Optimality and complexity for constrained optimization problems with nonconvex regularization

Wei Bian
Department of Mathematics, Harbin Institute of Technology, Harbin, China, bianweilvse520@163.com

Xiaojun Chen
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China, maxjchen@polyu.edu.hk

In this paper, we consider a class of constrained optimization problems where the feasible set is a general closed convex set and the objective function has a nonsmooth, nonconvex regularizer. Such regularizer includes widely used SCAD, MCP, logistic, fraction, hard thresholding and non-Lipschitz L_p penalties as special cases. Using the theory of the generalized directional derivative and the tangent cone, we derive a first order necessary optimality condition for local minimizers of the problem, and define the generalized stationary point of it. We show that the generalized stationary point is the Clarke stationary point when the objective function is Lipschitz continuous at this point, and satisfies the existing necessary optimality conditions when the objective function is not Lipschitz continuous at this point. Moreover, we prove the consistency between the generalized directional derivative and the limit of the classic directional derivatives associated with the smoothing function. Finally, we establish a lower bound property for every local minimizer and show that finding a global minimizer is strongly NP-hard when the objective function has a concave regularizer.

Key words: Constrained nonsmooth nonconvex optimization; optimality condition; generalized directional derivative; directional derivative consistency; numerical property
MSC2000 subject classification: Primary: 49K35, 90C26; secondary: 65K05, 90C46

---

1. Introduction In this paper, we consider the following constrained optimization problem

$$\min_{x\in\mathcal{X}} \quad f(x):=\Theta(x)+c(h(x)), \tag{1}$$

where $\Theta: R^n\to R$ and $c: R^m\to R$ are continuously differentiable, $h: R^n\to R^m$ is continuous, and $\mathcal{X}\subset R^n$ is a nonempty closed convex set. Of particular interest of this paper is when h is not convex, not differentiable, or even not Lipschitz continuous at some points, and f has at least one local minimizer over $\mathcal{X}$. Problem (1) includes many problems in practice. For instance, the following minimization problem

$$\min_{l\le x\le u, Ax\le b} \quad f(x):=\Theta(x)+\sum_{i=1}^m \varphi(\|D_i^T x\|_p^p) \tag{2}$$

is a special case of (1), where $l\in(R\cup\{-\infty\})^n, u\in(R\cup\{\infty\})^n, A\in R^{t\times n}, b\in R^t,\ D_i\in R^{n\times r}$, $p\in(0,1]$ and $\varphi: R_+\to R_+$ is continuous. Such problem arises from image restoration (Chan and Liang [16], Chen et al. [22], Nikolova et al. [44]), signal processing (Bruckstein et al. [12]), variable selection (Fan and Li [27], Huang et al. [33], Huang et al. [35], Zhang [55]), etc. Another special case of (1) is the following problem

$$\min_{x\in\mathcal{X}} \quad f(x):=\Theta(x)+\sum_{i=1}^m \varphi(\max\{\alpha_i-d_i^T x,0\}^p), \tag{3}$$

with $\alpha_i \in R$ and $d_i \in R^n$, which has attracted much interest in machine learning, wireless communication (Liu et al. [39, 40]), information theory, data analysis (Fan and Peng [28], Huber [34]), etc. Moreover, a number of constrained optimization problems can be reformulated as problem (1) by using the exact penalty method with nonsmooth or non-Lipschitz continuous penalty functions (Auslender [3]).

The generic nature of the first and second order optimality conditions in nonlinear programming are treated by Spingarn and Rockafellar [49]. When f is locally Lipschitz continuous and $\mathcal{X} = R^n$, x^* is called a Clarke stationary point of (1) if

$$f^\circ(x^*; v) \geq 0, \quad \forall v \in R^n, \tag{4}$$

where $f^\circ(x^*; v)$ is the Clarke generalized directional derivative of f at x^* in direction v (Clarke [24]), defined by

$$f^\circ(x^*; v) = \limsup_{y \to x^*, t \downarrow 0} \frac{f(y + tv) - f(y)}{t}. \tag{5}$$

From the following relation between the Clarke generalized directional derivative and Clarke subdifferential (Clarke [24])

$$\partial f(x) := \{s \in R^n : f^\circ(x; v) \geq v^T s, \quad \forall v \in R^n\} \quad \text{and} \quad f^\circ(x; v) = \max\{v^T s : \forall s \in \partial f(x)\},$$

condition (4) is equivalent to $0 \in \partial f(x^*)$.

For the constrained optimization, a Clarke stationary point of locally Lipschitz continuous function f over $\mathcal{X}$ is defined by the existence of $\xi \in \partial f(x^*)$ satisfying

$$\langle \xi, x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}. \tag{6}$$

Then, the Clarke generalized directional derivative in (5) is generalized by Jahn [36] and used in Audet and Dennis [2], Jahn [36] for Lipschitz constrained optimization problem, defined by

$$f^\circ(x^*; v; \mathcal{X}) = \limsup_{\substack{y \,\to\, x^*,\, y \,\in\, \mathcal{X} \\ t \,\downarrow\, 0,\, y + tv \,\in\, \mathcal{X}}} \frac{f(y + tv) - f(y)}{t}. \tag{7}$$

Based on the directional derivative in (7), Jahn [36, Theorem 3.46] presented a necessary optimality condition for a local minimizer x^* of f over $\mathcal{X}$, i.e.

$$f^\circ(x^*; x - x^*; \mathcal{X}) \geq 0, \quad \forall x \in \mathcal{X}. \tag{8}$$

When $\mathrm{int}(\mathcal{T}_{\mathcal{X}}(x^*)) \neq \emptyset$, the necessary optimality conditions in (6) and (8) are equivalent to

$$f^\circ(x^*; v; \mathcal{X}) \geq 0, \quad \forall v \in \mathcal{T}_{\mathcal{X}}(x^*), \tag{9}$$

where $\mathcal{T}_{\mathcal{X}}(x^*)$ is the tangent cone to $\mathcal{X}$ at x^*.

Due to the non-Lipschitz continuity of the objective function f in (1), the Clarke optimality conditions in (6), (8) or (9) cannot be directly applied to problem (1). For a proper, lower semi-continuous function $f : R^n \to R$ (possibly non-Lipschitz), the limiting (Mordukhovich) subdifferential and horizon subdifferential (Rockafellar and Wets [47]) are defined respectively as

$$\bar{\partial} f(x) = \{v : \exists x^k \xrightarrow{f} x,\ v^k \to v \text{ with } \liminf_{z \to x^k} \tfrac{f(z) - f(x^k) - \langle v^k, z - x^k \rangle}{\|z - x^k\|} \geq 0,\ \forall k\},$$

$$\partial^\infty f(x) = \{v : \exists x^k \xrightarrow{f} x,\ \lambda^k v^k \to v,\ \lambda^k \downarrow 0 \text{ with } \liminf_{z \to x^k} \tfrac{f(z) - f(x^k) - \langle v^k, z - x^k \rangle}{\|z - x^k\|} \geq 0,\ \forall k\},$$

where $\lambda^k \downarrow 0$ means $\lambda^k > 0$ and $\lambda^k \to 0$, and $x^k \xrightarrow{f} x$ means $x^k \to x$ and $f(x^k) \to f(x)$. Based on the following constraint qualification

$$-\partial^\infty f(\bar{x}) \bigcap N_{\mathcal{X}}(\bar{x}) = \{0\}, \tag{10}$$

for $\bar{x}$ to be a local minimizer of (1) it is necessary that

$$0 \in \bar{\partial} f(\bar{x}) + N_{\mathcal{X}}(\bar{x}) \tag{11}$$

(Rockafellar and Wets [47, Theorem 8.15]). The constraint qualification (10) holds naturally if f is locally Lipschitz continuous at $\bar{x}$ or $\bar{x}$ is an interior point of $\mathcal{X}$ (can be non-Lipschitz continuous at $\bar{x}$). When $\mathcal{X} = R^n$ and $c(h(x)) := \|x\|_p^p$ $(0 < p < 1)$, the affine scaled first and second order necessary optimality conditions for local minimizers of (1) are established in Chen et al. [23]. By using subspace techniques, Chen et al. [21] extended the first and second order necessary optimality conditions to problem (1) with $\mathcal{X} = R^n$ and $c(h(x)) := \|Dx\|_p^p$. Some other necessary optimality conditions for some special formats of (2) and (3) are studied in Bian and Chen [6, 8], Bian et al. [9], Ge et al. [30], Liu et al. [40]. However, the optimality conditions in Bian and Chen [6, 8], Bian et al. [9], Chen et al. [21, 23], Ge et al. [30], Liu et al. [40] are weaker than the Clarke optimality conditions given in (6), (8) and (9) for $p = 1$. The minimizers of problem (1) in practical problems are often the non-Lipschitz point and on the boundary of $\mathcal{X}$, which may lead to the unsatisfaction of quality (10). In this paper, we will derive a necessary optimality condition for the non-Lipschitz constrained optimization problem (1), which holds without the constraint qualification (10) and reduces to the Clarke optimality condition when the objective function in (1) is locally Lipschitz continuous at this point.

When f is locally Lipschitz continuous, from Theorem 9.61 and Corollary 8.47 (b) in Rockafellar and Wets [47], the subdifferential associated with a smoothing function

$$G_{\tilde{f}}(x) = \mathrm{con}\{v \,|\, \nabla_x \tilde{f}(x^k, \mu_k) \to v, \text{ for } x^k \to x, \ \mu_k \downarrow 0\},$$

is nonempty and bounded, and $\partial f(x) \subseteq G_{\tilde{f}}(x)$, where "con" denotes the convex hull. In Burke and Hoheisel [13], Burke et al. [14], Chen [19], Rockafellar and Wets [47], it is shown that many smoothing functions satisfy the gradient consistency

$$\partial f(x) = G_{\tilde{f}}(x). \tag{12}$$

The gradient consistency is an important property of the smoothing methods, which guarantees the convergence of smoothing methods with adaptive updating schemes of smoothing parameters to a stationary point of the original problem.

In this paper, we extend the directional derivative in Jahn [36] to the constrained optimization problem (1), whose objective function may not be Lipschitz continuous at some points. Using the extended directional derivative and the tangent cone, we derive a necessary optimality condition for local minimizers of problem (1), and define the generalized stationary point of (1). We show that the generalized stationary point is the Clarke stationary point when the objective function is Lipschitz continuous at this point defined in (6), (8) and (9), and satisfies the existing necessary optimality conditions when the objective function is not Lipschitz continuous at this point defined in Bian and Chen [6, 8], Bian et al. [9], Chen et al. [21, 23], Ge et al. [30], Liu et al. [40]. Moreover, we establish the consistency between the generalized directional derivative and the limit of the classic directional derivatives associated with the smoothing function. The directional derivative consistency guarantees the convergence of smoothing methods to a generalized stationary point of (1).

Problem (1) includes the regularized minimization problem as a special case when $\Theta(x)$ is a data fitting term and $c(h(x))$ is a regularization term (also called a penalty term in some articles). In sparse optimization, nonconvex non-Lipschitz regularization provides more efficient models to extract the essential features of solutions than the convex regularization (Bian and Chen [6], Chartrand and Staneva [17], Chen [19], Chen et al. [22], Fan and Li [27], Huang et al. [33], Huang et al. [35], Loh and Wainwright [41], Lu [42], Nikolova et al. [44], Wang et al. [53], Zhang [55]). The SCAD penalty function in Fan and Li [27] and the MCP function in Zhang [55] have various desirable properties in variable selection. Logistic and fraction penalty functions yield edge preservation in image restoration (Nikolova et al. [44]). The l_p norm penalty function with $0<p<1$ possesses the oracle property in statistics (Fan and Li [27], Knight and Fu [37]). Nonconvex regularized M-estimator is proved to have the statistical accuracy and prediction error estimation in Loh and Wainwright [41]. Moreover, the lower bound theory of the l_2-l_p regularized minimization problem in Chen et al. [22, 23], a special case of (1), states that the absolute value of each component of any local minimizer of the problem is either zero or greater than a positive constant. The lower bound theory not only helps us to distinguish zero and nonzero entries of coefficients in sparse high-dimensional approximation (Chartrand and Staneva [17], Huang et al. [33]), but also brings the restored image closed contours and neat edges (Chen et al. [22]). In this paper, we extend the lower bound theory of the l_2-l_p regularization minimization problem to problem (1) with constraints $\{x : Ax \le b\}$, which includes the most widely used models in statistics and sparse reconstruction as special cases. From the new bound theory, we can derive many interesting lower and upper bound results for different special problems. Moreover, we prove the strong NP-hardness of problem (1) via a special model of it, which generalizes the computational complexity results in Chen et al. [20], Ge et al. [30], Liu et al. [40]. Such extensions are not trivial because of the general constraints in problem (1). We show that the concavity of the regularization term is a key property for both the lower bound theory and the strong NP hardness of (1). The bound property gives the positive news of problem (1) in applications, while the strong NP-hardness indicates its negative aspect in numerical computation, which further illustrates the importance and necessity for presenting good necessary optimality conditions of (1).

The rest of this paper is organized as follows. In section 2, we first define a generalized directional derivative and present its properties. Next, we derive a necessary optimality condition for the local minimizers of problem (1), and prove the directional derivative consistency associated with smoothing functions. In section 3, we present the numerical properties of problem (1) with a special constraint from the bound property of its local minimizers and its computational complexity.

In our notation, $R_+ = [0,\infty)$, $R_{++} = (0,\infty)$ and $\mathbb{N} = \{1,2,\ldots\}$. For $x \in R^n$, $0<p<\infty$ and $\delta>0$, $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$, $B_\delta(x)$ means the open ball centered at x with radius δ. For a set $\Omega \subseteq R^n$, $\mathrm{int}(\Omega)$ means the interior of Ω, $\mathrm{cl}(\Omega)$ means the closure of Ω, $|\Omega|$ stands for its cardinality and and $P_\Omega[x] = \arg\min\{\|z-x\|_2 : z \in \Omega\}$ denotes the orthogonal projection from R^n to Ω. For locally Lipschitz continuous function $\phi : R^n \to R$, $\phi'(s+)$ and $\phi'(s-)$ indicate the derivative of ϕ at s on the right hand side and left hand side, respectively. For Π consisted by a class of column vectors of R^n, spanΠ indicates the subspace of R^n spanned by the elements in Π.

2. Optimality conditions Inspired by the generalized directional derivative and the tangent cone, we present a first order necessary optimality condition for local minimizers of the constrained optimization problem (1), which reduces to the Clarke necessary optimality condition at Lipschitz continuous point and to the necessary optimality conditions in the existing literatures (Bian and Chen [6, 8], Bian et al. [9], Chen et al. [21, 23], Ge et al. [30], Liu et al. [40]) at non-Lipschitz points. Thus, the generalized stationary point based on the derived necessary optimality condition provides a unified form for the stationary points of problem (1) with a continuous objective function. At the end of this section, we prove the directional directive consistency associated with smoothing functions, which gives some hints on how to find a generalized stationary points of (1) in numerical computation.

We suppose the function h in (1) has the following representation

$$h(x) := (h_1(D_1^T x), h_2(D_2^T x), \ldots, h_m(D_m^T x))^T \tag{13}$$

where $D_i \in R^{n\times r}$, $h_i (i=1,\ldots,m) : R^r \to R$ is continuous, but not necessarily Lipschitz continuous.

