Optimization under convexity constraints: are the finite element discretizations consistent ?

Hervé V.J. Le Meur[1]

[1]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France

March 17, 2016

Abstract

It is proved in [1] that the problem of minimizing a Dirichlet-like functional of the function u_h discretized with P_1 Finite Elements, under the constraint that u_h be convex, cannot converge. Here, we first improve this result by proving that non-convergence is due to the mesh refinment lack of richness, remains local and is true even for any mesh. Then, we investigate the consistency of various natural discretizations (P_1 and P_2) of second order constraints (subharmonicity and convexity) without discussing the convergence. We also numerically illustrate convergence of a method proposed in the literature that is simpler than existing methods.

Keywords:convexity; finite elements; interpolation; conformal approximation; minimization
subjclass:26B25, 52A41, 65K10, 65N30, 90C25, 91B24

1 INTRODUCTION

This paper is devoted to the numerical discretization of optimization with constraints on the second order derivative, namely problems of the form

$$\left\{ \begin{array}{l} \inf J(u), \\ \mathbf{D}^2 u \in K, \end{array} \right. \tag{1}$$

where J is a functional and K is a subset of the set of symmetric matrices. Such problems appear in various contexts, in particular in physics and economics.

Since Newton, the shape of minimal resistance has been a topic of interest. It is called the Newton's problem and is of the type of (1). With some additional assumptions, it amounts to looking for a concave function that minimizes a nonlinear functional. See for instance the original book [2], the historical survey in [3], more recent theoretical results in [4, 5, 6] or a numerical result [7].

The associated problem of discretizing convex functions or bodies is indeed wider than expected. For instance, the Alexandrov's problem (see [8] and [7]), the Cheeger's constant [7, 9] and the Newton's problem [7] can be numerically studied.

In economics, it suffices to remember that utility functions are concave to see the very wide applicability of these problems. For instance, in [10], the authors are interested in finding the minimum of a convex and quadratic functional J over the set of convex functions

$$\min_{u\in K} J(u) \text{ for } J(u) = \int_\Omega \left(\frac{1}{2}\nabla u^T\,\mathbf{C}\,\nabla u - \mathbf{x}\cdot\nabla u + (1-\alpha)u\right)\mathrm{d}\mathbf{x}, \tag{2}$$

where $0 \le \alpha \le 1$ and $\mathbf{C}$ is a (2,2) symmetric and positive-definite matrix and K is given by

$$K = \{u \in H^1(\Omega), u \ge 0, u_x \ge 0, u_y \ge 0, u \text{ convex }\}. \tag{3}$$

So J is strictly convex and the set K is convex. It is then easy to check that there exists a unique minimizer of J over K. The regularity of the solutions is studied in [11]. Note that, when $\alpha = 1$ and in a domain Ω with nonnegative coordinates, the problem degenerates: an exact convex solution exists, up to an additive constant

$$\nabla u(\mathbf{x}) = \mathbf{C}^{-1}\mathbf{x} \Rightarrow u(\mathbf{x}) = \mathbf{x}'\mathbf{C}^{-1}\mathbf{x}/2 - \mathrm{Cst}, \tag{4}$$

that can be fixed if we force u to vanish at a point (if $\alpha = 1$).

In [1], Choné and Le Meur exhibit an obstruction to the convergence of the mere *discretization* of this problem through conformal P_1 Finite Element (FE) method on regular meshes. They also extensively use the explicit solution (4) to illustrate the lack of convergence when a predicted condition is no more satisfied. Here "conformal" means that the *discretized* function is supposed to satisfy exactly the continuous constraint. As a consequence of this result, no optimization process that would use such conformal P_1 discretization of both the functional and the constraint may converge for any solution on these meshes. Yet approximation theory easily proves convergence of every separate discretization of these terms.

A possible approach to circumvent the mesh problem is, first, to test whether a sample of values on *given points* (and not *mesh*) may be associated to a convex function or body. Then one must construct the associated mesh (so, function dependent !) so as to interpolate. This was done in [12] for C^1 FE by Leung and Renka. But such a regularity is too restrictive for us. This article [12] reviews other papers, some of them being commented as false. It proves that the issue is not so simple.

The H_0^1 projection of a given function is addressed in [13] through the saddle-point method. In this article, the authors even weight the convexity constraint to tune the convergence to the solution. The computations appear to be more robust than expected by theory. An attempt of explanation is given in [14].

In [15], the authors describe and implement an algorithm that optimizes not in the set of discretized *and* convex (i.e. conformal) functions but in the set of the convex functions after discretization (so that they may be non-convex). More precisely they characterize, for a specific structured set of cartesian points, the image through the P_1 discretization of a (continuous) convex function. This yields such a huge number of constraints on the function that it may not be recommended.

In [7], Lachand-Robert and Oudet address the problem of discretizing a convex body. They use the parametric generalization in order to discretize convex functions' graphs. In the functional, they isolate the dependence on the point x, the unit normal ν at x and the signed distance $\phi = \nu.x$. Although they notice that the three variables are "somehow redundant", they implement a gradient like method, based on the variation of only $\phi = x.\nu$. They address both the Alexandrov's problem, the Cheeger's sets and the Newton's problem. This was revisited since in [9].

Separately, Aguilera and Morin [16] prove convergence of a Finite Difference (FD) "approximation using positive semidefinite programs and discrete Hessians". In [17], the same authors prove even convergence of a Finite Element (FE) discretization of the weak Hessian.

Since FD are included in FE for convenient meshes, these two articles seem to contradict both [1] and the present article. Indeed, they do not, but we discuss them below (section 5).

More recently, in [18], Ekeland and Moreno propose a non-local discretization that relies on the representation of any convex function as the supremum of affine functions (its minorants). So their "discrete" representation of continuous functions is conformal but non-local. The major drawback is that the complexity increases drastically with dimension, due to the nonlocality, but it works.

Even more recently, two very different means of resolution were proposed. Both are non-conformal in the sense that the discretized solution is not convex. In the first one, Mérigot and Oudet [19] propose to discretize the convexity constraint by under sampling it. More precisely, they prove convergence of their algorithm if the constraint is forced only on a subset of the sampling points. In the second one, Mirebeau [20] proves that for a given grid, he may provide a sequence of sets of (less than four) points on which it suffices to force convexity to ensure convexity at convergence on this given grid.

There seems to be a convergent opinion through [20], [19], [17] that indeed the number of constraints must be not too large. A deeper discussion of this crucial point is postponed to a forthcoming article.

The present article deals with two main types of non-convergence results. On the one hand, the non-convergence in L^2 for conformal P_1 Finite Element (FE) is revisited after [1] from a theoretical point of view. On the other hand, the consistency of discretized linear optimization over second order constraints is investigated for various discretizations (P_1 and P_2) and various constraints: linear (subharmonicity) or nonlinear (convexity). The motivation for convexity has been stressed. Subharmonicity is only studied as an intermediate step toward convexity because it is linear.

In Section 2, first we restate some already known results, then we prove that non-convergence is purely local and so is true for any mesh. In Section 3, we investigate the consistency of various P_1 FE discretizations of our model problem. Section 4 is devoted to the consistency of P_2 FE discretizations (strong or weak convexity). We discuss the articles [16, 17] that could seem to contradict [1] in Section 5 and we conclude in Section 6.

2 THEORETICAL RESULTS ON THE P_1 FEM

In the present section, we first recall some already known results and make them more explicit.

2.1 Already known results

In [1], various results are proved. Since the goal in the present subsection is to extend this study, we need to remind the reader of these results. We consider an open convex and bounded domain Ω of $\mathbb{R}^2$, but we use only $\Omega = (a,b)^2$ (and a and b positive) in applications. The Lemma 3 of [1] states:

Lemma 1 *A function u_h, P_1 in every triangle of Ω's mesh, is convex if and only if, for any pair of adjacent triangles*

$$(\mathbf{q}_2 - \mathbf{q}_1)\cdot\mathbf{n} \geq 0, \tag{5}$$

where $\mathbf{q}_1$ (resp. $\mathbf{q}_2$) is the (constant) gradient of u_h inside triangle 1 (resp. 2) and $\mathbf{n}$ is the unit normal pointing from triangle 1 to 2.

