SIAM J. Optim.
Vol. 30, No. 2, pp. 1610–1637

A NEW SEQUENTIAL OPTIMALITY CONDITION FOR CONSTRAINED NONSMOOTH OPTIMIZATION[*]

ELIAS S. HELOU[†], SANDRA A. SANTOS[‡], and LUCAS E. A. SIMÕES[‡]

Abstract. We introduce a sequential optimality condition for locally Lipschitz constrained nonsmooth optimization, verifiable just using derivative information, and which holds even in the absence of any constraint qualification. We present a practical algorithm that generates iterates either fulfilling the new necessary optimality condition or converging to stationary points of the infeasibility measure. A main feature of the devised algorithm is to allow a stronger control over the infeasibility of the iterates than usually obtained by exact penalty strategies, ensuring theoretical and practical advantages. Illustrative numerical experiments highlight the potentialities of the algorithm.

Key words. nonsmooth nonconvex optimization, constrained optimization, sequential optimality condition, constraint qualification

AMS subject classifications. 90C26, 90C30, 90C46

DOI. 10.1137/18M1228608

1. Introduction.

We consider constrained nonsmooth optimization problems of the form

$$\text{(P)} \qquad \begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \\ &\text{s.t. } \boldsymbol{c}(\boldsymbol{x}) \le \boldsymbol{0}, \end{aligned}$$

where both $f : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{c} : \mathbb{R}^n \to \mathbb{R}^p$ are locally Lipschitz continuous functions. Equality constraints of the type $\boldsymbol{h}(\boldsymbol{x}) = 0$ can also be easily incorporated in our framework (see Remark 2.8 for more details).

A common way of solving constrained optimization problems is to turn (P) into an unconstrained minimization by penalizing points $\boldsymbol{x} \in \mathbb{R}^n$ that violate the constraints; i.e., one seeks to find a solution to

$$\text{(Unc-P)} \qquad \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + \rho z(\boldsymbol{x}),$$

where ρ is a positive real number and $z : \mathbb{R}^n \to \mathbb{R}$ is a continuous function satisfying

$$\begin{cases} z(\boldsymbol{x}) = 0 & \text{if } \boldsymbol{c}(\boldsymbol{x}) \le \boldsymbol{0}, \\ z(\boldsymbol{x}) > 0 & \text{otherwise.} \end{cases}$$

Under certain conditions, and considering ρ to be large enough, the unconstrained minimization (Unc-P) and the original optimization (P) are equivalent [17]. This approach is known as *exact penalization* [33, 56], and it can be used to derive most of the theory behind constrained optimization for finite dimensions [18].

[*] Received by the editors November 26, 2018; accepted for publication (in revised form) December 18, 2019; published electronically June 16, 2020.
https://doi.org/10.1137/18M1228608
Funding: This work was partially funded by CNPq grants 310893/2019-4 and 302915/2016-8; FAPESP grants 2013/07375-0, 2018/24293-0, 2016/22989-2, and 2017/07265-0; and PRONEX Optimization.

[†] Institute of Mathematical Sciences and Computation, University of São Paulo, São Carlos - SP, 13566-590, Brazil (elias@icmc.usp.br).
[‡] Department of Applied Mathematics, University of Campinas, Campinas - SP, 13083-859, Brazil (sandra@ime.unicamp.br, simoes.lea@gmail.com).

For the majority of the optimization problems, the equivalence between (P) and (Unc-P) can only be achieved for finite ρ when z is a nondifferentiable function. However, when both objective and constraint functions are smooth in the entire domain, one usually does not want to trade the constrained smooth problem with a nonsmooth unconstrained one. Hence, many alternatives have been proposed over the years in order to overcome the nondifferentiable nature of (Unc-P). For instance, the augmented Lagrangian [1, 25] and the sequential quadratic programming (SQP) [11, 52] are widely known methods. However, when (P) already presents nonsmoothness, the nondifferentiability of z usually does not introduce any additional difficulty. For this reason, many methods developed for solving constrained nonsmooth optimization problems are based on exact penalty functions [27, 28, 44, 45, 54].

There is a strong connection between constraint qualifications (CQs) [6, 17, 52] for problem (P) and the existence of a finite penalty parameter that makes (Unc-P) equivalent to (P). Moreover, convergence results for exact penalization methods (and, more generally, for the majority of the optimization algorithms) are based on different kinds of CQs. However, checking the validity of constraint qualifications is not an easy task, which may justify the fact that most of the algorithms are not designed to test any kind of CQ even when theoretical convergence of the method relies on such conditions. As a result, practical necessary optimality conditions for (P) that do not depend on CQs are of great importance for establishing theoretically sound stopping criteria for any optimization method.

The *approximate Karush–Kuhn–Tucker* (AKKT) and the *complementary approximate Karush–Kuhn–Tucker* (CAKKT) conditions [3, 7] present themselves as reliable necessary optimality conditions for smooth optimization problems. Looking at auxiliary functions that approximate the exact penalty approach, the authors of both studies show that any solution $\boldsymbol{x}^*$ of the smooth optimization problem must satisfy the following sequential optimality condition:

$$\lim_{k\to\infty}\left\|\nabla f(\boldsymbol{x}^k)+\sum_{i=1}^{p}\nabla c_i(\boldsymbol{x}^k)\mu_i^k\right\|=0, \tag{1.1}$$

where $\|\boldsymbol{x}\|$ is the Euclidean norm of $\boldsymbol{x}\in\mathbb{R}^n$, $\{\boldsymbol{x}^k\}\subset\mathbb{R}^n$ is a sequence converging to $\boldsymbol{x}^*$, and $\{\boldsymbol{\mu}^k\}\subset\mathbb{R}^p_+$ is a sequence of vectors whose components μ_i^k must satisfy $\lim_{k\to\infty}\min\{-c_i(\boldsymbol{x}^k),\mu_i^k\}=0$, where $c_i:\mathbb{R}^n\to\mathbb{R}$ are the components of $\boldsymbol{c}:\mathbb{R}^n\to\mathbb{R}^p$ for each $i\in\{1,\dots,p\}$. A natural attempt to generalize (1.1) to the nonsmooth case is to use the subdifferential concept, i.e., to consider a sequence of elements

$$\{\boldsymbol{v}^k\}\subset\left(\partial f(\boldsymbol{x}^k)+\sum_{i=1}^{p}\partial c_i(\boldsymbol{x}^k)\mu_i^k\right),\quad\text{with}\ \lim_{k\to\infty}\|\boldsymbol{v}^k\|=0, \tag{1.2}$$

where $\partial g(\boldsymbol{x})$ stands for the Clarke subdifferential set of the function g at the point $\boldsymbol{x}$ [23]. Unfortunately, appropriate vectors from the subdifferential satisfying these conditions are, in general, not expected to be computed by numerical algorithms, which invalidates the use of (1.2) as a practical stopping criterion.

Using the fact that many real nonsmooth optimization problems are described by functions that are continuously differentiable in a full-measure subset of $\mathbb{R}^n$, the authors of [32] present a necessary optimality condition that relies upon the fact that, in many cases, the sets $\partial f(\boldsymbol{x}^k)$ and $\partial c_i(\boldsymbol{x}^k)$ can be traded by $\nabla f(\hat{\boldsymbol{x}}^k)$ and $\nabla c_i(\hat{\boldsymbol{x}}^k)$, where $\hat{\boldsymbol{x}}^k$ is a point sufficiently close to $\boldsymbol{x}^k$. However, if no additional hypothesis is assumed, one must have $\hat{\boldsymbol{x}}^k=\boldsymbol{x}^k$, which brings back the same issues discussed in the previous paragraph.

In this paper, we propose a new sequential optimality condition for constrained nonsmooth optimization problems that allows the user to work with derivatives even in the absence of any CQ. To show that our necessary optimality condition is practical, we present an algorithm that is proven to generate a sequence of points that either satisfies such a condition or converges to stationary points of the infeasibility measure. A main feature of the proposed algorithm is that it presents a stronger control over the infeasibility of the iterates than usual exact penalty methods, which has important practical consequences.

One of the main challenges of exact penalty methods is the selection of the initial value of the penalty parameter, and, as a consequence, several studies involving rules for updating the value of the penalty parameter can be found in the literature [14, 15, 16]. In case the penalty parameter value is chosen to be too large, the method may privilege the feasible region in detriment of optimality, which, in turn, can cause very short steps toward the optimal solution and/or may cause the penalized problem to be ill-conditioned. On the other hand, if the penalty parameter value is much smaller than the magnitude of the objective function at infeasible points, the method may rapidly be attracted by unconstrained minimizers that possess very low function values, preventing the user from obtaining a successful solution for reasonable values of ρ. Such a phenomenon is called *greediness* [9, 22] and may occur even if the user starts the method at a feasible point of the optimization problem.

Unlike some existing penalty methods, for which there is no simple way to compute a suitable initial value for the penalty parameter to easily confine the iterates into an almost feasible region, our algorithm allows the user to set a tolerance target value $\xi > 0$ for infeasibility. Outside this tolerance region, the method is indifferent to the objective function, which ensures the control of the infeasibility even at the initial iterations. As a consequence, it prevents our method from suffering from the greediness phenomenon.

The outline of the paper is as follows. Section 2 presents theoretical results supporting that every local minimizer of (P) must fulfill our proposed sequential optimality conditions. Section 3 shows the relation between our sequential optimality conditions and the AKKT and CAKKT conditions in the case where the objective and constraint functions of (P) are all smooth. In section 4, we state a general algorithm that produces a sequence of iterates fulfilling the proposed sequential optimality conditions whenever stationary points of the infeasibility measure are avoided. Section 5 brings numerical results that illustrate some of the important properties of our method. Finally, we leave section 6 for the conclusions of this study.

Along the manuscript, the following notations will be frequently used:

- $\|\cdot\|$ is the Euclidean norm.
- $\mathcal{B}(\boldsymbol{x}, \epsilon)$ is the Euclidean closed ball with center at $\boldsymbol{x}$ and radius ϵ.
- $\mathcal{P}(\boldsymbol{v} \,|\mathcal{X})$ is the Euclidean projection of $\boldsymbol{v}$ onto $\mathcal{X}$.
- $\mathbb{R}_+$ is the set of all nonnegative real numbers.
- $\mathbb{R}_+^*$ is the set of all strictly positive real numbers.
- $v_+ := \max\{v, 0\}$. If v is a vector, then $\boldsymbol{v}_+$ is also a vector where every entry is taken as the maximum between the respective entry of $\boldsymbol{v}$ and zero.
- $\operatorname{conv} \mathcal{X}$ denotes the convex hull of $\mathcal{X}$.
- $\operatorname{cl} \mathcal{X}$ denotes the closure of $\mathcal{X}$.
- $A + B$ represents the set $\{x + y : x \in A, y \in B\}$.
- $A \cdot B$ represents the set $\{xy : x \in A, y \in B\}$.
- $\boldsymbol{e}$ is the vector with one in all entries, with appropriate dimension.

2. Establishing new necessary optimality conditions.

We start this section by introducing the concept of a (Goldstein) ϵ-subdifferential set for any locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$ [36].

DEFINITION 2.1 (ϵ-subdifferential set, ϵ-subgradient, ϵ-stationary point). *The ϵ-subdifferential set of f at $\boldsymbol{x}$ is given by*

$$\partial_\epsilon f(\boldsymbol{x}) := \operatorname{conv} \partial f(\mathcal{B}(\boldsymbol{x}, \epsilon)).$$

Any $\boldsymbol{v} \in \partial_\epsilon f(\boldsymbol{x})$ is known as an ϵ-subgradient of f at $\boldsymbol{x}$. Moreover, if $\boldsymbol{0} \in \partial_\epsilon f(\boldsymbol{x})$, then we say that $\boldsymbol{x}$ is an ϵ-stationary point for f.

One of the greatest advantages of looking at ϵ-subdifferential sets instead of subdifferential sets is the possibility of seeing $\partial_\epsilon f(\boldsymbol{x})$ as a convex hull of the derivatives of f nearby $\boldsymbol{x}$. Indeed, for any locally Lipschitz continuous function f and any full-measure subset $\mathcal{D}^f$ of $\mathbb{R}^n$ such that f is differentiable at any point in $\mathcal{D}^f$—this subset always exists due to Rademacher's theorem [35, Theorem 3.1.6]—the following relations hold (see [19, 46]):

$$\mathcal{G}^f_\epsilon(\boldsymbol{x}) \subset \partial_\epsilon f(\boldsymbol{x}) \quad \text{and} \quad \partial_{\epsilon_1} f(\boldsymbol{x}) \subset \mathcal{G}^f_{\epsilon_2}(\boldsymbol{x}) \quad (0 \leq \epsilon_1 < \epsilon_2), \tag{2.1}$$

where $\mathcal{G}^f_\epsilon(\boldsymbol{x}) := \operatorname{cl} \operatorname{conv} \nabla f \left(\mathcal{B}(\boldsymbol{x}, \epsilon) \cap \mathcal{D}^f\right)$. Recent advances on practical tools [19, 20, 21] to approximate $\mathcal{G}^f_\epsilon(\boldsymbol{x})$ allow us to consider a new sequential optimality condition of practical use. The subgradients involved in this new necessary optimality condition are those associated with the objective function and the constraints related to the following index set:

$$\mathcal{I}_\epsilon(\boldsymbol{x}) := \{i : \exists \boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}, \epsilon) \text{ with } c_i(\boldsymbol{y}) \geq 0\}.$$

DEFINITION 2.2 (weak ϵ-approximate nonsmooth optimality condition). *A feasible point $\boldsymbol{x}^* \in \mathbb{R}^n$ of* (P) *is said to satisfy the weak ϵ-approximate nonsmooth optimality condition (weak ϵ-ANOC) if there exist sequences $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$, $\{\epsilon_k\} \subset \mathbb{R}^*_+$, $\{\boldsymbol{v}^k\} \subset \mathbb{R}^n$, and $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}^p_+$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$, $\epsilon_k \downarrow 0$, and $\|\boldsymbol{v}^k\| \to 0$, where*

$$\boldsymbol{v}^k \in \left(\mathcal{G}^f_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{i=1}^p \mu_i^k \mathcal{G}^{c_i}_{\epsilon_k}(\boldsymbol{x}^k)\right) \quad \textit{and} \quad i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0. \tag{2.2}$$

Notice that the above definition does not impose any control over the speed of the convergence $\epsilon_k \downarrow 0$. Not establishing a relation between the sequences $\{\epsilon_k\}$ and $\{\boldsymbol{x}^k\}$ allows $\epsilon_k \gg \|\boldsymbol{x}^k - \boldsymbol{x}^*\|$, which, in turn, means that the weak ϵ-ANOC may depart too much from the necessary condition that at least one generalized derivative of f at $\boldsymbol{x}^k$ must be close to a linear combination of the subgradients of the active constraints at $\boldsymbol{x}^k$. Figure 1 exemplifies this issue, showing that subgradients of far away inactive constraints may be considered in the linear combination.

Using the exact penalization approach for problem (P) and a reasoning similar to the one employed in Theorem 3.3 of [7], one can infer the weak ϵ-ANOC as a necessary optimality condition for nonsmooth problems. However, this strategy does not clarify the question of how fast the sequence $\{\epsilon_k\}$ must go to zero. To overcome this issue, we have applied a new penalization strategy that has its roots in [10, section 4.1].

The inspiring idea behind this new penalization is to trade the original nonsmooth problem with the minimization of a discontinuous function $\Theta_\rho : \mathbb{R}^n \to \mathbb{R}$,

$$\begin{cases} \Theta_\rho(\boldsymbol{x}) = f(\boldsymbol{x}) - \rho & \text{if } \boldsymbol{c}(\boldsymbol{x}) \leq \boldsymbol{0}, \\ \Theta_\rho(\boldsymbol{x}) \geq 0 & \text{otherwise,} \end{cases}$$

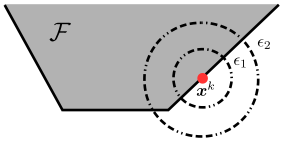


FIG. 1. *Illustration of how the value of ϵ in the ϵ-subdifferential can influence the activeness of the constraints.*

where $\rho \in \mathbb{R}$ plays a role different from the one presented in the exact penalization function. Here, the parameter ρ is any constant ensuring that $f(\boldsymbol{x}^*) - \rho < 0$, with $\boldsymbol{x}^*$ being a solution of (P). Like the strategy of solving constrained smooth optimization problems by the use of an exact penalty function, where a degree of difficulty is added by introducing a nondifferentiable term in order to avoid a constrained minimization, this approach also inserts one extra difficulty to the nonsmoothness of the original problem by using a discontinuous function. Therefore, we avoid using this idea directly.

