e_2^n \le \frac{v(C_n)-v(A_n)}{|C_n-A_n|} = \nabla[I_n(v)](M)\cdot e_2^n \le r_n\cdot e_2^n.$$

Since v is differentiable at M, sequences p_n, q_n, r_n all converge to $\nabla v(M)$ (*cf.* [5]), and we get:

$$\Big(\nabla[I_n(v)](M) - \nabla v(M)\Big) \cdot e_i^n \longrightarrow 0, \qquad i = 1, 2.$$

Since the triangulations are quasiuniform, it follows that $\nabla[I_n(v)(M)]$ converges to $\nabla v(M)$.

This ends the proof of the lemma. $\square$

Proof of Theorem 1. We recall that u_n is the projection of $u_f := \Delta^{(-1)} f$ onto $\mathcal{C}_n$ so that:

$$\forall n \in \mathbb{N}, \qquad \|\nabla u_n\|_{L^2(\Omega)} \leq \|\nabla u_f\|_{L^2(\Omega)} \tag{8}$$

Since u_n is a minimizer in $\mathcal{C}_n$, we have:

$$\forall v \in \mathcal{C} \cap H_0^1(\Omega), \qquad J(u_n) \leq J(I_n(v)). \tag{9}$$

Hence for all $\varepsilon > 0$, there exist $v_\varepsilon \in W_0^{1,\infty} \cap \mathcal{C}$ such that:

$$J(v_\varepsilon) \leq J(\overline{u}) + \varepsilon. \tag{10}$$

From Lemma 1, there exists $C > 0$ such that:

$$\|\nabla I_n(v_\varepsilon)\|_{L^\infty} \leq C \, \|\nabla v_\varepsilon\|_{L^\infty} \qquad \text{and} \qquad \nabla I_n(v_\varepsilon) \to \nabla v_\varepsilon \text{ a.e.}$$

Hence, by Lebesgue's Dominated Convergence Theorem, we get:

$$J(I_n(v_\varepsilon)) \to J(v_\varepsilon).$$

Taking (9) and (10) into account, and since ε is arbitrary, we deduce:

$$\limsup_{n\to\infty} J(u_n) \leq J(\overline{u}). \tag{11}$$

By (8), we may extract a subsequence, again labeled u_n and find some $u \in H_0^1$ such that u_n converges to u a.e. and strongly in $L^2(\Omega)$ and ∇u_n converges to ∇u weakly in $L^2(\Omega)$.

We will prove that u is convex. By definition, for all n, $u_n = I_n(v_n)$, for some $v_n \in \mathcal{C}$. Let us fix now some convex set $\omega \subset\subset \Omega$.

Let us show first that (∇v_n) is bounded in $L^\infty(\omega)$. If not, there would exist $x_n \in \omega$ such that, up to a subsequence, $|\nabla v_n(x_n)| \to +\infty$.
Up to subsequences, we may also assume that x_n is converging and:

$$d_n := \frac{\nabla v_n(x_n)}{|\nabla v_n(x_n)|} \longrightarrow d \in S^1.$$

Now, let $x_0 \in \Omega$ be such that, for n large enough:

$$(x_0 - x_n) \cdot d_n \geq \frac{1}{2}|x_0 - x_n|.$$

(Such a point exists since $(x_n) \subset \omega \subset\subset \Omega$ and d_n converges.) Since v_n is convex, we get:

$$|\nabla v_n(x_0)| \geq \frac{1}{2}\,|\nabla v_n(x_n)| \longrightarrow +\infty.$$

Let $e_n := \nabla v_n(x_0)/\,|\nabla v_n(x_0)|$ and, extracting subsequences, assume that it converges to $e \in S^1$. Define

$$\Sigma := \left\{p \in S^1 : \, p \cdot e \geq \frac{2}{3}\right\} \qquad \text{and} \qquad Q := (x_0 + \mathbb{R}_+\Sigma) \cap \Omega.$$

If n is large enough, then for all $p \in \Sigma$, $p \cdot e_n \geq \frac{1}{2}$. Then, for any $x \in Q$ (with $x = x_0 + tp$, $t > 0$, $p \in \Sigma$), the following holds:

$$|\nabla v_n(x)| \geq \nabla v_n(x) \cdot p \geq \nabla v_n(x_0) \cdot p \geq \frac{1}{2}\,|\nabla v_n(x_0)|\,.$$

Since the rightmost member tends to infinity independently of $x \in Q$, it follows that $\|\nabla v_n\|_{L^2(Q)} \to +\infty$. On the other hand, since the triangulations are quasiuniform and v_n is convex, this also implies $\|\nabla u_n\|_{L^2(\Omega)} \to +\infty$ which yields a contradiction with (8).

Hence, (∇v_n) is bounded in $L^\infty(\omega)$. Standard interpolate estimate yields that there exists a constant C_ω such that:

$$\|u_n - v_n\|_{L^\infty(\omega)} \leq C_\omega\, h_n\, \|\nabla v_n\|_{L^\infty(\omega)} \longrightarrow 0. \tag{12}$$

This implies in particular that v_n converges to u in $L^2(\omega)$; hence u is convex in ω. Since ω is arbitrary, u is convex in Ω.

