mathematics

MDPI

Article

Optimality Conditions for Group Sparse Constrained Optimization Problems

Wenying Wu † and Dingtao Peng *,†

School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China; wywu@gzu.edu.cn
* Correspondence: dtpeng@gzu.edu.cn
† These authors contributed equally to this work.

Abstract: In this paper, optimality conditions for the group sparse constrained optimization (GSCO) problems are studied. Firstly, the equivalent characterizations of Bouligand tangent cone, Clarke tangent cone and their corresponding normal cones of the group sparse set are derived. Secondly, by using tangent cones and normal cones, four types of stationary points for GSCO problems are given: T^B-stationary point, N^B-stationary point, T^C-stationary point and N^C-stationary point, which are used to characterize first-order optimality conditions for GSCO problems. Furthermore, both the relationship among the four types of stationary points and the relationship between stationary points and local minimizers are discussed. Finally, second-order necessary and sufficient optimality conditions for GSCO problems are provided.

Keywords: group sparse constrained optimization; tangent cone; normal cone; first-order optimality condition; second-order optimality condition

Citation: Wu, W.; Peng, D. Optimality Conditions for Group Sparse Constrained Optimization Problems. *Mathematics* **2021**, *9*, 84. https://doi.org/10.3390/math9010084

Received: 27 November 2020
Accepted: 29 December 2020
Published: 1 January 2021

1. Introduction

The sparsity of a vector means that few entries of the vector are non-zero, while the group sparsity of a vector means that non-zero entries or zero entries in the vector may have some group structures, that is, they appear in blocks in certain areas. A vector can be grouped according to the prior information of the group structure among the entries, and then each group is examined to see if they are zeros entirely. For example, genes of the same biological path can be regarded as a group in gene expression analysis, so when they are described by a vector, the vector has group sparsity. Since it was first proposed by Yuan and Lin [1] in 2006, the group sparse optimization has attracted much attention of researchers [2–5]. The aim of group sparse optimization is to seek a solution of group sparsity for a system. It is now known that group sparse optimization has broad applications in bioinformatics, pattern recognition, image restoration, neuroimaging and other fields [1,6–8]. For instance, we can restore the signal by use of group sparse optimization according to the prior information of its group sparse structure. Moreover, the stability of the recovery can be improved in the presence of noise while the accuracy of the recovery can be improved in the absence of noise [2]. In practical problems, it is more targeted to adopt the corresponding group sparse optimization model for problems with group sparse structure [9].

The general sparse constrained optimization has been researched by many authors and achieved a lot. Here we mention few of them. In [10], the authors proposed both concepts of restricted strong convexity and restricted strong smoothness to ensure the existence of unique solution for the sparse constrained optimization, and obtained the corresponding error bounds. In [11], the authors defined N^B-stationary point and N^C-stationary point for the sparse constrained optimization. Beck and Eldar [12] put forward three types of first-order necessary optimality conditions for sparse constraints optimization. One of them is the basic feasibility which is a generation of the necessary optimality condition for zero gradient in unconstrained optimization. Another one of them is the L-stationary

point which is based on the fixed point condition and can be used to derive the iterative hard thresholding algorithm for solving sparse constrained optimization problems. As we all know, Calamai and Moŕe [13] introduced T^B-stationary points and T^C-stationary points to describe the optimal conditions for general constrained optimization problems. Although N-stationary points, L-stationary points and T-stationary points are equivalent for convex optimization problems, they are not equivalent for sparse constrained optimization problems because of the non-convexity. In [14], the authors provided a description of the tangent cone and the normal cone of the sparse set, and then used to describe the first-order optimality condition and the second-order optimality condition, furthermore, they extended the results to the optimization problems subjected to sparse constraints and non-negative constraints. Chen, Pan, and Xiu [15] characterized the solutions of three kinds of sparse optimization problems and investigated the relationship among them. Recently, Bian and Chen [16] gave an exact continuous relaxation problem for the sparsity penalty optimization problem, and proposed a smoothing proximal gradient for the relaxation problem.