Later, Theorem 2.5 shows that the weak ϵ-ANOC is, indeed, a necessary optimality condition for problem (P) and also elucidates the matter of the speed of the convergence $\epsilon_k \downarrow 0$. Its proof relies on a sequence $\{\Psi_{\xi_k,\rho}\}$ of nonsmooth continuous functions $\Psi_{\xi_k,\rho} : \mathbb{R}^n \to \mathbb{R}$ approximating the discontinuous function Θ_ρ. Given a scalar $\xi \in \mathbb{R}^*_+$, we define

$$\Psi_{\xi,\rho}(\boldsymbol{x}) := \max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x})_+\|_1}{\xi}, 0\right\}[f(\boldsymbol{x}) - \rho] + \|\boldsymbol{c}(\boldsymbol{x})_+\|_1. \tag{2.3}$$

It is worth noticing that, since f and $\boldsymbol{c}$ are locally Lipschitz continuous functions, the map $\Psi_{\xi,\rho}$ is also locally Lipschitz continuous.

Preceding the result that establishes the weak ϵ-ANOC as a necessary optimality condition, we present two lemmas that will facilitate the proof of Theorem 2.5.

LEMMA 2.3. *Let $\boldsymbol{x} \in \mathbb{R}^n$, $(\epsilon, \xi) \in \mathbb{R}^*_+ \times \mathbb{R}^*_+$, $\rho \in \mathbb{R}$ with $\rho > f(\boldsymbol{x})$, and $\boldsymbol{v} \in \partial_\epsilon \Psi_{\xi,\rho}(\boldsymbol{x})$. Then, if $\epsilon > 0$ is such that $\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}, \epsilon) \Rightarrow \rho \geq f(\boldsymbol{y})$, there exists $\{\boldsymbol{x}^j\}_{j=1}^{n+1} \subset \mathcal{B}(\boldsymbol{x}, \epsilon)$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n+1}_+$, with $\boldsymbol{e}^T\boldsymbol{\lambda} = 1$, such that*

$$\boldsymbol{v} \in \sum_{j=1}^{n+1} \lambda_j \left(\max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^j)_+\|_1}{\xi}, 0\right\}\partial f(\boldsymbol{x}^j) + \sigma_j \partial\|\boldsymbol{c}(\boldsymbol{x}^j)_+\|_1\right), \tag{2.4}$$

where $\sigma_j \geq 1$, $j \in \{1, \ldots, n+1\}$.

Proof. Recalling that $\partial_\epsilon \Psi_{\xi,\rho}(\boldsymbol{x})$ is a convex set, it follows from Carathéodory's theorem [55, Theorem 2.29] that there exists $\{\boldsymbol{s}^j\}_{j=1}^{n+1} \subset \partial\Psi_{\xi,\rho}(\mathcal{B}(\boldsymbol{x}, \epsilon))$ and $\boldsymbol{\lambda} \in \mathbb{R}^{n+1}_+$, with $\boldsymbol{e}^T\boldsymbol{\lambda} = 1$, such that $\boldsymbol{v} = \sum_{j=1}^{n+1} \lambda_j \boldsymbol{s}^j$. Because of [23, Theorems 2.3.9 and 2.3.13] and [23, Proposition 2.3.3], we know that $\partial\Psi_{\xi,\rho}(\boldsymbol{x})$ is a subset of

$$\max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x})_+\|_1}{\xi}, 0\right\}\partial f(\boldsymbol{x}) + \left(1 + [\rho - f(\boldsymbol{x})]\operatorname{conv}\left\{\frac{1}{\xi}, 0\right\}\right) \cdot \partial\|\boldsymbol{c}(\boldsymbol{x})_+\|_1.$$

So, since, by hypothesis, $\rho - f(\boldsymbol{x}^j) \geq 0$, $j \in \{1, \ldots, n+1\}$, there exists $\{\boldsymbol{x}^j\}_{j=1}^{n+1} \subset \mathcal{B}(\boldsymbol{x}, \epsilon)$ such that (2.4) holds with $1 \leq \sigma_j \in \left(1 + [\rho - f(\boldsymbol{x}^j)]\operatorname{conv}\left\{\frac{1}{\xi}, 0\right\}\right)$. ◻

The next result presents sufficient conditions for the feasible point $\boldsymbol{x}^*$ to satisfy the weak ϵ-ANOC.

Lemma 2.4. *Suppose $\boldsymbol{x}^*$ is a feasible point of problem* (P), *$\{\boldsymbol{x}^k\}$ is a sequence converging to $\boldsymbol{x}^*$, $\{\zeta_k\}$ and $\{\xi_k\}$ are both real-valued sequences with $\zeta_k \downarrow 0$, $\xi_k \downarrow 0$, and $\zeta_k/\xi_k \to 0$, and ρ is a real value satisfying $f(\boldsymbol{x}^*)-\rho < 0$. If $\lim_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1/\xi_k$ exists and it is strictly less than one, and there exists a sequence $\{\boldsymbol{r}^k\} \subset \mathbb{R}^n$ such that $\|\boldsymbol{r}^k\| \to 0$ and $\boldsymbol{r}^k \in \partial_{\zeta_k}\Psi_{\xi_k,\rho}(\boldsymbol{x}^k)$ for all $k \in \mathbb{N}$, then $\boldsymbol{x}^*$ satisfies the weak ϵ-ANOC.*

Proof. Let us choose a sequence $\{\epsilon_k\}$ satisfying $\zeta_k < \epsilon_k$ for all $k \in \mathbb{N}$, with $\epsilon_k \downarrow 0$. Recalling Lemma 2.3, it follows that, for all large enough k, there exists $\{\boldsymbol{x}^{k,j}\}_{j=1}^{n+1} \subset \mathcal{B}(\boldsymbol{x}^k, \zeta_k)$ such that

$$\boldsymbol{r}^k \in \sum_{j=1}^{n+1} \lambda_j^k \left(\max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0\right\} \partial f(\boldsymbol{x}^{k,j}) + \sigma_j^k \partial \|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1 \right),$$

with $\sigma_j^k \geq 1$, $j \in \{1, \dots, n+1\}$. Therefore, for each j, there must exist $\boldsymbol{s}_f^{k,j} \in \partial f(\boldsymbol{x}^{k,j})$ and $\boldsymbol{s}_{c_i}^{k,j} \in \partial c_i(\boldsymbol{x}^{k,j})$ for each i, such that

$$\boldsymbol{r}^k = \sum_{j=1}^{n+1} \lambda_j^k \left(\max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0\right\} \boldsymbol{s}_f^{k,j} + \sigma_j^k \sum_{i=1}^{p} \Delta_i(\boldsymbol{x}^{k,j}) \boldsymbol{s}_{c_i}^{k,j} \right), \tag{2.5}$$

where

$$\Delta_i(\boldsymbol{x}) \in \begin{cases} \{1\} & \text{if } c_i(\boldsymbol{x}) > 0, \\ \operatorname{conv}\{1, 0\} & \text{if } c_i(\boldsymbol{x}) = 0, \\ \{0\}, & \text{if } c_i(\boldsymbol{x}) < 0. \end{cases}$$

Then, defining

$$\iota_k^f = \sum_{j=1}^{n+1} \lambda_j^k \max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0\right\} \quad \text{and} \quad \iota_k^{c_i} = \sum_{j=1}^{n+1} \lambda_j^k \sigma_j^k \Delta_i(\boldsymbol{x}^{k,j}), \tag{2.6}$$

one can see, for large values of k, that ι_k^f is strictly positive due to the fact that $\lim_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1/\xi_k$ is strictly less than one, $\|\boldsymbol{x}^k - \boldsymbol{x}^{k,j}\| \leq \zeta_k$, and $\zeta_k/\xi_k \to 0$. Hence, for k large enough, we have, for all i and j, that

$$0 \leq \tau_{k,j}^f := \frac{\lambda_j^k \max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0\right\}}{\iota_k^f} \quad \text{and} \quad 0 \leq \tau_{k,j}^{c_i} := \begin{cases} \frac{\lambda_j^k \sigma_j^k \Delta_i(\boldsymbol{x}^{k,j})}{\iota_k^{c_i}} & \text{if } \iota_k^{c_i} > 0, \\ 0 & \text{if } \iota_k^{c_i} = 0. \end{cases}$$

Additionally, the following holds: $\sum_{j=1}^{n+1} \tau_{k,j}^f = 1$ and

$$\sum_{j=1}^{n+1} \tau_{k,j}^{c_i} = \begin{cases} 1 & \text{if } \iota_k^{c_i} > 0 \\ 0 & \text{if } \iota_k^{c_i} = 0 \end{cases} \quad \text{for all } i \in \{1, \dots, p\}.$$

This together with (2.5) implies that $\boldsymbol{r}^k \in \left(\iota_k^f \partial_{\zeta_k} f(\boldsymbol{x}^k) + \sum_{i=1}^p \iota_k^{c_i} \partial_{\zeta_k} c(\boldsymbol{x}^k)\right)$. So, choosing $\boldsymbol{v}^k := \frac{1}{\iota_k^f}\boldsymbol{r}^k$, we obtain $\boldsymbol{v}^k \in \left(\partial_{\zeta_k} f(\boldsymbol{x}^k) + \sum_{i=1}^p \frac{\iota_k^{c_i}}{\iota_k^f} \partial_{\zeta_k} c(\boldsymbol{x}^k)\right)$. Now, notice that, because of the way $\iota_k^{c_i}$ and Δ_i were defined, it must follow that $i \notin \mathcal{I}_{\zeta_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k := \frac{\iota_k^{c_i}}{\iota_k^f} = 0$. Moreover, since $\|\boldsymbol{r}^k\| \to 0$, it yields $\|\boldsymbol{v}^k\| \to 0$ as well. So,

$$\boldsymbol{v}^k \in \left(\partial_{\zeta_k} f(\boldsymbol{x}^k) + \sum_{i=1}^{p} \mu_i^k \partial_{\zeta_k} c_i(\boldsymbol{x}^k)\right) \quad \text{and} \quad i \notin \mathcal{I}_{\zeta_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0. \tag{2.7}$$

Remembering the inclusions presented in (2.1) and that $\partial_{\epsilon_1} g(\boldsymbol{x}) \subset \partial_{\epsilon_2} g(\boldsymbol{x})$ for any nonsmooth function g and $0 \le \epsilon_1 \le \epsilon_2$, it yields $\boldsymbol{v}^k \in \left(\mathcal{G}^f_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{i=1}^p \mu_i^k \mathcal{G}^{c_i}_{\epsilon_k}(\boldsymbol{x}^k)\right)$ and $i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0$, which proves the statement. □

We are finally able to introduce the theorem that guarantees that the weak ϵ-ANOC is a necessary optimality condition.

THEOREM 2.5. *Let $\boldsymbol{x}^*$ be a local minimizer of* (P)*. Then, $\boldsymbol{x}^*$ satisfies the weak ϵ-ANOC.*

Proof. Since $\boldsymbol{x}^*$ is a local minimizer of (P), there must exist $\delta_1 > 0$ such that $f(\boldsymbol{x}) \ge f(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, \delta_1) \cap \mathcal{F}$, where $\mathcal{F}$ is the feasible set of (P). Moreover, setting ρ as any real number satisfying $\rho > f(\boldsymbol{x}^*)$, it yields that there exists $\delta_2 > 0$ with $\rho > f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, \delta_2)$. Choosing $\delta = \min\{\delta_1, \delta_2\}/2$, we define $\Phi_\xi, \rho : \mathbb{R}^n \to \mathbb{R}$ as

$$\Phi_{\xi,\rho}(\boldsymbol{x}) := \max\left\{1 - \frac{\max\{\|\boldsymbol{x} - \boldsymbol{x}^*\| - \delta, 0\}}{\xi}, 0\right\} \max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x})_+\|_1}{\xi}, 0\right\} [f(\boldsymbol{x}) - \rho] + \|\boldsymbol{c}(\boldsymbol{x})_+\|_1 + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2.$$

Since $\Phi_{\xi,\rho}$ is a coercive function for any $\xi > 0$, a global minimizer of this function always exists. So, taking $\{\xi_k\} \subset \mathbb{R}^*_+$ as any sequence satisfying $\xi_k \downarrow 0$, we consider an infinite sequence $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$ such that $\boldsymbol{x}^k \in \operatorname{argmin}_{x \in \mathbb{R}^n} \Phi_{\xi_k,\rho}(\boldsymbol{x})$.

Consequently, $\Phi_{\xi_k,\rho}(\boldsymbol{x}^k) \le \Phi_{\xi_k,\rho}(\boldsymbol{x}^*) = f(\boldsymbol{x}^*) - \rho < 0$, which yields

$$\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 < \xi_k \quad \text{and} \quad \|\boldsymbol{x}^k - \boldsymbol{x}^*\| < \xi_k + \delta. \tag{2.8}$$

This implies that the sequence $\{\boldsymbol{x}^k\}$ must be bounded. So, let $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ be a cluster point of $\{\boldsymbol{x}^k\}$, which means that there exists an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that $\boldsymbol{x}^k \to_{k \in \mathcal{K}} \hat{\boldsymbol{x}}$. We then define

$$0 \le \tau_k := \max\left\{1 - \frac{\max\{\|\boldsymbol{x}^k - \boldsymbol{x}^*\| - \delta, 0\}}{\xi_k}, 0\right\} \max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1}{\xi_k}, 0\right\} \le 1.$$

Therefore, it is possible to find an infinite index set $\tilde{\mathcal{K}} \subset \mathcal{K}$ such that $\tau_k \to \hat{\tau}$ for some $\hat{\tau} \in [0, 1]$. Hence, since $\|\boldsymbol{c}(\hat{\boldsymbol{x}})_+\|_1 = 0$ because of (2.8), we have

$$\begin{aligned}\Phi_{\xi_k,\rho}(\boldsymbol{x}^k) \le \Phi_{\xi_k,\rho}(\boldsymbol{x}^*) &\Rightarrow \lim_{k \in \tilde{\mathcal{K}}} \Phi_{\xi_k,\rho}(\boldsymbol{x}^k) \le \lim_{k \in \tilde{\mathcal{K}}} \Phi_{\xi_k,\rho}(\boldsymbol{x}^*) \\ &\Rightarrow \hat{\tau}[f(\hat{\boldsymbol{x}}) - \rho] + \frac{1}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*) - \rho.\end{aligned}$$

By (2.8), we know that $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\| \le \delta$. Then, because of the way we have defined δ, we see that $f(\hat{\boldsymbol{x}}) - \rho$ and $f(\boldsymbol{x}^*) - \rho$ are strictly negative numbers. Consequently, recalling that $\boldsymbol{x}^*$ is a local minimizer of (P), we must have $\hat{\tau} = 1$ and $\hat{\boldsymbol{x}} = \boldsymbol{x}^*$. Therefore, since $\hat{\boldsymbol{x}}$ and $\hat{\tau}$ are arbitrary cluster points of $\{\boldsymbol{x}^k\}$ and $\{\tau_k\}$, respectively, this means that any cluster point of $\{\boldsymbol{x}^k\}$ must be $\boldsymbol{x}^*$ and that any cluster point of $\{\tau_k\}$ must be 1. Since both sequences are bounded, it follows that $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and $\tau_k \to 1$. In particular, this last limit also ensures that $\lim_{k \to \infty} \frac{\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1}{\xi_k}$ exists and it is strictly less than one.