As reminded in [1], convexity of u can be defined dually as $-\langle \nabla u \otimes \nabla v\rangle \succeq 0$, where $\mathbf{A} \succeq 0$ means the matrix $\mathbf{A}$ is positive semi-definite, v is a nonnegative test function in $\mathcal{C}_0^\infty$ and

$$-\langle \nabla u \otimes \nabla v\rangle = -\begin{pmatrix} \langle u_x, v_x\rangle & \langle u_x, v_y\rangle \\ \langle u_y, v_x\rangle & \langle u_y, v_y\rangle \end{pmatrix}. \tag{6}$$

This definition is weak (in the sense of distributions) but it is also equivalent to a strong one when u is sufficiently regular. We will use a weaker definition where the trial and test functions will be in a (finite dimensionnal) subspace of $\mathcal{C}_0^\infty$: the set of some FE functions. Last we will report on [17] where an even weaker definition is used where the test functions are in a much smaller subspace than u_h. One can easily check that the weak definition of convexity (with $\mathcal{C}_0^\infty$ test functions) implies the weaker definition of convexity (with only the basis functions in P_1 associated to interior points as test functions) but they are not equivalent.

The proof of Lemma 1 is based on the following formula written for the P_1 function u_h:

$$\langle \frac{\partial^2 u_h}{\partial \mathbf{a}\partial \mathbf{b}}, \varphi\rangle = \sum_e (\mathbf{q}_2 - \mathbf{q}_1)\cdot\mathbf{n}\,(\mathbf{n}\cdot\mathbf{a})(\mathbf{n}\cdot\mathbf{b})\int_e \varphi(s)\mathrm{d}s, \tag{7}$$

where the summation is taken over all interior edges e of the mesh, $\mathbf{a}$ and $\mathbf{b}$ are two unit vectors, φ is a C^∞ function with compact support in Ω, $\mathbf{q}_1$ and $\mathbf{q}_2$ designate the two (constant) gradients of u_h in the two triangles that share the edge e and $\mathbf{n}$ is the unit normal from triangle 1 to triangle 2. This formula is right thanks to the property that the gradient of a continuous function across the edge has its tangential derivative along the edge continuous. Namely $(\mathbf{q}_2 - \mathbf{q}_1)\cdot\mathbf{t} = 0$ where $\mathbf{t}$ is the unit tangent vector. The proof also relies on the fact that for all unit vector $\mathbf{n}$, the bilinear form $(\mathbf{a},\mathbf{b}) \mapsto (\mathbf{n}\cdot\mathbf{a})(\mathbf{n}\cdot\mathbf{b})$ is positive semi-definite, whatever $\mathbf{n}$.

Through a proof very similar to the one of Theorem 4 in [1], one may prove a wider Theorem that even applies locally:

Theorem 1 *Let Ω be an open and convex subset of $\mathbb{R}^2$ and $\mathcal{T}_h$ a triangulation of Ω. If there exists an open and convex $\Omega' \subset \Omega$ such that the following property is satisfied in Ω' and for the mesh $\mathcal{T}_h$:*

$$(PM)\left\{\begin{array}{l} \exists(\mathbf{a},\mathbf{b}) \text{ two independent unit vectors such that } (\mathbf{n}\cdot\mathbf{a})(\mathbf{n}\cdot\mathbf{b}) \geq 0 \\ \text{for all } \mathbf{n} \text{ unit normal to an edge of } \mathcal{T}_h \bigcap \Omega', \end{array}\right. \tag{8}$$

then any function u_h convex and P_1 will satisfy the following equation in the sense of distributions on Ω':

$$\frac{\partial^2 u_h}{\partial \mathbf{a} \partial \mathbf{b}} \geq 0. \tag{9}$$

The proof relies on the fact that if a P_1 function u_h is convex, then by Lemma 1, the gradient's jumps are non-negative. As a consequence, in Formula (7), if u_h is convex, the scalar coefficients $(\mathbf{q}_2 - \mathbf{q}_1) \cdot \mathbf{n}$ are all non-negative. But separately, if condition (PM) is satisfied, $(\mathbf{n} \cdot \mathbf{a})(\mathbf{n} \cdot \mathbf{b}) \geq 0$ and so even the non-diagonal terms of the hessian $(\partial^2 u_h / (\partial \mathbf{a} \partial \mathbf{b}))$ are sign-constrained !

Notice first that property (PM) only depends on the geometry of the mesh and not on any function. Notice then that the property (PM) is at given h but, if it remains for a sequence of $h \to 0$ in $\Omega' \subset \Omega$, then the associated functions u_h given by Theorem 1 satisfy (9) even at the limit in $\Omega' \subset \Omega$. Yet such a property (9) is contradictory with the property of convexity. This is the key of the obstruction to the convergence stated in [1]. The next subsection is devoted to discussing the generality of (PM).

2.2 Is property (PM) frequent ?

In this subsection, we start from some elementary observations on particular meshes, then we argue on whether such meshes are frequent and state a Theorem. Three types of meshes may be considered that are depicted in Figure 1.

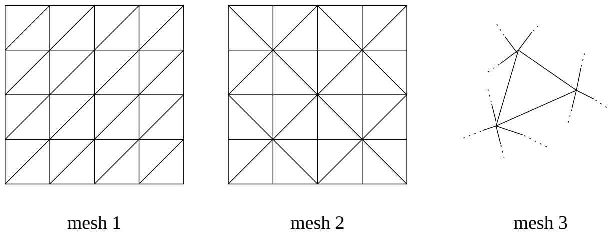

Figure 1: Three meshes.

Concerning mesh 1, three different unit normals (up to a multiplicative factor -1) exist in *all* the domain: $\{(0,1);(1,0);(-1/\sqrt{2},1/\sqrt{2})\}$. If we choose $\mathbf{a} = (-1,0)$ and $\mathbf{b} = (0,1)$, the values taken by $(\mathbf{n} \cdot \mathbf{a})(\mathbf{n} \cdot \mathbf{b})$ are $\{0;0;(1/\sqrt{2})(1/\sqrt{2})\}$. They are all nonnegative.

Concerning mesh 2, four types of unit normals can be found in *all* the domain: $\{(0,1);$ $(1,0);$ $(-1/\sqrt{2},1/\sqrt{2});$ $(1/\sqrt{2},1/\sqrt{2})\}$. If we choose $\mathbf{a} = (1,0)$ and $\mathbf{b} = (1/\sqrt{2},-1/\sqrt{2})$, the values taken by $(\mathbf{n} \cdot \mathbf{a})(\mathbf{n} \cdot \mathbf{b})$ are $\{1/\sqrt{2};0;(-1/\sqrt{2})(-1);0\}$. They are all nonnegative.

Any structured mesh like mesh 1 or 2, if refined while keeping the same structure, will satisfy (PM) with the very same $\mathbf{a}$ and $\mathbf{b}$ for any h. But what can be stated about a more general mesh like mesh 3 ?

We are going to prove the following Theorem.

Theorem 2 *Let Ω be an open and convex set of $\mathbb{R}^2$, $\Omega' \subset \Omega$ an open and convex subset and $\mathcal{T}_h$ a triangulation of Ω.*
For any given h and Ω', there exists $(\mathbf{a},\mathbf{b})$ such that (PM) is true in Ω' and for $\mathcal{T}_h$. Moreover, if the refinement process does not enrich the edges' directions of $\mathcal{T}_{h_n}$ in Ω', then the property (PM) is true for any $\mathcal{T}_{h_n}$ in Ω'.

Proof of Theorem 2
First, one must notice that the condition $(\mathbf{n} \cdot \mathbf{a})(\mathbf{n} \cdot \mathbf{b}) \geq 0$ is invariant through changing $\mathbf{n}$ into $-\mathbf{n}$. So, it does not depend on the choice of the hyperplane's unit normal.

Let us take a very general mesh like mesh 3 for which we may isolate a single triangle. Then, for each of the three unit normals (up to the multiplicative coefficient -1), having $(\mathbf{n} \cdot \mathbf{a})(\mathbf{n} \cdot \mathbf{b}) \geq 0$ is equivalent to having $\mathbf{a}$ and $\mathbf{b}$ in the *same* closed half plane whose normal is $\mathbf{n}$. Equivalently, we can take either $\mathbf{a}$ and $\mathbf{b}$ on the same side as $\mathbf{n}$ or on the opposite side. Gathering the conditions associated to each of the three edges, we are led to choosing both $\mathbf{a}$ and $\mathbf{b}$ in one of the six half cone that intersects none of the three hyperplanes. Such a configuration is depicted on Figure (2). The three unit normals are drawn and called n1, n2, n3 and their associated hyperplanes are called P1, P2, P3. So, one may easily find such a couple $(\mathbf{a}, \mathbf{b})$ for any given triangle.

Even better, whatever might be the finite number of edges for a larger Ω' and a given $\mathcal{T}_h$, one sees obviously that such a choice of $\mathbf{a}$ and $\mathbf{b}$ is easy since there always exists a finite number of hyperplanes. As a consequence, $\mathbf{a}$ and $\mathbf{b}$ can be choosen to be *both* in any of the cones which partition the whole unit circle. So the property (PM) is true even on any *given* and potentially unstructured mesh of a subdomain of Ω.