However, the above works are mainly for general sparse optimization problems. Due to the complexity of the group sparse structure, there still lacks of research on group sparse constrained optimization problems. If the group sparsity is a penalty in the objective function, Peng and Chen [17] studied the first-order and second-order optimality conditions for the relaxation problems for group sparse optimization problems, while Pan and Chen [18] used a capped folded concave function to approximate the group sparsity function and showed that the solution set of the continuous approximation problem and the set of group sparse solutions are same.

This paper focuses on the following group sparse constrained optimization (GSCO) problem, that is,

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad s.t.\ \|\mathbf{x}\|_{2,0} \leq k, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function or a twice continuously differentiable function, $\mathbf{x} \in \mathbb{R}^n$ is divided into m disjoint groups, denoted by $\mathbf{x} = (\mathbf{x}_1^\top, \cdots, \mathbf{x}_m^\top)^\top$ with $\mathbf{x}_i = (x_{i(1)}, \cdots, x_{i(n_i)})^\top \in \mathbb{R}^{n_i}$, $i = 1, \cdots, m$ and $\sum_{i=1}^m n_i = n, n_i \geq 1$, $\|\mathbf{x}\|_{2,0} := \sum_{i=1}^m \sharp\{\|\mathbf{x}_i\|_2 \neq 0\}$ counts the number of non-zero groups in $\mathbf{x}$, where $\|\mathbf{x}_i\|_2$ is the ℓ_2 vector norm of the ith group $\mathbf{x}_i$. Throughout this paper, for simplicity, $\|\cdot\|$ denotes the ℓ_2 vector norm. Let k be a positive integer with $k \leq m \leq n$, and $S := \{\mathbf{x} : \|\mathbf{x}\|_{2,0} \leq k\}$ be a group sparse set.

Problem (1) is called GSCO due to the group structure in its entries. When $m = n$ and $n_i = 1, i = 1, \cdots, m$, Problem (1) reduces to the standard sparse constrained optimization.

Problem (1) is non-convex, non-smooth, and non-Lipschitz, for which the optimality conditions are of the theoretical importance. It is the basis of analyzing and solving the problem. The optimality conditions for constrained optimization are closely related to tangent cones and normal cones of the constraint set. We will use Boligand tangent cone, Clarke tangent cone and the corresponding normal cones of the group sparse set to describe optimality conditions for Problem (1).

This paper is organized as follows. In Section 2, some basic notations and definitions are introduced. In Section 3, the equivalent expressions of Boligand tangent cone, Clarke tangent cone, and the corresponding normal cones of the group sparse constraint set S are given. In Section 4, first-order optimality conditions for Problem (1) based on the tangent cones and normal cones of S are provided. The relationship between stationary points and local minimizers of Problem (1) is also discussed. In Section 5, second-order necessary and sufficient optimality conditions for Problem (1) are given. At last, a brief concluding remark is given in Section 6.

2. Notations and Definitions

In this section, we introduce some notations and preliminaries including the definitions of Boligand tangent cone, Clarke tangent cone and their corresponding normal cones.

For any $\mathbf{x} = (\mathbf{x}_1^\top, \mathbf{x}_2^\top, \cdots, \mathbf{x}_m^\top)^\top \in \mathbb{R}^n$ with $\mathbf{x}_i \in \mathbb{R}^{n_i}$, the group support set of $\mathbf{x}$ is denoted by

$$\Gamma(\mathbf{x}) := \{i \in \{1, \cdots, m\} : \mathbf{x}_i \neq \mathbf{0}\},$$

$|\Gamma(\mathbf{x})|$ is the cardinality of the set $\Gamma(\mathbf{x})$, then $\|\mathbf{x}\|_{2,0} = |\Gamma(\mathbf{x})|$, which means $\|\mathbf{x}\|_{2,0}$ is the number of groups in $\mathbf{x}$ that have nonzero ℓ_2-norm.

For the n-dimensional real number space $\mathbb{R}^n$, $\mathbb{R}_{x_i}$ denotes the x_i coordinate axis, and $\mathbb{R}^2_{x_i x_j}$ denotes the $x_i O x_j$ coordinate plane. Let $e_i \in \mathbb{R}^n$ denote the n-dimensional vector in which the entries in ith group are all ones and the other entries are all zeros. Let $e_{ij}(i = 1, \cdots, m, j = 1, \cdots, n_i)$ denote the n-dimensional vector in which the jth entry of the ith group is one and the other entries are all zeros.