The equality $\hat{\boldsymbol{x}} = \boldsymbol{x}^*$ implies, for any sufficiently large $k \in \mathbb{N}$, that

$$\|\boldsymbol{x}^k - \boldsymbol{x}^*\| \le \delta/2 \Rightarrow \partial\Phi_{\xi_k,\rho}(\boldsymbol{x}^k) = \partial\Psi_{\xi_k,\rho}(\boldsymbol{x}^k) + \boldsymbol{x}^k - \boldsymbol{x}^*.$$

However, by the way we have defined $\boldsymbol{x}^k$, we know that $\mathbf{0} \in \partial\Phi_{\xi_k,\rho}(\boldsymbol{x}^k)$, which, for any $\zeta_k \downarrow 0$, gives

$$\boldsymbol{x}^* - \boldsymbol{x}^k \in \partial\Psi_{\xi_k,\rho}(\boldsymbol{x}^k) \Rightarrow \exists \boldsymbol{r}^k \in \partial_{\zeta_k}\Psi_{\xi_k,\rho}(\boldsymbol{x}^k) \text{ with } \|\boldsymbol{r}^k\| \to 0.$$

Due to Lemma 2.4, the statement is proven. □

The demonstration of Lemma 2.4 gives a clue about the matter of convergence of ϵ_k. Notice that, along the proof, we had to assume $\epsilon_k/\xi_k > \zeta_k/\xi_k \to 0$, where each ξ_k is a parameter for the nonsmooth function $\Psi_{\xi_k,\rho}$. This was necessary to ensure that the multiplier associated with $\partial_{\zeta_k} f(\boldsymbol{x}^k)$ is bounded away from 0 in (2.6), which consequently means that the objective function is always considered in the sequential optimality condition (2.7). However, since ξ_k is a parameter intrinsically associated with the function $\Psi_{\xi_k,\rho}$, any sequential optimality condition requiring $\epsilon_k/\xi_k \to 0$ would only be useful for algorithms based on $\Psi_{\xi_k,\rho}$, and, consequently, such a condition would be very restrictive. This issue can be overcome by noticing that $\boldsymbol{\mu}^k$ and the limit $\epsilon_k/\xi_k \to 0$ are linked inside the proof of Lemma 2.4. So, by defining a new sequential optimality condition that requires $\epsilon_k\|\boldsymbol{\mu}^k\|_\infty \to 0$, we introduce the ϵ-approximate nonsmooth optimality condition.

Definition 2.6 (ϵ-approximate nonsmooth optimality condition). *A feasible point $\boldsymbol{x}^* \in \mathbb{R}^n$ of* (P) *is said to satisfy the ϵ-approximate nonsmooth optimality condition (ϵ-ANOC) if there exist sequences $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$, $\{\epsilon_k\} \subset \mathbb{R}^*_+$, $\{\boldsymbol{v}^k\} \subset \mathbb{R}^n$, and $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}^p_+$ fulfilling the weak ϵ-ANOC at $\boldsymbol{x}^*$ and, moreover, $\epsilon_k\|\boldsymbol{\mu}^k\|_\infty \to 0$ holds.*

From the proof of Theorem 2.5, one can see that the ϵ-ANOC is also a necessary optimality condition for any local solution $\boldsymbol{x}^*$ of (P).

Theorem 2.7. *Let $\boldsymbol{x}^*$ be a local minimizer of* (P). *Then, $\boldsymbol{x}^*$ satisfies the ϵ-ANOC.*

Proof. Looking at the proof of Theorem 2.5, and since $\boldsymbol{x}^*$ is a local minimizer of (P), one can find a sequence $\{\boldsymbol{x}^k\}$ converging to $\boldsymbol{x}^*$, and the associated sequences $\{\epsilon_k\}$, $\{\xi_k\}$ satisfying $\epsilon_k \downarrow 0$, $\xi_k \downarrow 0$, and $\epsilon_k/\xi_k \to 0$, such that the assumptions of Lemma 2.4 hold, and (2.7) is also valid (where $\zeta_k < \epsilon_k$). Additionally, since $\epsilon_k/\xi_k \to 0$, we also have $\operatorname{conv}\{0, 1/\xi_k\}\epsilon_k \to 0$. Hence, recalling that, in (2.6), ι_k^f is bounded away from zero for large values of k, and that $\{\iota_k^{c_i}\epsilon_k\}$ approaches zero for all $i \in \{1,\dots,p\}$ (due to the fact that $\operatorname{conv}\{0, 1/\xi_k\}\epsilon_k \to 0$), it must follow that $\epsilon_k\mu_i^k = \epsilon_k(\iota_k^{c_i}/\iota_k^f) \to 0$ for all $i \in \{1,\dots,p\}$. This information together with the proof of Theorem 2.5 ensures the result. □

Remark 2.8. Although the sequential optimality conditions that we have proposed so far consider only inequality constraints, one can easily generalize them to nonsmooth problems that may include equality constraints. Indeed, if in problem (P) the feasible set is given by $\mathcal{F} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{c}(\boldsymbol{x}) \le \mathbf{0}, \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}\}$, where $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^l$, we can apply the concepts of the weak ϵ-ANOC and ϵ-ANOC to the problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) \le \mathbf{0},\ \boldsymbol{g}(\boldsymbol{x}) \le \mathbf{0},$$

where $\boldsymbol{g}(\boldsymbol{x}) = \max\{\boldsymbol{h}(\boldsymbol{x}), -\boldsymbol{h}(\boldsymbol{x})\}$. Considering (2.2) with the above equality-free optimization problem and using the definition of $\boldsymbol{g}$, one obtains a sequence $\{\boldsymbol{\lambda}^k\} \subset \mathbb{R}^l$, free of sign, that satisfies

$$\boldsymbol{v}^k \in \left(\mathcal{G}^f_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{i=1}^p \mu_i^k \mathcal{G}^{c_i}_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{j=1}^l \lambda_j^k \mathcal{G}^{h_j}_{\epsilon_k}(\boldsymbol{x}^k)\right) \quad \text{and} \quad i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0.$$

For the ϵ-ANOC, one only needs to replace $\epsilon_k\|\boldsymbol{\mu}^k\|_\infty \to 0$ by $\epsilon_k\|((\boldsymbol{\mu}^k)^T, (\boldsymbol{\lambda}^k)^T)\|_\infty \to 0$.

Although the ϵ-ANOC is more restrictive over the sequence $\{\epsilon_k\}$ than the weak ϵ-ANOC, it is not clear whether the former has practical advantages over the latter. The next example has the goal to show that the ϵ-ANOC is, in fact, stronger than the weak ϵ-ANOC.

Example 2.9. Let us consider the following nonsmooth optimization problem:

$$\min_x \ f(x) = \max\{x, x^2 - 2\} \quad \text{s.t.} \quad c_1(x) = x \leq 0, \ c_2(x) = -x^2 \leq 0.$$

We choose $\{x^k\} \subset \mathbb{R}$ as a sequence converging to $x^* = 0$ and also let $\{\epsilon_k\}$ be such that $0 < \epsilon_k < |x^k|$. So, $2 \notin \mathcal{I}_{\epsilon_k}(x^k)$. Then, for all k sufficiently large such that $|x^k| < 1/2$, and in case $\mu_2^k = 0$, we obtain

$$\mathcal{G}^f_{\epsilon_k}(x^k) + \mu_1^k \mathcal{G}^{c_1}_{\epsilon_k}(x^k) + \mu_2^k \mathcal{G}^{c_2}_{\epsilon_k}(x^k) = 1 + \mu_1^k \geq 1.$$

This shows that if $\{x^k\}$ fulfills the required conditions of the weak ϵ-ANOC, the sequence $\{\epsilon_k\}$ must satisfy $\epsilon_k \geq |x^k|$ for every k large enough. This guarantees that $2 \in \mathcal{I}_{\epsilon_k}(x^k)$. So, for sufficiently large values of ϵ_k, one can always set $\mu_1^k = 0$ and find μ_2^k large enough in order to have $\mathcal{P}\left(0 \mid \mathcal{G}^f_{\epsilon_k}(x^k) + \mu_1^k \mathcal{G}^{c_1}_{\epsilon_k}(x^k) + \mu_2^k \mathcal{G}^{c_2}_{\epsilon_k}(x^k)\right) \to 0$. The previous reasoning ensures that, although x^* is a local maximum point, it satisfies the weak ϵ-ANOC.

However, we state that x^* does not satisfy the stronger condition ϵ-ANOC. By contradiction, suppose x^* satisfies the ϵ-ANOC. Then, in particular, it must fulfill the weak ϵ-ANOC, which gives us $\epsilon_k \geq |x^k|$ for every large k. Also, notice that any element in $\mathcal{G}^{c_2}_{\epsilon_k}(x^k)$ is always greater than $-2(|x^k| + \epsilon_k)$. Combining this with $\epsilon_k \geq |x^k|$ gives $g^k_{c_2} \in \mathcal{G}^{c_2}_{\epsilon_k}(x^k) \Rightarrow g^k_{c_2} \geq -4\epsilon_k$. So, $v^k \in \left(\mathcal{G}^f_{\epsilon_k}(x^k) + \mu_1^k \mathcal{G}^{c_1}_{\epsilon_k}(x^k) + \mu_2^k \mathcal{G}^{c_2}_{\epsilon_k}(x^k)\right)$ implies

$$v^k = 1 + \mu_1^k + \mu_2^k g^k_{c_2} \geq 1 - 4\mu_2^k \epsilon_k. \tag{2.9}$$

Since we supposed that x^* satisfies the ϵ-ANOC, it follows that $\epsilon_k\|\boldsymbol{\mu}^k\|_\infty \to 0$. This last fact together with (2.9) shows that v^k cannot converge to zero, which evinces a contradiction with the statement that x^* satisfies the ϵ-ANOC.

We now prove that our proposed optimality conditions must be at least as strong as the necessary optimality condition presented in [34] (considering $D(\boldsymbol{x}^*) = \mathbb{R}^n$ in [34, Proposition 3.1]), which has its roots in the study presented in [41].

THEOREM 2.10. *If $\boldsymbol{x}^* \in \mathbb{R}^n$ is a feasible point for* (P) *satisfying the (weak) ϵ-ANOC, then the generalized Fritz John optimality conditions hold at $\boldsymbol{x}^*$; i.e., there must exist positive real values $\lambda_0, \lambda_1, \ldots, \lambda_p$, not all simultaneously zero, such that*

$$\mathbf{0} \in \lambda_0 \partial f(\boldsymbol{x}^*) + \sum_{i=1}^p \lambda_i \partial c_i(\boldsymbol{x}^*) \quad \textit{and} \quad \lambda_i c_i(\boldsymbol{x}^*) = 0 \textit{ for all } i \in \{1, \ldots, p\}. \tag{2.10}$$

Proof. For any feasible point $\boldsymbol{x}^* \in \mathbb{R}^n$ satisfying the (weak) ϵ-ANOC, there exist sequences $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$, $\{\epsilon_k\} \subset \mathbb{R}^*_+$, $\{\boldsymbol{v}^k\} \subset \mathbb{R}^n$, and $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}^p_+$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$, $\epsilon_k \downarrow 0$, and $\|\boldsymbol{v}^k\| \to 0$, with

$$\boldsymbol{v}^k = \left(\boldsymbol{v}^k_f + \sum_{i=1}^p \mu_i^k \boldsymbol{v}^k_{c_i}\right) \quad \text{and} \quad i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0, \tag{2.11}$$

where $\boldsymbol{v}_f^k \in \mathcal{G}_{\epsilon_k}^f(\boldsymbol{x}^k) \subset \partial_{\epsilon_k} f(\boldsymbol{x}^k)$ and $\boldsymbol{v}_{c_i}^k \in \mathcal{G}_{\epsilon_k}^{c_i}(\boldsymbol{x}^k) \subset \partial_{\epsilon_k} c_i(\boldsymbol{x}^k)$. Then one of two situations can happen: (a) the sequence $\{\boldsymbol{\mu}^k\}$ is bounded, or (b) the sequence $\{\boldsymbol{\mu}^k\}$ is unbounded.

Let us consider that (a) is the case. Since $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and all functions considered here are locally Lipschitz continuous functions, it follows that $\{\boldsymbol{v}_f^k\}$ and $\{\boldsymbol{v}_{c_i}^k\}$, $i \in \{1, \ldots, p\}$, are all bounded sequences. Therefore, there must exist $\boldsymbol{\mu}^* \in \mathbb{R}_+^p$, $\boldsymbol{v}_f^* \in \mathbb{R}^n$, $\boldsymbol{v}_{c_i}^* \in \mathbb{R}^n$ for $i \in \{1, \ldots, p\}$, and an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that $\boldsymbol{\mu}^k \to_{k\in\mathcal{K}} \boldsymbol{\mu}^*$, $\boldsymbol{v}_f^k \to_{k\in\mathcal{K}} \boldsymbol{v}_f^*$, and $\boldsymbol{v}_{c_i}^k \to_{k\in\mathcal{K}} \boldsymbol{v}_{c_i}^*$, $i \in \{1, \ldots, p\}$. So considering the result of [46, Lemma 3.2(iii)] for the auxiliary functions $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - {\boldsymbol{v}_f^*}^T \boldsymbol{x}$ and $\tilde{c}_i(\boldsymbol{x}) = c_i(\boldsymbol{x}) - {\boldsymbol{v}_{c_i}^*}^T \boldsymbol{x}$, $i \in \{1, \ldots, p\}$, it follows that $\boldsymbol{v}_f^* \in \partial f(\boldsymbol{x}^*)$ and $\boldsymbol{v}_{c_i}^* \in \partial c_i(\boldsymbol{x}^*)$, $i \in \{1, \ldots, p\}$. Therefore, recalling that $\|\boldsymbol{v}^k\| \to 0$, we get $\mathbf{0} \in (\partial f(\boldsymbol{x}^*) + \sum_{i=1}^p \mu_i^* \partial c_i(\boldsymbol{x}^*))$. Notice also that if $c_i(\boldsymbol{x}^*) < 0$ for some i, then $i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k)(\Rightarrow \mu_i^k = 0)$ for all $k \in \mathbb{N}$ sufficiently large, which yields $\mu_i^* = 0$. This ensures the result for case (a).

If (b) holds, then one only has to divide the equation that appears in (2.11) by $\max_i\{\mu_i^k\}$. This ensures that the new multipliers of $\boldsymbol{v}_{c_i}^k$, $i \in \{1, \ldots, p\}$, will be bounded and the multiplier of $\boldsymbol{v}_f^k$ will be $1/\max_i\{\mu_i^k\}$. Hence, the same reasoning used in case (a) can be employed, obtaining the Fritz John conditions with zero being the multiplier of $\partial f(\boldsymbol{x}^*)$. □

Considering the Mangasarian–Fromovitz CQ for nonsmooth problems (see [32, Basic Constraint Qualification] and [42]), which guarantees that (2.10) cannot be true if $\lambda_0 = 0$, and the generalized KKT condition, as defined below, the corollary presented in the sequence follows immediately from the theorem above.

Definition 2.11 (generalized KKT conditions [23, section 6.3]). *A feasible point $\boldsymbol{x}^* \in \mathbb{R}^n$ of* (P) *is said to satisfy the generalized KKT conditions if there exists $\boldsymbol{\mu}^* \in \mathbb{R}_+^p$ such that*

$$0 \in \left(\partial f(\boldsymbol{x}^*) + \sum_{i=1}^p \mu_i^* \partial c_i(\boldsymbol{x}^*) \right) \quad \textit{and} \quad \mu_i^* c_i(\boldsymbol{x}^*) = 0 \textit{ for all } i \in \{1, \ldots, p\}.$$

Corollary 2.12. *If a feasible point $\boldsymbol{x}^* \in \mathbb{R}^n$ for* (P) *satisfies the (weak) ϵ-ANOC and the nonsmooth Mangasarian–Fromovitz CQ holds at $\boldsymbol{x}^*$, then the generalized KKT condition also holds at $\boldsymbol{x}^*$.*

We believe that the results obtained in this section have great relevance in the nonsmooth optimization field, especially when one is concerned with convergence analysis of nonsmooth optimization methods. In the nonsmooth context, it is common to prove the convergence of an algorithm by showing that every accumulation point generated by the iterates of such a method is a stationary point of (Unc-P) (see, for example, [28, 44, 45]). This can be well justified when a property called *calmness* holds at a local minimizer $\boldsymbol{x}^*$ of the original problem (P).

Definition 2.13 (calmness [17, 37]). *Consider $\boldsymbol{x}^*$ a local minimizer of* (P). *We say that problem* (P) *is calm at $\boldsymbol{x}^*$ when one can find $\rho > 0$ sufficiently large such that $\boldsymbol{x}^*$ is also a locally optimal solution of* (Unc-P).