2.1. Tangent cone Since $\mathcal{X}$ is a nonempty closed convex subset of R^n, the tangent cone to $\mathcal{X}$ at $x \in \mathcal{X}$, denoted as $\mathcal{T}_{\mathcal{X}}(x)$, is the set consisting of all tangent vectors [47, Proposition 6.2], where we call a vector $v \in R^n$ a tangent vector to $\mathcal{X}$ at x, if there are a sequence $\{x^k\}$ of elements in $\mathcal{X}$ converging to x and a sequence $\{\lambda_k\}$ of positive numbers converging to 0 such that

$$v = \lim_{k\to\infty} \frac{x^k - x}{\lambda_k}.$$

Assumption 1. *Assume that $\mathcal{X}$ can be expressed by $\mathcal{X} = \mathcal{X}_1 \cap \mathcal{X}_2$ with a nonempty closed convex set $\mathcal{X}_1$, $\mathcal{X}_2 := \{x : Ax \le b\}$ and int$(\mathcal{X}_1) \cap \mathcal{X}_2 \neq \emptyset$, where $A \in R^{t\times n}$ and $b \in R^t$.*

Denote A_i^T and b_i the ith row of A and b, and let $\mathcal{C}_{\bar{x}} = \{i : A_i^T \bar{x} - b_i = 0\}$ for $\bar{x} \in \mathcal{X}_2$. Under Assumption 1, we can obtain the following properties of the tangent cones to $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}$.

Lemma 1. *Suppose Assumption 1 holds. Then the following statements hold.*
(1) $\mathcal{T}_{\mathcal{X}_2}(x) = \{v : A_i^T v \le 0,\ \forall i \in \mathcal{C}_{\bar{x}}\}$, $\forall x \in \mathcal{X}_2$;
(2) $\mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(x)) \cap \mathcal{T}_{\mathcal{X}_2}(x) \neq \emptyset$, $\forall x \in \mathcal{X}$;
(3) $\mathcal{T}_{\mathcal{X}_1 \cap \mathcal{X}_2}(x) = \mathcal{T}_{\mathcal{X}_1}(x) \cap \mathcal{T}_{\mathcal{X}_2}(x)$, $\forall x \in \mathcal{X}$.

Proof. (1) and (3) can be obtained by Borwein and Lewis [10, Corollary 6.3.7] and Rockafellar and Wets [47, Theorem 6.42] respectively.

Fix $\bar{x} \in \mathcal{X}$. By $\mathrm{int}(\mathcal{X}_1) \neq \emptyset$ and from Rockafellar and Wets [47, Example 6.22,Example 6.24], $\mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(\bar{x})) \neq \emptyset$ and $\hat{x} - \bar{x} \in \mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(\bar{x}))$ with $\hat{x} \in \mathrm{int}(\mathcal{X}_1)$. Using Assumption 1 and (1), we get $\hat{x} - \bar{x} \in \mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(\bar{x})) \cap \mathcal{T}_{\mathcal{X}_2}(\bar{x})$ with $\hat{x} \in \mathrm{int}(\mathcal{X}_1) \cap \mathcal{X}_2$. □

2.2. Generalized directional derivative

Definition 1. A function $\phi : R^n \to R$ is said to be Lipschitz continuous at(near) $x \in R^n$ if there exist positive numbers L_x and δ such that

$$|\phi(y) - \phi(z)| \le L_x \|y - z\|_2, \quad \forall\, y, z \in B_\delta(x).$$

Otherwise, ϕ is said to be not Lipschitz continuous at x.

For a fixed $\bar{x} \in R^n$, denote

$$\mathcal{I}_{\bar{x}} = \{i \in \{1,2,\ldots,m\} : h_i \text{ is not Lipschitz continuous at } D_i^T \bar{x}\}, \tag{14}$$

$$\mathcal{V}_{\bar{x}} = \{v : D_i^T v = 0,\ i \in \mathcal{I}_{\bar{x}}\}, \tag{15}$$

and define

$$h_i^{\bar{x}}(D_i^T x) := \begin{cases} h_i(D_i^T x) & i \notin \mathcal{I}_{\bar{x}} \\ h_i(D_i^T \bar{x}) & i \in \mathcal{I}_{\bar{x}}, \end{cases}$$

which is Lipschitz continuous at $D_i^T \bar{x}$, $i = 1,2,\ldots,m$. In particular, we let $\mathcal{V}_{\bar{x}} = R^n$ when $\mathcal{I}_{\bar{x}} = \emptyset$. And then we define

$$f_{\bar{x}}(x) := \Theta(x) + c(h_{\bar{x}}(x)), \tag{16}$$

with $h_{\bar{x}}(x) := (h_1^{\bar{x}}(D_1^T x), h_2^{\bar{x}}(D_2^T x), \dots, h_m^{\bar{x}}(D_m^T x))^T$.

We notice that the generalized directional derivative of f at $\bar{x} \in \mathcal{X}$ in direction v defined in (7) involves only the behavior of f around $\bar{x}$ in $\mathcal{X}$. Moreover, $f^\circ(\bar{x}; v; \mathcal{X})$ exists for any $v \in \mathcal{T}_{\mathcal{X}}(\bar{x})$ if f is Lipschitz continuous at $\bar{x}$. In particular, when $\bar{x} \in \mathrm{int}(\mathcal{X})$, $f^\circ(\bar{x}; v) = f^\circ(\bar{x}; v; \mathcal{X})$. However, the objective function f in (1) may not be Lipschitz continuous at some points, which implies that $f^\circ(\bar{x}; v; \mathcal{X})$ may not exist for some $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{T}_{\mathcal{X}}(\bar{x})$. So, we consider the generalized directional derivative of f over $\mathcal{X}$ in directions $v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \mathcal{V}_{\bar{x}}$ when f is not Lipschitz continuous at $\bar{x}$.

There are various generalized directional derivatives, such as the lower subderivative $d^- f(x)(v)$, upper subderivative $d^+ f(x)(v)$ and regular subderivative $\hat{d} f(x)(v)$ for the extended real-valued function defined in Rockafellar and Wets [47] and the Clarke directional derivative in (7) for real-valued function. When $f(x_1, x_2) = \sqrt{|x_1 - 1|} - \sqrt{|x_2|}$ and $\mathcal{X} = R^n$, which is an example modeled by the objective function in (1), $f^\circ(\bar{x}; v; \mathcal{X}) = -\frac{1}{2}$, but $d^+ f(\bar{x})(v) = +\infty$ for $\bar{x} = (1,1)$ and $v = (0,1) \in \mathcal{V}_{\bar{x}}$; $f^\circ(\bar{x}; v; \mathcal{X}) = \frac{1}{2}$, but $d^- f(\bar{x})(v) = \hat{d} f(\bar{x})(v) = -\infty$ for $\bar{x} = (2,0)$ and $v = (1,0) \in \mathcal{V}_{\bar{x}}$. This motivates us to use the Clarke generalized directional derivative with the constraints in (7) for problem (1) and we will prove the existence of $f^\circ(\bar{x}; v; \mathcal{X})$ for any $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \mathcal{V}_{\bar{x}}$.

Proposition 1. *For any $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \mathcal{V}_{\bar{x}}$,*

$$f^\circ(\bar{x}; v; \mathcal{X}) = \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ t \downarrow 0,\, y + tv \in \mathcal{X}}} \frac{f(y + tv) - f(y)}{t} \quad \textit{exists} \tag{17}$$

and equals $f^\circ_{\bar{x}}(\bar{x}; v; \mathcal{X})$ defined in (7).

Proof. Fix $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{V}_{\bar{x}}$. For $y \in R^n$ and $t > 0$, there exists z between $h(y)$ and $h(y + tv)$ such that

$$\begin{aligned} c(h(y+tv)) - c(h(y)) =& \nabla c(z)^T (h(y+tv) - h(y)) \\ =& \nabla c(z)^T (h_{\bar{x}}(y+tv) - h_{\bar{x}}(y)). \end{aligned}$$

Then,

$$\frac{f(y+tv) - f(y)}{t} = \frac{\Theta(y+tv) - \Theta(y) + \nabla c(z)^T (h_{\bar{x}}(y+tv) - h_{\bar{x}}(y))}{t}.$$

By the Lipschitz continuity of Θ and $h_{\bar{x}}$ at $\bar{x}$, there exist $\delta > 0$ and $L > 0$ such that $\left|\frac{f(y+tv)-f(y)}{t}\right| \leq L$, $\forall y \in B_\delta(\bar{x})$, $t \in (0, \delta)$. Thus, the generalized directional derivative of f at $\bar{x} \in \mathcal{X}$ in the direction $v \in \mathcal{V}_{\bar{x}}$ defined in (17) exists.

Let $\{y_j\}$ and $\{t_j\}$ be the sequences such that $y_j \in \mathcal{X}$, $t_j \downarrow 0$, $y_j \to \bar{x}$, $y_j + t_j v \in \mathcal{X}$ and

$$\lim_{j \to \infty} \frac{f(y_j + t_j v) - f(y_j)}{t_j} = f^\circ(\bar{x}; v; \mathcal{X}).$$

Using the Lipschitz continuity of $h_{\bar{x}}$ at $\bar{x}$ again, we can get the subsequences $\{y_{j_k}\} \subseteq \{y_j\}$ and $\{t_{j_k}\} \subseteq \{t_j\}$ such that

$$\lim_{k \to \infty} \frac{h_{\bar{x}}(y_{j_k} + t_{j_k} v) - h_{\bar{x}}(y_{j_k})}{t_{j_k}} \quad \text{exists.} \tag{18}$$

By the above analysis, then

$$\begin{aligned} f^\circ(\bar{x}; v; \mathcal{X}) =& \lim_{k \to \infty} \frac{f(y_{j_k} + t_{j_k} v) - f(y_{j_k})}{t_{j_k}} \\ =& \nabla \Theta(\bar{x}) + \nabla c(z)^T_{z = h(\bar{x})} \lim_{k \to \infty} \frac{h_{\bar{x}}(y_{j_k} + t_{j_k} v) - h_{\bar{x}}(y_{j_k})}{t_{j_k}}. \end{aligned} \tag{19}$$

By virtue of (17), we have

$$\begin{aligned} f_{\bar{x}}^{\circ}(\bar{x};v;\mathcal{X}) \geq & \lim_{k\to\infty} \frac{f_{\bar{x}}(y_{j_k}+t_{j_k}v)-f_{\bar{x}}(y_{j_k})}{t_{j_k}} \\ &= \nabla\Theta(\bar{x}) + \nabla c(z)^T_{z=h_{\bar{x}}(\bar{x})} \lim_{k\to\infty} \frac{h_{\bar{x}}(y_{j_k}+t_{j_k}v)-h_{\bar{x}}(y_{j_k})}{t_{j_k}}. \end{aligned} \tag{20}$$

Using $h(\bar{x})=h_{\bar{x}}(\bar{x})$, (19) and (20), we obtain $f_{\bar{x}}^{\circ}(\bar{x};v;\mathcal{X}) \geq f^{\circ}(\bar{x};v;\mathcal{X})$.

On the other hand, by extracting the sequences $\{y_{j_k}\}$ and $\{t_{j_k}\}$ such that the upper limit in (7) holds for $f_{\bar{x}}$ and the limit in (18) exists with them, similar to the above analysis, we find that $f^{\circ}(\bar{x};v;\mathcal{X}) \geq f_{\bar{x}}^{\circ}(\bar{x};v;\mathcal{X})$. Therefore, $f^{\circ}(\bar{x};v;\mathcal{X}) = f_{\bar{x}}^{\circ}(\bar{x};v;\mathcal{X})$. □

2.3. Necessary optimality condition Denote

$$\text{r-int}(\mathcal{T}_{\mathcal{X}}(x)) = \text{int}(\mathcal{T}_{\mathcal{X}_1}(x)) \cap \mathcal{T}_{\mathcal{X}_2}(x).$$

By Lemma 1 (2), r-int$(\mathcal{T}_{\mathcal{X}}(x))$ is not empty for any $x\in\mathcal{X}$.

For a vector $v\in\text{int}(\mathcal{T}_{\mathcal{X}_1}(x))$, there exists a scalar $\delta_1>0$ such that

$$y+tw\in\mathcal{X}_1, \quad \text{for all } y\in\mathcal{X}_1\cap B_{\delta_1}(x),\ w\in B_{\delta_1}(v) \text{ and } 0\leq t<\delta_1.$$

We often call $\text{int}(\mathcal{T}_{\mathcal{X}_1}(x))$ the hypertangent cone to $\mathcal{X}_1$ at x. Whenever $\text{int}(\mathcal{T}_{\mathcal{X}_1}(x)) \neq \emptyset$, $\text{cl}(\text{int}(\mathcal{T}_{\mathcal{X}_1}(x))) = \mathcal{T}_{\mathcal{X}_1}(x)$ and $\text{cl}(\text{r-int}(\mathcal{T}_{\mathcal{X}}(x))) = \mathcal{T}_{\mathcal{X}}(x)$ (Rockafellar [48]).

By Lemma 1 (1), for $v\in\mathcal{T}_{\mathcal{X}_2}(x)$, there is $\delta_2>0$ such that $y+tv\in\mathcal{X}_2$, $\forall y\in\mathcal{X}_2\cap B_{\delta_2}(x)$, $0\leq t<\delta_2$.

Therefore, for any vector $v\in\text{r-int}(\mathcal{T}_{\mathcal{X}}(x))$, there exists a scalar $\epsilon>0$ such that

$$y+tv\in\mathcal{X}, \quad \text{for all } y\in\mathcal{X}\cap B_{\epsilon}(x),\ 0\leq t<\epsilon. \tag{21}$$

Since f in (1) may not be locally Lipschitz continuous at some points, the calculus theory developed in Audet and Dennis [2] cannot be directly applied. The next lemma extends calculus results for the unconstrained case in Clarke [24] and the constrained case in Audet and Dennis [2].

Lemma 2. *For $\bar{x}\in\mathcal{X}$ and $v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\mathcal{V}_{\bar{x}}$, if r-int$(\mathcal{T}_{\mathcal{X}}(\bar{x}))\cap\mathcal{V}_{\bar{x}}\neq\emptyset$, then*

$$f^{\circ}(\bar{x};v;\mathcal{X}) = \lim_{\substack{w\,\to\,v \\ w\,\in\,\text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x}))\,\cap\,\mathcal{V}_{\bar{x}}}} f^{\circ}(\bar{x};w;\mathcal{X}).$$

Proof. By the locally Lipschitz continuity of $h_{\bar{x}}$, there are $\epsilon>0$ and $L_{\bar{x}}>0$ such that

$$\|h_{\bar{x}}(x)-h_{\bar{x}}(y)\|_2 \leq L_{\bar{x}}\|x-y\|_2, \quad \forall x,\, y\in B_{\epsilon}(\bar{x}). \tag{22}$$

Let $\{w_k\}\subseteq\text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x}))\cap\mathcal{V}_{\bar{x}}$ be a sequence of directions converging to a vector $v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\mathcal{V}_{\bar{x}}$.