From (11) and since u is convex, we deduce that $u = \overline{u}$. Since (u_n) is bounded in $H_0^1(\Omega)$ and $\overline{u}$ is the only cluster point of (u_n) in the weak topology of $H_0^1(\Omega)$, we deduce that the whole sequence (u_n) converges weakly to $\overline{u}$. On the other hand, (11) yields:

$$\|\nabla u_n\|_{L^2(\Omega)} \longrightarrow \|\nabla \overline{u}\|_{L^2(\Omega)}$$

and then (u_n) converges strongly to $\overline{u}$.

It remains to show that u_n converges uniformly to $\overline{u}$ on compact subsets. Let ω be any relatively compact open convex subset of Ω. We have

$$\|u_n - \overline{u}\|_{L^\infty(\omega)} \leq \|u_n - v_n\|_{L^\infty(\omega)} + \|v_n - \overline{u}\|_{L^\infty(\omega)}\,.$$

From (12), we just have to show that v_n converges to $\overline{u}$ in $L^\infty(\omega)$. Since we know that the sequence (v_n) is uniformly Lipschitz, this is a relatively

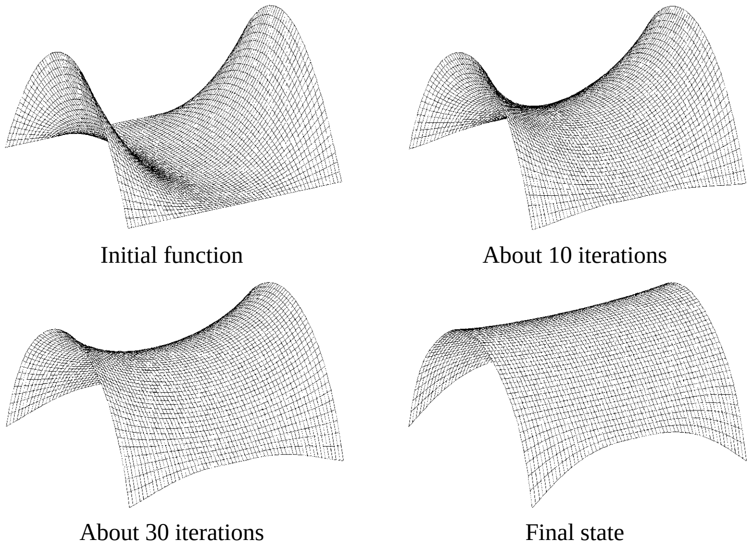

Fig. 3. Iterations of the algorithm. $H^1([0,1]^2)$ projection on $\mathcal{C}$ of $-x(1-x)(y-\frac{1}{2})^2$

compact sequence in $C^0(\overline{\omega})$ from Ascoli theorem. Since v_n has the same L^2 limit than u_n from (12), that is $\overline{u}$, the whole sequence converges to $\overline{u}$ in $C^0(\overline{\omega})$.

This ends the proof of the theorem. □

3 Characterization of cones $\mathcal{C}_n$ and the finite dimensional problems

In order to construct a numerical scheme for (5), we have to characterize more precisely the set $\mathcal{C}_n$. We are mainly interested in characterization in the form of a finite number of affine constraints on the values $z_i = u(M_i^n)$, since then the functional can be expressed as a quadratic form of the (z_i), using standard relations in numerical analysis. The minimization of a quadratic functional in a set defined as the intersection of a finite number of hyperplanes is called ‘quadratic programming’, and is very classical in the literature.

In the following, we will use some useful notations for points in $\mathbb{R}^2$. We note $[A,B,C] := \mathrm{co}\{A,B,C\}$ the closed triangle generated by three points. We recall that the area of this triangle is half the absolute value of

$$\begin{aligned}[A:B:C] &:= (B-A)\wedge(C-A)\\ &= (x_B-x_A)(y_C-y_A)-(x_C-x_A)(y_B-y_A).\end{aligned}$$

If this area is nonzero, we can define for all $M \in [A, B, C]$ its barycentric coordinates with respect to this triangle:

$$\alpha(M) := \left|\frac{[M:B:C]}{[A:B:C]}\right|, \ \beta(M) := \left|\frac{[M:C:A]}{[A:B:C]}\right|,$$
$$\gamma(M) := \left|\frac{[M:A:B]}{[A:B:C]}\right|.$$

They sum to 1 and

$$M = \alpha A + \beta B + \gamma C. \tag{13}$$

If $[A : B : C] = 0$, that is for instance $B \in [A, C]$, we extend this definition by setting for instance $\beta = 0$, $\alpha = AM/AC$, $\gamma = MC/AC$, so that (13) remains valid. (The barycentric coordinates are not unique in this degenerate case.)

