Available online at www.sciencedirect.com
SCIENCE DIRECT®

J. Math. Pures Appl. 82 (2003) 1527–1553

JOURNAL DE MATHÉMATIQUES PURES ET APPLIQUÉES

www.elsevier.com/locate/matpur

The obstacle problem for water tanks

J.F. Bonnans[a,*], R. Bessi Fourati[b], H. Smaoui[c]

[a] *Projet SYDOCO, INRIA, Domaine de Voluceau, B.P. 105, 78153 Rocquencourt, France*
[b] *ENIT, Tunis, Tunisia*
[c] *URMCS, ENIT, Tunis, Tunisia*

Received 18 June 2003

Abstract

In this paper we discuss the problem of computing and analyzing the static equilibrium of a nonrigid water tank. Specifically, we fix the amount of water contained in the tank, modelled as a membrane. In addition, there are rigid obstacles that constrain the deformation. This amounts to a nonconvex variational problem. We derive the optimality system and its interpretation in terms of equilibrium of forces. A second-order sensitivity analysis, allowing to compute derivatives of solutions and a second-order Taylor expansion of the cost function, is performed, in spite of the fact that the cost function is not twice differentiable. We also study the finite elements discretization, introduce a decomposition algorithm for the numerical computation of the solution, and display numerical results.

Résumé

Cet article discute le problème du calcul et de l’analyse de l’équilibre statique d’un réservoir d’eau souple. La quantité d’eau est fixée et la déformation obéit à l’équation des membranes. De plus, des obstacles rigides limitent la déformation. Ceci aboutit à un problème variationnel non-convexe. Nous obtenons le système d’optimalité et son interprétation en terme d’équilibre de forces. Une étude de sensibilité au second ordre permet le calcul des dérivées de la solution ainsi qu’un développement de Taylor au second ordre de la valeur, bien que le critère ne soit pas deux fois différentiable. Nous étudions aussi la discrétisation par éléments finis et nous introduisons un algorithme de décomposition pour le calcul numérique, et finalement nous donnons des résultats numériques.

Keywords: Obstacle problem; Nonconvex variational problems; Sensitivity analysis; Finite elements; Decomposition algorithms

[*] Corresponding author.

E-mail addresses: frederic.bonnans@inria.fr (J.F. Bonnans), fourati.mhenni@planet.tn (R. Bessi Fourati), hichem.smaoui@enit.rnu.tn (H. Smaoui).

0021-7824/$ – see front matter © 2003 Éditions scientifiques et médicales Elsevier SAS. All rights reserved.
doi:10.1016/j.matpur.2003.09.003

Mots-clés : Problème de l'obstacle ; Problèmes variationnels non convexes ; Analyse de sensibilité ; Éléments finis ; Algorithmes de décomposition

1. Introduction

Let Ω be a connected, bounded and open subset of $\mathbb{R}^n$, $n = 1$ or 2, with Lipschitz boundary $\partial\Omega$ (Ω is an interval if $n = 1$). Given $f \in L^2(\Omega)$, the classical obstacle problem reads as follows:

$$\operatorname*{Min}_{v\in K} \frac{1}{2}\int\limits_{\Omega} \big\|\nabla v(\omega)\big\|^2 \,\mathrm{d}\omega - \int\limits_{\Omega} f(\omega)v(\omega)\,\mathrm{d}\omega, \tag{OP$_f$}$$

where $K \subset H_0^1(\Omega)$ is the set of functions satisfying the constraint of nonpenetration with a certain obstacle, defined by:

$$K = \big\{v \in H_0^1(\Omega);\ v(\omega) \leqslant \Phi(\omega) \text{ a.e. on } \Omega\big\}. \tag{1}$$

Here Φ is a measurable extended value function $\Omega \to \mathbb{R} \cup \{+\infty\}$, such that the set K above is nonempty. This holds if $\Phi(\omega) = +\infty$ a.e., which is the case without obstacle, if $\Phi(\omega) \geqslant 0$ a.e., and also if Φ belongs to $H^1(\Omega)$, and is nonnegative on the boundary of Ω. This problem is perhaps the simplest example of a variational inequality, and has been the subject of numerous works. The starting point of the study of variational inequalities was Lions and Stampacchia [15]. Extension to various mechanical problems was made in Duvaut and Lions [9]. At the same time, Brézis [6] established various mathematical properties of the solutions of variational inequalities. Mignot [16] showed that polyhedricity of the feasible set allowed to perform a sensitivity analysis of solutions (see also Haraux [13]), the expression of which necessitates the concepts of capacity theory; see also the introduction to the subject [5, Section 6.4]. Two recent papers discuss the case when the field f is itself the result of a mechanical equilibrium. In Aissani, Chipot and Fouad [1] the membrane supports one or two heavy disks. Buttazzo and Wagner [7] consider the case of a support of a rigid body.

Another approach to the sensitivity analysis consists in studying the solutions of the optimality system rather than those of the minimization problem. Among abstract results that possibly apply, let us mention [5, Theorem 5.10], and Levy [14]. The latter computes proto-derivatives, which coincide with derivatives if the latter exist. This approach has the advantage of avoiding the second-order sufficient conditions. It has been applied to a nonlinear obstacle plate problem in Figueiredo and Leal [10].

The novelty in our study lies in the fact that, in addition to the given distributed forces field f, we take into account the weight of a given amount of water, filling the volume between the part of the tank that is below the water level, and the water level itself. The latter is of course an unknown of the problem. The mechanical potential to be minimized is a nonconvex function of vertical displacement and water level. This potential is to be minimized under the restriction that the volume of water is given. Although the potential and constraints are nonconvex and nonsmooth, we can establish existence of solutions, give a mechanical interpretation of the optimality system, and perform, under reasonable assumptions, a sensitivity analysis. Finally, we study a decomposition algorithm whose essential step is to solve at each iteration a classical obstacle problem, and display numerical results.

2. Setting and equivalent formulations

As said in the introduction, let Ω be a connected, bounded and open subset of $\mathbb{R}^n$, $n = 1$ or 2, with Lipschitz boundary $\partial\Omega$. Consider a membrane fixed at the boundary. Let $v(\omega)$ be the vertical displacement, positively oriented downward. Under the hypothesis of small deformation, we have that the potential of elastic deformation is $E_D(v) = \frac{1}{2}\int_\Omega \|\nabla v(\omega)\|^2 \,\mathrm{d}\omega$. Here and later, we assume physical constants to be equal to 1 for the sake of notational simplicity. The potential associated with a distributed forces field $f \in L^2(\Omega)$ (oriented downward) is $E_C(v) = -\int_\Omega f(\omega)v(\omega)\,\mathrm{d}\omega$. In addition, assume that the tank formed by the deformed membrane contains some water. If $h \in \mathbb{R}$ denotes the water level, the gravity potential associated with the water is

$$E_F(v,h) := -\frac{1}{2}\int_\Omega \big(v(\omega)+h\big)\big(v(\omega)-h\big)_+ \,\mathrm{d}\omega.$$

Indeed, the height of water is $(v(\omega)-h)_+$, hence, after integration we obtain the above expression. The mechanical potential is defined as the sum of the three potentials already discussed: $E(v,h) := E_D(v) + E_C(v) + E_F(v,h)$. The constraint over the volume $L > 0$ of water is

$$G(v,h) = \int_\Omega \big(v(\omega)-h\big)_+ \,\mathrm{d}\omega = L. \tag{2}$$

Let K be defined by (1). Taking $H_0^1(\Omega)$ as the space of displacement, we may formulate the problem of static equilibrium as the minimization of the mechanical potential, subject to the constraint of the volume of water (2) and to the obstacle constraint:

$$\min E(v,h); \qquad G(v,h) = L; \quad (v,h) \in K \times \mathbb{R}. \tag{3}$$

It may be more efficient to consider another formulation of this problem. Observe that, whenever the constraint is satisfied, the gravity potential associated with the water is such that $E_F(v,h) = -\frac{1}{2}\int_\Omega (v(\omega)-h)_+^2 \,\mathrm{d}\omega - hL$. Therefore, define the cost function,

$$J(v,h,L) := \frac{1}{2}\int_\Omega \big\|\nabla v(\omega)\big\|^2 \,\mathrm{d}\omega - \frac{1}{2}\int_\Omega \big(v(\omega)-h\big)_+^2 \,\mathrm{d}\omega - \int_\Omega f(\omega)v(\omega)\,\mathrm{d}\omega - hL.$$

A problem equivalent to (3) is

$$\operatorname*{Min}_{v,h} J(v,h,L); \qquad G(v,h) = L; \quad (v,h) \in K \times \mathbb{R}. \tag{P}$$

In the sequel, we will denote by $F(\mathrm{P})$, $S(\mathrm{P})$ and $\mathrm{val}(\mathrm{P})$ the set of feasible points, set of solutions, and value of an optimization problem; the value is the infimum of cost function over the feasible set. Let "meas" denote Lebesgue's measure. Surprisingly, we may "forget the constraint" if the cost function is maximized (instead of being minimized) with respect to h.

Proposition 2.1. *Given $L > 0$, an element $(\bar{v}, \bar{h})$ of $K \times \mathbb{R}$ is solution of* (P) *if and only if $\bar{v}$ is solution of the problem below*:

$$\operatorname*{Min}_{v\in K} \sup_{h\in\mathbb{R}} J(v, h, L). \tag{4}$$

Proof. We have that $h \to J(v, h, L)$ is a concave function, with continuous derivative $\int_\Omega (v(\omega) - h)_+ \,\mathrm{d}\omega - L$. We check in the lemma below that this derivative is continuous, and is equal to 0 for a unique value of h, denoted $h(v, L)$. This is precisely the value for which the constraint is satisfied; in other words, $h(v, L)$ is the height of water associated with deformation v and volume L. Therefore $\sup_{h\in\mathbb{R}} J(v, h, L) = J(v, h(v, L), L)$, from which the conclusion follows easily. □

The above result is related to the fact that $h(v, L)$ has an interpretation as a Lagrange multiplier, see Lemma 10.1. We denote by $\mathbb{R}_{++}$ the set of positive real numbers.