Thus, under the calmness hypothesis, the statement that $\boldsymbol{x}^*$ is a locally optimal solution of (Unc-P) becomes a necessary optimality condition. However, this is not a legitimate necessary optimality condition, in the sense that one needs to assume more than just the fact that $\boldsymbol{x}^*$ is a local minimizer of (P). That is not the case for the (weak) ϵ-ANOC. By the results of Theorems 2.5 and 2.7, we have shown that the (weak) ϵ-ANOC at $\boldsymbol{x}^*$ is a legitimate necessary optimality condition. Therefore, the results exposed here give alternative forms of dealing with the convergence analysis of nonsmooth optimization methods even in the absence of calmness.

We end our theoretical analysis with a brief section presenting the relations between the ϵ-ANOC and the sequential optimality conditions AKKT and CAKKT [3, 7] in the smooth context.

3. ϵ-ANOC applied to smooth optimization. The goal of this section is to show that, under mild assumptions, the scheme from Figure 2 holds true when the objective and constraints functions of (P) are all differentiable. For completeness, we define below the concepts of AKKT and CAKKT for the following smooth optimization problem:

$$\text{(P-smooth)} \qquad \min_{\boldsymbol{x}\in\mathbb{R}^n} \ f(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{c}(\boldsymbol{x}) \leq \boldsymbol{0}, \ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{c} : \mathbb{R}^n \to \mathbb{R}^p$, and $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^l$ are all functions of class C^1.

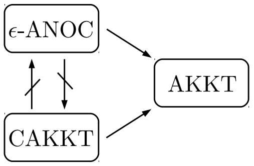


FIG. 2. *Scheme of relations between the ϵ-ANOC and the well-known sequential optimality conditions in the smooth field.*

DEFINITION 3.1 (AKKT, CAKKT [3, 7]). *A feasible point $\boldsymbol{x}^*$ of* (P-smooth) *is an AKKT point if there exist sequences $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$, $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}^p_+$, and $\{\boldsymbol{\lambda}^k\} \subset \mathbb{R}^l$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$, $\lim_{k\to\infty} \min\{-c_i(\boldsymbol{x}^k), \mu_i^k\} = 0$ for all $i \in \{1, \ldots, p\}$, and*

$$\lim_{k\to\infty} \left\| \nabla f(\boldsymbol{x}^k) + \sum_{i=1}^{p} \mu_i^k \nabla c_i(\boldsymbol{x}^k) + \sum_{j=1}^{l} \lambda_j^k \nabla h_j(\boldsymbol{x}^k) \right\| = 0.$$

If, in addition, $\lim_{k\to\infty} \mu_i^k c_i(\boldsymbol{x}^k) = 0$ and $\lim_{k\to\infty} \lambda_j^k h_j(\boldsymbol{x}^k) = 0$ for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, l\}$, then we call $\boldsymbol{x}^*$ a CAKKT point.

First, we prove that every feasible point $\boldsymbol{x}^*$ of any smooth optimization problem satisfying the ϵ-ANOC must also be an AKKT point.

THEOREM 3.2. *Suppose the functions f, $\boldsymbol{c}$, and $\boldsymbol{h}$ in* (P-smooth) *have locally Lipschitz continuous derivatives. If x^* is a feasible point of* (P-smooth) *and satisfies the ϵ-ANOC, then x_* is also an AKKT point.*

Proof. Since x^* satisfies the ϵ-ANOC, there must exist sequences $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$, $\{\epsilon_k\} \subset \mathbb{R}^*_+$, $\{\boldsymbol{v}^k\} \subset \mathbb{R}^n$, $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}^p_+$, and $\{\boldsymbol{\lambda}^k\} \subset \mathbb{R}^l$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$, $\epsilon_k \downarrow 0$, and $\|\boldsymbol{v}^k\| \to 0$, where $\epsilon_k \|((\boldsymbol{\mu}^k)^T, (\boldsymbol{\lambda}^k)^T)\|_\infty \to 0$, $i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0$, and

$$\boldsymbol{v}^k \in \left(\mathcal{G}^f_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{i=1}^{p} \mu_i^k \mathcal{G}^{c_i}_{\epsilon_k}(\boldsymbol{x}^k) + \sum_{j=1}^{l} \lambda_j^k \mathcal{G}^{h_j}_{\epsilon_k}(\boldsymbol{x}^k) \right). \tag{3.1}$$

Since the derivatives of the involved functions are locally Lipschitz continuous, it follows that

$$\boldsymbol{v}^k = \nabla f(\boldsymbol{x}^k) + \sum_{i=1}^{p} \mu_i^k \nabla c_i(\boldsymbol{x}^k) + \sum_{j=1}^{l} \lambda_j^k \nabla h_j(\boldsymbol{x}^k) + \boldsymbol{r}_f^k + \sum_{i=1}^{p} \mu_i^k \boldsymbol{r}_{c_i}^k + \sum_{j=1}^{l} \lambda_j^k \boldsymbol{r}_{h_j}^k,$$

where $\boldsymbol{r}_f^k$, $\boldsymbol{r}_{c_i}^k$, and $\boldsymbol{r}_{h_j}^k$ are all error vectors of order ϵ_k. Since $\|\boldsymbol{v}^k\| \to 0$, $\epsilon_k \downarrow 0$, and $\epsilon_k \|((\boldsymbol{\mu}^k)^T, (\boldsymbol{\lambda}^k)^T)\|_\infty \to 0$, we get

$$\nabla f(\boldsymbol{x}^k) + \sum_{i=1}^{p} \mu_i^k \nabla c_i(\boldsymbol{x}^k) + \sum_{j=1}^{l} \lambda_j^k \nabla h_j(\boldsymbol{x}^k) \to \boldsymbol{0}. \tag{3.2}$$

Additionally, because $i \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k) \Rightarrow \mu_i^k = 0$, one can see that $\min\{-c_i(\boldsymbol{x}^k), \mu_i^k\} \to 0$ for all $i \in \{1, \ldots, p\}$. This together with (3.2) fulfills the conditions of an AKKT point, which completes the proof. □

Remark 3.3. It is worth noticing that the reciprocal is not true, i.e., AKKT does not imply the ϵ-ANOC in the smooth case. Indeed, considering the optimization problem given in Example 2.9, $x^* = 0$ is an AKKT point (around x^* the optimization problem is smooth, which justifies applying the AKKT concept at this point). However, we have shown that x^* does not satisfy the ϵ-ANOC, which guarantees that AKKT$\not\Rightarrow$ ϵ-ANOC. Hence, our new sequential optimality condition is also stronger than AKKT in the smooth context.

We now proceed with two examples showing that CAKKT and the ϵ-ANOC are not connected to each other.

Example 3.4. Let us consider the following constrained optimization problem:

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = x_1 - x_2 \quad \text{s.t.} \;\; x_1 \geq 0, \; x_2 \geq 0, \; x_1 x_2 \leq 0.$$

Although the point $\boldsymbol{x}^* = \boldsymbol{0}$ is not a local minimizer for the problem, if one chooses $\boldsymbol{x}^k = (1/k, -1/k)^T$, $\mu_1^k = \mu_2^k = 0$, and $\mu_3^k = k$, one can see that $\boldsymbol{x}^*$ is a CAKKT point [4]. However, we state that $\boldsymbol{x}^*$ does not fulfill the ϵ-ANOC.

Indeed, suppose, by contradiction, that $\boldsymbol{x}^*$ satisfies the ϵ-ANOC. Recalling the arguments used to go from (3.1) to (3.2), and since the objective and constraint functions are all smooth and have locally Lipschitz continuous derivatives, there must exist $\{\boldsymbol{x}^k\} = \{(x_1^k, x_2^k)^T\}$ and $\{\boldsymbol{\mu}^k\} \subset \mathbb{R}_+^3$ with $\epsilon_k \|\boldsymbol{\mu}^k\|_\infty \to 0$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$ and

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} + \mu_1^k \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \mu_2^k \begin{bmatrix} 0 \\ -1 \end{bmatrix} + \mu_3^k \begin{bmatrix} x_2^k \\ x_1^k \end{bmatrix} \to \boldsymbol{0}.$$

By the second line of the above relation, we have that, for large values of k, the following must hold: $x_1^k > 0$ and $\mu_3^k > 1/(2x_1^k)$. Since $\epsilon_k \mu_3^k \to 0$, it yields $\epsilon_k = o(x_1^k)$. Hence, $1 \notin \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k)$ and, consequently, $\mu_1^k = 0$ for large values of k. So, by the first line, we obtain $\mu_3^k x_2^k \to -1$, which ensures $x_2^k < 0$ and $\mu_3^k > 1/(2|x_2^k|)$ for k sufficiently large. Since μ_3^k can only be different from zero if $3 \in \mathcal{I}_{\epsilon_k}(\boldsymbol{x}^k)$, the value ϵ_k must be greater than $\min\{x_1^k, |x_2^k|\}$ for large values of k. This contradicts the fact that $\epsilon_k \|\boldsymbol{\mu}^k\|_\infty \to 0$, which allows us to conclude that $\boldsymbol{x}^*$ does not satisfy the ϵ-ANOC.

Example 3.5. We now consider the smooth optimization problem

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \frac{(x_2 - 2)^2}{2} \quad \text{s.t.} \;\; x_1 = 0, \; x_1 x_2 = 0.$$

As shown in [7], the point $\boldsymbol{x}^* = (0,1)^T$ is not a CAKKT point. However, if we choose $\boldsymbol{x}^k = (1/k, 1)^T$, $\boldsymbol{\lambda}^k = (-k, k)^T$, and $\epsilon_k = 1/k^2$, we obtain that the ϵ-ANOC is fulfilled as defined in Remark 2.8. Therefore, $\boldsymbol{x}^*$ satisfies the ϵ-ANOC.

The above examples show that, in fact, the ϵ-ANOC is a new legitimate (i.e., that does not require any kind of CQ) necessary optimality condition even in the case where all functions are smooth. In addition, because the ϵ-ANOC implies the AKKT condition in the smooth scenario, very general CQs may be used together with the ϵ-ANOC to ensure convergence to a KKT point—for instance, the *cone-continuity property* (CCP) [5] is a valid CQ.

Finally, before we proceed to the next section, it is worth establishing a relation between the ϵ-ANOC and the very recently sequential optimality condition called *positive approximate KKT* (PAKKT) [2]. Indeed, Example 3.5 can be used to show that ϵ-ANOC does not imply PAKKT. The other implication is yet to be investigated.

4. A practical algorithm.

The goal of this section is to present an algorithm that is able to generate a sequence of iterates satisfying the sequential optimality conditions introduced in the last section. Looking at relation (2.2) and the proof of Theorem 2.5, one can easily come up with the following algorithm: given $\xi_k \downarrow 0$ and $\rho \in \mathbb{R}$ sufficiently large, whenever possible, set

$$\boldsymbol{x}^k \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} \Psi_{\xi_k,\rho}(\boldsymbol{x}). \tag{4.1}$$

Although conceptually simple, the above procedure cannot be used in a practical method. For optimization problems involving nonconvex functions, an implementable algorithm usually guarantees that the method will only find stationary points for the problem, not minimizers. Moreover, the set of minimizers of the function $\Psi_{\xi_k,\rho}$ may be empty; therefore, this procedure may not be well defined.

We thus present Algorithm 4.1 (`PACNO` - penalized algorithm for constrained nonsmooth optimization), where we avoid these issues by using a solution of a relaxed version of (4.1) (see Remark 4.7 for a more detailed discussion about this algorithm) where we trade the global minimization for the easier problem of Step 1 of Algorithm 4.1, which is concerned only with approximate stationary points. This is sufficient to ensure that Algorithm 4.1 generates a sequence of iterates that verifies the ϵ-ANOC—Theorem 4.6 presents a mathematical proof of this fact whenever $\theta_\epsilon < \theta_\xi$. Therefore, the ϵ-ANOC is not explicitly verified along the execution of Algorithm 4.1, since Step 1 implicitly ensures this condition.

A matter of concern here is how the condition required in Step 1 of Algorithm 4.1 can be guaranteed in practice. Since the definition of the set $\mathcal{G}_\epsilon^f(\boldsymbol{x})$ [20], which culminated in the nonsmooth optimization method called *gradient sampling* (GS) [21, 46], finding an element that satisfies (4.2) has turned from being only an idealized step to a procedure that can be practically executed by a computer; see section 5 for numerical experiments. The GS method is a monotone decreasing algorithm that works by sampling points around ϵ-neighborhoods of its current iterate $\boldsymbol{y}^k$ in order to have local gradient information of the function Υ that the user seeks to minimize. Under the assumption that Υ is Lipschitz continuous, it is possible to show that, with probability one, all the sampled points will lie in the differentiable set of such a function. Therefore, by finding the vector of minimum norm over the convex hull of these gradients, the method tries to approximate the element $\mathcal{P}(\mathbf{0} \mid \mathcal{G}_\epsilon^\Upsilon(\boldsymbol{y}^k))$. The algorithm stops when it has found an iterate $\boldsymbol{y}^k$ at which $\|\mathcal{P}(\mathbf{0} \mid \mathcal{G}_\epsilon^\Upsilon(\boldsymbol{y}^k))\|$ is less than a pre-established tolerance $\nu > 0$. Consequently, when one sets $\Upsilon := \Psi_{\xi_k,\rho_k}$,

Algorithm 4.1. Penalized algorithm for constrained nonsmooth optimization.

```
Step 0. Choose $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\rho_1 \in [0,+\infty)$, $M \in (0,+\infty)$, $\{\theta_\xi, \theta_\epsilon, \theta_\nu\} \subset (0,1)$, $\omega \in (0,1]$, $\{\xi_1, \epsilon_1, \nu_1\} \subset (0,+\infty)$, $\xi_{\text{opt}} \in [0,\xi_1]$, $\epsilon_{\text{opt}} \in [0,\epsilon_1]$, and $\nu_{\text{opt}} \in [0,\nu_1]$.
        Set $k = 1$.
Step 1. Define
        (4.2)  $\mathcal{S}^k_\Psi := \{\boldsymbol{x} : \Psi_{\xi_k,\rho_k}(\boldsymbol{x}) \le \Psi_{\xi_k,\rho_k}(\boldsymbol{x}^{k-1})$ and $\|\mathcal{P}(\boldsymbol{0} \mid \partial_{\epsilon_k}\Psi_{\xi_k,\rho_k}(\boldsymbol{x}))\| \le \nu_k\}$.
        If $\mathcal{S}^k_\Psi \neq \emptyset$, then set $\boldsymbol{x}^k$ as any point in $\mathcal{S}^k_\Psi$. Otherwise, set $\boldsymbol{x}^k$ as any point $\boldsymbol{x}$ satisfying $\|\boldsymbol{c}(\boldsymbol{x})_+\|_1 \le \|\boldsymbol{c}(\boldsymbol{x}^{k-1})_+\|_1$ and $\|\mathcal{P}(\boldsymbol{0} \mid \partial_{\epsilon_k}\|\boldsymbol{c}(\boldsymbol{x})_+\|_1)\| \le \nu_k$, and go to Step 3.
Step 2. If $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 < \xi_{\text{opt}}$, then
            terminate if                        $\epsilon_k \le \epsilon_{\text{opt}}$ and $\nu_k \le \nu_{\text{opt}}$,
            set $\xi_{k+1} = \xi_k$ and go to Step 5    otherwise.
Step 3. If $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 > \omega\xi_k$ with $\epsilon_k \le \epsilon_{\text{opt}}$ and $\nu_k \le \nu_{\text{opt}}$, then stop. The iterate $\boldsymbol{x}^k$ is close to a stationary point of the infeasibility measure.
Step 4. Set $\xi_{k+1} = \xi_k - (1-\theta_\xi)[\xi_k - \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1]_+$.
Step 5. If $f(\boldsymbol{x}^k) - \rho_k \le -M$, then set $\rho_{k+1} = \rho_k$. Otherwise, proceed with $\rho_{k+1} = \rho_k + 2(M + [f(\boldsymbol{x}^k) - \rho_k]_+)$.
Step 6. Set $\epsilon_{k+1} = \theta_\epsilon\epsilon_k$, $\nu_{k+1} = \theta_\nu\nu_k$, and $k \leftarrow k+1$. Go back to Step 1.
```

$\epsilon := \epsilon_k$, and $\nu := \nu_k$, the requirement (4.2) is achieved, with probability one, in a finite number of iterations of the GS method. For more details on the functioning of this algorithm, we guide the reader to [19, 21, 46]. It should be mentioned that although the GS method fits well with Step 1, it is not the only technique that can be applied to perform this step; for example, the recent GRANSO algorithm [27] may also be used; see the numerical experiments in section 5. Therefore, the PACNO algorithm should not necessarily be viewed as a GS-like technique.