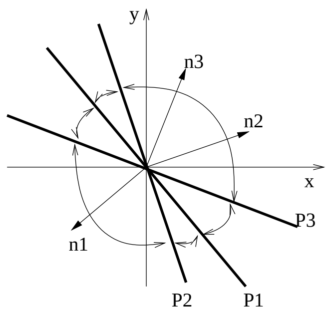

Figure 2: Unit normals and the associated hyperplanes in case of a triangle.

Could a refinement process preserve the property (PM) for a given couple $(\mathbf{a}, \mathbf{b})$, a given Ω' and any $h \to 0$?

In case of any general given mesh $\mathcal{T}_h$ we have just proved that one may find $\mathbf{a}$ and $\mathbf{b}$, should they suit the property only in Ω'. Then, for instance, if the refinement process is such that every triangle in Ω' is refined into four homothetic subtriangles, it is obvious that no more edges' direction will be provided. As a consequence, the very same $\mathbf{a}$ and $\mathbf{b}$ will then suit property (PM) in Ω' for *any* refined mesh and *any* h. The proof is complete.

□

In a sense to be defined, the set of refinement processes that enable (PM) is open and non-empty.

2.3 More information on the lack of convergence

In [1], the authors state that the "conformal method may not converge for some limit function" because the second derivative of the limit is forced to satisfy an unnatural condition. In this subsection, we give a more precise result in L^2 and still use our Lemma 1 that enables to identify the convexity of u_h and the non-negativity of its gradients' jumps (for P_1 FE). This enables us to state our main Theorem:

Theorem 3 *Let Ω be an open and convex domain in $\mathbb{R}^2$, Ω' an open and convex subdomain of Ω and a family of meshes $(\mathcal{T}_h)_{h\to 0^+}$. If the property (PM) is satisfied in Ω' and for the meshes $(\mathcal{T}_h)_{h\to 0}$, then there exists $\varepsilon > 0$ and a $\mathcal{C}^\infty$ convex function u_{exact} such that*

$$\min_{u_h \in P_1 \text{ and } u_h \text{ convex}} \mid u_h - u_{exact} \mid_{L^2} \geq \varepsilon.$$

This Theorem uses that if Property (PM) is satisfied in a domain Ω for a family of meshes $(\mathcal{T}_h)_{h\to 0^+}$, should it be only locally in Ω', there will be two unit vectors $\mathbf{a}, \mathbf{b}$ such that (9) holds. Given those $\mathbf{a}, \mathbf{b}$, there exists a function u_{exact} that may not be the limit in L^2 of *any* sequence of functions both P_1 and convex (in a strong definition) on *this* family of meshes. The improvement of this Theorem with respect to [1] is that the obstruction is local. As a consequence we may apply our obstruction on *any* mesh and not only regular ones.

We need some more preliminary results and definitions before proving this Theorem 3 concerning P_1 FE.

Definition 1 *Let Ω be an open subset of $\mathbb{R}^2$, and $u \in \mathcal{C}^0(\Omega)$. For any $\mathbf{x} = (x, y) \in \Omega$ and $\{\mathbf{a}, \mathbf{b}\}$ two independent unit vectors, there exist two positive numbers $\{\alpha_0, \beta_0\}$ such that*

$$\forall\alpha,\ 0 \leq \alpha \leq \alpha_0, \forall\beta,\ 0 \leq \beta \leq \beta_0,$$
$$\mathbf{x} + \alpha\mathbf{a} + \beta\mathbf{b} = (x + \alpha a_1 + \beta b_1, y + \alpha a_2 + \beta b_2) \in \Omega.$$

For such $(\alpha_0, \beta_0, \mathbf{a}, \mathbf{b})$, one defines:

$$\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u) = \left(u(\mathbf{x} + \alpha_0\mathbf{a} + \beta_0\mathbf{b}) - u(\mathbf{x} + \alpha_0\mathbf{a}) - u(\mathbf{x} + \beta_0\mathbf{b}) + u(\mathbf{x})\right)/(\alpha_0\beta_0).$$

If $\alpha_0 = \beta_0 \to 0^+$, then $\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u) \to \partial^2 u/(\partial\mathbf{a}\partial\mathbf{b})$. Then $\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u)$ may be considered as a kind of double primitive of $\partial^2 u/(\partial\mathbf{a}\partial\mathbf{b})$. We are going to prove that such a quantity $\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u)$ will be overconstrained by the mere discretization and the limit u will not satisfy the right sign of the second derivative.

We want now to define an explicit solution depending on some parameters. A well-chosen combination of these parameters will provide a non-approximable function. For instance, let $\Omega = (1, 2)^2$ and the problem (2, 3) for a positive definite real matrix $\mathbf{C}$ be such that:

$$\mathbf{C} = \begin{pmatrix} \mu_2 & \rho \\ \rho & \mu_1 \end{pmatrix}. \tag{10}$$

Then there exists an exact solution thanks to (4):

$$u_{exact} = \frac{1}{\mu_1\mu_2 - \rho^2}\left(\mu_1(x^2 - 1)/2 + \mu_2(y^2 - 1)/2 - \rho(xy - 1)\right) \text{ in } \Omega. \tag{11}$$

The function u_{exact} is such that it is zero at the corner $(1, 1)$ of the domain Ω chosen but it can easily be generalized to other domains. Simple computations prove the following formula:

$$\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact}) = \frac{1}{\mu_1\mu_2 - \rho^2}\left(\mu_1 a_1 b_1 + \mu_2 a_2 b_2 - \rho(a_1 b_2 + a_2 b_1)\right) = \mathbf{a}'\mathbf{C}^{-1}\mathbf{b}, \tag{12}$$

where a_1, a_2, b_1, b_2 are the components of $\mathbf{a}, \mathbf{b}$. It is then useful to state the following Lemma.

Lemma 2 *Let* $\mathbf{a}, \mathbf{b}$ *be two given independent unit vectors in* $\mathbb{R}^2$. *Let* $\Omega \subset \mathbb{R}^2$ *and* $\eta > 0$ *given. Then, there exists* (μ_1, μ_2, ρ) *such that* $\mu_1 \geq 0, \mu_2 \geq 0$, $\mu_1\mu_2 - \rho^2 \geq 0$ *and*

$$\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact}) = \mathbf{a}^T\mathbf{C}^{-1}\mathbf{b} \leq -\eta \text{ in } \Omega, \tag{13}$$

provided $(\alpha_0, \beta_0) \in \mathbb{R}^{+*2}$ *are such that there exists* $\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact})$.

Roughly speaking, Lemma 2 states that for any independent $\mathbf{a}, \mathbf{b}$, one may find a positive definite matrix $\mathbf{C}^{-1}$ such that the associated u_{exact} satisfies

$$\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact}) = \mathbf{a}^T\mathbf{C}^{-1}\mathbf{b} < 0.$$

Proof of Lemma 2
From the formula (12), one sees that it is sufficient to find a positive definite symmetric bilinear form ($\mathbf{C}^{-1}$) such that, for independent $\mathbf{a}, \mathbf{b}$ given, $\mathbf{a}^T\mathbf{C}^{-1}\mathbf{b} < 0$. We will build up $\mathbf{C}^{-1}$ from its eigenvalues and eigenvectors.

Let $\mathbf{e}_1$ be a unit vector between $\mathbf{a}$ and $\mathbf{b}$ (normalized mean for instance). Then let $\mathbf{e}_2$ be a unit vector normal to $\mathbf{e}_1$. With the same notations for $\mathbf{a}$ and $\mathbf{b}$ in this new basis as above, $a_1b_1 > 0$ and $a_2b_2 < 0$.

Let now $\mathbf{C}^{-1}$ be a matrix in the canonical basis with the eigenvectors $\mathbf{e}_1, \mathbf{e}_2$ and the associated positive eigenvalues λ_1, λ_2. Obviously, $\mathbf{C}$ is positive semi-definite. Then $\mathbf{a}^T\mathbf{C}^{-1}\mathbf{b} = \lambda_1 a_1 b_1 + \lambda_2 a_2 b_2$ which may be less than $-\eta$ given, for an appropriate choice of λ_1, λ_2. Then the (μ_1, μ_2, ρ) are the coefficients of the matrix $\mathbf{C}$ in the canonical basis as written in (10).

□

We have now obtained all what is needed to start the following proof.