Lemma 2.2. (i) *Given $(v, L) \in L^2(\Omega) \times \mathbb{R}_{++}$, there exists a unique $h = h(v, L)$ such that $G(v, h) = L$. In addition, the function $h(v, L)$, with domain $L^2(\Omega) \times \mathbb{R}_{++}$, is convex, locally Lipschitz, nondecreasing* (*respectively nonincreasing*) *function of v* (*respectively L*), *and satisfies, if $h_i = h(v_i, L_i)$, for $i = 1, 2$*:

$$\begin{aligned}&\mathrm{meas}\big(\{v_2 \geqslant h_2; v_1 \geqslant h_1\}\big)|h_2 - h_1| \\ &\quad \leqslant |L_2 - L_1| + 2\,\mathrm{meas}(\Omega)^{1/2}\|v_2 - v_1\|_{L^2(\Omega)}.\end{aligned} \tag{5}$$

(ii) *The function $h(v, L)$ has a directional derivative δh in direction $(\delta v, \delta L)$ determined by the relation*:

$$\int_{\{v=h\}} (\delta v - \delta h)_+ + \int_{\{v>h\}} (\delta v - \delta h) = \delta L. \tag{6}$$

(iii) *The restriction of $h(v, L)$ to $H_0^1(\Omega) \times \mathbb{R}_{++}$ is Fréchet directionally differentiable. More precisely, $\delta h(\delta v, \delta L)$ denoting the solution of* (6), *we have that*

$$h(v + \delta v, L + \delta L) = h(v, L) + \delta h(\delta v, \delta L) + \mathrm{o}\big(\|\delta v\|_{H_0^1(\Omega)} + |\delta L|\big). \tag{7}$$

Proof. (i) Let $(v, L) \in L^2(\Omega) \times \mathbb{R}_+$. Using Lebesgue's dominated convergence theorem, it is easily checked that the real function

$$h \to g(h) := \int_{\Omega} \big(v(\omega) - h\big)_+ \, \mathrm{d}\omega - L$$

is continuous, nonincreasing, and varies over $\mathbb{R}$ from $+\infty$ to $-L$. It follows that g has at least one zero, say $\bar{h}$, and the set of zeroes is an interval. In addition, by Lebesgue's dominated convergence theorem, we have that $g(h)$ has directional derivatives, whose expression is

$$g'(h, \delta h) = -\min(\delta h, 0)\,\mathrm{meas}\{v = h\} - \delta h\,\mathrm{meas}\{v > h\}. \tag{8}$$

Since $L > 0$, and hence, $\mathrm{meas}\{v > \bar{h}\} > 0$, we have that $g'(h, \delta h)$ is nonzero when $\delta h \neq 0$, of sign opposite to the one of δh. This implies uniqueness of the zero of g. Denote the latter by $h(v, L)$. Since g is nonincreasing, $h(v, L)$ is a nondecreasing (respectively nonincreasing) function of v (respectively L). Let us prove that this function is convex. Let v_1 and v_2 belong to $L^2(\Omega)$, and L_1 and L_2 be two positive numbers. Set $h_i = h(v_i, L_i)$, for $i = 1, 2$. Let $\alpha \in (0, 1)$, and set $v = \alpha v_1 + (1-\alpha)v_2$, $L = \alpha L_1 + (1-\alpha)L_2$. Since G is convex, we have that

$$L = \alpha G(v_1, h_1) + (1-\alpha)G(v_2, h_2) \geqslant G\big(v, \alpha h_1 + (1-\alpha)h_2\big).$$

Since G is a nonincreasing function of its second argument, convexity of $h(v, L)$ follows. Being convex, $h(v, L)$ is locally Lipschitz. Let us prove the estimate (5). Assume for instance that $h_2 \geqslant h_1$. We have:

$$\begin{aligned} L_2 - L_1 = {} & \int_{\{v_2 \geqslant h_2;\, v_1 \geqslant h_1\}} \big(v_2(\omega) - v_1(\omega) - h_2 + h_1\big)\,\mathrm{d}\omega + \int_{\{v_2 \geqslant h_2;\, v_1 < h_1\}} \big(v_2(\omega) - h_2\big)\,\mathrm{d}\omega \\ & - \int_{\{v_2 < h_2;\, v_1 \geqslant h_1\}} \big(v_1(\omega) - h_1\big)\,\mathrm{d}\omega. \end{aligned}$$

We may majorize the last term by 0, and for the two others we have:

$$\begin{aligned} & \int_{\{v_2 \geqslant h_2;\, v_1 \geqslant h_1\}} \big(v_2(\omega) - v_1(\omega) - h_2 + h_1\big)\,\mathrm{d}\omega \\ & \quad \leqslant \mathrm{meas}\big(\{v_2 \geqslant h_2;\, v_1 \geqslant h_1\}\big)(h_1 - h_2) + \mathrm{meas}(\Omega)^{1/2}\|v_2 - v_1\|_{L^2(\Omega)}, \end{aligned}$$

and

$$\begin{aligned} \int_{\{v_2 \geqslant h_2;\, v_1 < h_1\}} \big(v_2(\omega) - h_2\big)\,\mathrm{d}\omega & \leqslant \int_{\{v_2 \geqslant h_2;\, v_1 < h_1\}} \big(v_2(\omega) - v_1(\omega)\big)\,\mathrm{d}\omega \\ & \leqslant \mathrm{meas}(\Omega)^{1/2}\|v_2 - v_1\|_{L^2(\Omega)}. \end{aligned}$$

Combining the previous inequalities, we obtain (5).

(ii) Fix $(\delta v, \delta L) \in L^2(\Omega) \times \mathbb{R}$, and let δh be such that (6) holds. Then

$$G(v + t\delta v, h + t\delta h) = L + t\delta L + \mathrm{o}(t).$$

Since $h(\cdot\,,\cdot)$ is locally Lipschitz, we have that $h(v + t\delta v, L + t\delta L) = h + t\delta h + \mathrm{o}(t)$. Relation (6) follows.

(iii) For the sake of notational simplicity we prove the result when $\delta L = 0$. Assume that (7) does not hold, and hence, there exist $\varepsilon > 0$ and a sequence $v_k \to v$ in $H_0^1(\Omega)$ such that

$$\bigl|h(v_k, L) - h(v, L) - \delta h(v_k - v, 0)\bigr| \geqslant \varepsilon \|v_k - v\|_{H_0^1(\Omega)}. \tag{9}$$

We may write $v_k = v + t_k w_k$, with w_k of unit norm in $H_0^1(\Omega)$ and $t_k \to 0$. Extracting, if necessary, a subsequence, we may assume that w_k weakly converges to some $\overline{w}$ in $H_0^1(\Omega)$. Note that $\overline{w}$ is of norm at most one, and may be equal to 0. We have that w_k strongly converges to $\overline{w}$ in $L^2(\Omega)$. Since $h(\cdot\,,\cdot)$ is a Lipschitz function and has directional derivatives, it is also Hadamard directionally differentiable, and has continuous directional derivatives, see, e.g., [5, Proposition 2.49]. It follows that

$$h(v + t_k w_k, L) = h(v, L) + t_k \delta h(\overline{w}, 0) + \mathrm{o}(t_k) = h(v, L) + t_k \delta h(w_k, 0) + \mathrm{o}(t_k)$$

in contradiction with (9). □

Denote $F(v) := J(v, h(v, L), L)$, where $(v, L) \in L^2(\Omega) \times \mathbb{R}_{++}$. A problem equivalent to (P) is what we will call the *reduced problem*:

$$\operatorname*{Min}_{v} F(v); \quad v \in K. \tag{RP}$$

3. Existence and basic properties

In this section we will establish the existence of solutions of problem (RP). The hard point is to check coerciveness of the cost function in the sense that, for any $L > 0$, $F(v) \to +\infty$ when $\|v\|_{H_0^1(\Omega)} \to +\infty$. Since Ω has a Lipschitz boundary, we have the following Sobolev inclusion (e.g., Gilbarg and Trudinger [11, Theorem 7.26])

$$H^1(\Omega) \subset L^4(\Omega), \quad \text{with compact injection.} \tag{10}$$

Since the dimension is at most 2, the compact inclusion $H^1(\Omega) \subset L^p(\Omega)$ holds for all p in $[2, \infty[$. However, only (10) is used in our proofs, and not other property of the boundary. Indeed, up to Section 6 (including it) we only use the compact injection $H^1(\Omega) \subset L^2(\Omega)$.

Lemma 3.1. *For all $\varepsilon > 0$, there exists $C_\varepsilon > 0$ such that, for all $v \in H^1(\Omega)$, one has*:

$$\int_\Omega v^2(\omega)\,\mathrm{d}\omega \leqslant \varepsilon \int_\Omega \bigl\|\nabla v(\omega)\bigr\|^2\,\mathrm{d}\omega + C_\varepsilon \biggl(\int_\Omega \bigl|v(\omega)\bigr|\,\mathrm{d}\omega\biggr)^2. \tag{11}$$

Proof. If the conclusion were false, there would exist $\varepsilon > 0$ and a sequence v_k in $H^1(\Omega)$ such that $\int_\Omega (v_k(\omega))^2\,\mathrm{d}\omega = 1$, and

$$1 > \varepsilon \int_\Omega \|\nabla v_k(\omega)\|^2\,\mathrm{d}\omega + k\left(\int_\Omega v_k(\omega)\,\mathrm{d}\omega\right)^2. \tag{12}$$

Clearly v_k is bounded in $H^1(\Omega)$, and has a weak limit point $\bar{v}$. By (10), the latter satisfies $\int \bar{v}^2(\omega)\,\mathrm{d}\omega = 1$, as well as $(\int_\Omega |\bar{v}(\omega)|\,\mathrm{d}\omega)^2 = 0$, which is impossible. □

Since Ω is bounded, Poincaré's inequality holds, i.e., there exists $c_P > 0$ such that $\|v\|_{L^2(\Omega)} \leqslant c_P(\int_\Omega \|\nabla v(\omega)\|^2\,\mathrm{d}\omega)^{1/2}$, for all $v \in H_0^1(\Omega)$. Therefore, we endow $H_0^1(\Omega)$ with the norm $\|v\|_{H_0^1(\Omega)} := (\int_\Omega \|\nabla v(\omega)\|^2\,\mathrm{d}\omega)^{1/2}$.

Proposition 3.2. (i) *For all* $\varepsilon > 0$, C_ε *denoting the constant in Lemma* 3.1, *for all* $(v, L) \in H_0^1(\Omega) \times \mathbb{R}_{++}$, *the following inequality holds*:

$$F(v) \geqslant \left(\frac{1}{2} - c_P \varepsilon^{1/2}\right)\|v\|^2_{H_0^1(\Omega)} - c_P\big(\|f\|_{L^2(\Omega)} + C_\varepsilon^{1/2} L\big)\|v\|_{H_0^1(\Omega)}. \tag{13}$$

(ii) *If the reduced problem* (RP) *is feasible, its set of solutions is nonempty, weakly closed, and uniformly bounded whenever* (f, L) *varies in a bounded subset of* $L^2(\Omega) \times \mathbb{R}_{++}$.

(iii) *If* $f_k \to \bar{f}$ *weakly in* $L^2(\Omega)$, $L_k \to \bar{L}$ *in* $\mathbb{R}_{++}$, *and if* v_k *is a solution of problem* (RP) *with* $f = f_k$ *and* $L = L_k$, *then any weak limit point* $\bar{v}$ *of* v_k *is a strong limit point, and is solution of problem* (RP) *with* $f = \bar{f}$ *and* $L = \bar{L}$.

Proof. (i) We have that $F(v) = \frac{1}{2}\|v\|^2_{H_0^1(\Omega)} - \int_\Omega f(\omega)v(\omega)\,\mathrm{d}\omega - F_1(v, L)$, where

$$\begin{aligned} F_1(v) = \frac{1}{2}\int_\Omega \big(v(\omega) + h\big)\big(v(\omega) - h\big)_+\,\mathrm{d}\omega &\leqslant \int_\Omega v(\omega)\big(v(\omega) - h\big)_+\,\mathrm{d}\omega \\ &\leqslant \|v\|_{L^2(\Omega)}\big\|(v - h)_+\big\|_{L^2(\Omega)} \leqslant c_P\|v\|_{H_0^1(\Omega)}\big\|(v - h)_+\big\|_{L^2(\Omega)}. \end{aligned} \tag{14}$$

Applying Lemma 3.1, for any $\varepsilon > 0$, we obtain:

$$\begin{aligned} \int_\Omega \big(v(\omega) - h\big)_+^2\,\mathrm{d}\omega &\leqslant \varepsilon\int_\Omega \big\|\nabla\big(v(\omega) - h\big)_+\big\|^2\,\mathrm{d}\omega + C_\varepsilon\left(\int_\Omega \big(v(\omega) - h\big)_+\,\mathrm{d}\omega\right)^2 \\ &\leqslant \varepsilon\int_\Omega \big\|\nabla v(\omega)\big\|^2\,\mathrm{d}\omega + C_\varepsilon L^2, \end{aligned}$$

and hence, since $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for all nonnegative a and b,

$$\left\|(v-h)_+\right\|_{L^2(\Omega)} \leqslant \varepsilon^{1/2}\|v\|_{H_0^1(\Omega)} + C_\varepsilon^{1/2} L. \tag{15}$$

We also have, denoting again by c_P the constant in Poincaré's inequality,

$$\int_\Omega f(\omega)v(\omega)\,\mathrm{d}\omega \leqslant \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leqslant c_P\|f\|_{L^2(\Omega)}\|v\|_{H_0^1(\Omega)}. \tag{16}$$

We obtain (13) by combining (14), (15) and (16).