By $\{w_k\}\subseteq\text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x}))$ and (21), there exists $\epsilon_k>0$ such that $x+tw_k\in\mathcal{X}$ whenever $x\in\mathcal{X}\cap B_{\epsilon_k}(\bar{x})$ and $0\leq t<\epsilon_k$. Then, for all w_k, it gives

$$\begin{aligned} f^{\circ}(\bar{x};v;\mathcal{X}) = & \limsup_{\substack{x\,\to\,\bar{x},\, x\,\in\,\mathcal{X} \\ t\,\downarrow\,0,\, x\,+\,tv\,\in\,\mathcal{X}}} \frac{f(x+tv)-f(x)}{t} \\ = & \limsup_{\substack{x\,\to\,\bar{x},\, x\,\in\,\mathcal{X} \\ t\,\downarrow\,0,\, x\,+\,tv\,\in\,\mathcal{X} \\ x\,+\,tw_k\,\in\,\mathcal{X}}} \frac{f(x+tv)-f(x)}{t} \\ = & \limsup_{\substack{x\,\to\,\bar{x},\, x\,\in\,\mathcal{X} \\ t\,\downarrow\,0,\, x\,+\,tv\,\in\,\mathcal{X} \\ x\,+\,tw_k\,\in\,\mathcal{X}}} \frac{f(x+tw_k)-f(x)}{t} + \frac{f(x+tv)-f(x+tw_k)}{t}. \end{aligned} \tag{23}$$

Let $\delta > 0$ be such that $x + tw_k \in B_\epsilon(\bar{x})$ for any $x \in B_\delta(\bar{x})$, $0 \le t < \delta$ and $k \in \mathbb{N}$. By the Lipschitz property in (22), we have

$$\left\| \frac{h_{\bar{x}}(x+tv) - h_{\bar{x}}(x+tw_k)}{t} \right\|_2 \le L_{\bar{x}} \|v - w_k\|_2, \quad \forall x \in B_\delta(\bar{x}), 0 < t < \delta, k \in \mathbb{N}.$$

From the mean value theorem, there exists z between $h(x+tv)$ and $h(x+tw_k)$ such that

$$\begin{aligned} & f(x+tv) - f(x+tw_k) \\ =& \Theta(x+tv) - \Theta(x+tw_k) + \nabla c(z)^T (h(x+tv) - h(x+tw_k)) \\ =& \Theta(x+tv) - \Theta(x+tw_k) + \nabla c(z)^T (h_{\bar{x}}(x+tv) - h_{\bar{x}}(x+tw_k)). \end{aligned}$$

Then, for any $x \in B_\delta(\bar{x})$ and $t \in (0, \delta)$, we have

$$\left| \frac{f(x+tv) - f(x+tw_k)}{t} \right| \le L_\Theta \|v - w_k\|_2 + L_c L_{\bar{x}} \|v - w_k\|_2,$$

where $L_\Theta = \sup\{\|\nabla\Theta(y)\|_2 : y \in B_\epsilon(\bar{x})\}$ and $L_c = \sup\{\|\nabla c(z)_{z=h(y)}\|_2 : y \in B_\epsilon(\bar{x})\}$.

Thus, (23) implies

$$\begin{aligned} f^\circ(\bar{x}; w_k; \mathcal{X}) - & L_\Theta \|v - w_k\|_2 - L_c L_{\bar{x}} \|v - w_k\|_2 \le f^\circ(\bar{x}; v; \mathcal{X}) \\ & \le f^\circ(\bar{x}; w_k; \mathcal{X}) + L_\Theta \|v - w_k\|_2 + L_c L_{\bar{x}} \|v - w_k\|_2, \forall k \in \mathbb{N}. \end{aligned}$$

As k goes to infinity, the above inequality gives $f^\circ(\bar{x}; v; \mathcal{X}) = \lim_{k\to\infty} f^\circ(\bar{x}; w_k; \mathcal{X})$. Since $\{w_k\}$ is an arbitrary sequence in r-int$(\mathcal{T}_\mathcal{X}(\bar{x})) \cap \mathcal{V}_{\bar{x}}$ converging to v, we obtain the result in this lemma. $\square$

Note that the limit in Lemma 2 is not necessarily true when r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))$ is empty, even for the case $\mathcal{V}_{\bar{x}} = R^n$. A similar example can be given following the idea in Audet and Dennis [2, Example 3.10]. Some other conditions on $\mathcal{X}$ to ensure that the set consisting of the vectors v satisfying (21) is not empty can be used to obtain the limit in Lemma 2. A sufficient condition for (21) is int$(\mathcal{X}_1) \cap \mathcal{X}_2 \neq \emptyset$. Based on Lemmas 1-2, the following theorem gives the main theoretical result of this section.

THEOREM 1. *Suppose the function h in (1) has the form in (13) and Assumption 1 holds for $\mathcal{X}$. If x^* is a local minimizer of (1) and* r-int$(\mathcal{T}_\mathcal{X}(x^*)) \cap \mathcal{V}_{x^*} \neq \emptyset$, *then $f^\circ(x^*; v; \mathcal{X}) \ge 0$ for every direction $v \in \mathcal{T}_\mathcal{X}(x^*) \cap \mathcal{V}_{x^*}$.*

PROOF. Suppose x^* is a local minimizer of f over $\mathcal{X}$ and let $w \in$ r-int$(\mathcal{T}_\mathcal{X}(x^*)) \cap \mathcal{V}_{x^*}$.

Since $w \in$ r-int$(\mathcal{T}_\mathcal{X}(x^*))$, by (21), there exists $\epsilon > 0$ such that

$$x + tw \in \mathcal{X}, \forall x \in \mathcal{X} \cap B_\epsilon(x^*), 0 \le t \le \epsilon.$$

And there exist $\bar{\epsilon} \in (0, \epsilon]$ and $L_{x^*} > 0$ such that $f(x^*) \le f(x)$, $\forall x \in \mathcal{X} \cap B_\epsilon(x^*)$, and

$$\|h_{x^*}(x) - h_{x^*}(y)\|_2 \le L_{x^*} \|x - y\|_2, \ \forall x, y \in \mathcal{X} \cap B_\epsilon(x^*). \tag{24}$$

Then, we can choose $\delta \in (0, \bar{\epsilon}]$ such that $x, x+tw, x^*+tw \in B_\epsilon(x^*) \cap \mathcal{X}, \forall x \in B_{t^2}(x^*) \cap \mathcal{X}$, $0 \le t < \delta$. By (24), for all $x \in B_{t^2}(x^*) \cap \mathcal{X}$, $0 < t < \delta$, we obtain

$$\left\| \frac{h_{x^*}(x+tw) - h_{x^*}(x^*+tw)}{t} - \frac{h_{x^*}(x) - h_{x^*}(x^*)}{t} \right\|_2 \le 2L_{x^*} \frac{\|x - x^*\|_2}{t} \le 2L_{x^*} t.$$

Thus,

$$\lim_{\substack{x \in B_{t^2}(x^*) \cap \mathcal{X} \\ x + tw \in \mathcal{X}, t \downarrow 0}} \frac{h_{x^*}(x+tw) - h_{x^*}(x^*+tw)}{t} - \frac{h_{x^*}(x) - h_{x^*}(x^*)}{t} = 0. \tag{25}$$

From the mean value theorem, there exist z_1 between $h(x)$ and $h(x+tw)$, and z_2 between $h(x^*)$ and $h(x^*+tw)$ such that

$$\begin{aligned}&|(c(h(x+tw))-c(h(x)))-(c(h(x^*+tw))-c(h(x^*)))|\\=&|\nabla c(z_1)^T(h(x+tw)-h(x))-\nabla c(z_2)^T(h(x^*+tw)-h(x^*))|\\=&|\nabla c(z_1)^T(h_{x^*}(x+tw)-h_{x^*}(x))-\nabla c(z_2)^T(h_{x^*}(x^*+tw)-h_{x^*}(x^*))|.\end{aligned} \tag{26}$$

Using (25), (26) and the continuous differentiability of Θ, we have

$$\begin{aligned}&\lim_{\substack{x\in B_{t^2}(x^*)\cap\mathcal{X}\\ x+tw\in\mathcal{X},t\downarrow 0}}\left[\frac{f(x+tw)-f(x)}{t}-\frac{f(x^*+tw)-f(x^*)}{t}\right]\\=&\nabla c(z)^T_{z=h(x^*)}\lim_{\substack{x\in B_{t^2}(x^*)\cap\mathcal{X}\\ x+tw\in\mathcal{X},t\downarrow 0}}\left[\frac{h_{x^*}(x+tw)-h_{x^*}(x)}{t}-\frac{h_{x^*}(x^*+tw)-h_{x^*}(x^*)}{t}\right]\\=&0.\end{aligned}$$

Thus,

$$\limsup_{\substack{x\to x^*,x\in\mathcal{X}\\ t\downarrow 0,x+tw\in\mathcal{X}}}\left[\frac{f(x+tw)-f(x)}{t}-\frac{f(x^*+tw)-f(x^*)}{t}\right]\geq 0. \tag{27}$$

By $f(x^*+tw)-f(x^*)\geq 0$ for $0\leq t<\bar{\epsilon}$, (27) implies

$$f^\circ(x^*;w;\mathcal{X})=\limsup_{\substack{x\to x^*,x\in\mathcal{X}\\ t\downarrow 0,x+tw\in\mathcal{X}}}\frac{f(x+tw)-f(x)}{t}\geq 0.$$

By Lemma 2, we can give that $f^\circ(x^*;v;\mathcal{X})\geq 0$ for any $v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$. □

Based on Theorem 1, we give a new definition of the generalized stationary point of problem (1).

Definition 2. $x^*\in\mathcal{X}$ is said to be a generalized stationary point of (1), if r-int$(\mathcal{T}_{\mathcal{X}}(x^*))\cap\mathcal{V}_{x^*}=\emptyset$ or $f^\circ(x^*;v;\mathcal{X})\geq 0$ for every $v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$.

$\mathcal{T}_{\mathcal{X}}(x)\cap\mathcal{V}_x$ is nonempty for any $x\in\mathcal{X}$, but $f^\circ(x^*;v;\mathcal{X})\geq 0$, $\forall v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$ is not necessarily true at the local minimizer x^* of f over $\mathcal{X}$ when r-int$(\mathcal{T}_{\mathcal{X}}(x^*))\cap\mathcal{V}_{x^*}=\emptyset$. For example, if $f(x_1,x_2)=\sqrt{|x_1-2|}+4\sqrt{|x_2-1|}$ and $\mathcal{X}=\{(x_1,x_2):(x_1-1)^2+(x_2-2)^2\leq 1\}$, then $f^\circ(x^*;v;\mathcal{X})=-\frac{1}{2}$, where $x^*=(1,1)^T$ is the global minimizer of f over $\mathcal{X}$ and $v=(1,0)^T\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$. So Definition 2 is reasonable and robust.

We call f Lipschitz continuous at $\bar{x}\in\mathcal{X}$ in direction v, if there exist $L>0$ and $\varepsilon>0$ such that

$$|f(\bar{x}+tv)-f(\bar{x})|\leq Lt\|v\|_2,\quad \forall t\in(0,\varepsilon).$$

When Assumption 1 holds, if f is not Lipschitz continuous at $\bar{x}\in\mathcal{X}$ in direction $x-\bar{x}$ for any $x\in\mathcal{X}$, r-int$(\mathcal{T}_{\mathcal{X}}(x))\cap\mathcal{V}_x=\emptyset$ or $\mathcal{T}_{\mathcal{X}}(x)\cap\mathcal{V}_x=\{0\}$, which implies that $\bar{x}$ is a trivial generalized stationary point of (1).

Corollary 1. *Suppose the function h in (1) has the form in (13) and Assumption 1 holds. Then the following statements hold.*

(1) When f is Lipschitz continuous at $x^*\in\mathcal{X}$, x^* is a generalized stationary point of (1) defined in Definition 2 if and only if it is a Clarke stationary point of (1).

(2) When $\mathcal{X}:=\{x:Ax\leq b\}$ with $A\in R^{t\times n}$ and $b\in R^t$, x^* is a generalized stationary point of (1) defined in Definition 2 if and only if

$$f^\circ(x^*;v;\mathcal{X})\geq 0,\quad \textit{for every } v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}.$$

Proof. When f is Lipschitz continuous at $x^* \in \mathcal{X}$, by r-int$(\mathcal{T}_{\mathcal{X}}(x)) \neq \emptyset$ and $\mathcal{V}_{x^*} = R^n$, x^* is a generalized stationary point of (1) if and only if

$$f^\circ(x^*; v; \mathcal{X}) \geq 0 \quad \text{for every } v \in \mathcal{T}_{\mathcal{X}}(x^*),$$

which means that x^* is a Clarke stationary point of (1).

When $\mathcal{X} := \{x : Ax \leq b\}$, by letting $\mathcal{X}_1 = R^n$ and $\mathcal{X}_2 = \{x : Ax \leq b\}$, we find that r-int$(\mathcal{T}_{\mathcal{X}}(x^*)) = \mathcal{T}_{\mathcal{X}_2}(x^*)$ and then $0 \in$ r-int$(\mathcal{T}_{\mathcal{X}}(x^*)) \cap \mathcal{V}_{x^*}$. So statement (2) is true. $\square$

Remark 1. Suppose h_i is regular in $\{D_i^T x : x \in \mathcal{X} \backslash \mathcal{N}_i\}$, where

$$\mathcal{N}_i = \{x \in \mathcal{X} : h_i \text{ is not Lipschitz continuous at } D_i^T x\},\ i = 1, 2, \ldots, m.$$

Then, the statement in Theorem 1 holds with $\tilde{\mathcal{V}}_{x^*}$ instead of $\mathcal{V}_{x^*}$, where

$$\begin{aligned} \tilde{\mathcal{V}}_{x^*} = \{v : & \text{for any } i \in \mathcal{I}_{x^*}, \text{ there exists } \delta > 0 \text{ such that} \\ & h_i(D_i^T(x^* + tv)) = h_i(D_i^T x^*) \text{ holds for all } 0 \leq t \leq \delta\}. \end{aligned} \tag{28}$$

Since $\mathcal{V}_{x^*} \subseteq \tilde{\mathcal{V}}_{x^*}$, the generalized stationary point defined in Definition 2 can be more robust with $\tilde{\mathcal{V}}_{x^*}$ instead of $\mathcal{V}_{x^*}$ for some cases. For example, if f is modeled by (2), $\tilde{\mathcal{V}}_{x^*} = \mathcal{V}_{x^*}$; however, if f is defined as in (3),

$$\tilde{\mathcal{V}}_{x^*} = \{v : d_i^T v \geq 0,\ \forall\, i \in \{i : d_i^T x^* = \alpha_i\}\}, \tag{29}$$

which includes $\mathcal{V}_{x^*} = \{v : d_i^T v = 0,\ \forall i \in \{i : d_i^T x^* = \alpha_i\}\}$ as a proper subset.

Other necessary optimality conditions for the special cases of (1) are studied by Bian and Chen [6, 8], Bian et al. [9], Chen et al. [21], Ge et al. [30], Liu et al. [40]. Bian and Chen [8] considered a special case of (1) modeled by (2) with $l = (-\infty)^n$, $u = \infty^n$ and $r = 1$, that is

$$\begin{aligned} \min \quad & f(x) := \Theta(x) + \sum_{i=1}^m \varphi(|d_i^T x|^p) \\ \text{s.t.} \quad & x \in \mathcal{X} := \{x : Ax \leq b\}. \end{aligned} \tag{30}$$

Recall $\mathcal{C}_{\bar{x}} = \{i : A_i^T \bar{x} - b_i = 0\}$. If $\bar{x} \in \mathcal{X}$ is a local minimizer of (30), there exists a nonnegative vector $\gamma \in R^{|\mathcal{C}_{\bar{x}}|}$ such that

$$Z_{\bar{x}}^T (\nabla\Theta(\bar{x}) + \sum_{i \notin \mathcal{I}_{\bar{x}}} p\varphi'(s)_{s = |d_i^T \bar{x}|^p} |d_i^T \bar{x}|^{p-1} \text{sign}(d_i^T \bar{x}) d_i + \sum_{i \in \mathcal{C}_{\bar{x}}} \gamma_i A_i) = 0, \tag{31}$$

where $\mathcal{I}_{\bar{x}} = \{i : d_i^T \bar{x} = 0\}$ and $Z_{\bar{x}}$ is a matrix whose columns form an orthogonal basis of the null space of $\{d_i : d_i^T \bar{x} = 0\}$ (Bian and Chen [8]). Most recently, Liu et al. [40] considered a special case of (1) modeled by (3) with $\varphi(t) := t$ and $\mathcal{X} := \{x : Ax \leq b\}$. They called $\bar{x} \in \mathcal{X}$ a KKT point of (3), if there exists a nonnegative vector $\bar{\lambda} \in R^{|\mathcal{I}_{\bar{x}}|}$ such that

$$\bar{x} = P_{\mathcal{X}}(\bar{x} - \nabla L(\bar{x}, \bar{\lambda})), \tag{32}$$

where $L(x, \lambda) = \Theta(x) + \sum_{i \in \mathcal{J}_{\bar{x}}} (\alpha_i - d_i^T x)^p + \sum_{i \in \mathcal{I}_{\bar{x}}} \lambda_i (\alpha_i - d_i^T x)$ with $\mathcal{J}_{\bar{x}} = \{i : \alpha_i - d_i^T \bar{x} > 0\}$ and $\mathcal{I}_{\bar{x}} = \{i : \alpha_i - d_i^T \bar{x} = 0\}$.