For a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, let

$$[\nabla f(\mathbf{x})]_i := ([\nabla f(\mathbf{x})]_{i(1)}, \cdots, [\nabla f(\mathbf{x})]_{i(n_i)})^\top, \quad \nabla f(\mathbf{x}) := ([\nabla f(\mathbf{x})]_1^\top, \cdots, [\nabla f(\mathbf{x})]_m^\top)^\top,$$

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

The following example shows that the group sparse structure is different from the sparse structure.

Example 1. *Let $\mathbf{x} = (x_1, x_2, x_3)^\top$ be a 3-dimensional vector. We show the different ways of grouping and the corresponding group sparsity of $\mathbf{x}$ as follows.*

(1) When $\mathbf{x} = (x_1, x_2, x_3)^\top, n_1 = n_2 = n_3 = 1$,
if $\|\mathbf{x}\|_{2,0} = 0$, then $\mathbf{x} = \mathbf{0}$;
if $\|\mathbf{x}\|_{2,0} = 1$, then $\mathbf{x} \in \{\mathbf{x}|x_1 \in \mathbb{R}\backslash\{0\}, x_2 = x_3 = 0\} \bigcup \{\mathbf{x}|x_2 \in \mathbb{R}\backslash\{0\}, x_1 = x_3 = 0\} \bigcup \{\mathbf{x}|x_3 \in \mathbb{R}\backslash\{0\}, x_1 = x_2 = 0\}$;
if $\|\mathbf{x}\|_{2,0} = 2$, then $\mathbf{x} \in \{\mathbf{x}|x_1, x_2 \in \mathbb{R}\backslash\{0\}, x_3 = 0\} \bigcup \{\mathbf{x}|x_1, x_3 \in \mathbb{R}\backslash\{0\}, x_2 = 0\} \bigcup \{\mathbf{x}|x_2, x_3 \in \mathbb{R}\backslash\{0\}, x_1 = 0\}$;
if $\|\mathbf{x}\|_{2,0} = 3$, then $\mathbf{x} \in \{\mathbf{x}|x_1, x_2, x_3 \in \mathbb{R}\backslash\{0\}\}$.

(2) When $\mathbf{x} = (x_1, (x_2, x_3))^\top, n_1 = 1, n_2 = 2$,
if $\|\mathbf{x}\|_{2,0} = 0$, then $\mathbf{x} = \mathbf{0}$;
if $\|\mathbf{x}\|_{2,0} = 1$, then $\mathbf{x} \in \{\mathbf{x}|x_1 \in \mathbb{R}\backslash\{0\}, x_2 = x_3 = 0\} \bigcup \{\mathbf{x}|x_1 = 0, (x_2, x_3)^\top \in \mathbb{R}^2\backslash\{\mathbf{0}\}\}$;
if $\|\mathbf{x}\|_{2,0} = 2$, then $\mathbf{x} \in \{\mathbf{x}|x_1 \in \mathbb{R}\backslash\{0\}, (x_2, x_3)^\top \in \mathbb{R}^2\backslash\{\mathbf{0}\}\}$.

(3) When $\mathbf{x} = ((x_1, x_2, x_3))^\top, n_1 = 3$,
if $\|\mathbf{x}\|_{2,0} = 0$, then $\mathbf{x} = \mathbf{0}$;
if $\|\mathbf{x}\|_{2,0} = 1$, then $\mathbf{x} \in \{\mathbf{x}|(x_1, x_2, x_3)^\top \neq \mathbf{0}\}$.

In the end of this section, we will introduce the definition of Bouligand tangent cone, Clarke tangent cone and their corresponding normal cones [19].