(ii) We prove uniform boundedness of solutions. Whenever (f, L) varies in a bounded subset of $L^2(\Omega) \times \mathbb{R}_{++}$, taking ε small enough, we have by (13) an inequality of the form $F(v) \geqslant \frac{1}{4}\|v\|^2_{H_0^1(\Omega)} - a\|v\|_{H_0^1(\Omega)}$, where $a > 0$ does not depend on (v, f, L). On the other hand, choosing a bounded feasible solution $v_0 \in K$, with associated height $h(v_0, L)$, it is easily checked that

$$F(v_0) \leqslant \frac{1}{2}\int_\Omega \left\|\nabla v_0^2(\omega)\right\| \mathrm{d}\omega + c_P\|f\|_{L^2(\Omega)}\|v\|_{H_0^1(\Omega)} + L\left|h(v_0, L)\right|.$$

Since the function $h(v_0, \cdot)$ is bounded on bounded sets, $F(v_0)$ is uniformly upper bounded whenever (f, L) is bounded. Since $F(v) \leqslant F(v_0)$, combining with (i), we obtain uniform boundedness of solutions.

By (10), $G(v, h)$, and hence $h(v, L)$ is weakly continuous. It follows that $F(v)$ is weakly lower semicontinuous, hence the set of solutions is weakly closed. Feasibility of (RP) and weak semi continuity of its cost function, as well as weak closedness of the feasible set, combined with uniform boundedness of solutions implies existence of at least one solution.

(iii) This is an easy consequence of (ii), the strong convergence of the subsequence of v_k being due to the fact that convergence of the cost function and weak convergence of its arguments implies the strong convergence. The reason is that the cost function is the sum of the square of $H_0^1(\Omega)$ norm and a weakly continuous term. □

4. When is the cost function convex?

Using the rules for directional derivatives of locally Lipschitz functions, we have that $F(v)$ has directional derivatives:

$$F'(v)\delta v = \int_\Omega \nabla v(\omega) \cdot \nabla \delta v(\omega)\,\mathrm{d}\omega - \int_\Omega \left[\left(v(\omega) - h(v, L)\right)_+ + f(\omega)\right]\delta v(\omega)\,\mathrm{d}\omega. \tag{17}$$

This expression takes into account the fact that, by Proposition 2.1, the directional derivative of $J(v, h, L)$ with respect to h is, when $h = h(v, L)$, equal to 0. Observe that the directional derivative is linear and continuous with respect to δv, and hence, F is

Gâteaux differentiable. In addition, this Gâteaux derivative is continuous, since $h(v, L)$ is a continuous function, and hence, F is continuously differentiable. We denote by DF the derivative of F. Similarly, it is easily checked that the second-order directional derivative of F in direction δv, defined as,

$$D^2F(v)(\delta v, \delta v) := \lim_{t \downarrow 0} \frac{F(v + t\delta v) - F(v) - DF(v)\delta v}{\frac{1}{2}t^2},$$

has the following expression, where δh is the directional derivative of $h(v, L)$ in direction δv:

$$D^2F(v)(\delta v, \delta v) = \int\limits_{\Omega} \big\|\nabla \delta v(\omega)\big\|^2 \,\mathrm{d}\omega - \int\limits_{\{v > h(v,L)\}} \big(\delta v(\omega) - \delta h\big)^2 \,\mathrm{d}\omega$$
$$- \int\limits_{\{v = h(v,L)\}} \big(\delta v(\omega) - \delta h\big)_+^2 \,\mathrm{d}\omega.$$

It is clear that F is convex iff $D^2F(v)(\delta v, \delta v) \geqslant 0$, for all v and δv in $H_0^1(\Omega)$. The cost function F is not always convex, as the following example shows.

Example 4.1. Let $L = 1$, and let $v_0 \in H^1(\Omega)_+$ be such that $\int_\Omega v_0(\omega)\,\mathrm{d}\omega = 1$, whence $h(v_0, L) = 0$. Then $DF(v_0)v_0 = \int_\Omega \|\nabla v_0(\omega)\|^2\,\mathrm{d}\omega - \int_\Omega v_0^2(\omega)\,\mathrm{d}\omega$. For $v = 0$ we have $h(0, 1) = -\operatorname{meas}(\Omega)^{-1}$, and $DF(0)v_0 = -\operatorname{meas}(\Omega)^{-1}$. It follows that

$$\big(DF(v_0) - DF(0)\big)v_0 = \int\limits_{\Omega} \big\|\nabla v_0(\omega)\big\|^2 - \int\limits_{\Omega} v_0^2(\omega)\,\mathrm{d}\omega + \operatorname{meas}(\Omega)^{-1}.$$

Assume that $\Omega =]0, m[$ so that $\operatorname{meas}(\Omega) = m$, and take $v_0(\omega) := c\sin(\Pi x/m)$. That $\int_\Omega v_0(\omega)\,\mathrm{d}\omega = 1$ implies $c = \pi/(2m)$. We have that $\int_\Omega v_0^2(\omega)\,\mathrm{d}\omega = \pi^2/(8m)$, while $\int_\Omega \|\nabla v_0(\omega)\|^2\,\mathrm{d}\omega = \pi^4/(8m^3)$. Hence

$$\big(DF(v_0) - DF(0)\big)v_0 = \pi^4/\big(8m^3\big) - \big(\pi^2/8 - 1\big)m^{-1}$$

is negative, and therefore F is not convex, if m is large enough.

Examples of nonconvexity of the cost for bidimensional problems are discussed in [3]. It can be suspected that the cost function F is convex whenever Ω is “small enough”, since in that case the first term in the expression of $D^2F(v)(\delta v, \delta v)$ should dominate the two others. For proving such results we recall the following notions. A classical result of functional analysis (e.g., [8, Vol. 5, p. 120]) is that the positive amount

$$\nu_0(\Omega) := \inf_{\substack{v \in H_0^1(\Omega) \\ v \neq 0}} \frac{\int_\Omega \|\nabla v(\omega)\|^2\,\mathrm{d}\omega}{\|v\|_{L^2(\Omega)}^2} \tag{18}$$

is in fact the smallest eigenvalue of $-\Delta$, where Δ is the Laplacian operator in $H_0^1(\Omega) \cap H^2(\Omega)$, and that the eigenvector $w_0 \neq 0$ is unique (up to multiplication by a scalar), nonzero and of constant sign, say positive, over Ω.

For any open and connected subset $\widehat{\Omega} \subset \Omega$, let (see, e.g., [8, Vol. 3, p. 926] $V(\widehat{\Omega}) := \{v \in H^1(\widehat{\Omega});\ \int_{\widehat{\Omega}} v(\omega)\,\mathrm{d}\omega = 0\}$ denote the set of functions over $\widehat{\Omega}$ with square integrable gradient and zero mean, and set:

$$\nu_1(\widehat{\Omega}) := \inf_{\substack{v \in V(\widehat{\Omega}) \\ v \neq 0}} \frac{\int_{\widehat{\Omega}} \|\nabla v(\omega)\|^2\,\mathrm{d}\omega}{\|v\|^2_{L^2(\widehat{\Omega})}}. \tag{19}$$

If the injection of $H^1(\widehat{\Omega})$ into $L^2(\widehat{\Omega})$ is compact, it is not difficult to check that $\nu_1(\widehat{\Omega}) > 0$, and that there exists a nonzero eigenvector $\widehat{w} \in V(\widehat{\Omega})$ solution of

$$-\Delta \widehat{w} = \nu_1(\widehat{\Omega})\widehat{w} \quad \text{in } \widehat{\Omega}, \qquad \frac{\partial \widehat{w}}{\partial n} = 0 \quad \text{on } \partial\widehat{\Omega}, \tag{20}$$

where $\partial \cdot / \partial n$ denotes the normal derivative.

Lemma 4.2. (i) *The function* $\eta : \mathbb{R} \to \mathbb{R}_+$ *defined by*:

$$\eta(\delta h) := \int\limits_{\{v > h(v,L)\}} \big(\delta v(\omega) - \delta h\big)^2\,\mathrm{d}\omega + \int\limits_{\{v = h(v,L)\}} \big(\delta v(\omega) - \delta h\big)_+^2\,\mathrm{d}\omega,$$

is convex and attains its minimum when $\delta h = h'$, *where we denote in this lemma* $h' := h'((v,L),(\delta v,0))$.

(ii) *The cost function* F *is convex* (*respectively strongly convex*) *over* $H_0^1(\Omega)$ *whenever* $\nu_0(\Omega) \geqslant 1$ *or* $\nu_1(\Omega) \geqslant 1$ (*respectively* $\nu_0(\Omega) > 1$ *or* $\nu_1(\Omega) > 1$).

Proof. (i) The function η is easily seen to be convex, and therefore attains its minimum when its derivative vanishes, i.e., when $\delta h = h'$.

(ii) Using $\eta(0) \geqslant \eta(h')$, which follows from (i), get:

$$\begin{aligned} D^2F(v)(\delta v, \delta v) &\geqslant \|\delta v\|^2_{H_0^1(\Omega)} - \int\limits_{\{v > h(v,L)\}} \delta v^2(\omega)\,\mathrm{d}\omega - \int\limits_{\{v = h(v,L)\}} \delta v_+^2(\omega)\,\mathrm{d}\omega \\ &\geqslant \int\limits_{\Omega} \big\|\nabla \delta v(\omega)\big\|^2\,\mathrm{d}\omega - \int\limits_{\Omega} \delta v^2(\omega)\,\mathrm{d}\omega \end{aligned}$$

which proves convexity if $\nu_0(\Omega) \geqslant 1$ (respectively strong convexity if $\nu_0(\Omega) > 1$). The proof of the case when $\nu_1(\Omega) \geqslant 1$ (respectively $\nu_1(\Omega) > 1$) is similar, using $\eta(h_0) \geqslant \eta(h')$, where h_0 is such that $\int_\Omega (\delta v(\omega) - h_0)\,\mathrm{d}\omega = 0$. □

5. First-order optimality conditions

The first step consists in obtaining primal first-order optimality conditions. The *cone of feasible directions*, and the *cone of tangent directions* to K at $\bar{v} \in K$ are defined, respectively, as

$$\mathcal{R}_K(\bar{v}) := \bigl\{t(v-\bar{v});\ t>0,\ v \in K\bigr\}; \qquad T_K(\bar{v}) := \operatorname{cl} \mathcal{R}_K(\bar{v}), \tag{21}$$

where by cl we mean the closure in $H_0^1(\Omega)$. The next lemma is a consequence of a classical result (see, e.g., [5, Lemma 3.7]), and hence, we skip the proof.

Lemma 5.1. *Let $\bar{v}$ be a local solution of* (RP). *Then*

$$DF(\bar{v})\delta v \geqslant 0, \quad \textit{for all } \delta v \in T_K(\bar{v}). \tag{22}$$

Let $H^{-1}(\Omega)$ denote the topological dual of $H_0^1(\Omega)$. The normal cone (in the sense of convex analysis) to K at $v \in K$ is

$$N_K(v) := \bigl\{\lambda \in H^{-1}(\Omega);\ \langle \lambda, v' - v\rangle \leqslant 0,\ \text{for all } v' \in K\bigr\}.$$

We sometimes need the following regularity assumption on the domain Ω and obstacle:

$$\text{For every } f \in L^2(\Omega),\ \text{the solution of } (\mathrm{OP}_f) \text{ belongs to } H^2(\Omega). \tag{23}$$

This holds if $\partial\Omega$ is of class C^2, and under various hypotheses on the obstacle Φ, see [6, Chapter I].