A first characterization of cone $\mathcal{C}_n$ is given by:

Theorem 2 *Let $u \in E_n$ and $z_i = u(M_i^n)$. Then $u \in \mathcal{C}_n$ if and only if, for all $i, j, k, l \in (N_n)^4$ such that $M_i^n \in [M_j^n, M_k^n, M_l^n]$, we have*

$$z_i \leq \alpha(M_i^n) z_j + \beta(M_i^n) z_k + \gamma(M_i^n) z_l \tag{14}$$

where α, β, γ are the barycentric coordinates in $[M_j^n, M_k^n, M_l^n]$.

Note that even if the barycentric coordinates are not unique (for instance if $k = l$), they all give the same right member in (14).

Proof. If $u = I_n(v)$, with $v \in \mathcal{C}$, we have $z_i = v(M_i^n)$. Since v is convex, (14) follows immediately.

Let us prove that (14) implies that $u \in \mathcal{C}_n$. Consider $P_i := (x_i^n, y_i^n, z_i) \in \mathbb{R}^3$, Q_0 the convex hull of $\{P_i\}_{i \in N_n}$ in $\mathbb{R}^3$ and

$$Q := \bigcup_{t \geq 0} (Q_0 + t e_3)$$

where $e_3 := (0, 0, 1)$. It is easy to check that Q is a closed convex unbounded subset of $\mathbb{R}_3$ whose extremal points are included in $\{P_i\}_{i \in N_n}$ and having the graph property: there exists a function $v : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ such that

$$Q = \left\{(x, y, z) \in \mathbb{R}^3 : z \geq v(x, y)\right\}.$$

Notice that v is convex since its epigraph Q is convex. Moreover, if $D \subset \overline{\Omega}$ is the projection of Q_0 onto $\mathbb{R}^2$, we see that D is the union of all triangles of T_n, and that $v(M)$ is finite if and only if $M \in D$. The restriction of v to D can be written as the maximum of a finite number of affine functions (since Q_0 is a polyhedron); hence, there exists a convex function $w \in \mathcal{C}$ such that $v \equiv w$ in D.

We claim that $P_i \in \partial Q$ for all i. For if not, we can find i such that P_i is an interior point of Q, that is $z_i > v(x_i^n, y_i^n)$. Since the point $(x_i^n, y_i^n, v(x_i^n, y_i^n))$ is in ∂Q, it belongs to a two-dimensional face of Q; from Caratheodory's theorem, it belongs to a triangle of extremal points $[P_j, P_k, P_l]$; in particular by projection, $M_i^n \in [M_j^n, M_k^n, M_l^n]$. Notice that we have $v(x_j^n, y_j^n) = z_j$ and similar relations hold for k, l, since P_j, P_k, P_l are extremal. Let α, β, γ some barycentric coordinates for M_i^n in $[M_j^n, M_k^n, M_l^n]$; we then have:

$$\alpha z_j + \beta z_k + \gamma z_l = v(x_i^n, y_i^n) < z_i$$

and this contradicts (14).

Hence $z_i = v(M_i^n) = w(M_i^n)$ for all i. This implies that $u = I_n(w) \in \mathcal{C}_n$ and the proof is complete. □

As a consequence, problem (5) turns out to be finite dimensional quadratic programming problem for which the set of linear constraints is given by the previous proposition. Unfortunately, the number of constraints in (14) is of order $O(k_n^4)$, which is very large. But there are plenty of redundancies in those relations:

Theorem 3 *Under the same assumptions than in Theorem 2,* $u \in \mathcal{C}_n$ *if and only if (14) is satisfied for all indices* (i, j, k, l) *such that* $M_i^n \in [M_j^n, M_k^n, M_l^n]$ *and*

$$\forall p \notin \{i, j, k, l\}, \qquad M_p^n \notin [M_j^n, M_k^n, M_l^n]. \tag{15}$$

Moreover, this characterization is optimal in the following sense: if the indices (i_0, j_0, k_0, l_0) *satisfy (15), there exists* $u_0 \in E_n$ *such that* $u_0 \notin \mathcal{C}_n$*,* u_0 *satisfy (14) for all indices* $(i, j, k, l) \neq (i_0, j_0, k_0, l_0)$ *satisfying (15).*

We will give the proof below, but let us first give some consequences. We note that, if M_j^n, M_k^n, M_l^n are non aligned, the additional condition (15) expresses that M_i^n is the only vertex of T_n in the non-extremal points of the triangle. Up to permutations, the only other case is $k = l$, and then (15) expresses that M_i^n is the only vertex in (M_j^n, M_k^n).