For problems (30) and (3), the next proposition shows the expressions of the optimality conditions in Bian and Chen [8] and Liu et al. [40] with the generalized directional derivative and tangent cone.

Proposition 2. *(1) When problem (1) reduces to problem (30), for $\bar{x} \in \mathcal{X}$, there holds*

$$f^\circ(\bar{x}; v; \mathcal{X}) \geq 0,\ \forall v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \tilde{\mathcal{V}}_{\bar{x}} \quad \Longleftrightarrow \quad \text{(31) holds with a nonnegative vector } \gamma \in R^{|\mathcal{C}_{\bar{x}}|};$$

(2) When problem (1) reduces problem (3) with $\mathcal{X} := \{x : Ax \leq b\}$, for $\bar{x} \in \mathcal{X}$, there holds

$$f^\circ(\bar{x}; v; \mathcal{X}) \geq 0,\ \forall v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \tilde{\mathcal{V}}_{\bar{x}} \quad \Longleftrightarrow \quad \text{(32) holds with a nonnegative vector } \bar{\lambda} \in R^{|\mathcal{I}_{\bar{x}}|}.$$

Proof. First, we establish the direction " $\Longleftarrow$ " in statement (1). From the definition of $f_{\bar{x}}$ in (16) for (30), we can easily verify that

$$\nabla f_{\bar{x}}(\bar{x}) = \nabla\Theta(\bar{x}) + \sum_{i\notin\mathcal{I}_{\bar{x}}} p\varphi'(s)_{s=|d_i^T\bar{x}|^p}|d_i^T\bar{x}|^{p-1}\mathrm{sign}(d_i^T\bar{x})d_i.$$

If $\bar{x}$ satisfies (31) with a nonnegative vector $\gamma\in R^{|\mathcal{C}_{\bar{x}}|}$, by the definition of $Z_{\bar{x}}$, we obtain

$$\nabla f_{\bar{x}}(\bar{x}) + \sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i \in \mathrm{span}\{d_i : i\in\mathcal{I}_{\bar{x}}\},$$

which implies the existence of a nonnegative vector $\kappa\in R^{|\mathcal{I}_{\bar{x}}|}$ such that

$$\nabla f_{\bar{x}}(\bar{x}) + \sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i = \sum_{i\in\mathcal{I}_{\bar{x}}}\kappa_i d_i.$$

Since $\mathcal{T}_{\mathcal{X}}(\bar{x}) = \{v : A_i^T v\leq 0, \forall i\in\mathcal{C}_{\bar{x}}\}$, $\tilde{\mathcal{V}}_{\bar{x}} = \{v : d_i^T v = 0, \forall i\in\mathcal{I}_{\bar{x}}\}$ and $\gamma\geq 0$, for any $v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\tilde{\mathcal{V}}_{\bar{x}}$, by Proposition 1, it gives

$$f^\circ(\bar{x};v;\mathcal{X}) = \langle\nabla f_{\bar{x}}(\bar{x}), v\rangle = \langle\sum_{i\in\mathcal{I}_{\bar{x}}}\kappa_i d_i - \sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i, v\rangle \geq 0.$$

Then, we prove the direction " $\Longrightarrow$ " in statement (1). Denote $S = \{x : d_i^T x = 0, i\in\mathcal{I}_{\bar{x}}\}$. By $\tilde{\mathcal{V}}_{\bar{x}} = \mathcal{T}_S(\bar{x})$ and Proposition 1, $f^\circ(\bar{x};v;\mathcal{X})\geq 0$, $\forall v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\tilde{\mathcal{V}}_{\bar{x}}$, implies

$$\langle\nabla f_{\bar{x}}(\bar{x}), v\rangle \geq 0,\ \forall v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\mathcal{T}_S(\bar{x}). \tag{33}$$

Since $\mathcal{T}_{\mathcal{X}\cap S}(\bar{x})\subseteq\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\mathcal{T}_S(\bar{x})$, (33) gives

$$-\nabla f_{\bar{x}}(\bar{x})\in\mathcal{N}_{\mathcal{X}\cap S}(\bar{x}).$$

From the definition of $\mathcal{X}$ and S, $\mathcal{N}_{\mathcal{X}\cap S}(\bar{x})\subseteq\mathcal{N}_{\mathcal{X}}(\bar{x})+\mathcal{N}_S(\bar{x}) = \{\sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i : \gamma_i\geq 0\} + \{\sum_{i\in\mathcal{I}_{\bar{x}}}\lambda_i d_i : \lambda_i\in R\}$, which ensures the existence of a nonnegative vector $\gamma\in R^{|\mathcal{C}_{\bar{x}}|}$ such that

$$-\nabla f_{\bar{x}}(\bar{x}) - \sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i \in \{\sum_{i\in\mathcal{I}_{\bar{x}}}\lambda_i d_i : \lambda_i\in R\}.$$

Thus,

$$Z_{\bar{x}}^T(\nabla f_{\bar{x}}(\bar{x}) + \sum_{i\in\mathcal{C}_{\bar{x}}}\gamma_i A_i) = 0.$$

Next, we show the second statement in this proposition. For problem (3), $\tilde{\mathcal{V}}_{\bar{x}} = \{v : d_i^T v\geq 0,\ \forall i\in\mathcal{I}_{\bar{x}}\}$.

Suppose $\bar{x}\in\mathcal{X}$ satisfies (32) with $\bar{\lambda}\geq 0$, by the projection inequality, it holds

$$\langle\nabla\Theta(\bar{x}) - \sum_{i\in\mathcal{J}_{\bar{x}}}p(\alpha_i - d_i^T\bar{x})^{p-1}d_i - \sum_{i\in\mathcal{I}_{\bar{x}}}\bar{\lambda}_i d_i, x-\bar{x}\rangle \geq 0,\ \forall x\in\mathcal{X}, \tag{34}$$

which implies

$$\langle\nabla\Theta(\bar{x}) - \sum_{i\in\mathcal{J}_{\bar{x}}}p(\alpha_i - d_i^T\bar{x})^{p-1}d_i - \sum_{i\in\mathcal{I}_{\bar{x}}}\bar{\lambda}_i d_i, v\rangle \geq 0,\ \forall v\in\mathcal{T}_{\mathcal{X}}(\bar{x}).$$

By the definition of $\tilde{\mathcal{V}}_{\bar{x}}$ and $\mathcal{I}_{\bar{x}}$, we obtain

$$\langle\nabla\Theta(\bar{x}) - \sum_{i\in\mathcal{J}_{\bar{x}}}p(\alpha_i - d_i^T\bar{x})^{p-1}d_i, v\rangle \geq 0,\ \forall v\in\mathcal{T}_{\mathcal{X}}(\bar{x})\cap\tilde{\mathcal{V}}_{\bar{x}}. \tag{35}$$

Proposition 1 gives

$$f^\circ(\bar{x};v;\mathcal{X}) = \langle \nabla\Theta(\bar{x}) - \sum_{i\in\mathcal{J}_{\bar{x}}} p(\alpha_i - d_i^T\bar{x})^{p-1}d_i, v\rangle. \tag{36}$$

Then, (35) and (36) guarantee

$$f^\circ(\bar{x};v;\mathcal{X}) \geq 0,\ \forall v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \tilde{\mathcal{V}}_{\bar{x}}.$$

Conversely, by (36), similar to the analysis for the first statement, $f^\circ(\bar{x};v;\mathcal{X}) \geq 0, \forall v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \tilde{\mathcal{V}}_{\bar{x}}$, implies

$$-\nabla\Theta(\bar{x}) + \sum_{i\in\mathcal{J}_{\bar{x}}} p(\alpha_i - d_i^T\bar{x})^{p-1}d_i \in \mathcal{N}_{\mathcal{X}}(\bar{x}) + \mathcal{N}_S(\bar{x}),$$

where $\mathcal{N}_S(\bar{x}) = \{\sum_{i\in\mathcal{I}_{\bar{x}}} \lambda_i(-d_i) : \lambda_i \geq 0\}$ with $S = \{x : \alpha_i - d_i^T x \leq 0, \forall i \in \mathcal{I}_{\bar{x}}\}$. Then, there exists a nonnegative vector $\bar{\lambda} \in R^{|\mathcal{I}_{\bar{x}}|}$ such that

$$-\nabla\Theta(\bar{x}) + \sum_{i\in\mathcal{J}_{\bar{x}}} p(\alpha_i - d_i^T\bar{x})^{p-1}d_i + \sum_{i\in\mathcal{I}_{\bar{x}}} \bar{\lambda}_i d_i \in \mathcal{N}_{\mathcal{X}}(\bar{x}).$$

Thus,

$$\langle \nabla\Theta(\bar{x}) + \sum_{i\in\mathcal{J}_{\bar{x}}} p(\alpha_i - d_i^T\bar{x})^{p-1}d_i - \sum_{i\in\mathcal{I}_{\bar{x}}} \bar{\lambda}_i d_i, x - \bar{x}\rangle \geq 0,\ \forall x \in \mathcal{X},$$

which implies $\bar{x}$ satisfies (32) with a nonnegative vector $\bar{\lambda}$. □

When problem (1) reduces to (30) and (3), by Remark 1, the necessary optimality condition given in Theorem 1 holds with $\tilde{\mathcal{V}}_{x^*}$ instead of $\mathcal{V}_{x^*}$. And from Corollary 1, the optimality conditions in Bian and Chen [8] and Liu et al. [40] are equivalent to the necessary optimality conditions given in this paper for problems (30) and (3). Similarly, the generalized stationary point defined in Definition 2 reduces to the first order necessary optimality conditions given or used in Bian and Chen [6], Bian et al. [9], Chen et al. [21, 23], Ge et al. [30] for the special cases of (2) with $0 < p < 1$. Thus, Definition 2 provides a uniform version of the existing necessary optimality conditions for the Lipschitz and non-Lipschitz problems modeled by (1). In computation, by the properties of generalized directional derivative and tangent cone, some closed forms can be derived from our optimality condition for different special cases. However, when $p = 1$ in (2) or (3), a generalized stationary point defined in Definition 2 is a Clarke stationary point, while the scaled stationary point, first-order stationary point, scaled KKT point and KKT point in Bian and Chen [6, 8], Bian et al. [9], Chen et al. [21, 23], Ge et al. [30], Liu et al. [40] are not necessarily Clarke stationary points. All of these not only show the robustness of the optimality conditions given in this paper, but also illustrate the superiority of the generalized direction derivative defined in (17) for studying the optimality conditions of problem (1).

For a locally Lipschitz function f, we have

$$\partial f(x) = \mathrm{con}\bar{\partial} f(x),$$

where ∂ is the Clarke subdifferential, $\bar{\partial}$ is the limiting (Mordukhovich) subdifferential and " con" denotes the convex hull (Rockafellar and Wets [47, Theorem 8.49]). Thus a Clarke stationary point defined in (6) is necessary but not sufficient to be a limiting (Mordukhovich) stationary point defined in (11) when f is locally Lipschitz continuous. Some interesting results on this topic can also be found in Burke et al. [15]. When the constraint qualification (10) holds, the necessary optimality condition given in this paper is also weaker than condition (11). However, the constraint qualification (10) is likely to be unsatisfied for many non-Lipschitz optimization problems and $\bar{\partial} f(x)$ may be empty at some non-Lipschitz points of f.

2.4. Directional derivative consistency In this subsection, we show that the generalized directional derivative of f defined in (17) can be represented by the limit of a sequence of directional derivatives of a smoothing function of f. This property is important for the development of numerical algorithms in nonsmooth optimization.

Definition 3. Let $\phi : R^n \to R$ be a continuous function. We call $\tilde{\phi} : R^n \times [0,\infty) \to R$ a smoothing function of ϕ, if $\tilde{\phi}(\cdot,\mu)$ is continuously differentiable for any fixed $\mu > 0$ and $\lim_{z\to x,\mu\downarrow 0} \tilde{\phi}(z,\mu) = \phi(x)$ holds for any $x \in R^n$.

When ϕ is Lipschitz continuous at $\bar{x}$, we call the gradient consistency associated with the smoothing function $\tilde{\phi}$ holds at $\bar{x}$, if

$$\{\lim_{z\to\bar{x},\mu\downarrow 0} \nabla_x \tilde{\phi}(z,\mu)\} \subseteq \partial\phi(\bar{x}). \tag{37}$$

Some conditions to guarantee (37) can be found in Chen [19], Burke and Hoheisel [13], Burke et al. [14], Rockafellar and Wets [47].

Let $\tilde{h}(x,\mu) := (\tilde{h}_1(D_1^T x,\mu), \tilde{h}_2(D_2^T x,\mu), \ldots, \tilde{h}_m(D_m^T x,\mu))^T$, where $\tilde{h}_i$ is a smoothing function of h_i in (13) and the gradient consistency associated with $\tilde{h}_i$ holds at its Lipschitz continuous points. Then $\tilde{f}(x,\mu) := \Theta(x) + c(\tilde{h}(x,\mu))$ is a smoothing function of f.

Since $\tilde{f}(\cdot,\mu)$ is continuously differentiable for any fixed $\mu > 0$, the generalized directional derivative of it with respect to x can be given by

$$\tilde{f}^\circ(x,\mu;v;\mathcal{X}) = \limsup_{\substack{y\to x,\, y\in\mathcal{X}\\ t\downarrow 0,\, y+tv\in\mathcal{X}}} \frac{\tilde{f}(y+tv,\mu) - \tilde{f}(y,\mu)}{t} = \langle \nabla_x \tilde{f}(x,\mu), v\rangle. \tag{38}$$

Theorem 2. *Suppose the function h in (1) has the form in (13), and h_i is continuously differentiable at $D_i^T x$ for $x \in \mathcal{X}\backslash\mathcal{N}_i$ with $\mathcal{N}_i = \{x :\ h_i$ is not Lipschitz continuous at $D_i^T x\}$, $i \in \{1,2,\ldots,m\}$, then*

$$\lim_{\substack{x_k\in\mathcal{X},\\ x_k\to\bar{x},\,\mu_k\downarrow 0}} \langle \nabla_x \tilde{f}(x_k,\mu_k), v\rangle = f^\circ(\bar{x};v;\mathcal{X}),\ \forall v \in \mathcal{V}_{\bar{x}}. \tag{39}$$

Proof. Let x_k be a sequence in $\mathcal{X}$ converging to $\bar{x}$ and $\{\mu_k\}$ be a positive sequence converging to 0. For $v \in \mathcal{V}_{\bar{x}}$, by the differentiability of $\tilde{f}(x_k,\mu_k)$, we have

$$\langle \nabla_x \tilde{f}(x_k,\mu_k), v\rangle = \langle \nabla\Theta(x_k), w\rangle + \langle \nabla c(z)_{z=\tilde{h}(x_k,\mu_k)}, \nabla_x \tilde{h}(x_k,\mu_k)^T v\rangle, \tag{40}$$

where

$$\nabla_x \tilde{h}(x_k,\mu_k)^T v = (\nabla_x \tilde{h}_1(D_1^T x_k,\mu_k)^T v, \ldots, \nabla_x \tilde{h}_m(D_m^T x_k,\mu_k)^T v)^T.$$

For $i \in \mathcal{I}_{\bar{x}}$, by $v \in \mathcal{V}_{\bar{x}}$, we obtain $D_i^T v = 0$, then $\nabla_x \tilde{h}_i(D_i^T x_k,\mu_k)^T v = \nabla_z \tilde{h}_i(z,\mu_k)^T_{z=D_i^T x_k} D_i^T v = 0$. Define

$$\tilde{h}_i^{\bar{x}}(D_i^T x,\mu) := \begin{cases} \tilde{h}_i(D_i^T x,\mu) & i \notin \mathcal{I}_{\bar{x}}, \\ \tilde{h}_i(D_i^T \bar{x},\mu) & i \in \mathcal{I}_{\bar{x}}, \end{cases} \quad i = 1,2,\ldots,m.$$

Denote $\tilde{h}_{\bar{x}}(x,\mu) = (\tilde{h}_1^{\bar{x}}(D_1^T x,\mu), \tilde{h}_2^{\bar{x}}(D_2^T x,\mu), \ldots, \tilde{h}_m^{\bar{x}}(D_m^T x,\mu))^T$. Then,

$$\nabla_x \tilde{h}(x_k,\mu_k)^T v = \nabla_x \tilde{h}_{\bar{x}}(x_k,\mu_k)^T v.$$

Thus, coming back to (40), we obtain

$$\begin{aligned} \langle \nabla_x \tilde{f}(x^k,\mu_k), v\rangle =& \langle \nabla\Theta(x_k), v\rangle + \langle \nabla c(z)_{z=\tilde{h}(x_k,\mu_k)}, \nabla_x \tilde{h}_{\bar{x}}(x_k,\mu_k)^T v\rangle \\ =& \langle \nabla\Theta(x_k) + \nabla_x \tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)}, v\rangle. \end{aligned} \tag{41}$$

Since h_i is continuously differentiable at $D_i^T \bar{x}$ for $i \notin \mathcal{I}_{\bar{x}}$ and $h_{\bar{x}}(\bar{x}) = h(\bar{x})$, we obtain

$$\begin{aligned} &\lim_{k\to\infty} \nabla\Theta(x_k) + \nabla_x \tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)} \\ =&\nabla\Theta(\bar{x}) + \nabla h_{\bar{x}}(\bar{x})\nabla c(z)_{z=h(\bar{x})} = \nabla f_{\bar{x}}(\bar{x}), \end{aligned} \tag{42}$$

where $f_{\bar{x}}$ is defined in (16).

