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

Computation of second-order directional stationary points for group sparse optimization

Dingtao Peng · Xiaojun Chen

Abstract We consider a nonconvex and nonsmooth group sparse optimization problem where the penalty function is the sum of compositions of a folded concave function and the ℓ_2 vector norm for each group variable. We show that under some mild conditions a first-order directional stationary point is a strict local minimizer that fulfils the first-order growth condition, and a second-order directional stationary point is a strong local minimizer that fulfils the second-order growth condition. In order to compute second-order directional stationary points, we construct a twice continuously differentiable smoothing problem and show that any accumulation point of the sequence of second-order stationary points of the smoothing problem is a second-order directional stationary point of the original problem. We give numerical examples to illustrate how to compute a second-order directional stationary point by the smoothing method.

Keywords Group sparse optimization; nonconvex and nonsmooth optimization; composite folded concave penalty; directional stationary point; smoothing method

MSC(2010) 90C26 · 90C46

1 Introduction

Let $\mathbf{x} = (\mathbf{x}_1^\top, \cdots, \mathbf{x}_K^\top)^\top \in \mathbb{R}^n$ with $\mathbf{x}_i = (x_{i(1)}, \cdots, x_{i(d_i)})^\top \in \mathbb{R}^{d_i}$, $d_i \geq 1$, $\sum_{i=1}^{K} d_i = n$. We consider the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) := \mathcal{L}(\mathbf{x}) + \sum_{i=1}^{K} \varphi(\|\mathbf{x}_i\|), \tag{1.1}$$

where $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function, and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a concave penalty function satisfying the following properties: (i) φ is locally Lipschitz continuous and non-decreasing on $[0, \infty)$ with $\varphi(0) = 0$ and $\varphi(t) > 0$ for $t > 0$; (ii) $\varphi'(0+) > 0$. Throughout this paper, $\|\cdot\|$ denotes the ℓ_2 vector norm.

The paper is dedicated to Professor Ya-Xiang Yuan on the occasion of his 60th birthday.

Dingtao Peng, School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China.
E-mail: dingtaopeng@126.com
Xiaojun Chen ✉, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China.
E-mail: maxjchen@polyu.edu.hk

· This paper is partially supported by the Hong Kong Research Grant Council PolyU, PolyU153000/17P, NSFC (11861020), the Growth Project of Education Department of Guizhou Province for Young Talents in Science and Technology ([2018]121), and the Foundation for Selected Excellent Project of Guizhou Province for High-level Talents Back from Overseas ([2018]03)

In practice, many loss functions are twice continuously differentiable, for example, square loss function $\mathcal{L}(\mathbf{x}) = \frac{1}{2m}\|A\mathbf{x}-b\|^2$, exponential loss function $\mathcal{L}(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m}\exp(-b_j(\mathbf{a}_j^\top\mathbf{x}))$, and logistic loss function

$$\mathcal{L}(\mathbf{x}) = -\frac{1}{m}\sum_{j=1}^{m}\left\{b_j\log(1+\exp(-\mathbf{a}_j^\top\mathbf{x})) + (1-b_j)\log(1+\exp(\mathbf{a}_j^\top\mathbf{x}))\right\},$$

where $\mathbf{b}\in\mathbb{R}^m$, $A=(\mathbf{a}_1,\cdots,\mathbf{a}_m)^\top\in\mathbb{R}^{m\times n}$.

Problem (1.1) is called group sparse optimization due to the group structure in its variable. When $K=n$ and $d_1=\cdots=d_n=1$, problem (1.1) reduces to the standard sparse optimization which is aimed to find a sparse solution to minimize the function $\mathcal{L}(\mathbf{x})$. Sparse optimization has attracted considerable attention in signal processing, machine learning and statistics in recent years. To yield a sparse solution, a penalty term is often used. Tibshirani [28] suggested using the ℓ_1 penalty to obtain a sparse vector of regression coefficients in linear regression problem, which results in a convex optimization problem, called Lasso, and can be solved by many efficient algorithms. However, Fan and Li [12,13] pointed out that the solution of the ℓ_1 penalized optimization does not possess some good statistical properties such as unbiasedness and oracle property. Fan and Li [12,13] then proposed a folded concave penalty and showed that there exists a local solution with the desired statistical properties for the resulting non-convex optimization. Till now, many specific folded concave penalty functions are widely used in signal reconstruction, image restoration, and variable selection, for example, logarithm penalty [12], fraction penalty [25], hard thresholding penalty (HT-P) [6,20], capped ℓ_1 penalty (CapL1) [36], minimax concave penalty (MCP) [35], smoothly clipped absolute deviation (SCAD) [12].

Although there exist some local minimizers with good statistical properties for a folded concave penalized optimization, how to find such local minimizers has not been addressed satisfactorily. Fan, Xue and Zou [14] proposed a local linear approximation algorithm to obtain an oracle solution with an initial point being sufficiently close to the true solution. In [23], the authors developed a concept of subspace second-order optimality which is related to subspace optimality in [3,4,9,10], and showed that under some conditions the stationary point of subspace second-order optimality can be an oracle solution with high probability. In 1985, Yuan [33] studied convergence of trust region algorithms to a first-order d(irectional)-stationary point of nonsmooth optimization. Recently, [1,26] adopted a first-order d(irectional)-stationary point for optimality, and showed that a first-order d-stationary point must be one of other stationary points using the first-order information of the objective function. Moreover, [7,27] proposed the concept of second-order directional derivatives and the concept of second-order d(irectional)-stationary points, and showed that under some mild conditions second-order d-stationary points can fulfil the second-order growth condition. However, how to compute second-order directional derivatives and second-order d-stationary points is unknown for problem (1.1).

Group sparse problem was studied by many authors, e.g., see [11,15,16,17,18,19,24, 30,32,34,37]. It has wide applications in statistics, machine learning, and computational biology such as joint covariate selection [16,17,34,37], multi-task learning [19,32], and gene finding [15,24]. Most of the literatures use group ℓ_1 penalty which yields group Lasso model. Huang and Zhang [17] showed that group Lasso is superior to standard Lasso for strongly group-sparse signals. In consideration of the good performance of folded concave penalties comparing to ℓ_1 penalty for standard sparse optimization, some authors used group folded concave penalties such as group SCAD [5,22,29], group MCP [5,22,29], $\ell_q(\ell_p)$ $(0\le q\le 1\le$

p) [15] and $\ell_0(\ell_2)$ [19] for group sparse problems. However, these works only used first-order information of objective functions which is weaker than second-order information.

In this paper, we will provide a deep analysis of the second-order directional stationarity for folded concave penalized group sparse optimization. Our main contributions are presented as follows.

In Section 2, by virtue of an explicit formula for computing the directional derivative of the objective function, we show that under some mild conditions a first-order d-stationary point of problem (1.1) is a strict local minimizer that fulfils the first-order growth condition.

In Section 3, we provide an explicit formula for computing the second-order directional derivative, and show that under some mild conditions a second-order d-stationary point of problem (1.1) is a strong local minimizer that fulfils the second-order growth condition. Moreover, we establish lower bounds of the ℓ_2 vector norm of nonzero groups of second-order d-stationary points of problem (1.1). These lower bounds are important for theoretical analysis and numerical algorithms.

In Section 4, we construct a twice continuously differentiable smoothing approximation for the nonsmooth objective function in problem (1.1), and show that any accumulation point of the sequence of second-order stationary points of the smoothing problem is a second-order d-stationary point of the original problem. This result provides a theoretic basis for computing second-order d-stationary points of problem (1.1) using the gradient and Hessian of the smoothing function.

Notations. For any $\widehat{\mathbf{x}} \in \mathbb{R}^n$ and the groups $\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_K$, denote

$$\begin{gathered}
I(\widehat{\mathbf{x}}) := \{i \in \{1, \cdots, K\} : \|\widehat{\mathbf{x}}_i\| \neq \mathbf{0}\}, \quad J_i(\widehat{\mathbf{x}}) := \{j \in \{1, \cdots, d_i\} : \widehat{x}_{i(j)} \neq 0\} \text{ for } i \in I(\widehat{\mathbf{x}}), \\
i \notin I(\widehat{\mathbf{x}}) \text{ if } i \in \{1, \cdots, K\} \setminus I(\widehat{\mathbf{x}}), \quad j \notin J_i(\widehat{\mathbf{x}}) \text{ if } i \in I(\widehat{\mathbf{x}}) \text{ and } j \in \{1, \cdots, d_i\} \setminus J_i(\widehat{\mathbf{x}}), \\
[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i := ([\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(1)}, \cdots, [\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(d_i)})^\top, \quad \nabla \mathcal{L}(\widehat{\mathbf{x}}) := ([\nabla \mathcal{L}(\widehat{\mathbf{x}})]_1^\top, \cdots, [\nabla \mathcal{L}(\widehat{\mathbf{x}})]_K^\top)^\top,
\end{gathered}$$

where $\widehat{x}_{i(j)} \in \mathbb{R}$ denotes the jth entry in $\widehat{\mathbf{x}}_i$ and $[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(j)}$ denotes the jth entry in $[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i$.

2 First-order d-stationary points

This section provides the local optimality and some properties of first-order d-stationary points of problem (1.1).

2.1 **Local optimality of first-order d-stationary points**

Let us introduce the concept of first-order d-stationary points [1,7,26,27].

Definition 2.1 $\widehat{\mathbf{x}} \in \mathbb{R}^n$ *is called a first-order d-stationary point of problem (1.1) if the directional derivative satisfies*

$$f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) := \lim_{t \downarrow 0} \frac{f(\widehat{\mathbf{x}} + t(\mathbf{x} - \widehat{\mathbf{x}})) - f(\widehat{\mathbf{x}})}{t} \geq 0, \quad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{2.1}$$

According to [1,26], first-order d-stationary points are sharper than lifted stationary points, critical points, and C-stationary points for the local optimality. It is known that first-order d-stationary points have the following locally optimal properties.

Theorem 2.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous and directionally differentiable at $\widehat{\mathbf{x}} \in \mathbb{R}^n$. The following two statements hold:*

(i) If $\widehat{\mathbf{x}}$ is a local minimizer of f, then $\widehat{\mathbf{x}}$ is a first-order d-stationary point of f.

(ii) $\widehat{\mathbf{x}}$ is a strict local minimizer that fulfils the first-order growth condition, i.e., there exists a neighborhood $\mathcal{W}$ of $\widehat{\mathbf{x}}$ and a positive number δ such that

$$f(\mathbf{x}) \geq f(\widehat{\mathbf{x}}) + \delta\|\mathbf{x} - \widehat{\mathbf{x}}\|, \ \forall\ \mathbf{x} \in \mathcal{W}, \tag{2.2}$$

if and only if $\widehat{\mathbf{x}}$ satisfies that

$$f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) > 0, \ \forall\ \mathbf{x} \in \mathbb{R}^n \setminus \{\widehat{\mathbf{x}}\}. \tag{2.3}$$

If f is differentiable at $\mathbf{x}$, then $f'(\mathbf{x}; \mathbf{z}) = \langle \nabla f(\mathbf{x}), \mathbf{z} \rangle$. Inequality (2.3) does not hold at any differentiable point of f, but it may hold at some non-differentiable points of f. Many local minimizers of problem (1.1) are non-differentiable points of f, which makes conclusion (ii) of Theorem 2.2 very interesting. For example, let $f(t) = t^2 + \log(1 + |t|)$, then $f'(0; s) = |s| > 0$ $(s \neq 0)$, and $f(t) \geq |t|$ for any $t \in \mathbb{R}$.

To have a clear presentation, we denote the ℓ_2 vector norm as a function

$$m(\mathbf{u}) := \|\mathbf{u}\| = \left(\sum_{j=1}^{d_i} u_j^2 \right)^{\frac{1}{2}}, \quad \forall\ \mathbf{u} \in \mathbb{R}^{d_i},\ i \in \{1, \cdots, K\}. \tag{2.4}$$

Although the dimensions of the vectors may be different, we believe that it will not cause any confusion according to the context.

Since $m(\mathbf{u})$ is differentiable at all points except $\mathbf{u} = \mathbf{0}$, we have that for any $\mathbf{u}, \mathbf{w} \in \mathbb{R}^{d_i}$,

$$m'(\mathbf{u}; \mathbf{w}) = \lim_{t \downarrow 0} \frac{\|\mathbf{u} + t\mathbf{w}\| - \|\mathbf{u}\|}{t} = \begin{cases} \|\mathbf{w}\|, & \text{if } \mathbf{u} = \mathbf{0}, \\ \dfrac{\langle \mathbf{u}, \mathbf{w} \rangle}{\|\mathbf{u}\|}, & \text{if } \mathbf{u} \neq \mathbf{0}. \end{cases} \tag{2.5}$$

2.2 First-order d-stationary points of problem (1.1)

In this subsection, we use an explicit formula of directional derivative to provide sufficient and necessary conditions for first-order d-stationary points of problem (1.1).

Our analysis is based on a difference-of-convex (DC) form of the penalty function so that the directional derivative of the objective function in (1.1) can be explicitly expressed.

Assumption (A1): The penalty function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a DC function given by

$$\varphi(t) \triangleq g(t) - h(t), \text{ with } h(t) \triangleq \max_{1 \leq \nu \leq \bar{\nu}} \{h_\nu(t)\} \text{ for some integer } \bar{\nu} \geq 1, \tag{2.6}$$

where g and h_ν $(1 \leq \nu \leq \bar{\nu})$ are convex and differentiable in $t \in (0, \infty)$ with $g'(0) := g'(0+)$ and $h'_\nu(0) := h'_\nu(0+)$ for $1 \leq \nu \leq \bar{\nu}$.

