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 Clarke tangent cone, we derive a first order necessary optimality condition for local minimizers of the problem, and define the generalized stationary point of it. The generalized stationary point is the Clarke stationary point when the objective function is Lipschitz continuous at this point, and the scaled stationary point when the objective function is not Lipschitz continuous at this point. We prove the consistency between the generalized directional derivative and the limit of the classic directional derivatives associated with the smoothing function. Moreover we show that finding a global minimizer of such optimization problems is strongly NP-hard and establish positive lower bounds for the absolute value of nonzero entries in every local minimizer of the problem if the regularizer is concave in an open set.

Key words: Constrained nonsmooth nonconvex optimization; directional derivative consistency; optimality condition; lower bound theory; complexity
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. Problem (1) includes many problems in practice. For instance, the following minimization problem

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

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 [12], Chen et al. [17], Nikolova et al. [36]), signal processing (Bruckstein et al. [9]), variable selection (Fan and Li [22], Huang et al. [27], Huang et al. [29], Zhang [44]), etc. Another special case of (1) is the following problem

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

where $\alpha_i \in R$ and $m_i \in R^n$, which has attracted much interest in machine learning, wireless communication (Liu et al. [33]), information theory, data analysis (Fan and Peng [23], Huber [28]),

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 in Spingarn and Rockafellar [39]. When $\mathcal{X} = R^n$ and $c(h(x)) = \|x\|_p^p$ $(0 < p < 1)$, the affine scaled first and second order necessary conditions for local minimizers of (1) are established in Chen et al. [18]. By using subspace techniques, Chen et al. [16] extended the first and second order necessary conditions to $c(h(x)) = \|Dx\|_p^p$ with $D \in R^{m\times n}$. However, the optimality conditions in Chen et al. [16, 18] are weaker than the Clarke optimality conditions in Clarke [19] for $p = 1$. In this paper, we will derive a necessary optimality condition for the non-Lipschitz constrained optimization problem (1), which reduces to the Clarke optimality condition when the objective function in (1) is locally Lipschitz continuous.

A point x^* is called a Clarke stationary point of f if f is locally Lipschitz at x^* and there is $V \in \partial f(x^*)$ such that

$$(V, x - x^*) \geq 0, \quad \forall x \in X, \tag{4}$$

where

$$\partial f(x) = \mathrm{con}\{v \,|\, \nabla f(y) \to v,\ f \text{ is differentiable at } y,\ y \to x\}$$

is the Clarke subdifferential of f and "con" denotes the convex hull. From Theorem 9.61 and (b) of Corollary 8.47 in Rockafellar and Wets [38], 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 $\partial f(x) \subseteq G_{\tilde{f}}(x)$. In Burke and Hoheisel [10], Burke et al. [11], Chen [14], Rockafellar and Wets [38], it is shown that many smoothing functions satisfy the gradient consistency

$$\partial f(x) = G_{\tilde{f}}(x). \tag{5}$$

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.

Due to the non-Lipschitz continuity of the objective function f, Clarke optimality condition (4) cannot be applied to (1). In Jahn [30], Jahn introduced a directional derivative for Lipschitz constrained optimization problems

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

which is equal to the Clarke generalized directional derivative at the interior points of $\mathcal{X}$. In this paper, we extend the directional derivative in Jahn [30] to the non-Lipschitz constrained optimization problem (1). Using the extended directional derivative and the Clarke tangent cone, we derive necessary optimality conditions. The new optimality conditions are equivalent to the optimality conditions in Bian et al. [8], Chen et al. [16, 18], when the objective function is not Lipschitz continuous, and to the Clarke optimality condition (4) when the objective function is Lipschitz continuous. 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 [13], Chen [14], Chen et al. [17], Fan and Li [22], Huang et al. [27], Huang et al. [29], Loh and Wainwright [34], Lu [35], Nikolova et al. [36], Wang et al. [42], Zhang [44]). The SCAD penalty function in Fan and Li [22] and the MCP function in Zhang [44] have various desirable properties in variable selection. Logistic and fraction penalty functions yield edge preservation in image restoration (Nikolova et al. [36]). The l_p norm penalty function with $0 < p < 1$ owns the oracle property in statistics (Fan and Li [22], Knight and Fu [31]). Nonconvex regularized M-estimator is proved to have the statistical accuracy and prediction error estimation in Loh and Wainwright [34]. Moreover, the lower bound theory of the l_2-l_p regularized minimization problem in Chen et al. [17, 18], 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 [13], Huang et al. [27]), but also brings the restored image closed contours and neat edges (Chen et al. [17]). In this paper, we extend the lower bound theory of the l_2-l_p regularization minimization problem to problems (2) and (3) with $0 < p \leq 1$ which include the most widely used models in statistics and sparse reconstruction. Moreover, we extend the complexity results of the l_2-l_p regularization minimization problem in Chen et al. [15] to problem (2) with a concave function φ and $0 < p \leq 1$. We show that the concavity of penalty functions is a key property for both the lower bound theory and the strong NP hardness. Such extension of the lower bound theory and complexity is not trivial because of the general constraints and weak conditions on φ.

The rest of the paper is organized as follows. In section 2, we first define a generalized directional derivative and present its properties. Next, we derive necessary optimality conditions for a local minimizer of problem (1), and prove the directional derivative consistency associated with smoothing functions. In section 3, we present the computational complexity and the lower bound theory of problem (2).

In our notation, $R_+ = [0, \infty)$ and $R_{++} = (0, \infty)$. 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 closed convex subset $\Omega \subseteq R^n$, $\mathrm{int}(\Omega)$ means the interior of Ω, $\mathrm{cl}(\Omega)$ means the closure of Ω and $\mathrm{m}(\Omega)$ denotes the element in Ω with the smallest Euclidean norm. $P_{\mathcal{X}}[x] = \arg\min\{\|z - x\|_2 : z \in \mathcal{X}\}$ denotes the orthogonal projection from R^n to $\mathcal{X}$. $\mathbb{N}_{++} = \{1, 2, \ldots\}$.

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 is equivalent to the Clarke necessary condition for locally Lipschitz optimization problems and stronger than the necessary optimality conditions for the non-Lipschitz optimization problems in the existing literature. At the end of this section, we prove the directional directive consistency associated with smoothing functions

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

$$h(x) := (h_1(D_1^T x), h_2(D_2^T x), \ldots, h_m(D_m^T x))^T \tag{6}$$

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. 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

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

Otherwise, ϕ is said to be not Lipschitz continuous at(near) $x \in R^n$.

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

$$\mathcal{I}_{\bar{x}} = \{i \in \{1,2,\ldots,m\} : h_i \text{ is not Lipschitz continuous at } D_i^T \bar{x}\}, \tag{7}$$

$$\mathcal{V}_{\bar{x}} = \{v : D_i^T v = 0,\ i \in \mathcal{I}_{\bar{x}}\}, \tag{8}$$

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 $D_i^T \bar{x}$, $i = 1,2,\ldots,m$. Specially, we let $\mathcal{V}_{\bar{x}} = R^n$ when $\mathcal{I}_{\bar{x}} = \emptyset$. And then we let

$$f_{\bar{x}}(x) = \Theta(x) + c(h_{\bar{x}}(x)), \tag{9}$$

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

The function $f_{\bar{x}}(x)$ is Lipschitz continuous at $\bar{x}$ and $f_{\bar{x}}(\bar{x}) = f(\bar{x})$. The generalized directional derivative in Clarke [19] of $f_{\bar{x}}$ at $\bar{x}$ in the direction $v \in R^n$ is defined as

$$f_{\bar{x}}^{\circ}(\bar{x}; v) = \limsup_{y \to \bar{x}, t \downarrow 0} \frac{f_{\bar{x}}(y + tv) - f_{\bar{x}}(y)}{t}. \tag{10}$$

Specially, when f is regular,

$$f_{\bar{x}}^{\circ}(\bar{x}; v) = f_{\bar{x}}'(\bar{x}; v) = \lim_{t \downarrow 0} \frac{f_{\bar{x}}(\bar{x} + tv) - f_{\bar{x}}(\bar{x})}{t}.$$

The generalized directional derivative in (10) is generalized in Jahn [30] and used in Audet and Dennis [2], Jahn [30] for locally Lipschitz constrained optimization. The generalization motives us to use the following generalized directional derivative of $f_{\bar{x}}$ at $\bar{x} \in \mathcal{X}$ in the direction $v \in R^n$

$$f_{\bar{x}}^{\circ}(\bar{x}; v) = \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ t \downarrow 0,\, y + tv \in \mathcal{X}}} \frac{f_{\bar{x}}(y + tv) - f_{\bar{x}}(y)}{t}. \tag{11}$$

The definitions in (10) and (11) coincide when $\bar{x} \in \operatorname{int}(\mathcal{X})$.