These conditions appear to be very simple for a structured mesh:

Corollary 4 *Assume that for some* n, T_n *is a structured mesh, that is,* $\{M_i^n\}_{i \in N_n} = \Omega \cap \mathbb{Z}^2 =: \mathcal{M}$. *Let* $u \in E_n$*; we extend it to* $\mathbb{Z}^2$*, defining* $u(\alpha) = +\infty$ *for all* $\alpha \in \mathbb{Z}^2 \setminus \mathcal{M}$. *Then* $u \in \mathcal{C}_n$ *if and only if, for all* $\alpha = (\alpha_1, \alpha_2) \in \mathcal{M}$ *we have:*

$$\begin{aligned} &\forall (\sigma_1, \sigma_2) \in \{-1, +1\}^2, \\ &u(\alpha) \leq \tfrac{1}{3}[u(\alpha_1 + \sigma_1, \alpha_2) + u(\alpha_1, \alpha_2 + \sigma_2) + u(\alpha_1 - \sigma_1, \alpha_2 - \sigma_2)] \end{aligned} \tag{16}$$

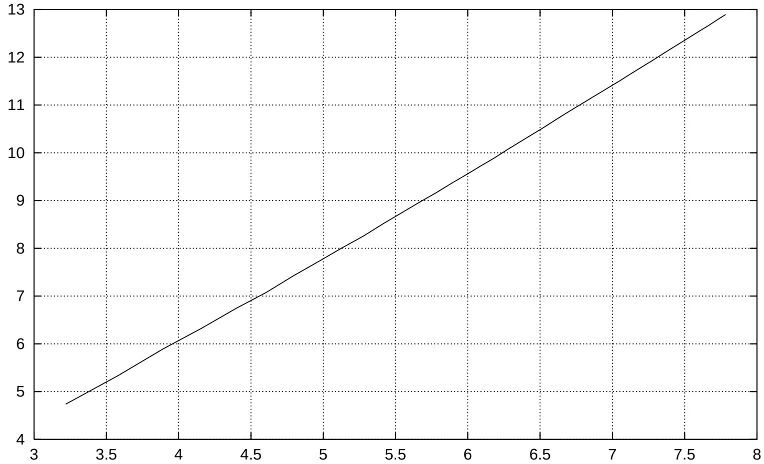

Fig. 4. Graph of the number of constraints C with respect to the total number of points N in the mesh, in log-log scale. A law in the form $C \simeq N^{1.8}$ appears

and

$$u(\alpha) \leq \frac{1}{2}[u(\alpha+\beta) + u(\alpha-\beta)] \tag{17}$$

for all $\beta \in \mathbb{Z}^2$ *such that either* $\beta = (0,1)$, *or* $\beta = (1,0)$, *or* $\beta = (\pm k, l)$ *where* $k \geq 1$, $l \geq 1$ *are integers satisfying* $k \wedge l = 1$.

This follows from Theorem 3 by observing that, if three points of $\mathbb{Z}^2$ are aligned, but their segment does not contain any other point of $\mathbb{Z}^2$, then they have the form $\alpha, \alpha+\beta, \alpha-\beta$ with β as described in the corollary. And if a triangle of $\mathbb{Z}^2$ contains only $\alpha \in \mathbb{Z}^2$ in its interior or boundary (except for the vertices), it has the form $[(\alpha_1+\sigma_1, \alpha_2), (\alpha_1, \alpha_2+\sigma_2), (\alpha_1-\sigma_1, \alpha_2-\sigma_2)]$ with $(\sigma_1, \sigma_2) \in \{-1, +1\}^2$; in this case, the barycentric coordinates of α are equal.

Hence, for a structured mesh with k_n vertices, the number of constraints is of order $k_n^{1.8}$ approximately (see Fig. 4).

Proof of Theorem 3. In this proof, n is constant; we drop upperscript n for simplicity. We note

$$G := \Big\{(i,j,k,l) \in N^4 : \ M_i \in [M_j, M_k, M_l] \setminus \{M_j, M_k, M_l\}\Big\}.$$

We assume that (14) is satisfied for all indices $(i,j,k,l) \in G$ such that (15) is satisfied. We would like to prove it for all other $(i,j,k,l) \in G$. The case $k = l$ (corresponding to one-dimensional simplexes) is easy: the restriction of a convex function on a line has a monotone derivative, so the property (14) has just to be verified for three consecutive points on the same line.

Hence we just have to consider two-dimensional simplexes. In order to shorten notations, we will write

$$[i:j:k] := (M_j - M_i) \wedge (M_k - M_i).$$

We note for further reference the algebraic identity, valid for all (i,j,k,l,p):

$$[i:p:j]\,[i:k:l] + [i:j:k]\,[p:l:i] = [i:k:p]\,[i:l:j]. \tag{18}$$

Let $G_0 \subset G$ be the set of indices (i,j,k,l) satisfying (14). If we assume that $G_0 \neq G$, then we can find $(i,j,k,l) \in G \setminus G_0$ such that $[M_j, M_k, M_l]$ has the smaller area. We then have (assuming that $[M_j, M_k, M_l]$ is direct):

$$[j:k:l]z_i > [i:k:l]z_j + [i:l:j]z_k + [i:j:k]z_l. \tag{19}$$

By assumption, (15) is not satified for (i,j,k,l): hence, there exists $p \notin \{i,j,k,l\}$ such that $M_p \in [M_j, M_k, M_l]$. We can even assume that $M_p \in [M_i, M_k, M_l]$ for instance, up to a permutation of the indices j,k,l. Since the area of $[M_i, M_k, M_l]$ is smaller than the area of $[M_j, M_k, M_l]$, we have $(p,i,k,l) \in G_0$:

$$[i:k:l]z_p \leq [p:k:l]z_i + [p:l:i]z_k + [p:i:k]z_l. \tag{20}$$

Since $M_p \in [M_i, M_k, M_l]$, and

$$M_i \in [M_j, M_k, M_l] = [M_p, M_k, M_l] \cup [M_j, M_p, M_l] \cup [M_j, M_k, M_p]$$

we must have $M_i \in [M_j, M_p, M_l]$ or $M_i \in [M_j, M_k, M_p]$. We assume the latter, the other case is similar. We can assume that $[i:k:p] \neq 0$ (in that case it is positive), since if M_i, M_k, M_p are aligned, they must also be aligned with M_l. And then this reduces to the case of one-dimensional simplexes.