Definition 1 ([19]). *Let $\Omega \subseteq \mathbb{R}^n$ be an arbitrary nonemepty set. The Bouligand tangent cone $T_\Omega^B(\widehat{\mathbf{x}})$, the Clarke tangent cone $T_\Omega^C(\widehat{\mathbf{x}})$ and their corresponding normal cone $N_\Omega^B(\widehat{\mathbf{x}})$ and $N_\Omega^C(\widehat{\mathbf{x}})$ to the set Ω at the point $\widehat{\mathbf{x}} \in \Omega$ are defined as follows.*

(1) Bouligand tangent cone:

$$T_\Omega^B(\widehat{\mathbf{x}}) := \left\{\mathbf{d} \in \mathbb{R}^n : \exists\{\mathbf{x}^t\} \subset \Omega, \lim_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}, \exists \lambda_t \geq 0, t \in \mathbb{N}, s.t. \lim_{t\to\infty} \lambda_t(\mathbf{x}^t - \widehat{\mathbf{x}}) = \mathbf{d}\right\};$$

(2) Fréchet normal cone:

$$N_\Omega^B(\widehat{\mathbf{x}}) := [T_\Omega^B(\widehat{\mathbf{x}})]^\circ = \left\{\mathbf{u} \in \mathbb{R}^n : \langle \mathbf{u}, \mathbf{z}\rangle \leq 0, \forall \mathbf{z} \in T_\Omega^B(\widehat{\mathbf{x}})\right\};$$

(3) Clarke tangent cone:

$$T_{\Omega}^{C}(\widehat{\mathbf{x}}) := \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{c} \forall \{\mathbf{x}^t\} \subset \Omega, \lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}, \forall \{\lambda_t\} \subset \mathbb{R}_+, \lim\limits_{t\to\infty} \lambda_t = 0, \exists \{\mathbf{d}^t\} \subset \mathbb{R}^n, \\ s.t. \lim\limits_{t\to\infty} \mathbf{d}^t = \mathbf{d} \text{ and } \mathbf{x}^t + \lambda_t \mathbf{d}^t \in \Omega, t \in \mathbb{N} \end{array} \right\};$$

(4) Clarke normal cone:

$$N_{\Omega}^{C}(\widehat{\mathbf{x}}) := [T_{\Omega}^{C}(\widehat{\mathbf{x}})]^{\circ} = \left\{ \mathbf{u} \in \mathbb{R}^n : \langle \mathbf{u}, \mathbf{z} \rangle \leq 0, \forall \mathbf{z} \in T_{\Omega}^{C}(\widehat{\mathbf{x}}) \right\}.$$

3. Tangent Cones and Normal Cones of the Group Sparse Set S

Tangent cones and normal cones are widely used to describe optimality conditions for constrained optimization problems [19]. The following two theorems give the equivalent characterizations of Bouligand tangent cone, Clarke tangent cone and their corresponding normal cones to the group sparse constraint set S.

Theorem 1. *For any $\widehat{\mathbf{x}} \in S$, the Boligand tangent cone $T_S^B(\widehat{\mathbf{x}})$ and Fréchet normal cone $N_S^B(\widehat{\mathbf{x}})$ to the group sparse set S at the point $\widehat{\mathbf{x}}$ has the following equivalent expressions:*

$$\begin{aligned} T_S^B(\widehat{\mathbf{x}}) &= \{\mathbf{d} \in \mathbb{R}^n : \|\mathbf{d}\|_{2,0} \leq k, \|\widehat{\mathbf{x}} + \gamma \mathbf{d}\|_{2,0} \leq k, \forall \gamma \in \mathbb{R}\} \\ &= \bigcup_{J \in \Theta(\widehat{\mathbf{x}})} \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin J\} \\ &= \bigcup_{J \in \Theta(\widehat{\mathbf{x}})} \operatorname{span}\{e_{ij}, i \in J, j = 1, \cdots, n_i\}; \end{aligned}$$

$$N_S^B(\widehat{\mathbf{x}}) = \begin{cases} \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})\} = \operatorname{span}\{e_{ij}, i \notin \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}, & \|\widehat{\mathbf{x}}\|_{2,0} = k, \\ \{\mathbf{0}\}, & \|\widehat{\mathbf{x}}\|_{2,0} < k, \end{cases}$$

where $\Gamma(\widehat{\mathbf{x}}) = \{i \in \{1, \cdots, m\} : \widehat{\mathbf{x}}_i \neq \mathbf{0}\}$, $\Theta(\widehat{\mathbf{x}}) = \{J \subseteq \{1, \cdots, m\} : \Gamma(\widehat{\mathbf{x}}) \subseteq J, |J| = k\}$, $\mathbf{d}_i \in \mathbb{R}^{n_i}$ is the ith group of $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{u}_i \in \mathbb{R}^{n_i}$ is the ith group of $\mathbf{u} \in \mathbb{R}^n$.

