Noname manuscript No.
(will be inserted by the editor)

Optimality Conditions for Locally Lipschitz Optimization with l_0-Regularization

Hui Zhang[1] · **Lili Pan**[2] · **Naihua Xiu**[1]

Received: date / Accepted: date

Abstract This paper mainly investigates the **l**ocally **L**ipschitz **o**ptimization **p**roblem (LLOP) with l_0-regularization in a finite dimensional space, which is generally NP-hard but highly applicable in statistics, compressed sensing and deep learning. First, we introduce two classes of stationary points for this problem: subdifferential-stationary point and proximal-stationary point. Secondly, based on these two concepts, we analyze the first-order necessary/sufficient optimality conditions for the LLOP with l_0-regularization. Finally, we present two examples to illustrate the validity of the proposed optimality conditions.

Keywords locally Lipschitz optimization · l_0-regularization · subdifferential-stationary point · proximal-stationary point · application

1 Introduction

We consider the following **l**ocally **L**ipschitz **o**ptimization **p**roblem (LLOP) with l_0-regularization:

$$\min_{x\in\mathbb{R}^n} \quad f(x) + \lambda\|x\|_0, \tag{1}$$

Hui Zhang
18118011@bjtu.edu.cn

Lili Pan
panlili1979@163.com

Naihua Xiu
nhxiu@bjtu.edu.cn

1. Department of Mathematics, School of Science, Beijing Jiaotong University, Beijing 100044, People's Republic of China

2. Department of Mathematics, Shandong University of Technology, Zibo 255049, People's Republic of China

where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a locally Lipschitz continuous function (not necessarily smooth nor convex) and $\lambda > 0$ is a regularization parameter. $\|x\|_0$ counts the number of nonzero elements in the vector x and is a lower semicontinuous function. Problem (1) has wide applications in linear and nonlinear compressed sensing [5, 6, 13, 14, 16, 25], robust linear regression [23], censored regression [4, 10, 30, 33], and deep learning [1, 15, 18, 20, 21, 26, 34, 35].

In general, due to the combinatorial nature of the l_0-regularization term, finding a global minimizer of sparse optimization problems is generally NP-hard ([28]), and continuous optimization theory is usually not applicable to combinatorial optimization problems. However, based on the special structure of the regular term, there are some interesting research topics. Particularly, people have explored the optimality conditions for some special cases of Problem (1).

Some related studies in the literature state the optimality conditions of Problem (1) when f is continuously differentiable. Blumensath et al. [6, 7] discussed Problem (1) for the specific case where $f(x) = \|Ax - b\|_2^2$, and presented an optimality condition based on the concept of fixed points. They then designed an effective algorithm to solve the problem. Furthermore, Lu et al. [24, 25] extended the concept of fixed points to the case when f is generally differentiable, and proposed a first-order optimality condition (KKT condition) for these problems using subspace technique under the Robinson constraint qualification from variational analysis. More recently, Beck et al. [3] introduced and analyzed three kinds of first-order necessary optimality conditions for the existence of solutions when f is generally differentiable: L-stationarity, support optimality, and partial coordinate-wise optimality.

Much less is known about Problem (1) when f is nonsmooth. Bian [4] studied a special case of Problem (1) where f is a convex function through an exact continuous relaxation problem with local minimizers that are the same as those of original problem under some mild assumptions. For a relatively more general case, Guo et al., in a recent work [17] established a first-order necessary condition (KKT condition) by extending the standard quasi-normality when the object function is a locally Lipschitz continuous function f, in addition to an extended valued lower semicontinuous function under equality and inequality constraints.

Optimality conditions play an important theoretical role in the research on optimization problems. Motivated by the observations above, this paper analyzes optimality conditions for Problem (1). The contributions of this paper are summarized as follow: (i) We present two kinds of stationary points for Problem (1) in the sense of a subdifferential and a proximal operator, and state the relationships between each stationary points of Problem (1) [in Section 3]. (ii) We propose the first-order sufficient and necessary optimality conditions for Problem (1) [in Section 3]. (iii) Two examples are given to illustrate the effectiveness of the proposed optimality conditions [in Section 4].

Our notation is standard. The set of all n-dimensional positive vectors is denoted by $\mathbb{R}^n_{++}$ and $\overline{\mathbb{R}} := (-\infty, \infty]$. For $x = (x_1, ..., x_n)^\top \in \mathbb{R}^n$, $\|x\|_0 = |\Gamma_x|$ with $\Gamma_x := \{i \in \{1, 2, ..., n\} \mid x_i \neq 0\}$, and index set $\overline{\Gamma}_x$ is the complement of index set Γ_x. Moreover, the non-negative part of x is denoted by $(x)_+ := \max\{x, 0\}$, and $\|x\|_p := (\sum_{i=1}^{n} |x_i|^p)^{1/p}$ for any $p > 0$. For $t \in \mathbb{R}$, $sign$ is the sign operator:

$$sign(t) = \begin{cases} 1, & \text{if } t > 0, \\ [-1, 1], & \text{if } t = 0, \\ -1, & \text{if } t < 0. \end{cases}$$

Furthermore, I_m is a unit matrix, and $U(x, \delta)$ is a neighborhood around x with constant $\delta(> 0)$. For $X = (x_{i,j}) \in \mathbb{R}^{m\times n}$, $\|X\|_F := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2}$. For any set $\Omega \subseteq \mathbb{R}^n$, the convex hull of Ω in $\mathbb{R}^n$ is denoted by $co\Omega$.