Theorem 5.2. *Let $\bar{v}$ be a local solution of problem* (RP), *and denote by $\bar{h}$ the associated height. Then there exists $\lambda \in N_K(v)$ such that*

$$-\Delta\bar{v} - (\bar{v} - \bar{h})_+ + \lambda = f \quad \textit{in } \Omega. \tag{24}$$

If in addition (23) *holds, then $\bar{v} \in H^2(\Omega)$ and $\lambda \in L^2(\Omega)$.*

Proof. Let $\lambda := -DF(\bar{v})$ (element of $H^{-1}(\Omega)$). Then $DF(\bar{v}) + \lambda = 0$ and, by (22), $\lambda \in N_K(v)$. It follows also from (22) that $\bar{v}$ is a solution of the obstacle problem (without water) with the field $\hat{f} := f + (\bar{v} - h)_+$. Therefore, (23) implies that $\bar{v} \in H^2(\Omega)$. In that case, $\lambda = -DF(\bar{v}) = \Delta\bar{v} + (\bar{v} - \bar{h})_+ + f$ belongs to $L^2(\Omega)$. □

We now discuss some consequences of the theorem.

Lemma 5.3. *Assume that $f \geqslant 0$ and $\Phi \geqslant 0$ a.e. Let v be a solution of problem* (RP). *Then $v \geqslant 0$ a.e.*

Proof. Since v is solution of the obstacle problem with the nonnegative field force $f + (v - h)_+$, and the obstacle is nonnegative, the conclusion follows from Brézis [6, Corollary I.5]. □

We say that $v \in K$ is a *stationary point* of problem (RP) if (24) is satisfied, for some $\lambda \in N_K(v)$, called the associated multiplier.

Proposition 5.4. *Assume that $\Phi = +\infty$. If $\bar{v}$ is a stationary point, with associated height $\bar{h}$, then*

$$2F(\bar{v}) + \bar{h}L + \int_{\Omega} f(\omega)\bar{v}(\omega)\,\mathrm{d}\omega = 0. \tag{25}$$

In particular, if $f = 0$ a.e., then $F(\bar{v}) = -\frac{1}{2}\bar{h}L$, and hence, solutions of (RP) *are the stationary points with largest height of water.*

Proof. It suffices to multiply (24) by $\bar{v}$, and integrate over Ω, to obtain (25), from which the conclusion follows. □

Remark 5.5. Let us highlight the dependence of (RP) over L by denoting it (RP$_L$) in this remark. Whenever f is identically 0 and $\Phi = +\infty$, it is easily checked that $S(\mathrm{RP}_{tL}) = tS(\mathrm{RP}_L)$, for all $t > 0$. Hence the height of water is proportional to L, and $\mathrm{val}(\mathrm{RP}_L) = -aL^2$, where a is a constant of the same sign as $\bar{h}$.

As shown in the following proposition, if $f = 0$ and the obstacle is not present, then the sign of height of water depends only on the amount $\nu_0(\Omega)$ defined in (18).

Proposition 5.6. *Assume that $f = 0$ and $\Phi = +\infty$ a.e. Let $\bar{v}$ be a solution of* (RP), *with associated height $\bar{h}$. Then $\bar{h} \leqslant 0$ iff $\nu_0(\Omega) \geqslant 1$.*

Proof. Let w_0 be the positive eigenvector of unit norm in $L^2(\Omega)$ associated with $\nu_0(\Omega)$, and $\alpha > 0$ be such that $\alpha \int_\Omega w_0(\omega)\,\mathrm{d}\omega = L$. With αw_0 is associated a zero height of water. By Proposition 5.4, $-\frac{1}{2}\bar{h}L \leqslant F(\alpha w_0) = \alpha^2(\nu_0(\Omega) - 1)$. It follows that, if $\bar{h} \leqslant 0$, then $\nu_0(\Omega) \geqslant 1$. Conversely, assume that $\nu_0(\Omega) \geqslant 1$. Let $v \in H_0^1(\Omega)$. Denoting $h = h(v, L)$, get

$$\begin{aligned} F(v) &= \frac{1}{2}\int_{\Omega} \left\|\nabla v(\omega)\right\|^2 \mathrm{d}\omega - \frac{1}{2}\int_{\{v>h\}} \big(v(\omega) - h\big)\big(v(\omega) + h\big)\,\mathrm{d}\omega \\ &\geqslant \frac{1}{2}\int_{\Omega} \left\|\nabla v(\omega)\right\|^2 \mathrm{d}\omega - \frac{1}{2}\int_{\{v>h\}} v^2(\omega)\,\mathrm{d}\omega \geqslant 0 \end{aligned} \tag{26}$$

which proves that val(RP) is nonnegative; since $\mathrm{val(RP)} = -\frac{1}{2}\bar{h}L$ by Remark 6.6(ii), $\bar{h}$ is nonpositive. □

For one-dimensional problems, all computations can be carried out explicitly; see [3].

6. Tangent and normal cone; polyhedricity

In order to state second-order sufficient conditions, and to perform a sensitivity analysis, we need the concept below. We say that K, defined in (1), is *polyhedric* at $\bar{v} \in K$ if, for any $\mu \in N_K(\bar{v})$, the following holds:

$$T_K(\bar{v}) \cap \mu^{\perp} = \mathrm{cl}\big(\mathcal{R}_K(\bar{v}) \cap \mu^{\perp}\big). \tag{27}$$

If this holds for every $\bar{v} \in K$, we say that K is polyhedric. If Φ is identically zero, the next proposition is a particular case of Mignot [16], see also [5, Theorem 3.58].

Proposition 6.1. *The set K is polyhedric at $\bar{v}$.*

Proof. Since $T_K(v) \supset \mathcal{R}_K(v)$, and the left-hand side of (27) is closed, we have that the right-hand side is included in the left-hand side. Let us prove the converse. Given $w \in H_0^1(\Omega)$, set $w_- := \min(0, w)$, and $w_+ := \max(0, w)$. Observe first that if $w \in \mathcal{R}_K(v)$, then $w_+ \in \mathcal{R}_K(v)$, since if $v + tw \in K$ for a given $t > 0$, we have that

$$v + tw_+ = v + \max(tw, 0) = \max(v + tw, v) \leqslant \Phi \quad \text{a.e.} \tag{28}$$

Assume now that $w \in T_K(v)$, then w is the limit of a sequence $w_n \in \mathcal{R}_K(v)$, and hence, $w_+ = \lim_n (w_n)_+$ is limit of elements of $\mathcal{R}_K(v)$. We have proved that $w \in T_K(v)$ implies $w_+ \in T_K(v)$. Let $\mu \in N_K(v)$. Since $w_+ \in T_K(v)$, and also $-w_+ \in \mathcal{R}_K(v)$, we have that $w_+ \perp \mu$. Finally, let $w \in T_K(\bar{v}) \cap \mu^{\perp}$, where $\mu \in N_K(v)$. Since $w_+ \perp \mu$, we have that $w_- \perp \mu$. Let $\widehat{w}_n$ be a sequence in $\mathcal{R}_K(v)$ converging to w_+. Then $(\widehat{w}_n)_+$ also converges to w_+ and, by the above claims, belongs to $\mathcal{R}_K(v) \cap \mu^{\perp}$. Therefore, $w_- + (\widehat{w}_n)_+$ belongs to $\mathcal{R}_K(v) \cap \mu^{\perp}$ and converges to w. The conclusion follows. □

We need some classical results, see, e.g., [5, Section 6.4] and references therein. Denote by $M(\Omega)$ the set of locally finite Borel measures, which is the dual, for an appropriate topology, of the space $C_{00}(\Omega)$ of continuous functions with compact support in Ω. Let $M(\Omega)_+$ be the set of nonnegative locally finite Borel measures. Also, denote by $H_0^1(\Omega)_+$ the set of functions in $H_0^1(\Omega)$ that are nonnegative a.e., and by $H^{-1}(\Omega)_+$ the set:

$$H^{-1}(\Omega)_+ := \big\{\mu \in H^{-1}(\Omega);\ \langle \mu, v\rangle \geqslant 0,\ \text{for all } v \in H_0^1(\Omega)\big\}.$$

A set $A \subset \Omega$ is said to be of *null capacity* if there exists a sequence $u_k \to 0$ in $H_0^1(\Omega)$, such that for each k, $u_k \geqslant 1$ over a neighborhood of A. It is easily checked that a set of null capacity has zero measure, but the converse is false. Let $v \in H_0^1(\Omega)$. Then v is in fact a class of functions under the relation of being equal a.e.; in this class there exists an element that is continuous except on a set of null capacity, called the quasi-representative.

Lemma 6.2. *Let $\mu \in H^{-1}(\Omega)_+$. Then μ has a unique extension, also denoted μ, from $H_0^1(\Omega) \cap C_{00}(\Omega)$ to $C_{00}(\Omega)$. The latter belongs to $M(\Omega)_+$. In addition, if $f \in H_0^1(\Omega)$, with quasi-representative $\tilde{f}$, then $\tilde{f} \in L^1(\mu)$, and*

$$\int_\Omega \tilde{f}(\omega)\,\mathrm{d}\mu(\omega) = \langle \mu, f \rangle_{H^{-1}(\Omega), H_0^1(\Omega)}. \tag{29}$$

In the sequel we identify functions of $H_0^1(\Omega)$ with their quasi-representatives. We say that a property is true quasi-everywhere, or q.e., if it is true everywhere except on a set of null capacity.

Proposition 6.3. *Let $v \in K$. If $\Phi \in H_0^1(\Omega)$, then the following equalities hold*:

$$N_K(\bar{v}) = \bigl\{\mu \in H^{-1}(\Omega)_+;\ \mu\bigl(\{\bar{v} < \Phi\}\bigr) = 0\bigr\}, \tag{30}$$

$$T_K(\bar{v}) = \bigl\{v \in H_0^1(\Omega);\ v \leqslant 0 \text{ q.e. on } \{\bar{v} = \Phi\}\bigr\}. \tag{31}$$

Proof. It is clear that $\mu \in N_K(v)$ is equivalent to $\sigma_K(\mu) = \langle \mu, v\rangle$, where the support function σ_K is defined by $\sigma_K(\mu) := \sup\{\langle \mu, w\rangle;\ w \in K\}$. If $\Phi \in H_0^1(\Omega)$, then $\Phi \in K$, and hence,

$$\sigma_K(\mu) = \begin{cases} \langle \mu, \Phi\rangle & \text{if } \mu \in H^{-1}(\Omega)_+, \\ +\infty & \text{otherwise.} \end{cases} \tag{32}$$

In that case, we have that $\mu \in N_K(v)$ iff $\mu \in H^{-1}(\Omega)_+$ and $\langle \mu, \Phi - v\rangle = 0$. Since $\mu \in H^{-1}(\Omega)_+$ and $\Phi - v \geqslant 0$, this is equivalent to (30). For proving (31) we use the fact that a Borel set $A \subset \Omega$ has null capacity iff $\mu(A) = 0$, for all $\mu \in H^{-1}(\Omega)_+$ (see, e.g., Lemma 6.55 in [5]). Therefore, v is in the r.h.s. of (31) iff each $\mu \in H^{-1}(\Omega)_+$ with support in $\{v = \Phi\}$ is such that

$$\langle \mu, v - \bar{v}\rangle = \int_{\{v=\Phi\}\cap\{v \leqslant \bar{v}\}} (v - \bar{v})\,\mathrm{d}\mu \leqslant 0. \tag{33}$$

This is the characterization of $T_K(\bar{v})$, since the latter is the polar cone of $N_K(v)$. □

We have seen in Proposition 5.4 that stationary points satisfy a certain integral relation, if $\Phi = +\infty$. Let us extend this kind of result to the case when the obstacle is active.