Thus,

$$\begin{aligned} f^\circ(\bar{x};v;\mathcal{X}) = f^\circ_{\bar{x}}(\bar{x};v;\mathcal{X}) &= \langle \nabla f_{\bar{x}}(\bar{x}), v\rangle \\ &= \langle \lim_{k\to\infty} \nabla\Theta(x_k) + \nabla_x \tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)}, v\rangle \\ &= \lim_{k\to\infty}\langle \nabla_x \tilde{f}(x^k,\mu_k), v\rangle, \end{aligned} \tag{43}$$

where the first equation uses Proposition 1, the third uses (42) and the fourth uses (41). □

Now we give another consistency result on the subspace $\mathcal{V}_x$.

Lemma 3. *Let $\{x_k\}$ be a sequence in $\mathcal{X}$ with a limit point $\bar{x}$. For $w\in\mathcal{V}_{\bar{x}}$, there exists a sequence $\{x_{k_l}\}\subseteq\{x_k\}$ such that $w\in\mathcal{V}_{x_{k_l}}$, $\forall l\in\mathbb{N}$.*

Proof. If this lemma is not true, then there is $K\in\mathbb{N}$ such that

$$w\notin\mathcal{V}_{x_k},\quad \forall k\geq K.$$

By the definition of $\mathcal{V}_{x_k}$, there exists $i_k\in\mathcal{I}_{x_k}$ such that

$$D_{i_k}^T w\neq 0,\quad \forall k\geq K.$$

From $\mathcal{I}_{x_k}\subseteq\{1,2,\ldots,m\}$, there exist $j\in\{1,2,\ldots,m\}$ and a subsequence of $\{x_k\}$, denoted as $\{x_{k_l}\}$, such that $j\in\mathcal{I}_{x_{k_l}}$ and $D_j^T w\neq 0$.

Note that $j\in\mathcal{I}_{x_{k_l}}$ implies h_j is not Lipschitz continuous at $D_j^T x_{k_l}$. Since the non-Lipschitz points of h_j is a closed subset of R^n, h_j is also not Lipschitz continuous at $D_j^T\bar{x}$, which means $j\in\mathcal{I}_{\bar{x}}$. By $w\in\mathcal{V}_{\bar{x}}$, we obtain $D_j^T w=0$, which leads to a contradiction. Therefore, the statement in this lemma holds. □

For $\bar{x}\in\mathcal{X}$, from the definitions of r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))$ and $\mathcal{V}_{\bar{x}}$, r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))\cap\mathcal{V}_{\bar{x}}\neq\emptyset$ implies r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))\cap\mathcal{V}_{\bar{x}}\cap B_1(0)\neq\emptyset$. Based on the consistency results given in Theorem 2 and Lemma 3, we give a corollary to show the generalized stationary point consistency of the smoothing functions.

Corollary 2. *Suppose the function h in (1) has the form in (13). Let $\{\epsilon_k\}$ and $\{\mu_k\}$ be positive sequences converging to 0. With the conditions on h in Theorem 2, if $x^k\in\mathcal{X}$ satisfies*

$$\text{r-int}(\mathcal{T}_\mathcal{X}(x^k))\cap\mathcal{V}_{x^k}=\emptyset \quad \textit{or} \quad \langle\nabla_x\tilde{f}(x^k,\mu_k),v\rangle\geq-\epsilon_k,\ \forall v\in\mathcal{T}_\mathcal{X}(x^k)\cap\mathcal{V}_{x^k}\cap B_1(0), \tag{44}$$

then any accumulation point of $\{x^k\}$ is a generalized stationary point of (1).

Proof. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Then, there exists a subsequence of $\{x^k\}$ (also denoted as $\{x^k\}$) converging to $\bar{x}$, i.e. $\lim_{k\to\infty}x_k=\bar{x}$.

Without loss of generality, we suppose r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))\cap\mathcal{V}_{\bar{x}}\neq\emptyset$. If not, by Definition 2, the statement in this corollary holds naturally.

First, we will show that there is a subsequence of $\{x^k\}$ (also denoted as $\{x^k\}$) such that r-int$(\mathcal{T}_\mathcal{X}(x^k))\cap\mathcal{V}_{x^k}\neq\emptyset$. Denote $\bar{v}\in$ r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))\cap\mathcal{V}_{\bar{x}}$. By Lemma 3, we can suppose $\bar{v}\in\mathcal{V}_{x^k}$. $\bar{v}\in$ int$(\mathcal{T}_{\mathcal{X}_1}(\bar{x}))$ guarantees the existence of $K\in\mathbb{N}$ such that $\bar{v}\in$ int$(\mathcal{T}_{\mathcal{X}_1}(x^k))$, $\forall k\geq K$. Since $\mathcal{C}_{x^k}$ and $\mathcal{C}_{\bar{x}}$ are the subsets of $\{1,2,\ldots,t\}$, by their definitions, there exists a subsequence of $\{x^k\}$ (also denoted as $\{x^k\}$) such that $\mathcal{C}_{x^k}\subseteq\mathcal{C}_{\bar{x}}$. From Lemma 1 (1), we obtain $\mathcal{T}_{\mathcal{X}_2}(\bar{x})\subseteq\mathcal{T}_{\mathcal{X}_2}(x^k)$. Thus, there is a subsequence of $\{x^k\}$ (also denoted as $\{x^k\}$) such that $\bar{v}\in$ r-int$(\mathcal{T}_\mathcal{X}(x^k))\cap\mathcal{V}_{x^k}$.

For $w\in$ r-int$(\mathcal{T}_\mathcal{X}(\bar{x}))\cap\mathcal{V}_{\bar{x}}\cap B_1(0)$, from Lemma 3, we can suppose

$$w\in\mathcal{V}_{x_k},\quad\forall k\in\mathbb{N}.$$

By $w \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x}))$, there exists $\epsilon > 0$ such that

$$x + sw \in \mathcal{X}, \quad \forall x \in \mathcal{X} \cap B_\epsilon(\bar{x}),\ 0 \le s \le \epsilon. \tag{45}$$

Since x_k converges to $\bar{x}$, there exists $K \in \mathbb{N}$ such that $x_k \in \mathcal{X} \cap B_\epsilon(\bar{x})$, $\forall k \ge K$. By (45), we have $x_k + sw \in \mathcal{X}$, $\forall k \ge K$, $0 \le s \le \epsilon$. From the convexity of $\mathcal{X}$, we obtain $w \in \mathcal{T}_{\mathcal{X}}(x_k)$.

From Theorem 2, we have $f^\circ(\bar{x}; w; \mathcal{X}) \ge 0$. Then, for any $\rho > 0$, we have

$$\begin{aligned} f^\circ(\bar{x}; \rho v; \mathcal{X}) = & \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ t \downarrow 0,\, y + t\rho v \in \mathcal{X}}} \frac{f(y + t\rho v) - f(y)}{t} \\ =& \rho \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ s \downarrow 0,\, y + sv \in \mathcal{X}}} \frac{f(y + sv) - f(y)}{s} = \rho f^\circ(\bar{x}; v; \mathcal{X}) \ge 0. \end{aligned} \tag{46}$$

Thus, $f^\circ(\bar{x}; v; \mathcal{X}) \ge 0$ for every $v \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x})) \cap \mathcal{V}_{\bar{x}} \cap B_1(0)$ implies $f^\circ(\bar{x}; v; \mathcal{X}) \ge 0$ for every $v \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x})) \cap \mathcal{V}_{\bar{x}}$. By Lemma 2, it is easy to verify that $f^\circ(\bar{x}; v; \mathcal{X}) \ge 0$ holds for any $v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \mathcal{V}_{\bar{x}}$, which means that $\bar{x}$ is a generalized stationary point of (1). □

From Corollary 2, it is easy to prove that any accumulation point of the generalized stationary points of $\min_{x\in\mathcal{X}} \tilde{f}(x, \mu_k)$ defined by the gradient of $\tilde{f}(x, \mu)$ is a generalized stationary point of (1) defined in Definition 2 as μ_k tends to 0.

Remark 2. Similar to the proof for Theorem 2 and Corollary 2, the consistence result in them holds with $\tilde{\mathcal{V}}_x$ instead of $\mathcal{V}_x$, which is defined in (28) and comes from the necessary optimality condition for the minimizers of (1) in Remark 1.

Remark 3. The gradient consistency of $\tilde{h}_i$ at its Lipschitz continuous points implies

$$\{\lim_{z \to \bar{x}, \mu \downarrow 0} \nabla\Theta(z) + \nabla_x \tilde{h}_{\bar{x}}(z, \mu) \nabla c(y)_{y = \tilde{h}_{\bar{x}}(z,\mu)}\} \subseteq \partial f_{\bar{x}}(\bar{x}). \tag{47}$$

Since $f_{\bar{x}}$ is Lipschitz continuous at $\bar{x}$, it gives

$$f^\circ_{\bar{x}}(\bar{x}; w; R^n) = \max\{\langle \xi, w\rangle : \xi \in \partial f_{\bar{x}}(\bar{x})\}. \tag{48}$$

Similar to the calculation in (43), by (47) and (48), we obtain

$$\begin{aligned} f^\circ(\bar{x}; w; R^n) = f^\circ_{\bar{x}}(\bar{x}; w; R^n) &= \max\{\langle \xi, w\rangle : \xi \in \partial f_{\bar{x}}(\bar{x})\} \\ &\ge \limsup_{k\to\infty} \langle \nabla\Theta(x_k) + \nabla_x \tilde{h}_{\bar{x}}(x_k, \mu_k) \nabla c(z)_{z = \tilde{h}(x_k, \mu_k)}, w\rangle \\ &= \limsup_{k\to\infty} \langle \nabla_x \tilde{f}(x^k, \mu_k), w\rangle. \end{aligned}$$

Thus, when $\mathcal{X} = R^n$, if x^k satisfies $\langle \nabla_x \tilde{f}(x^k, \mu_k), v\rangle \ge -\epsilon_k$ for every $v \in \mathcal{T}_{\mathcal{X}}(x^k) \cap \mathcal{V}_{x^k} \cap B_1(0)$, the conclusions in Theorem 2 and Corollary 2 can be true without the continuous differentiability of h_i at $D_i^T x$ for $x \in \mathcal{X} \backslash \mathcal{N}_i$, $\forall i \in \{1, 2, \ldots, m\}$.

In particular, when the function h in f has the form

$$h(x) := (h_1(d_1^T x), h_2(d_2^T x), \ldots, h_m(d_m^T x))^T$$

with $d_i \in R^n$, by Clarke [24, Theorem 2.3.9 (i)], the regularity of $h_i(d_i^T x)$ in $\mathcal{X} \backslash \mathcal{N}_i$ is a sufficient condition for the statement in Theorem 2 and Corollary 2.

The statement that x^k is an approximate stationary point of $\tilde{f}(x, \mu_k)$ over $\mathcal{X}$, i.e.

$$\langle \nabla_x \tilde{f}(x^k, \mu_k), v\rangle \ge -\epsilon_k \quad \text{ for every } v \in \mathcal{T}_{\mathcal{X}}(x^k) \cap B_1(0)$$

is a sufficient condition for (44). Corollary 2 shows that one can find a generalized stationary point of (1) by using the approximate stationary points of $\min_{x\in\mathcal{X}} \tilde{f}(x, \mu)$ defined by the gradient of $\tilde{f}(x, \mu)$.

Since $\tilde{f}(\cdot,\mu)$ is continuously differentiable for any fixed $\mu>0$, many numerical algorithms can find a stationary point of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu)$ (Beck and Teboulle [4], Curtis and Overton [25], Levitin and Polyak [38], Nocedal and Wright [46], Ye [54]). In what follows, we use one example to show the validity of the first order necessary optimality condition in Theorem 1 and the consistency result given in Corollary 2.

Example 1. Consider the following minimization problem

$$\begin{array}{ll}\min & f(x):=(x_1+2x_2-1)^2+\lambda_1\sqrt{\max\{x_1+x_2+1,0\}}+\lambda_2\sqrt{|x_2|},\\ \text{s.t.} & x\in\mathcal{X}:=\{x\in R^2: -1\le x_1,x_2\le 1\}.\end{array} \tag{49}$$

This problem is an example of (1) with $\Theta(x)=(x_1+2x_2-1)^2$, $c(y)=\lambda_1y_1+\lambda_2y_2$, $h_1(D_1^Tx)=\sqrt{\max\{x_1+x_2+1,0\}}$ and $h_2(D_2^Tx)=\sqrt{|x_2|}$, where $D_1=(1,1)^T$, $D_2=(0,1)^T$.

Define the smoothing function of f as

$$\tilde{f}(x,\mu)=(x_1+2x_2-1)^2+\lambda_1\sqrt{\psi(x_1+x_2+1,\mu)}+\lambda_2\sqrt{\theta(x_2,\mu)},$$

with $\psi(s,\mu)=\frac{1}{2}(s+\sqrt{s^2+4\mu^2})$ and $\theta(s,\mu)=\begin{cases}|s| & |s|>\mu,\\ \dfrac{s^2}{2\mu}+\dfrac{\mu}{2} & |s|\le\mu.\end{cases}$

Here, we use the classical projected algorithm with Armijo line search to find an approximate generalized stationary point of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu)$. There exists $\alpha>0$ such that $\bar{x}-P_{\mathcal{X}}[\bar{x}-\alpha\nabla_x\tilde{f}(\bar{x},\mu)]=0$ if and only if $\bar{x}$ is a generalized stationary point of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu)$, which is also a Clarke stationary point of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu)$ for any fixed $\mu>0$. We call x^k an approximate stationary point of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu_k)$, if there exists $\alpha_k>0$ such that $\|x^k-P_{\mathcal{X}}[x^k-\alpha_k\nabla_x\tilde{f}(x^k,\mu_k)]\|_2\le\mu_k$, which can be found in a finite number of iterations by the analysis in Bertsekas [5].