2 Technical results

In this section, for the sake of convenience, some definitions and results are given [11, 27, 31, 32].

To present the stationary point in the form of subdifferential, we need to introduce the definition of a subdifferential. Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with a point $\bar{x} \in \mathbb{R}^n$ at which $f(\bar{x})$ is finite. According to [31, Definition 8.3] and [11, 12], for a vector $v \in \mathbb{R}^n$, we say:

(i) v is a regular subgradient of f at $\bar{x}$, written $v \in \hat{\partial} f(\bar{x})$, if

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x}\rangle + o(\|x - \bar{x}\|);$$

(ii) v is a limiting (or Mordukhovich) subgradient of f at $\bar{x}$, written $v \in \partial f(\bar{x})$, if there are sequences $x^\nu \to_f \bar{x}$ and $v^\nu \in \hat{\partial} f(x^\nu)$ with $v^\nu \to v$;

(iii) v is a horizon subgradient of f at $\bar{x}$, written $v \in \partial^\infty f(\bar{x})$, if the same holds as in (ii), except that instead of $v^\nu \to v$, we have $\lambda^\nu v^\nu \to v$ for some sequence $\lambda^\nu \to 0$, or $v^\nu \to dir v$(or $v = 0$);

(iv)([11, Corollary on page 97], [12, line 12-14 on page 5]) v is a Clarke subgradient of f at $\bar{x}$, written $v \in \partial^C f(\bar{x})$, if

$$\langle v, \xi\rangle \leq f^o(\bar{x}; \xi), \forall \xi \in \mathbb{R}^n,$$

where $f^o(\bar{x}; \xi)$ is the Clarke directional derivative of f at $\bar{x}$ in the direction ξ ([11, Theorem 2.9.1]), defined as follows

$$f^o(\bar{x}; \xi) := \lim_{\epsilon \downarrow 0} \limsup_{\substack{y \to x \\ t \downarrow 0}} \inf_{\xi \in v + \epsilon B} \frac{f(y + t\xi) - f(y)}{t}.$$

The next lemma significantly simplifies the inclusion relationship among subdifferentials outlined above.

Lemma 1 [11, 27, 31] Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a lower semicontinuous (l.s.c.) function, and let $x \in \mathbb{R}^n$ be a point at which f is finite, then

$$\hat{\partial} f(x) \subset \partial f(x) \subset \partial^C f(x).$$

To proceed, we introduce properties of the Clarke subdifferential and the limiting subdifferential, including existence and boundedness, respectively.

Lemma 2 [11, Section 1.2] Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with $x \in domf$. If f is a locally Lipschitz continuous function around x with modulus $L_f(x) \geq 0$, then we have

$$\|v\| \leq L_f(x), \ \forall v \in \partial^C f(x).$$

Lemma 3 [27, Theorem 1.22] Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ with $x \in domf$. Then f is locally Lipschitz around x with modulus $L_f(x) \geq 0$ if and only if $\partial^\infty f(x) = \{0\}$. In this case $\partial f(x) \neq \emptyset$, and for a fixed Lipschitz modulus $L_f(x)$, we have

$$\|v\| \leq L_f(x), \ \forall v \in \partial f(x).$$

Using the definitions of subdifferentials, Le [22] gives expressions of three subdifferentials of $\|x\|_0$.

Lemma 4 [22, Theorems 1 and 2] For $\|x\|_0 : \mathbb{R}^n \to \overline{\mathbb{R}}$, the regular subdifferential, limiting subdifferential, Clarke subdifferential and horizon subdifferential have the following expressions

$$\hat{\partial}\|x\|_0 = \partial\|x\|_0 = \partial^C\|x\|_0 = \partial^\infty\|x\|_0 = \left\{ v \in \mathbb{R}^n \;\middle|\; v_i \begin{cases} = 0, i \in \Gamma_x, \\ \in \mathbb{R}, i \in \overline{\Gamma}_x. \end{cases} \right\} \quad (2)$$

Clarke [12] stated the calculus rules of the limiting subdifferential.

Lemma 5 [12, proposition 1.5] Let f and g be extended-valued lower semicontinuous functions finite at x, with $\partial^\infty f(x) \cap (-\partial^\infty g(x)) = \{0\}$. Then we have

$$\partial(f + g)(x) \subset \partial f(x) + \partial g(x).$$

To end this section, we will review the definitions of the Moreau envelope and the proximal operator in variational analysis [31].

Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper l.s.c. function and let $\beta \in \mathbb{R}_{++}$. The Moreau envelope of g of parameter β is defined by

$$e_\beta g(x) := \min_{w \in \mathbb{R}^n} \{ \frac{1}{2\beta}\|w - x\|^2 + g(w) \} \leq g(x),$$

and the proximal operator of f of parameter β is defined by

$$Prox^B_{\beta g}(x) := \operatorname*{argmin}_{w \in \mathbb{R}^n} \{ \frac{1}{2\beta}\|w - x\|^2 + g(w) \}.$$

For convenience, let $Prox_{\beta g}(x)$ refer to an element in $Prox^B_{\beta g}(x)$. The proximal operator, which is extensively studied, plays an important role in designing

optimization algorithms.

