CHAPITRE 3: NUMERICAL TREATMENT OF VARIATIONAL INEQUALITIES

We saw in the last chapters methods to solve constrained optimization problems, where the constraints take the form of equalities. In this chapter, we consider the case where the constraints are given by inequalities.

1. Example

Given an integer d, let Ω be an open space of $\mathbb{R}^d$. A standard example is given by a hanging cable, that may have a contact with an obstacle. The system is assumed to satisfy on Ω

$$\begin{aligned} -\Delta u - \lambda &= f \\ u &\geq h \\ \lambda &\geq 0, \end{aligned}$$

where $u = u(x)$ is the ordinate of the cable, $h = h(x)$ is the ordinate of the obstacle, $\lambda = \lambda(x)$ is reaction force exerted on the obstacle and $f = f(x) = -mg$, is the gravity with m the lineic masse of the cable and $g = 9.81$.

We see that this system has two unknowns, namely, u and λ.

2. Formalization

2.1. **Convex optimization.** Consider now the following abstract problem:

$$\min_{u \in K} J(u),$$

where K is a convex set of an Hilbert space. We have the following (well-known) result.

Theorem 1. *Assume that J is convex, Lower semi-continuous and coercive (in the sense that $J(u) \to +\infty$ when $\|u\| \to +\infty$). Then J admits a minimum.*

In the case of the obstacle problem introduced in Section 1, we have

$$J(u) = \frac{1}{2} a(u,u) - f(u),$$

with $a(u,v) := \int_\Omega \nabla u \cdot \nabla v$, $f(v) := \int_\Omega fv$ and $K := \{v \in H, \forall \eta \in M, (v-h, \eta) \geq 0\}$, where M is a convex closed cone in H. Such a definition of K is generally associated with the notation $K - h = M^+$. This definition follows from the fact that a vector v has only positive components if and only if $(v, e_i) \geq 0$ for all $i \in \mathbb{N}$, where e_i is a unitary basis vector of H. In this way, $M = \text{Span}_+(e_i, i \in \mathbb{N})$, where Span_+ denotes the set of linear combination with positive coefficients.

2.2. **Optimality conditions.** In the literature, one meets two ways to write the optimality conditions[1].

[1]This was also the case of equality constraints, see Chapter 1.

2.2.1. *Direct formulation.* The first formulation follows from the next reasoning. If u is a minimum of J in K, then, for all δu such that $u + \delta u \in K$, we have

$$J(u + \delta u) \geq J(u).$$

Assuming that J is differentiable, we get

$$(\nabla J(u), \delta u) \geq 0.$$

Setting $v := u + \delta u$, we find that this condition is equivalent to

$$(\nabla J(u), v - u) \geq 0, \tag{1}$$

and which should hold for all $v \in K$.

This formulation is in practice difficult to implement, since the set K may be not easy to discretize. Hence, we consider a Lagrange Multiplier approach.

2.2.2. *Lagrange multiplier approach.* In this approach, we set $\lambda := \nabla J(u)$ and express the optimality condition by characterizing λ.

First note that if M is a closed convex cone, then so as M^+. We start with a preliminary result.

Lemma 1. *If M is a closed convex cone, then $M^{++} = M$.*

Proof. We first prove that $M \subset M^{++}$. Indeed, let $\eta \in M$, and $v \in M^+$. By definition of M^+, we have $(v, \eta) \geq 0$ for all $v \in M^+$ and $\eta \in M$. It follows that $\eta \in M^{++}$, hence $M \subset M^{++}$.