Corollary 6.4. *Assume that $\Phi \in H_0^1(\Omega)$. Let $\bar{v}$ be a stationary point of problem* (RP) *and λ its associated multiplier. Denote by $\bar{h}$ the associated height. Then* (*the duality product below being in the $H_0^1(\Omega)$ space*)

$$2F(\bar{v}) + \bar{h}L + \int_\Omega f(\omega)\bar{v}(\omega)\,\mathrm{d}\omega + \langle \lambda, \Phi\rangle = 0. \tag{34}$$

Proof. Multiplying (24) by $\bar{v}$ and integrating over Ω, we obtain, after some elementary computations:

$$2F(\bar{v}) + \bar{h}L + \int_{\Omega} f(\omega)\bar{v}(\omega)\,\mathrm{d}\omega + \langle \lambda, \bar{v}\rangle = 0. \tag{35}$$

It remains to observe that $\langle \lambda, \Phi - \bar{v}\rangle = 0$. Indeed this quantity is nonpositive since $\Phi \in K$ and $\lambda \in N_K(\bar{v})$. On the other hand, $\lambda \in H^{-1}(\Omega)_+$, while $\bar{v} \leqslant \Phi$, hence this amount is nonnegative. The conclusion follows. □

Remark 6.5. The conclusion still holds if we assume only that v and Φ are continuous. In that case we apply Lemma 5.1 with δv a smooth function with support in the set $\{v < \Phi\}$. Therefore, the support of the measure λ belongs to $\{v = \Phi\}$, and hence, $\int_\Omega \lambda(\omega)(\Phi(\omega) - \bar{v}(\omega))\,\mathrm{d}\omega = 0$ still holds, from which (34) follows.

Remark 6.6. Under the assumptions of the above corollary, since $\lambda \geqslant 0$ a.e., if $\Phi \geqslant 0$ a.e., the last term in (34) is nonnegative, and hence, if $\Phi \in H_0^1(\Omega)$, we have that $2F(\bar{v}) + \bar{h}L + \int_\Omega f(\omega)\bar{v}(\omega)\,\mathrm{d}\omega \leqslant 0$, with equality if $\bar{v} < \Phi$ quasi everywhere.

7. Second-order optimality conditions

Although the cost function is not twice differentiable, it is possible to state second-order necessary or sufficient conditions for optimality, thanks to the following pseudo-Taylor expansion in the lemma below.

Lemma 7.1. *Let $H : L^4(\Omega) \to \mathbb{R}$ be defined by $H(v) := \frac{1}{2}\int_\Omega v_+^2(\omega)\,\mathrm{d}\omega$. Then the expansion below holds, for all v and z in $L^4(\Omega)$:*

$$\begin{aligned} H(v+z) = H(v) &+ \int_{\Omega} v_+(\omega)z(\omega)\,\mathrm{d}\omega + \frac{1}{2}\int_{\{v=0\}} z_+^2(\omega)\,\mathrm{d}\omega + \frac{1}{2}\int_{\{v>0\}} z^2(\omega)\,\mathrm{d}\omega \\ &+ \mathrm{o}\big(\|z\|_{L^4(\Omega)}^2\big). \end{aligned} \tag{36}$$

Proof. Let us set:

$$\begin{aligned} A := H(v+z) - H(v) &- \int_{\Omega} v_+(\omega)z(\omega)\,\mathrm{d}\omega \\ &- \frac{1}{2}\int_{\{v=0\}} z_+^2(\omega)\,\mathrm{d}\omega - \frac{1}{2}\int_{\{v>0\}} z^2(\omega)\,\mathrm{d}\omega. \end{aligned} \tag{37}$$

We have to check the equality $A = \mathrm{o}(\|z\|_{L^4(\Omega)}^2)$. Observe that

$$A = \int\limits_{\{v<0;\, v+z>0\}} \big(v(\omega)+z(\omega)\big)^2\,\mathrm{d}\omega - \int\limits_{\{v>0;\, v+z<0\}} \big(v(\omega)+z(\omega)\big)^2\,\mathrm{d}\omega, \tag{38}$$

and hence, denoting by ξ_z the indicator function of $\{v<0;\ v+z>0\}$, and using the Cauchy–Schwarz inequality, get

$$A \leqslant \int\limits_{\{v<0;\, v+z>0\}} z^2(\omega)\,\mathrm{d}\omega \leqslant \|\xi_z\|_{L^2(\Omega)}\|z\|^2_{L^4(\Omega)}. \tag{39}$$

By Lebesgue's dominated convergence theorem, if $z\to 0$ in $L^4(\Omega)$, $\xi_z \to 0$ in $L^2(\Omega)$. With (38), it follows that $A \leqslant \mathrm{o}(\|z\|^2_{L^4(\Omega)})$. The opposite inequality can be obtained in the same manner. □

Thanks to the above lemma, we are able to state a pseudo-Taylor expansion for the cost function of problem (RP). We remind that the expression of DF is given in (17). Given $v \in H_0^1(\Omega)$, define $Q_v : H_0^1(\Omega) \to \mathbb{R}$ by:

$$Q_v(z) := \int\limits_{\Omega} \big\|\nabla z(\omega)\big\|^2\,\mathrm{d}\omega - \int\limits_{\{v=h(v,L)\}} \hat{z}_+^2(\omega)\,\mathrm{d}\omega - \int\limits_{\{v>h(v,L)\}} \hat{z}^2(\omega)\,\mathrm{d}\omega, \tag{40}$$

where $\hat{z} \in H^1(\Omega)$ is defined by $\hat{z}(\omega) := z(\omega) - \delta h$, δh being the directional derivative of $h(v,L)$ at $\bar{v}$ in direction z (whose expression is given in Lemma 2.2).

Lemma 7.2. *Let $\bar{v}$ and z belong to $H_0^1(\Omega)$, and denote $\bar{h}$ the height associated with $\bar{v}$. Then the following expansion holds*:

$$F(v+z) = F(v) + DF(v)z + \frac{1}{2}Q_v(z) + \mathrm{o}\big(\|z\|^2_{H_0^1(\Omega)}\big). \tag{41}$$

Proof. Given $(v,L) \in H_0^1(\Omega)\times\mathbb{R}_{++}$, set $h = h(v,L)$. Let $D_h J(v,h,L)$ denote the partial derivative of J with respect to h. By Proposition 2.1, $D_h J(v,h,L) = 0$. Combining with Lemma 7.1, we have that, for every $(z,\delta h) \in H_0^1(\Omega)\times\mathbb{R}$:

$$\begin{aligned} J(v+z, h+\delta h, L) = J(v,h,L) + D_v J(v,h,L)z + \frac{1}{2}\int\limits_{\Omega} \big\|\nabla z(\omega)\big\|^2\,\mathrm{d}\omega \\ - \frac{1}{2}\int\limits_{\{v=h\}} \big(z(\omega)-\delta h\big)_+^2\,\mathrm{d}\omega - \frac{1}{2}\int\limits_{\{v>h\}} \big(z(\omega)-\delta h\big)^2\,\mathrm{d}\omega \\ + \mathrm{o}\big(\|z\|^2_{H_0^1(\Omega)} + (\delta h)^2\big). \end{aligned}$$

The result follows by combining with (7). □

We can now state the second-order necessary conditions for local optimality. The *cone of critical directions* is defined by:

$$C(\bar{v}) := \big\{z \in T_K(\bar{v});\ DF(\bar{v})z = 0\big\}. \tag{42}$$

In the analysis we also use the *cone of feasible critical directions*,

$$\mathcal{C}(\bar{v}) := \big\{z \in \mathcal{R}_K(\bar{v});\ DF(\bar{v})z = 0\big\}. \tag{43}$$

Since the set K is *polyhedric* at $\bar{v}$, we know that $C(\bar{v})$ is the closure of $\mathcal{C}(\bar{v})$.

Theorem 7.3. *Let $\bar{v}$ be a local solution of* (RP), *and $\bar{h}$ the associated height. Then $Q_{\bar{v}}(z) \geqslant 0$, for all critical direction z.*

Proof. Let z be a feasible critical direction. By local optimality of $\bar{v}$, and using Lemma 7.2, get $0 \leqslant \lim_{t\downarrow 0}(\frac{1}{2}t^2)^{-1}(F(\bar{v}+tz) - F(\bar{v})) = Q_{\bar{v}}(z)$. Since $Q_{\bar{v}}(\cdot)$ is continuous, we also have that $Q_{\bar{v}}(\cdot)$ is nonnegative over the closure of $\mathcal{C}(\bar{v})$; the latter being equal to $C(\bar{v})$ since K is polyhedric, the conclusion follows. □

We now turn to the second-order sufficient conditions for local optimality. A first step is the following lemma:

Lemma 7.4. *The positively homogeneous form of second-order $Q_{\bar{v}}(\cdot)$, stated in* (40), *is an extended Legendre form in the sense of* [5, Section 3.3], *i.e., is weakly lower semi continuous and such that, if a sequence z_k weakly converges to z in $H_0^1(\Omega)$, and $Q_{\bar{v}}(z_k) \to Q_{\bar{v}}(z)$, then $z_k \to z$ strongly in $H_0^1(\Omega)$.*

Proof. We can write $Q_{\bar{v}}(z)$ as $\|z\|^2_{H_0^1(\Omega)} + q(z)$, where $q(\cdot)$ is, by (10) and since $h(v, L)$ has continuous directional derivatives, continuous for the weak topology. Therefore, if z_k weakly converges to z in $H_0^1(\Omega)$, and $Q_{\bar{v}}(z_k) \to Q_{\bar{v}}(z)$, then $\|z_k\|^2_{H_0^1(\Omega)} \to \|z\|^2_{H_0^1(\Omega)}$, which in turn implies $z_k \to z$ in $H_0^1(\Omega)$, as was to be proved. □

Theorem 7.5. *Let $\bar{v} \in K$, and let $\bar{h}$ be the associated height. Assume the following second-order sufficient condition*: *for every nonzero critical direction z, $Q_{\bar{v}}(z) > 0$. Then $\bar{v}$ is a local solution of* (RP), *satisfying the quadratic growth condition*: *there exists $\alpha > 0$ such that, for all $v' \in K$*:

$$F(v') \geqslant F(\bar{v}) + \alpha\|v' - \bar{v}\|^2_{H_0^1(\Omega)} + \mathrm{o}\big(\|v' - \bar{v}\|^2_{H_0^1(\Omega)}\big). \tag{44}$$

Proof. Although this is a variant of the proof of Theorem 3.63 combined with Proposition 3.74 of [5], it is useful to give a direct argument. If the conclusion were false, there would exist sequences $v_k \to \bar{v}$ in $H_0^1(\Omega)$, and $\varepsilon_k \downarrow 0$, such that

$$F(v_k) < F(\bar{v}) + \varepsilon_k\|v_k - \bar{v}\|^2_{H_0^1(\Omega)}. \tag{45}$$

Set $t_k := \|v_k - \bar{v}\|_{H_0^1(\Omega)}$, and $\delta v_k := t_k^{-1}(v_k - \bar{v})$. Then $\|\delta v_k\|_{H_0^1(\Omega)} = 1$, and $v_k = \bar{v} + t_k \delta v_k$. Extracting if necessary a subsequence, we may assume that δv_k has a weak limit δv; obviously $\delta v \in T_K(\bar{v})$. From a first-order expansion of F in (45), we deduce that $DF(\bar{v})\delta v \leqslant 0$, and hence, δv is a critical direction. Since $DF(\bar{v})\delta v_k \geqslant 0$ by the first-order optimality conditions, we have by Lemma 7.2:

$$\begin{aligned} F(v_k) = F(\bar{v} + t_k\delta v_k) &= F(\bar{v}) + t_k DF(\bar{v})\delta v_k + \frac{1}{2}t_k^2 Q_{\bar{v}}(\delta v_k) + \mathrm{o}\big(t_k^2\big), \\ &\geqslant F(\bar{v}) + \frac{1}{2}t_k^2 Q_{\bar{v}}(\delta v_k) + \mathrm{o}\big(t_k^2\big). \end{aligned}$$

Combining with (45), obtain $Q_{\bar{v}}(\delta v_k) \leqslant \mathrm{o}(1)$. Since $Q_{\bar{v}}(\cdot)$ is an extended Legendre form, it follows that $Q_{\bar{v}}(\delta v) \leqslant 0$, with equality implying $\delta v_k \to \delta v$ strongly. In the latter case δv is a nonzero critical direction such that $Q_{\bar{v}}(\delta v) \leqslant 0$: this contradicts the second-order sufficient conditions. Similarly, by the second-order necessary conditions, $Q_{\bar{v}}(\delta v) < 0$ is impossible. We have obtained the desired contradiction. □

Note that, by Lemma 4.2(ii), the second-order sufficient optimality condition trivially holds if $\nu_0(\Omega) > 1$ or $\nu_1(\Omega) > 1$.