Again, we must have $(i,j,k,p) \in G_0$ since the area of $[M_j, M_k, M_p]$ is smaller than the area of $[M_j, M_k, M_l]$:

$$[p:j:k]z_i \leq [i:k:p]z_j + [i:p:j]z_k + [i:j:k]z_p. \tag{21}$$

Mutliplying this relation by $[i:k:l]$ and using (20), we get

$$\begin{aligned}
&[i:k:l]\,[p:j:k]z_i \\
&\quad \leq [i:k:l]\,[i:k:p]z_j + [i:k:l]\,[i:p:j]z_k + \\
&\qquad\qquad + [i:j:k]\Big([p:k:l]z_i + [p:l:i]z_k + [p:i:k]z_l\Big) \\
&\quad \leq [i:k:l]\,[i:k:p]z_j + [i:k:p]\,[i:l:j]z_k + \\
&\qquad\qquad + [i:j:k]\,[i:k:p]z_l + [i:j:k]\,[p:k:l]z_i
\end{aligned}$$

taking (18) into account.

Exchanging k and i in (18), we get:

$$[p:j:k]\,[i:k:l] - [i:j:k]\,[p:k:l] = [i:k:p]\,[j:k:l]$$

so that the preceding inequality can be rewritten:

$$[i:k:p]\,[j:k:l]z_i \le [i:k:p]\,\Big([i:k:l]z_j + [i:l:j]z_k + [i:j:k]z_l\Big).$$

This contradicts (19) since $[i:k:p] > 0$. Hence, we have $G_0 = G$ and the proof of the first part is complete.

We now prove our assertion on the optimality. If $(i_0, j_0, k_0, l_0) \in G$ satisfies (15), we can find a compact convex set K such that

$$K \cap \{M_i\}_{i=1,\ldots,k_n} = \{M_{i_0}, M_{j_0}, M_{k_0}, M_{l_0}\}$$

and M_{i_0} is an interior point of K. Let v be any convex function satisfying

$$v \equiv 0 \text{ in } K \qquad \text{and} \qquad v > 0 \text{ in } \mathbb{R}^2 \setminus K.$$

For instance the function $v(M) = \mathrm{dist}(M, K)$ is convenient. Since v is continuous and K is compact,

$$\min_{p \notin \{i_0, j_0, k_0, l_0\}} v(M_p) > 0.$$

Hence there exists $\varepsilon > 0$ such that for all indices $(j,k,l) \neq (j_0, k_0, l_0)$ (up to permutations) satisfying $(i_0, j, k, l) \in G$, we have

$$\varepsilon \le \alpha v(M_j) + \beta v(M_k) + \gamma v(M_l) \tag{22}$$

where α, β, γ are the barycentric coordinates of M_{i_0} in $[M_j, M_k, M_l]$ (indeed, $\max(\alpha, \beta, \gamma)$ is bounded from below as (j,k,l) changes, since M_{i_0} is an interior point of K).

Consider the function $u \in E_n$ satisfying $u(M_p) = v(M_p)$ for all $p \neq i_0$ and $u(M_{i_0}) = \varepsilon$. If $(i,j,k,l) \in G$, with $i \neq i_0$, we have

$$\begin{aligned} u(M_i) = v(M_i) &\le \alpha(M_i)v(M_j) + \beta(M_i)v(M_k) + \gamma(M_i)v(M_l) \\ &\le \alpha(M_i)u(M_j) + \beta(M_i)u(M_k) + \gamma(M_i)u(M_l) \end{aligned}$$

using the convexity of v and $u \ge v$ at every node. Hence u satisfies (14) for these indices.

Also if $(j,k,l) \neq (j_0, k_0, l_0)$ (up to permutations) satisfies $(i_0, j, k, l) \in G$, then u satisfies (14) from the definition of ε in (22).

We conclude that u satisfies all constraints with indices $(i,j,k,l) \neq (i_0, j_0, k_0, l_0)$, but it is not in $\mathcal{C}_n$ since

$$u(M_{i_0}) = \varepsilon > 0 = \alpha(M_{i_0})v(M_{j_0}) + \beta(M_{i_0})v(M_{k_0}) + \gamma(M_{i_0})v(M_{l_0}).$$

This ends the proof of the proposition. $\square$

4 Convexification and projection

Let u_0 be some function in $H_0^1(\Omega)$ and $f = \Delta u_0$. We will prove in this section that the solution $\Pi_{\mathcal{C}}(u_0)$ of the minimization problem in $K = H_0^1(\Omega)$ with $J = J_f$ (as defined in (2)) is, in general, different from u_0^{**} (which is the minimizer of J_{u_0} defined in (3) on K_{u_0}).