The proximal operator of $\|x\|_0$ [8, Proposition 2.4] can be calculated as

$$Prox^B_{\beta\lambda\|\cdot\|_0}(x) = \left\{ v \in \mathbb{R}^n \;\middle|\; v_i = \begin{cases} 0, & \text{if } |x_i| < \sqrt{2\beta\lambda}, \\ 0 \text{ or } x_i, & \text{if } |x_i| = \sqrt{2\beta\lambda}, \\ x_i, & \text{if } |x_i| > \sqrt{2\beta\lambda}. \end{cases} \right\} \tag{3}$$

3 Main results

In this section, we study the first-order necessary and sufficient optimality conditions for Problem (1).

First, we give the definitions of the subdifferential-stationary point (also called KKT point [17]) and the proximal-stationary point (also called L-stationary point [3]) of Problem (1) based on the existence of a subdifferential and features of proximal operator.

Definition 1 Let $x^* \in \mathbb{R}^n$.
(i) We say that x^* is a subdifferential-stationary (S-stationary) point of Problem (1), if

$$0 \in \partial f(x^*) + \lambda \partial \|x^*\|_0. \tag{4}$$

(ii) We say that x^* is a weak proximal-stationary (P-stationary) point of Problem (1), if there exists a constant $\theta > 0$, a vector $p \in \partial f(x^*)$ such that

$$x^* \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p), \text{ for all } \beta \in (0, \theta). \tag{5}$$

If (5) holds for any $p \in \partial f(x^*)$, then we call x^* a P-stationary point of Problem (1).

Remark: When f is smooth, the concepts of both a weak P-stationary point and a P-stationary point are the same.

Based on the definition of an S-stationary point, the next theorem presents first-order necessary and sufficient optimality conditions for local optimal solutions of Problem (1).

Theorem 1 Let $x^* \in \mathbb{R}^n$ be a local optimal solution of Problem (1). Then x^* is an S-stationary point of Problem (1). If f is convex at x^*, then the converse holds.

Proof : Suppose x^* is a local optimal solution of Problem (1). Using Theorem 10.1 in [31], we have

$$0 \in \partial(f(x^*) + \lambda\|x^*\|_0). \tag{6}$$

Because $f(x)$ is locally Lipschitz around x^*, it follows from Lemma 3 that

$$\partial^\infty f(x^*) = \{0\}. \tag{7}$$

Combining (7) and Lemma 4, we have $\partial^\infty f(x^*) \cap -\lambda\partial^\infty\|x^*\|_0 = \{0\}$. Hence, the qualification condition introduced in Lemma 5 holds. Using (6) and Lemma 5, we have

$$0 \in \partial(f(x^*) + \lambda\|x^*\|_0) \subseteq \partial f(x^*) + \lambda\partial\|x^*\|_0.$$

Hence x^* is an S-stationary point of Problem (1).

Conversely, because x^* is an S-stationary point of Problem (1), we have

$$0 \in \partial f(x^*) + \lambda \partial \|x^*\|_0.$$

Thus, for any $\frac{v^*}{\lambda} \in \partial \|x^*\|_0$ with $v_i^* \in \mathbb{R}$ for $i \in \overline{\Gamma}_{x^*}$, $v_i^* = 0$ for $i \in \Gamma_{x^*}$, one has

$$0 \in \partial f(x^*) + v^*. \tag{8}$$

Hence, there exists a vector $\xi^* \in \partial f(x^*)$ such that

$$0 = \xi^* + v^*. \tag{9}$$

To proceed, we first prove x^* is a globally optimal solution of the following convex programming problem with linear equality constraints:

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & x \in S^* := \{x \mid x_i = 0,\ i \in \overline{\Gamma}_{x^*}\}. \end{array} \tag{10}$$

Suppose there exists an $\hat{x} \in S^*$ such that $f(\hat{x}) < f(x^*)$. Because f is convex at x^*, it follows from the properties of the subdifferential of a convex function that

$$\langle \xi, \hat{x} - x^* \rangle \leq f(\hat{x}) - f(x^*) < 0, \text{ for all } \xi \in \partial f(x^*). \tag{11}$$

For any $x \in S^*$, it's easy to see that

$$\langle v^*, x - x^* \rangle = 0. \tag{12}$$

Then, for any $x \in S^*$, we derive from (9) and (12) that

$$0 = \langle \xi^*, x - x^* \rangle + \langle v^*, x - x^* \rangle = \langle \xi^*, x - x^* \rangle.$$

Taking this into account, it follows from $\hat{x} \in S^*$ that $\langle \xi^*, \hat{x} - x^* \rangle = 0$, which contradicts (11) and verifies that x^* is a global minimizer of (10). Thus, for any $x \in S^*$, there has

$$f(x^*) \leq f(x). \tag{13}$$

Because $\|x\|_0$ can only take the integer values $0, 1, ..., m$, there exists a $\delta_1 = \min\{\frac{|x_i^*|}{2}, i = 1, ..., n\}$ such that

$$\|x_i\|_0 \geq \|x_i^*\|_0,\ \forall x \in U(x^*, \delta_1). \tag{14}$$

The proof is divided into two cases.
Case (i): $x \in S^*$ and $x \in U(x^*, \delta_1)$. It follows from (13) and (14) that