PROPOSITION 1. *For any $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{V}_{\bar{x}}$,*

$$f^{\circ}(\bar{x}; v) = \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ t \downarrow 0,\, y + tv \in \mathcal{X}}} \frac{f(y + tv) - f(y)}{t} \quad \textit{exists} \tag{12}$$

and equals to $f_{\bar{x}}^{\circ}(\bar{x}; v)$ defined in (11).

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 $h_i^{\bar{x}}$ 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 (12) exists.

Let $\{y_n\}$ and $\{t_n\}$ be the sequences such that $y_n \in \mathcal{X}$, $t_n \downarrow 0$, $y_n \to \bar{x}$, $y_n + t_n v \in \mathcal{X}$ and the upper limit in (12) holds. Using the Lipschitz continuity of $h_i^{\bar{x}}$ at $\bar{x}$ again, we can get the subsequences $\{y_{n_k}\} \subseteq \{y_n\}$ and $\{t_{n_k}\} \subseteq \{t_n\}$ such that

$$\lim_{k\to\infty} \frac{h_{\bar{x}}(y_{n_k} + t_{n_k} v) - h_{\bar{x}}(y_{n_k})}{t_{n_k}} \quad \text{exists.} \tag{13}$$

By the above analysis, then

$$\begin{aligned} f^\circ(\bar{x}; v) &= \lim_{k\to\infty} \frac{f(y_{n_k} + t_{n_k} v) - f(y_{n_k})}{t_{n_k}} \\ &= \nabla\Theta(\bar{x}) + \nabla c(z)_{z=h(\bar{x})} \lim_{k\to\infty} \frac{h_{\bar{x}}(y_{n_k} + t_{n_k} v) - h_{\bar{x}}(y_{n_k})}{t_{n_k}}. \end{aligned} \tag{14}$$

By virtue of (11), we have

$$\begin{aligned} f_{\bar{x}}^\circ(\bar{x}; v) &\geq \lim_{k\to\infty} \frac{f_{\bar{x}}(y_{n_k} + t_{n_k} v) - f_{\bar{x}}(y_{n_k})}{t_{n_k}} \\ &= \nabla\Theta(\bar{x}) + \nabla c(z)_{z=h_{\bar{x}}(\bar{x})} \lim_{k\to\infty} \frac{h_{\bar{x}}(y_{n_k} + t_{n_k} v) - h_{\bar{x}}(y_{n_k})}{t_{n_k}}. \end{aligned} \tag{15}$$

Using $h(\bar{x}) = h_{\bar{x}}(\bar{x})$, (14) and (15), we obtain $f_{\bar{x}}^\circ(\bar{x}; v) \geq f^\circ(\bar{x}; v)$.

On the other hand, by extracting the sequences $\{y_{n_k}\}$ and $\{t_{n_k}\}$ such that the upper limit in (11) holds and the limit in (13) exists with them, similar to the above analysis, we find that $f^\circ(\bar{x}; v) \geq f_{\bar{x}}^\circ(\bar{x}; v)$.

Therefore, $f^\circ(\bar{x}; v) = f_{\bar{x}}^\circ(\bar{x}; v)$. □

Notice that the generalized directional derivative of f at $\bar{x} \in \mathcal{X}$ in the direction $v \in \mathcal{V}_{\bar{x}}$ defined in (12) involves only the behavior of f at $\bar{x}$ in the hyperplane $\mathcal{V}_{\bar{x}}$.

2.2. Clarke tangent cone Since $\mathcal{X}$ is a nonempty closed convex subset of R^n, the distance function related to $\mathcal{X}$ is a nonsmooth, Lipschitz continuous function, defined by

$$d_{\mathcal{X}}(x) = \min\{\|x - y\|_2 : y \in \mathcal{X}\}.$$

The Clarke tangent cone to $\mathcal{X}$ at $x \in \mathcal{X}$, denoted as $\mathcal{T}_{\mathcal{X}}(x)$, is defined by

$$\mathcal{T}_{\mathcal{X}}(x) = \{v \in R^n : d_{\mathcal{X}}^\circ(x; v) = 0\}.$$

ASSUMPTION 1. *Assume that $\mathcal{X} = \mathcal{X}_1 \cap \mathcal{X}_2$ and $\operatorname{int}(\mathcal{X}_1) \cap \mathcal{X}_2 \neq \emptyset$, where $\mathcal{X}_1 \subseteq R^n$ is a nonempty closed convex set and $\mathcal{X}_2 = \{x \mid Ax = b\}$ with $A \in R^{t\times n}, b \in R^t$.*

Under Assumption 1, we can obtain the following properties of the Clarke tangent cones to $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{X}$.

LEMMA 1. *The following statements hold.*
(1) $\operatorname{int}(\mathcal{T}_{\mathcal{X}_1}(x)) \neq \emptyset$, $\forall x \in \mathcal{X}_1$;
(2) $\mathcal{T}_{\mathcal{X}_2}(x) = \operatorname{cl}\{\lambda(c - x) : c \in \mathcal{X}_2, \lambda \geq 0\}$, $\forall x \in \mathcal{X}_2$;
(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) Fix $x \in \mathcal{X}_1$ and denote $\hat{x} \in \operatorname{int}(\mathcal{X}_1)$. Let $\epsilon > 0$ be a constant such that $\hat{x} + B_\epsilon(0) \subseteq \operatorname{int}(\mathcal{X}_1)$. We shall show that $\hat{x} - x + B_\epsilon(0) \subseteq \mathcal{T}_{\mathcal{X}_1}(x)$, and hence $\hat{x} - x \in \operatorname{int}(\mathcal{T}_{\mathcal{X}_1}(x))$.

By the convexity of $\mathcal{X}_1$, $d_{\mathcal{X}_1}(x)$ is a convex function and for $v \in \hat{x} - x + B_\epsilon(0)$, we notice that

$$x + tv \in (1 - t)x + t(\hat{x} + B_\epsilon(0)) \subseteq \mathcal{X}_1, \quad \forall x \in \mathcal{X}_1,\ 0 \leq t \leq 1.$$

Then,

$$d^{\circ}_{\mathcal{X}_1}(x;v) = d'_{\mathcal{X}_1}(x;v) = \lim_{\lambda \downarrow 0} \frac{d_{\mathcal{X}_1}(x+\lambda v) - d_{\mathcal{X}_1}(x)}{\lambda} = 0,$$

which confirms that $v \in \mathcal{T}_{\mathcal{X}_1}(x)$.

(2) Since $\mathcal{X}_2$ is defined by a class of affine equalities, we have $\mathcal{T}_{\mathcal{X}_2}(x) = \mathrm{cl}\{\lambda(c-x) : c \in \mathcal{X}_2, \lambda \geq 0\}$.

(3) By $\mathrm{int}(\mathcal{X}_1) \cap \mathcal{X}_2 \neq \emptyset$, $0 \in \mathrm{int}(\mathcal{X}_1 - \mathcal{X}_2)$, then $\mathcal{T}_{\mathcal{X}_1 \cap \mathcal{X}_2}(x) = \mathcal{T}_{\mathcal{X}_1}(x) \cap \mathcal{T}_{\mathcal{X}_2}(x)$ (Aubin and Cellina [1, pp.141]). □

Since $\mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(x)) \neq \emptyset$, for a vector $v \in \mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(x))$, there exists a scalar $\epsilon > 0$ such that

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

We often call $\mathrm{int}(\mathcal{T}_{\mathcal{X}_1}(x))$ the hypertangent cone to $\mathcal{X}_1$ at x.