The inverse inclusion $M^{++} \subset M$ is slightly more difficult. Assume that $M^{++} \not\subset M$, and let $x \in M^{++}$ with $x \notin M$. Since M is convex and closed, one can separate[2] the singleton $\{x\}$ and the set M. This means that there exists $\varepsilon > 0$ and $p \in H$ such that for all $y \in M$

$$(p, y) \geq \varepsilon + (p, x). \tag{2}$$

Since M is assumed to be a cone, this implies that $(p, y) \geq 0$ for all $y \in M$. Indeed, if $(p, y) \leq 0$, then $(p, \alpha y) \to -\infty$ when $\alpha \to +\infty$, which contradicts (2). It follows that $p \in M^+$. Taking $y = 0$ in (2), gives $-\varepsilon \geq (p, x)$, which contradicts $x \in M^{++}$. □

We then need another result.

Lemma 2. *Using the previous notation, we have*

$$(\lambda, u - h) = 0. \tag{3}$$

Proof. Because of the direct formulation (1), we have

$$(\nabla J(u), v - u) \geq 0,$$

and which holds for all $v \in K$. We set successively $v = h \in K$ and $v = 2u - h$ which also belongs to K, since $v - h = 2(u - h) \in M^+$ (here, we use that M^+ is a cone). We get respectively

$$(\lambda, u - h) \leq 0,$$

and

$$(\lambda, u - h) \geq 0,$$

hence the result. □

[2]Here, we use the Hahn-Banach theorem, which, in the case of Hilbert spaces, is easier to prove.

We are now in a position to state and prove our main result.

Theorem 2. *With the previous notation, $\lambda \in M$.*

Proof. We have

$$(\lambda, v-h) = (\nabla J(u), v-u) + (\nabla J(u), u-h) = (\nabla J(u), v-u) \geq 0,$$

where the second equality follows from Lemma 2 and the last inequality corresponds to the direct formulation (1). Since $K - h = M^+$, this means that $\lambda \in M^{++}$. Because of Lemma 1, $\lambda \in M$. □

We need to complete the previous condition by the so-called complementary conditions. These one correspond to a stronger version of Lemma 2.

Consider again an Hilbert basis $(e_i)_{i\in\mathbb{N}}$, such that $K-h = M^+ = \mathrm{Span}_+(e_i)$, and define the corresponding dual basis[3] $(f_i)_{i\in\mathbb{N}}$ such that $M = \mathrm{Span}(f_i)$ and $(e_i, f_j) = \delta_{i,j}$. Given $i \in \mathbb{N}$, we can now successively substitute $v = u + (u_i - h_i)e_i \in M^+$ and $v = u + \sum_{j\neq i}(u_j - h_j)e_j \in M^+$ (since $v \in K$) in (1) . We get:

$$(\lambda_i, u_i - h_i) \geq 0$$

and

$$(\lambda_, u_i - h_i) \leq 0,$$

so that (3) actually holds "component-wise" !

We finally obtain the optimality system:

$$\begin{aligned} \nabla J(u) - \lambda &= 0 \\ \lambda &\in M \\ u - h &\in M^+ \\ \lambda_i(u_i - h_i) &= 0. \end{aligned}$$

3. Solution Algorithm: The primal-Dual Active Set Strategy

At the discrete level, the previous system reads

$$AU - \Lambda = F \tag{4}$$

$$\Lambda \geq 0 \tag{5}$$

$$U - H \geq 0 \tag{6}$$

$$\Lambda_i(U_i - H_i) = 0. \tag{7}$$

Here, we assume the vectors to be of size N, where is a fixed given integer. Dealing with inequalities is possible numerically, but we rather focus on a method where this system is transformed into a system of equalities.

3.1. Reformulation in terms of a non-smooth system of equalities.

We introduce the function

$$\Phi(U, \Lambda) = \Lambda - \max(0, \Lambda - c(U - H)).$$

Theorem 3. *The system (4–7) is equivalent to*

$$AU - \Lambda = F \tag{8}$$

$$\Phi(U, \Lambda) = 0. \tag{9}$$

[3]Such a basis is usually called "biorthogonal basis" when $(e_i)_{i\in\mathbb{N}}$ is orthogonal.