Consequently, our group sparse optimization model (1.1) is rewritten as

$$\min_{\mathbf{x} \in \mathbb{R}^n} \ f(\mathbf{x}) := \mathcal{L}(\mathbf{x}) + \sum_{i=1}^{K} \big[g(\|\mathbf{x}_i\|) - h(\|\mathbf{x}_i\|) \big]. \tag{2.7}$$

From the literatures (e.g., [1,21]), we know that several folded concave penalty functions can be formulated as DC functions satisfying Assumption (A1), such as logarithm penalty, fraction penalty, CapL1, HTP, MCP and SCAD. In particular, as given in [1] we have the following expressions:

CapL1: $\varphi^{\text{CapL1}}(t) = g^{\text{CapL1}}(t) - h^{\text{CapL1}}(t)$ with

$$g^{\text{CapL1}}(t) = \frac{\lambda t}{\alpha}, \quad h^{\text{CapL1}}(t) = \max\left\{0, \frac{\lambda t}{\alpha} - \lambda\right\}, \quad (\alpha > 0, \lambda > 0);$$

MCP: $\varphi^{\text{MCP}}(t) = g^{\text{MCP}}(t) - h^{\text{MCP}}(t)$ with

$$g^{\text{MCP}}(t) = \lambda t, \quad h^{\text{MCP}}(t) = \begin{cases} \frac{t^2}{2\alpha}, & \text{if } 0 \le t \le \alpha\lambda, \\ \lambda t - \frac{\alpha\lambda^2}{2}, & \text{if } \quad t > \alpha\lambda, \end{cases} \quad (\alpha > 1, \lambda > 0);$$

SCAD: $\varphi^{\text{SCAD}}(t) = g^{\text{SCAD}}(t) - h^{\text{SCAD}}(t)$ with

$$g^{\text{SCAD}}(t) = \lambda t, \quad h^{\text{SCAD}}(t) = \begin{cases} 0, & \text{if } 0 \le t \le \lambda, \\ \frac{(t-\lambda)^2}{2(\alpha-1)}, & \text{if } \lambda < t \le \alpha\lambda, \\ \lambda t - \frac{(\alpha+1)\lambda^2}{2}, & \text{if } \quad t > \alpha\lambda, \end{cases} \quad (\alpha > 2, \lambda > 0).$$

Theorem 2.3 *Under Assumption (A1), the directional derivative of the objective function f in (1.1) has the following form*

$$\begin{aligned} f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) &= \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} g'(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ &\quad - \sum_{i=1}^{K} \max_{\nu_i \in \mathcal{A}_i(\widehat{\mathbf{x}}_i)} h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \end{aligned} \tag{2.8}$$

with $\mathcal{A}_i(\widehat{\mathbf{x}}_i) = \{\nu_i \in \{1, \cdots, \bar{\nu}\} : h_{\nu_i}(\|\widehat{\mathbf{x}}_i\|) = h(\|\widehat{\mathbf{x}}_i\|)\}$ *and*

$$m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = \begin{cases} \|\mathbf{x}_i\|, & \text{if } i \notin I(\widehat{\mathbf{x}}), \\ \frac{\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \rangle}{\|\widehat{\mathbf{x}}_i\|}, & \text{if } i \in I(\widehat{\mathbf{x}}_i). \end{cases} \tag{2.9}$$

Proof Under Assumption (A1), problem (1.1) can be written as (2.7). Since $h : \mathbb{R}_+ \to \mathbb{R}_+$ and $m : \mathbb{R}^{d_i} \to \mathbb{R}_+$ are both convex, $h \circ m : \mathbb{R}^{d_i} \to \mathbb{R}_+$ is directionally differentiable. According to the chain rule for directional derivatives and the differentiability of each h_ν, for $i = 1, \cdots, K$, we have

$$(h \circ m)'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = h'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) = \max_{\nu_i \in \mathcal{A}_i(\widehat{\mathbf{x}}_i)} h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|)\, m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i).$$

Since $\mathcal{L}$ and g are differentiable, we obtain the directional derivative at $\widehat{\mathbf{x}}$ for $\mathbf{x} - \widehat{\mathbf{x}}$ in (2.8). □

The following lemma shows that at any first-order d-stationary point of (1.1), the entries of the gradient of the loss function $\mathcal{L}$ for $i \in I(\widehat{\mathbf{x}})$ can be presented by the derivatives of g and h_{ν_i}.

Lemma 2.4 *Suppose Assumption (A1) holds. Let $\widehat{\mathbf{x}} \in \mathbb{R}^n$ be a first-order d-stationary point of problem (1.1). Then for $i \in I(\widehat{\mathbf{x}})$, we have*

$$[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(j)} = 0, \quad \forall\, j \notin J_i(\widehat{\mathbf{x}}), \tag{2.10}$$

and

$$|[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(j)}| = \frac{|g'(\|\widehat{\mathbf{x}}_i\|) - h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|)| \cdot |\widehat{x}_{i(j)}|}{\|\widehat{\mathbf{x}}_i\|}, \quad \forall\, j \in J_i(\widehat{\mathbf{x}}),\ \forall\, \nu_i \in \mathcal{A}_i(\widehat{\mathbf{x}}_i). \tag{2.11}$$

Proof From Theorem 2.3, we have

$$\begin{aligned}
&\langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} \left[g'(\|\widehat{\mathbf{x}}_i\|) - h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|) \right] m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\
\geq\ & \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} g'(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) - \sum_{i=1}^{K} \max_{\nu_i \in \mathcal{A}_i(\widehat{x}_i)} h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\
\geq\ & 0, \quad \forall\, \mathbf{x} \in \mathbb{R}^n, \qquad (2.12)
\end{aligned}$$

where $\nu_i \in \mathcal{A}_i(\widehat{\mathbf{x}}_i)$, $i = 1, \ldots, K$, and $m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$ is given by (2.9). It is obvious that inequality (2.12) also holds for any $\mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}}) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_i = \mathbf{0} \text{ whenever } i \notin I(\widehat{\mathbf{x}})\ \}$. This combining with formula (2.9) yields that

$$\sum_{i \in I(\widehat{\mathbf{x}})} \left\langle [\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i + \frac{\left[g'(\|\widehat{\mathbf{x}}_i\|) - h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|) \right]}{\|\widehat{\mathbf{x}}_i\|} \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \right\rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}}).$$

According to the arbitrariness of $\mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}})$, we obtain

$$[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i + \frac{g'(\|\widehat{\mathbf{x}}_i\|) - h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|)}{\|\widehat{\mathbf{x}}_i\|} \widehat{\mathbf{x}}_i = \mathbf{0}, \quad \forall\, i \in I(\widehat{\mathbf{x}}). \qquad (2.13)$$

Therefore, we have

$$[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(j)} = 0, \quad \forall\, i \in I(\widehat{\mathbf{x}}),\ j \notin J_i(\widehat{\mathbf{x}}),$$

and

$$\frac{|g'(\|\widehat{\mathbf{x}}_i\|) - h'_{\nu_i}(\|\widehat{\mathbf{x}}_i\|)| \cdot |\widehat{x}_{i(j)}|}{\|\widehat{\mathbf{x}}_i\|} = |[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_{i(j)}|, \quad \forall\, i \in I(\widehat{\mathbf{x}}),\ j \in J_i(\widehat{\mathbf{x}}).$$

The conclusion is obtained. □

By applying Lemma 2.4 to CapL1, MCP and SCAD, we can get the following lower bounds of the ℓ_2 vector norm of nonzero groups of first-order d-stationary points, whose proof is omitted.

Corollary 2.5 *Suppose there exists a nondecreasing function $C : \mathbb{R} \to \mathbb{R}_+$ such that $\|\nabla \mathcal{L}(\mathbf{x})\| \leq C(\mathcal{L}(\mathbf{x}))$ for any $\mathbf{x} \in \mathbb{R}^n$. Let $\widehat{\mathbf{x}} \in \mathbb{R}^n$ be a first-order d-stationary point of problem (1.1), and $\mathbf{x}^0 \in \mathbb{R}^n$ be a point such that $\mathcal{L}(\widehat{\mathbf{x}}) \leq \mathcal{L}(\mathbf{x}^0)$, then the following statements hold:*

(i) For CapL1, if $\frac{\lambda}{\alpha} > C(\mathcal{L}(\mathbf{x}^0))$, then either $\|\widehat{\mathbf{x}}_i\| = 0$ or $\|\widehat{\mathbf{x}}_i\| \geq \alpha$, $i = 1, \cdots, K$.

(ii) For MCP, if $\lambda > C(\mathcal{L}(\mathbf{x}^0))$, then either $\|\widehat{\mathbf{x}}_i\| = 0$ or $\|\widehat{\mathbf{x}}_i\| \geq \alpha\lambda - \alpha \cdot C(\mathcal{L}(\mathbf{x}^0)) > 0$, $i = 1, \cdots, K$.

(iii) For SCAD, if $\lambda > C(\mathcal{L}(\mathbf{x}^0))$, then either $\|\widehat{\mathbf{x}}_i\| = 0$ or $\|\widehat{\mathbf{x}}_i\| \geq \alpha\lambda - (\alpha - 1) \cdot C(\mathcal{L}(\mathbf{x}^0)) > \lambda$, $i = 1, \cdots, K$.

Remark 2.6 *The existence of the nondecreasing function $C : \mathbb{R} \to \mathbb{R}_+$ means that the norm of the gradient $\nabla \mathcal{L}(\mathbf{x})$ can be bounded by the function value $\mathcal{L}(\mathbf{x})$ via $C(\cdot)$. This condition can be easily satisfied, for example, for the square loss function $\mathcal{L}(\mathbf{x}) = \frac{1}{2m}\|A\mathbf{x} - b\|^2$, $C(t) = \|A\|_2 \sqrt{\frac{2}{m} t}$ meets the requirements since*

$$\|\nabla \mathcal{L}(\mathbf{x})\| = \frac{1}{m} \|A^\top (A\mathbf{x} - b)\| \leq \frac{\|A\|_2}{m} \|A\mathbf{x} - b\| = \|A\|_2 \sqrt{\frac{2}{m} \mathcal{L}(\mathbf{x})}.$$

When φ is the difference of two differentiable convex functions in $(0,\infty)$, such as φ^{MCP} and φ^{SCAD}, we have the following corollary, which will be used in Theorem 4.5 to derive the consistency of the second-order stationary point.

Corollary 2.7 *Suppose Assumption (A1) holds with $\bar{\nu}=1$, that is, $\varphi = g - h$ where g, h are both convex and differentiable in $(0,\infty)$. Let $\widehat{\mathbf{x}} \in \mathbb{R}^n$ be a first-order d-stationary point of problem (1.1), then the following statements hold:*

(i) $f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i, \mathbf{x}_i\rangle + \varphi'(0)\|\mathbf{x}_i\|\Big]$ *for any* $\mathbf{x}\in\mathbb{R}^n$.

(ii) $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| \le \varphi'(0)$ *whenever* $i \notin I(\widehat{\mathbf{x}})$.

(iii) $f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = 0$ *implies* $\mathbf{x}_i = \mathbf{0}$ *whenever* $i\notin I(\widehat{\mathbf{x}})$ *and* $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| < \varphi'(0)$.

Proof (i) From Theorem 2.3, $f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}})$ has the following form

$$\begin{aligned} f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = &\sum_{i\in I(\widehat{\mathbf{x}})} \left\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i + \frac{[g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|)]}{\|\widehat{\mathbf{x}}_i\|}\widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \right\rangle \\ &+ \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i, \mathbf{x}_i\rangle + (g'(0)-h'(0))\|\mathbf{x}_i\|\Big]. \end{aligned} \tag{2.14}$$

Since $\widehat{\mathbf{x}}$ is a first-order d-stationary point of problem (1.1), equation (2.13) holds with $h_{\nu_i} = h$. Hence (2.14) can be simplified as

$$f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i, \mathbf{x}_i\rangle + \varphi'(0)\|\mathbf{x}_i\|\Big],$$

where $\varphi'(0) = g'(0) - h'(0) > 0$.

(ii) Since $\widehat{\mathbf{x}}$ is a first-order d-stationary point of problem (1.1), $f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) \ge 0$ for all $\mathbf{x}\in\mathbb{R}^n$, that is,

$$f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i, \mathbf{x}_i\rangle + \varphi'(0)\|\mathbf{x}_i\|\Big] \ge 0, \quad \forall\, \mathbf{x}\in\mathbb{R}^n. \tag{2.15}$$

For each fixed $i\notin I(\widehat{\mathbf{x}})$, if we take $\check{\mathbf{x}}_i = -[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i$ and the other entries of $\check{\mathbf{x}}$ are all zeros, then we get

$$f'(\widehat{\mathbf{x}};\check{\mathbf{x}}-\widehat{\mathbf{x}}) = \|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| \cdot \Big[\varphi'(0) - \|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\|\Big] \ge 0. \tag{2.16}$$

If $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| = 0$, then $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| = 0 < \varphi'(0)$. If $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| > 0$, then from (2.16), we obtain $\varphi'(0) \ge \|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\|$.

(iii) It follows from (i), (ii) and Cauchy-Schwartz inequality that

$$\begin{aligned} f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) &= \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\langle [\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i, \mathbf{x}_i\rangle + \varphi'(0)\|\mathbf{x}_i\|\Big] \\ &\ge \sum_{i\notin I(\widehat{\mathbf{x}})} \Big[\varphi'(0) - \|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\|\Big]\|\mathbf{x}_i\| \ge 0. \end{aligned}$$

Hence, if $f'(\widehat{\mathbf{x}};\mathbf{x}-\widehat{\mathbf{x}}) = 0$, it must hold that $\|\mathbf{x}_i\| = 0$ whenever $i\notin I(\widehat{\mathbf{x}})$ and $\|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| < \varphi'(0)$. □

3 Second-order d-stationary points

In this section, we provide second-order optimality conditions for problem (1.1) using second-order directional derivatives.

3.1 Local optimality of second-order d-stationary points

Second-order directional derivatives for nonsmooth functions have been studied by many authors (e.g., see [2,7,27,31]) with different definitions for one direction or two directions. In this paper, we use the definition of the second-order directional derivative for one direction in [7,27] to define the second-order d-stationary point of problem (1.1). We show that second-order d-stationary points of problem (1.1) are local minimizers fulfilling the second-order growth condition under some mild conditions.