We should also explain our strategy for updating ξ_k in Step 4. Such a procedure ensures that, when PACNO finds an iterate $\boldsymbol{x}^k$ that possesses an infeasibility measure less than the target value ξ_k, the next value ξ_{k+1} will lie between ξ_k and $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1$; see Figure 3. However, when $\boldsymbol{x}^k$ is an iterate at which $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1$ is greater than ξ_k, it makes no sense to decrease the target value of infeasibility. Therefore, Step 4 sets $\xi_{k+1} = \xi_k$ in this last case.

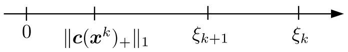


FIG. 3. *An illustration of the update that appears inside Step* 4 *of Algorithm* 4.1 *whenever* $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 < \xi_k$.

Another parameter value that plays a key role in our algorithm is ρ_k. The difference $f(\boldsymbol{x}^k) - \rho_k$ should be kept sufficiently negative. In case PACNO performs an iteration in which this condition does not hold, Step 5 tries to predict a value for ρ_{k+1} in order to ensure the desired condition for the next iteration.

To proceed with our analysis, first we need to show that the set $\mathcal{S}^k_\Psi$ defined in (4.2) will eventually be nonempty and, moreover, that the sequence $\{\rho_k\}$ generated by the algorithm will stabilize at a sufficiently large value. To guarantee that $\mathcal{S}^k_\Psi \neq \emptyset$, we need to consider one extra assumption.

Assumption 4.1. There exists $\alpha > 0$ such that $\mathcal{F}_\alpha := \{x \in \mathbb{R}^n : \|\boldsymbol{c}(\boldsymbol{x})_+\|_1 \leq \alpha\}$ is compact.

Since we are dealing with possibly nonconvex functions, it is not sufficient to ask for the compactness of the feasible set, but, instead, one needs to ask for the compactness of the perturbed feasible set $\mathcal{F}_\alpha$. This is required to exclude some pathological functions like the one illustrated in Figure 4, where the feasible set is compact but there is no $\alpha > 0$ ensuring the compactness of $\mathcal{F}_\alpha$. Nevertheless, the assumption can be easily ensured if the feasible points are additionally restricted to box constraints. Noticing that unbounded solutions do not occur for the majority of real problems, the user may include sufficiently large artificial box constraints to enforce the condition.

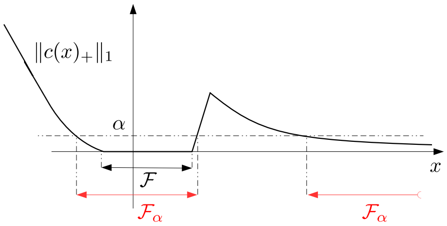


FIG. 4. *An illustration of a function that does not have a compact $\mathcal{F}_\alpha$. The set $\mathcal{F}$ stands for the feasible region. Notice that one cannot find $\alpha > 0$ that ensures the compactness of $\mathcal{F}_\alpha$.*

We start by showing a technical lemma that will be helpful in the subsequent results.

LEMMA 4.2. *Let $\{\boldsymbol{x}^k\}$ be a sequence of iterates generated by Algorithm* 4.1 *with $\xi_{opt} = \epsilon_{opt} = \nu_{opt} = 0$. Then, there must exist a sequence $\{\delta_k\} \subset \mathbb{R}^*_+$ converging to zero such that $\xi_k \leq \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 + \delta_k$. Moreover, if* $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 = 0$, *then* $\xi_k \downarrow 0$.

Proof. Since $\xi_{\text{opt}} = \epsilon_{\text{opt}} = \nu_{\text{opt}} = 0$, the algorithm will have infinitely many iterations. Moreover, notice that $\{\xi_k\}$ is a positive bounded monotone decreasing sequence, which ensures that such a sequence must converge to some positive real value $\bar{\xi}$. Hence, since ξ_k is always updated by Step 4 (when $\xi_{\text{opt}} = 0$), we have

$$\begin{aligned}\xi_{k+1} = \xi_k - (1-\theta_\xi)[\xi_k - \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1]_+ &\Rightarrow (1-\theta_\xi)[\xi_k - \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1]_+ = \xi_k - \xi_{k+1}\\ &\Rightarrow \lim_{k\to\infty}[\xi_k - \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1]_+ = 0.\end{aligned}$$

In other words, there exists $\{\delta_k\} \subset \mathbb{R}^*_+$, with $\delta_k \to 0$, such that $\xi_k \leq \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 + \delta_k$. So, if $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 = 0$, it is straightforward to see that $\xi_k \downarrow 0$. □

The next lemma is useful to prove the result that, eventually, the set $\mathcal{S}^k_\Psi$ will have at least one element. It gives a sufficient condition for the existence of a global minimizer of $\Psi_{\xi,\rho}$.

LEMMA 4.3. *Let ρ be a real number such that*

$$\min_{\boldsymbol{x}\in\mathcal{F}_{\alpha/2}} f(\boldsymbol{x}) - \rho < 0$$

and $\xi \in (0, \alpha]$, with α being a strictly positive real number satisfying Assumption 4.1. *Then, the function $\Psi_{\xi,\rho}$ defined in* (2.3) *has a global minimizer.*

Proof. Consider any fixed $\xi \in (0, \alpha]$. Let us first show that $\Psi_{\xi,\rho}$ is bounded from below. By contradiction, let us suppose the opposite, i.e., we can find a sequence $\{\boldsymbol{x}^k\}$ such that $\Psi_{\xi,\rho}(\boldsymbol{x}^k) \leq -k$. Since $-k$ is a negative number, we must have, for all $k \in \mathbb{N}$, that $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \leq \xi$, since, otherwise, we would have $\Psi_{\xi,\rho}(\boldsymbol{x}^k) = \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \geq 0$. Because $\xi \in (0, \alpha]$, this means that $\boldsymbol{x}^k \in \mathcal{F}_\alpha$ for all $k \in \mathbb{N}$. However, $\Psi_{\xi,\rho}$ is a continuous function, which yields that it must assume a minimum value on $\mathcal{F}_\alpha$. This is a contradiction with the fact that $\Psi_{\xi,\rho}(\boldsymbol{x}^k) \leq -k$ for all $k \in \mathbb{N}$. Therefore, the function $\Psi_{\xi,\rho}$ must be bounded.

Now, let us then consider a sequence $\{\boldsymbol{z}^k\}$ such that $\Psi_{\xi,\rho}(\boldsymbol{z}^k) \to \inf_{x\in\mathbb{R}^n} \Psi_{\xi,\rho}(\boldsymbol{x})$. By hypothesis, we know that there exists $\hat{\boldsymbol{x}} \in \mathcal{F}_{\alpha/2}$ such that $f(\hat{\boldsymbol{x}}) - \rho < 0$. This implies that

$$\inf_{\boldsymbol{x}\in\mathbb{R}^n} \Psi_{\xi,\rho}(\boldsymbol{x}) \leq \Psi_{\xi,\rho}(\hat{\boldsymbol{x}}) \leq \|\boldsymbol{c}(\hat{\boldsymbol{x}})_+\|_1.$$

Therefore, for any large $k \in \mathbb{N}$, we must have $\Psi_{\xi,\rho}(\boldsymbol{z}^k) \leq 2\|\boldsymbol{c}(\hat{\boldsymbol{x}})_+\|_1$. Since, $\hat{\boldsymbol{x}} \in \mathcal{F}_{\alpha/2}$, it implies that $\Psi_{\xi,\rho}(\boldsymbol{z}^k) \leq \alpha$. Recalling that $\xi \leq \alpha$, we must have $\boldsymbol{z}^k \in \mathcal{F}_\alpha$ for any large $k \in \mathbb{N}$. So, because $\mathcal{F}_\alpha$ is compact, one can ensure that there exists a subsequence of $\{\boldsymbol{z}^k\}$ converging to $\overline{\boldsymbol{z}} \in \mathcal{F}_\alpha$ such that $\Psi_{\xi,\rho}(\overline{\boldsymbol{z}}) = \inf_{x\in\mathbb{R}^n} \Psi_{\xi,\rho}(\boldsymbol{x})$, which ends the proof. □

The following result is a technical lemma, and it proves that $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1$ will be related with the value ξ_k and the infeasibility measure of the past iteration. It also ensures that our method will not suffer from the greediness phenomenon.

LEMMA 4.4. *Let $k \in \mathbb{N}$ be any iteration of Algorithm* 4.1. *Then,*

$$\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \leq \max\left\{\xi_k, \|\boldsymbol{c}(\boldsymbol{x}^{k-1})_+\|_1\right\}.$$

Proof. By contradiction, suppose that there exists an iteration k such that

$$\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 > \max\left\{\xi_k, \|\boldsymbol{c}(\boldsymbol{x}^{k-1})_+\|_1\right\}. \tag{4.3}$$

Then, because of Step 1, we have $\boldsymbol{x}^k \in \mathcal{S}^k_\Psi$. Therefore, $\Psi_{\xi_k,\rho_k}(\boldsymbol{x}^k) \leq \Psi_{\xi_k,\rho_k}(\boldsymbol{x}^{k-1})$. Moreover, because of Step 5, one can see that $f(\boldsymbol{x}^{k-1}) - \rho_k < 0$, which implies

$$\Psi_{\xi_k,\rho_k}(\boldsymbol{x}^k) \leq \Psi_{\xi_k,\rho_k}(\boldsymbol{x}^{k-1}) \leq \|\boldsymbol{c}(\boldsymbol{x}^{k-1})_+\|_1.$$

Thus, by (4.3), we have $\Psi_{\xi_k,\rho_k}(\boldsymbol{x}^k) < \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1$. Consequently, the above inequality can only be true in the case that $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 < \xi_k$, which contradicts (4.3). □

The next result tells us that when the sequence of iterates approaches the feasible set, then, eventually, the set $\mathcal{S}^k_\Psi$ will be nonempty and the sequence $\{\rho_k\}$ will be constant for large values of $k \in \mathbb{N}$.

LEMMA 4.5. *Let $\{\boldsymbol{x}^k\}$ be a sequence of iterates generated by Algorithm* 4.1 *with $\xi_{opt} = \epsilon_{opt} = \nu_{opt} = 0$. If $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 = 0$, then there exist a sufficiently large $\hat{k} \in \mathbb{N}$ and $\overline{\rho} \in \mathbb{R}$ such that, for all $k \geq \hat{k}$, we have*

$$\mathcal{S}^k_\Psi \neq \emptyset \quad \textit{and} \quad \rho_k = \overline{\rho}.$$

Proof. First, let us prove that $\mathcal{S}^k_\Psi \neq \emptyset$. Since $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 = 0$, we must have an iteration $\hat{k}$ such that $\|\boldsymbol{c}(\boldsymbol{x}^{\hat{k}})_+\|_1 \leq \alpha/2$, where α is the positive real number satisfying Assumption 4.1. Consequently, by Step 5 of the algorithm, one can see that

$$\min_{\boldsymbol{x}\in\mathcal{F}_{\alpha/2}} f(\boldsymbol{x}) - \rho_k \leq f(\boldsymbol{x}^{\hat{k}}) - \rho_k < 0 \quad \text{ for any } k > \hat{k}.$$

Additionally, Lemma 4.2 tells us that $\xi_k \to 0$. So, by Lemma 4.3, we see, for any large $k \in \mathbb{N}$, that $\mathcal{S}_\Psi^k \neq \emptyset$.

Let us now prove that there exists $\overline{\rho} \in \mathbb{R}$ such that $\rho_k = \overline{\rho}$ for any large $k \in \mathbb{N}$. By contradiction, let us assume that the statement is false. This means, by the way the algorithm was designed, that there exists an infinite index set $\mathcal{K} \subset \mathbb{N}$ such that

$$f(\boldsymbol{x}^k) - \rho_k > -M \quad \text{for all } k \in \mathcal{K}, \tag{4.4}$$

implying that $\rho_k \to \infty$. Because $\liminf_{k\to\infty} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 = 0$, it means that there exists an iteration $\hat{k}$ satisfying $\|\boldsymbol{c}(\boldsymbol{x}^{\hat{k}})_+\|_1 \leq \alpha$ and $\xi_{\hat{k}} \leq \alpha$ (because of Lemma 4.2). Consequently, due to Lemma 4.4 and recalling that $\{\xi_k\}$ is a monotone decreasing sequence, we have

$$\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \leq \alpha, \text{ i.e., } \boldsymbol{x}^k \in \mathcal{F}_\alpha \quad \text{for all } k \geq \hat{k}.$$

However, by hypothesis, $\mathcal{F}_\alpha$ is compact, which, together with the assumption $\rho_k \to \infty$, implies

$$\rho_k \geq \max_{\boldsymbol{x}\in\mathcal{F}_\alpha} f(\boldsymbol{x}) + M \geq f(\boldsymbol{x}^k) + M \quad \text{for any large } k \in \mathbb{N}. \tag{4.5}$$

This is a contradiction with (4.4). Therefore, there must exist $\overline{\rho}$ such that $\rho_k = \overline{\rho}$ for any large $k \in \mathbb{N}$. □

We are now ready to present the convergence theorem of Algorithm 4.1. Because the nonsmooth optimization problem in hand might involve nonconvex functions, the result ensures that a cluster point $\boldsymbol{x}^*$ of the iterate sequence $\{\boldsymbol{x}^k\}$ can be a stationary point for the infeasibility measure $\|\boldsymbol{c}(\cdot)_+\|_1$ or satisfy the ϵ-ANOC.

THEOREM 4.6. *If $\{\boldsymbol{x}^k\}$ is the sequence of iterates generated by Algorithm* 4.1 *with $\xi_{opt} = \epsilon_{opt} = \nu_{opt} = 0$ and $\theta_\epsilon \in (0, \theta_\xi)$, then, given any infinite index set $\mathcal{K} \subset \mathbb{N}$ such that $\boldsymbol{x}^k \to_{k\in\mathcal{K}} \boldsymbol{x}^*$, for some $\boldsymbol{x}^* \in \mathbb{R}^n$, one of the following statements must be true:*

(a) *$\boldsymbol{x}^*$ satisfies $0 \in \partial\|\boldsymbol{c}(\boldsymbol{x}^*)_+\|_1$.*

(b) *$\boldsymbol{x}^*$ is a feasible point for* (P)*, and it satisfies the ϵ-ANOC. Moreover, there exists $\overline{\rho} \in \mathbb{R}$ such that $\rho_k = \overline{\rho}$ for any large $k \in \mathbb{N}$.*

Proof. We start by noticing that, since $\xi_{\text{opt}} = \epsilon_{\text{opt}} = \nu_{\text{opt}} = 0$, Algorithm 4.1 in fact generates an infinite sequence of points $\{\boldsymbol{x}^k\} \subset \mathbb{R}^n$. Moreover, ξ_k is always updated by Step 4 (since $\xi_{\text{opt}} = 0$), and hence $\xi_{k+1} = \xi_k - (1-\theta_\xi)[\xi_k - \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1]_+ \geq \theta_\xi \xi_k$. Therefore, because we suppose $\theta_\epsilon \in (0, \theta_\xi)$, it yields

$$\frac{\epsilon_{k+1}}{\xi_{k+1}} \leq \eta \frac{\epsilon_k}{\xi_k} \leq \eta^2 \frac{\epsilon_{k-1}}{\xi_{k-1}} \leq \cdots \leq \eta^k \frac{\epsilon_1}{\xi_1}, \text{ where } \eta = \frac{\theta_\epsilon}{\theta_\xi} < 1,$$

which ensures $\epsilon_k/\xi_k \to 0$. We then divide the proof in the following cases:

(i) $\liminf_{k\in\mathcal{K}} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1/\xi_k \geq 1$, and there exists an infinite index set $\hat{\mathcal{K}} \subset \mathcal{K}$ such that $\mathcal{S}_\Psi^k = \emptyset$ for all $k \in \hat{\mathcal{K}}$;

(ii) $\liminf_{k\in\mathcal{K}} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1/\xi_k \geq 1$ and $\mathcal{S}_\Psi^k \neq \emptyset$ for all $k \in \mathcal{K}$ sufficiently large;

(iii) $\liminf_{k\in\mathcal{K}} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1/\xi_k < 1$.