$$f(x^*) + \lambda \|x^*\|_0 \leq f(x) + \lambda \|x\|_0. \tag{15}$$

Case (ii): $x \notin S^*$ and $x \in U(x^*, \delta_1)$. For any $x \notin S^*$, there exists an $i_0 \in \overline{\Gamma}_{x^*}$ with $x_{i_0} \neq 0$. This implies $\|x_{i_0}\|_0 = 1$ but $\|x_{i_0}^*\|_0 = 0$. Using (14), we have

$$\|x\|_0 \geq \|x^*\|_0 + 1. \tag{16}$$

Because the function $f(x)$ is locally Lipschitz continuous around x^*, so there exist constant $L_f(x^*)$ and $\delta_2 = \frac{\lambda}{L_f(x^*)}$ such that

$$|f(x) - f(x^*)| \leq L_f(x^*)||x - x^*|| \leq L_f(x^*)\frac{\lambda}{L_f(x^*)} = \lambda, \ \forall x \in U(x^*, \delta_2). \quad (17)$$

Taking $\delta = \min\{\delta_1, \delta_2\}$ and combining (16) and (17), we derive that for any $x \in U(x^*, \delta)$,

$$\begin{aligned} f(x^*) + \lambda\|x^*\|_0 &\leq f(x^*) + \lambda(\|x\|_0 - 1) \\ &\leq (f(x) + \lambda) + \lambda\|x\|_0 - \lambda = f(x) + \lambda\|x\|_0. \end{aligned} \quad (18)$$

Summarizing (15) and (18), we establish that x^* is a local minimizer of Problem (1). The proof is completed. □

Remark: Based on Theorem 1 and lemma 1, if $x^* \in \mathbb{R}^n$ is a local optimal solution of Problem (1), then $0 \in \partial^C f(x^*) + \lambda\partial^C\|x^*\|_0$, which is called Clark-stationary point.

Next, we discuss the relationship between the S-stationary point and the weak P-stationary point of Problem (1).

Theorem 2 $x^* \in \mathbb{R}^n$ is an S-stationary point of Problem (1) if and only if x^* is a weak P-stationary point of Problem (1), where

$$\theta := \min\{1, \frac{\min\{(x_i^*)^2, \ i \in \Gamma_{x^*}\}}{2\lambda}\}, \lambda > \kappa := \frac{1}{2}L_f^2(x^*).$$

Proof : Because x^* is an S-stationary point of Problem (1), it follows from Definition 1 that $0 \in \partial f(x^*) + \lambda\partial\|x^*\|_0$. This implies the existence of $p \in \partial f(x^*)$ and $u \in \partial\|x^*\|_0$ such that $0 = p + \lambda u$, and thus, using Lemma 4, $p_i = 0, i \in \Gamma_{x^*}$.

It follows from (3) that

$$Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p) = \left\{ v \in \mathbb{R}^n \ \middle| \ v_i = \begin{cases} 0, & \text{if } |x_i^* - \beta p_i| < \sqrt{2\beta\lambda}, \\ 0 \text{ or } x_i^* - \beta p_i, & \text{if } |x_i^* - \beta p_i| = \sqrt{2\beta\lambda}, \\ x_i^* - \beta p_i, & \text{if } |x_i^* - \beta p_i| > \sqrt{2\beta\lambda}. \end{cases} \right\}$$

For $i \in \Gamma_{x^*}$, i.e., $x_i^* \neq 0$, combining $p_i = 0$ and $0 < \beta \leq \theta$, we have

$$|x_i^* - \beta p_i| = |x_i^*| \geq \min\{|x_i^*|, \ i \in \Gamma_{x^*}\} \geq \sqrt{2\theta\lambda} \geq \sqrt{2\beta\lambda}.$$

For $i \in \overline{\Gamma}_{x^*}$, i.e., $x_i^* = 0$, we combine $|p_i| \leq \|p\| \leq L_f(x^*) \leq \sqrt{2\lambda}$ with $0 < \beta < \theta \leq 1$ to obtain $|x_i^* - \beta p_i| = |\beta p_i| \leq \sqrt{\beta}|p_i| \leq \sqrt{2\beta\lambda}$.
To sum up, we conclude that $x^* \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$, which implies x^* is a weak P-stationary point and complete the proof of the "only if" part.

To verify the converse implication, because x^* is a weak P-stationary point of Problem (1), there exists a $p \in \partial f(x^*)$ such that

$$x_i^* = \begin{cases} 0, & \text{if } |x_i^* - \beta p_i| < \sqrt{2\beta\lambda}, \\ 0 \text{ or } x_i^* - \beta p_i, & \text{if } |x_i^* - \beta p_i| = \sqrt{2\beta\lambda}, \\ x_i^* - \beta p_i, & \text{if } |x_i^* - \beta p_i| > \sqrt{2\beta\lambda}. \end{cases}$$

For $i \in \Gamma_{x^*}$, combing the formula above with $x_i^* \neq 0$, we have $x_i^* = x_i^* - \beta p_i$. Thus $p_i = 0, i \in \Gamma_{x^*}$. For $i \in \overline{\Gamma}_{x^*}$, using $x_i^* = 0$, it is easy to see that $|x_i^* - \beta p_i| < \sqrt{2\beta\lambda}$, which implies $p_i < \sqrt{\frac{2\lambda}{\beta}}$. Thus, x^* is an S-stationary point of Problem (1). The proof is completed. □

To end this section, we present the relationship between the global optimal solution and the P-stationary point of Problem (1). Let us start with the following simple assumption.