Definition 3.1 *[7,27] Let $\theta : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous and directionally differentiable function, and $\widehat{\mathbf{x}}, \mathbf{z} \in \mathbb{R}^n$. If the limit*

$$\lim_{\mathbf{y}\to\mathbf{z},\ t\downarrow 0} \frac{\theta(\widehat{\mathbf{x}} + t\mathbf{y}) - \theta(\widehat{\mathbf{x}}) - t\theta'(\widehat{\mathbf{x}}; \mathbf{y})}{\frac{1}{2}t^2} \tag{3.1}$$

exists, it is called the second-order directional derivative of θ at $\widehat{\mathbf{x}}$ for $\mathbf{z}$, denoted by $\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z})$. If for every $\mathbf{z} \in \mathbb{R}^n$, $\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z})$ exists, θ is called twice directionally differentiable at $\widehat{\mathbf{x}}$.

Indeed, to say that limit (3.1) exists and equals $\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z})$ is to say that whenever $\mathbf{x}^\nu$ converges to $\widehat{\mathbf{x}}$ from the direction of $\mathbf{z}$, in the sense that $[\mathbf{x}^\nu - \widehat{\mathbf{x}}]/t^\nu \to \mathbf{z}$ for some choice of $t^\nu \downarrow 0$, one has

$$\frac{\theta(\mathbf{x}^\nu) - \theta(\widehat{\mathbf{x}}) - \theta'(\widehat{\mathbf{x}}; \mathbf{x}^\nu - \widehat{\mathbf{x}})}{\frac{1}{2}(t^\nu)^2} \to \theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}).$$

Clearly, if limit (3.1) exists, then

$$\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}) = \lim_{t\downarrow 0} \frac{\theta(\widehat{\mathbf{x}} + t\mathbf{z}) - \theta(\widehat{\mathbf{x}}) - t\theta'(\widehat{\mathbf{x}}; \mathbf{z})}{\frac{1}{2}t^2}.$$

It is obvious that if θ is twice directionally differentiable at $\widehat{\mathbf{x}}$, then for any $\mathbf{z} \in \mathbb{R}^n$ there exists $\delta > 0$ such that

$$\theta(\widehat{\mathbf{x}} + t\mathbf{y}) = \theta(\widehat{\mathbf{x}}) + t\theta'(\widehat{\mathbf{x}}; \mathbf{y}) + \frac{1}{2}t^2\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}) + o(t^2), \quad \forall\ t \in (0, \delta) \text{ and } \forall\ \mathbf{y} \in \mathcal{N}(\mathbf{z}, \delta),$$

and particularly

$$\theta(\widehat{\mathbf{x}} + t\mathbf{z}) = \theta(\widehat{\mathbf{x}}) + t\theta'(\widehat{\mathbf{x}}; \mathbf{z}) + \frac{1}{2}t^2\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}) + o(t^2), \quad \forall\ t \in (0, \delta).$$

Moreover, if θ is twice differentiable at $\widehat{\mathbf{x}}$, then

$$\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}) = \langle \nabla^2\theta(\widehat{\mathbf{x}})\mathbf{z}, \mathbf{z}\rangle, \quad \forall\ \mathbf{z} \in \mathbb{R}^n.$$

From [7,27], we also know that if θ is convex and twice directionally differentiable at $\widehat{\mathbf{x}}$, then

$$\theta^{(2)}(\widehat{\mathbf{x}}; \mathbf{z}) \geq 0, \quad \forall\ \mathbf{z} \in \mathbb{R}^n.$$

For a vector-valued function $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ with component functions Φ_i for $i = 1, \cdots, m$, $\Phi^{(2)}(\mathbf{x}; \mathbf{z})$ is defined to be the m-vector with components $\Phi_i^{(2)}(\mathbf{x}; \mathbf{z})$ for $i = 1, \cdots, m$.

Lemma 3.2 *Let $\varrho : \mathbb{R}^m \to \mathbb{R}$ be locally Lipschitz continuous at $\Phi(\mathbf{x}) \in \mathbb{R}^m$, and $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous at $\mathbf{x} \in \mathbb{R}^n$, then the composite function $\theta = \varrho \circ \Phi : \mathbb{R}^n \to \mathbb{R}$ is twice directionally differentiable at $\mathbf{x}$ under either one of the following three conditions:*

*(a) ϱ is semismoothly differentiable at $\Phi(\mathbf{x})$ (**i.e.**, ϱ is differentiable near $\Phi(\mathbf{x})$ and $\nabla\varrho$ is semismooth at $\Phi(\mathbf{x})$), and Φ is twice directionally differentiable at $\mathbf{x}$.*

(b) ϱ is twice directionally differentiable at $\Phi(\mathbf{x})$ and Φ is piecewise affine near $\mathbf{x}$.

(c) ϱ is piecewise affine near $\Phi(\mathbf{x})$ and Φ is twice directionally differentiable at $\mathbf{x}$.

Moreover, we have, for all $\mathbf{z} \in \mathbb{R}^n$,

$$\theta^{(2)}(\mathbf{x};\mathbf{z}) = \Phi'(\mathbf{x};\mathbf{z})^\top (\nabla\varrho)'(\Phi(\mathbf{x});\Phi'(\mathbf{x};\mathbf{z})) + \nabla\varrho(\Phi(\mathbf{x}))^\top \Phi^{(2)}(\mathbf{x};\mathbf{z}), \text{ if (a) holds}; \tag{3.2}$$

$$\theta^{(2)}(\mathbf{x};\mathbf{z}) = \varrho^{(2)}(\Phi(\mathbf{x});\Phi'(\mathbf{x};\mathbf{z})), \text{ if (b) holds}; \tag{3.3}$$

and

$$\theta^{(2)}(\mathbf{x};\mathbf{z}) = \varrho'(\Phi(\mathbf{x});\Phi^{(2)}(\mathbf{x};\mathbf{z})), \text{ if (c) holds.} \tag{3.4}$$

Proof Conclusions (3.2) and (3.3) have been proved in [7, Prop. 3.2]. It is easy to prove conclusion (3.4) under condition (c) by noting that $\varrho'(\mathbf{u};\mathbf{v})$ exists and $\varrho^{(2)}(\mathbf{u};\mathbf{v}) = 0$ at any point $\mathbf{u}$ for any direction $\mathbf{v}$ when ϱ is piecewise affine. $\square$

Definition 3.3 *[7] Let $\theta : \mathbb{R}^n \to \mathbb{R}$ be twice directionally differentiable at $\widehat{\mathbf{x}} \in \mathbb{R}^n$. $\widehat{\mathbf{x}}$ is called a second-order d-stationary point of θ if $\widehat{\mathbf{x}}$ is a first-order d-stationary point of θ, and for any $\mathbf{z} \in \mathbb{R}^n$,*

$$\theta'(\widehat{\mathbf{x}};\mathbf{z}) = 0 \quad \textit{implies} \quad \theta^{(2)}(\widehat{\mathbf{x}};\mathbf{z}) \geq 0.$$

According to [7, Theorem 1] and [27, Theorem 13.24], if θ is twice directionally differentiable, then second-order d-stationary points of θ have the following locally optimal properties.

Proposition 3.4 *Let $\theta : \mathbb{R}^n \to \mathbb{R}$ be twice directionally differentiable at $\widehat{\mathbf{x}} \in \mathbb{R}^n$. The following two statements hold:*

(i) If $\widehat{\mathbf{x}} \in \mathbb{R}^n$ is a local minimizer of θ, then $\widehat{\mathbf{x}}$ is a second-order d-stationary point of θ.

*(ii) $\widehat{\mathbf{x}} \in \mathbb{R}^n$ is a **strong** local minimizer of θ, **i.e.**, there exist a neighborhood $\mathcal{W}$ of $\widehat{\mathbf{x}}$ and a scalar $\delta > 0$ such that*

$$\theta(\mathbf{x}) \geq \theta(\widehat{\mathbf{x}}) + \delta\|\mathbf{x} - \widehat{\mathbf{x}}\|^2, \quad \forall\, \mathbf{x} \in \mathcal{W},$$

if and only if $\widehat{\mathbf{x}}$ is a first-order d-stationary point of θ and satisfies that for any $\widehat{\mathbf{x}} \neq \mathbf{x} \in \mathbb{R}^n$,

$$\theta'(\widehat{\mathbf{x}};\mathbf{x} - \widehat{\mathbf{x}}) = 0 \quad \textit{implies} \quad \theta^{(2)}(\widehat{\mathbf{x}};\mathbf{x} - \widehat{\mathbf{x}}) > 0.$$

In the following parts, we will use the second-order directional derivative of ℓ_2 vector norm function. Recall that $m(\mathbf{u}) = \|\mathbf{u}\|$, and that

$$m'(\mathbf{u};\mathbf{v}) = \lim_{t\downarrow 0} \frac{\|\mathbf{u} + t\mathbf{v}\| - \|\mathbf{u}\|}{t} = \begin{cases} \|\mathbf{v}\|, \text{ if } \mathbf{u} = \mathbf{0}, \\ \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\|\mathbf{u}\|}, \text{ if } \mathbf{u} \neq \mathbf{0}, \end{cases} \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathbb{R}^{d_i}.$$

It is easy to know that $m(\cdot)$ is twice differentiable at all points except $\mathbf{u} = \mathbf{0}$, and that

$$\begin{aligned} m^{(2)}(\mathbf{u};\mathbf{w}) &= \lim_{\mathbf{v}\to\mathbf{w},\ t\downarrow 0} \frac{\|\mathbf{u} + t\mathbf{v}\| - \|\mathbf{u}\| - t m'(\mathbf{u};\mathbf{v})}{\frac{1}{2}t^2} \\ &= \begin{cases} 0, & \text{if } \mathbf{u} = \mathbf{0}, \\ \frac{(\|\mathbf{u}\|\|\mathbf{w}\|)^2 - |\langle \mathbf{u},\mathbf{w}\rangle|^2}{\|\mathbf{u}\|^3}, & \text{if } \mathbf{u} \neq \mathbf{0}, \end{cases} \quad \forall\, \mathbf{u}, \mathbf{w} \in \mathbb{R}^{d_i}. \end{aligned} \tag{3.5}$$

3.2 Second-order sufficient and necessary conditions for problem (1.1)

To study second-order d-stationary points of problem (1.1), we need the following assumption.

Assumption (A2) The penalty function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a DC function given by

$$\varphi(t) \triangleq g(t) - h(t), \tag{3.6}$$

where g is affine in $t \in [0, \infty)$ with $g'(0) := g'(0+)$, and h is convex and semismoothly differentiable in $t \in (0, \infty)$ with $h'(0) := h'(0+)$.

We can easily check that several folded concave penalty functions satisfy Assumption (A2), such as logarithm penalty, fraction penalty, HTP, MCP and SCAD.

In general the second-order directional derivative of a function is not easy to compute. The following lemma provides an explicit formula for computing the second-order directional derivative of the objective function of problem (1.1).

Lemma 3.5 *Under Assumption (A2), the second-order directional derivative of the objective function f in (1.1) has the following form*

$$\begin{aligned} f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) = \quad & \langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} \Big[g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|) \Big] m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ & - \sum_{i=1}^{K} m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) H'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)), \end{aligned} \tag{3.7}$$

where $m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$ is given by (2.9),

$$m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = \begin{cases} 0, & \text{if } i \notin I(\widehat{\mathbf{x}}), \\ \frac{(\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\| \|\widehat{\mathbf{x}}_i\|)^2 - |\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \rangle|^2}{\|\widehat{\mathbf{x}}_i\|^3}, & \text{if } i \in I(\widehat{\mathbf{x}}), \end{cases} \tag{3.8}$$

and $H(t) := h'(t)$ for any $t \in [0, +\infty)$.

Proof Since $\mathcal{L}$ is twice continuously differentiable, $\mathcal{L}^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) = \langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle$. Since g is affine in $[0, \infty)$ with $g'(0) = g'(0+)$, $(g \circ m)'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = g'(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$, $g^{(2)}(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) = 0$. By Lemma 3.2,

$$\begin{aligned} (g \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= g'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) + g^{(2)}(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ &= g'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \end{aligned}$$

for $i = 1, \cdots, K$.

Since h is semismoothly differentiable in $(0, \infty)$ with $h'(0) = h'(0+)$, h is twice directionally differentiable and $(h \circ m)'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = h'(\|\widehat{\mathbf{x}}_i\|) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$. By Lemma 3.2,

$$\begin{aligned} (h \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= h'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) + h^{(2)}(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ &= h'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) + H'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \end{aligned}$$

for $i = 1, \cdots, K$.

Then we have

$$\begin{aligned} f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) = & \langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} \Big[g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|) \Big] m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \\ & - \sum_{i=1}^{K} m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) H'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)), \end{aligned}$$

where $m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$ and $m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)$ are given by (2.5) and (3.5) respectively. $\square$

From Definition 3.3, Proposition 3.4 and Lemma 3.5, we obtain the following theorem.

Theorem 3.6 *Suppose Assumption (A2) holds and* $\widehat{\mathbf{x}} \in \mathbb{R}^n$ *is a first-order d-stationary point of problem (1.1), then the following two statements hold with* $f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}})$ *and* $f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}})$ *given by (2.8) and (3.7) respectively.*

(i) $\widehat{\mathbf{x}}$ *is a second-order d-stationary point of problem (1.1) if and only if for any* $\mathbf{x} \in \mathbb{R}^n$, $f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) = 0$ *implies* $f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) \geq 0$.

(ii) $\widehat{\mathbf{x}}$ *is a strong local minimizer of problem (1.1) if and only if for any* $\widehat{\mathbf{x}} \neq \mathbf{x} \in \mathbb{R}^n$, $f'(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) = 0$ *implies* $f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x} - \widehat{\mathbf{x}}) > 0$.

The following theorem shows that the second-order directional derivative at a second-order d-stationary point can be simplified and is nonnegative on a special set.