8. Sensitivity analysis

It is possible to perform a sensitivity analysis with respect to the volume of water L and the field of forces f; for the sake of simplicity we will only study the dependence of solutions with respect to L. For that reason we denote the cost function as $F(v, L) = J(v, h(v, L), L)$, and the minimization problem as

$$\operatorname*{Min}_{v \in K} F(v, L), \tag{P$_L$}$$

its value being denoted val(L). Denote also by $S_+(\mathrm{P}_L)$ (respectively $S_-(\mathrm{P}_L)$) the set of solutions of (P_L) with *maximum* (*minimum*) height of water. Similarly, let $(z, \ell) \in H_0^1(\Omega) \times \mathbb{R}$. Let δh denote in this section the directional derivative of $h(v, L)$ at (v, L) in direction (z, ℓ), solution of (6). Let

$$Q_{v,L}(z, \ell) := \int_{\Omega} \big\|\nabla z(\omega)\big\|^2 \,\mathrm{d}\omega - \int_{\{\bar{v}=\bar{h}\}} \hat{z}_+^2(\omega)\,\mathrm{d}\omega - \int_{\{\bar{v}>\bar{h}\}} \hat{z}^2(\omega)\,\mathrm{d}\omega - 2\ell\delta h, \tag{46}$$

where $\hat{z}(\omega) := z(\omega) - \delta h$. An easy variant of the proof of Lemma 7.2 allows to prove that

$$\begin{aligned} F(v + z, L + \ell) = F(v, L) + D_v F(v, L)z - h(v, L)\ell + \frac{1}{2}Q_{v,L}(z, \ell) \\ + \mathrm{o}\big(\|z\|_{H_0^1(\Omega)}^2 + \ell^2\big). \end{aligned} \tag{47}$$

Denote the critical cone as

$$C(v,L) := \{z \in T_K(v);\ D_v F(v,L)z = 0\}. \tag{48}$$

Consider the subproblem associated with $\bar{v} \in K$ and $L > 0$:

$$\operatorname*{Min}_{z \in C(\bar{v},L)} Q_{\bar{v},L}(z,\ell). \tag{SP$_\ell$}$$

Below $\mathrm{s}(\ell)$ denotes the sign of ℓ, with value 1 (respectively -1) if ℓ is positive (respectively negative). Note that $\mathrm{val}(\mathrm{SP}_0) = 0$ in view of the second-order necessary optimality condition, and for $\ell \neq 0$, due to positive homogeneity,

$$\mathrm{val}(\mathrm{SP}_\ell) = \ell^2 \,\mathrm{val}(\mathrm{SP}_{\mathrm{s}(\ell)}); \qquad S(\mathrm{SP}_\ell) = |\ell| S(\mathrm{SP}_{\mathrm{s}(\ell)}). \tag{49}$$

Theorem 8.1. (i) *When $\ell \to 0$, the weak limit points of solutions of* $(\mathrm{P}_{L+\ell})$, *for $\ell > 0$* (*respectively $\ell < 0$*) *are strong limit points, and belong to $S_+(\mathrm{P}_L)$* (*respectively $S_-(\mathrm{P}_L)$*). *In addition, the following expansion of value function holds*:

$$\mathrm{val}(L+\ell) = \mathrm{val}(L) - \hat{h}\ell + \mathrm{o}(\ell), \tag{50}$$

where $\hat{h}$ is the maximum (*respectively minimum*) *height of water among all solutions of (P_L) if $\ell > 0$* (*respectively $\ell < 0$*).

(ii) *Assume that $\ell > 0$* (*respectively $\ell < 0$*), *and that $S_+(\mathrm{P}_L)$* (*respectively $S_-(\mathrm{P}_L)$*) *has a unique element $\bar{v}$ satisfying the second-order sufficient condition. Then, if $v_\ell \in S(\mathrm{SP}_{L+\ell})$, we have that*

$$\|v_\ell - \bar{v}\|_{H_0^1(\Omega)} = \mathrm{O}(\ell), \tag{51}$$

and the following expansion holds for the value function:

$$\mathrm{val}(L+\ell) = \mathrm{val}(L) - \hat{h}\ell + \frac{1}{2}\mathrm{val}(\mathrm{SP}_{\mathrm{s}(\ell)})\ell^2 + \mathrm{o}(\ell^2). \tag{52}$$

In addition, any weakly convergent subsequence in $H_0^1(\Omega)$ of $(v_{L+\ell} - v_L)/\ell$ is in fact strongly convergent, and its limit is solution of $(\mathrm{SP}_{\mathrm{s}(\ell)})$. *If* $(\mathrm{SP}_{\mathrm{s}(\ell)})$ *has a unique solution $\bar{z}$, then the following expansion of solutions holds*:

$$v_{L+\ell} = v_L + |\ell|\bar{z} + \mathrm{o}(\ell). \tag{53}$$

Proof. (i) Assume for instance that $\ell > 0$, and let $\hat{h}$ denote the maximum height of water. Since the set of solutions is a nonempty, weakly closed and bounded subset of $H_0^1(\Omega)$, $S_+(\mathrm{P}_L)$ is itself nonempty, weakly closed and bounded. Given $v \in S_+(\mathrm{P}_L)$, we have with (47) and (49),

$$\mathrm{val}(L+\ell) \leqslant F(v, L+\ell) = F(v,L) - \hat{h}\ell + \mathrm{o}(\ell). \tag{54}$$

It remains to prove the converse inequality. Take a sequence $\ell_k \downarrow 0$, along which $\lim_k(\mathrm{val}(L+\ell_k)-\mathrm{val}(L))/\ell_k$ attains the smallest possible value, say Δ. By (54), $\Delta \leqslant -\hat{h}$. Let $v_k \in S(\mathrm{P}_{L+\ell_k})$. Extracting a subsequence if necessary, we may assume that v_k has a weak limit point $\bar{v} \in K$. Passing to the limit in the inequality

$$F(v_k, L+\ell_k) \leqslant F(v, L+\ell_k), \quad \text{for all } v \in K, \tag{55}$$

thanks to the l.s.c. of F, we deduce that $\bar{v} \in S(\mathrm{P}_L)$. Taking $v = \bar{v}$ in (55), we obtain $\limsup_k F(v_k, L+\ell_k) \leqslant F(\bar{v}, L)$, which since F is l.s.c. implies $F(v_k, L+\ell_k) \to F(\bar{v}, L)$. In view of the expression of F, this implies $v_k \to \bar{v}$ in $H_0^1(\Omega)$. Since F is continuously, and hence strictly differentiable, we have that

$$\Delta \geqslant \lim_k \frac{F(v_k, L+\ell_k) - F(v_k, L)}{\ell_k} = -\bar{h}, \tag{56}$$

where $\bar{h}$ is the height of water associated with $\bar{v}$, and hence, $\Delta \geqslant -\bar{h}$. Since $\bar{h} \leqslant \hat{h}$, this implies $\Delta = \bar{h} = \hat{h}$, and also that each (strong) limit point of v_k is solution of $S_+(\mathrm{P}_L)$, as was to be proved.

(ii) Assume for instance that $\ell > 0$. Note that, by the second-order sufficient condition, a minimizing sequence of (SP_1) is bounded. Since the cost function is l.s.c. and the feasible set is weakly closed, this implies that $S(\mathrm{SP}_1)$ is nonempty and bounded. Since K is polyhedric, for any $\varepsilon > 0$, there exists $z_\varepsilon \in C(v, L) \cap \mathcal{R}_K(v)$ that is an ε-solution of (SP_1). It follows that, for $\ell > 0$ small enough,

$$\begin{aligned}\mathrm{val}(L+\ell) &\leqslant F(\bar{v}+\ell z_\varepsilon, L+\ell) = F(\bar{v}, L) - \hat{h}\ell + \frac{1}{2}Q_{\bar{v},L}(z_\varepsilon, 1)\ell^2 + \mathrm{o}\big(\ell^2\big) \\ &\leqslant \mathrm{val}(L) - \hat{h}\ell + \frac{1}{2}\big(\mathrm{val}(\mathrm{SP}_1) + \varepsilon\big)\ell^2 + \mathrm{o}\big(\ell^2\big).\end{aligned} \tag{57}$$

Since ε can be arbitrarily small we deduce that

$$\mathrm{val}(L+\ell) \leqslant F(\bar{v}, L) - \hat{h}\ell + \frac{1}{2}\,\mathrm{val}(\mathrm{SP}_1)\ell^2 + \mathrm{o}\big(\ell^2\big). \tag{58}$$

We will prove the converse inequality and (51). Given any sequence $\ell_k \downarrow 0$, by (i), the associated sequence $v_k \in S(\mathrm{SP}_{L+\ell_k})$ converges to $\bar{v}$. Let $v_\ell \in S(\mathrm{P}_{L+\ell})$. In view of the expansion (47) and the second-order sufficient condition (Theorem 7.5), setting $z_\ell := v_\ell - \bar{v}$, we get an estimate of the form:

$$F(\bar{v}+z_\ell, L+\ell) \geqslant \mathrm{val}(L) + D_v F(v, L)z_\ell - \hat{h}\ell + \frac{1}{2}\alpha\|z_\ell\|^2_{H_0^1(\Omega)} - \beta\|z_\ell\|_{H_0^1(\Omega)}|\ell|, \tag{59}$$

for some $\beta > 0$. Combining with (58), we deduce that $\|z_\ell\|_{H_0^1(\Omega)} = \mathrm{O}(|\ell|)$, which proves (51).