Assumption 1 Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a local Lipschitz continuous function. Then for any $x, y \in \mathbb{R}^n$, there exists an $L > 0$ such that

$$f(x) \leq f(y) + \langle q, x - y\rangle + \frac{L}{2}\|x - y\|^2 \text{ for any } q \in \partial f(y). \tag{19}$$

Remark: Similar to the proof of [2, Proposition A.24], if f is a Lipschitz continuous function on $\mathbb{R}^n$, and satisfies subgradient Lipschitz with L, i.e.,

$$\|p - q\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^n, \forall p \in \partial f(x), \ \forall q \in \partial f(y),$$

then this assumption is naturally true.

Under some cases of f, the above assumption is true. For example, let $f := (x_+ - 1)^2$ in $\mathbb{R}^n$. Apparently, it is locally lipschitz with

$$\partial f(x) := \begin{cases} \{0\}, & \text{if } x < 0, \\ \{0, -2\}, & \text{if } x = 0, \\ \{2(x-1)\}, & \text{if } x > 0. \end{cases}$$

For all $x, y \in \mathbb{R}$, it is easy to calculate that (19) holds for any $q \in \partial f(y)$ under

$$\begin{array}{ll} \text{case (i)}: \ x \geq 0, y > 0. \ L_1 = 2; & \text{case (ii)}: \ x \leq 0, y < 0. \ L_2 = 0; \\ \text{case (iii)}: \ x > 0, y < 0. \ L_3 = 0; & \text{case (iv)}: \ x < 0, y > 0. \ L_4 = 2; \\ \text{case (v)}: \ x < 0, y = 0. \ L_5 = 0; & \text{case (vi)}: \ x > 0, y = 0. \ L_6 = 2. \end{array}$$

Taking $L = \max\{L_i, i = 1, 2, ..., 6\} = 2$, then we establish that Assumption 1 holds.

Relying on the above assumption, we now show that any global optimal solution of Problem (1) is a P-stationary point for any $\beta \in (0, \frac{1}{L})$.

Theorem 3 Consider Problem (1) with $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ satisfying Assumption 1. If x^* be a global optimal solution of Problem (1), then for any $p \in \partial f(x^*)$, we have
(i) x^* is a P-stationary point, i.e., $x^* \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$, $\forall \beta \in (0, \frac{1}{L})$.
(ii) The set $Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$ contains exactly one element.

Proof : We observe that the set $Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$ is nonempty for any $p \in \partial f(x^*)$. Thus, we just need to prove (ii), where one element is x^*. Suppose, to the contrary, that there exists an element

$$z \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p),$$

which is different from x^*. Letting

$$S_L(x, y) := f(y) + \langle q, x - y\rangle + \frac{L}{2}\|x - y\|^2 \text{ for all } q \in \partial f(y),$$

it follows from Assumption 1 that

$$f(x^*) - f(z) \geq f(x^*) - S_L(z, x^*). \tag{20}$$

Then for all $p \in \partial f(x^*)$, we have

$$\begin{aligned} S_{\frac{1}{\beta}}(z, x^*) + \lambda\|z\|_0 &= f(x^*) + \langle p, z - x^*\rangle + \frac{1}{2\beta}\|z - x^*\|^2 + \lambda\|z\|_0 \\ &= \frac{1}{2\beta}\left\|z - (x^* - \beta p)\right\|^2 + \lambda\|z\|_0 + f(x^*) - \frac{\beta}{2}\|p\|^2 \\ &= \frac{1}{\beta}\Big(\frac{1}{2}\left\|z - (x^* - \beta p)\right\|^2 + \beta\lambda\|z\|_0\Big) + f(x^*) - \frac{\beta}{2}\|p\|^2. \end{aligned}$$

Together with the fact that z is an optimal solution of $Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$, we obtain

$$\begin{aligned} S_{\frac{1}{\beta}}(z, x^*) + \lambda\|z\|_0 &\leq \frac{1}{\beta}\Big(\frac{1}{2}\left\|x^* - (x^* - \beta p)\right\|^2 + \beta\lambda\|x^*\|_0\Big) + f(x^*) - \frac{\beta}{2}\|p\|^2 \\ &= S_{\frac{1}{\beta}}(x^*, x^*) + \lambda\|x^*\|_0 \\ &= f(x^*) + \lambda\|x^*\|_0. \end{aligned} \tag{21}$$

Notably,

$$S_L(z, x^*) = S_{\frac{1}{\beta}}(z, x^*) - \frac{\frac{1}{\beta} - L}{2}\|x^* - z\|^2. \tag{22}$$

Subsequently, from (20), (22) and (21), we can easily derive

$$\begin{aligned} f(z) + \lambda\|z\|_0 &\leq S_L(z, x^*) + \lambda\|z\|_0 \\ &= S_{\frac{1}{\beta}}(z, x^*) + \lambda\|z\|_0 - \frac{\frac{1}{\beta} - L}{2}\|x^* - z\|^2 \\ &\leq f(x^*) + \lambda\|x^*\|_0 - \frac{\frac{1}{\beta} - L}{2}\|x^* - z\|^2, \end{aligned}$$

which contradicts the optimality of x^*. Hence, x^* is the only vector in the set $Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p)$. The proof is completed. □

Beck et al. in [3] posted that x^* is the unique vector in $Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta\nabla f(x^*))$ for

$$\min_{x\in\mathbb{R}^n} \; f(x) + \lambda\|x\|_0,$$

when f is a strongly smooth function. In Theorem 3, we extend this result to Problem (1). Furthermore, the main advantages of the P-stationary point are summarized as follows: (i) Based on Theorems 2 and 3, this conception of P-stationary point is stronger than that of the S-stationary point for Problem (1). (ii) The P-stationary point uses the proximal operator of $\|x\|_0$, which has a closed-form solution and is thus easy to compute. (iii) Based on Theorem 3, we can design efficient algorithms for Problem (1). Furthermore, a few algorithms have been proposed for smooth optimization problems with l_0-regularization, e.g., [3, 6, 7].