Theorem 3.7 *Under Assumption (A2), let* $\widehat{\mathbf{x}} \in \mathbb{R}^n$ *be a second-order d-stationary point of problem (1.1), and*

$$\mathcal{X}(\widehat{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_i = \mathbf{0} \text{ whenever } i \notin I(\widehat{\mathbf{x}})\}, \tag{3.9}$$

then for any $\mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}})$,

$$\langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i \in I(\widehat{\mathbf{x}})} \Big[(g \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) - (h \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \Big] \geq 0,$$

where for $i \in I(\widehat{\mathbf{x}})$,

$$\begin{aligned}
(g \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= g'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i), \\
(h \circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) H'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i)) \\
&\quad + h'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i), \\
m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= \frac{\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \rangle}{\|\widehat{\mathbf{x}}_i\|}, \\
m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) &= \frac{(\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\| \|\widehat{\mathbf{x}}_i\|)^2 - |\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \rangle|^2}{\|\widehat{\mathbf{x}}_i\|^3}, \\
H(t) &= h'(t) \text{ for any } t \in (0, \infty).
\end{aligned}$$

Proof Since $\widehat{\mathbf{x}}$ is a second-order d-stationary point of problem (1.1), it is also a first-order d-stationary point of problem (1.1), which means

$$\langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} \big(g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|) \big) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) \geq 0, \quad \forall\, \mathbf{x} \in \mathbb{R}^n.$$

By the same argument in the proof of (2.13), we have

$$[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i + \frac{g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|)}{\|\widehat{\mathbf{x}}_i\|} \widehat{\mathbf{x}}_i = \mathbf{0}, \quad \forall\, i \in I(\widehat{\mathbf{x}}).$$

Therefore, we get

$$\sum_{i \in I(\widehat{\mathbf{x}})} \left\langle [\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i + \frac{g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|)}{\|\widehat{\mathbf{x}}_i\|} \widehat{\mathbf{x}}_i, \mathbf{x}_i - \widehat{\mathbf{x}}_i \right\rangle = 0, \quad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{3.10}$$

For any $\mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}})$, by (2.9), (3.10) and direct computation, we obtain

$$\langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{x} - \widehat{\mathbf{x}} \rangle + \sum_{i=1}^{K} \big(g'(\|\widehat{\mathbf{x}}_i\|) - h'(\|\widehat{\mathbf{x}}_i\|) \big) m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i - \widehat{\mathbf{x}}_i) = 0,$$

that is, $f'(\widehat{\mathbf{x}}; \mathbf{x}-\widehat{\mathbf{x}})=0$, which together with that $\widehat{\mathbf{x}}$ is a second-order d-stationary point of problem (1.1) yields that $f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x}-\widehat{\mathbf{x}}) \geq 0$. From (2.9), (3.8) and $\mathbf{x} \in \mathcal{X}(\widehat{\mathbf{x}})$, we have that for $i \notin I(\widehat{\mathbf{x}})$, $m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) = m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i)=0$, and that for $i \in I(\widehat{\mathbf{x}})$,

$$\begin{aligned} m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) &= \frac{\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i-\widehat{\mathbf{x}}_i\rangle}{\|\widehat{\mathbf{x}}_i\|}, \\ m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) &= \frac{(\|\mathbf{x}_i-\widehat{\mathbf{x}}_i\|\|\widehat{\mathbf{x}}_i\|)^2 - |\langle \widehat{\mathbf{x}}_i, \mathbf{x}_i-\widehat{\mathbf{x}}_i\rangle|^2}{\|\widehat{\mathbf{x}}_i\|^3}, \\ (g\circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) &= g'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i), \\ (h\circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) &= m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) H'(\|\widehat{\mathbf{x}}_i\|; m'(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i)) \\ &\quad + h'(\|\widehat{\mathbf{x}}_i\|) m^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i), \end{aligned}$$

where $H(t) = h'(t)$ for any $t \in (0,\infty)$. Hence we get

$$\begin{aligned} &\langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x}-\widehat{\mathbf{x}}), \mathbf{x}-\widehat{\mathbf{x}}\rangle + \sum_{i\in I(\widehat{\mathbf{x}})} \Big[(g\circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) - (h\circ m)^{(2)}(\widehat{\mathbf{x}}_i; \mathbf{x}_i-\widehat{\mathbf{x}}_i) \Big] \\ &= f^{(2)}(\widehat{\mathbf{x}}; \mathbf{x}-\widehat{\mathbf{x}}) \geq 0. \end{aligned}$$

The proof is finished. □

3.3 Lower bound theory of second-order d-stationary points

In this subsection, we analyze the lower bound of the ℓ_2 vector norm of nonzero groups of second-order d-stationary points of problem (1.1). We will see that the second-order lower bounds are tighter than the corresponding first-order lower bounds. At first, we give a useful lemma which provides an upper bound for the second-order directional derivative of the penalty function h at any second-order d-stationary point.

Lemma 3.8 *Under Assumption (A2), let $\widehat{\mathbf{x}} \in \mathbb{R}^n$ be a second-order d-stationary point of problem (1.1), then*

$$\langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_i\rangle \geq \|\widehat{\mathbf{x}}_i\|^2 \cdot \max\{H'(\|\widehat{\mathbf{x}}_i\|; 1), -H'(\|\widehat{\mathbf{x}}_i\|; -1)\}, \quad \forall\, i \in I(\widehat{\mathbf{x}}),$$

where $\nabla_i^2 \mathcal{L}(\mathbf{x})$ denotes the principal submatrix of $\nabla^2 \mathcal{L}(\mathbf{x})$ corresponding to the group $\mathbf{x}_i$.

Proof For each fixed $i \in I(\widehat{\mathbf{x}})$, let $\mathbf{x}^1, \mathbf{x}^2 \in \mathbb{R}^n$ be taken as

$$\mathbf{x}^1_{i'} = \begin{cases} 2\widehat{\mathbf{x}}_i, & \text{if } i'=i, \\ \widehat{\mathbf{x}}_{i'}, & \text{if } i' \neq i, \end{cases} \qquad \mathbf{x}^2_{i'} = \begin{cases} \mathbf{0}, & \text{if } i'=i, \\ \widehat{\mathbf{x}}_{i'}, & \text{if } i' \neq i. \end{cases}$$

Then it is easy to check that $\mathbf{x}^1, \mathbf{x}^2 \in \mathcal{X}(\widehat{\mathbf{x}})$ which has been defined by (3.9). By Theorem 3.7, we have

$$\begin{aligned} &\langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x}^\eta-\widehat{\mathbf{x}}), \mathbf{x}^\eta-\widehat{\mathbf{x}}\rangle + \sum_{i'\in I(\widehat{\mathbf{x}})} \Big[(g\circ m)^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'}-\widehat{\mathbf{x}}_{i'}) - (h\circ m)^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'}-\widehat{\mathbf{x}}_{i'}) \Big] \\ &\geq 0, \quad \eta = 1,2, \end{aligned} \tag{3.11}$$

where, according to the definitions of $\mathbf{x}^1$ and $\mathbf{x}^2$ as well as formulas (2.9) and (3.8),

$$\begin{aligned}
\langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}})(\mathbf{x}^\eta - \widehat{\mathbf{x}}), \mathbf{x}^\eta - \widehat{\mathbf{x}}\rangle &= \langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_i\rangle, \quad \eta = 1,2,\\
m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^1_{i'} - \widehat{\mathbf{x}}_{i'}) &= \begin{cases} \|\widehat{\mathbf{x}}_i\|, & \text{if } i' = i,\\ 0, & \text{if } i' \neq i,\end{cases} \quad m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^2_{i'} - \widehat{\mathbf{x}}_{i'}) = \begin{cases} -\|\widehat{\mathbf{x}}_i\|, & \text{if } i' = i,\\ 0, & \text{if } i' \neq i,\end{cases}\\
m^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) &= 0, \quad \forall\ i' = 1, \cdots, K, \quad \eta = 1,2,\\
(g \circ m)^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) &= g'(\|\widehat{\mathbf{x}}_{i'}\|) m^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) = 0, \quad \eta = 1,2,\\
(h \circ m)^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) &= m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) H'(\|\widehat{\mathbf{x}}_{i'}\|; m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}))\\
&\quad + h'(\|\widehat{\mathbf{x}}_{i'}\|) m^{(2)}(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'})\\
&= m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'}) H'(\|\widehat{\mathbf{x}}_{i'}\|; m'(\widehat{\mathbf{x}}_{i'}; \mathbf{x}^\eta_{i'} - \widehat{\mathbf{x}}_{i'})), \quad \eta = 1,2,\\
H(t) &= h'(t) \text{ for any } t \in (0, \infty).
\end{aligned}$$

Therefore, by taking the above terms into inequality (3.11), we get

$$\langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}})\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_i\rangle \geq \|\widehat{\mathbf{x}}_i\| \cdot \max\{H'(\|\widehat{\mathbf{x}}_i\|; \|\widehat{\mathbf{x}}_i\|), -H'(\|\widehat{\mathbf{x}}_i\|; -\|\widehat{\mathbf{x}}_i\|)\}, \quad \forall\ i \in I(\widehat{\mathbf{x}}).$$

By the positive homogeneity of $H'(\|\widehat{\mathbf{x}}_i\|; \cdot)$ and $\|\widehat{\mathbf{x}}_i\| > 0$, we derive the desired result. □

Theorem 3.9 *Suppose Assumption (A2) holds and there exists $M > 0$ such that $\|\nabla^2 \mathcal{L}(\mathbf{x})\|_2 \leq M$ for all $\mathbf{x} \in \mathbb{R}^n$. Let $\widehat{\mathbf{x}} \in \mathbb{R}^n$ be a second-order d-stationary point of problem (1.1), then the following statements hold:*

(i) For MCP, if $M < \frac{1}{\alpha}$, then either $\|\widehat{\mathbf{x}}_i\| = 0$ or $\|\widehat{\mathbf{x}}_i\| > \alpha\lambda$, $i = 1, \cdots, K$.

(ii) For SCAD, if $M < \frac{1}{\alpha - 1}$, then either $\|\widehat{\mathbf{x}}_i\| < \lambda$ or $\|\widehat{\mathbf{x}}_i\| > \alpha\lambda$, $i = 1, \cdots, K$.

(iii)For SCAD, suppose, in addition, there exists a nondecreasing function $C : \mathbb{R} \to \mathbb{R}_+$ such that $\|\nabla \mathcal{L}(\mathbf{x})\| \leq C(\mathcal{L}(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^n$. If there exists $\mathbf{x}^0 \in \mathbb{R}^n$ satisfying $\mathcal{L}(\mathbf{x}^0) \geq \mathcal{L}(\widehat{\mathbf{x}})$, $\varphi'(0) > C(\mathcal{L}(\mathbf{x}^0))$, and $\frac{1}{\alpha-1} > M$, then either $\|\widehat{\mathbf{x}}_i\| = 0$ or $\|\widehat{\mathbf{x}}_i\| > \alpha\lambda$, $i = 1, \cdots, K$.

Proof Since $\|\nabla^2 \mathcal{L}(\mathbf{x})\|_2 \leq M$ for all $\mathbf{x} \in \mathbb{R}^n$, we have

$$\langle \nabla^2 \mathcal{L}(\mathbf{x})\mathbf{z}, \mathbf{z}\rangle \leq M\|\mathbf{z}\|^2, \quad \forall\ \mathbf{x}, \mathbf{z} \in \mathbb{R}^n. \tag{3.12}$$

(i) For MCP: recall that $H^{\mathrm{MCP}}(t) = (h^{\mathrm{MCP}})'(t) = \begin{cases} \frac{t}{\alpha}, & \text{if } 0 \leq t \leq \alpha\lambda,\\ \lambda, & \text{if } t > \alpha\lambda,\end{cases}$ we have

$$(H^{\mathrm{MCP}})'(t; 1) = \begin{cases} \frac{1}{\alpha}, & \text{if } 0 \leq t < \alpha\lambda,\\ 0, & \text{if } t \geq \alpha\lambda,\end{cases} \quad (H^{\mathrm{MCP}})'(t; -1) = \begin{cases} -\frac{1}{\alpha}, & \text{if } 0 < t \leq \alpha\lambda,\\ 0, & \text{if } t > \alpha\lambda.\end{cases}$$

Assume, on the contrary, that $0 < \|\widehat{\mathbf{x}}_i\| \leq \alpha\lambda$, then

$$(H^{\mathrm{MCP}})'(\|\widehat{\mathbf{x}}_i\|; 1) \leq -(H^{\mathrm{MCP}})'(\|\widehat{\mathbf{x}}_i\|; -1) = \frac{1}{\alpha}.$$

From Lemma 3.8 and (3.12), we have

$$M \geq -(H^{\mathrm{MCP}})'(\|\widehat{\mathbf{x}}_i\|; -1) = \frac{1}{\alpha},$$

which contradicts the condition $M < \frac{1}{\alpha}$. Therefore, we have $\|\widehat{\mathbf{x}}_i\| > \alpha\lambda$ for any $i \in I(\widehat{\mathbf{x}})$.

(ii) For SCAD: recall that $H^{\rm SCAD}(t) = (h^{\rm SCAD})'(t) = \begin{cases} 0, & \text{if } 0 \le t \le \lambda, \\ \frac{t-\lambda}{\alpha-1}, & \text{if } \lambda < t \le \alpha\lambda, \\ \lambda, & \text{if } \quad t > \alpha\lambda, \end{cases}$ we have

$$(H^{\rm SCAD})'(t;1) = \begin{cases} 0, & \text{if } t \in [0,\lambda) \cup [\alpha\lambda, +\infty), \\ \frac{1}{\alpha-1}, & \text{if } t \in [\lambda, \alpha\lambda), \end{cases}$$
$$(H^{\rm SCAD})'(t;-1) = \begin{cases} 0, & \text{if } t \in (0,\lambda] \cup (\alpha\lambda, +\infty), \\ -\frac{1}{\alpha-1}, & \text{if } t \in (\lambda, \alpha\lambda]. \end{cases}$$

Assume, on the contrary, that $\lambda \le \|\widehat{\mathbf{x}}_i\| \le \alpha\lambda$, then

$$\max\{(H^{\rm SCAD})'(\|\widehat{\mathbf{x}}_i\|;1), -(H^{\rm SCAD})'(\|\widehat{\mathbf{x}}_i\|;-1)\} = \frac{1}{\alpha-1}.$$

From Lemma 3.8 and (3.12), we have

$$M \ge \frac{1}{\alpha-1},$$

which contradicts the condition $M < \frac{1}{\alpha-1}$. Therefore, we have either $\|\widehat{\mathbf{x}}_i\| < \lambda$ or $\|\widehat{\mathbf{x}}_i\| > \alpha\lambda$.