Assume now that the sequence $(\mathrm{val}(L+\ell_k) - \mathrm{val}(L) + \bar{h}L)/\ell_k^2$ attains its smallest possible value. By (51), $z_k := (v_k - \bar{v})/\ell_k$ is bounded. Extracting if necessary a subsequence, we may assume that it has a weak limit $\bar{z}$. Since $z_k \in \mathcal{R}_K(\bar{v})$, $\bar{z} \in T_K(\bar{v})$. Using (50), obtain $D_v F(\bar{v}, L)\bar{z} \leqslant 0$. It follows that $\bar{z} \in C(\bar{v}, L)$. Since $Q_{\bar{v},L}(\cdot, \ell)$ is l.s.c., we have with (47) that

$$\begin{aligned}\operatorname{val}(L+\ell_k) = F(\bar{v}+\ell_k z_k, L+\ell_k) &= F(\bar{v}, L) - \bar{h}\ell_k + \frac{1}{2}Q_{\bar{v},L}(z_k, 1)\ell_k^2 + \mathrm{o}\big(\ell_k^2\big)\\ &\geqslant F(\bar{v}, L) - \bar{h}\ell_k + \frac{1}{2}Q_{\bar{v},L}(\bar{z}, 1)\ell_k^2 + \mathrm{o}\big(\ell_k^2\big)\\ &\geqslant F(\bar{v}, L) - \bar{h}\ell_k + \frac{1}{2}\operatorname{val}(\mathrm{SP}_1)\ell_k^2 + \mathrm{o}\big(\ell_k^2\big),\end{aligned} \tag{60}$$

which combined with (58) implies (52), as well as $\bar{z} \in S(\mathrm{SP}_{\mathrm{s}(\ell)})$, as was to be proved. □

9. Numerical approximation of solutions

In this section we give a basic discussion of the discretization of problem (RP) in the case when Ω is a convex polygon of $\mathbb{R}^2$ (although in our numerical results we deal also with the case when Ω is a disc). A basic reference for the numerical analysis of variational inequalities is the book by Glowinski et al. [12]. These authors deal with convex problems. Here, due to nonconvexity, we have to rely on the local analysis for obtaining error estimates. Consider a family of regular triangulation of Ω. That is, with each $\varepsilon > 0$ we associate a finite family $\mathcal{T}_\varepsilon$ of triangles whose union is equal to Ω, and such that (i) the intersection of two of these triangles is either empty, or is a vertex, or a common side, (ii) the diameter of each triangle is not larger than ε, and (iii) if r_ε denotes the smallest radius of the circle inscribed in a triangle, then $\lim_{\varepsilon\downarrow 0} r_\varepsilon/\varepsilon > 0$. Denote by V_ε the finite-dimensional space of continuous functions that are affine on each triangle, and vanish on $\partial\Omega$; we have that $V_\varepsilon \subset H_0^1(\Omega)$. Let $K_\varepsilon := K \cap V_\varepsilon$. We will study the approximate reduced problem (to be compared to problem (RP), stated in Section 2),

$$\operatorname*{Min}_{v} F(v); \quad v \in K_\varepsilon. \tag{RP$_\varepsilon$}$$

In this section we assume that K_ε is an approximation of K in the following sense (same hypothesis as in [12, Section 4.3]):

$$\begin{cases}\text{(i)} & \text{every } v \in K \text{ is a strong limit of } v_\varepsilon \in K_\varepsilon,\\ \text{(ii)} & \text{any weak limit point of } v_\varepsilon \in K_\varepsilon \text{ belongs to } K.\end{cases} \tag{61}$$

Point (ii) always holds since $K_\varepsilon \subset K$, and K is closed and convex. Point (i) holds, for instance, if Φ is continuous, and nonnegative on a neighborhood of $\partial\Omega$.

Theorem 9.1. (i) *The set of solutions of* (RP$_\varepsilon$) *is nonempty, and uniformly bounded* (*for* $\varepsilon > 0$ *small enough*), *and the following inequalities hold*:

$$\operatorname{val}(\mathrm{RP}) \leqslant \operatorname{val}(\mathrm{RP}_\varepsilon) \leqslant \operatorname{val}(\mathrm{RP}) + \mathrm{o}(1). \tag{62}$$

(ii) *Let* $\bar{v}$ *be a solution of* (RP). *Then*

$$\bigl|\mathrm{val}(\mathrm{RP}_\varepsilon) - \mathrm{val}(\mathrm{RP})\bigr| \leqslant \mathrm{O}\bigl(\mathrm{dist}(\bar{v}, K_\varepsilon)\bigr). \tag{63}$$

(iii) *Any weak limit point* $\bar{v}$ *of* $v_\varepsilon \in S(\mathrm{RP}_\varepsilon)$ *is a strong limit, and belongs to* $S(\mathrm{RP})$.

Proof. Let $\hat{v} \in S(\mathrm{RP})$, and let $\hat{v}_\varepsilon$ be the orthogonal projection of $\hat{v}$ onto K_ε (in the space $H_0^1(\Omega)$). Denote by L_F a Lipschitz constant of F near $\hat{v}$. We have that

$$\mathrm{val}(\mathrm{RP}_\varepsilon) \leqslant F(\hat{v}_\varepsilon) \leqslant F(\hat{v}) + L_F\|\hat{v}_\varepsilon - \hat{v}\|_{H_0^1(\Omega)} = \mathrm{val}(\mathrm{RP}) + L_F\|\hat{v}_\varepsilon - \hat{v}\|_{H_0^1(\Omega)}. \tag{64}$$

By (61)(i), $\|\hat{v}_\varepsilon - \hat{v}\|_{H_0^1(\Omega)} \to 0$. The second inequality in (62) follows, while the first is due to the fact that (RP) and $(\mathrm{RP}_\varepsilon)$ have the same cost function, whereas $F(\mathrm{RP}) \supset F(\mathrm{RP}_\varepsilon)$. Combining with the lower estimate of F in (13), and standard arguments on bounded minimizing sequences, it follows that the set of solutions of $(\mathrm{RP}_\varepsilon)$ is nonempty and, for $\varepsilon > 0$ small enough, uniformly bounded. Relation (63) is a consequence of (64) and (i). In addition, any weak limit $\bar{v}$ is such that $F(\bar{v})$ is the limit of the corresponding sequence $F(v_{\varepsilon_k})$, which in view of the expression of F implies that the subsequence strongly converges; this proves (iii). □

Corollary 9.2. *Assume the problem to be without obstacle, and* (RP) *to have a unique solution* $\bar{v} \in H^2(\Omega)$. *Let* v_ε *denote a solution of problem* $(\mathrm{RP}_\varepsilon)$. *Then* $v_\varepsilon \to \bar{v}$ *in* $H_0^1(\Omega)$. *If in addition* $\bar{v}$ *satisfies the second-order sufficient condition, then* $\|v_\varepsilon - \bar{v}\|_{H_0^1(\Omega)} = \mathrm{O}(\varepsilon^{1/2})$.

Proof. The first statement is a consequence of Theorem 9.1(iii). Since there is no obstacle, a classical result is that the distance of $\bar{v}$ to K_ε (in the norm of $H_0^1(\Omega)$) is $\mathrm{O}(\varepsilon)$. By Theorem 7.5, if $v_\varepsilon \in S(\mathrm{RP}_\varepsilon)$, we have that for some $\alpha > 0$,

$$\mathrm{val}(\mathrm{RP}_\varepsilon) = F(v_\varepsilon) \geqslant \mathrm{val}(\mathrm{RP}) + \alpha\|v_\varepsilon - \bar{v}\|^2_{H_0^1(\Omega)} + \mathrm{o}\bigl(\|v_\varepsilon - \bar{v}\|^2_{H_0^1(\Omega)}\bigr). \tag{65}$$

Combining this with (63), the conclusion follows. □

Remark 9.3. (i) This type of proof allows to obtain the same conclusion (under the assumption of a unique solution $\bar{v}$ satisfying the second-order sufficient condition) if the obstacle is such that the distance (in the norm of $H_0^1(\Omega)$) from $\bar{v}$ to K_ε is still $\mathrm{O}(\varepsilon)$. This is the case, for instance, if Φ is constant and nonnegative, since the operation of taking the punctual minimum of two functions is Lipschitz in $H_0^1(\Omega)$.

(ii) The result is to be compared with the $\mathrm{O}(\varepsilon^{1/2})$ error estimate obtained for the standard obstacle problem in [12, Proposition 4.1], whereas for the Laplace equation we have an $\mathrm{O}(\varepsilon)$ error estimate, see [18]. It would be interesting to identify specific situations when the $\mathrm{O}(\varepsilon)$ error estimate holds for the problem studied in this paper. This probably requires some strong form of second-order sufficient conditions as those presented in [5].

10. Decomposition algorithms

In this section we discuss how to solve the discretized problem $(\mathrm{RP}_\varepsilon)$. There are several ways to do this. If the obstacle is present, it may be convenient to approximate the constraint $v \leqslant \Phi$, for instance by upper bounds on the value of the deformation $v \in V_\varepsilon$ only at the nodes of the triangulation. This upper bound may be the value of Φ at these nodes, or an average value of Φ in a neighboring region. Or we may keep the constraint $v \leqslant \Phi$ everywhere, which means that we have to solve a semiinfinite programming problem (see, e.g., [5, Section 5.4]). In this paper we will not go into the details of discretization of the constraint, but rather discuss how to design a decomposition algorithm for solving the problem. If the discretized problem has upper bounds only at nodes of the triangulation, then it reduces to the minimization of a continuously differentiable cost function with upper bounds on the variables. There are efficient algorithms for this, even for large scale problems, such as limited memory quasi-Newton algorithms with projections, and interior-point algorithms, see, e.g., Bertsekas [2], Bonnans et al. [4], or Nocedal and Wright [17]. However, in view of the integration of such algorithms in the software for mechanical design, it may be desirable to state an algorithm whose essential step is to solve a classical obstacle problem. Such an algorithm is already available in many of these softwares. Another desirable property is that the algorithm behaves well when the discretization parameter ε vanishes. A favorable situation is when the algorithm makes sense for the original (nondiscretized) problem, if we can prove that, for small ε, the sequence computed by the algorithm applied to $(\mathrm{RP}_\varepsilon)$ is close to the one for problem (RP). Such a property is not easy to prove. In this section we will design an algorithm which at least makes sense for the original problem. To this end, consider the following reformulation of problem (RP):

$$\operatorname*{Min}_{v,g} \mathcal{F}(v,g); \quad v \in K;\ g \in \mathcal{K}, \tag{RFRP}$$

where we set $\mathcal{K} = \{g \in L^2(\Omega)_+;\ \int_\Omega g(\omega)\,\mathrm{d}\omega = L\}$, and

$$\mathcal{F}(v,g) = \frac{1}{2}\int_\Omega \big\|\nabla v(\omega)\big\|^2\,\mathrm{d}\omega - \int_\Omega f(\omega)v(\omega)\,\mathrm{d}\omega - \int_\Omega \left(v(\omega) - \frac{1}{2}g(\omega)\right)g(\omega)\,\mathrm{d}\omega.$$

In this formulation, $g(\omega)$ is the amount of water at the vertical of point $\omega \in \Omega$, that clearly is nonnegative and whose integral must equal L. This means that we allow the height of water to vary over Ω. The average level of water at point $\omega \in \Omega$ is $v(\omega) - \frac{1}{2}g(\omega)$. The last term of $\mathcal{F}(v,g)$ represents therefore the potential energy associated with the water. Note that $\mathcal{F}$ is a convex function of each of its two variables, but not of (v,g) together in general. Let us compute the minimum over g, for a given v.

Lemma 10.1. *Given $v \in H_0^1(\Omega)$, the minimum over $g \in \mathcal{K}$ is attained at the unique point $\gamma(v) := (v - h(v,L))_+$, and the associated Lagrange multiplier is $h(v,L)$.*

Proof. The problem of minimization over g is strongly convex and is feasible for any positive value of L. Therefore there exists a unique minimum, characterized by the existence of a Lagrange multiplier λ, such that g attains the minimum over $L^2(\Omega)_+$ of the Lagrangian function:

$$-\int_{\Omega}\left(v(\omega)-\frac{1}{2}g(\omega)\right)g(\omega)\,\mathrm{d}\omega+\lambda\left(\int_{\Omega}g(\omega)\,\mathrm{d}\omega-L\right). \tag{66}$$

The minimum is attained over $L^2(\Omega)_+$ at the unique point $(v-\lambda)_+$. In view of the linear constraint it appears that $\lambda=h(v,L)$. The result follows. □

Substituting this expression of $\gamma(v)$ and using the linear constraint, we obtain that $F(v)=\mathcal{F}(v,\gamma(v))$. Therefore it is equivalent to minimize either F over K, or $\mathcal{F}$ over $K\times\mathcal{K}$. We remind that the obstacle problem (OP_f) was defined in Section 1. We now consider the relaxation algorithm, that consists in minimizing alternatively over each variable:

Relaxation algorithm RA.