4 Applications

In this section, we apply the developed optimality conditions to several well-known applications.

Example 1: Least Absolute Deviation Compressed Sensing Problem (LADCS) [13, 16, 23].

Compressed sensing(CS) is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems. CS is presented as

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|^2+\lambda\|x\|_0,$$

where $A=[A_1,A_2,...,A_m]^\top\in\mathbb{R}^{m\times n}$ is the sensing matrix, and $b\in\mathbb{R}^m$ is the given measurement. To increase the robustness of the CS, CS can convert into the following LADCS problem (for example, [23]):

$$\min_{x\in\mathbb{R}^n} \quad \|Ax-b\|_1+\lambda\|x\|_0. \tag{23}$$

Apparently, Problem (23) is a special case of Problem (1) when $f(x)=\|Ax-b\|_1$ with $L_f(x)=\sum\limits_{i=1}^{m}\|A_i\|_1$.

Because $\|Ax-b\|_1$ is a convex function, we have

$$\partial\|Ax-b\|_1=\sum_{i=1}^{m} sign(A_i^\top x-b_i)A_i.$$

Thus, we give the following explicit expressions for stationary points of Problem (23) according to Definition 1.
(i) S-stationary points of Problem (23):

$$0\in\sum_{i=1}^{m} sign(A_i^\top x-b_i)A_i+\partial\|x^*\|_0.$$

(ii) Weak P-stationary points of Problem (23): $\exists p \in \sum_{i=1}^{m} sign(A_i^\top x - b_i)A_i$ such that

$$x^* \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p) \text{ for all } \beta \in (0, \min\{1, \frac{\min\{(x_i^*)^2,\ i \in \Gamma_{x^*}\}}{2\lambda}\}),$$

where $\lambda \in (\frac{1}{2}L_f^2(x^*), \infty)$ and $L_f(x^*) = \sum_{i=1}^{m} \|A_i\|_1$.

Notice that $\|Ax - b\|_1$ is a convex function, it follows from Theorems 1 and 2 that the S-stationary points, weak P-stationary points and local minimum points of Problem (23) are equivalent.

Example 2: ReLU Deep Learning [15, 35].

A multi-layer neural network with the ReLu activation function is a widely used model in deep learning. For simplicity, taking m independent and identically distributed samples $A_i \in \mathbb{R}^n$ and $b_i \geq 0,\ i = 1, 2, ..., m$, we present a 1-layer model in deep learning,

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \left((A_i^\top x)_+ - b_i\right)^2 + \lambda\|x\|_0, \tag{24}$$

where $(x)_+ := \max\{x, \mathbf{0}\}$ is call Rectified Linear Unit (ReLU) function [19]. Obviously, Problem (24) is a special case of Problem (1) when we take $f(x) = \sum_{i=1}^{m}((A_i^\top x)_+ - b_i)^2$. Problem (24) is also an optimization model in censored regression problem [4, 30].

It's not hard to see that

$$\begin{aligned}\Theta_i(x) := \partial((A_i^\top x)_+ - b_i)^2 &= \partial((A_i^\top x)_+)^2 - 2b_i\partial\max\{A_i^\top x, 0\} + \partial b_i^2 \\ &= \begin{cases} \{0, 2A_i\left(A_i^{\mathrm{T}}x - b_i\right)\}, & \text{if } A_i^\top x = 0, \\ \{0\}, & \text{if } A_i^\top x < 0, \\ \{2A_i\left(A_i^{\mathrm{T}}x - b_i\right)\}, & \text{if } A_i^\top x > 0. \end{cases}\end{aligned}$$

So we have $\Theta(x) := \partial\sum_{i=1}^{m}((A_i^\top x)_+ - b_i)^2 = \sum_{i=1}^{m}\partial((A_i^\top x)_+ - b_i)^2 = \sum_{i=1}^{m}\Theta_i(x)$. Furthermore, for any $x^* \in \mathbb{R}^n$, it follows from

$$\sum_{i=1}^{m}\Theta_i(x) \leq 2\sum_{i=1}^{m}\|A_i\|\|A_i^{\mathrm{T}}x - b_i\| \leq 2\|A\|_F^2\|x^*\| + 2\sum_{i=1}^{m}|b_i|\|A_i\|$$

that there exists a constant

$$L_f(x^*) := 2\|A\|_F^2\|x^*\| + 2\sum_{i=1}^{m}|b_i|\|A_i\|$$

such that $\|p\| \leq L_f(x^*),\ \forall p \in \Theta(x^*)$.

Furthermore, let us consider the following proposition, which shows that $\sum_{i=1}^{m}((A_i^\top x)_+ - b_i)^2$ satisfy Assumption 1 with a constant $L > 0$.