(iii) Since $\widehat{\mathbf{x}}$ is a second-order d-stationary point of problem (1.1), it is also a first-order d-stationary point of problem (1.1). Combining (ii) with Corollary 2.5 (iii), we derive the desired result. □

Remark 3.10 *The condition in Theorem 3.9 means that the operator $\nabla^2\mathcal{L}(\mathbf{x})$ has an uniform bound M on $\mathbb{R}^n$. We can easily check that $\mathcal{L}(\mathbf{x}) = \frac{1}{2m}\|A\mathbf{x}-b\|^2$ satisfies this condition since $\|\nabla^2\mathcal{L}(\mathbf{x})\|_2 = \frac{\|A^\top A\|_2}{m} = \frac{\|A\|_2^2}{m}$.*

4 Smoothing functions and consistency of stationary points

As we have seen, first-order and second-order d-stationary points have good locally optimal properties. How to compute such points is an interesting and challenging problem. Smooth approximations are widely used in optimization and scientific computing, e.g., see [8,9,10]. In this section, we construct a twice continuously differentiable smoothing function of the objective function f of problem (1.1), and show that the first-order and second-order d-stationary points of problem (1.1) can be obtained via the first-order and second-order stationary points of the smoothing problem. We should notice that in problem (1.1), the term $\varphi(\|\widehat{\mathbf{x}}_i\|)$ is a composite of two nonsmooth functions φ and $\|\cdot\|$. Using the special structure of these two functions, our smoothing function can be easily constructed.

For $\mu \in (0,\infty)$ and $m(\mathbf{u}) = \|\mathbf{u}\|$, let

$$\widetilde{m}_\mu(\mathbf{u}) = \sqrt{\|\mathbf{u}\|^2 + \mu}, \quad \forall\ \mathbf{u} \in \mathbb{R}^{d_i}, \tag{4.1}$$

then $\widetilde{m}_\mu(\mathbf{u})$ is always positive and twice continuously differentiable with

$$\nabla\widetilde{m}_\mu(\mathbf{u}) = \frac{\mathbf{u}}{\sqrt{\|\mathbf{u}\|^2+\mu}}, \quad \nabla^2\widetilde{m}_\mu(\mathbf{u}) = \frac{(\|\mathbf{u}\|^2+\mu)\mathbf{I} - \mathbf{u}\mathbf{u}^\top}{(\|\mathbf{u}\|^2+\mu)^{3/2}}, \tag{4.2}$$

and

$$0 < \widetilde{m}_\mu(\mathbf{u}) - m(\mathbf{u}) = \sqrt{\|\mathbf{u}\|^2+\mu} - \|\mathbf{u}\| \le \mu^{\frac{1}{2}}, \tag{4.3}$$

where $\mathbf{I}$ denotes the identity matrix. One can also check that $\widetilde{m}_\mu(\mathbf{u})$ satisfies the following three properties:

(i) $\lim\limits_{\mathbf{v}\to\mathbf{u},\mu\downarrow 0} \widetilde{m}_\mu(\mathbf{v}) = m(\mathbf{u})$ for all $\mathbf{u}\in\mathbb{R}^{d_i}$;

(ii) (Consistency or weak consistency of directional derivatives)

$$\lim_{\mathbf{v}\to\mathbf{u},\mu\downarrow 0}\langle\nabla\widetilde{m}_\mu(\mathbf{v}),\mathbf{w}\rangle = \langle\nabla m(\mathbf{u}),\mathbf{w}\rangle = m'(\mathbf{u};\mathbf{w}),\quad \forall\ \mathbf{0}\neq\mathbf{u}\in\mathbb{R}^{d_i},\ \forall\ \mathbf{w}\in\mathbb{R}^{d_i},\tag{4.4}$$

$$\begin{aligned}\limsup_{\mathbf{v}\to\mathbf{0},\mu\downarrow 0}\langle\nabla\widetilde{m}_\mu(\mathbf{v}),\mathbf{w}\rangle &= \limsup_{\mathbf{v}\to\mathbf{0},\mu\downarrow 0}\frac{\langle\mathbf{v},\mathbf{w}\rangle}{\sqrt{\|\mathbf{v}\|^2+\mu}} = \limsup_{t\downarrow 0,\mu\downarrow 0}\frac{t\|\mathbf{w}\|^2}{\sqrt{t^2\|\mathbf{w}\|^2+\mu}}\\ &= \limsup_{t\downarrow 0,\mu\downarrow 0}\frac{\|\mathbf{w}\|^2}{\sqrt{\|\mathbf{w}\|^2+\frac{\mu}{t^2}}} = \|\mathbf{w}\| = m'(\mathbf{0};\mathbf{w}),\ \forall\ \mathbf{w}\in\mathbb{R}^{d_i};\end{aligned}\tag{4.5}$$

(iii) (Consistency or weak consistency of second-order directional derivatives)

$$\begin{aligned}\lim_{\mathbf{v}\to\mathbf{u},\mu\downarrow 0}\langle\nabla^2\widetilde{m}_\mu(\mathbf{v})\mathbf{w},\mathbf{w}\rangle &= \langle\nabla^2 m(\mathbf{u})\mathbf{w},\mathbf{w}\rangle\\ &= m^{(2)}(\mathbf{u};\mathbf{w}),\quad\forall\ \mathbf{0}\neq\mathbf{u}\in\mathbb{R}^{d_i},\ \forall\ \mathbf{w}\in\mathbb{R}^{d_i},\end{aligned}\tag{4.6}$$

$$\begin{aligned}\liminf_{\mathbf{v}\to\mathbf{0},\mu\downarrow 0}\langle\nabla^2\widetilde{m}_\mu(\mathbf{v})\mathbf{w},\mathbf{w}\rangle &= \liminf_{\mathbf{v}\to\mathbf{0},\mu\downarrow 0}\frac{\|\mathbf{v}\|^2\|\mathbf{w}\|^2-(\mathbf{v}^\top\mathbf{w})^2+\mu\|\mathbf{w}\|^2}{(\|\mathbf{v}\|^2+\mu)^{3/2}}\\ &= \liminf_{\mathbf{v}\to\mathbf{0},\mu\downarrow 0}\frac{\|\mathbf{w}\|^2}{\left(\frac{\|\mathbf{v}\|^2}{\mu^{2/3}}+\mu^{1/3}\right)^{3/2}}\\ &= 0 = m^{(2)}(\mathbf{0};\mathbf{w}),\quad\forall\ \mathbf{w}\in\mathbb{R}^{d_i}.\end{aligned}\tag{4.7}$$

Under Assumption (A2), h is semismoothly differentiable in $(0,\infty)$. If h is not twice continuously differentiable in $(0,\infty)$, for each $\mu>0$, let $\widetilde{h}_\mu$ be a twice continuously differentiable function in $(0,\infty)$ such that

$$\lim_{s\to t,\mu\downarrow 0}\widetilde{h}_\mu(s) = h(t),\quad \lim_{s\to t,\mu\downarrow 0}\widetilde{h}'_\mu(s) = h'(t),\ \lim_{s\downarrow 0,\mu\downarrow 0}\widetilde{h}'_\mu(s) = h'(0+),\tag{4.8}$$

$$\liminf_{s\to t,\mu\downarrow 0}\widetilde{h}''_\mu(s) = \min\{H'(t;1),-H'(t;-1)\},\ \text{and}\tag{4.9}$$

$$\limsup_{s\to t,\mu\downarrow 0}\widetilde{h}''_\mu(s) = \max\{H'(t;1),-H'(t;-1)\}.\tag{4.10}$$

Note that if h is twice continuously differentiable at $t>0$, then $H'(t;1) = -H'(t;-1) = h''(t)$.

For example, in MCP,

$$h^{\mathrm{MCP}}(t) = \begin{cases}\frac{t^2}{2\alpha}, & \text{if } 0\le t\le\alpha\lambda,\\ \lambda t-\frac{\alpha\lambda^2}{2}, & \text{if } \quad t>\alpha\lambda,\end{cases} = \lambda t-\lambda\int_0^t\left(1-\frac{\tau}{\alpha\lambda}\right)_+ \mathrm{d}\tau\quad(\alpha>1,\lambda>0).$$

Let

$$\widetilde{h}^{\mathrm{MCP}}_\mu(t) = \lambda t-\frac{\lambda}{2}\int_0^t\left[\left(\left(1-\frac{\tau}{\alpha\lambda}\right)^2+\mu\right)^{1/2}+\left(1-\frac{\tau}{\alpha\lambda}\right)\right]\mathrm{d}\tau,\tag{4.11}$$

then one can check that for each $\mu > 0$, $\widetilde{h}_\mu^{\mathrm{MCP}}$ is twice continuously differentiable in $t \in (0, \infty)$ with

$$(\widetilde{h}_\mu^{\mathrm{MCP}})'(t) = \lambda - \frac{\lambda}{2}\left[\left(\left(1-\frac{t}{\alpha\lambda}\right)^2+\mu\right)^{1/2} + \left(1-\frac{t}{\alpha\lambda}\right)\right],$$

$$(\widetilde{h}_\mu^{\mathrm{MCP}})''(t) = \frac{1}{2\alpha}\left[\frac{1-\frac{t}{\alpha\lambda}}{\sqrt{\left(1-\frac{t}{\alpha\lambda}\right)^2+\mu}}+1\right],$$

and satisfies the following three properties:

(i) $\lim_{s\to t,\mu\downarrow 0} \widetilde{h}_\mu^{\mathrm{MCP}}(s) = h^{\mathrm{MCP}}(t)$ for all $t \in [0,\infty)$;

(ii) $\lim_{s\to t,\mu\downarrow 0} (\widetilde{h}_\mu^{\mathrm{MCP}})'(s) = (h^{\mathrm{MCP}})'(t)$ for all $t \in (0,\infty)$, and $\lim_{s\downarrow 0,\mu\downarrow 0} (\widetilde{h}_\mu^{\mathrm{MCP}})'(s) = (h^{\mathrm{MCP}})'(0+)$;

(iii) For any $t \in (0, \alpha\lambda)\bigcup(\alpha\lambda, \infty)$,

$$\begin{aligned}\lim_{s\to t,\mu\downarrow 0}(\widetilde{h}_\mu^{\mathrm{MCP}})''(s) &= \lim_{s\to t,\mu\downarrow 0}\frac{1}{2\alpha}\left[\frac{1-\frac{s}{\alpha\lambda}}{\sqrt{\left(1-\frac{s}{\alpha\lambda}\right)^2+\mu}}+1\right]\\ &= \lim_{s\to t,\mu\downarrow 0}\frac{1}{2\alpha}\left[\frac{\mathrm{sign}(1-\frac{s}{\alpha\lambda})}{\sqrt{1+\frac{\mu}{(1-\frac{s}{\alpha\lambda})^2}}}+1\right] = \begin{cases}\frac{1}{\alpha}, & \text{if } 0<t<\alpha\lambda,\\ 0, & \text{if } \quad t>\alpha\lambda,\end{cases}\\ &= (H^{\mathrm{MCP}})'(t;1) = -(H^{\mathrm{MCP}})'(t;-1) = (h^{\mathrm{MCP}})''(t);\end{aligned}$$

for $t = \alpha\lambda$,

$$\liminf_{s\to t,\mu\downarrow 0}(\widetilde{h}_\mu^{\mathrm{MCP}})''(s) = \liminf_{s\to t,\mu\downarrow 0}\frac{1}{2\alpha}\left[\frac{\mathrm{sign}(1-\frac{s}{\alpha\lambda})}{\sqrt{1+\frac{\mu}{(1-\frac{s}{\alpha\lambda})^2}}}+1\right] = 0 = (H^{\mathrm{MCP}})'(t;1), \text{and}$$

$$\limsup_{s\to t,\mu\downarrow 0}(\widetilde{h}_\mu^{\mathrm{MCP}})''(s) = \limsup_{s\to t,\mu\downarrow 0}\frac{1}{2\alpha}\left[\frac{\mathrm{sign}(1-\frac{s}{\alpha\lambda})}{\sqrt{1+\frac{\mu}{(1-\frac{s}{\alpha\lambda})^2}}}+1\right] = \frac{1}{\alpha} = -(H^{\mathrm{MCP}})'(t;-1).$$

Now, under Assumption (A2), we have a twice continuously differentiable approximation $\widetilde{f}_\mu(\mathbf{x})$ of the objective function $f(\mathbf{x})$ in problem (1.1),

$$\widetilde{f}_\mu(\mathbf{x}) = \mathcal{L}(\mathbf{x}) + \sum_{i=1}^{K}\Big[g\circ\widetilde{m}_\mu(\mathbf{x}_i) - \widetilde{h}_\mu\circ\widetilde{m}_\mu(\mathbf{x}_i)\Big],$$

with $\lim_{\mathbf{z}\to\mathbf{x},\mu\downarrow 0}\widetilde{f}_\mu(\mathbf{z}) = f(\mathbf{x})$ for any $\mathbf{x}\in\mathbb{R}^n$. It should be noted that although $g(\|\cdot\|)-\widetilde{h}_\mu(\|\cdot\|)$ is not differentiable at $\mathbf{x}_i = \mathbf{0}$, $g\circ\widetilde{m}_\mu(\cdot)-\widetilde{h}_\mu\circ\widetilde{m}_\mu(\cdot)$ is twice continuously differentiable at any point $\mathbf{x}_i\in\mathbb{R}^{d_i}$ since $\widetilde{m}_\mu(\mathbf{x}_i)$ is always strictly positive for any $\mu>0$. Consequently, $\widetilde{f}_\mu(\cdot)$ is twice continuously differentiable at any point $\mathbf{x}\in\mathbb{R}^n$. Thus we obtain a twice continuously differentiable optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n}\ \widetilde{f}_\mu(\mathbf{x}). \tag{4.12}$$

By the standard definitions for twice differentiable optimization problems, $\widehat{\mathbf{x}}^\mu$ is called a first-order stationary point of problem (4.12) if $\nabla\widetilde{f}_\mu(\widehat{\mathbf{x}}^\mu) = \mathbf{0}$; and $\widehat{\mathbf{x}}^\mu$ is called a second-order stationary point of problem (4.12) if $\nabla\widetilde{f}_\mu(\widehat{\mathbf{x}}^\mu) = \mathbf{0}$ and

$$\langle\nabla^2\widetilde{f}_\mu(\widehat{\mathbf{x}}^\mu)\mathbf{z},\mathbf{z}\rangle \ge 0, \quad \forall\ \mathbf{z}\in\mathbb{R}^n. \tag{4.13}$$

Let $\{\widehat{\mathbf{x}}^{\mu_k}\}$ denote a sequence of first-order or second-order stationary points of problem (4.12) with $\mu_k > 0,\ k = 1, 2, \cdots$, and $\mu_k \to 0$ as $k \to \infty$. We will investigate the accumulation points of $\{\widehat{\mathbf{x}}^{\mu_k}\}$.