And by Lemma 1 (2), we have $x + tv \in \mathcal{X}_2$, $\forall x \in \mathcal{X}_2$, $t \geq 0$, $v \in \mathcal{T}_{\mathcal{X}_2}(x)$.

2.3. Necessary optimality condition Denote

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

Since f is not assumed to be locally Lipschitz continuous, the calculus theory developed in Audet and Dennis [2] cannot be directly applied to f. The next lemma extends calculus results for the unconstrained case in Clarke [19] and the constrained case in Audet and Dennis [2].

For any $x \in \mathcal{X}$, from $0 \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(x)) \cap \mathcal{V}_x$, we know $\text{r-int}(\mathcal{T}_{\mathcal{X}}(x)) \cap \mathcal{V}_x \neq \emptyset$.

LEMMA 2. *For $\bar{x} \in \mathcal{X}$ and $v \in \mathcal{T}_{\mathcal{X}}(\bar{x}) \cap \mathcal{V}_{\bar{x}}$,*

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

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

$$\|h_{\bar{x}}(x) - h_{\bar{x}}(y)\|_2 \leq L_{\bar{x}} \|x - y\|_2, \quad \forall x, y \in B_\epsilon(\bar{x}). \tag{16}$$

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 $\{w_k\} \subseteq \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x}))$, let $\epsilon_k > 0$ be such that $x + tw_k \in \mathcal{X}_1$ whenever $x \in \mathcal{X} \cap B_{\epsilon_k}(\bar{x})$ and $0 \leq t < \epsilon_k$. By Lemma 1 (2), we obtain $x + tv \in \mathcal{X}_2, x + tw_k \in \mathcal{X}_2$, $\forall t \geq 0$, $x \in \mathcal{X}$. Then, for all w_k, it gives

$$\begin{aligned} f^{\circ}(\bar{x};v) &= \limsup_{\substack{x \to \bar{x}, x \in \mathcal{X} \\ t \downarrow 0, x + tv \in \mathcal{X}}} \frac{f(x+tv) - f(x)}{t} \\ &= \limsup_{\substack{x \to \bar{x}, x \in \mathcal{X} \\ t \downarrow 0, x + tv \in \mathcal{X} \\ x + tw_k \in \mathcal{X}}} \frac{f(x+tv) - f(x)}{t} \\ &= \limsup_{\substack{x \to \bar{x}, x \in \mathcal{X} \\ t \downarrow 0, x + tv \in \mathcal{X} \\ x + tw_k \in \mathcal{X}}} \frac{f(x+tw_k) - f(x)}{t} + \frac{f(x+tv) - f(x+tw_k)}{t}. \end{aligned} \tag{17}$$

Let $\delta > 0$ be such that $x + tw_k \in B_\epsilon(\bar{x})$ for any $x \in B_\delta(\bar{x})$, $0 \leq t < \delta$ and $k \in \mathbb{N}_{++}$. By the Lipschitz condition in (16), we have

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

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 $x\in B_\delta(\bar{x}), 0\le t<\delta$, we have

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

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

Thus, (17) implies

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

As k goes to infinity, the above inequality follows $f^\circ(\bar{x};v)=\lim_{k\to\infty}f^\circ(\bar{x};w_k)$. 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. □

Note that the above lemma is not necessarily true when r-int$(\mathcal{T}_\mathcal{X}(x))$ is empty. A similar example can be given following the idea in Audet and Dennis [2, Example 3.10]. That is why we put the assumption $\text{int}(\mathcal{X}_1)\cap\mathcal{X}_2\neq\emptyset$ at the beginning of this section. Based on Lemmas 1-2, the following theorem gives the main theoretical result of this section.

THEOREM 1. *If x^* is a local minimizer of (1), then $f^\circ(x^*,v)\ge 0$ 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$(\mathcal{T}_\mathcal{X}(x^*))\cap\mathcal{V}_{x^*}$.

There exist $\epsilon>0$ and $L_{x^*}>0$ such that $f(x^*)\le f(x)$, and

$$\|h_{x^*}(x)-h_{x^*}(y)\|_2\le L_{x^*}\|x-y\|_2,\ \forall x,y\in\mathcal{X}\cap B_\epsilon(x^*). \tag{18}$$

Since $w\in\text{int}(\mathcal{T}_{\mathcal{X}_1}(x^*))$, there exists $\bar{\epsilon}\in(0,\epsilon]$ such that

$$x+tw\in\mathcal{X}_1, \forall x\in\mathcal{X}_1\cap B_{\bar{\epsilon}}(x^*), 0\le t\le\bar{\epsilon}.$$

By Lemma 1 (2), $x+tw\in\mathcal{X}$, $\forall x\in\mathcal{X}\cap B_{\bar{\epsilon}}(x^*)$, $0\le t<\bar{\epsilon}$. And 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 (18), for all $x\in B_{t^2}(x^*)\cap\mathcal{X}$, $0<t<\delta$, we obtain

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

Thus,

$$\lim_{\substack{x\in B_{t^2}(x^*)\cap\mathcal{X}\\ x+tw\in\mathcal{X}, t\downarrow 0}}\frac{h_{x^*}(x+tw)-h_{x^*}(x^*+tw)}{t}-\frac{h_{x^*}(x)-h_{x^*}(x^*)}{t}=0. \tag{19}$$

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

$$\begin{aligned} &|\frac{c(h(x+tw))-c(h(x))}{t}-\frac{c(h(x^*+tw))-c(h(x^*))}{t}|\\ =&|\frac{\nabla c(z_1)^T(h(x+tw)-h(x))}{t}-\frac{\nabla c(z_2)^T(h(x^*+tw)-h(x^*))}{t}|\\ =&|\frac{\nabla c(z_1)^T(h_{x^*}(x+tw)-h_{x^*}(x))}{t}-\frac{\nabla c(z_2)^T(h_{x^*}(x^*+tw)-h_{x^*}(x^*))}{t}|. \end{aligned} \tag{20}$$

By (19), (20) and the continuous differentiability of Θ, we have

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

Thus,

$$\limsup_{\substack{x\to x^*,\,x\in\mathcal{X}\\ t\downarrow 0,\,x+tw\in\mathcal{X}}}\Big[\frac{f(x+tw)-f(x)}{t}-\frac{f(x^*+tw)-f(x^*)}{t}\Big]\geq 0. \tag{21}$$

By $f(x^*+tw)-f(x^*)\geq 0$ for $0\leq t<\bar{\epsilon}$, (21) implies

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

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

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

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

It is worth noting that a generalized stationary point x^* is a Clarke stationary point of problem (1) when f is Lipschitz continuous at x^*.

REMARK 1. Suppose $h_i(D_i^Tx)$ is regular in $\mathcal{X}\backslash\mathcal{N}_i$, where

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

For $\bar{x}\in\mathcal{X}$, the regularity assumption allows us to define $\mathcal{V}_{\bar{x}}$ by

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

which is a bigger set than $\mathcal{V}_{\bar{x}}$ given in (8). Hence a generalized stationary point defined in Definition 2 can be more robust with this $\mathcal{V}_{\bar{x}}$. For example, if f is defined as in (3), $\mathcal{I}_{\bar{x}}=\{i\in\{1,2,\ldots,m\}:m_i^T\bar{x}=\alpha_i\}$ and we can let

$$\mathcal{V}_{\bar{x}}=\{v:m_i^Tv\geq 0,\ \forall\, i\in\mathcal{I}_{\bar{x}}\},$$

which includes $\{v:m_i^Tv=0,\ \forall i\in\mathcal{I}_{\bar{x}}\}$ as a proper subset.

We notice that a generalized stationary point defined in Definition 2 is a scaled stationary point defined in Bian and Chen [6], Bian et al. [8], Chen et al. [16], Ge et al. [25] for the special cases of (2) with $0<p<1$. Moreover it is stronger than a scaled stationary point for the Lipschitz case, since it is a Clarke stationary point but a scaled stationary point is not necessarily a Clarke stationary point for the Lipschitz optimization problem.