Suppose (i) holds. By Step 1, it follows that $\left\|\mathcal{P}\left(\mathbf{0} \mid \partial_{\epsilon_k}\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1\right)\right\| \to_{k\in\hat{\mathcal{K}}} 0$. Consequently, since $\epsilon_k \to 0$, [46, Lemma 3.2(iii)] guarantees that $\mathbf{0} \in \partial\|\boldsymbol{c}(\boldsymbol{x}^*)_+\|_1$.

Assume now that case (ii) holds. This yields

$$\max\left\{1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1}{\xi_k}, 0\right\} \underset{k\in\mathcal{K}}{\to} 0. \tag{4.6}$$

Additionally, by the way Algorithm 4.1 updates the value ρ_k, we must have $\rho_k \geq f(\boldsymbol{x}^*) + M$ for any large iteration k. Recalling $\nu_k \downarrow 0$, $\epsilon_k \downarrow 0$ (because $\nu_{\text{opt}} = 0$ and $\epsilon_{\text{opt}} = 0$), $\mathcal{S}_\Psi^k \neq \emptyset$ for any large iteration $k \in \mathcal{K}$, and the result of Lemma 2.3, it follows, for any large $k \in \mathcal{K}$, that there exist $\{\boldsymbol{x}^{k,j}\}_{j=1}^{n+1} \subset \mathcal{B}(\boldsymbol{x}^k, \epsilon_k)$, and $\boldsymbol{\lambda}^k \in \mathbb{R}^{n+1}$ with $\boldsymbol{e}^T \boldsymbol{\lambda}^k = 1$, such that there exists $\boldsymbol{r}^k \in \partial \Psi_{\xi_k, \rho_k}(\boldsymbol{x}^k)$ with $\|\boldsymbol{r}^k\| \to 0$ and

$$\boldsymbol{r}^k \in \sum_{j=1}^{n+1} \lambda_j^k \left(\max\left\{ 1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0 \right\} \partial f(\boldsymbol{x}^{k,j}) + \sigma_j^k \partial \|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1 \right), \tag{4.7}$$

where $\sigma_j^k \geq 1$, $j \in \{1, \ldots, n+1\}$. Due to (4.6), $\|\boldsymbol{x}^k - \boldsymbol{x}^{k,j}\| \leq \epsilon_k$, and $\epsilon_k / \xi_k \to 0$, we get, for any $j \in \{1, \ldots, n+1\}$,

$$\max\left\{ 1 - \frac{\|\boldsymbol{c}(\boldsymbol{x}^{k,j})_+\|_1}{\xi_k}, 0 \right\} \underset{k \in \mathcal{K}}{\to} 0.$$

The above limit together with (4.7) tells us that $\| \mathcal{P}(\boldsymbol{0} \mid \partial_{\epsilon_k} \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1)\| \to_{k \in \mathcal{K}} 0$. So, since $\epsilon_k \downarrow 0$, we obtain $\boldsymbol{0} \in \partial \|\boldsymbol{c}(\boldsymbol{x}^*)_+\|_1$ due to [46, Lemma 3.2(iii)].

Finally, we assume the validity of case (iii), and we claim $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \to 0$. Indeed, because $\{\xi_k\}$ is a monotone decreasing and bounded sequence, it must follow that $\xi_k \downarrow \bar{\xi}$ for some $\bar{\xi} \geq 0$. According to Lemma 4.2, there exists $\{\delta_k\} \subset \mathbb{R}_+^*$ satisfying $\delta_k \to 0$ such that

$$\xi_k \leq \|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 + \delta_k \Rightarrow 1 \leq \liminf_{k \in \mathcal{K}} \frac{\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1}{\xi_k} + \liminf_{k \in \mathcal{K}} \frac{\delta_k}{\xi_k} \Rightarrow \liminf_{k \in \mathcal{K}} \frac{\delta_k}{\xi_k} > 0.$$

This yields $\bar{\xi} = 0$. However, by the way we have designed our algorithm, this only happens if there exists an infinite index set $\hat{\mathcal{K}} \subset \mathbb{N}$ with $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \to_{k \in \hat{\mathcal{K}}} 0$. Therefore, given any $\tau > 0$, there exists $\hat{k}$ such that $\|\boldsymbol{c}(\boldsymbol{x}^{\hat{k}})_+\|_1 < \tau$ and $\xi_{\hat{k}} < \tau$. Because of Lemma 4.4, this ensures $k \geq \hat{k} \Rightarrow \|\boldsymbol{c}(\boldsymbol{x}^k)_+\| \leq \tau$. Since $\tau > 0$ is arbitrary, we obtain $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 \to 0$, which guarantees that $\boldsymbol{x}^*$ is a feasible point.

Then, Lemma 4.5 ensures that, for all k large enough, we have $\mathcal{S}_\Psi^k \neq \emptyset$ and $\rho_k = \overline{\rho}$ for some $\overline{\rho} \in \mathbb{R}$. Notice that Step 5 ensures $f(\boldsymbol{x}^*) - \overline{\rho} \leq -M$. This information, together with Step 1, Lemma 2.4, the limit $\epsilon_k / \xi_k \to 0$, and the reasoning used in the proof of Theorem 2.7, proves that the cluster point $\boldsymbol{x}^*$ satisfies the ϵ-ANOC. ∎

Remark 4.7. Some comments regarding Algorithm 4.1 are in order:

(a) The technique based on sampling points to approximate $\mathcal{G}_\epsilon(\boldsymbol{x})$ that originated the method known as GS [21, 46] and its recent variants [29, 30, 38, 39, 40, 50] have shown to be effective tools for minimizing nonsmooth and nonconvex functions. The good results of those methods ensure that Step 1 is not just an idealized step, but it can be performed in practice.[1] In addition, BFGS techniques, which are not guaranteed to converge but work well in practice, may be used as well [27, 48]. Both possibilities were considered in section 5.

(b) Looking carefully at the proof of Theorem 4.6, we can understand the importance of Step 3 inside Algorithm 4.1. When $\|\boldsymbol{c}(\boldsymbol{x}^k)_+\|_1 / \xi_k$ is greater than or approaching one—cases (i) and (ii) inside the proof—this implies that we are losing information about the objective function along the iterations, meaning that the method is tending to a stationary point of the infeasibility measure. Therefore, this suggests that one should choose $\omega \approx 1$.

[1] Assuming that f and c_i, $i \in \{1, \ldots, p\}$, are continuously differentiable in full-measure open subsets of $\mathbb{R}^n$, all hypotheses required by the convergence theory of the GS method are satisfied.

(c) For the case that problem (P) satisfies *calmness* [17] and ρ is large enough, every local minimizer of (P) is also a local minimizer of $\Psi_{\xi,\rho}$ for every ξ small enough (this can be easily proven by following the same reasoning used in the proof of [23, Proposition 6.4.3]). This ensures that, in many cases, PACNO will find a good approximation to the solution of the problem without needing to bring ξ_k down to zero.

(d) When Algorithm 4.1 is applied to a smooth optimization problem, the statement of Theorem 4.6 remains the same.

5. Numerical results.

This section has the goal to illustrate different properties of PACNO. Subsection 5.1 is devoted to showing that, since our method is based on a legitimate necessary optimality condition (i.e., our convergence result does not depend on any kind of CQ), PACNO may achieve good convergence to the solution of the nonsmooth optimization problem even in the absence of calmness. Additionally, we exhibit a practical example in which PACNO prevents the greediness phenomenon. Subsection 5.2 reveals that our method may converge even when the optimization problem does not satisfy all of our convergence hypotheses. Finally, subsection 5.3 aims to illustrate that Step 1 of PACNO may accept different solvers to find a stationary point of $\Psi_{\xi,\rho}$. The tests were performed in a notebook DELL Latitude 7490, processor Intel Core i7-8650U, CPU 2.11GHz, with 16GB RAM (64-bit) using MATLAB R2018a.

5.1. Nonsmooth optimization in the absence of calmness.

In the previous sections, we have discussed the theoretical benefits of applying the penalization strategy used in $\Psi_{\xi,\rho}$ over the exact penalization approach to produce a practical necessary optimality condition. However, one can wonder whether our penalization idea has any practical advantage when compared to the standard exact penalization procedure. Aiming to elucidate this matter, we present a simple nonsmooth optimization problem:

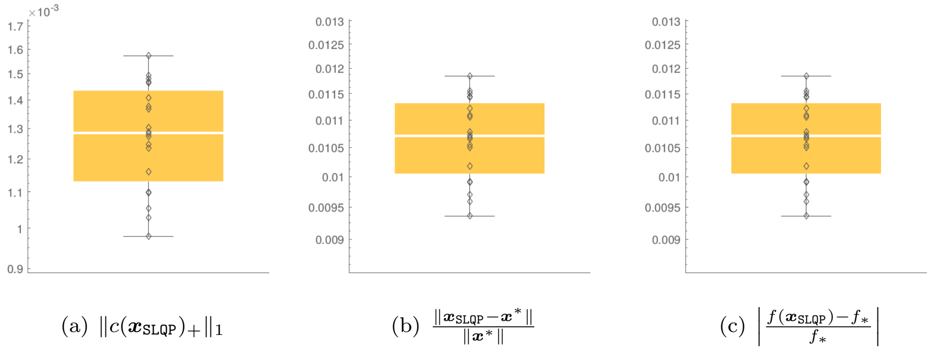
$$\min_{x_1,x_2} f(x_1,x_2) = \max\{x_1^3 - x_2, x_2\} \quad \text{s.t.} \quad c(x_1,x_2) = (x_1-10)^2 \le 0. \tag{5.1}$$

The optimal solution of this problem is given by $\boldsymbol{x}^* = (10, 500)^T$ with its respective optimal value $f_* = 500$. Due to our choice of the constraint function that defines the feasible set $\mathcal{F} = \{(x_1,x_2) \in \mathbb{R}^2 : x_1 = 10\}$, the calmness property does not hold at the optimal point. This has a great impact on the behavior of methods that are based on the exact penalization approach, since this implies that one cannot solve the original problem with a finite penalty parameter [17]. In case this parameter value is allowed to go to infinity, this usually produces a sequence of points that has a poor precision on the optimal function value.

To illustrate how a method based on exact penalization behaves in the absence of calmness, we have solved the nonsmooth problem (5.1) using the SLQP-GS[2] (version 1.3) [28] in its default mode, with the exception that we have allowed the SLQP-GS algorithm to run 10^4 iterations in order to observe the optimal value precision that such a method is able to reach. In a similar way, the GS algorithm [21, 46] was chosen to be our internal solver for Step 1—a version that we call $\texttt{PACNO}_{\texttt{GS}}$—and we have used the following parameter values for $\texttt{PACNO}_{\texttt{GS}}$: $\rho_1 = 0$, $M = 10$, $\theta_\xi = 0.5$, $\theta_\epsilon = 0.25$, $\theta_\nu = 0.5$, $\omega = 0.99$, $\xi_1 = \epsilon_1 = \nu_1 = 1$, and $\xi_{\text{opt}} = \epsilon_{\text{opt}} = \nu_{\text{opt}} = 10^{-8}$. In addition, for each call of GS, the sampling radii were set first as $\min\{10\epsilon_k, 1\}$ with optimality

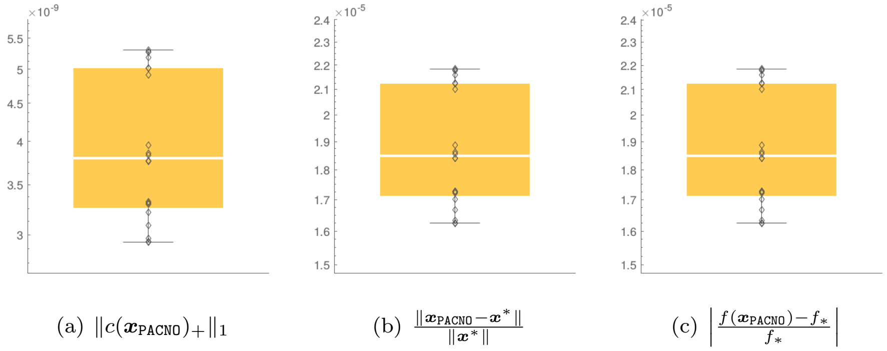
[2] The code is available online from http://coral.ise.lehigh.edu/frankecurtis/software/.

tolerance given by $\min\{10\nu_k, 1\}$ and then $\min\{\epsilon_k, 1\}$ as the final sampling radius with optimality tolerance $\min\{\nu_k, 1\}$. The results are shown in Figures 5 and 6, so that one can follow the reached distributions of the inviability measure in (a) and of the relative errors ((b) in the domain and (c) in the image space).



(a) $\|c(\boldsymbol{x}_{\texttt{SLQP}})_+\|_1$ (b) $\frac{\|\boldsymbol{x}_{\texttt{SLQP}} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{SLQP}}) - f_*}{f_*}\right|$

FIG. 5. *Boxplots of the results of* 20 *runs (depicted as* $\diamond$*) of the* `SLQP-GS` *method with different initial points chosen in a box* $[-5, 5]^2$ *centered at the optimal solution of problem* (5.1). *The last iterate obtained by the* `SLQP-GS` *method is represented by* $\boldsymbol{x}_{\texttt{SLQP}}$.

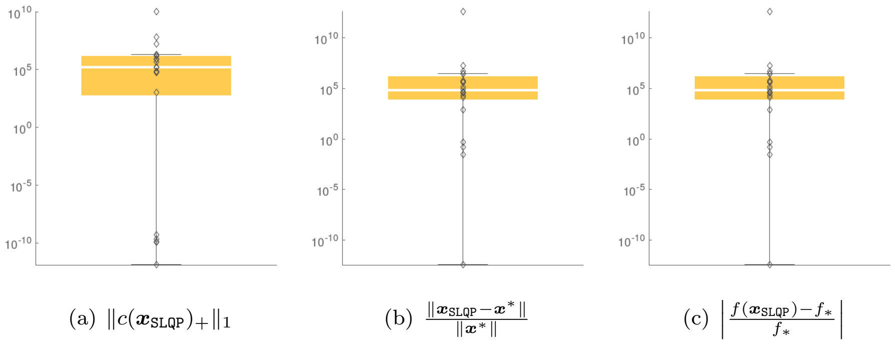


(a) $\|c(\boldsymbol{x}_{\texttt{PACNO}})_+\|_1$ (b) $\frac{\|\boldsymbol{x}_{\texttt{PACNO}} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{PACNO}}) - f_*}{f_*}\right|$

FIG. 6. *Boxplots of the result of* 20 *runs (depicted as* $\diamond$*) of the* $\texttt{PACNO}_{\texttt{GS}}$ *method with different initial points chosen in a box* $[-5, 5]^2$ *centered at the optimal solution of problem* (5.1). *The last iterate obtained by the* $\texttt{PACNO}_{\texttt{GS}}$ *method is represented by* $\boldsymbol{x}_{\texttt{PACNO}}$.

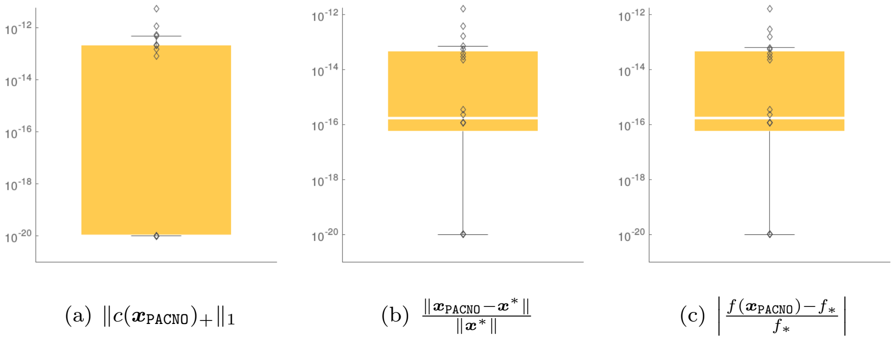
Because the `SLQP-GS` method relies on exact penalization and, additionally, its convergence theory is established only when the nonsmooth problem satisfies the calmness property at the solution, it was expected that the precision achieved by the method regarding the optimal function value would not be satisfactory. On the other hand, our algorithm possesses a convergence theory even in the absence of calmness, and, as anticipated, the precision obtained related to the optimality measure is considerably better. In addition, the $\texttt{PACNO}_{\texttt{GS}}$ algorithm is able to keep the iterates closer to the feasible set.