Theorem 4.1 *(Consistency of first-order stationary points) Suppose Assumption (A2) holds. Let $\{\widehat{\mathbf{x}}^{\mu_k}\}$ be a sequence of first-order stationary points of problem (4.12) with $\mu = \mu_k$. Then any accumulation point of $\{\widehat{\mathbf{x}}^{\mu_k}\}$ is a first-order d-stationary point of problem (1.1).*

Proof Let $\widehat{\mathbf{x}}$ be an accumulation point of $\{\widehat{\mathbf{x}}^{\mu_k}\}$. Without loss of generality, we may assume that $\{\widehat{\mathbf{x}}^{\mu_k}\}$ converges to $\widehat{\mathbf{x}}$.

Since $\widehat{\mathbf{x}}^{\mu_k}$ is a first-order stationary point of problem (4.12) with $\mu = \mu_k$, then

$$\nabla \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}) = \nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}) + \begin{pmatrix} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k}) - \widetilde{h}'_\mu \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k})\right] \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k}) \\ \vdots \\ \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k}) - \widetilde{h}'_\mu \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k})\right] \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k}) \end{pmatrix} = \mathbf{0}.$$

Therefore, for any $\mathbf{z} \in \mathbb{R}^n$ we have

$$\begin{aligned} 0 &= \langle \nabla \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}), \mathbf{z} \rangle \\ &= \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}), \mathbf{z} \rangle + \sum_{i=1}^{K} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle. \end{aligned} \tag{4.14}$$

Let $k \to \infty$, then we get $\mu_k \to 0$ and $\widehat{\mathbf{x}}^{\mu_k} \to \widehat{\mathbf{x}}$, consequently, $\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \to m(\widehat{\mathbf{x}}_i)$, $g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \to g' \circ m(\widehat{\mathbf{x}}_i)$ and $\widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \to h' \circ m(\widehat{\mathbf{x}}_i)$. Moreover, from (4.4) and (4.5), we have

$$\lim_{k\to\infty} \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle = m'(\widehat{\mathbf{x}}_i; \mathbf{z}_i) \quad \text{if } \widehat{\mathbf{x}}_i \neq \mathbf{0},$$

and

$$\limsup_{k\to\infty} \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle = m'(\widehat{\mathbf{x}}_i; \mathbf{z}_i) \quad \text{if } \widehat{\mathbf{x}}_i = \mathbf{0}.$$

By the condition $\varphi'(0) := \varphi'(0+) = g'(0+) - h'(0+) > 0$, we know that $g' \circ m(\widehat{\mathbf{x}}_i) - h' \circ m(\widehat{\mathbf{x}}_i) > 0$ if $\widehat{\mathbf{x}}_i = \mathbf{0}$. Hence when k is sufficiently large, $g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) > 0$ for the index i such that $\widehat{\mathbf{x}}_i = \mathbf{0}$. From (4.14), we derive that for any $\mathbf{z} \in \mathbb{R}^n$,

$$\begin{aligned} 0 &= \lim_{k\to\infty} \langle \nabla \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}), \mathbf{z} \rangle \\ &= \lim_{k\to\infty} \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}), \mathbf{z} \rangle + \lim_{k\to\infty} \sum_{i=1}^{K} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle \\ &= \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{z} \rangle + \lim_{k\to\infty} \sum_{i:\ \widehat{\mathbf{x}}_i \neq \mathbf{0}} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle \\ &\quad + \lim_{k\to\infty} \sum_{i:\ \widehat{\mathbf{x}}_i = \mathbf{0}} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle \\ &\le \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{z} \rangle + \sum_{i:\ \widehat{\mathbf{x}}_i \neq \mathbf{0}} \lim_{k\to\infty} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \cdot \lim_{k\to\infty} \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle \\ &\quad + \sum_{i:\ \widehat{\mathbf{x}}_i = \mathbf{0}} \lim_{k\to\infty} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right] \cdot \limsup_{k\to\infty} \langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i \rangle \end{aligned}$$

$$
\begin{aligned}
&= \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{z} \rangle + \sum_{i:\ \widehat{\mathbf{x}}_i \neq \mathbf{0}} \Big[g' \circ m(\widehat{\mathbf{x}}_i) - h' \circ m(\widehat{\mathbf{x}}_i) \Big] m'(\widehat{\mathbf{x}}_i; \mathbf{z}_i) \\
&\quad + \sum_{i:\ \widehat{\mathbf{x}}_i = \mathbf{0}} \Big[g' \circ m(\widehat{\mathbf{x}}_i) - h' \circ m(\widehat{\mathbf{x}}_i) \Big] m'(\widehat{\mathbf{x}}_i; \mathbf{z}_i) \\
&= \langle \nabla \mathcal{L}(\widehat{\mathbf{x}}), \mathbf{z} \rangle + \sum_{i=1}^{K} \Big[g' \circ m(\widehat{\mathbf{x}}_i) - h' \circ m(\widehat{\mathbf{x}}_i) \Big] m'(\widehat{\mathbf{x}}_i; \mathbf{z}_i) \\
&= f'(\widehat{\mathbf{x}}; \mathbf{z}),
\end{aligned}
$$

which shows that $\widehat{\mathbf{x}}$ is a first-order d-stationary point of problem (1.1). $\square$

Before discussing the consistency of second-order stationary points, we first study the property of second-order stationary points of the smoothing problem (4.12).

Lemma 4.2 *Under Assumption (A2), let $\widehat{\mathbf{x}}^{\mu_k} \in \mathbb{R}^n$ be a second-order stationary point of problem (4.12) with $\mu = \mu_k$, then the following two statements hold for $i = 1, \cdots, K$:*

(i) $\|[\nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})]_i\| = \Big| g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \Big| \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|}{\sqrt{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k}}.$

(ii) $\widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^4}{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k} \leq \langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}) \widehat{\mathbf{x}}_i^{\mu_k}, \widehat{\mathbf{x}}_i^{\mu_k} \rangle$
$+ \Big[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \Big] \frac{\mu_k \|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k)^{\frac{3}{2}}}.$

Proof (i) Since $\widehat{\mathbf{x}}^{\mu_k}$ is a second-order stationary point of problem (4.12), we have

$$
\nabla \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}) = \nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}) + \begin{pmatrix} \Big[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k}) \Big] \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_1^{\mu_k}) \\ \vdots \\ \Big[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k}) \Big] \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_K^{\mu_k}) \end{pmatrix} = \mathbf{0},
$$

where, according to (4.2), $\nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) = \frac{\widehat{\mathbf{x}}_i^{\mu_k}}{\sqrt{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k}}$ for $i = 1, \cdots, K$. Therefore, we get

$$
\|[\nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})]_i\| = \Big| g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \Big| \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|}{\sqrt{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k}}.
$$

(ii) Since $\widehat{\mathbf{x}}^{\mu_k}$ is a second-order stationary point of problem (4.12), we know that $\nabla^2 \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k})$ is positive semi-definite, and then $\langle \nabla^2 \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}) \mathbf{z}, \mathbf{z} \rangle \geq 0$ for any $\mathbf{z} \in \mathbb{R}^n$. For each fixed $i = 1, \cdots, K$, let $\bar{\mathbf{z}}_i = \widehat{\mathbf{x}}_i^{\mu_k}$ and other entries of $\bar{\mathbf{z}}$ are all zeros, then we get

$$
\begin{aligned}
0 &\leq \langle \nabla^2 \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}) \bar{\mathbf{z}}, \bar{\mathbf{z}} \rangle \\
&= \langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}) \widehat{\mathbf{x}}_i^{\mu_k}, \widehat{\mathbf{x}}_i^{\mu_k} \rangle + \Big[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \Big] \cdot \langle \nabla^2 \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \widehat{\mathbf{x}}_i^{\mu_k}, \widehat{\mathbf{x}}_i^{\mu_k} \rangle \\
&\quad - \widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \Big[\langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \widehat{\mathbf{x}}_i^{\mu_k} \rangle \Big]^2,
\end{aligned}
\tag{4.15}
$$

where, according to (4.2) and $\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) = \sqrt{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k}$,

$$
\Big[\langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \widehat{\mathbf{x}}_i^{\mu_k} \rangle \Big]^2 = \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^4}{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k}, \ \langle \nabla^2 \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \widehat{\mathbf{x}}_i^{\mu_k}, \widehat{\mathbf{x}}_i^{\mu_k} \rangle = \frac{\mu_k \|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k)^{\frac{3}{2}}}.
$$

Thus, from (4.15), we obtain

$$\begin{aligned}&\widetilde{h}''_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^4}{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2+\mu_k}\\ \le\ &\langle\nabla_i^2\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\widehat{\mathbf{x}}_i^{\mu_k},\widehat{\mathbf{x}}_i^{\mu_k}\rangle+\left[g'\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})-\widetilde{h}'_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\right]\frac{\mu_k\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2+\mu_k)^{\frac{3}{2}}}.\end{aligned}$$

The proof is completed. □

Now we begin to discuss the consistency of second-order stationary points. If h is twice differentiable in $(0,\infty)$, such as logarithm penalty and fraction penalty, there is no need to smooth h, but $h^{\rm MCP}$ and $h^{\rm SCAD}$ are not twice differentiable in $(0,\infty)$. In the following part, we focus on that h is not twice differentiable in $(0,\infty)$.

Assumption (A3) Under Assumption (A2),

$$D(h):=\{t\in(0,\infty): h \text{ is not twice differentiable at } t\} \tag{4.16}$$

has finite many points. In this case, we denote $l_h:=\min\{t:t\in D(h)\}$, $L_h:=\max\{t:t\in D(h)\}$.

We can easily check that several penalty functions satisfy Assumption (A3), such as MCP ($l_h=L_h=\alpha\lambda$) and SCAD ($l_h=\lambda, L_h=\alpha\lambda$). We also observe that the values of l_h and L_h are highly consistent with the corresponding lower bounds obtained in Corollary 2.5 and Theorem 3.9. Since g is affine in Assumption (A2), we know $\varphi=g-h$ is also not twice differentiable at t for $t\in D(h)$.

Lemma 4.3 *Suppose Assumption (A3) holds and the following four conditions hold.*

(a) There exists a nondecreasing function $C:\mathbb{R}\to\mathbb{R}_+$ such that $\|\nabla\mathcal{L}(\mathbf{x})\|\le C(\mathcal{L}(\mathbf{x}))$ for any $\mathbf{x}\in\mathbb{R}^n$.

(b) There exists $\mathbf{x}^0\in\mathbb{R}^n$ satisfying $\varphi'(0)>C(\mathcal{L}(\mathbf{x}^0))$.

(c) There exists $M>0$ such that $\|\nabla^2\mathcal{L}(\mathbf{x})\|_2\le M$ for all $\mathbf{x}\in\mathbb{R}^n$.

(d) If $l_h=L_h$, it holds that $\inf\limits_{t\in(0,L_h]}\max\{H'(t;1),-H'(t;-1)\}>M$; if $l_h<L_h$, it holds that $\inf\limits_{t\in(0,l_h]}\varphi'(t)\ge\varphi'(0)$ and that $\inf\limits_{t\in(l_h,L_h]}\max\{H'(t;1),-H'(t;-1)\}>M$, where $H(t)=h'(t)$.

Let $\{\widehat{\mathbf{x}}^{\mu_k}\}$ be a sequence of second-order stationary points of problem (4.12) with $\mu=\mu_k$ satisfying $\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\le\mathcal{L}(\mathbf{x}^0)$, and $\widehat{\mathbf{x}}$ be any accumulation point of $\{\widehat{\mathbf{x}}^{\mu_k}\}$, then the following three statements hold:

(i) $\|\nabla\mathcal{L}(\widehat{\mathbf{x}})\|<\varphi'(0)$.

(ii) $\min\limits_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}\|\widehat{\mathbf{x}}_i\|>L_h$.

(iii) For any subsequence $\{\widehat{\mathbf{x}}^{\mu_k}\}_{k\in\mathcal{K}}$ converging to $\widehat{\mathbf{x}}$, we have

$$\varGamma^{\mu_k}:=\{i\in\{1,\cdots,K\}:\|\widehat{\mathbf{x}}_i^{\mu_k}\|\le\frac{L_h}{2}\}=\{i\in\{1,\cdots,K\}:\|\widehat{\mathbf{x}}_i\|=0\}:=\varGamma$$

for all sufficiently large $k\in\mathcal{K}$,

Proof Without loss of generality, we may assume that $\{\widehat{\mathbf{x}}^{\mu_k}\}$ converges to $\widehat{\mathbf{x}}$.

(i) By Condition (a) and $\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\le\mathcal{L}(\mathbf{x}^0)$, we have

$$\|\nabla\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\|\le C(\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}))\le C(\mathcal{L}(\mathbf{x}^0)).$$

Then it follows from the continuity of $\nabla\mathcal{L}(\cdot)$, $\widehat{\mathbf{x}}^{\mu_k}\to\widehat{\mathbf{x}}$ and Condition (b) that

$$\|\nabla\mathcal{L}(\widehat{\mathbf{x}})\|\le C(\mathcal{L}(\mathbf{x}^0))<\varphi'(0).$$

The first conclusion is proved.