Choose initial iterate $x_0=(0,0)^T$ and let the iteration be terminated when $\mu_k\le 10^{-6}$. For different values of λ_1 and λ_2 in (49), the simulation results are listed in Table 1, where f^* indicates the optimal function value of (49). In what follows, we will show that the accumulation points in Table 1 are the generalized stationary points of (49) defined in Definition 2. Moreover, it is interesting that these points are global minimizers of (49).

λ_1	λ_2	accumulation point x^*	$\mathcal{I}_{x^*}$	$\mathcal{V}_{x^*}$	$f(x^*)$	f^*
8	2	$(-1.000,0.000)^T$	$\{1\}$	$\{v=(a,-a)^T:a\in R\}$	4.000	4.000
0.1	0.2	$(0.982,0.000)^T$	$\{2\}$	$\{v=(a,0)^T:a\in R\}$	0.141	0.141
0.5	0.1	$(-1.000,0.962)^T$	$\emptyset$	R^2	0.594	0.594

Table 1. Simulation results in Example 1

When $\lambda_1=8$ and $\lambda_2=2$, since $h_2(D_2^Tx)$ is continuously differentiable at x^*, for $v\in\mathcal{V}_{x^*}$, by $h_1(D_1^T(x^*+tv))=h_1(D_1^Tx^*)$, $\forall t>0$, we obtain

$$f^\circ(x^*;v;\mathcal{X})=\langle\nabla\Theta(x^*)+\lambda_2h_2'(D_2^Tx^*)D_2,v\rangle=-4v_1-550.473v_2,$$

where $v_1=-v_2$ by $v\in\mathcal{V}_{x^*}$, and $v_1\in R_+$ by $v\in\mathcal{T}_{\mathcal{X}}(x^*)$. Then, $f^\circ(x^*;v;\mathcal{X})\ge 0$, $\forall v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$, which means that $(-1.000,0.000)^T$ is a generalized stationary point of (49). Similarly,

- when $\lambda_1=0.1$ and $\lambda_2=0.2$:

$$f^\circ(x^*;v;\mathcal{X})=\langle\nabla\Theta(x^*)+\lambda_1h_1'(D_1^Tx^*)D_1,v\rangle=-0.036v_2,$$

where $v_2=0$ by $v\in\mathcal{T}_{\mathcal{X}}(x^*)\cap\mathcal{V}_{x^*}$;

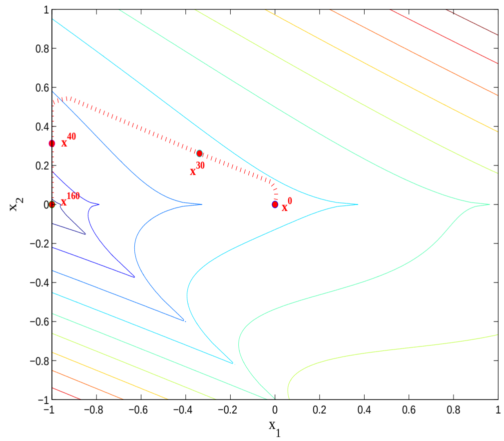

FIGURE 1. Trajectory of x^k in Example 1 with $\lambda_1 = 8$ and $\lambda_2 = 2$

- when $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$:

$$f^\circ(x^*; v; \mathcal{X}) = \langle \nabla\Theta(x^*) + \lambda_1 h_1'(D_1^T x^*)D_1 + \lambda_2 h_2'(D_2^T x^*)D_2, v\rangle = 0.102 v_1,$$

where $v_1 \in R_+$ by $v \in \mathcal{T}_{\mathcal{X}}(x^*) \cap \mathcal{V}_{x^*}$.
This gives $f^\circ(x^*; v; \mathcal{X}) \geq 0$, for all $v \in \mathcal{T}_{\mathcal{X}}(x^*) \cap \mathcal{V}_{x^*}$. Furthermore, the trajectory of x^k of the smoothing algorithm for (49) with $\lambda_1 = 8$ and $\lambda_2 = 2$ is pictured in Figure 1 with the isolines of f in $\mathcal{X}$.

From Example 1, we find that the proposed algorithm with the classical projected algorithm for finding the approximated generalized stationary point can find a global minimizer of problem (49), which is of course a generalized stationary point of it. Finding global minimizers via generalized stationary points for problm (1) is an interesting problem for further study.

3. Numerical properties In this section, we focus on the numerical properties of problem (1) with the lower bound property of its local minimizers and its computational complexity.

It is known that problem (2) with $\mathcal{X} = R^n$, $\varphi(t) := t$ and $p \in (0,1)$ enjoys lower bound property (Chen et al. [23]) but is strongly NP-hard (Chen et al. [20]). Owning to the better numerical properties of different penalty functions or regularization terms in various applications and the importance of the lower bound property in practice, we will explore a wider bound property and the computational complexity of (1) with $\mathcal{X} := \{x : Ax \leq b\}$, where $A = (A_1, \ldots, A_t)^T \in R^{t\times n}$ with $A_i \in R^n$, $i = 1, 2, \ldots, t$, and $b = (b_1, b_2, \ldots, b_t)^T \in R^t$. In this section, we suppose the function h in (1) can be represented as

$$h(x) := (h_1(d_1^T x), h_2(d_2^T x), \ldots, h_m(d_m^T x))^T, \tag{50}$$

with $d_i \in R^n$.

3.1. Bound property For $x \in \mathcal{X}$, let $\mathcal{C}_x = \{j \in \{1, 2, \ldots, t\} : A_j^T x - b_j = 0\}$ be the set of active inequality constraints at x, $\mathcal{L}_x$ be the index set such that h_i is not LC^1 at $d_i^T x$ for $i \in \mathcal{L}_x$, where we call a function $\phi : R^n \to R$ is LC^1 (or $C^{1,1}$) at x, if ϕ is continuously differentiable and its gradient is locally Lipschitz continuous at x (Hiriart-Urruty et al. [32]). For a function $\phi : R^n \to R$ of LC^1, $\partial^2\phi(x)$ is the generalized Hessian matrix of ϕ at x, which is the generalized derivative of $\nabla\phi(x)$ in Clarke's sense (Clarke [24]), i.e.

$$\partial^2\phi(x) = \text{con}\{M : \exists x^k \to x \text{ with } \phi \text{ twice differentiable at } x^k \text{ and } \nabla^2\phi(x^k) \to M\ \}.$$

Let $\mathcal{M}$ be the set of local minimizers of problem (1).

To show the local minimizers of problem (1) own the lower bound property, we need the following assumptions.

ASSUMPTION 2. *There exists $\beta \geq 0$ such that $\sup_{x\in\mathcal{M}} \|\nabla^2\Theta(x)\|_2 \leq \beta$.*

ASSUMPTION 3. *Function h in (1) has the form in (50), the points at which h_i is not LC^1 are finite and $\{\eta \in \partial^2 h_i(y)_{y=d_i^T x} : x \in \mathcal{X}\}$ is non-positive, $i = 1, 2, \ldots, m$.*

ASSUMPTION 4. *Function c is of LC^1 on its domain, $\nabla c(s)_{s=h(x)} \geq 0$, $\forall x \in \mathcal{X}$, and all elements in $\{C \in \partial^2 c(s)_{s=h(x)} : x \in \mathcal{X}\}$ is negative semi-definite.*

REMARK 4. When $\Theta(x) := \|Hx - \omega\|_2^2$ or $\Theta(x) := \log(\|Hx - \omega\|_2^2 + 1)$ with $H \in R^{s\times n}$ and $\omega \in R^s$, Assumption 2 holds naturally with $\beta = 2\|H^T H\|_2$. And the boundedness of $\mathcal{M}$ also guarantees Assumption 2.

Assumptions 3 and 4 for $c(h(x))$ are also satisfied by many models. For example, let $c(y) = \sum_{i=1}^m y_i$, $h_i(z) = \varphi(|z|^p)$ or $h_i(z) = \varphi(\max\{0, z\}^p)$, $i = 1, 2, \ldots, m$, then $c(h(x))$ with $0 < p \leq 1$ satisfies Assumptions 3 and 4 when φ is with one of the following expressions

- soft thresholding penalty function (Tibshirani [51]): $\varphi_1(s) = \lambda s$,
- logistic penalty function (Nikolova et al. [44]): $\varphi_2(s) = \lambda \log(1 + as)$,
- fraction penalty function (Nikolova et al. [44]): $\varphi_3(s) = \lambda \frac{as}{1+as}$,
- hard thresholding penalty function (Fan [26]): $\varphi_4(s) = \lambda^2 - (\lambda - s)_+^2$,
- smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li [27]):

$$\varphi_5(s) = \lambda \int_0^s \min\{1, \frac{(a - t/\lambda)_+}{a - 1}\}\mathrm{d}t,$$

- minimax concave penalty (MCP) function (Zhang [55]):

$$\varphi_6(s) = \lambda \int_0^s (1 - \frac{t}{a\lambda})_+ \mathrm{d}t,$$

with $\lambda > 0$ and $a > 0$.

THEOREM 3. *Suppose Assumptions 2-4 hold. Then, for any $a \in R$ and $i \in \{1, 2, \ldots, m\}$, if there exists $\alpha_i > 0$ such that $\nabla_i c(s)_{s=h(x)} \geq \alpha_i$, $\forall x \in \mathcal{X}$, then there exist $\theta_i > 0$ and $\kappa_i > 0$ such that any local minimizer x^* of (1) with $\mathcal{X} := \{x : Ax \leq b\}$ satisfies*
(1) $|h_i''(a+)| > \kappa_i \Longrightarrow$ either $d_i^T x^* \geq a + \theta_i$ or $d_i^T x^* \leq a$;
(2) $|h_i''(a-)| > \kappa_i \Longrightarrow$ either $d_i^T x^* \leq a - \theta_i$ or $d_i^T x^* \geq a$.

PROOF. By $\mathcal{C}_x \subseteq \{1, 2, \ldots, t\}$ and $\mathcal{L}_x \subseteq \{1, 2, \ldots, m\}$, we divide $\mathcal{M}$ into the finite disjoint sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s$, such that for each $\mathcal{M}_i$, $\mathcal{C}_x$ and $\mathcal{L}_x$ are the same for all elements $x \in \mathcal{M}_i$.

Then, $\mathcal{W}_x = \{v : d_i^T v = 0 \,\text{for}\, i \in \mathcal{L}_x \,\text{and}\, A_k^T v \leq 0 \,\text{for}\, k \in \mathcal{C}_x\}$ is the same for all elements x in each set $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s$. So, we let $\mathcal{C}^k$, $\mathcal{L}^k$ and $\mathcal{W}^k$ denote $\mathcal{C}_x$, $\mathcal{L}_x$ and $\mathcal{W}_x$ for $x \in \mathcal{M}_k$, $k = 1, 2, \ldots, s$, respectively.

For $\bar{i} \in \{1, 2, \ldots, m\}$, we first prove the statement (1) holds for any $x^* \in \mathcal{M}_1$. If $\bar{i} \in \mathcal{L}^1$, then statement (1) holds naturally by Assumption 3. Next, we consider it for $\bar{i} \notin \mathcal{L}^1$.

Let $\bar{x} \in \mathcal{M}_1$ and define $h^{\bar{x}}(x) := (h_1^{\bar{x}}(d_1^T x), h_2^{\bar{x}}(d_2^T x), \ldots, h_m^{\bar{x}}(d_m^T x))$ with

$$h_i^{\bar{x}}(d_i^T x) = \begin{cases} h_i(d_i^T x) & i \notin \mathcal{L}^1 \\ h_i(d_i^T \bar{x}) & i \in \mathcal{L}^1. \end{cases}$$

Then, there exists $\delta > 0$ such that

$$\begin{aligned} f(\bar{x}) &= \min\{\Theta(x) + c(h(x)) : \|x - \bar{x}\|_2 \leq \delta,\, Ax \leq b\} \\ &= \min\{\Theta(x) + c(h^{\bar{x}}(x)) : \|x - \bar{x}\|_2 \leq \delta,\, Ax \leq b,\, d_i^T x = d_i^T \bar{x} \text{ for } i \in \mathcal{L}^1\}, \end{aligned}$$

which implies that $\bar{x}$ is a local minimizer of the following constrained minimization problem

$$\begin{array}{ll} \min & f_{\bar{x}}(x) := \Theta(x) + c(h^{\bar{x}}(x)) \\ \text{s.t.} & Ax \leq b,\, d_i^T x = d_i^T \bar{x},\ i \in \mathcal{L}^1. \end{array} \tag{51}$$

Since $f_{\bar{x}}$ is LC^1 at $\bar{x}$, by the second order necessary optimality condition for the minimizers of (51) given by Hiriart-Urruty et al. [32, Corollary 3.1] and the finite sum rule in Clarke [24, Proposition 2.3.3], for every $v \in \mathcal{W}^1$, there exist $C^v \in \partial(\nabla c(s))_{s=h(\bar{x})}$ and $\eta_i^v \in \partial^2 h_i^{\bar{x}}(z)_{z=d_i^T\bar{x}}$ for $i \notin \mathcal{L}^1$ such that

$$v^T\nabla^2\Theta(\bar{x})v + v^T\nabla h^{\bar{x}}(\bar{x})C^v(\nabla h^{\bar{x}}(\bar{x}))^T v + \sum_{i\notin\mathcal{L}^1}\nabla_i c(s)_{s=h(\bar{x})}\eta_i^v|d_i^T v|^2 \geq 0, \tag{52}$$

where the constraint qualification can be ignored for (51) (Sun and Yuan [50, Definition 8.2.8, Corollary 8.2.9]).

By Assumptions 2-4, we obtain

$$-\alpha_{\bar{i}}\eta_{\bar{i}}^v|d_{\bar{i}}^T v|^2 \leq v^T\nabla^2\Theta(\bar{x})v \leq \beta\|v\|_2^2, \quad \forall v \in \mathcal{W}^1. \tag{53}$$

Without loss of generality, we suppose that there are infinite elements in $\mathcal{M}_1$, which means that there exists two elements in $\mathcal{M}_1$ with different values of $d_{\bar{i}}^T x$, and the same values of $d_i^T x$ for $i \in \mathcal{L}^1$, denoted as $\bar{x}$ and $\hat{x}$. If not, the result in statement (1) holds naturally.

Consider the following constrained convex optimization

$$\begin{array}{ll} \min & \|v\|_2^2 \\ \text{s.t.} & v \in \mathcal{W}_{\bar{i}}^1 = \{v : d_{\bar{i}}^T v = 1 \,\text{and}\, v \in \mathcal{W}^1\}. \end{array} \tag{54}$$

By $\frac{1}{d_{\bar{i}}^T\bar{x} - d_{\bar{i}}^T\hat{x}}(\bar{x}-\hat{x}) \in \mathcal{W}_{\bar{i}}^1$, unique existence of the optimal solution of (54) is guaranteed, denoted by $v_{\bar{i}}^1$. Let $v = v_{\bar{i}}^1$ in (53), then we have

$$-\eta_{\bar{i}}^v \leq \kappa_{\bar{i}}, \tag{55}$$

with $\kappa_{\bar{i}} = \beta\|v_{\bar{i}}^1\|_2^2/\alpha_{\bar{i}}$.