Proof of Theorem 3.
Since we assume (PM) is satisfied in Ω' for $(\mathcal{T}_h)_{h\to 0}$ and $(\mathbf{a}, \mathbf{b})$, Theorem 1 enables us to claim that any P_1 function u_h satisfies (9) in the sense of distributions in Ω':

$$0 \ \leq \ \langle \frac{\partial^2 u_h}{\partial \mathbf{a}\partial \mathbf{b}}, \Psi\rangle_\Omega = \langle u_h, \frac{\partial^2 \Psi}{\partial \mathbf{a}\partial \mathbf{b}}\rangle_\Omega, \tag{14}$$

for any nonnegative $\Psi \in \mathcal{C}_0^\infty(\Omega')$. This property obviously remains for open and convex subdomains of Ω'.

Should it be needed, one could decrease $\alpha_0, \beta_0 > 0$ and find a subdomain $\Omega'' \subset \Omega'$ such that for any nonnegative $\varphi \in \mathcal{C}_0^\infty(\Omega'')$, the function

$$\Psi: \ \mathbf{x} \mapsto \int_0^1 \int_0^1 \varphi(\mathbf{x} - t\alpha_0\mathbf{a} - t'\beta_0\mathbf{b})\, \mathrm{d}t\, \mathrm{d}t', \tag{15}$$

is nonnegative and in $\mathcal{C}_0^\infty(\Omega')$. So the function Ψ is eligible for equation (14). The second derivative of Ψ may be computed:

$$\begin{aligned}\frac{\partial^2 \Psi}{\partial \mathbf{a}\partial \mathbf{b}} &= \left(\varphi(\mathbf{x} - \alpha_0\mathbf{a} - \beta_0\mathbf{b}) - \varphi(\mathbf{x} - \beta_0\mathbf{b}) - \varphi(\mathbf{x} - \alpha_0\mathbf{a}) + \varphi(\mathbf{x})\right)/(\alpha_0\beta_0)\\ &= \phi_{(\alpha_0,\beta_0,-\mathbf{a},-\mathbf{b})}(\varphi).\end{aligned}$$

So we have, for any nonnegative $\varphi \in \mathcal{C}_0^\infty(\Omega'')$:

$$\begin{aligned}0 \leq \langle \frac{\partial^2 u_h}{\partial \mathbf{a}\partial \mathbf{b}}, \Psi\rangle_\Omega &= \langle u_h, \frac{\partial^2 \Psi}{\partial \mathbf{a}\partial \mathbf{b}}\rangle_\Omega\\ &= \langle u_h, \phi_{(\alpha_0,\beta_0,-\mathbf{a},-\mathbf{b})}(\varphi)\rangle_{\Omega'}\\ &= \langle \phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_h), \varphi\rangle_{\Omega'}.\end{aligned} \tag{16}$$

Separately, given $\mathbf{a}, \mathbf{b}, \eta > 0$, Lemma 2 enables us to claim there exists a convex quadratic function u_{exact} such that (13) is true for $\eta > 0$ given. Moreover, because of (16), since the open set $\Omega'' \subset \Omega'$ is of non-zero measure, and $\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact})$ is a constant known by (12), we have

$$\langle \phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_h - u_{exact}), \varphi \rangle_{\Omega''} \geq -\phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_{exact}) \int_\Omega \varphi.$$

As a consequence, for convenient Ω'', we are led to

$$\langle \phi_{(\alpha_0,\beta_0,\mathbf{a},\mathbf{b})}(u_h - u_{exact}), \varphi \rangle_{\Omega''} \geq \eta \int_\Omega \varphi, \tag{17}$$

for any *nonnegative* $\varphi \in \mathcal{C}_0^\infty(\Omega'')$ and u_h convex and P_1. This last inequality contradicts any possible convergence in $L^2(\Omega'')$ of a sequence u_h of convex functions P_1 in $(\mathcal{T}_h)_h$ to u_{exact} given by Lemma 2.

□

The previous Theorem adds one more argument to the need of convenient numerical methods to discretize the constraint of convexity.

3 The P_1 FEM

The non-convergence of the convexity problem (2, 3) discretized by conformal P_1 Finite Elements is proved and numerically illustrated in [1]. It is even proved to apply to non-regular meshes in the subsection 2.3. We give here a very different argument by studying the consistency of various discretizations of convexity.

One could wonder why study consistency. The reason is that when one discretizes convexity (or even its linear form of subharmonicity), there exist meshes and FE on which the discretization is not consistent. As a consequence, any proof of convergence must exclude at least these cases. The fact that the FE-discretized Dirichlet functional and the FE-discretized Hessian matrix converge separately is not sufficient. Already [21] claimed convergence, but was later proved to be wrong in [1].

3.1 Use of the gradients' jumps for convexity

Given a sufficiently smooth function u, one may interpolate it in P_1 FE as $u_h = \sum_{i=1}^N u_{hi}\phi_i(x)$. If the mesh is structured as in Figure 3, should it be local, one may compute the gradients' jumps accross the edges as functions of the values at the various involved nodes. Then one may compute the series expansion of the sampled values u_i from the exact initial function u. The jumps between triangles 1 and 2, 2 and 3 and last 3 and 4 are respectively:

$$\begin{array}{lll} \text{Jump}(1/2) = & \frac{+u_6-u_7-u_1+u_2}{h} & = (u_{xx}+u_{xy})_{x=x_1,y=y_1}h + O(h^2), \\ \text{Jump}(2/3) = & \frac{-u_1+u_7-u_2+u_3}{h} & = (u_{xy}+u_{yy})_{x=x_1,y=y_1}h + O(h^2), \\ \text{Jump}(3/4) = & \sqrt{2}\frac{-u_1+u_2-u_3+u_4}{h} & = -\sqrt{2}(u_{xy})_{x=x_1,y=y_1}h + O(h^2), \end{array} \tag{18}$$

where we denote u_x the partial derivative with respect to x and u_{xx} the second order derivative with respect to x. When $h \to 0$, instead of forcing the solution of a problem to be convex, we force it to some mesh-dependent combination of its second order derivatives to have a sign or another. In addition, this combination is meaningless since for the convex limit function, it may have whatever sign. We will say that such a discretization is not consistent.

Below, we investigate the consistency of various other discretizations.

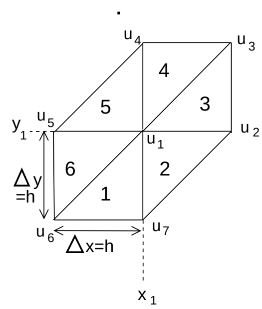

Figure 3: Local shape of the mesh close to (x_1, y_1).

3.2 Use of a weak P_1 version

We use here a weak P_1 definition of convexity which is similar to the one of [17], except that their trial basis and test basis functions are different. We use the same trial and test P_1 basis. Their method is fully discussed and illustrated in Section 5.

3.2.1 The subharmonicity constraint

Subharmonicity is an interesting property simpler than convexity since it is only linear. A weak definition of subharmonicity ($\Delta u \geq 0$) is, for any test function ϕ_i in the discrete basis ($\otimes$ is defined in (6)):

$$\langle \Delta u_h, \phi_i \rangle = \mathrm{Tr}\ \langle \mathbf{D}^2 u_h, \phi_i \rangle = -\mathrm{Tr}\ \langle \nabla u_h \otimes \nabla \phi_i \rangle \geq 0. \tag{19}$$

One may then prove the following Proposition:

Proposition 3.1 *The weak P_1 discretization of the subharmonicity constraint on a mesh like in Figure 3 is consistent:*

$$\langle \Delta u_h, \phi_i \rangle = -4u_1 + u_2 + u_5 + u_7 + u_4 = (u_{xx} + u_{yy})_{x=x_1, y=y_1} h^2 + O(h^3), \tag{20}$$

where i is the node at the center of the cell's group (x_1, y_1).

The proof is very easy and left to the reader. Indeed, it is only the discretization of the Laplacian which is known to be consistent and even convergent !

In order to test this discretization, we used the Matlab package `optim` to minimize the functional $\int_\Omega |\nabla u|^2/2 + \nabla u_{exact} \cdot \nabla u$, where $f = \Delta u_{exact}$, over the set of subharmonic functions. The initial condition is $x(x-1)y(y-1)$. The exact solution of this H_0^1 projection is $u_{exact} = x^2y^2 - (x^4 + y^4)/6$ and its Laplacian vanishes. The convergence with the procedure `quadprog` of quadratic programming can be seen on Figure 4. It is quite satisfactory.

3.2.2 The convexity constraint

One may give a weak definition of convexity:

$$\mathrm{Tr}\ \langle \mathbf{D}^2 u_h, \phi_i \rangle \geq 0 \text{ and } \det \langle \mathbf{D}^2 u_h, \phi_i \rangle \geq 0, \tag{21}$$

for any ϕ_i, basis function of the FEM at the node i. One may then prove the following Proposition:

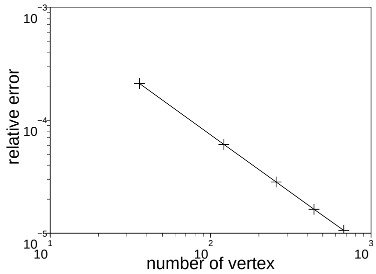

Figure 4: convergence of P_1 FE in case of a subharmonicity constraint.