2.4. Directional derivative consistency In this subsection, we show that the generalized directional derivative of f defined in (12) can be represented by the limit of a sequence of directional derivatives of a smoothing function of f. This property is important for development of numerical algorithms for nonconvex non-Lipschitz constrained optimization problems.

DEFINITION 3. (Chen [14]) Let $g:R^n\to R$ be a continuous function. We call $\tilde{g}:R^n\times[0,\infty)\to R$ a smoothing function of g, if $\tilde{g}(\cdot,\mu)$ is continuously differentiable for any fixed $\mu>0$ and $\lim_{z\to x,\mu\downarrow 0}\tilde{g}(z,\mu)=g(x)$ holds for any $x\in R^n$.

Let $\tilde{h}(x,\mu)=(\tilde{h}_1(D_1^Tx,\mu),\tilde{h}_2(D_2^Tx,\mu),\ldots,\tilde{h}_m(D_m^Tx,\mu))^T$, where $\tilde{h}_i$ is a smoothing function of h_i in (6). Then $\tilde{f}(x,\mu):=\Theta(x)+c(\tilde{h}(x,\mu))$ is a smoothing function of f.

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

$$\tilde{f}^\circ(x,\mu;v)=\limsup_{\substack{y\to x,\,y\in\mathcal{X}\\ t\downarrow 0,\,y+tv\in\mathcal{X}}}\frac{\tilde{f}(y+tv,\mu)-\tilde{f}(y)}{t}=\langle\nabla_x\tilde{f}(x,\mu),v\rangle. \tag{22}$$

THEOREM 2. *Suppose h_i is continuously differentiable in $\mathcal{X}\backslash\mathcal{N}_i$, $\forall i\in\{1,2,\ldots,m\}$, where $\mathcal{N}_i=\{x:\ h_i$ is not Lipschitz continuous at $D_i^Tx\}$, then*

$$\lim_{\substack{x_k\in\mathcal{X},\\ x_k\to x,\,\mu_k\downarrow 0}}\langle\nabla_x\tilde{f}(x_k,\mu_k),v\rangle=f^\circ(x;v),\ \forall v\in\mathcal{V}_x. \tag{23}$$

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 $w\in\mathcal{V}_{\bar{x}}$, by the closed form of $\nabla_x\tilde{f}(x_k,\mu_k)$, we have

$$\begin{aligned}&\langle\nabla_x\tilde{f}(x_k,\mu_k),w\rangle\\ =&\langle\nabla\Theta(x_k),w\rangle+\langle\nabla_x\tilde{h}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},w\rangle\\ =&\langle\nabla\Theta(x_k),w\rangle+\langle\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},\nabla_x\tilde{h}(x_k,\mu_k)^Tw\rangle,\end{aligned} \tag{24}$$

where

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

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

Define

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

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

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

Thus, coming back to (24), we obtain

$$\begin{aligned}\langle\nabla_x\tilde{f}(x^k,\mu_k),w\rangle=&\langle\nabla\Theta(x_k),w\rangle+\langle\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},\nabla_x\tilde{h}_{\bar{x}}(x_k,\mu_k)^Tw\rangle\\ =&\langle\nabla\Theta(x_k),w\rangle+\langle\nabla_x\tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},w\rangle\\ =&\langle\nabla\Theta(x_k)+\nabla_x\tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},w\rangle.\end{aligned} \tag{25}$$

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

$$\begin{aligned}&\lim_{k\to\infty}\nabla\Theta(x_k)+\nabla_x\tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)}\\ =&\nabla\Theta(\bar{x})+\nabla h_{\bar{x}}(\bar{x})\nabla c(z)_{z=h(\bar{x})}=\nabla f_{\bar{x}}(\bar{x}),\end{aligned} \tag{26}$$

where $f_{\bar{x}}$ is defined in (9).

Thus,

$$\begin{aligned}f^\circ(\bar{x},w)=f^\circ_{\bar{x}}(\bar{x},w)&=\langle\nabla f_{\bar{x}}(\bar{x}),w\rangle\\ &=\langle\lim_{k\to\infty}\nabla\Theta(x_k)+\nabla_x\tilde{h}_{\bar{x}}(x_k,\mu_k)\nabla c(z)_{z=\tilde{h}(x_k,\mu_k)},w\rangle\\ &=\lim_{k\to\infty}\langle\nabla_x\tilde{f}(x^k,\mu_k),w\rangle,\end{aligned} \tag{27}$$

where the first equation uses Proposition 1, the third uses (26) and the fourth uses (25). □

Now we give another consistency result on 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

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

By $\mathcal{I}_{x_k} \subseteq \{1, 2, \ldots, m\}$, 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 a contradiction. Therefore, the statement in this lemma holds. □

Based on the consistency results given in Theorem 2 and Lemma 3, the next corollary shows the generalized stationary point consistency of the smoothing functions.

COROLLARY 1. *Let $\{\epsilon_k\}$ and $\{\mu_k\}$ be positive sequences converging to 0. With the conditions on h in Theorem 2, if x^k satisfies $\langle \nabla_x \tilde{f}(x^k, \mu_k), v \rangle \geq -\epsilon_k$ for every $v \in \mathcal{T}_{\mathcal{X}}(x^k) \cap \mathcal{V}_{x^k} \cap B_1(0)$, then any accumulation point of $\{x^k\} \subseteq \mathcal{X}$ is a generalized stationary point of (1).*

PROOF. Let $\bar{x}$ be an accumulation point of $\{x^k\}$. Without loss of generality, we suppose $\lim_{k\to\infty} x_k = \bar{x}$.

For $w \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x})) \cap \mathcal{V}_{\bar{x}} \cap B_1(0)$, 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

$$x + sw \in \mathcal{X}, \quad \forall x \in \mathcal{X} \cap B_\epsilon(\bar{x}),\ 0 \leq s \leq \epsilon. \tag{28}$$

Since x_k is converging to $\bar{x}$, there exists $K \in \mathbb{N}_{++}$ such that $x_k \in \mathcal{X} \cap B_\epsilon(\bar{x})$, $\forall k \geq K$. By (28), we have $x_k + sw \in \mathcal{X}$, $\forall k \geq K$, $0 \leq s \leq \epsilon$. From the convexity of $\mathcal{X}$, we obtain $w \in \mathcal{T}_{\mathcal{X}}(x_k)$.

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

$$\begin{aligned} f^\circ(\bar{x}; \rho v) = & \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ t \downarrow 0,\, y + t\rho v \in \mathcal{X}}} \frac{f(y + t\rho v) - f(y)}{t} \\ = & \rho \limsup_{\substack{y \to \bar{x},\, y \in \mathcal{X} \\ s \downarrow 0,\, y + sv \in \mathcal{X}}} \frac{f(y + sv) - f(y)}{s} = \rho f^\circ(\bar{x}; v) \geq 0. \end{aligned} \tag{29}$$

Thus, $f^\circ(\bar{x}; v) \geq 0$ for every $v \in \text{r-int}(\mathcal{T}_{\mathcal{X}}(\bar{x})) \cap \mathcal{V}_{\bar{x}} \cap B_1(0)$ implies $f^\circ(\bar{x}; v) \geq 0$ 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 $f^\circ(\bar{x}, v) \geq 0$ 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). □

REMARK 2. Suppose the gradient consistency associated with the smoothing function $\tilde{h}_i$ holds at its Lipschitz continuous points, that is

$$\{\lim_{z \to x, \mu \downarrow 0} \nabla_x \tilde{h}_i(D_i^T x, \mu)\} \subseteq \partial h_i(D_i^T x), \quad \forall x \in \mathcal{X},\ i \notin \mathcal{I}_x, \tag{30}$$

then

$$\{\lim_{z\to x,\mu\downarrow 0}\nabla\Theta(z)+\nabla\tilde{h}_x(x,\mu)\nabla c(z)_{z=h_x(x)}\}\subseteq\partial f_x(x),\ \forall x\in\mathcal{X}. \tag{31}$$

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

$$f^\circ_{\bar{x}}(\bar{x},v)=\max\{\langle\xi,v\rangle:\xi\in\partial f_{\bar{x}}(\bar{x})\}. \tag{32}$$

Similar to the calculation in (27), by (31) and (32), we obtain

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

Thus, the conclusion in Corollary 1 can be true with (30), which is weaker than the strict differentiability of h_i in $\mathcal{X}\backslash\mathcal{N}_i$, $i\in\{1,2,\ldots,m\}$. Some conditions can be found in Clarke [19] to ensure (30). Specially, when the function h in f is with the form

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

with $d_i\in R^n$, by Clarke [19, Theorem 2.3.9 (i)], the regularity of $h_i(d_i^Tx)$ in $\mathcal{X}\backslash\mathcal{N}_i$ is a sufficient condition for the statement in Theorem 2.