Proof. (i) According to the definition of Bouligand tangent cone, we have

$$T_S^B(\widehat{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \exists \{\mathbf{x}^t\} \subseteq S, \lim_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}, \exists \lambda_t \geq 0, t \in \mathbb{N}, s.t. \lim_{t\to\infty} \lambda_t (\mathbf{x}^t - \widehat{\mathbf{x}}) = \mathbf{d}\}.$$

Firstly, we prove that $T_S^B(\widehat{\mathbf{x}}) = H(\widehat{\mathbf{x}}) := \{\mathbf{d} \in \mathbb{R}^n : \|\mathbf{d}\|_{2,0} \leq k, \|\widehat{\mathbf{x}} + \gamma \mathbf{d}\|_{2,0} \leq k, \forall \gamma \in \mathbb{R}\}$. For any $\mathbf{d} \in T_S^B(\widehat{\mathbf{x}})$, there exists $\{\mathbf{x}^t\} \subseteq S$ such that $\lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}$, then

$$\Gamma(\widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t) \text{ and } |\Gamma(\widehat{\mathbf{x}})| \leq |\Gamma(\mathbf{x}^t)|$$

for any sufficiently large t. It follows from $\mathbf{x}^t \in S$ that

$$|\Gamma(\mathbf{x}^t)| = \|\mathbf{x}^t\|_{2,0} \leq k.$$

Since $\mathbf{d} = \lim\limits_{t\to\infty} \lambda_t (\mathbf{x}^t - \widehat{\mathbf{x}})$ with $\lambda_t \geq 0$, then $\Gamma(\mathbf{d}) \subseteq \Gamma(\mathbf{x}^t - \widehat{\mathbf{x}})$. Due to $\Gamma(\widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t)$, we obtain

$$\Gamma(\mathbf{d}) \subseteq \Gamma(\mathbf{x}^t - \widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t).$$

Therefore,

$$\|\mathbf{d}\|_{2,0} = |\Gamma(\mathbf{d})| \leq |\Gamma(\mathbf{x}^t)| \leq k. \tag{2}$$

According to $\Gamma(\widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t)$ and $\Gamma(\mathbf{d}) \subseteq \Gamma(\mathbf{x}^t)$, then $\Gamma(\widehat{\mathbf{x}} + \gamma\mathbf{d}) \subseteq \Gamma(\mathbf{x}^t)$ for any $\gamma \in \mathbb{R}$. Hence we get

$$\|\widehat{\mathbf{x}} + \gamma\mathbf{d}\|_{2,0} = |\Gamma(\widehat{\mathbf{x}} + \gamma\mathbf{d})| \leq |\Gamma(\mathbf{x}^t)| \leq k. \tag{3}$$

Combining (2) with (3), we get $T_S^B(\widehat{\mathbf{x}}) \subseteq H(\widehat{\mathbf{x}})$.

Conversely, for any $\mathbf{d} \in H(\widehat{\mathbf{x}})$, take any sequence $\{\lambda_t\}$ such that $\lambda_t > 0$ and $\lambda_t \to \infty$, let $\mathbf{x}^t = \widehat{\mathbf{x}} + \frac{\mathbf{d}}{\lambda_t}$, then $\lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}$. Since $\|\widehat{\mathbf{x}} + \gamma\mathbf{d}\|_{2,0} \leq k$ for any $\gamma \in \mathbb{R}$, we get

$$\|\mathbf{x}^t\|_{2,0} = \|\widehat{\mathbf{x}} + \frac{\mathbf{d}}{\lambda_t}\|_{2,0} \leq k,$$

which means $\{\mathbf{x}^t\} \subseteq S$. It follows from $\lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}$ that $\Gamma(\widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t)$. Hence we obtain

$$\|\widehat{\mathbf{x}}\|_{2,0} = |\Gamma(\widehat{\mathbf{x}})| \leq |\Gamma(\mathbf{x}^t)| = \|\mathbf{x}^t\|_{2,0} \leq k.$$