Proposition 3.2 *The weak P_1 discretization of the convexity constraint (21) on a mesh as in Figure 3 is consistent as can be seen from (20) and ($\otimes$ is defined in (6)):*

$$\begin{aligned} \det\langle \mathbf{D}^2 u_h, \phi_i\rangle &= \det\langle \nabla u_h \otimes \nabla \phi_i\rangle = \\ &= (-2u_1+u_2+u_5)(u_7+u_4-2u_1) - (-u_7+u_3-u_4+u_6-u_5-u_2+2u_1)^2/4 \\ &= (u_{xx}u_{yy}-u_{xy}u_{yx})_{x=x_1,y=y_1} h^4 + O(h^5). \end{aligned} \tag{22}$$

where i is the node at the center of the cell's group (point of coordinates (x_1, y_1).

This discretization (21) for the convexity contraint is consistent although the numerical treatment of the linear and nonlinear constraints should raise inaccuracies since they are of very different orders of magnitude. A very natural question would be to investigate its convergence. Such a study is postponed to an other article.

4 THE P_2 FEM

In the present section, we investigate the consistency of various discretizations through P_2 FEM of two second order derivative constraints: subharmonicity and convexity.

First, we interpolate a continuous function u to a P_2 function $u_h = \sum_i u_i \phi_i(x)$ in a domain Ω meshed with triangles as in Figure 5. Here, the index i denotes both vertices and edge midpoints. Then we compute the discretized version of the second order term constrained to be nonnegative (various versions are treated) and compute its series expansion. If forcing it to be nonnegative amounts to forcing the continuous limit function to the correct constraint, then we claim the discretization is consistent. Otherwise it is inconsistent.

For the whole section, we assume the mesh is (at least locally) structured around the point of coordinates (x_1, y_1). The local numbering of triangles is depicted in Figure 5. The node (x_1, y_1) is localy numbered 1 in every triangle and the local numbering of nodes is depicted in triangle 3 of Figure 5.

4.1 Gradients' jumps for the convexity constraint

In a way similar to the P_1 case, one may prove for P_2 functions u_h:

$$\langle \frac{\partial^2 u_h}{\partial \mathbf{a} \partial \mathbf{b}}, \varphi\rangle = \sum_e \int_e (\mathbf{q}_2(s) - \mathbf{q}_1(s)) \cdot \mathbf{n}\, \varphi(s) \mathrm{d}s\, (\mathbf{n}\cdot\mathbf{a})(\mathbf{n}\cdot\mathbf{b}) + \sum_K \frac{\partial^2 u_h}{\partial \mathbf{a}\partial\mathbf{b}}|_K \int_K \varphi,$$

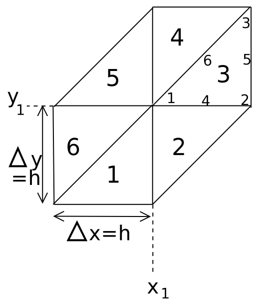

Figure 5: Local shape of the mesh close to (x_1, y_1). Local numbering in triangle 3.

for any e interior edge of the mesh, $\mathbf{a}$ and $\mathbf{b}$ are two unit normal vectors, φ is a C_0^∞ function with compact support in Ω. By taking φ localized along the edge e, one may state that this definition of convexity (with $\mathcal{C}_0^\infty$ test functions) implies the non-negativity of the gradients' jumps.

In order to test the consistency of such a discretization, we compute the gradients' jumps accross the edges common to triangles 1 and 2, 2 and 3 and last 3 and 4. Of course, they are not constant since they are P_1 FE. After an exact computation and a series expansion (details left to the reader), one may state the following Proposition.

Proposition 4.1 *The discretization of the convexity constraint with the jump of the gradients between triangles* 1 *and* 2*,* 2 *and* 3*,* 3 *and* 4 *of a* P_2 *function on a mesh like in Figure 5 gives terms*

$$\begin{aligned} \textit{Jump(1/2)} &= (u_{xxy}+u_{xyy})_{x=x_1,y=y_1}(y-y_1+h/2)h/2+O(h^3);\ y\in[y_1-h,y_1] \\ \textit{Jump(2/3)} &= (u_{xxy}+u_{xyy})_{x=x_1,y=y_1}(x-x_1-h/2)h/2+O(h^3);\ x\in[x_1,x_1+h] \\ \textit{Jump(3/4)} &= -(u_{xxy}+u_{xyy})_{x=x_1,y=y_1}(x-x_1-h/2)\sqrt{2}h/2+O(h^3);\ x\in[x_1,x_1+h]. \end{aligned} \tag{23}$$

Such a discretization is non-consistent.

Since for instance between triangles 1 and 2, $(y-y_1+h/2)$ changes sign (but remains $O(h)$), one deduces from (23) that forcing the non-negativity of the gradients' jump along the edge forces the limit function to satisfy the non-natural equality condition $u_{xxy}+u_{xyy}=0$!

So the weak $\mathcal{C}_0^\infty$ (or strong !) definition of convexity implies the non-negativity of gradient's jumps which is non-consistent and so must be rejected. We check below that the weak P_2 definition may (subsection 4.3) or maynot be consistent (subsection 4.2).

4.2 Weak P_2 version of the second derivative at a vertex

We define the weak $P_2 - P_2$ (*i.e.* with P_2 trial and test functions) version of the Hessian matrix at vertex i as:

$$\langle \mathbf{D}^2 u_h, \phi_i\rangle = -\langle \nabla u_h \otimes \nabla \phi_i\rangle, \tag{24}$$

where $\otimes$ is defined in (6). Strictly speaking, this may not provide a correct weak version of non-negativity since ϕ_i changes sign. Anyway, one may state the following Proposition which proof is left to the reader:

Proposition 4.2 *The discretization of the linear part of the convexity constraint according to (24) of a P_2 function on a mesh as in Figure 5 ($h = \Delta x = \Delta y$) and a P_2 test function centered at the vertex (x_1, y_1) gives:*

$$Tr\ (\langle \mathbf{D}^2 u_h, \phi_i \rangle) = -(u_{xxxx} + u_{yyyy})_{x=x_1, y=y_1} * h^4/48 + O(h^5). \quad (25)$$

Such a discretization of an inequality constraint is non-consistent.

It appears that such a discretization is not even consistent for the linear part of the constraint. More precisely, while one could believe one forces the solution to be subharmonic, indeed, one forces it to be such that $u_{xxxx} + u_{yyyy} \leq 0$! The full nonlinear convexity constraint on the same vertices may only work worse.

One must notice that the fourth order of the expansion in (25) is meaningful. Indeed, on the one hand the three midpoints quadrature is exact for P_2 functions in a triangle. On the other hand the basis function for a vertex vanishes on these midpoints. As a result, the order two term is identically zero. So the first non-zero term is the fourth one and the constraint is of fourth order too.

4.3 Weak P_2 version of the second derivative at an edge midpoint

We define the weak P_2 version of the second derivative at an edge midpoint (denoted by index j) in a way similar to (24). Indeed, the function ϕ_i in (24) is replaced by the P_2 basis function ϕ_j associated to an edge midpoint indexed by j.

Unlike in subsection 3.2, the basis functions (of the edge midpoints) are non-negative. So it is *a priori* an admissible weak formulation of semi-definite positiveness.

4.3.1 The subharmonicity constraint

Let us assume we discretize the constraint $\Delta u \geq 0$ by

$$\int_\Omega \nabla u_h \cdot \nabla \phi_j \leq 0, \quad (26)$$

for all j index of an edge midpoint interior to Ω. One may then state a Proposition (which proof is left to the reader):

Proposition 4.3 *The discretization of the linear part of the convexity constraint according to (26) on a mesh as in Figure 5 ($h = \Delta x = \Delta y$) gives the same series expansion, whether the edge is vertical, horizontal or diagonal:*

$$Tr\ (\langle \mathbf{D}^2 u_h, \phi_j \rangle) = (\Delta u)_{x=x_1, y=y_1} h^2/3 + O(h^3), \quad (27)$$

for any j index of an edge midpoint interior to Ω. Such a discretization of an inequality constraint is consistent.

Various questions remain. If we discretize the subharmonicity (or the convexity) constraint only at edge midpoints, is it enough constraints or not ? More generaly, the amount of constraints compared to the amount of degrees of freedom (dof) should be discussed. The relevance of this question seems to be the conclusion of most recent articles like [20, 19, 17].