Corollary 1 shows that one can find a generalized stationary point of (1) by using the approximate first order optimality condition of $\min_{x\in\mathcal{X}}\tilde{f}(x,\mu)$. Since $\tilde{f}(x,\mu)$ 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 [20], Levitin and Polyak [32], Nocedal and Wright [37], Ye [43]). We use one example to show the validity of the first order necessary optimality condition and the consistency result given in this section.

Example 1. Consider the following minimization problem

$$\begin{aligned} &\min\ f(x):=(x_1+2x_2-1)^2+\lambda_1\sqrt{\max\{x_1+x_2+1,0\}}+\lambda_2\sqrt{|x_2|},\\ &\text{s.t.}\ \ x\in\mathcal{X}=\{x\in R^2:-1\leq x_1,x_2\leq 1\}. \end{aligned} \tag{33}$$

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

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 $\|x^k-P_{\mathcal{X}}[x^k-\alpha_k\nabla_x\tilde{f}(x^k,\mu_k)]\|_2\leq\alpha_k\mu_k$, which can be found in a finite number of iterations by the analysis in Bertsekas [5].

Choose the initial iterate $x_0=(0,0)^T$. For different values of λ_1 and λ_2 in (33), the simulation results are listed in Table 1, where f^* indicates the optimal function value of (33), where the iteration is terminated when $\mu_k\leq 10^{-6}$.

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

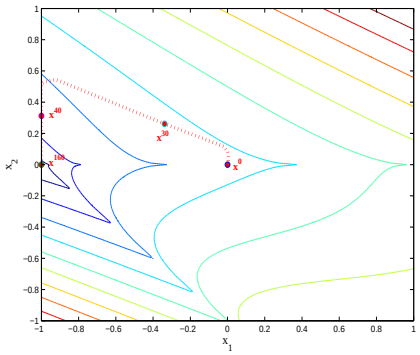

FIGURE 1. Trajectory of x^k in Example 1 with $\lambda_1 = 8$ and $\lambda_2 = 2$

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

$$\begin{aligned} f^\circ(x^*; v) = & \limsup_{\substack{y \to x^*, y \in \mathcal{X} \\ t \downarrow 0, y + tv \in \mathcal{X}}} \frac{\Theta(y+tv) - \Theta(y) + \lambda_2 h_2(D_2^T(y+tv)) - \lambda_2 h_2(D_2^T y)}{t} \\ = & \langle \nabla \Theta(x^*) + \lambda_2 h_2'(D_2^T x^*) D_2, v \rangle = -4v_1 - 550.473 v_2, \end{aligned}$$

where $v_1 = -v_2$ by $v \in \mathcal{V}_{x^*}$, and $v_1 \in R_+$ by $x_1^* = -1.000$ and the condition $x^* + tv \in \mathcal{X}$ in $f^\circ(x^*; v)$. Then, $f^\circ(x^*; v) \geq 0$, $\forall v \in \mathcal{V}_{x^*}$, which means that $(-1.000, 0.000)^T$ is a generalized stationary point of (33). Similarly,

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

$$f^\circ(x^*; v) = \langle \nabla \Theta(x^*) + \lambda_1 h_1'(D_1^T x^*) D_1, v \rangle = -0.036 v_2,$$

where $v_2 = 0$ by $v \in \mathcal{V}_{x^*}$;

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

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

where $v_1 \in R_+$ by $x_1^* = -1.000$.

This gives $f^\circ(x^*; v) \geq 0$, for all $v \in \mathcal{V}_{x^*}$. Thus, the accumulation points in Table 1 are generalized stationary points of (33) with different values of λ_1 and λ_2. Furthermore, the trajectory of x^k of the smoothing algorithm for (33) with $\lambda_1 = 8$, $\lambda_2 = 2$ are pictured in Figure 1 with the isolines of f in $\mathcal{X}$.

3. Nonconvex regularization In this section, we focus on problem (2) with the function φ satisfying the following assumption.

ASSUMPTION 2. *Assume that $\varphi : R_+ \to R_+$ with $\varphi(0) = 0$ is continuously differentiable, non-decreasing and concave on $(0, \infty)$, and φ' is locally Lipschitz continuous on R_{++}.*

The function $\varphi(t)=t$ satisfies Assumption 2. It is known that problem (2) with $\mathcal{X}=R^n$, $\varphi(t)=t$ and $p\in(0,1)$ is strongly NP hard but enjoys lower bound theory. However, the complexity and lower bound theory of problem (2) with a general convex set $\mathcal{X}$ and the class of functions φ satisfying Assumption 2 have not been studied. In this section, we show that the key condition for the complexity and lower bound theory is that the function $\varphi(z^p)$ is strictly concave in an open interval.

3.1. Computational complexity In this subsection, we will show the strong NP-hardness of the following problem

$$\min \quad \|Hx-c\|_2^2+\sum_{i=1}^{n}\varphi(|x_i|^p), \tag{34}$$

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

LEMMA 4. *$\varphi(|s|^p)+\varphi(|t|^p)\ge\varphi(|s+t|^p)$, $\forall s,t\in R$.*

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

First, we give two preliminary results for proving the strong NP-hardness of (34) with $0<p\le 1$. The first is for $p=1$ and the second is for $0<p<1$.

LEMMA 5. *Suppose φ is strictly concave and twice continuously differentiable on $[\tau_1,\tau_2]$ with $\tau_1>0$ and $\tau_2>\tau_1$. There exists $\bar\gamma>0$ such that when $\gamma>\bar\gamma$ and $p=1$, the minimization problem*

$$\min_{z\in R} \quad g(z)=\gamma|z-\tau_1|^2+\gamma|z-\tau_2|^2+\varphi(|z|^p), \tag{35}$$

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

PROOF. Since φ is twice continuously differentiable in $[\tau_1,\tau_2]$, there exists $\alpha>0$ such that $0\le\varphi'(s)\le\alpha$ and $-\alpha\le\varphi''(s)\le 0$, $\forall s\in[\tau_1,\tau_2]$. Let $\bar\gamma=\max\{\frac{\alpha}{2(\tau_2-\tau_1)},\frac{\alpha}{4}\}$ and suppose $\gamma>\bar\gamma$.

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

To minimize $g(z)$ on $[0,\tau_2]$, we check its first derivative

$$g'(z)=2\gamma(z-\tau_1)+2\gamma(z-\tau_2)+\varphi'(z),\quad 0<z\le\tau_2.$$

When $0<z\le\tau_1$, $g'(z)=4\gamma z-2\gamma\tau_1-2\gamma\tau_2+\varphi'(z)\le 2\gamma\tau_1-2\gamma\tau_2+\alpha<0$, which means that $g(z)$ is strictly decreasing on $[0,\tau_1]$. Therefore, the minimum point of $g(z)$ must lie within $(\tau_1,\tau_2]$.

Consider solving $g'(z)=2\gamma(z-\tau_1)+2\gamma(z-\tau_2)+\varphi'(z)=0$ on $(\tau_1,\tau_2]$. Calculate $g''(z)=4\gamma+\varphi''(z)>0$. And we have $g'(\tau_2)=2\gamma\tau_2-2\gamma\tau_1+\varphi'(\tau_2)>0$, $g'(\tau_1)<0$. Therefore, there exists a unique $\bar z\in(\tau_1,\tau_2)$ such that $g'(\bar z)=0$, which is the unique global minimum point of $g(z)$ in R. □

For the case that $0<p<1$, we need a weaker condition on φ to obtain a similar result as in Lemma 5.