1. Choose $v^0\in K$; $k:=0$.
2. Compute $g^k:=\gamma(v^k)$, and set $f_k:=g^k+f$.
3. Compute v^{k+1}, solution of (OP_{f_k}).
4. $k:=k+1$; go to step 2.

Theorem 10.2. *The sequence* (v^k,g^k) *is bounded in* $H_0^1(\Omega)\times H^1(\Omega)$*, and every weak limit-point* $(\bar v,\bar g)$ *of this sequence is a strong limit-point, such that* $\bar g=\gamma(\bar v)$*. In addition,* $\bar v$ *satisfies the first-order optimality conditions of* (RP).

Proof. By definition of g^k and step 3, we have that, for $k\geqslant 1$,

$$F\big(v^{k+1}\big)=\mathcal{F}\big(v^{k+1},g^{k+1}\big)\leqslant\mathcal{F}\big(v^{k+1},g^k\big)\leqslant\mathcal{F}\big(v^k,g^k\big)=F\big(v^k\big). \tag{67}$$

Since $F(v^k)$ is nonincreasing, by Proposition 3.2, the sequence v^k is bounded in $H_0^1(\Omega)$. Let us prove that g^k is bounded in $H^1(\Omega)$. Denote by $v^\sharp$ the solution of (OP_f). Since $f_k\geqslant f$, we have that $v^{k+1}\geqslant v^\sharp$, for all k, see [6, Corollary I.5]. This, by Lemma 2.2, implies that $h^k:=h(v^k,L)\geqslant h^\sharp:=h(v^\sharp,L)$ for all k. Therefore, by well-known properties of the maximum of two functions in $H^1(\Omega)$,

$$\begin{aligned}\|g^k\|_{H^1(\Omega)}&=\big\|\big(v^k-h^k\big)_+\big\|_{H^1(\Omega)}\leqslant\big\|\big(v^k-h^\sharp\big)_+\big\|_{H^1(\Omega)}\\&\leqslant\big\|\big(v^k-h^\sharp\big)\big\|_{H^1(\Omega)}\leqslant\big\|v^k\big\|_{H^1(\Omega)}+\big|h^\sharp\big|\,\mathrm{meas}(\Omega)^{1/2}.\end{aligned} \tag{68}$$

This proves that g^k is bounded in $H^1(\Omega)$. Since $\mathcal{F}$ is a quadratic function of v, its Hessian being the identity, and $\mathcal{F}(\cdot\,,g^k)$ attains its minimum over K at v^{k+1}, we have that

$$\mathcal{F}\big(v^{k+1},g^k\big)+\frac{1}{2}\big\|v^{k+1}-v^k\big\|^2_{H_0^1(\Omega)}\leqslant\mathcal{F}\big(v^k,g^k\big)\leqslant\mathcal{F}\big(v^k,g^{k-1}\big). \tag{69}$$

Since $\mathcal{F}(v^{k+1}, g^k)$ is bounded from below, the previous inequality implies that $\|v^{k+1} - v^k\| \to 0$ in $H_0^1(\Omega)$. Let $(\bar{v}, \bar{g})$ be the weak limit of (v^k, g^k), for $k \in N$, an infinite subset of $\mathbb{N}$. Since $\|v^{k+1} - v^k\| \to 0$ in $H_0^1(\Omega)$, we have that g^{k-1} has for the subsequence N the same limit $\bar{g}$. By (10), we have the strong limits of v^k and v^{k-1} in $L^2(\Omega)$. Passing to the limit, thanks to the weak l.s.c. of the elastic energy, we obtain $\mathcal{F}(\bar{v}, \bar{g}) \leqslant \mathcal{F}(v, \bar{g})$, for all $v \in K$. This means that $\bar{v}$ is solution of the obstacle problem $(\mathrm{OP}_{\bar{g}+f})$, proving that $\bar{v}$ satisfies the first-order optimality conditions of (RP). Let us prove the strong convergence. By step 3 of the algorithm, $\mathcal{F}(v^{k+1}, g^k) \leqslant \mathcal{F}(\bar{v}, g^k)$. Passing to the limit, we obtain that $\mathcal{F}(v^{k+1}, g^k) \to \mathcal{F}(\bar{v}, \bar{g})$, which implies convergence of the elastic energy, and therefore strong convergence of v^k in $H_0^1(\Omega)$. Since $h(v, L)$, and hence $\gamma(v)$, are continuous functions, this implies strong convergence of g^k in $H^1(\Omega)$ too. □

11. Numerical results

We have implemented the decomposition algorithm, setting the bound constraints only at the nodes of the triangulation. Then a quadratic program has to be solved at each iteration. For this we use the function 'quadprog' of Matlab, with option PCG (preconditioned conjugate gradients). The stopping criterion is based on the variation of cost function. Setting $S_k = \mathcal{F}(v^{k+1}, g^k)$, we stop if $|S_k - S_{k-1}| + |S_{k-1} - S_{k-2}| \leqslant \varepsilon$. In our tests we have used $\varepsilon = 0.0001$.

We consider the case when Ω is a disc with center 0 and radius r, whose triangulation is as in Fig. 1. The number of elements is $p^2 n_T$, and the size of the rigidity matrix is of order $N = \frac{1}{2}p(p-1)n_T + 1$. Here n_T is the number of sectors into which the disk is equally divided, while p is the number of rings. We use $r = 10$, $L = 10$, and $n_T = p = 8$. We display the results for the cases with or without obstacles in Fig. 2. Without obstacle the algorithm needs 9 iterations and the height is $h = 2.2162$. We next add the obstacle Φ given by $\Phi(\omega) = (\omega_1)^2 + (\omega_2)^2 + 2$. Then only 8 iterations are needed, and $h = 1.5899$.

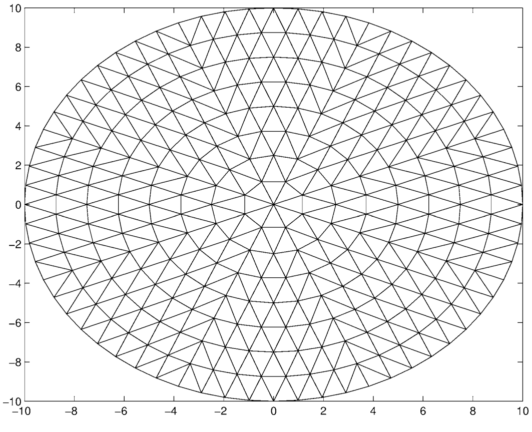
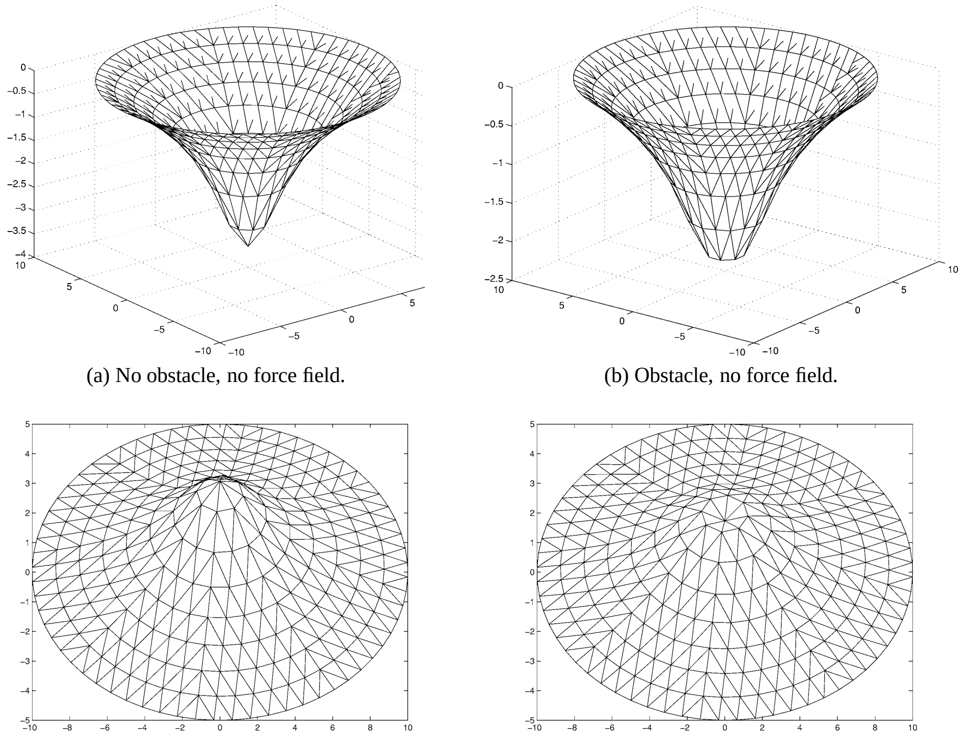


Fig. 1. Triangulation of circular domain.



(a) No obstacle, no force field.

(b) Obstacle, no force field.

(c) Case (a), view from below.

(d) Case (b), view from below.

Fig. 2. Numerical results.

References

[1] A. Aissani, M. Chipot, S. Fouad, On deformation of an elastic wire by one or two heavy disks, Arch. Math. 76 (2001) 467–480.

[2] D.P. Bertsekas, Nonlinear Programming, 2nd Edition, Athena Scientific, Belmont, MA, 1982.

[3] R. Bessi-Fourati, Equilibre d'une membrane contenant de l'eau, PhD thesis, ENIT, Tunis, 2003, in preparation.

[4] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, C. Sagastizábal, Numerical Optimization, Springer-Verlag, Berlin, 2002.

[5] J.F. Bonnans, A. Shapiro, Perturbation Analysis of Optimization Problems, Springer-Verlag, Berlin, 2000.

[6] H. Brézis, Problèmes unilatéraux, J. Math. Pures Appl. 51 (1972) 1–168.

[7] G. Buttazzo, A. Wagner, On the optimal shape of a rigid body supported by an elastic membrane, Nonlinear Anal. 39 (2000) 47–63.

[8] R. Dautray, J.L. Lions, Analyse Mathématique et Calcul Numérique pour les Sciences et Techniques, Masson, Paris, 1984–1985.

[9] G. Duvaut, J.L. Lions, Les Inéquations en Mécanique et en Physique, Dunod, Paris, 1972.

[10] I.N. Figueiredo, C.F. Leal, Sensitivity analysis of a nonlinear obstacle plate problem, ESAIM Control Optim. Calc. Var. 7 (2002) 135–155, electronic.

[11] D. Gilbarg, N.S. Trudinger, Elliptic Partial Differential Equations of Second Order, Springer-Verlag, Berlin, 1983.

[12] R. Glowinski, J.-L. Lions, R. Trémolières, Numerical Analysis of Variational Inequalities, in: Stud. Math. Appl., Vol. 8, North-Holland, Amsterdam, 1981; translated from the French edition: Dunod, Paris, 1976.

[13] A. Haraux, How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities, J. Math. Soc. Japan 29 (1977) 615–631.

[14] A.B. Levy, Sensitivity of solutions to variational inequalities on Banach spaces, SIAM J. Control Optim. 38 (1999) 50–60.

[15] J.L. Lions, G. Stampacchia, Variational inequalities, Comm. Pure Appl. Math. 20 (1967) 493–519.

[16] F. Mignot, Contrôle dans les inéquations variationnelles elliptiques, J. Funct. Anal. 22 (1976) 130–185.

[17] J. Nocedal, S.J. Wright, Numerical Optimization, Springer-Verlag, New York, 1999.

[18] P.A. Raviart, J.M. Thomas, Introduction à l'Analyse des Equations aux Dérivées Partielles, Masson, Paris, 1983.