4.3.2 The convexity constraint

Like in the subharmonic case, we take a weak P_2 version of the continuous nonlinear constraint det $D^2u \geq 0$ with (nonnegative) test functions associated to every edge midpoint j in the interior. One may then easily prove the following Proposition:

Proposition 4.4 *The discretization of the nonlinear part of the convexity constraint on a mesh as in Figure 5 ($h = \Delta x = \Delta y$) gives:*

$$\det(\langle \mathbf{D}^2 u_h, \phi_j \rangle) = (u_{xx}u_{yy} - u_{xy}u_{yx})_{x=x_1, y=y_1} h^4/9 + O(h^5), \quad (28)$$

for any j index of an edge midpoint interior to Ω. Such a discretization of an inequality constraint is consistent.

So the convergence order of the linear constraint is two while the nonlinear one's is four. Such a discrepancy between the orders of convergence of those two constraints should be managed in numerical simulation using both constraints.

5 Discussion on the Aguilera and Morin's articles

Let us notice that [1] seems not to have been known of the authors of [16, 17].

Roughly speaking, the first article [16] proves convergence of the FD discretization of problems like ours. The second [17] proves convergence of the FE discretization of the same problems but the authors imagine the clever trick of using different basis for the trial and the test functions.

5.1 Finite Differences

In [16], Aguilera and Morin deal with FD and "prove convergence under very general conditions, even when the continuous solution is not smooth". Their proof relies on approximation theory and on a weak definition of convexity, namely the FD approximation of the Hessian matrix is forced to be positive. They stress that their FD-convexity is not equivalent to a continuous convexity. As a consequence, their approximation is *not conformal*. So there is no contradiction with [1]. In addition, one must notice that the number of constraints is roughly twice the number of interior points. It is large, but grows only linearly with the degrees of freedom. They also provide numerical experiments.

Unfortunately the numerical experiments are not very convincing. Concerning the monopolist problem in 2D, they claim that "the error in the L^∞ norm is smaller or approximately equal to h" (p. 27). Yet, their Table 1 provides errors that we draw in a loglog scale in the left part of Figure 6. The order of convergence is not clear.

For the monopolist problem in 3D, the authors notice that "the L^∞ error is not converging to zero with order $O(h)$" (p. 29). Since they only prove convergence and do not forecast any order, there is no contradiction. But when their Table 2 is drawn in a loglog scale (see Figure 6 right), convergence seems not to be reached yet. Moreover the execution time grows faster than polynomialy both for 2D and 3D monopolist.

As reminded by the authors, FD discretization gives the very same matrices as the P_1 FE (both for trial and test functions), on a mesh like our mesh 1 on Figure 1 (except boundary conditions irrelevant here). In this sense, one may claim that their FD discretization is somehow equivalent to a P_1 (for trial) - P_1 (for test) discretization which was proved to be non-consistent. Indeed, the next article [17] gives one example (Example 3.7 p. 3150) of P_1 discretization and a weak definition of convexity (with P_1 trial and test functions, so alike FD discretization), where " (u_h) does not converge to u as $h \to 0$" (p. 3151 and Figure 3.1).

Discussing the way these results on FD are coherent with the results on FE is postponed to a forthcoming article.

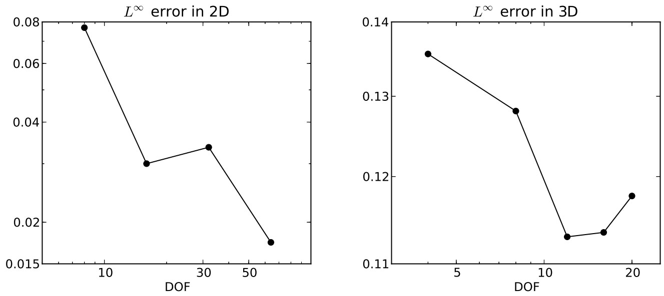

Figure 6: The monopolist numerical solution error versus DOF

5.2 Finite Elements

In [17], Aguilera and Morin prove convergence of the discretization of the full problem: not only the approximation of the Hessian, but also of the functional together.

Especially, they define u_h, interpolation of u on a given mesh with respect to the trial functions basis $\{\phi_r^h\}_r$ as $u_h(x) = \sum_{r\in I_{trial}^h} u_{hr}\phi_r^h(x)$. They also use a (possibly) different FE basis for the test functions $\{\varphi_s^h\}_{s\in I_{test}^h}$. Then, the discrete Hessian matrix is defined weakly by $H_s^h u_h = -\left(\langle \partial_i u_h, \partial_j \varphi_s^h\rangle\right)_{i,j}$ and FE convexity is defined as:

$$H_s^h u_h = -\left(\sum_{r\in I_{trial}^h} u_{hr}\langle \partial_i \phi_r^h, \partial_j \varphi_s^h\rangle\right)_{i,j} \succeq 0,$$

for all φ_s^h in the test functions basis, where they denote $\mathbf{A} \succeq 0$ if $\mathbf{A}$ is positive semi-definite. Notice that the matrix $(\langle \partial_i \phi_r^h, \partial_j \varphi_s^h\rangle)_{r,s}$ is a rectangle if the trial and test basis do not have the same number of functions !

Then their main result states convergence of the discretization once one assumes at least (p. 3147):

C.2 There exists a linear operator $\mathcal{I}^h$ with values in the discretization space V_h (the interpolant), an integer $m \geq 2$, and a constant C independent of u and h such that

$$\| u - \mathcal{I}^h u \|_{H^1(\Omega)} \leq Ch^{m-1} \| u \|_{H^m(\Omega)},$$

C.3 The basis test-functions are such that

$$\varphi_s^h(x) \geq 0 \text{ for all } x \in \Omega, \text{ all } h > 0, \text{and all } s.$$

The condition C.3 merely states that the test functions, that are supposed to approximate positive $\mathcal{C}_0^\infty$ functions, remain nonnegative. So the basis functions are at most P_1 or a subset of P_k.

The condition C.2 is very classical and states that the trial space is P_{m-1}. In addition, the proof requires $m > 2$. In other words, the trial functions must at least be P_2.

This is seen by the authors who say they "need to elaborate" on the condition $m > 2$. On their p. 3150, they notice the FE discretized Hessian ($P_1 - P_1$ and so $m = 2$) expands into the FD scheme (for one specific mesh !) at first order and so they refer to their FD article to conclude that in the case of some specific meshes, their Theorem "also holds for $m = 2$" (or $P_1 - P_1$ FE).

Yet, the very next example they provide is $P_1 - P_1$ and they "report numerical evidence supporting the necessity of assuming $m > 2$ in [their] Theorem 3.6" (p. 3150). They stress that "since u is convex, the projection u_h should converge to u as $h \to 0$, but this is not the case in this example" (p. 3150) and "(u_h) does not converge to u as $h \to 0$" (p. 3151). They summarize this result by saying "although there is some sort of super-convergence for some meshes, for general meshes [...] FE-convex piecewise *linear* [m=2!] function may not suffice". Discussing how this numerical example articulates with [16] is postponed to a forthcoming article.

All this is even complexified when the authors cope with numerical experiments of $P_2 - P_2$ FE discretization. As they notice, the FE test functions in the basis are not all nonnegative. So they "considered the usual piecewise *linear* nodal basis for the vertices" and "the usual quadratic bubbles" for the edges' midpoints (p. 3152).

Their table of convergence (Table 5.1) is transformed into loglog graphic in Figure 7. The last simulation with *adaptive* refinement gives a good point, but the uniform-refinement simulations give increasing errors. Again convergence is not obvious. So as to make this convergence obvious, we illustrate their method in Section 5.4.

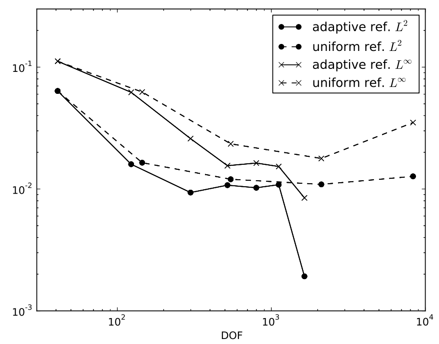

Figure 7: Error according to the refinement either automatic (solid lines) or uniform (dashed lines) and the norm (L^2 or L^∞) versus Degrees of Freedom (DOF)

5.3 Are [16, 17] contradictory with [1] ?

The answer is no. But it deserves to be explained.