LEMMA 6. *Suppose φ is twice continuously differentiable on $[\tau_1^p,\tau_2^p]$ with $\tau_2>\tau_1>0$. There exists $\bar\gamma>0$ such that when $\gamma>\bar\gamma$ and $0<p<1$, the minimization problem (35) has a unique solution $z^*\in(\tau_1,\tau_2)$.*

PROOF. First, there exists $\alpha>0$ such that $0\le\varphi'(s)\le\alpha$ and $-\alpha\le\varphi''(s)\le 0$, $\forall s\in[\tau_1^p,\tau_2^p]$. Let $\gamma>\bar\gamma$, where

$$\bar\gamma=\max\{\frac{2\varphi((\frac{\tau_1+\tau_2}{2})^p)}{(\tau_2-\tau_1)^2},\frac{p\alpha\tau_1^{p-1}}{2(\tau_2-\tau_1)},\frac{\alpha\tau_1^{2p-2}+\alpha\tau_1^{p-2}}{4}\}.$$

Similar to the analysis in Lemma 5, the minimum point of $g(z)$ must lie within $[0,\tau_2]$. When $z\in[0,\tau_1]$, $g(z)\geq\gamma(\tau_2-\tau_1)^2$, then by $\gamma>\frac{2\varphi((\frac{\tau_1+\tau_2}{2})^p)}{(\tau_2-\tau_1)^2}$, we have

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

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

To minimize $g(z)$ on $(\tau_1,\tau_2]$, we check its first derivative. By $\gamma>\frac{p\alpha\tau_1^{p-1}}{2(\tau_2-\tau_1)}$, we have $g'(\tau_1)=2\gamma(\tau_1-\tau_2)+p\varphi'(\tau_1^p)\tau_1^{p-1}<0$, and by $\varphi'\geq 0$, we get $g'(\tau_2)=2\gamma(\tau_2-\tau_1)+p\varphi'(\tau_2^p)\tau_2^{p-1}>0$. Now we consider the solution of the constrained equation

$$g'(z)=2\gamma(z-\tau_1)+2\gamma(z-\tau_2)+p\varphi'(z^p)z^{p-1}=0,\quad z\in(\tau_1,\tau_2].$$

We calculate that $g''(z)=4\gamma+p^2\varphi''(z^p)z^{2p-2}+p(p-1)\varphi'(z^p)z^{p-2}>0$ since $\gamma>\frac{\alpha\tau_1^{2p-2}+\alpha\tau_1^{p-2}}{4}$. Combining it with $g'(\tau_1)<0$ and $g'(\tau_2)>0$, 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. □

Since φ' is locally Lipschitz continuous in R_{++}, φ' is continuously differentiable almost everywhere in R_{++}. If φ is strictly concave in $(\underline{\tau},\bar{\tau})$ with $\bar{\tau}>\underline{\tau}>0$, there exist $\tau_1>0$ and $\tau_2>\tau_1$ with $[\tau_1^p,\tau_2^p]\subseteq(\underline{\tau},\bar{\tau})$ such that φ is strictly concave and twice continuously differentiable on $[\tau_1^p,\tau_2^p]$. Thus, the strict concavity of φ in an open interval of R_+ is sufficient for the existence of $[\tau_1^p,\tau_2^p]$ with $\tau_2>\tau_1>0$ such that φ is strictly concave and twice continuously differentiable on it. And there is no other condition needed to guarantee the supposition of φ in Lemma 6.

THEOREM 3. *1. Minimization problem (34) is strongly NP-hard for any given $0<p<1$.*
2. If φ is strongly concave in an open interval of R_+, then minimization problem (34) is strongly NP-hard for $p=1$.

PROOF. Now we present a polynomial time reduction from the well-known strongly NP-hard partition problem (Garey and Johnson [24]) to problem (34). 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?

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 (34):

$$\begin{aligned}\min_x\quad P(x)=&\sum_{j=1}^m|\sum_{i=1}^n\alpha_i x_{ij}-\beta|^2+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_1|^2\\&+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{i=1}^n(\sum_{j=1}^m\varphi(|x_{ij}|^p)),\end{aligned}\tag{36}$$

where the parameters τ_1, τ_2 and γ satisfy the suppositions in Lemma 5 for $p=1$ and them in Lemma 6 for $0<p<1$.

From Lemma 4, we have

$$\begin{aligned}&\min_x P(x)\\&\geq\min_{x_{ij}}\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_1|^2+\gamma\sum_{i=1}^n|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{i=1}^n(\sum_{j=1}^m\varphi(|x_{ij}|^p))\\&=\sum_{i=1}^n\left(\min_{x_{ij}}\gamma|\sum_{j=1}^m x_{ij}-\tau_1|^2+\gamma|\sum_{j=1}^m x_{ij}-\tau_2|^2+\sum_{j=1}^m\varphi(|x_{ij}|^p)\right)\\&\geq\sum_{i=1}^n\min_z\gamma|z-\tau_1|^2+\gamma|z-\tau_2|^2+\varphi(|z|^p).\end{aligned}\tag{37}$$

		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 2. Parameters for different potential functions in Remark 3

By Lemmas 5-6 and the strict concavity of $\varphi(z^p)$ on $[\tau_1,\tau_2]$, we can always choose one of x_{ij} to be $z^*(\neq 0)$ and the others are 0 for any $i=1,2,\ldots,n$ such that the last inequality in (37) becomes to be an equality and

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

Now we claim that there exists an equitable partition to the partition problem if and only if the optimal value of (36) equals to $ng(z^*)$. 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 $ng(z^*)$, 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 $\sum_{i=1}^n \alpha_i x_{ij}-\beta=0$ holds 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. □

REMARK 3. Many penalty functions satisfy the conditions in Lemma 5 and Lemma 6, such as the logistic penalty function in Nikolova et al. [36], fraction penalty function in Nikolova et al. [36], hard thresholding penalty function in Fan [21], SCAD function in Fan and Li [22] and MCP function in Zhang [44]. The soft thresholding penalty function in Huang et al. [27], Tibshirani [40] only satisfies the conditions in Lemma 6. Here, we list the formulations of these penalty functions below. For φ_2 and φ_3, all choices of τ_1 and τ_2 in R_{++} with $\tau_1<\tau_2$ satisfy the conditions in Lemma 5 and Lemma 6. For the other four penalty functions, the optional parameters of τ_1 and τ_2 are given in Table 2.

- soft thresholding penalty function: $\varphi_1(s)=\lambda s$,
- logistic penalty function : $\varphi_2(s)=\lambda\log(1+as)$,
- fraction penalty function: $\varphi_3(s)=\lambda\frac{as}{1+as}$,
- hard thresholding penalty function: $\varphi_4(s)=\lambda^2-(\lambda-s)_+^2$,
- smoothly clipped absolute deviation (SCAD) penalty function:

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

- minimax concave penalty (MCP) function:

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

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

3.2. Lower bound theory In this subsection, we will establish the lower bound theory for local minimizers of (2) with a special constraint, that is

$$\begin{aligned}\min\quad & f(x):=\Theta(x)+\sum_{i=1}^m \varphi(\|D_i^Tx\|_p^p)\\ \text{s.t.}\quad & x\in\mathcal{X}=\{x:Ax\leq b\},\end{aligned}\tag{38}$$

where $D_i = (D_{i1}, \ldots, D_{ir})$ with $D_{ij} \in R^n$, $j = 1, 2, \ldots, r$, $A = (A_1, \ldots, A_q)^T \in R^{q\times n}$ with $A_i \in R^n$, $i = 1, 2, \ldots, q$, and $b = (b_1, b_2, \ldots, b_q)^T \in R^q$.

Denote $\mathcal{M}$ the set of all local minimizers of (38). In this subsection, we suppose that there exists $\beta > 0$ such that $\sup_{x\in\mathcal{M}} \|\nabla^2\Theta(x)\|_2 \le \beta$.

For $x \in \mathcal{X}$, let $\mathcal{I}_{ac}(x) = \{i \in \{1, 2, \ldots, q\} : A_i^T x - b_i = 0\}$ be the set of active inequality constraints at x.