Proof. We define the set of indices

$$\begin{aligned}\mathcal{I} =&\{i \in \{1,\dots,N\}, \Lambda_i - c(U_i - H_i) \leq 0\}\\ \mathcal{A} =&\{i \in \{1,\dots,N\}, \Lambda_i - c(U_i - H_i) > 0\}.\end{aligned}$$

Let us prove that (8–9) implies (4–7). Assume i belongs to $\mathcal{I}$. Then $\Lambda_i - c(U_i - H_i) \leq 0$, so that (9) gives $\Lambda_i = 0$, hence (5) and (7). Considering again $\Lambda_i - c(U_i - H_i) \leq 0$, we get $-c(U_i - H_i) \leq 0$, so that (6) holds. Similar reasoning when i belongs to $\mathcal{A}$ give the result in this case.

Finally, we leave to the reader the (easy) proof that (4–7) implies (8–9). □

3.2. **Solution algorithm.** The algorithm we now consider is simply the Newton's method, applied to the system (8–9). This leads to the iteration

$$\begin{aligned} A\delta U^{k+1} - \delta\Lambda^{k+1} =& \quad -AU^k + \Lambda^k + F\\ \partial_U\Phi(U^k,\Lambda^k)\delta U^{k+1} + \partial_\Lambda\Phi(U^k,\Lambda^k)\delta\Lambda^{k+1} =& \quad \Phi(U^k,\Lambda^k),\end{aligned}$$

where

$$\delta U^{k+1} = U^{k+1} - U^k,\ \delta\Lambda^{k+1} = \Lambda^{k+1} - \Lambda^k,$$

so that (10) is equivalent to

$$AU^{k+1} - \Lambda^{k+1} = F$$

Next, we define

$$\begin{aligned}\mathcal{I}^k =&\{i \in \{1,\dots,N\}, \Lambda_i^k - c(U_i^k - H_i^k) \leq 0\},\\ \mathcal{A}^k =&\{i \in \{1,\dots,N\}, \Lambda_i^k - c(U_i^k - H_i^k) > 0\}.\end{aligned}$$

Focusing on (10), we see that if i belongs to $\mathcal{I}^k$, $\Phi_i(U^k,\Lambda^k) = \Lambda_i^k$, then we get

$$\delta\Lambda^{k+1}_{|\mathcal{I}^k} = -\Lambda^k_{|\mathcal{I}^k},$$

whereas if i belongs to $\mathcal{A}^k$, $\Phi_i(U^k,\Lambda^k) = c(U_i^k - H_i^k)$, meaning that

$$c\delta U^{k+1}_{|\mathcal{A}^k} = H_{|\mathcal{A}^k} - cU^k_{|\mathcal{A}^k}.$$

Summarizing, we see that (10) leads to the update

$$\Lambda^{k+1}_{|\mathcal{I}^k} = 0,\ U^{k+1}_{|\mathcal{A}^k} = H_{|\mathcal{A}^k}.$$

The other parts of Λ^{k+1} and U^{k+1} are finally determined by solving the system

$$(A_{\mathcal{I}^k}| - Id_{\mathcal{A}^k}) \begin{pmatrix} U^{k+1}_{|\mathcal{I}^k} \\ \Lambda^{k+1}_{|\mathcal{I}^k} \end{pmatrix} = F + \left(A_{|\mathcal{A}^k}\right) \begin{pmatrix} H_{|\mathcal{A}^k} \end{pmatrix}.$$

We conclude with two remarks.

Remark 1. *Considering the initial problem, we see that we originally have only u as an unknown. In this way, introducing a Lagrange multiplier doubles the number of unknowns. But we see that, at the end, the solution algorithm only requires the inversion of a system whose size equals the one of u.*

Remark 2. *A good exercise consists in studying how this algorithm reads when the initial problem is nonlinear. Such a study reveals that no additional iteration need to be introduced when applying the Newton's iteration.*