Theorem 5 *We have $u_0^{**} = \Pi_{\mathcal{C}}(u_0)$ if and only if*

$$\langle \Delta u_0^{**}, u_0 - u_0^{**} \rangle = 0. \tag{23}$$

As a consequence, in dimension 1, we always have $u_0^{**} = \Pi_{\mathcal{C}}(u_0)$ since (23) is always satisfied. However, this is not true in higher dimensions (see remark hereafter).

Proof. Assume first that u_0^{**} is a minimizer of J_f in $H_0^1(\Omega)$. Then for all convex v with $v = 0$ in $\partial\Omega$, $\left\langle J_f'(u_0^{**}), v - u_0^{**} \right\rangle \geq 0$. Taking $v = 0$ here yields:

$$\langle \Delta(u_0 - u_0^{**}), u_0^{**} \rangle = \langle \Delta u_0^{**}, u_0 - u_0^{**} \rangle \leq 0.$$

Since $u_0 - u_0^{**} \geq 0$ and Δu_0^{**} is a nonnegative measure, we exactly get (23).

Conversely assume (23), then for all $h \in H_0^1(\Omega)$ such that $u_0^{**} + h$ is convex, the following holds:

$$\begin{aligned} J_f(u_0^{**} + h) - J_f(u_0^{**}) &\geq \left\langle J_f'(u_0^{**}), h \right\rangle \\ &= \int_\Omega \nabla u_0^{**} \cdot \nabla h + h \Delta u_0 \\ &= \langle \Delta(u_0 - u_0^{**}), h \rangle \end{aligned}$$

writing $h = v - u_0^{**}$ with $v \in H_0^1(\Omega) \cap \mathcal{C}$ in the latter and using (23) yields:

$$J_f(v) - J_f(u_0^{**}) \geq \langle \Delta(u_0 - u_0^{**}), v \rangle = \langle \Delta v, u_0 - u_0^{**} \rangle \geq 0$$

which proves that u_0^{**} minimizes J_f over $H_0^1(\Omega) \cap \mathcal{C}$. □

Remark. Condition (23) indicates that for almost every $x \in \Omega$, we have either $u_0(x) = u_0^{**}(x)$ or $\nabla u_0^{**} \equiv \text{const.}$ in a neighborhood of x. For instance, if Ω is the unit ball of $\mathbb{R}^2$, and $u_0(x) = |x|^3 - |x|$, we have $u_0^{**} = \min(-\frac{2}{3\sqrt{3}}, u_0)$ and (23) is satisfied.

On the other hand, if $u_0(x) = \sqrt{|x|} - 1$, then $u_0^{**}(x) = |x| - 1$ and (23) is not satisfied.

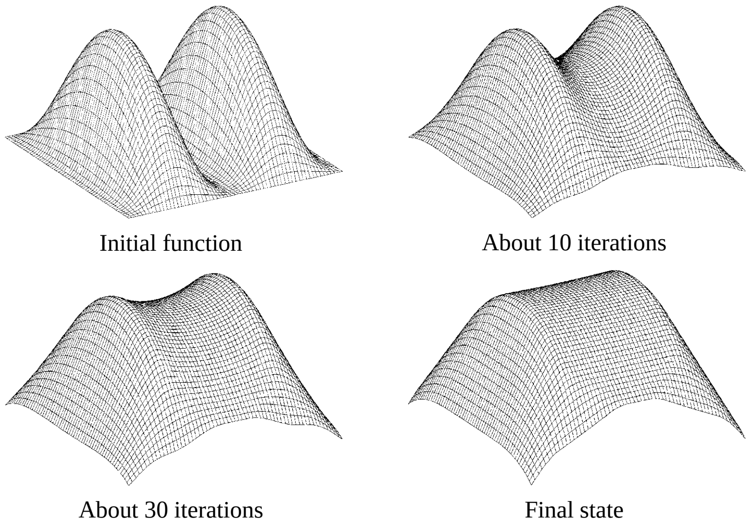

Fig. 5. Iterations of the algorithm. $H^1([0,1]^2)$ projection on $\mathcal{C}$ of $-x(1-x)\sin^2(2\pi y)$