THEOREM 4. *Let $p = 1$ in (38). There exist constants $\theta > 0$ and $\nu_1 > 0$ such that if $|\varphi''(0+)| > \nu_1$, then any local minimizer x^* of (38) satisfies*

$$\text{either } \|D_i^T x^*\|_1 = 0 \text{ or } \|D_i^T x^*\|_1 \ge \theta, \ \forall i \in \{1, 2, \ldots, m\}.$$

PROOF. We divide $\mathcal{M}$ into the finite disjoint sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s$ such that all element x in each set have the same following values:
(i) sign values $\text{sign}(D_{it}^T x)$ for $i = 1, 2, \ldots, m$, $t = 1, 2, \ldots, r$;
(ii) index values $\mathcal{I}_{ac}(x)$ and $\mathcal{I}_x = \{i \in \{1, 2, \ldots, m\} : D_i^T x = 0\}$.

First, we will prove that there exist $\theta_{1,1} > 0$ and $\kappa_{1,1}$ such that

$$\text{either } \|D_1^T x\|_1 = 0 \text{ or } \|D_1^T x\|_1 \ge \theta_{1,1}, \quad \forall x \in \mathcal{M}_1, \tag{39}$$

when $|\varphi''(0+)| > \beta\kappa_{1,1}$.

Specially, if the values of $\|D_1^T x\|_1$ are same for all $x \in \mathcal{M}_1$, then the statement in (39) holds naturally. In what follows, we suppose that there are at least two elements in $\mathcal{M}_1$ with different values of $\|D_1^T x\|_1$.

Suppose $\bar{x} \in \mathcal{M}_1$ is a local minimizer of minimization problem (38) satisfying $\|D_1^T \bar{x}\| \ne 0$. Then, there exists $\delta > 0$ such that

$$\begin{aligned} f(\bar{x}) &= \min\{\Theta(x) + \sum_{i=1}^m \varphi(\|D_i^T x\|_1) : \|x - \bar{x}\|_2 \le \delta, \ Ax \le b\} \\ &= \min\{\Theta(x) + \sum_{i\notin\mathcal{I}_{\bar{x}}} \varphi(\|D_i^T x\|_1) : \|x - \bar{x}\|_2 \le \delta, \ Ax \le b, \ D_i^T x = 0 \text{ for } i \in \mathcal{I}_{\bar{x}}\}, \end{aligned}$$

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

$$\begin{aligned} \min \quad & f_{\bar{x}}(x) := \Theta(x) + \sum_{i\notin\mathcal{I}_{\bar{x}}} \varphi(\|D_i^T x\|_1) \\ \text{s.t.} \quad & Ax \le b, \ D_i^T x = 0, \ i \in \mathcal{I}_{\bar{x}}. \end{aligned} \tag{40}$$

Since φ' is locally Lipschitz continuous in R_{++}, by the second order optimality necessary condition, there exists $\xi_i \in \partial(\varphi'(s))_{s=\|D_i^T\bar{x}\|_1}$ such that

$$v^T\nabla^2\Theta(\bar{x})v + \sum_{i\notin\mathcal{I}_{\bar{x}}} \xi_i \left(\sum_{t\in\{1,2,\ldots,r\}} \text{sign}(D_{it}^T\bar{x}) D_{it}^T v \right)^2 \ge 0, \ \forall v \in \mathcal{V}_{\bar{x}}, \tag{41}$$

where

$$\mathcal{V}_{\bar{x}} = \{v : D_j^T v = 0 \text{ for } j \in \mathcal{I}_{\bar{x}} \text{ and } A_k^T v = 0 \text{ for } k \in \mathcal{I}_{ac}(\bar{x})\}. \tag{42}$$

By $\xi_i \le 0$, $i = 1, 2, \ldots, m$, (41) gives

$$-\xi_1 \left(\sum_{t\in\{1,2,\ldots,r\}} \text{sign}(D_{1t}^T\bar{x}) D_{1t}^T v \right)^2 \le \|\nabla^2\Theta(\bar{x})\|_2 \|v\|_2^2, \quad \forall v \in \mathcal{V}_{\bar{x}}. \tag{43}$$

Fix $x \in \mathcal{M}_1$. For $c \in R^r$ with $c_i \in \{-1, 0, 1\}$, consider the following constrained convex minimization problem

$$\begin{array}{ll} \min & \|v\|_2^2 \\ \text{s.t.} & v \in \mathcal{V}_{x,c} = \{v : \sum\limits_{t\in\{1,\dots,r\}} c_t D_{1t}^T v = 1 \text{ and } v \in \mathcal{V}_x\}. \end{array} \tag{44}$$

When $\mathcal{V}_{x,c} \neq \emptyset$, unique existence of the optimal solution of (44) is guaranteed, denoted by $v_{x,c}$. Take all possible choices of nonzero vectors $c \in R^r$ with $c_i \in \{-1, 0, 1\}$ such that $\mathcal{V}_{x,c} \neq \emptyset$, which are finite, and we define

$$\kappa_{1,1} = \max \|v_{x,c}\|_2^2,$$

which is a positive number and same for all elements in $\mathcal{M}_1$ from the decomposition method for $\mathcal{M}$.

Since there is another element in $\mathcal{M}_1$, denoted as $\hat{x}$, such that $\|D_1^T \bar{x}\|_1 \neq \|D_1^T \hat{x}\|_1$, then

$$\tilde{v} = \frac{1}{\|D_1^T \bar{x}\|_1 - \|D_1^T \hat{x}\|_1}(\bar{x} - \hat{x}) \in \mathcal{V}_{\bar{x},c}.$$

Thus, the unique solution of (44) exists in this case and (43) holds with it, which follows

$$-\xi_1 \leq \beta \kappa_{1,1}.$$

If $|\varphi''(0+)| > \beta\kappa_{1,1}$, let

$$\theta_{1,1} = \inf\{t > 0 : \varphi''(t) \text{ exists and } \varphi''(t) \geq \beta\kappa_{1,1}\}, \tag{45}$$

by the upper semicontinuity of $\partial(\varphi'(t)))$ on R_{++}, we obtain that

$$\|D_1^T \bar{x}\|_1 \geq \theta_{1,1}.$$

By the randomicity of $\bar{x} \in \mathcal{M}_1$ satisfying $\|D_1^T \bar{x}\|_1 \neq 0$ in the above analysis, (43) implies

$$-\xi_1 \left(\sum_{t\in\{1,2,\dots,r\}} \operatorname{sign}(D_{1t}^T x) D_{1t}^T v \right)^2 \leq \|\nabla^2 \Theta(x)\|_2 \|v\|_2^2, \quad \forall v \in \mathcal{V}_x,$$

holds for any $x \in \mathcal{M}_1$ satisfying $\|D_1^T x\|_1 \neq 0$. Since $\kappa_{1,1}$ is same for all elements in $\mathcal{M}_1$, the statement in (39) holds.

Similarly, for any $i = 1, \dots, m$, $j = 1, \dots, s$, there exist $\theta_{i,j} > 0$ and $\kappa_{i,j} > 0$ such that

$$\text{either } \|D_i^T x\|_1 = 0 \text{ or } \|D_i^T x\|_1 \geq \theta_{i,j}, \quad \forall x \in \mathcal{M}_j,$$

when $|\varphi''(0+)| > \beta\kappa_{i,j}$.

Therefore, we can complete the proof for this theorem with $\nu_1 = \max\{\beta\kappa_{i,j} : i = 1, \dots, m, j = 1, \dots, s\}$ and $\theta = \min\{\theta_{i,j} : i = 1, \dots, m, j = 1, \dots, s\}$. □

If there exists constant $\nu_1 > 0$ such that $|\varphi''(0+)| \geq \nu_1$, by the concavity of φ and $\varphi' \geq 0$, there must exist $\nu_p > 0$ such that $\varphi'(0+) \geq \nu_p$. However, the converse does not hold. The following theorem presents the lower bound theory for the case that $0 < p < 1$ using the existence of $\nu_p > 0$ such that $\varphi'(0+) \geq \nu_p$.