Proposition 1 For $f(x) = \sum_{i=1}^{m}((A_i^\top x)_+ - b_i)^2$, Assumption 1 holds with $L = 2\|A\|_F^2$.

Proof : For any $x, y \in \mathbb{R}^n$, it is not hard to know that (19) holds for any $q \in \partial f(y)$ under following four cases.
Case (i): $A_i^T y = 0, A_i^\top x > 0$. It is easy to see that $\partial f(y) = \left\{0, 2A_i^\top \left(A_i^\top x - b_i\right)\right\}$. We need to check $\left(A_i^\top x - b_i\right)^2 \leq b_i^2 + \langle q, x - y\rangle + \frac{L_1}{2}\|x - y\|^2$. For $p = 0$, we have $\left(A_i^\top x\right)^2 - 2b_i A_i^T x \leq \langle q, x - y\rangle + \frac{L_{11}}{2}\|x - y\|^2$. It follows from $\left(A_i^\top x\right)^2 \leq \|A_i\|^2\|x\|^2$ and $\left(A_i^\top x\right)^2 = x^\top A_i A_i^\top x = (x - y)^\top A_i A_i^\top (x - y)$ that $\|A_i\|^2\|x - y\|^2 \leqslant \frac{L_{11}}{2}\|x - y\|^2$, which means $L_{11} = 2\|A_i\|^2$. For $p = A_i^\top \left(A_i^\top x - b_i\right)$, it is not hard to check that $\left(A_i^\top x\right)^2 - 2b_i A_i^T x \leq 2(A_i^T x)^2 - 2b_i A_i^T x + \frac{L_{12}}{2}\|x - y\|^2$ is naturally established for any $L_{12} \geq 0$. Thus $L_1 = \max\{L_{11}, L_{12}\} = 2\|A_i\|^2$.
Case (ii): $A_i^T y = 0, A_i^\top x \leq 0$. We have $L_2 = 0$.
Case (iii): $A_i^\top y < 0$. we have $\partial f(y) = 0$. For $A_i^\top x > 0$, using $\left(A_i^\top x\right)^2 - 2b_i A_i^T x \leq \langle q, x - y\rangle + \frac{L_{31}}{2}\|x - y\|^2$, we get $L_{31} = 2\|A_i\|^2$. For $A_i^\top x \leq 0$, it is not hard to check that $0 \leq 0 + \frac{L_{32}}{2}\|x - y\|^2$ is naturally established for any $L_{32} \geq 0$. Hence, $L_3 = \max\{L_{31}, L_{32}\} = 2\|A_i\|^2$.
Case (iv): $A_i^\top y > 0$. we have $\partial f(y) = 2A_i\left(A_i^\top x - b_i\right)$. For $A_i^\top x > 0$, it follows from $0 \leq (A_i^\top y)^2 + (A_i^\top x)^2 - 2A_i^\top y A_i^\top x + \frac{L_{41}}{2}\|x - y\|^2$ that $L_{41} = 2\|A_i\|^2$. For $A_i^\top x \leq 0$, it is not hard to check that $0 \leq (A_i^\top y)^2 + 2(A_i^\top x)^2 - 2b_i A_i^\top x - 2A_i^\top y A_i^\top x + \frac{L_{42}}{2}\|x - y\|^2$ is naturally established for any $L_{42} \geq 0$. Thus, we have $L_4 = \max\{L_{41}, L_{42}\} = 2\|A_i\|^2$.

Taking $L = \sum_{i=1}^{m}(\max\{L_i, i = \{1, 2, 3, 4\}) = 2\|A\|_F^2$, then we have the conclusion. □

Based Definition 1 and Proposition 1, we give the following explicit expressions for stationary points of Problem (24).
(i) S-stationary points of Problem (24):

$$0 \in \Theta(x^*) + \partial\|x^*\|_0,$$

(ii) Weak P-stationary points of Problem (24): exists $p \in \Theta(x^*)$ such that

$$x^* \in Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p) \text{ for all } \beta \in (0, \min\{1, \frac{\min\{(x_i^*)^2,\ i \in \Gamma_{x^*}\}}{2\lambda}\}),$$

where $\lambda \in \left(2\left(\|A\|_F^2\|x^*\| + \sum_{i=1}^{m}|b_i|\|A_i\|\right)^2, \infty\right)$.
(iii) P-stationary points of Problem (24): for any $p \in \Theta(x^*)$ such that

$$x^* = Prox^B_{\beta\lambda\|\cdot\|_0}(x^* - \beta p) \text{ for all } \beta \in (0, \frac{1}{L}), \tag{25}$$

where $L = 2\|A\|_F^2$.

Hence, we find that local minimizer of Problem (24) are S-stationary points

and weak P-stationary points of Problem (24) using Theorems 1 and 2. Furthermore, based Theorem 3, we obtain that the global optimal solutions of Problem (24) are P-stationary points. Furthermore, the characterization of a P-stationary point presents a much easier way to design fast numerical algorithms to search. This idea has been applied in some fast Newton algorithms for problems with smooth objective and sparsity constraint [36–39].

5 Conclusions

In this paper, we have introduced and analyzed the S-stationary point and P-stationary point of locally Lipschitz continuous optimization problem with l_0-regularization (1) by exploring the characteristic of $\|x\|_0$. Furthermore, different from the relaxation methods, our optimality conditions are for the original Problem (1) and are sufficient and necessary under some suitable conditions. In the future, based on the special structure of $\|x\|_0$ and the definition of P-stationary point, we will design some rapid and efficient algorithms for Problem (1), especially, for the deep learning. Furthermore, our approach can address the problem with sparsity constraint based on the connection between solutions of the optimization problems with sparsity constraint and l_0-regularization [9, 29, 40].