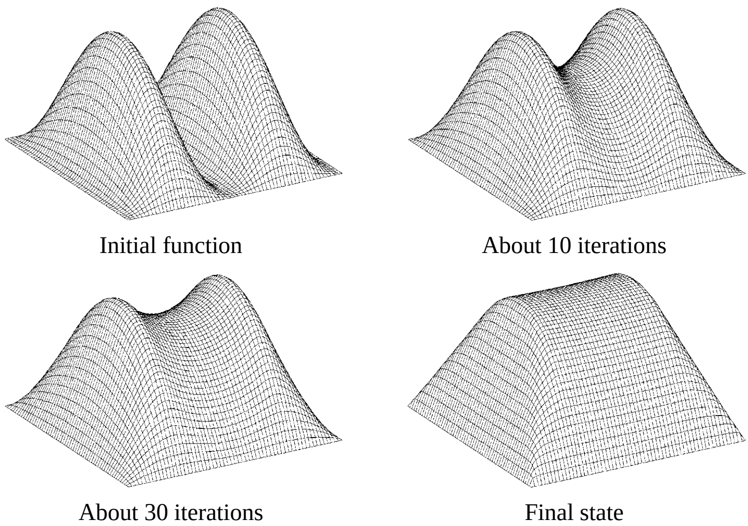

Fig. 6. Iterations of the algorithm. $H^1_0([0,1]^2)$ projection on $\mathcal{C}$ of the same function than in Fig. 5

5 Numerical solution

5.1 Algorithm

We consider a structured triangulation of the unit square $[0,1] \times [0,1] = \overline{\Omega}$. Any function $u \in \mathcal{C}_n$ can be written

$$u = \sum_{i=1}^{k_n} u_i w_i,$$

where (w_i) is the standard basis of E_n ($w_i(M_j) = \delta_{ij}$). In what follows, u will represent both the function of E_n and the vector of its components in this basis. The stiffness matrix $A = (a_{ij})$ is defined by

$$a_{ij} = \int_\Omega \nabla w_i \cdot \nabla w_j.$$

Let m be the number of constraints. The set of feasible states (see corollary 4) can be written

$$\mathcal{C}_n = \{u \in E_n\ ,\ \ Cu \leq 0\}\,, \tag{24}$$

where C is a $m \times k_n$ matrix.

We finally end up with a classical quadratic programming problem:

$$\begin{array}{l} \text{Find } u \in \mathcal{C}_n \text{ such that} \\ J(u) = \frac{1}{2}(Au, u) + (b, u) = \min_{v \in \mathcal{C}_n} J(v) \end{array} \tag{25}$$

We propose to solve this problem by a Uzawa-like algorithm [1]. The initial problem is replaced by the following: Find a saddle point for the Lagrangian defined for $(u, \lambda) \in \mathbb{R}^{k_n} \times \mathbb{R}^m_+$ by

$$L(v, \lambda) = \frac{1}{2}(Av, u) + (b, v) + (\lambda, Cv). \tag{26}$$

We denote by Π^+ the projection onto $\mathbb{R}^m_+$.

$$\lambda = (\lambda_i)_{1 \leq i \leq m} \in \mathbb{R}^m \longmapsto \Pi^+ \lambda = (\max(\lambda_i, 0))_{1 \leq i \leq m}.$$

As detailed in [1], an iterate of the algorithm is

$$u^k = A^{-1}(b - C^{\mathrm{T}} \lambda^k) \tag{27}$$

$$\lambda^{k+1} = \Pi^+ \left(\lambda^k + \rho\, C u^k\right) = \Pi^+ \left(\lambda^k + \rho\, C A^{-1}(b - C^{\mathrm{T}} \lambda^k)\right), \tag{28}$$

where $\rho > 0$ is the step parameter (see next section).

5.2 Numerical parameters

Weighting of the constraints From a theorical point of view, the initial problem remains the same if any constraint (row of C) is multiplied by any positive number, whereas the behaviour of the algorithm is likely to vary. Indeed, matrix C can be replaced by DC, where D is a diagonal $m \times m$ matrix with positive elements. Note that, as problem (26) does not admit a unique solution in λ since $m > k_n$, it may change completely the behaviour of the sequence (λ^k).

The choice we propose here is based on the following heuristic: given a field $u \in E_n$, we would like the m-dimensional vector DCu to be related to the distance between u and the set of convex functions. Using notations of corollary 4, we define $\delta_{\alpha\beta}(u)$ for any $u \in E_n$ by

$$\delta_{\alpha\beta}(u) = \frac{2u(\alpha) - u(\alpha-\beta) - u(\alpha+\beta)}{2\,|\beta|^2}, \tag{29}$$

where $\beta = (\beta_1, \beta_2) \in \mathbb{Z}^2$, $|\beta|^2 = \beta_1^2 + \beta_2^2$, and

$$\Delta_{\alpha\sigma}(u) = \frac{3u(\alpha) - u(\alpha_1+\sigma_1, \alpha_2) - u(\alpha_1, \alpha_2+\sigma_2) - u(\alpha-\sigma)}{4}, \tag{30}$$

where $\sigma = (\sigma_1, \sigma_2) \in \{-1, +1\}^2$, and we introduce the number

$$\eta(u) = \max(0, \max_{\alpha,\beta} \delta_{\alpha\beta}(u), \max_{\alpha,\sigma} \Delta_{\alpha\sigma}(u)). \tag{31}$$

We have the following proposition:

Proposition 3 *A function $u \in E_n$ is in $\mathcal{C}_n$ if and only if $\eta(u) \leq 0$. Furthermore, for any norm $\|\ \|$ on E_n, there exists a constant K such that,*

$$\forall u \in E_n, \qquad \mathrm{dist}(u, \mathcal{C}_n) = \inf_{v \in \mathcal{C}_n} \|u - v\| \leq K k_n \eta(u). \tag{32}$$

Proof. The first part is a direct consequence of corollary 4: all constraints have been multiplied by a positive number.