(ii) We consider an arbitrary nonzero group of $\widehat{\mathbf{x}}$, say $\widehat{\mathbf{x}}_i \neq \mathbf{0}$. Since $\mu_k \to 0$ and $\widehat{\mathbf{x}}_i^{\mu_k} \to \widehat{\mathbf{x}}_i$, it follows from (4.3) and (4.8) that

$$\begin{aligned} &\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \to m(\widehat{\mathbf{x}}_i) = \|\widehat{\mathbf{x}}_i\| \neq 0, \ \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|}{\sqrt{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2+\mu_k}} \to 1, \text{ and} \\ &\left| g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right| \to |g' \circ m(\widehat{\mathbf{x}}_i) - h' \circ m(\widehat{\mathbf{x}}_i)| = \varphi'(\|\widehat{\mathbf{x}}_i\|) \geq 0. \end{aligned} \tag{4.17}$$

As a consequence of Lemma 4.2 (i), (4.17) and $\|[\nabla \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})]_i\| \to \|[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i\|$, we get

$$\|[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i\| = \varphi'(\|\widehat{\mathbf{x}}_i\|). \tag{4.18}$$

From Lemma 4.2 (ii) and Condition (c), we derive

$$\begin{aligned} &\widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^4}{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k} \\ &\leq \langle \nabla_i^2 \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k}) \widehat{\mathbf{x}}_i^{\mu_k}, \widehat{\mathbf{x}}_i^{\mu_k} \rangle + \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right] \frac{\mu_k \|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k)^{\frac{3}{2}}} \\ &\leq M \|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right] \frac{\mu_k \|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k)^{\frac{3}{2}}}. \end{aligned}$$

Since $\|\widehat{\mathbf{x}}_i^{\mu_k}\| \to \|\widehat{\mathbf{x}}_i\| > 0$, when k is sufficiently large the above inequality can be simplified as

$$\widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \frac{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2}{\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k} \leq M + \frac{\left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right] \mu_k}{(\|\widehat{\mathbf{x}}_i^{\mu_k}\|^2 + \mu_k)^{\frac{3}{2}}}.$$

Let $k \to 0$ in the above inequality. By (4.10) and (4.17), we obtain

$$\max\{H'(\|\widehat{\mathbf{x}}_i\|; 1), -H'(\|\widehat{\mathbf{x}}_i\|; -1)\} = \limsup_{k \to \infty} \widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \leq M. \tag{4.19}$$

To verify the second conclusion, let us consider two cases.

Case 1: $l_h = L_h$. In this case, assume, on the contrary, that $\|\widehat{\mathbf{x}}_i\| \leq L_h$. Then by the first part of Condition (d), we obtain

$$\max\{H'(\|\widehat{\mathbf{x}}_i\|; 1), -H'(\|\widehat{\mathbf{x}}_i\|; -1)\} \geq \inf_{t \in (0, L_h]} \max\{H'(t; 1), -H'(t; -1)\} > M,$$

which is in contradiction with (4.19). Hence, we must have $\|\widehat{\mathbf{x}}_i\| > L_h$.

Case 2: $l_h < L_h$. In this case, assume at first that $\|\widehat{\mathbf{x}}_i\| \leq l_h$. Then by the second part of Condition (d), we obtain

$$\varphi'(\|\widehat{\mathbf{x}}_i\|) \geq \inf_{t \in (0, l_h]} \varphi'(t) \geq \varphi'(0).$$

But equality (4.18) and Conclusion (i) yield that

$$\varphi'(\|\widehat{\mathbf{x}}_i\|) = \|[\nabla \mathcal{L}(\widehat{\mathbf{x}})]_i\| \leq \|\nabla \mathcal{L}(\widehat{\mathbf{x}})\| < \varphi'(0),$$

which is a contradiction. Hence, we must have $\|\widehat{\mathbf{x}}_i\| > l_h$.

Secondly, assume that $l_h < \|\widehat{\mathbf{x}}_i\| \leq L_h$. Then by the second part of Condition (d), we obtain

$$\max\{H'(\|\widehat{\mathbf{x}}_i\|; 1), -H'(\|\widehat{\mathbf{x}}_i\|; -1)\} \geq \inf_{t \in (l_h, L_h]} \max\{H'(t; 1), -H'(t; -1)\} > M,$$

which is in contradiction with inequality (4.19). Hence, we must have $\|\widehat{\mathbf{x}}_i\| > L_h$.

Taken together, we have shown that $\|\widehat{\mathbf{x}}_i\| > L_h$ whenever $\widehat{\mathbf{x}}_i \neq 0$, which means $\min_{i:\widehat{\mathbf{x}}_i \neq 0} \|\widehat{\mathbf{x}}_i\| > L_h$.

(iii) Let a subsequence $\{\widehat{\mathbf{x}}^{\mu_k}\}_{k\in\mathcal{K}} \to \widehat{\mathbf{x}}$, then $\{\widehat{\mathbf{x}}_i^{\mu_k}\}_{k\in\mathcal{K}} \to \widehat{\mathbf{x}}_i$ for each $i = 1, \cdots, K$. Suppose $i \in \Gamma$, then $\widehat{\mathbf{x}}_i = 0$. Since $\{\|\widehat{\mathbf{x}}_i^{\mu_k}\|\}_{k\in\mathcal{K}} \to \|\widehat{\mathbf{x}}_i\| = 0$, we have $\|\widehat{\mathbf{x}}_i^{\mu_k}\| < \frac{L_h}{2}$ for all sufficiently large $k \in \mathcal{K}$. That is, $i \in \Gamma^{\mu_k}$ for any sufficiently large $k \in \mathcal{K}$, which shows $\Gamma \subset \Gamma^{\mu_k}$. Now we suppose $i \in \Gamma^{\mu_k}$, then $\|\widehat{\mathbf{x}}_i^{\mu_k}\| \le \frac{L_h}{2}$. If $i \notin \Gamma$, then $\widehat{\mathbf{x}}_i \neq 0$, therefore $\|\widehat{\mathbf{x}}_i\| > L_h$ according to (ii). It follows from $\{\widehat{\mathbf{x}}_i^{\mu_k}\}_{k\in\mathcal{K}} \to \widehat{\mathbf{x}}_i$ that $\|\widehat{\mathbf{x}}_i^{\mu_k}\| > \frac{L_h}{2}$ for any sufficiently large $k \in \mathcal{K}$. This contradiction shows $i \in \Gamma$, thus $\Gamma^{\mu_k} \subset \Gamma$ for all sufficiently large $k \in \mathcal{K}$. Therefore, $\Gamma^{\mu_k} = \Gamma$ for all sufficiently large $k \in \mathcal{K}$. □

Remark 4.4 *Condition (d) in Lemma 4.3 is very important to ensure the lower bound given by Conclusion (ii) when h is differentiable but not twice differentiable in* $(0,\infty)$*. We can see that MCP and SCAD meet this condition. In fact, for MCP,* $l_h = L_h = \alpha\lambda$ $(\alpha > 1)$*, then* $\inf_{t\in(0,L_h]} \max\{H'(t;1), -H'(t;-1)\} = \frac{1}{\alpha} > M$ *whenever* α *is taken such that* $1 < \alpha < \frac{1}{M}$*; and for SCAD,* $l_h = \lambda < \alpha\lambda = L_h$ $(\alpha > 2)$*, then* $\inf_{t\in(0,l_h]} \varphi'(t) = \lambda = \varphi'(0)$ *and* $\inf_{t\in(l_h,L_h]} \max\{H'(t;1), -H'(t;-1)\} = \frac{1}{\alpha-1} > M$ *whenever* α *is taken such that* $2 < \alpha < \frac{1}{M}+1$.

Theorem 4.5 *(Consistency of second-order stationary points) Under the conditions of Lemma 4.3, let* $\{\widehat{\mathbf{x}}^{\mu_k}\}$ *be a sequence of second-order stationary points of problem (4.12) with* $\mu = \mu_k$ *satisfying* $\mathcal{L}(\mathbf{x}^{\mu_k}) \le \mathcal{L}(\mathbf{x}^0)$*, then any accumulation point of* $\{\widehat{\mathbf{x}}^{\mu_k}\}$ *is a second-order d-stationary point of problem (1.1).*

Proof Without loss of generality, we may assume that $\{\widehat{\mathbf{x}}^{\mu_k}\}$ converges to $\widehat{\mathbf{x}}$. Since $\widehat{\mathbf{x}}^{\mu_k}$ is a second-order stationary point of problem (4.12) with $\mu = \mu_k$, we have

$$\nabla \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k}) = \mathbf{0} \quad \text{and} \quad \langle \nabla^2 \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z}, \mathbf{z}\rangle \ge 0, \quad \forall\, \mathbf{z} \in \mathbb{R}^n.$$

According to Theorem 4.1, $\widehat{\mathbf{x}}$ is a first-order d-stationary point of problem (1.1), that is, $f'(\widehat{\mathbf{x}};\mathbf{z}) \ge 0$ for any $\mathbf{z} \in \mathbb{R}^n$.

In the following arguments, we only consider such $\mathbf{z} \in \mathbb{R}^n$ that makes $f'(\widehat{\mathbf{x}};\mathbf{z}) = 0$. According to Lemma 4.3 (i), it holds that $\max_{i:\widehat{\mathbf{x}}_i=\mathbf{0}} \|[\nabla\mathcal{L}(\widehat{\mathbf{x}})]_i\| \le \|\nabla\mathcal{L}(\widehat{\mathbf{x}})\| < \varphi'(0)$. By virtue of this inequality and Corollary 2.7 (iii), it yields from $f'(\widehat{\mathbf{x}};\mathbf{z}) = 0$ that $\mathbf{z}_i = \mathbf{0}$ whenever $\widehat{\mathbf{x}}_i = \mathbf{0}$.

By using $\mathbf{z}_i = \mathbf{0}$ whenever $\widehat{\mathbf{x}}_i = \mathbf{0}$, we have

$$\begin{aligned}
0 \le\ & \langle \nabla^2 \widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z}, \mathbf{z}\rangle \\
=\ & \langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z}, \mathbf{z}\rangle \\
& + \sum_{i=1}^{K} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right] \cdot \langle \nabla^2 \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\mathbf{z}_i, \mathbf{z}_i\rangle \\
& - \sum_{i=1}^{K} \widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \left[\langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i\rangle \right]^2 \\
=\ & \langle \nabla^2 \mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z}, \mathbf{z}\rangle + \sum_{i:\widehat{\mathbf{x}}_i \neq \mathbf{0}} \left[g' \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \right] \langle \nabla^2 \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\mathbf{z}_i, \mathbf{z}_i\rangle \\
& - \sum_{i:\widehat{\mathbf{x}}_i \neq \mathbf{0}} \widetilde{h}''_{\mu_k} \circ \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) \cdot \left[\langle \nabla \widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i\rangle \right]^2.
\end{aligned} \tag{4.20}$$

According to Lemma 4.3 (ii), we have $\min\limits_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}} \|\widehat{\mathbf{x}}_i\| > L_h$. Under Assumption (A3), this inequality means that h is twice continuously differentiable at each $\|\widehat{\mathbf{x}}_i\|$ whenever $\widehat{\mathbf{x}}_i \neq \mathbf{0}$.

Since for each i, $\lim\limits_{k\to\infty} \widehat{\mathbf{x}}_i^{\mu_k} = \widehat{\mathbf{x}}_i$, $\lim\limits_{k\to\infty} \widehat{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) = m(\widehat{\mathbf{x}}_i) = \|\widehat{\mathbf{x}}_i\|$,

$$\lim_{k\to\infty}\Big[g'\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\Big] = g'\circ m(\widehat{\mathbf{x}}_i) - h'\circ m(\widehat{\mathbf{x}}_i),$$

and for $\widehat{\mathbf{x}}_i \neq \mathbf{0}$,

$$\lim_{k\to\infty}\langle\nabla\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}), \mathbf{z}_i\rangle = m'(\widehat{\mathbf{x}}_i;\mathbf{z}_i),$$

$$\lim_{k\to\infty}\langle\nabla^2\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\mathbf{z}_i, \mathbf{z}_i\rangle = m^{(2)}(\widehat{\mathbf{x}}_i;\mathbf{z}_i),$$

$$\lim_{k\to\infty}\widetilde{h}''_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) = H'(\|\widehat{\mathbf{x}}_i\|;1) = -H'(\|\widehat{\mathbf{x}}_i\|;-1),$$

then from (4.20), we obtain

$$\begin{aligned}
0 &\le \lim_{k\to\infty}\langle\nabla^2\widetilde{f}_{\mu_k}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z},\mathbf{z}\rangle \\
&= \lim_{k\to\infty}\langle\nabla^2\mathcal{L}(\widehat{\mathbf{x}}^{\mu_k})\mathbf{z},\mathbf{z}\rangle \\
&\quad + \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}\lim_{k\to\infty}\Big[g'\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}) - \widetilde{h}'_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\Big]\lim_{k\to\infty}\langle\nabla^2\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\mathbf{z}_i,\mathbf{z}_i\rangle \\
&\quad - \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}\lim_{k\to\infty}\widetilde{h}''_{\mu_k}\circ\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k})\cdot\Big[\lim_{k\to\infty}\langle\nabla\widetilde{m}_{\mu_k}(\widehat{\mathbf{x}}_i^{\mu_k}),\mathbf{z}_i\rangle\Big]^2 \\
&= \langle\nabla^2\mathcal{L}(\widehat{\mathbf{x}})\mathbf{z},\mathbf{z}\rangle + \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}\Big[g'\circ m(\widehat{\mathbf{x}}_i) - h'\circ m(\widehat{\mathbf{x}}_i)\Big]m^{(2)}(\widehat{\mathbf{x}}_i;\mathbf{z}_i) \\
&\quad - \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}H'(\widehat{\mathbf{x}}_i;1)[m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i)]^2 \\
&= \langle\nabla^2\mathcal{L}(\widehat{\mathbf{x}})\mathbf{z},\mathbf{z}\rangle + \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}\Big[g'\circ m(\widehat{\mathbf{x}}_i) - h'\circ m(\widehat{\mathbf{x}}_i)\Big]m^{(2)}(\widehat{\mathbf{x}}_i;\mathbf{z}_i) \\
&\quad - \sum_{i:\widehat{\mathbf{x}}_i\neq\mathbf{0}}H'(\widehat{\mathbf{x}}_i;m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i))m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i) \\
&= \langle\nabla^2\mathcal{L}(\widehat{\mathbf{x}})\mathbf{z},\mathbf{z}\rangle + \sum_{i=1}^K\Big[g'\circ m(\widehat{\mathbf{x}}_i) - h'\circ m(\widehat{\mathbf{x}}_i)\Big]m^{(2)}(\widehat{\mathbf{x}}_i;\mathbf{z}_i) - \sum_{i=1}^K H'(\widehat{\mathbf{x}}_i;m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i))m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i) \\
&= f^{(2)}(\widehat{\mathbf{x}};\mathbf{z}),
\end{aligned}$$

where the third equality is due to

$$H'(\widehat{\mathbf{x}}_i;1)[m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i)]^2 = -H'(\widehat{\mathbf{x}}_i;-1)[m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i)]^2 = H'(\widehat{\mathbf{x}}_i;m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i))m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i)$$

when $\|\widehat{\mathbf{x}}_i\| > L_h$, and the fourth equality is due to

$$\mathbf{z}_i = \mathbf{0},\ m'(\widehat{\mathbf{x}}_i,\mathbf{z}_i) = m^{(2)}(\widehat{\mathbf{x}}_i;\mathbf{z}_i) = 0$$

when $\widehat{\mathbf{x}}_i = \mathbf{0}$.