Both [17] and [1] use a dual definition of convexity and have conclusions that could seem to be contradictory. The only difference is that [17] uses different basis for trial and test functions:

$$-\langle \nabla u_h \otimes \nabla \varphi_s^h \rangle \succeq 0,$$

while [1] uses the same basis for trial and test functions, both in P_1:

$$-\langle \nabla u_h \otimes \nabla \phi_h \rangle \succeq 0.$$

Both discretizations are weak in a sense, but the one of [17] has less test functions than trial functions and so less contraints than the number of Degrees Of Freedom (DOF). Indeed $m > 2$ means that the trial basis is rather large (at least P_2) and condition C.3 that the test basis is at most P_1. The matrix of constraints is rectangle and there are less constraints than the unknowns whereas the discretization of [1] uses the same basis for trial and test and so involves square matrices.

It is then not surprising that the problem discretized with the weak $P_1 - P_1$ definition in [1] is overconstrained : there are as many constraints as DOF ! Indeed it appears also from the literature that the amount of constraints must be not too large. The article [20] defines a stencil of constraints that is coarser than the grid and [19] defines a subgrid of points on which the constraint is applied. Then it seems coherent that the FE proof of [17] requires *more* DOF than the number of constraints.

Our Proposition 4.2 proved in subsection 3.2 states that if we discretize u_h with P_2 FE and use a P_2 set of test functions (but then some test functions change sign), then the discretization of the linear part of the constraint is non-consistent. It proves that the proof of convergence of FEM by [17], which fails if the test functions change sign (as in P_2 and higher), maynot be improved on that point.

If we use the consistency test for some discretizations, we can compute the first non-zero term in the expansion of

$$-\langle \nabla u_h \otimes \nabla \varphi_s^h \rangle,$$

for $P_1 - P_1$ discretizations and report these computations in Table 1 (lower part) on various meshes depicted in Table 1 (upper part). It appears that some discretizations are not consistent with convexity (two on the left) while others are (two on the right).

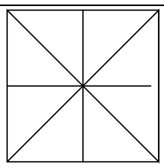	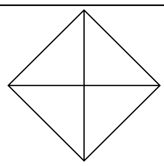	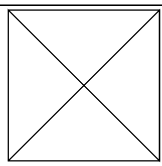	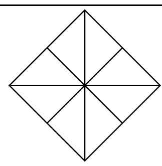
$h^2 \begin{pmatrix} u_{xx} & 2u_{xy} \\ 2u_{xy} & u_{yy} \end{pmatrix}$	$h^2 \begin{pmatrix} u_{xx} & 0 \\ 0 & u_{yy} \end{pmatrix}$	$4h^2/3 \begin{pmatrix} u_{xx} & u_{xy} \\ u_{xy} & u_{yy} \end{pmatrix}$	$2h^2/3 \begin{pmatrix} u_{xx} & u_{xy} \\ u_{xy} & u_{yy} \end{pmatrix}$

Table 1: Consistency for various meshes ($P_1 - P_1$)

5.4 Does the $P2 - P_1$ FEM converge ?

The convergence is proved in [17] for P_k ($k \geq 2$) trial functions and P_1 test functions with additionnal assumptions. In order to ensure the numerical convergence, we write a Python code to mesh a square $\Omega = [1, 2]^2$, discretize the functional (2), and discretize the nonlinear constraints of inequality (convexity). Then we use the `minimize` function in `scipy` to minimize the functional. The execution time are meaningless since they rely on the level of precision we ask and we only want to illustrate convergence.

In a very first step, we look for the solution to (2,3,10), which exact solution is given in (11). It is not surprising that, *whatever the mesh*, we can get a relative error of 10^{-12}. It only justifies that a quadratic function may be approximated by a P_2 numerical code.

In order to justify convergence, one needs a more complex convex function to find.

5.4.1 Convergence to an exact solution

We choose

$$u_{exact}(x) = \exp((x_1 + x_2)/2) - e,$$

and take the functional

$$\int_\Omega \mid \nabla u \mid^2 + \Delta u_{exact} u \, \mathrm{d}x - \int_{\partial\Omega} \frac{\partial u_{exact}}{\partial n} u,$$

which is also, up to a constant, $\int_\Omega \mid \nabla(u - u_{exact}) \mid^2$. The function is constrained to be convex and be such that $u(1,1) = 0$, which makes sense since the function is convex and so regular enough for evaluating at $(1,1)$. The solution is convex, but we are going to prove that its P_2 interpolate is *not* convex.

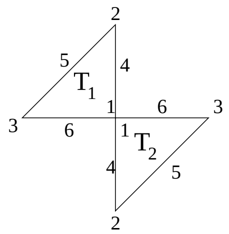

Figure 8: Local numerotation

The geometrical argument is that we P_2-interpolate a function of $x_1 + x_2$ on a mesh like mesh 1 in Figure 1. Let us take the P_2 interpolate of u_{exact} on the two triangles depicted in Figure 8, close to the point (x_1, y_1) at the center as before. Let us prove now that the P_2 interpolate $u_h^{P_2} = \Pi_{P_2} u \in P_2$ is strictly convex inside T_2. Along the segment between the point numbered 1 in T_2 and the point 5 in T_2, one has:

$$\begin{aligned} u_h^{P_2}(x_1 + \alpha h, y_1 - \alpha h) = \; & 2u_1^{T_2}(-2\alpha + 1)(-2\alpha + 1/2) + 2u_2^{T_2}\alpha(-\alpha + 1/2) \\ & +2u_3^{T_2}\alpha(\alpha - 1/2) + 4u_4^{T_2}\alpha(-2\alpha + 1) \\ & +4u_5^{T_2}\alpha^2 + 4u_6^{T_2}\alpha(-2\alpha + 1) \\ = \; & e^{(x_1+y_1)/2}\big(\alpha^2(12 + 2e^{-h/2} + 2e^{h/2} - 8e^{-h/4} - 8e^{h/4}) \\ & +\alpha(-6 - e^{-h/2} - e^{h/2} + 4e^{-h/4} + 4e^{h/4}) + 1\big), \end{aligned}$$

where $u_i^{T_2}$ is the value of the interpolate at the DOF locally numbered i in triangle T_2. These values are the same as those of u_{exact} by definition of the P_2 interpolate and enable to write the last equality. It is then easy to check that this function is strictly convex for $\alpha \in (0, 1/2)$. The same computations in the triangle T_1 gives for $\alpha \in (0, 1/2)$

$$\begin{aligned} u_h^{P_2}(x_1 - \alpha h, y_1 + \alpha h) = \; & e^{(x_1+y_1)/2}\big(\alpha^2(12 + 2e^{-h/2} + 2e^{h/2} - 8e^{-h/4} - 8e^{h/4}) \\ & +\alpha(-6 - e^{-h/2} - e^{h/2} + 4e^{-h/4} + 4e^{h/4}) + 1\big). \end{aligned}$$

Since both $u_h^{P_2,T_2}$ and $u_h^{P_2,T_1}$ are strictly convex in their domains and have the same value on three aligned points, the function $u_h^{P_2}$ maynot be convex at least on the segment between the points numbered 5 in T_2 and the point numbered 5 in T_1 (see Figure 8).

Despite the non-convexity of the P_2 interpolate of the exact solution (which is convex), the convergence of the FEM method proposed and proved in [17] is clearly illustrated in Figure 9 where structured and unstructured meshes are used.

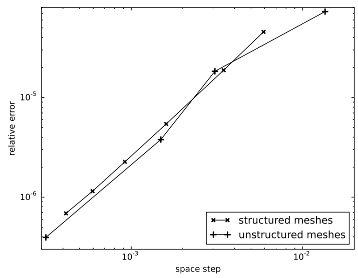

Figure 9: Convergence of the FEM $P_2 - P_1$ method

The order of convergence seems to be the same on structured and unstructured meshes. It is roughly 1.6.

5.4.2 Can this method be used ?

In order to use this method justified theoretically (in [17]) and numerically (above), we look for a solution to the problem of the monopolist in $\Omega = [1,2]^2$:

$$\min_{\substack{u \text{ convex} \\ u \geq 0}} \int_\Omega |\nabla u|^2 - x \cdot \nabla u + u \, \mathrm{d}x.$$

There exists no explicit solution to this problem but some properties of the solution are known (see [10]).

The computed P_2 function on a structured mesh 25×25, its gradient and its second eigenvalue look similar to the ones computed in [20]. The graph of the full P_2 function would be difficult to understand because of the number of DOF. That is why we only use the P_1 information from u_h in Figure 10 (left).

So as to display the gradient, one may compute the gradient $\int_\Omega \nabla u_h \phi_i$ for i associated to an interior point, with a P_2 function u_h, and $\phi_i \in P_2$. But since the quadrature uses the three mid-points, it is right only up to P_2 functions, and so the integral being over a P_3 function is not exact. Yet, since the $P_1 - P_2$ discretization, which is exact, gives the same graphical result with less points, we consider the Figure to be acceptable. The result may be seen in Figure 10 (right).