Let h be the mesh size, which verifies $h^2 \simeq 1/k_n$ for n large. Let us now define $\Lambda \in \mathcal{C}_n$ as the interpolate of the quadratic function $(x, y) \longmapsto x^2 + y^2$. A straightforward calculation shows that

$$\delta_{\alpha\beta}(u + \frac{\eta}{h^2}\Lambda) \leq 0 \qquad \forall \alpha,\ \beta, \tag{33}$$

and

$$\Delta_{\alpha\sigma}(u + \frac{\eta}{h^2}\Lambda) \leq 0 \qquad \forall \alpha,\ \sigma, \tag{34}$$

so that $u + \frac{\eta}{h^2}\Lambda \in \mathcal{C}_n$. (Actually $\frac{\eta}{h^2}$ is the smallest number τ such that $u + \tau\Lambda \in \mathcal{C}_n$.) Therefore,

$$\text{(35)}\quad \operatorname{dist}(u, \mathcal{C}_n) = \inf_{v\in\mathcal{C}_n} \|u - v\| \leq \left\| u - (u + \frac{\eta}{h^2}\Lambda) \right\| = \frac{\eta}{h^2} \|\Lambda\|,$$

which ends the proof, with $K = \|\Lambda\|$. □

The matrix C we used in computations is the algebraic form of the scaled constraints $\delta_{\alpha\beta}(u) \leq 0$, $\Delta_{\alpha\sigma}(u) \leq 0$, for all α, β, and σ.

Choice of ρ

Let α be the smallest eigenvalue of A. It is shown in [1] that the algorithm converges for any $\rho \in (0, \rho_c^0)$, with $\rho_c^0 = 2\alpha/\|C\|^2$. We propose a sharper upperbound $\rho_c^1 \geq \rho_c^0$,

$$\text{(36)}\qquad \rho_c^1 = \frac{2}{\|CA^{-1}C^{\mathrm{T}}\|},$$

for which convergence can be established as well. This critical value ρ_c^1 can be estimated numerically, using the fact that $CA^{-1}C^{\mathrm{T}}$ is symmetric, so that the 2-norm is the spectral radius.

It turns out to be much larger than ρ_c^0, leading to a faster converging algorithm. Furthermore, ρ_c^1 appears to be close to optimal, as the algorithm diverges as soon as $\rho \geq 1.1 \times \rho_c^1$.

Remark. The inequality $\rho_c^1 \gg \rho_c^0$ is due to

$$\text{(37)}\qquad \|CA^{-1}C^{\mathrm{T}}\| \ll \|C\|\,\|A^{-1}\|\,\|C^{\mathrm{T}}\|,$$

which is closely related to the nature of the problem we solve. More precisely, as the rows of C correspond to second order derivatives, $C^{\mathrm{T}}C$ is spectrally close to the discrete bilaplacian operator, which is basically A^2. Let us denote by $0 < \alpha_1 < \ldots < \alpha_{k_n}$ the eigenvalues of A. The smallest eigenvalue verifies $\alpha_1 \sim 2\pi^2 h^2$, where h is the mesh size, and α_{k_n} is a $O(1)$. Considering the idealized situation $C^{\mathrm{T}}C = A^2$, one can check easily that the spectrum of $CA^{-1}C^{\mathrm{T}}$ is the spectrum of A plus the eigenvalue 0 with multiplicity $m - k_n$, so that $\|CA^{-1}C^{\mathrm{T}}\|$ is $O(1)$. Besides the right-hand side of (37) is $\alpha_{k_n}^2/\alpha_1$, which is a $O(1/h^2)$. We therefore have

$$\text{(38)}\qquad \rho_c^0 = O(h^2), \qquad \rho_c^1 = O(1).$$

We checked numerically that these estimates hold for the actual matrices, and not only in the simplified situation $C^{\mathrm{T}}C = A^2$.

Some results of the computations are shown in Figs. 1–3, 5–6. For the clearness of the pictures, the z-axis is directed downward in these figures.

References

1. P. G. Ciarlet, Introduction à l'analyse numérique matricielle et à l'optimisation, Collection mathématiques appliquées pour la Maîtrise, Paris: Masson 1982
2. P. Choné, Étude de quelques problèmes variationnels intervenant en géométrie riemannienne et en économie mathématique, PhD Thesis, Univ. Toulouse I (1999)
3. I. Ekeland, R. Temam, Convex Analysis and Variational Problems. Amsterdam: North-Holland 1972
4. J.-C. Rochet, P. Choné, Ironing, Sweeping and Multidimensional screening. Econometrica **66**, 783–826 (1998)
5. R. T. Rockafellar, Convex Analysis. Princeton: Princeton University Press 1970