If $|h_{\bar{i}}''(a+)| > \kappa_{\bar{i}}$, by Assumption 3, which means $h_{\bar{i}}''(a+) < -\kappa_{\bar{i}}$, let

$$\theta_{\bar{i}} = \inf\{t > 0 : h_{\bar{i}}''(a+t)\,\text{exists and}\; h_{\bar{i}}''(a+t) \geq -\kappa_{\bar{i}}\}. \tag{56}$$

By the upper semicontinuity of $\partial(h_{\bar{i}}'(t))$ around $d_{\bar{i}}^T\bar{x}$ and $\eta_{\bar{i}}^v \in \partial(h_{\bar{i}}'(s))_{s=d_{\bar{i}}^T\bar{x}}$, (55) implies

$$\text{either } d_{\bar{i}}^T\bar{x} \geq a + \theta_{\bar{i}} \text{ or } d_{\bar{i}}^T\bar{x} \leq a.$$

By the randomicity of $\bar{x} \in \mathcal{M}_1$ and the invariance of $\kappa_{\bar{i}}^1$ and $\theta_{\bar{i}}^1$ for all elements in $\mathcal{M}_1$, the statement in (1) holds for the elements in $\mathcal{M}_1$. Similar analysis can be given for $\mathcal{M}_2$,...,$\mathcal{M}_s$. Therefore, we can complete the proof for statement (1).

Statement (2) can be shown by using the same arguments. □

From the proof for Theorem 3, we find that the concavity of $c(h(x))$ is the key condition to guarantee the results in Theorem 3. Based on the results in Theorem 3, we study the following special case of (1)

$$\min_{Ax\leq b} \quad f(x) := \Theta(x) + \sum_{i=1}^m \varphi(\max\{\alpha_i - d_i^T x, 0\}^p), \tag{57}$$

with $\alpha_i \in R$ and $d_i \in R^n$, $i = 1, 2, \ldots, m$.

Assumption 5. *$\varphi : R_+ \to R_+$ is continuously differentiable, non-decreasing and concave on R_{++}, and φ' is locally Lipschitz continuous on R_{++}.*

By Assumption 5, we can obtain the following lower bound results for problem (57).

COROLLARY 3. *Suppose Θ satisfies Assumption 2 and φ satisfies Assumption 5.*
(1) For $0<p<1$, if $\varphi'(0+)>0$, then there exists a constant $\theta>0$ such that any local minimizer x^* of (57) satisfies

$$\text{either } \alpha_i - d_i^T x^* \geq \theta \text{ or } \alpha_i - d_i^T x^* \leq 0,\ \forall i \in \{1,2,\ldots,m\}; \tag{58}$$

(2) For $p=1$, there exist constants $\theta>0$ and $\kappa>0$ such that if $|\varphi''(0+)|>\kappa$, then any local minimizer x^* of (57) satisfies (58).

PROOF. Define $c(y)=\sum_{i=1}^m y_i$ and $h_i(d_i^T x)=\varphi(\max\{\alpha_i - d_i^T x, 0\}^p)$, $i=1,2,\ldots,m$. It is easy to verify that functions c and h_i, $i=1,2,\ldots,m$, satisfy Assumptions 3 and 4. For $i=1,2,\ldots,m$, we have $h_i''(\alpha_i-)=-\infty$ when $0<p<1$, and $h_i''(\alpha_i-)=\varphi''(0+)$ when $p=1$, by Theorem 3 with $a=\alpha_i$, which implies the statements in this corollary. $\square$

When problem (1) reduces to the following special case:

$$\min_{Ax\leq b} \quad f(x):=\Theta(x)+\sum_{i=1}^m \varphi(|d_i^T x|^p). \tag{59}$$

Denote $c(y)=\sum_{i=1}^m y_i$ and $h_i(d_i^T x)=\varphi(|d_i^T x|^p)$, $i=1,2,\ldots,m$. Suppose φ satisfies Assumption 5, then $h_i''(0+)=h_i''(0-)=-\infty$ for (59) with $0<p<1$, and $h_i''(0+)=h_i''(0-)=\varphi''(0+)$ for (59) with $p=1$. Therefore, the results in Corollary 3 hold for problem (59) when (58) is replaced by

$$\text{either } |d_i^T x^*|\geq\theta \text{ or } |d_i^T x^*|=0,\ \forall i\in\{1,2,\ldots,m\}. \tag{60}$$

If there exists a constant $\kappa>0$ such that $|\varphi''(0+)|\geq\kappa$, by the concavity of φ and $\varphi'\geq 0$, $\varphi'(0+)>0$ holds obviously. However, the converse does not hold. So both (57) and (59) can own the lower bound property based on the weaker condition for $0<p<1$ than for $p=1$, which shows the superiority of the non-Lipschitz regularization in sparse reconstruction. For the potential functions in Remark 5, only φ_2, φ_3, φ_4 and φ_6 may meet the conditions in Corollary 3 for $p=1$, but all of them satisfy the conditions in Corollary 3 for $0<p<1$.

Moreover, the bound result in Theorem 3 can also be extended to the problem modeled by

$$\min_{Ax\leq b} \quad f(x):=\Theta(x)+\sum_{i=1}^m \varphi_i(\|D_i^T x\|_p^p), \tag{61}$$

with $0<p\leq 1$ and $D_i\in R^{n\times r}$, $i=1,2,\ldots,m$. Also with the conditions in Corollary 3, we can obtain the following estimation

$$\text{either } \|D_i^T x^*\|_p\geq\theta \text{ or } \|D_i^T x^*\|_p=0,\ \forall i\in\{1,2,\ldots,m\}. \tag{62}$$

Similar extension can also be done for

$$\min_{Ax\leq b} \quad f(x):=\Theta(x)+\sum_{i=1}^m \varphi_i(\|\max\{0, D_i^T x\}\|^p).$$

The lower bound result in Chen et al. [23] is meaningful, since it not only indicates the existence of the lower bound property for any local minimizers of the unconstrained l_2-l_p minimization with $0<p<1$, but also presents a lower bound value. Due to the generality of the considered model in (1), Theorem 3 only proves the existence of the bound property for problem (1) with general affine inequality constraints. However, following the proof of Theorem 3, we can obtain explicit values of κ_i and θ_i in Theorem 3 in the following two steps.

(i) find the points at which h_i is not LC^1;

(ii) for all possible cases of $\mathcal{W}_x$, which is finite, if $\{v : v \in \mathcal{W}_x,\, d_i^T v \neq 0\}$ is empty, find all possible points in $\mathcal{M}$ with this $\mathcal{W}_x$, otherwise, evaluate an upper bound of

$$\inf\{\|v\|_2^2/|d_i^T v|^2 : v \in \mathcal{W}_x,\, d_i^T v \neq 0\}. \tag{63}$$

So we can derive explicit bounds for some special cases. Specially, an upper bound of (63) can be obtained by finding an element in $\{v : v \in \mathcal{W}_x,\, d_i^T v \neq 0\}$.

For example, consider the following special case of problem (1)

$$\min_{x\in\mathcal{X}} \quad f(x) := \|Hx - w\|_2^2 + \sum_{i=1}^{n} \varphi(x_i^p) \tag{64}$$

with $H \in R^{s\times n}$, $\omega \in R^s$, $0 < p \le 1$ and $\mathcal{X} = \{x : x \ge 0,\ e^T x = 1\}$, where e denotes the vector whose elements are all 1 and φ is defined by one of the potential functions in Remark 5. Model (64) is often used in portfolio selections(Brodie et al. [11], Chen et al. [18]). For this example, β in Assumption 2 can be defined by $2\|H^T H\|_2$. When $\beta = 0$, by $\phi' \ge 0$ and the concave property of ϕ, it is easy to verify that the column vectors of the n-dimensional identity matrix are the solutions of (64). In what follows, we consider the case $\beta > 0$.

For $x \in \mathcal{X}$, $\mathcal{L}_x = \{i : x_i = 0\}$ and $\mathcal{W}_x$ in the proof of Theorem 3 can be expressed by

$$\mathcal{W}_x = \{v : e^T v = 0, \text{ and } v_i = 0 \,\text{for}\, x_i = 0\}.$$

Suppose $\bar{x} \in \mathcal{M}$. Denote

$$\mathcal{M}_1 = \{x \in \mathcal{M} : x \text{ has only one nonzero element}\}.$$

If $\bar{x} \in \mathcal{M}_1$, owning to the constraint $e^T\bar{x} = 1$, we get that

$$\text{either } \bar{x}_i = 0 \text{ or } \bar{x}_i = 1,\ \forall i \in \{1, 2, \ldots, n\}. \tag{65}$$

On the other hand, if $\bar{x} \notin \mathcal{M}_1$, then $\{v : v \in \mathcal{W}_{\bar{x}},\, v_i \neq 0\} \neq \emptyset$ for any $i \notin \mathcal{L}_{\bar{x}}$, and 2 is an upper bound of (63), which gives that $2\beta \ge \kappa_i$, $i = 1, 2, \ldots, n$. Thus, when $0 < p < 1$, any local minimizer x^* of (64) satisfies

$$\text{either } x_i^* = 0 \text{ or } x_i^* \ge \theta,\ \forall i \in \{1, 2, \ldots, n\}, \tag{66}$$

where $\theta = \min\{1, \inf\{t > 0 : \min_{\xi\in\partial^2\varphi(t^p)} \xi \ge -2\beta\}\}$ is a positive constant. On the other hand, when $p = 1$, then any local minimizer of (64) satisfies (66) under one of the following condition:

- $\varphi := \varphi_2$ and $\lambda a^2 > 2\beta$;
- $\varphi := \varphi_3$ and $2\lambda a^2 > 2\beta$;
- $\varphi := \varphi_4$ and $2 > 2\beta$;
- $\varphi := \varphi_6$ and $\frac{1}{a} > 2\beta$.

Specially, if we can find a continuous function $g : R_+ \to R_+$ and a positive constant $\tau \in (0, 1]$ such that $g(0+) < -2\beta$ and $g(t) \ge \min_{\xi\in\partial^2\varphi(t^p)} \xi$, $\forall t \in (0, \tau]$, then it is clear that

$$\theta \ge \min\left\{\tau, \inf\{t > 0 : g(t) \ge -2\beta\}\right\}.$$

Thus, we can give explicit values of θ in the bound property (66) for the different cases, which are shown in Table 2.

In this subsection, we prove the existence of the lower bound for local minimizers of problem (1) under some appropriate conditions, which shows the superiority of problem (1) for sparse solutions. We also show how to find an explicit value of the lower bound for problem (1) with some widely used potential functions.

φ	θ	
	$0<p<1$	$p=1$
φ_1	$\min\{(\frac{2\beta}{\lambda p(1-p)})^{\frac{1}{p-2}},1\}$	-
φ_2	$\min\{(\frac{2(1+a)^2\beta}{\lambda ap(1-p)+\lambda a^2p})^{\frac{1}{2p-2}},1\}$	$\min\{((\frac{\lambda a^2}{2\beta})^{\frac{1}{2}}-1)/a,1\}$
φ_3	$\min\{(\frac{2(1+a)^3\beta}{\lambda ap(1-p)+\lambda a^2p(1+p)})^{\frac{1}{2p-2}},1\}$	$\min\{((\frac{2\lambda a^2}{2\beta})^{\frac{1}{3}}-1)/a,1\}$
φ_4	$\min\{(\frac{2\beta}{\lambda p(1-p)+2p^2})^{\frac{1}{2p-2}},(\frac{\lambda}{2})^{\frac{1}{p}},1\}$	$\min\{\lambda,1\}$
φ_5	$\min\{(\frac{2\beta}{\lambda p(1-p)})^{\frac{1}{p-2}},\lambda^{\frac{1}{p}},1\}$	-
φ_6	$\min\{(\frac{4a\beta}{\lambda ap(1-p)+2p^2})^{\frac{1}{2p-2}},(\frac{a\lambda}{2})^{\frac{1}{p}},1\}$	$\min\{a\lambda,1\}$

TABLE 2. Explicit values of θ in the bound property (66) with $\beta=2\|H^TH\|_2$

3.2. Computational complexity In terms of the above positive result for problem (1), we will present a negative result for it in this part. From complexity theory perspective, an NP-hard optimization problem with a polynomially bounded objective function does not admit a polynomial-time algorithm, and a strongly NP-hard optimization problem with a polynomially bounded objective function does not even admit a fully-polynomial-time approximation scheme, unless P=NP (Vazirani [52]). Regarding the computational complexity, the minimization problem with l_0 norm was proved to be NP-hard by Natarajan [43]. Due to the strictness of the conditions for equivalently relaxing l_0 norm to l_1, one considered the l_p regularization with $0<p<1$ for the problem with l_0 norm. However, Ge et al. [30] showed that the l_p $(0<p<1)$ norm minimization with affine equality constraints is also strongly NP-hard, and so is its smoothed version. Then, the strong NP-hardness of unconstrained l_q-l_p problem with $q\geq 1$ and $0\leq p<1$ was shown by Chen et al. [20]. Most recently, Liu et al. [40] extended this statement to problem (3) with $\varphi(t):=t$ and a polyhedral set $\mathcal{X}$ by its special case with $\Theta(x)=0$, $\varphi(t):=t$ and $\mathcal{X}=R^n$. However, to the best of our knowledge, the computational complexity of (1), particularly modeled by (2) and (3) with $0<p\leq 1$, remains an open problem. In what follows, we will show the strong NP-hardness of problem (1) via the following special model of it

$$\min \quad \|Hx-\omega\|_2^2+\sum_{i=1}^n\varphi(|x_i|^p), \tag{67}$$

where $H\in R^{s\times n}$, $\omega\in R^s$ and $0<p\leq 1$.

LEMMA 4. *When $\varphi:R_+\to R_+$ is non-decreasing and concave on R_+, then*

$$\varphi(|s|^p)+\varphi(|t|^p)\geq\varphi(|s+t|^p),\quad \forall s,t\in R,\ 0<p\leq 1.$$

PROOF. Define $\psi(\alpha)=\varphi(\alpha+|s|^p)-\varphi(\alpha)$ on R_+. Then from the concavity of φ, $\psi'(\alpha+)=\varphi'_+((\alpha+|s|^p)+)-\varphi'_+(\alpha+)\leq 0$, which implies $\psi(|t|^p)\leq\psi(0)$. Thus, $\varphi(|t|^p+|s|^p)\leq\varphi(|t|^p)+\varphi(|s|^p)$. Since $|t+s|^p\leq|t|^p+|s|^p$ and φ is non-decreasing on R_+, we obtain $\varphi(|t+s|^p)\leq\varphi(|t|^p)+\varphi(|s|^p)$. $\square$

First, we give one preliminary result for proving the strong NP-hardness of (67) with $0<p\leq 1$.

LEMMA 5. *Let $0<p\leq 1$ and suppose $\varphi:R_+\to R_+$ is non-decreasing and concave on R_+. If ϕ, $(\phi(s):=\varphi(s^p))$, is twice continuously differentiable on $[\tau_1,\tau_2]$ with $0<\tau_1<\tau_2$, then there exists $\bar\gamma>0$ such that when $\gamma>\bar\gamma$, the minimization problem*

$$\min_{z\in R}\quad g(z)=\gamma|z-\tau_1|^2+\gamma|z-\tau_2|^2+\varphi(|z|^p) \tag{68}$$

has a unique solution $z^*\in(\tau_1,\tau_2)$.

PROOF. Since φ is twice continuously differentiable on $[\tau_1^p,\tau_2^p]$, there exists $\alpha>0$ such that $0\leq\varphi'(s)\leq\alpha$ and $-\alpha\leq\varphi''(s)\leq 0$, $\forall s\in[\tau_1^p,\tau_2^p]$.

Note that $g(z)>g(0)=\gamma\tau_1^2+\gamma\tau_2^2$ for all $z<0$, and $g(z)>g(\tau_2)=\gamma(\tau_2-\tau_1)^2+\varphi(\tau_2^p)$ for all $z>\tau_2$. Then, the minimum point of $g(z)$ must lie within $[0,\tau_2]$.