Nevertheless, one can easily represent the feasible set $\mathcal{F}$ in a manner that the calmness property is satisfied. Indeed, we can consider the constraint $|x_1 - 10| \leq 0$ instead of $(x_1 - 10)^2 \leq 0$ (see Figures 7 and 8). For this new optimization problem, the penalty parameter in the exact penalization approach no longer must go to infinity. However, methods based on exact penalization may experience another undesirable



(a) $\|c(\boldsymbol{x}_{\texttt{SLQP}})_+\|_1$ (b) $\frac{\|\boldsymbol{x}_{\texttt{SLQP}}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{SLQP}})-f_*}{f_*}\right|$

FIG. 7. *Boxplots of the result of* 20 *runs (depicted as* $\diamond$*) of the* `SLQP-GS` *method with different initial points chosen in a box* $[-5,5]^2$ *centered at the optimal solution of problem* (5.1)*. The last iterate obtained by the* `SLQP-GS` *method is represented by* $\boldsymbol{x}_{\texttt{SLQP}}$*.*



(a) $\|c(\boldsymbol{x}_{\texttt{PACNO}})_+\|_1$ (b) $\frac{\|\boldsymbol{x}_{\texttt{PACNO}}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{PACNO}})-f_*}{f_*}\right|$

FIG. 8. *Boxplots of the result of* 20 *runs (depicted as* $\diamond$*) of the* $\texttt{PACNO}_{\texttt{GS}}$ *method with different initial points chosen in a box* $[-5,5]^2$ *centered at the optimal solution of problem* (5.1)*. The last iterate obtained by the* $\texttt{PACNO}_{\texttt{GS}}$ *method is represented by* $\boldsymbol{x}_{\texttt{PACNO}}$*.*

behavior: by initializing the method with an inappropriate value for the penalty parameter, such methods may present the greediness phenomenon. This anomaly can be seen when one tries to solve this new nonsmooth problem with the `SLQP-GS` method. In many runs, due to the initial value of the penalty parameter in the standard configuration of the `SLQP-GS` algorithm, the method gives excessive importance to the objective function in the first iterations, which carries the iterates away from the feasible set, preventing the method from converging even when one allows a large number of iterations to be performed (10^4 iterations). However, it is worth mentioning that if the user sets a better scaled penalty parameter value, the `SLQP-GS` algorithm will easily converge to the solution point. In contrast, the $\texttt{PACNO}_{\texttt{GS}}$ algorithm is able to reach the solution without needing to bring the parameter ξ down to zero (in the 20 runs that were performed, the mean value of ξ was kept above 10^{-3}), and, due to Lemma 4.4, the good behavior of $\texttt{PACNO}_{\texttt{GS}}$ is not subjected to a tuned value of ξ.

5.2. Bilevel optimization via a nonsmooth approach. One of the most usual ways to solve a bilevel optimization problem [24] is to consider a mathematical problem with equilibrium constraints (MPEC) instead of the original multilevel instance. However, there are situations in which this approach may generate an MPEC that is not equivalent to the original problem. For example, the authors of [31] show that the minimization

$$\min_{x,\boldsymbol{y}} \; x \quad \text{s.t. } x \geq 0, \, \boldsymbol{y} \in \Lambda(x),$$

with $\Lambda(x) := \operatorname{argmin}_{\boldsymbol{y}}\{y_1 : y_1^2 - y_2 \leq x, y_1^2 + y_2 \leq 0\}$, may generate an MPEC that does not possess an optimal solution when one considers the KKT conditions of the minimization problem related to the definition of $\Lambda(x)$. Indeed, consider the following nonsmooth single level optimization problem in which the first three constraints are associated with the KKT conditions of the implicit optimization problem presented in $\Lambda(x)$:

$$\begin{array}{lll} \min\limits_{x,\boldsymbol{y},\boldsymbol{\lambda}} & x & \\ \text{s.t.} & |2\lambda_1 y_1 + 2\lambda_2 y_1 + 1| & \leq \; 0, \\ & |-\lambda_1 + \lambda_2| & \leq \; 0, \\ & \min\{-y_1^2 + y_2 + x, \lambda_1\}^2 + \min\{-y_1^2 - y_2, \lambda_2\}^2 & \leq \; 0, \\ & -x & \leq \; 0. \end{array} \tag{5.2}$$

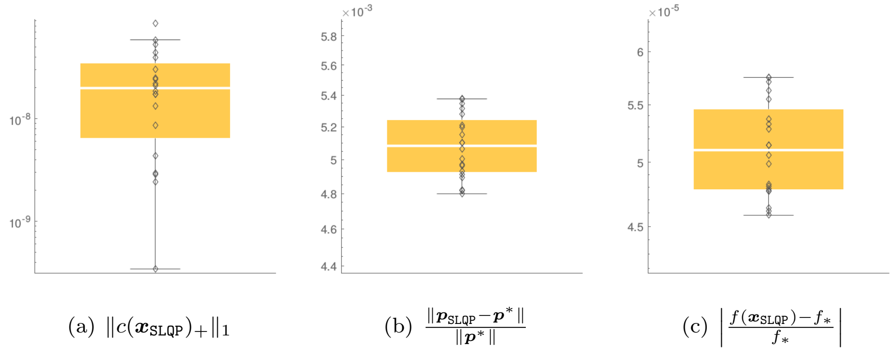
One can see that, for every $x > 0$, it is always possible to find $\boldsymbol{y}(x)$ and $\boldsymbol{\lambda}(x)$ such that $(x, \boldsymbol{y}(x), \boldsymbol{\lambda}(x))$ is a feasible point. Although the objective function value of this problem converges to zero when $x \to 0$, the feasible set is empty for $x = 0$. This indicates that the MPEC instance (5.2) does not possess an optimal point, and, consequently, it cannot recover the original solution $\boldsymbol{p}^* := (x^*, y_1^*, y_2^*)^T = (0, 0, 0)^T$.

This type of problem does not fit into our theoretical assumptions, but there are reasons to believe that our approach may present a good behavior when applied to such a problem. Notice that, although the feasible set is empty when $x = 0$, slight perturbations on the constraints produce a nonsmooth optimization problem that accepts points for which the first coordinate is zero. Since the proposed sequential optimality conditions allow the use of first-order information at infeasible points (see (2.2)), one can expect that the $\texttt{PACNO}_{\texttt{GS}}$ method (as defined in the previous subsection) might obtain a good approximation to the original solution of the bilevel optimization problem.

To verify our expectations, we have solved (5.2) using the $\texttt{PACNO}_{\texttt{GS}}$ algorithm. Furthermore, we have also used the SLQP-GS algorithm as a way to have some comparative results. In the same way that it was done in the previous subsection, we allowed the SLQP-GS method to run 10^4 iterations in order to seek the best precision that this method can achieve. The results are shown in Figures 9 and 10.

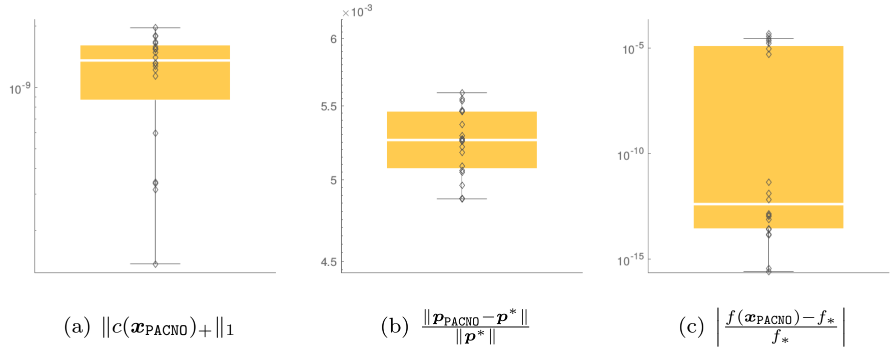
Regarding the infeasibility measure, both methods are able to reach values close to zero—although, in the limit, the problem is infeasible. In contrast, when one looks to the optimality precision achieved by the algorithms, the $\texttt{PACNO}_{\texttt{GS}}$ method presents a clear advantage in many runs. Because the problem is infeasible, the SLQP-GS algorithm must bring the penalty parameter to a value that overshadows the objective function, giving too much importance to the infeasibility measure term. As a consequence, the method cannot substantially improve the optimality measure. On the other hand, since the $\texttt{PACNO}_{\texttt{GS}}$ algorithm occults the objective function when the iterates are too far from feasibility, the method is able to achieve high precision in the optimal value for many runs (13 out of 20).

5.3. The kissing number problem. In $\mathbb{R}^n$, how many nonoverlapping spheres can touch or kiss, simultaneously, another sphere of the same size? This quantity, known as *kissing number*, and here denoted by κ_n, is closely related to finding bounds for spherical codes and sphere packings [26]. Apparently dating back to 1694, when



(a) $\|c(\boldsymbol{x}_{\texttt{SLQP}})_+\|_1$ (b) $\frac{\|\boldsymbol{p}_{\texttt{SLQP}}-\boldsymbol{p}^*\|}{\|\boldsymbol{p}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{SLQP}})-f_*}{f_*}\right|$

FIG. 9. *Boxplots of the results of* 20 *runs (depicted as* $\diamond$*) of the* `SLQP-GS` *method with different initial points chosen in a box* $[-5,5]^5$ *centered at* $(0,0,0,50,50)^T$*. The primal variables (i.e., the first three coordinates) of the last iterate obtained by the* `SLQP-GS` *method are represented by* $\boldsymbol{p}_{\texttt{SLQP}}$*, whereas the complete vector is represented by* $\boldsymbol{x}_{\texttt{SLQP}}$*.*



(a) $\|c(\boldsymbol{x}_{\texttt{PACNO}})_+\|_1$ (b) $\frac{\|\boldsymbol{p}_{\texttt{PACNO}}-\boldsymbol{p}^*\|}{\|\boldsymbol{p}^*\|}$ (c) $\left|\frac{f(\boldsymbol{x}_{\texttt{PACNO}})-f_*}{f_*}\right|$

FIG. 10. *Boxplots of the results of* 20 *runs (depicted as* $\diamond$*) of the* `PACNO`$_{\texttt{GS}}$ *method with different initial points chosen in a box* $[-5,5]^5$ *centered at* $(0,0,0,50,50)^T$*. The primal variables (i.e., the first three coordinates) of the last iterate obtained by the* `PACNO`$_{\texttt{GS}}$ *method are represented by* $\boldsymbol{p}_{\texttt{PACNO}}$*, whereas the complete vector is represented by* $\boldsymbol{x}_{\texttt{PACNO}}$*.*

Isaac Newton and James Gregory disputed whether κ_3 was 12 or 13, finding kissing numbers is still an open problem in most of the dimensions, for which just lower and upper bounds are known. Recent research has pursued improvements upon these bounds (see, e.g., [8, 49, 51]). For further details about the problem, including historical and mathematical related developments, we refer the reader to the review [53] and the survey [12].

Using known values, or bounds, of κ_n provides challenging instances for testing nonlinear programming algorithms [13, 16, 47]. Indeed, for a given pair of integers (n,p) and a given radius r, the following nonsmooth and nonconvex formulation aims at finding the centers $\boldsymbol{w}^i \in \mathbb{R}^n$, $i=1,\dots,p$, of the spheres that touch a sphere centered at the origin:

$$\max \min_{i \neq j} \ \|\boldsymbol{w}^i - \boldsymbol{w}^j\| \quad \text{s.t.} \quad \|\boldsymbol{w}^i\| = 2r, \ \ i=1,\dots,p.$$

Such a problem is equivalent to

$$\min \max_{i \neq j} \ (\boldsymbol{w}^i)^T \boldsymbol{w}^j \quad \text{s.t.} \quad \|\boldsymbol{w}^i\| = 2r, \ \ i=1,\dots,p, \tag{5.3}$$

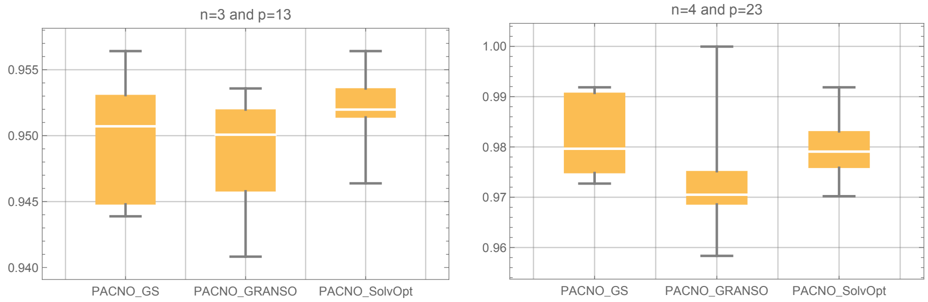


FIG. 11. *Boxplots of the maximum minimum distances obtained for* $(3, 13)$ *and* $(4, 23)$*.*

having $n \times p$ variables, which may be arranged as $\mathrm{vec}(\boldsymbol{w}^1 \cdots \boldsymbol{w}^p) \in \mathbb{R}^{np}$. To be addressed by Algorithm 4.1, each equality of (5.3) is turned into two inequalities, amounting to $2p$ inequality constraints. Setting the radius $r = 1/2$, for each choice of (n, p), 20 initial points were randomly and uniformly sampled in the box $[-2, 2]^{np}$. We have addressed nine instances, with (n, p) among the pairs: $(3, 11)$, $(3, 12)$, $(3, 13)$, $(4, 23)$, $(4, 24)$, $(4, 25)$, $(5, 39)$, $(5, 40)$, and $(5, 41)$.

The corresponding instances were solved by Algorithm 4.1 using three possible solvers for computing the current approximation in Step 1, namely `GS` [21, 46], `GRANSO` [27], and `SolvOpt` [43], respectively, referred to as $\texttt{PACNO}_{\texttt{GS}}$, $\texttt{PACNO}_{\texttt{GRANSO}}$, and $\texttt{PACNO}_{\texttt{SolvOpt}}$.

The algorithmic parameters were set as follows: $\rho_1 = 10$, $M = 5$, $\theta_\xi = 0.5$, $\theta_\epsilon = 0.25$, $\theta_\nu = 0.5$, $\omega = 0.99$, $\xi_1 = 10$, $\epsilon_1 = \nu_1 = 10^{-2}$, $\xi_{\mathrm{opt}} = 10^{-8}$, $\epsilon_{\mathrm{opt}} = 10^{-6}$, and $\nu_{\mathrm{opt}} = 10^{-6}$. We have used tight tolerances aiming to reach accurate and better solutions. The maximum allowed number of iterations to be performed by the three inner solvers was set to 10^4. Concerning their specific stopping criteria, we have made the following choices: for each call of `GS`, the sampling radii were set first as $\min\{10\epsilon_k, 1\}$ with optimality tolerance given by $\min\{10\nu_k, 1\}$ and then $\min\{\epsilon_k, 1\}$ as the final sampling radius with optimality tolerance $\min\{\nu_k, 1\}$; for `GRANSO`, the stationarity tolerance was $\tau_\diamond = \min\{\nu_k, 10^{-12}\}$, and the violation tolerance was $\tau_v = \min\{\epsilon_k, 10^{-6}\}$; and for `SolvOpt`, the tolerances for the relative error in the domain and in the image space were set as $10^{-6}/\sqrt{n \cdot p}$ and $10^{-8}/\sqrt{n \cdot p}$, respectively.

The results for instances $(3, 11)$ and $(3, 12)$ were quite stable, without variability between the smallest and the largest values for the maximum minimum distances. In fact, the value 1.05146222 has been attained, no matter the initialization and the inner solver used. For $(3, 13)$, however, we have observed not only variability among the 20 initializations, but also slight differences among the reached values, as depicted in the distributions of Figure 11 and detailed in Table 1. It is worth mentioning that the values attained by $\texttt{PACNO}_{\texttt{SolvOpt}}$ for (3,13) are even better than those reported in [16], corresponding to results obtained by differentiable solvers that addressed a smooth reformulation of (5.3).