THEOREM 5. *Let $0 < p < 1$ in (38). If there exists $\nu_p > 0$ such that $\varphi'(0+) \geq \nu_p$, then there exists a constant $\theta > 0$ such that any local minimizer x^* of (38) satisfies*

$$\textit{either } \|D_i^T \bar{x}\|_p = 0 \textit{ or } \|D_i^T \bar{x}\|_p \geq \theta, \ \forall i \in \{1, 2, \dots, m\}.$$

PROOF. We divide $\mathcal{M}$ by the method in Theorem 4 and we will also prove that there exists $\theta_{1,1}>0$ such that

$$\text{either } \|D_1^T x\|_p = 0 \text{ or } \|D_1^T x\|_p \geq \theta_{1,1}, \quad \forall x \in \mathcal{M}_1. \tag{46}$$

Specially, if the values of $\|D_1^T x\|_p$ are same for all $x \in \mathcal{M}_1$, then the statement in (46) holds naturally. In what follows, we also suppose that there are at least two elements in $\mathcal{M}_1$ with different values of $\|D_1^T x\|_p$.

Similar to the analysis in Theorem 4, $\bar{x}$ is a local minimizer of minimization problem (38) satisfying $\|D_1^T \bar{x}\|_p \neq 0$ implies that $\bar{x}$ is a local minimizer of the minimization problem

$$\begin{array}{ll} \min & f_{\bar{x}}(x) := \Theta(x) + \sum\limits_{i \notin \mathcal{I}_{\bar{x}}} \varphi(\|D_i^T x\|_p^p) \\ \text{s.t.} & Ax \leq b,\ D_i^T x = 0,\ i \in \mathcal{I}_{\bar{x}}. \end{array} \tag{47}$$

By the second order optimality necessary condition for the minimizers of (47), there exists $\xi_i \in \partial(\varphi'(s))_{s=\|D_i^T \bar{x}\|_p^p}$ such that

$$\begin{aligned} & v^T \nabla^2 \Theta(\bar{x}) v + \sum_{i \notin \mathcal{I}_{\bar{x}}} \xi_i \left(\sum_{t \in T_i} p |D_{it}^T \bar{x}|^{p-1} \text{sign}(D_{it}^T \bar{x}) D_{it}^T v \right)^2 \\ & + \sum_{i \notin \mathcal{I}_{\bar{x}}} p(p-1) \varphi'(s)_{s=\|D_i^T \bar{x}\|_p^p} \left(\sum_{t \in T_i} |D_{it}^T \bar{x}|^{p-2} (D_{it}^T v)^2 \right) \geq 0,\ \forall v \in \mathcal{V}_{\bar{x}}, \end{aligned}$$

where $\mathcal{V}_{\bar{x}}$ is same as in (42) and $T_i = \{t \in \{1,2,\ldots,r\} : D_{it}^T \bar{x} \neq 0\}$, $i = 1,2,\ldots,m$. Then, by $\xi_i \leq 0$, $\forall i = 1,2,\ldots,m$ and $\|D_1^T \bar{x}\|_p \neq 0$, we obtain

$$p(1-p) \varphi'(s)_{s=\|D_1^T \bar{x}\|_p^p} \left(\sum_{t \in T_i} |D_{1t}^T \bar{x}|^{p-2} (D_{1t}^T v)^2 \right) \leq v^T \nabla^2 \Theta(\bar{x}) v, \forall v \in \mathcal{V}_{\bar{x}}. \tag{48}$$

Fix $x \in \mathcal{M}_1$. For $t \in T_1$, consider the following constrained convex optimization

$$\begin{array}{ll} \min & \|v\|_2^2 \\ \text{s.t.} & v \in \mathcal{V}_{x,t} = \{v : D_{1t}^T v = 1 \text{ and } v \in \mathcal{V}_x\}. \end{array} \tag{49}$$

When $\mathcal{V}_{x,t} \neq \emptyset$, unique existence of the optimal solution of (49) is guaranteed, denoted by $v_{x,t}$. Take all possible choices of $t \in T_1$ such that $\mathcal{V}_{x,c} \neq \emptyset$, which are finite, and we define

$$\kappa_{1,1} = \max \|v_{x,t}\|_2^2,$$

which is also a positive number same for all elements in $\mathcal{M}_1$.

Since there is another element in $\mathcal{M}_1$, denoted as $\hat{x}$, such that $\|D_1^T \bar{x}\|_p \neq \|D_1^T \hat{x}\|_p$. Then, there exists $t_1 \in \{1,2,\ldots,r\}$ such that $D_{1t_1}^T \bar{x} \neq D_{1t_1}^T \hat{x}$. Thus,

$$\tilde{v} = \frac{1}{D_{1t_1}^T \bar{x} - D_{1t_1}^T \hat{x}} (\bar{x} - \hat{x}) \in \mathcal{V}_{\bar{x},t_1},$$

which follows the existence of the unique solution of (49) exists with $t = t_1$, denoted as $v_{\bar{x},t_1}^*$.

By the decomposition method for $\mathcal{M}$, we have $\text{sign}(D_{1t}^T \bar{x}) = \text{sign}(D_{1t}^T \hat{x})$, which implies $t_1 \in T_1$. Let $v = v_{\bar{x},t_1}^*$ in (48), by $\varphi' \geq 0$, we have

$$p(1-p) \varphi'(s)_{s=\|D_1^T \bar{x}\|_p^p} |D_{1t_1}^T \bar{x}|^{p-2} \leq \beta \kappa_{1,1}. \tag{50}$$

$|D^T_{1t_1}\bar{x}| \le \|D^T_1\bar{x}\|_p$ implies $|D^T_{1t_1}\bar{x}|^{p-2} \ge \|D^T_1\bar{x}\|^{p-2}_p$, then (50) gives

$$p(1-p)\varphi'(s)_{s=\|D^T_1\bar{x}\|^p_p}\|D^T_1\bar{x}\|^{p-2}_p \le \beta\kappa_{1,1}. \tag{51}$$

By the concavity of φ, $\lim_{t\to\infty}\varphi'(s)_{s=t^p}t^{p-2} \le \lim_{t\to\infty}\varphi'(1)t^{p-2} = 0$. From $\varphi'(0+) \ge \nu_2$, $\lim_{t\downarrow 0}\varphi'(t^p)t^{p-2} = +\infty$. Let

$$\theta_{1,1} = \inf\{t>0 : \varphi'(t^p)t^{p-2} = \frac{\beta\kappa_{1,1}}{p(1-p)}\},$$

which is an existent number larger than 0. Therefore, (51) implies

$$\|D^T_1\bar{x}\|_p \ge \theta_{1,1}.$$

Similar to the analysis in Theorem 4, the statement in this theorem holds. □

REMARK 4. For the other cases, such as the regularization term is given by $\sum_{i=1}^m \varphi_i(\max\{d_i^T x, 0\}^p)$ with $d_i \in R^n$, the lower bound theory in Theorems 4-5 can also be guaranteed under the same conditions. Moreover, the lower bound theories in Theorems 4-5 can also be extended to the more general case with the objective function $f(x) := \Theta(x) + \sum_{i=1}^m \varphi_i(\|D_i^T x\|_p^p)$.

All the potential functions in Remark 3 satisfy the conditions in Theorem 5, but only φ_2, φ_3, φ_4 and φ_6 may meet the conditions in Theorem 4 under some conditions on the parameters, which shows the superiority of the non-Lipschitz regularization in sparse reconstruction.

While our paper [7](2014) was under review, we became aware of an independent line of related work on computational complexity by Ge et al. [26](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 (12) and the Clarke tangent cone. The generalized stationary point that satisfies the first order necessary optimality condition is a Clarke stationary point when the objective function f is locally Lipschitz continuous near this point, and a scaled stationary point if f is non-Lipschitz 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 stationary point of problem (1). Computational complexity and lower bound theory of problem (1) are also studied to illustrate the negative and positive news of the concave penalty function in applications.

Acknowledgments. 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, Ye Y (2015) Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.* 149: 301-327

[9] 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

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

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

[12] Chan RH and 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

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

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

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

[16] 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

[17] 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

[18] 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

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

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

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

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

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

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

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

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

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

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

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

[30] Jahn J (1996) *Introduction to the Nonlinear Optimization* (Springer-Verlag, Berlin)

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

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

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

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

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

[36] 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

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

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

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

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

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

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

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

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