By $g(\frac{\tau_1+\tau_2}{2})=\frac{\gamma}{2}(\tau_1-\tau_2)^2+\varphi((\frac{\tau_1+\tau_2}{2})^p)$, and $g(z)\geq\gamma(\tau_2-\tau_1)^2$, $\forall z\in[0,\tau_1]$, when $\gamma>\frac{2\varphi((\frac{\tau_1+\tau_2}{2})^p)}{(\tau_2-\tau_1)^2}$,

$$g(z)>g(\frac{\tau_1+\tau_2}{2}),\ \forall z\in[0,\tau_1].$$

Thus, the minimum point of $g(z)$ must lie within $(\tau_1,\tau_2]$.

To minimize $g(z)$ on $(\tau_1,\tau_2]$, we check its first and second derivatives. If $\gamma>\frac{p\alpha\tau_1^{p-1}}{2(\tau_2-\tau_1)}$, we have $g'(\tau_1)=2\gamma(\tau_1-\tau_2)+p\varphi'(\tau_1^p)\tau_1^{p-1}<0$. By $\varphi'(\tau_2^p)\geq 0$, we get $g'(\tau_2)=2\gamma(\tau_2-\tau_1)+p\varphi'(\tau_2^p)\tau_2^{p-1}>0$. And we calculate that $g''(z)=4\gamma+p^2\varphi''(z^p)z^{2p-2}+p(p-1)\varphi'(z^p)z^{p-2}>0$ when $\gamma>\frac{p^2\alpha\tau_1^{2p-2}+p(1-p)\alpha\tau_1^{p-2}}{4}$. Thus, when

$$\gamma>\bar{\gamma}:=\max\{\frac{2\varphi((\frac{\tau_1+\tau_2}{2})^p)}{(\tau_2-\tau_1)^2},\frac{2\varphi((\frac{\tau_1+\tau_2}{2})^p)}{(\tau_2-\tau_1)^2},\frac{p\alpha\tau_1^{p-1}}{2(\tau_2-\tau_1)}\},$$

by $g'(\tau_1)<0$, $g'(\tau_2)>0$ and $g''(z)>0$, $\forall z\in(\tau_1,\tau_2)$, there exists a unique $\bar{z}\in(\tau_1,\tau_2)$ such that $g'(\bar{z})=0$, which is the unique global minimizer of $g(z)$ in R. □

THEOREM 4. *Suppose φ satisfies Assumption 5 and ϕ ($\phi(s)=\varphi(s^p)$) is strongly concave on an open interval of R_+, then minimization problem (67) is strongly NP-hard for any given $0<p\leq 1$.*

PROOF. Now we present a polynomial time reduction from the well-known strictly NP-hard partition problem (Garey and Johnson [29]) to problem (67). The 3-partition problem can be described as follows: given a multiset S of $n=3m$ integers $\{a_1,a_2,\ldots,a_n\}$ with sum mb, is there a way to partition S into m disjoint subsets $S_1,S_2,\ldots,S_m$, such that the sum of the numbers in each subset is equal?

If ϕ is strictly concave on an open interval of R_+, denoted as $(\underline{\tau},\bar{\tau})$ with $\bar{\tau}>\underline{\tau}>0$, by the locally Lipschitz continuity of φ' on R_{++}, there exist $\tau_1>0$ and $\tau_2>\tau_1$ with $[\tau_1,\tau_2]\subseteq(\underline{\tau},\bar{\tau})$ such that ϕ is strictly concave and twice continuously differentiable on $[\tau_1,\tau_2]$.

Given an instance of the partition problem with $a=(a_1,a_2,\ldots,a_n)^T\in R^n$. We consider the following minimization problem in form (67):

$$\begin{aligned}\min_x\quad P(x)=&\sum_{j=1}^m|\sum_{i=1}^n a_ix_{ij}-b|^2+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_1|^2\\&+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{i=1}^n(\sum_{j=1}^m\varphi(|x_{ij}|^p)),\end{aligned}\tag{69}$$

where parameter γ satisfies the supposition in Lemma 5.

From Lemma 4, we have

$$\begin{aligned}&\min_x P(x)\\\geq&\min_{x_{ij}}\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_1|^2+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{i=1}^n(\sum_{j=1}^m\varphi(|x_{ij}|^p))\\=&\sum_{i=1}^n\left(\min_{x_{ij}}\gamma|\sum_{j=1}^m x_{ij}-\tau_1|^2+\gamma|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{j=1}^m\varphi(|x_{ij}|^p)\right)\\\geq&\sum_{i=1}^n\min_z\gamma|z-\tau_1|^2+\gamma|z-\tau_2|^2+\varphi(|z|^p).\end{aligned}\tag{70}$$

		φ_1	φ_4		φ_5			φ_6	
$p=1$	τ_1	none	$(0,\lambda)$		$(\lambda,a\lambda)$			$(0,a\lambda)$	
	τ_2	none	(τ_1,λ)		$(\tau_1,a\lambda)$			$(\tau_1,a\lambda)$	
$0<p<1$	τ_1	$(0,\infty)$	$(0,\lambda)$	(λ,∞)	$(0,\lambda)$	$(\lambda,a\lambda)$	$(a\lambda,\infty)$	$(0,a\lambda)$	$(a\lambda,\infty)$
	τ_2	(τ_1,∞)	(τ_1,λ)	(τ_1,∞)	(τ_1,λ)	$(\tau_1,a\lambda)$	(τ_1,∞)	$(\tau_1,a\lambda)$	(τ_1,∞)

TABLE 3. Optimal parameters for different potential functions in Remark 5

To make the last inequality of (70) hold with equality, by Lemma 5, we can always choose one of x_{ij} to be $z^*(\neq 0)$ and the others are 0 for any $i=1,2,\dots,n$. Then,

$$P(x)\geq ng(z^*).$$

Now we claim that there exists an equitable partition to the partition problem if and only if the optimal value of (69) equals $ng(z^*)$. First, if S can be evenly partitioned into m sets, then we define $x_{ik}=z^*$, $x_{ij}=0$ for $j\neq k$ if a_i belongs to S_k. These x_{ij} provide an optimal solution to $P(x)$ with optimal value $ng(z^*)$. On the other hand, if the optimal value of $P(x)$ is $ng(z^*)$, by the strict concavity of ϕ on $[\tau_1,\tau_2]$, then in the optimal solution, for each i, there is only one element in $\{x_{ij}:1\leq j\leq m\}$ is nonzero. And we must also have $\sum_{i=1}^n a_i x_{ij}-b=0$ for any $1\leq j\leq m$, which implies that there exists a partition to set S into m disjoint subsets such that the sum of the numbers in each subset is equal. Thus this theorem is proved. $\square$

Theorem 4 implies the strong NP-hardness of problem (1), since (67) is a special case of it. Through the problems considered in Chen et al. [20], Ge et al. [30], Liu et al. [40] can also imply the strong NP-hardness of problem (1), they cannot imply the strong NP-hardness of (2) with $0<p\leq 1$, when φ is defined by φ_2, φ_3, φ_4, φ_5 or φ_6 in Remark 4.

REMARK 5. The conditions in Theorem 4 are satisfied by many penalty functions for $0<p\leq 1$, such as the logistic penalty function in Nikolova et al. [44], fraction penalty function in Nikolova et al. [44], hard thresholding penalty function in Fan [26], SCAD function in Fan and Li [27] and MCP function in Zhang [55], while these conditions are satisfied by the soft thresholding penalty function in Tibshirani [51], Huang et al. [33] only for $0<p<1$. For φ_2 and φ_3 in Remark 4, all choices of τ_1 and τ_2 in R_{++} with $\tau_1<\tau_2$ satisfy the conditions in Theorem 4. For the other four penalty functions in Remark 4, the optional parameters of τ_1 and τ_2 are given in Table 3.

While our paper Bian and Chen [7](2014) was under review, we became aware of an independent line of related work on computational complexity by Ge et al. Ge et al. [31](2015). Our contribution is different in that we show that the concavity of penalty functions is a key property not only for the strong NP-hardness but also for the nice lower bound theory.

4. Conclusions In Theorem 1, we derive a first order necessary optimality condition for local minimizers of problem (1) based on the new generalized directional derivative (17) and the tangent cone. The generalized stationary point that satisfies the given necessary optimality condition is a Clarke stationary point when the objective function f is locally Lipschitz continuous at this point, and satisfies the first order necessary optimality condition given or used in Bian and Chen [6], Bian et al. [9], Bian and Chen [8], Chen et al. [21], Ge et al. [30], Liu et al. [40] if f is not Lipschitz continuous at the point. Moreover, in Theorem 2 we establish the directional derivative consistency associated with smoothing functions and in Corollary 1 we show that the consistency guarantees the convergence of smoothing algorithms to a generalized stationary point of problem (1). For problem (1) with special constraints, the lower bound property of its local minimizers and its computational complexity are also studied to illustrate the positive and negative news of it with concave regularization in applications.

Acknowledgments We would like to thank Prof. Marc Teboulle, Associate Editor and two anonymous referees for their insightful and constructive comments, which help us to enrich the content and improve the presentation of the results in this paper. The work in the present paper was supported by the NSF foundation (11101107,11471088) of China, HIT.BRETIII.201414, PIRS of HIT No.A201402, and partly by Hong Kong Research Grant Council grant (PolyU5001/12P).

References

[1] Aubin JP, Cellina A (1984) *Differential Inclusion: Set-Valued Maps and Viability Theory* (Springer-Verlag, Berlin)

[2] Audet C, Dennis Jr. JE (2006) Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.* 17: 188-217

[3] Auslender A (1997) How to deal with the unbounded in optimization: theory and algorithms. *Math. Program.* 79: 3-18

[4] Beck A, Teboulle M (2012) Smoothing and first order methods: a unified framework. *SIAM J. Optim.* 22: 557-580

[5] Bertsekas DP (1976) On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Autom. Control* AC-21: 174-184

[6] Bian W, Chen X (2013) Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization. *SIAM J. Optim.* 23: 1718-1741

[7] Bian W, Chen X (2014) Optimality conditions and complexity for non-Lipschitz constrained optimization problems. *Preprint,* Department of Applied Mathematics, The Hong Kong Polytechnic University

[8] Bian W, Chen X (2015) Linearly constrained non-Lipschitz optimization for image restoration. *SIAM J. Imaging Sci.* 8: 2294-2322

[9] Bian W, Chen X, Ye Y (2015) Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.* 149: 301-327

[10] Borwein JM, Lewis AS (2000) *Convex Analysis and Nonlinear Optimization. Theory and Examples* (Springer-Verlag, New York)

[11] Brodie J, Daubechies I, De Mol C, Giannone D, Loris I (2009) Sparse and Stable Markowitz portfolios. *Proc. Nat. Acad. Sci.* 106: 12267-12272

[12] Bruckstein AM, Donoho DL, Elad M (2009) From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review* 51: 34-81

[13] Burke JV, Hoheisel T (2013) Epi-convergent smoothing with applications to convex composite functions. *SIAM J. Optim.* 23: 1457-1479

[14] Burke JV, Hoheisel T, Kanzow C (2013) Gradient consistency for integral-convolution smoothing. *Set-Valued Var. Anal.* 21: 359-376

[15] Burke JV, Lewies AS, Overton ML (2013) Approximating subdifferentials by random samplying of gradients. *Math. Oper. Res.* 27: 567-584

[16] Chan RH, Liang HX (2014) Half-quadratic algorithm for l_p-l_q problems with applications to TV-l_1 image restoration and compressive sensing. *Springer Lecture Notes in Computer Science* 8293: 78-103

[17] Chartrand R, Staneva V (2008) Restricted isometry properties and nonconvex compressive sensing. *Inverse Probl.* 24: 1-14

[18] Chen C, Li X, Tolman C, Wang S, Ye Y (2014) Sparse portfoloi selection via quasi-norm regularization. ArXiv preprint, arXiv: 1312.6350

[19] Chen X (2012) Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.* 134: 71-99

[20] Chen X, Ge D, Wang Z, Ye Y (2014) Complexity of unconstrained L_2-L_p minimization. *Math. Program.* 143: 371-383

[21] Chen X, Niu L, Yuan Y (2013) Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization. *SIAM J. Optim.* 23: 1528-1552

[22] Chen X, Ng M, Zhang C (2012) Nonconvex l_p regularization and box constrained model for image restoration. *IEEE Trans. Image Processing* 21: 4709-4721

[23] Chen X, Xu F, Ye Y (2010) Lower bound theory of nonzero entries in solutions of l_2-l_p minimization. *SIAM J. Sci. Comput.* 32: 2832-2852

[24] Clarke FH (1983) *Optimization and Nonsmooth Analysis* (John Wiley, New York)

[25] Curtis FE, Overton ML (2012) A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization. *SIAM J. Optim.* 22: 474-500

[26] Fan J (1997) Comments on 'Wavelets in statistics: a review' by A. Antoniadis. *Stat. Method. Appl.* 6: 131-138

[27] Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96: 1348-1360

[28] Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* 32: 928-961

[29] Garey MR, Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness* (W. H. Freeman & Co Ltd, New York)

[30] Ge D, Jiang X, Ye Y (2011) A note on the complexity of L_p minimization. *Math. Program.* 21: 1721-1739

[31] Ge D, Wang Z, Ye Y, Yin, H (2015) Strong NP-hardness result for regularized L_q-minimization problems with concave penalty functions. *Preprint.*

[32] Hiriart-Urruty J-B, Strodiot J-J, Nguyen VH (1984) Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. *Appl. Math. Optim.* 11: 43-56.

[33] Huang J, Horowitz JL, Ma S (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* 36: 587-613

[34] Huber P (1981) *Robust Estimation* (Wiley, New York)

[35] Huang J, Ma S, Xue H, Zhang C (2009) A group bridge approach for variable selection. *Biometrika* 96: 339-355

[36] Jahn J (1996) *Introduction to the Theory of Nonlinear Optimization* (Springer, Berlin)

[37] Knight K, Fu WJ (2000) Asymptotics for lasso-type estimators. *Ann. Stat.* 28: 1356-1378

[38] Levitin ES, Polyak BT (1966) Constrained minimization problems. *USSR. Comput. Math. Math. Phys.* 6: 1-50

[39] Liu YF, Dai YH, Ma S (2013) Joint power and admission control via linear programming deflation. *IEEE Trans. Signal Processing* 61: 1327-1338

[40] Liu YF, Ma S, Dai YH, Zhang SZ (2015) A smoothing SQP framework for a class of composite L_q minimization over polyhedron. *Math. Program.* DOI: 10.1007/s10107-015-0939-5

[41] Loh P, Wainwright MJ (2014) Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* 1: 1-56

[42] Lu Z (2014) Iterative reweighted minimization methods for l_p regularized unconstrained nonlinear programming. *Math. Program.* 147: 277-307

[43] Natarajan BK (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.* 24: 227-234

[44] Nikolova M, Ng MK, Zhang S, Ching WK (2008) Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.* 1: 2-25

[45] Nikolova M (2005) Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *Multiscale Model. Simul.* 4: 960991

[46] Nocedal J, Wright SJ (2006) *Numerical Optimization* (Springer, New York)

[47] Rockafellar RT, Wets R J-B (1998) *Variational Analysis* (Springer, Berlin)

[48] Rockafellar RT (1970) *Convex Analysis* (Princeton University Press, Princeton)

[49] Spingarn JE, Rockafellar RT (1979) The genaric nature of optimality conditions in nonlinear programming. *Math. Oper. Res.* 4: 425-430

[50] Sun WY, Yuan Y (2006) *Optimization Theory and Methods: Nonlinear Programming* (Springer, United States)

[51] Tibshirani R (1996) Shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* 58: 267-288

[52] Vazirani V (2003) *Approximation Algorithms* (Springer, Berlin)

[53] Wang Z, Liu H, Zhang T (2014) Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Stat.* 42: 2164-2201

[54] Ye Y (1997) *Interior Point Algorithms: Theory and Analysis* (John Wiley & Sons, Inc., New York)

[55] Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 38: 894-942