As the time demanded by `GS` increases significantly with the problem dimension, instance $(4, 23)$ was the largest one addressed by the three inner solvers, and from $(4, 24)$ onwards, just `GRANSO` and `SolvOpt` were considered. The range of the CPU demanded by the inner solvers can be put in perspective by means of Figure 12, in which $\log_{10}$ scaled values are presented. The plots evince not only the larger amount of time required by `GS` for solving instance $(4, 23)$, in comparison with the other two solvers, but also the wider variability reached by $\texttt{PACNO}_{\texttt{GRANSO}}$ in comparison with

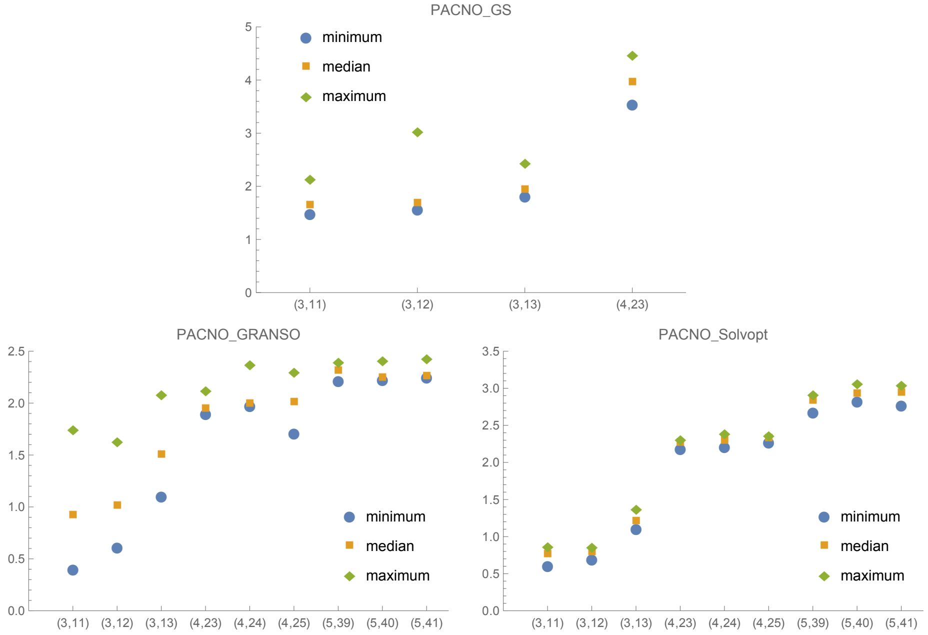


FIG. 12. *Range of the CPU time (in seconds, and* $\log_{10}$ *scaled) demanded by* `PACNO` *for solving each instance, according with the inner solver:* `GS` *(top),* `GRANSO` *(bottom/left), and* `SolvOpt` *(bottom/right).*

$\mathtt{PACNO}_{\mathtt{SolvOpt}}$ for each instance.

Taking the values reported in [16] as a reference, we have also noticed that, as the problem dimension increases, the quality of the obtained results slightly deteriorates (cf. Table 2). Moreover, the results reached by $\mathtt{PACNO}_{\mathtt{SolvOpt}}$ were usually better than those obtained by $\mathtt{PACNO}_{\mathtt{GRANSO}}$, at the price of demanding more CPU time for being computed, as shown in Figure 12. The investigation of suitable parameter settings aiming to increase the attained maximum minimum distances will be the subject of future research.

TABLE 1
Distribution values of the maximum minimum distances between two centers.

Instance	Algorithm	Minimum	Median	Maximum	Average
$(3, 13)$	$\mathtt{PACNO}_{\mathtt{GS}}$	0.9438814	0.9507097	0.9564136	0.9494221
	$\mathtt{PACNO}_{\mathtt{GRANSO}}$	0.9408229	0.9500742	0.9535789	0.9490796
	$\mathtt{PACNO}_{\mathtt{SolvOpt}}$	0.9463817	0.9519782	0.9564136	0.9522653
$(4, 23)$	$\mathtt{PACNO}_{\mathtt{GS}}$	0.9727178	0.9796446	0.9918580	0.9815232
	$\mathtt{PACNO}_{\mathtt{GRANSO}}$	0.9583368	0.9705020	0.9999590	0.9723568
	$\mathtt{PACNO}_{\mathtt{SolvOpt}}$	0.9702008	0.9790741	0.9918580	0.9797924

6. Conclusion.

We have proposed two sequential optimality conditions for a wide class of nonsmooth optimization problems. Both the weak ϵ-ANOC and the ϵ-ANOC are legitimate necessary optimality conditions in the sense that they do not require any kind of CQ to hold. In addition, when our stronger optimality condition is taken to the smooth context, we were able to prove that the ϵ-ANOC is stronger than the AKKT condition and, moreover, that CAKKT and ϵ-ANOC are not connected

TABLE 2
Distribution values of the maximum minimum distances between two centers (cont.).

Instance	Algorithm	Minimum	Median	Maximum	Average
$(4, 24)$	$\texttt{PACNO}_{\texttt{GRANSO}}$	0.9434813	0.9578637	0.9730925	0.9579984
	$\texttt{PACNO}_{\texttt{SolvOpt}}$	0.9597049	0.9716582	0.9828751	0.9704148
$(4, 25)$	$\texttt{PACNO}_{\texttt{GRANSO}}$	0.9325337	0.9428726	0.9541111	0.9436839
	$\texttt{PACNO}_{\texttt{SolvOpt}}$	0.9474955	0.9554975	0.9617487	0.9552285
$(5, 39)$	$\texttt{PACNO}_{\texttt{GRANSO}}$	0.9431839	0.9561418	0.9647305	0.9550583
	$\texttt{PACNO}_{\texttt{SolvOpt}}$	0.9531385	0.9653934	0.9758015	0.9653913
$(5, 40)$	$\texttt{PACNO}_{\texttt{GRANSO}}$	0.9382251	0.9510471	0.9576960	0.9497185
	$\texttt{PACNO}_{\texttt{SolvOpt}}$	0.9451599	0.9615852	0.9727113	0.9604777
$(5, 41)$	$\texttt{PACNO}_{\texttt{GRANSO}}$	0.9320542	0.9453193	0.9547505	0.9439547
	$\texttt{PACNO}_{\texttt{SolvOpt}}$	0.9295390	0.9515950	0.9599363	0.9507821

to each other—classes of optimization problems in which the ϵ-ANOC may present a practical advantage over the CAKKT condition is currently under study. Finally, we exhibited a practical algorithm able to generate both sequential optimality conditions as well as illustrative numerical results that highlight the potentialities of the devised algorithm.

Acknowledgments. We would like to thank Michael L. Overton for hosting the third author during his internship at New York University, Courant Institute, where much of this work was done, and also for his valuable comments. We are also indebted to Frank E. Curtis for the algorithmic suggestions given in the early days of this study, as well as to C. Yalçin Kaya for sharing with us the code of `AUGPEN` introduced in [16] and for making us realize the potential use of `SolvOpt`. In addition, we would like to thank the two anonymous referees who provided insightful comments that helped us to improve our presentation.

REFERENCES

[1] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, *Augmented Lagrangian methods under the constant positive linear dependence constraint qualification*, Math. Program., 111 (2008), pp. 5–32.

[2] R. Andreani, N. S. Fazzio, M. L. Schuverdt, and L. D. Secchin, *A sequential optimality condition related to the quasi-normality constraint qualification and its algorithmic consequences*, SIAM J. Optim., 29 (2019), pp. 743–766, https://doi.org/10.1137/17M1147330.

[3] R. Andreani, G. Haeser, and J. M. Martínez, *On sequential optimality conditions for smooth constrained optimization*, Optimization, 60 (2011), pp. 627–641.

[4] R. Andreani, G. Haeser, L. D. Secchin, and P. J. S. Silva, *New sequential optimality conditions for mathematical problems with complementarity constraints and algorithmic consequences*, Optimization Online, (2018), pp. 1–30.

[5] R. Andreani, J. M. Martínez, A. Ramos, and P. J. S. Silva, *A cone-continuity constraint qualification and algorithmic consequences*, SIAM J. Optim., 26 (2016), pp. 96–110, https://doi.org/10.1137/15M1008488.

[6] R. Andreani, J. M. Martínez, and M. L. Schuverdt, *On the relation between constant positive linear dependence condition and quasinormality constraint qualification*, J. Optim. Theory Appl., 125 (2005), pp. 473–483.

[7] R. Andreani, J. M. Martínez, and B. F. Svaiter, *A new sequential optimality condition for constrained optimization and algorithmic consequences*, SIAM J. Optim., 20 (2010), pp. 3533–3554, https://doi.org/10.1137/090777189.

[8] C. Bachoc and F. Vallentin, *New upper bounds for kissing numbers from semidefinite programming*, J. Amer. Math. Soc., 21 (2008), pp. 909–924.

[9] E. G. Birgin, E. V. Castelani, A. L. M. Martinez, and J. M. Martínez, *Outer trust-region method for constrained optimization*, J. Optim. Theory Appl., 150 (2011), pp. 142–155.

[10] E. G. Birgin, N. Krejić, and J. M. Martínez, *On the minimization of possibly discontinuous*

functions by means of pointwise approximations, Optim. Lett., 11 (2017), pp. 1623–1637.

[11] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 2006.

[12] P. Boyvalenkov, S. Dodunekov, and O. Musin, *A survey on the kissing numbers*, Serdica Math. J., 38 (2012), pp. 507–522.

[13] R. S. Burachik, W. P. Freire, and C. Y. Kaya, *Interior epigraph directions method for nonsmooth and nonconvex optimization via generalized augmented Lagrangian duality*, J. Global Optim., 60 (2014), pp. 501–529.

[14] R. S. Burachik and C. Y. Kaya, *An update rule and a convergence result for a penalty function method*, J. Ind. Manag. Optim., 3 (2007), pp. 381–398.

[15] R. S. Burachik and C. Y. Kaya, *A deflected subgradient method using a general augmented Lagrangian duality with implications on penalty methods*, in Variational Analysis and Generalized Differentiation in Optimization and Control, Springer, New York, 2010, pp. 109–132.

[16] R. S. Burachik and C. Y. Kaya, *An augmented penalty function method with penalty parameter updates for nonconvex optimization*, Nonlinear Anal., 75 (2012), pp. 1158–1167.

[17] J. V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497, https://doi.org/10.1137/0329027.

[18] J. V. Burke, *An exact penalization viewpoint of constrained optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998, https://doi.org/10.1137/0329054.

[19] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. A. Simões, *Gradient Sampling Methods for Nonsmooth Optimization*, preprint, https://arxiv.org/abs/1804.11003, 2018.

[20] J. V. Burke, A. S. Lewis, and M. L. Overton, *Approximating subdifferentials by random sampling of gradients*, Math. Oper. Res., 27 (2002), pp. 567–584.

[21] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779, https://doi.org/10.1137/030601296.

[22] E. V. Castelani, A. L. M. Martinez, J. M. Martínez, and B. F. Svaiter, *Addressing the greediness phenomenon in nonlinear programming by means of proximal augmented Lagrangians*, Comput. Optim. Appl., 46 (2010), pp. 229–245.

[23] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990, https://doi.org/10.1137/1.9781611971309.

[24] B. Colson, P. Marcotte, and G. Savard, *An overview of bilevel optimization*, Ann. Oper. Res., 153 (2007), pp. 235–256.

[25] A. R. Conn, N. I. M. Gould, and P. L. Toint, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572, https://doi.org/10.1137/0728030.

[26] J. H. Conway and N. J. C. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.

[27] F. E. Curtis, T. Mitchell, and M. L. Overton, *A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles*, Optim. Methods Softw., 32 (2017), pp. 148–181.

[28] F. E. Curtis and M. L. Overton, *A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization*, SIAM J. Optim., 22 (2012), pp. 474–500, https://doi.org/10.1137/090780201.

[29] F. E. Curtis and X. Que, *An adaptive gradient sampling algorithm for non-smooth optimization*, Optim. Methods Softw., 28 (2013), pp. 1302–1324.

[30] F. E. Curtis and X. Que, *A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees*, Math. Program. Comput., 7 (2015), pp. 399–428.

[31] S. Dempe and J. Dutta, *Is bilevel programming a special case of a mathematical program with complementarity constraints?*, Math. Program., 131 (2012), pp. 37–48.

[32] J. Dutta, K. Deb, R. Tulshyan, and R. Arora, *Approximate KKT points and a proximity measure for termination*, J. Global Optim., 56 (2013), pp. 1463–1499.

[33] I. Eremin, *The penalty method in convex programming*, Soviet Math. Dokl., 8 (1966), pp. 459–462.

[34] G. Fasano, G. Liuzzi, S. Lucidi, and F. Rinaldi, *A linesearch-based derivative-free approach for nonsmooth constrained optimization*, SIAM J. Optim., 24 (2014), pp. 959–992, https://doi.org/10.1137/130940037.

[35] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, 1969.

[36] A. A. Goldstein, *Optimization of Lipschitz continuous functions*, Math. Programming, 13

(1977), pp. 14–22.

[37] L. Guo, G.-H. Lin, and J. J. Ye, *Second-order optimality conditions for mathematical programs with equilibrium constraints*, J. Optim. Theory Appl., 158 (2013), pp. 33–64.

[38] E. S. Helou, S. A. Santos, and L. E. A. Simões, *On the differentiability check in gradient sampling methods*, Optim. Methods Softw., 31 (2016), pp. 983–1007.

[39] E. S. Helou, S. A. Santos, and L. E. A. Simões, *On the local convergence analysis of the gradient sampling method for finite max-functions*, J. Optim. Theory Appl., 175 (2017), pp. 137–157.

[40] E. S. Helou, S. A. Santos, and L. E. A. Simões, *A fast gradient and function sampling method for finite-max functions*, Comput. Optim. Appl., 71 (2018), pp. 673–717.

[41] J.-B. Hiriart-Urruty, *On optimality conditions in nondifferentiable programming*, Math. Programming, 14 (1978), pp. 73–86.

[42] J.-B. Hiriart-Urruty, *Refinements of necessary optimality conditions in nondifferentiable programming* I, Appl. Math. Optim., 5 (1979), pp. 63–82.

[43] F. Kappel and A. V. Kuntsevich, *An implementation of Shor's r-algorithm*, Comput. Optim. Appl., 15 (2000), pp. 193–205.

[44] K. C. Kiwiel, *An exact penalty function algorithm for non-smooth convex constrained minimization problems*, IMA J. Numer. Anal., 5 (1985), pp. 111–119.

[45] K. C. Kiwiel, *Exact penalty functions in proximal bundle methods for constrained convex nondifferentiable minimization*, Math. Programming, 52 (1991), pp. 285–302.

[46] K. C. Kiwiel, *Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization*, SIAM J. Optim., 18 (2007), pp. 379–388, https://doi.org/10.1137/050639673.

[47] N. Krejić, J. M. Martínez, M. P. Mello, and E. A. Pilota, *Validation of an augmented Lagrangian algorithm with a Gauss-Newton Hessian approximation using a set of hard-spheres problems*, Comput. Optim. Appl., 16 (2000), pp. 247–263.

[48] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., 141 (2013), pp. 135–163.

[49] L. Liberti, *Mathematical programming bounds for kissing numbers*, in Optimization and Decision Science: Methodologies and Applications, A. Sforza and C. Sterle, eds., Springer, Cham, 2017, pp. 213–222.

[50] M. Loreto, H. Aponte, D. Cores, and M. Raydan, *Nonsmooth spectral gradient methods for unconstrained optimization*, EURO J. Comput. Optim., 5 (2017), pp. 529–553.

[51] F. C. Machado and F. M. de Oliveira Filho, *Improving the semidefinite programming bound for the kissing number by exploiting polynomial symmetry*, Exp. Math., 27 (2018), pp. 362–369, https://doi.org/10.1080/10586458.2017.1286273.

[52] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed., Springer-Verlag, New York, 2006.

[53] F. Pfender and G. M. Ziegler, *Kissing numbers, sphere packings, and some unexpected proofs*, Notices Amer. Math. Soc., 51 (2004), pp. 873–883.

[54] E. Polak, D. Q. Mayne, and Y. Wardi, *On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems*, SIAM J. Control Optim., 21 (1983), pp. 179–203, https://doi.org/10.1137/0321010.

[55] R. T. Rockafellar, R. Wets, and M. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 1998.

[56] W. I. Zangwill, *Non-linear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.