As a summary, we have shown that $\widehat{\mathbf{x}}$ is a first-order d-stationary point of problem (1.1) and that for any $\mathbf{z}\in\mathbb{R}^n$, $f'(\widehat{\mathbf{x}};\mathbf{z}) = 0$ implies $f^{(2)}(\widehat{\mathbf{x}};\mathbf{z}) \ge 0$. Therefore, $\widehat{\mathbf{x}}$ is a second-order d-stationary point of problem (1.1). $\square$

Now, we use an example of problem (1.1) to illustrate how to compute a second-order directional stationary point by the smoothing method.

Example 4.1. Consider the following problem

$$\min_{x_1,x_2\in\mathbb{R}} f(x_1,x_2) := \frac{1}{2}(x_1+x_2-1)^2 + \varphi^{\mathrm{MCP}}(|x_1|) + \varphi^{\mathrm{MCP}}(|x_2|), \tag{4.21}$$

where the parameters in φ^{MCP} satisfy $\alpha > 1$ and $\lambda > 0$. In Tables 1,2,3, we present the sets of the first-order d-stationary points, second-order d-stationary points, local minimizers, and global minimizers of (4.21) with different parameters. From the tables, we can see the relation between these sets for problem (4.21):

first-order d-stationary $\Leftarrow$ second-order d-stationary $\Leftrightarrow$ local minimizer $\Leftarrow$ global minimizer

For example, when $0 < \alpha\lambda \le \frac{1}{2}$, $\lambda < 1$, let $\bar{x} := (\bar{x}_1, \bar{x}_2)^\top = (1+\alpha\lambda, -\alpha\lambda)^\top$, then $\bar{x}_1 + \bar{x}_2 = 1$, $|\bar{x}_1| \ge \alpha\lambda$, and $|\bar{x}_2| \ge \alpha\lambda$. It is easy to check that for any $d := (d_1, d_2)^\top \in \mathbb{R}^2$, $f'(\bar{x}; d) = 0$ and

$$\begin{aligned} f^{(2)}(\bar{x}; d) &= (d_1, d_2)\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}(d_1, d_2)^\top + \begin{cases} -\frac{d_2^2}{\alpha}, & d_2 > 0, \\ 0, & d_2 \le 0, \end{cases} \\ &= \begin{cases} (d_1+d_2)^2 - \frac{d_2^2}{\alpha}, & d_2 > 0, \\ (d_1+d_2)^2, & d_2 \le 0. \end{cases} \end{aligned}$$

Since it cannot ensure $f^{(2)}(\bar{x}; d) \ge 0$ for any d, $\bar{x}$ is a first-order d-stationary point but not a second-order d-stationary point of (4.21).

Table 1 First-order d-stationary points of (4.21)

parameters		first-order d-stationary points
$0 < \alpha\lambda \le \frac{1}{2}$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert \ge \alpha\lambda, \lvert x_2\rvert \ge \alpha\lambda\}$
$\frac{1}{2} < \alpha\lambda \le 1$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top,\ (\frac{\alpha(1-\lambda)}{2\alpha-1}, \frac{\alpha(1-\lambda)}{2\alpha-1})^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert \ge \alpha\lambda, \lvert x_2\rvert \ge \alpha\lambda\}$
$\alpha\lambda > 1$	$\lambda < 1$	$(\frac{\alpha(1-\lambda)}{\alpha-1}, 0)^\top,\ (0, \frac{\alpha(1-\lambda)}{\alpha-1})^\top,\ (\frac{\alpha(1-\lambda)}{2\alpha-1}, \frac{\alpha(1-\lambda)}{2\alpha-1})^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert \ge \alpha\lambda, \lvert x_2\rvert \ge \alpha\lambda\}$
$\alpha\lambda > 1$	$\lambda \ge 1$	$(0,0)^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert \ge \alpha\lambda, \lvert x_2\rvert \ge \alpha\lambda\}$

Table 2 Second-order d-stationary points / local minimizers of (4.21)

parameters		second-order d-stationary points / local minimizers
$0 < \alpha\lambda \le \frac{1}{2}$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert > \alpha\lambda, \lvert x_2\rvert > \alpha\lambda\}$
$\frac{1}{2} < \alpha\lambda \le 1$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert > \alpha\lambda, \lvert x_2\rvert > \alpha\lambda\}$
$\alpha\lambda > 1$	$\lambda < 1$	$(\frac{\alpha(1-\lambda)}{\alpha-1}, 0)^\top,\ (0, \frac{\alpha(1-\lambda)}{\alpha-1})^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert > \alpha\lambda, \lvert x_2\rvert > \alpha\lambda\}$
$\alpha\lambda > 1$	$\lambda \ge 1$	$(0,0)^\top,\ \{(x_1,x_2)^\top : x_1+x_2=1, \lvert x_1\rvert > \alpha\lambda, \lvert x_2\rvert > \alpha\lambda\}$

To test the smoothing method and the consistency theory of stationary points, we use the smoothing trust region Newton (STRN) method proposed in [9] with an initial point $(1,1)^\top$ to solve problem (4.21) where the smoothing function of h^{MCP} is taken $\tilde{h}_\mu^{\mathrm{MCP}}$ as (4.11). The numerical results are listed in Table 4, where f^* means the global minimum of (4.21), and $\bar{x}$ is the output solution of the STRN method. Table 4 shows that $\bar{x}$ is a second-order d-stationary point and a global minimizer of problem (4.21).

Table 3 Global minimizers of (4.21)

parameters		global minimizers
$0 < \alpha\lambda \le \frac{1}{2}$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top$
$\frac{1}{2} < \alpha\lambda \le 1$	$\lambda < 1$	$(1,0)^\top,\ (0,1)^\top$
$\alpha\lambda > 1$	$\lambda < 1$	$(\frac{\alpha(1-\lambda)}{\alpha-1}, 0)^\top,\ (0, \frac{\alpha(1-\lambda)}{\alpha-1})^\top$
$\alpha\lambda > 1$	$\lambda \ge 1$	$(0,0)^\top$

Table 4 Numerical results of the STRN method for (4.21) with different values of α and λ

α	λ	global minimizers	f^*	output solution $\bar{x}$	$f(\bar{x})$
$\alpha = 2$	$\lambda = 0.25$	$(1,0)^\top,\ (0,1)^\top$	0.0625	$(1,0)^\top$	0.0625
$\alpha = 1.5$	$\lambda = 0.5$	$(1,0)^\top,\ (0,1)^\top$	0.1875	$(1,0)^\top$	0.1875
$\alpha = 3$	$\lambda = 0.5$	$(0.75,0)^\top,\ (0,0.75)^\top$	0.3125	$(0.75,0)^\top$	0.3125
$\alpha = 2$	$\lambda = 1$	$(0,0)^\top$	0.5	$(0,0)^\top$	0.5

5 Concluding remarks

This paper shows that the first-order and second-order d-stationary points of folded concave penalized group sparse optimization problem (1.1) are local minimizers fulfilling the first-order and second-order growth conditions respectively under some mild conditions. Moreover, we construct a twice continuously differentiable smoothing approximation for the nonsmooth objective function, and show that any accumulation point of the sequence of second-order stationary points of the smoothing problem is a second-order d-stationary point of problem (1.1). The result provides a theoretic basis for computing first-order and second-order d-stationary points of the problem by using the gradient and Hessian of smoothing functions. Our results can be used for developing second-order algorithms for folded concave penalized group sparse optimization problems, and verifying the optimality of numerical solutions obtained by any algorithms. A simple example shows the validity of our theory and numerical method.

Acknowledgements. The authors would like to thank two referees for their helpful comments.

References

1. Ahn, M., Pang, J.-S., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. **27**, 1637-1665 (2017)
2. Ben-Tal, A., Zowe, J.: Necessary and sufficient optimality conditions for a class of nonsmooth minimization problems. Math. Program. **24**, 70-91 (1982)
3. Bian, W., Chen, X.: Optimality and complexity for constrained optimization problems with nonconvex regularization. Math. Oper. Res. **42**, 1063-1084 (2017)
4. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. Math. Program. **149**, 301-327 (2015)
5. Breheny, P., Huang, J.: Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat. Comput. **25**, 173-187 (2015)
6. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34-81 (2009)
7. Chang, T.H., Hong, M., Pang, J.-S.: Local minimizers and second-order conditions in composite piecewise programming via directional derivatives. arxiv.org/abs/1709.05758 (2017)
8. Chen, X.: Smoothing methods for nonsmooth, novonvex minimization. Math. Program. **134**, 71-99 (2012)
9. Chen, X., Niu L., Yuan Y.: Optimality conditions and a smoothing trust region Newton method for non-Lipschitz optimization. SIAM J. Optim. **23**, 1528-1552 (2013)

10. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization. SIAM J. Sci. Comput. **32**, 2832-2852 (2010)
11. Eren Ahsen, M., Vidyasagar, M.: Error bounds for compressed sensing algorithms with group sparsity: A unified approach. Appl. Comput. Harmon. Anal. **43**, 212-232 (2017)
12. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Stat. Assoc. **96**, 1348-1360 (2001)
13. Fan, J., Li, R.: Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the International Congress of Mathematicians **3**, 595-622, Madrid, Spain (2006)
14. Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**, 819-849 (2014)
15. Hu, Y., Li, C., Meng, K., Qin, J., Yang, X.: Group sparse optimization via $\ell_{p,q}$ regularization. J. Mach. Learn. Res. **18**, 1-52 (2017)
16. Huang, J., Ma, S., Xue, H., Zhang, C.H.: A group bridge approach for variable selection. Biometrika **96**, 339-355 (2009)
17. Huang, J., Zhang, T.: The benefit of group sparsity. Ann. Stat. **38**, 1978-2004 (2010)
18. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. Proceedings of the 26th International Conference on Machine Learning, 433-440, Montreal, Canada (2009)
19. Jiao, Y., Jin, B., Lu, X.: Group sparse recovery via the $\ell_0(\ell_2)$ penalty: theory and algorithm. IEEE Trans. Signal Process. **65**, 998-1012 (2017)
20. Knight, K., Fu, W. J.: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356-1378 (2000)
21. Le Thi, H.A., Pham Dinh, T., Le, H.M., Vo, X.T.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**, 26-46 (2015)
22. Lee, S., Oh, M., Kim, Y.: Sparse optimization for nonconvex group penalized estimation. J. Stat. Comput. Simul. **86**, 597-610 (2016)
23. Liu, H., Yao, T., Li, R., Ye, Y.: Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. Math. Program. **166**, 207-240 (2017)
24. Meier, L., van de Geer, S., Bühlmann, P.: The group Lasso for logistic regression. J. R. Stat. Soc. B **70**, 53-71 (2008)
25. Nikolova, M., Ng, M.K., Zhang, S., Ching, W.: Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. SIAM J. Imag. Sci. **1**, 2-25 (2008)
26. Pang, J.-S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Math. Oper. Res. **42**, 95-118 (2017)
27. Rochafellar, R.T., Wets, R.J.-B.: Variational Analysis (3rd). Springer-Verlag, Berlin, 2009
28. Tibshirani, R.: Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. B **58**, 267-288 (1996)
29. Wei, F., Zhu, H.: Group coordinate descent algorithms for nonconvex penalized regression. Comput. Stat. Data Anal. **56**, 316-326 (2012)
30. Yang, E., Lozano, A.C.: Sparse + group-sparse dirty models: statistical guarantees without unreasonable conditions and a case for non-convexity. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 3911-3920, Sydney, Australia (2017)
31. Yang, X.Q.: On second-order directional derivatives. Nonlin. Anal. TMA **26**, 55-66, 1996.
32. Yang, Y., Zou, H.: A fast unified algorithm for solving group-lasso penalize learning problems. Stat. Comput. **25**, 1129-1141 (2015)
33. Y. Yuan, Conditions for convergence of trust region algorithms for nonsmooth optimization. Math. Program. **31**, 220-228 (1985)
34. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B **68**, 49-67 (2006)
35. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894-942 (2010)
36. Zhang, T.: Analysis of multi-stage convex relaxation for sparse regularization. J. Mach. Learn. Res. **11**, 1081-1107 (2010)
37. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**, 3468-3497 (2009)