In Figure 11, one may see the second eigenvalue of the Hessian, computed as a P_2 function tested on P_1 functions and its level sets on the right. The first eigenvalue is not significant because of numerical artifacts. They are computed as the two roots per interior vertex of:

$$-\begin{pmatrix} \int_\Omega \partial u_h/\partial x \, \partial\phi_i/\partial x & \partial u_h/\partial x \, \partial\phi_i/\partial y \\ \int_\Omega \partial u_h/\partial y \, \partial\phi_i/\partial x & \partial u_h/\partial y \, \partial\phi_i/\partial y \end{pmatrix}.$$

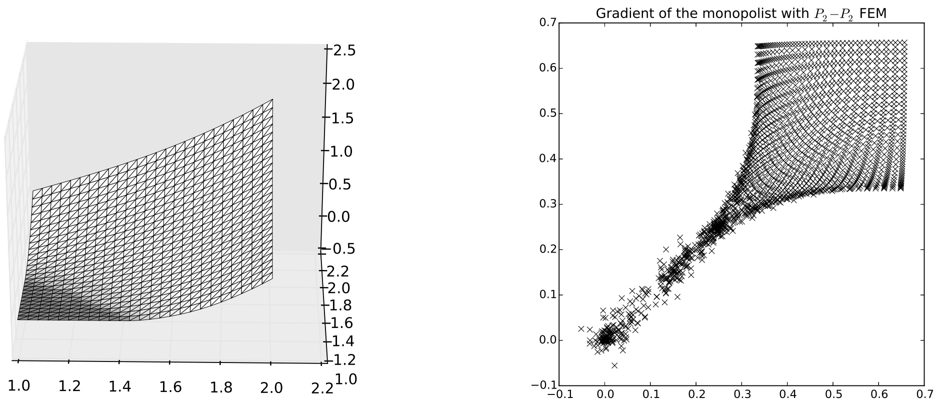

Figure 10: The monopolist's graph and its gradient

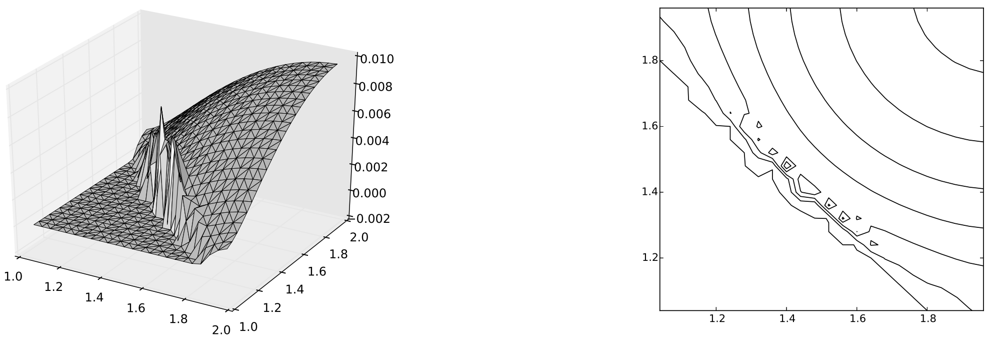

Figure 11: The eigenvalues

6 CONCLUSION

In this article we prove (Theorem 3) that the P_1 discretization of a function that satisfies a strong definition of convexity (with $\mathcal{C}_0^\infty$ test functions), which is equivalent to the gradients' jumps positivity (for P_1 FE), and so is conformal, leads to an additional constraint on the limit function. While [1] dealt mainly with regular meshes, our result extends explicitely to any mesh. The error of such a discretization of the constraint does not vanish with the space step. We justify that it is localized where the additional constraint on the limit function is not satisfied. The condition for existence of such a counter-example requires information both from the mesh and its refinement but is general.

In addition, among the P_1 discretization of u with a strong (section 3.1) or weak (section 3.2) definition of convexity, not all are consistent. The definition of consistency is very similar to the one of partial differential equations and it is used above to discriminate likely discretizations and unlikely ones. But this does not guarantee good numerical results even when the discretization is consistent. We also prove on a structured mesh that the P_1 discretization of u and the use of P_1 test functions for convexity (say $P_1 - P_1$ convexity) may or maynot be consistent (Table 1).

We also test the gradients' jumps of P_2 functions (strong discretization with $\mathcal{C}_0^\infty$ test functions in subsection 4.1) and various weak P_2 discretizations for u_h and for some P_2 test functions (subsections 4.2 and 4.3). Some of them are consistent and some are not. We also discuss the literature, but let a deeper study of how [16, 17] interact with [1] to a forthcoming article. In a last subsection, we numerically illustrate the convergence of [17]'s FEM where the trial and test basis are different. Even if other methods exist, this FEM is maybe one of the simplest. Comparisons of these methods would be very interesting.

Acknowledgments

The author thanks G. Carlier for drawing his attention to the problem of subharmonicity approximation.

References

[1] P. Choné and H.V.J. Le Meur (2001). Non-convergence result for conformal approximation of variational problems subject to a convexity constraint. *Numer. Funct. Anal. Optim.* 22 no. 5-6:529–547.

[2] I. Newton (1686). *Philosophiae Naturalis Principia Mathematica.*

[3] H.H. Goldstine (1980). *A history of the calculus of variations from the 17th to the 19th century* Springer-Verlag, Heidelberg.

[4] G. Buttazzo, V. Ferone and B. Kawohl (1995). Minimum problems over sets of concave functions and related questions. *Math. Nachr.* 173:71-89.

[5] T. Lachand-Robert and M.A. Peletier (2001). An Example of Non-convex Minimization and an Application to Newton's Problem of the Body of Least Resistance, *Ann. Inst. H. Poincaré Anal. Non Linéaire* 18, no. 2:179–198.

[6] P.L. Lions (1998). Identification du cône dual des fonctions convexes et applications. (French. English, French summary) [Identification of the dual cone of convex functions and applications] *C. R. Acad. Sci. Paris Sér. I Math.* 326 no. 12, 1385–1390.

[7] T. Lachand-Robert and É. Oudet (2005). Minimizing within convex bodies using a convex hull method, *SIAM Journal on Control and Optimization* Vol. 16 Number 2:368-379.

[8] G. Carlier (2004). On a theorem of Alexandrov. *J. Nonlinear Convex Anal.* 5, no. 1:49–58.

[9] G. Carlier, M. Comte and G. Peyré (2009). Approximation of maximal Cheeger sets by projection, *M2AN Math. Model. Numer. Anal.* 43(1):139-150.

[10] J.-C. Rochet and P. Choné (1998). Ironing, sweeping and multidimensionnal screening, *Econometrica* **66**:783-826.

[11] G. Carlier and T. Lachand-Robert (2001). Regularity of solutions for some variational problems subject to a convexity constraint. *Comm. Pure Appl. Math.* 54 no. 5:583–594.

[12] N.K. Leung and R.J. Renka (1999). C^1 convexity-preserving interpolation of scattered data. *SIAM J. Sci. Comput.* 20 , no. 5:1732–1752.

[13] G. Carlier, T. Lachand-Robert and B. Maury (1999). H^1-projection into the set of convex functions: a saddle-point formulation. *CEMRACS 1999 (Orsay)*, 277–289 (electronic), *ESAIM Proc., 10, Soc. Math. Appl. Indust.*, Paris.

[14] B. Maury (2003). Version continue de l'algorithme d'Uzawa. (French) [Continuous version of the Uzawa algorithm] *C. R. Math. Acad. Sci.* Paris 337 , no. 1:31–36.

[15] G. Carlier, T. Lachand-Robert and B. Maury (2001). A numerical approach to variational problems subject to convexity constraint, *Numer. Math.* 88, no. 2:299–318.

[16] N.E. Aguilera and P. Morin (2008). Approximating optimization problems over convex functions, *Numer. Math.* 111(1):1-34

[17] N.E. Aguilera and P. Morin (2009). On convex functions and the finite element method, *SIAM J. Numer. Anal.* 47(4):3139-3157

[18] I. Ekeland and S. Moreno-Bromberg (2010). An algorithm for computing solutions of variational problems with global convexity constraints, *Numer. Math.* 115(1):45-69.

[19] Q. Mérigot and É. Oudet (2014). Handling convexity-like constraints in variational problems, *SIAM J. Numer. Anal.*, 52 (5):2466–2487.

[20] J.M. Mirebeau (2015). Adaptive, Anisotropic and Hierarchical cones of Discrete Convex functions, *Numer. Math.* pp. 1–47.

[21] B. Kawohl and C. Schwab, Convergent finite elements for a class of nonconvex variationnal problems, *IMA Journal of Numerical Analysis* (1998) **18**:133-149.