Acknowledgements

The authors would like to thank the associate editor and two anonymous referees for their constructive comments, which have significantly improved the quality of the paper. This work is supported by the National Natural Science Foundation of China (No.11971052) and (No.11801325).

References

1. Allen-Zhu Z, Hazan E. Variance reduction for faster non-convex optimization. *in Proceedings of the 33rd International Conference on Machine Learning*, 699-707.(2016)
2. Bertsekas D P. Nonlinear Programming. *2nd ed,Athena Scientific*, Belmont, MA, (1999)
3. Beck A and Hallak N. Proximal mapping for symmetric penalty and sparsity. *SIAM Journal on Optimization*, 28(1): 496-527.(2018)
4. Bian W. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. Submitted: 1-25.(2019)
5. Blumensath T. Compressed sensing with nonlinear observations and related nonlinear optimization problems. *IEEE Transactions on Information Theory*, 59(6):3466-3474.(2013)
6. Blumensath T, Davies M E. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629-654.(2008)
7. Blumensath T, Davies M E. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265-274.(2009)
8. Chen Y Q, Xiu N H, Peng D T. Global solutions of non-Lipschitz S_2CS_p minimization over the positive semidefinite cone. Optimization Letters, 8(7):2053-2064.(2013)
9. Chen X J, Pan L L, Xiu N H. Relationship between three sparse optimization problems for multivariate regression. Submitted: 1-32.(2019)

10. Chib S. Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51(1-2):79-99.(1992)
11. Clarke F H. Optimization and nonsmooth analysis, *Wiley.*(1983)
12. Clarke F H. Methods of dynamic and nonsmooth optimization, *CBMS-NSF Regional conference series in Applied mathmatics, SIAM Publications, Philadelphia*, 57. (1989)
13. Candès E J, Tao T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(42):4203-4215.(2005)
14. Chen X J, Ge D D, Wang Z Z, et al. Complexity of unconstrained L_2-L_p minimization. *Mathematical Programming*, 143(1-2):371-383.(2014)
15. Cui Y , Pang J S , Sen B. Composite Difference-Max Programs for Modern Statistical Estimation Problems. *SIAM Journal on Optimization*, 28(4):3344-3374. (2018)
16. Donoho D L. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289-1306.(2006)
17. Guo L, Ye J J. Necessary optimality conditions and exact penalization for non-Lipschitz nonlinear programs. *Mathematical Programming*, 168(1-2):571C598.(2018)
18. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *Artificial Intelligence and Statistics*. 15:315-323.(2011)
19. Hinton G E. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. *International Conference on International Conference on Machine Learning. Omnipress.* (2010)
20. Hossein R, Ajmal M, Mubarak S. Learning a deep model for human action recognition from novel viewpoints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):667-681.(2017)
21. Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Computer Science.*(2014)
22. Le H Y, Generalized subdifferentials of the rank function. *Optimization Letters*, 7(4):731-743.(2013)
23. Liu J, Cosman P C, Rao B D. Robust linear regression via l_0 regularization. *IEEE Transactions on Signal Processing*, 66(3):698-713.(2017)
24. Lu Z S, Zhang Y. Sparse Approximation via Penalty Decomposition Methods. *SIAM Journal on Optimization*, 23(4):2448-2478.(2013)
25. Lu Z S. Iterative reweighted minimization methods for l_p-regularized unconstrained nonlinear programming. *Mathematical Programming*, 147(1-2):277-307.(2014)
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 521:436-444.(2015)
27. Mordukhovich B S. Variational analysis and application. *Springer.*(2018)
28. Natarajan B K . Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227-234.(1995)
29. Nikolova M. Relationship between the optimal solutions of least squares regularized with l_0-norm and constrained by k-sparsity. *Applied and Computational Harmonic Analysis*, 41(1):237-265.(2016)
30. Powell J L. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303-325.(1984)
31. Rockafellar R T. Wets R J. Variational analysis. *Springer.*(1998)
32. Rockafellar R T. Convex analysis. *Princeton University Press.*(1970)
33. Thorarinsdottir T L, Gneiting T. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 173(2):371-388.(2010)
34. Witten I H, Frank E. Data mining: practical machine learning tools and techniques with java implementations. *Morgan Kaufmann Publishers.*(2000)
35. Yu D, Deng L. Automatic Speech Recognition: A Deep Learning Approach, Signals and Communications Technology, *Springer.* (2015)
36. Yuan X T, Liu Q S. Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4122-4129. (2014)
37. Yuan X T, Liu Q S. Newton-type greedy selection methods for l_0-constrained minimization. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2437-2450.(2017)

38. Wang R, Xiu N H, Zhou S L. Fast newton method for sparse logistic regression. *arXiv preprint arXiv:1901.02768.*(2019)
39. Zhou S L, Xiu N H, Qi H D. Global and quadratic convergence of Newton hard-thresholding pursuit. *arXiv preprint arXiv:1901.02763.*(2019)
40. Zhang N, Li Q. On optimal solutions of the constrained l_0 regularization and its penalty problem. *Inverse Problems*, 33(2).(2017)