From $\mathbf{x}^t = \widehat{\mathbf{x}} + \frac{\mathbf{d}}{\lambda_t}$, we get

$$\lim_{t\to\infty} \lambda_t(\mathbf{x}^t - \widehat{\mathbf{x}}) = \mathbf{d}.$$

Hence we have $\mathbf{d} \in T_S^B(\widehat{\mathbf{x}})$, which means $T_S^B(\widehat{\mathbf{x}}) \supseteq H(\widehat{\mathbf{x}})$.

The above proof yields $T_S^B(\widehat{\mathbf{x}}) = H(\widehat{\mathbf{x}})$.

It is easy to prove that

$$H(\widehat{\mathbf{x}}) = \bigcup_{J\in\Theta(\widehat{\mathbf{x}})} \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin J\} = \bigcup_{J\in\Theta(\widehat{\mathbf{x}})} \text{span}\{e_{ij}, i \in J, j = 1, \cdots, n_i\}.$$

(ii) According to the definition of Fréchet normal cone,

$$N_S^B(\widehat{\mathbf{x}}) = [T_S^B(\widehat{\mathbf{x}})]^\circ \quad = \quad \{\mathbf{u} \in \mathbb{R}^n : \langle \mathbf{u}, \mathbf{d}\rangle \leq 0, \forall \mathbf{d} \in T_S^B(\widehat{\mathbf{x}})\}.$$

For any $\mathbf{u} \in N_S^B(\widehat{\mathbf{x}})$ and any $\mathbf{d} \in T_S^B(\widehat{\mathbf{x}})$, it must hold $\langle \mathbf{u}, \mathbf{d}\rangle \leq 0$.

If $\|\widehat{\mathbf{x}}\|_{2,0} = k$, we have

$$\langle \mathbf{u}, \mathbf{d}\rangle = \sum_{i\in\Gamma(\widehat{\mathbf{x}})} \langle \mathbf{u}_i, \mathbf{d}_i\rangle + \sum_{i\notin\Gamma(\widehat{\mathbf{x}})} \langle \mathbf{u}_i, \mathbf{d}_i\rangle.$$

Since $\mathbf{d} \in T_S^B(\widehat{\mathbf{x}}) = \bigcup_{J\in\Theta(\widehat{\mathbf{x}})}\{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin J\}$, for any $J \in \Theta(\widehat{\mathbf{x}})$, we have $\Gamma(\widehat{\mathbf{x}}) \subseteq J$ and $\mathbf{d}_i = \mathbf{0}, i \notin J$. Thus we have $\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})$, $\sum\limits_{i\notin\Gamma(\widehat{\mathbf{x}})} \langle \mathbf{u}_i, \mathbf{d}_i\rangle = 0$, and then

$$\langle \mathbf{u}, \mathbf{d}\rangle = \sum_{i\in\Gamma(\widehat{\mathbf{x}})} \langle \mathbf{u}_i, \mathbf{d}_i\rangle \leq 0,$$

which, together with the arbitrariness of $\mathbf{d}_i \in \mathbb{R}^{n_i}$ for $i \in \Gamma(\widehat{\mathbf{x}})$, implies $\mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})$. Therefore, $N_S^B(\widehat{\mathbf{x}}) = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})\}$. It is easy to prove that $\{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})\} = \text{span}\{e_{ij}, i \notin \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}$.

If $\|\widehat{\mathbf{x}}\|_{2,0} < k$, for any $J \in \Theta(\widehat{\mathbf{x}})$, it holds

$$\langle \mathbf{u}, \mathbf{d}\rangle = \sum_{i\in J} \langle \mathbf{u}_i, \mathbf{d}_i\rangle + \sum_{i\notin J} \langle \mathbf{u}_i, \mathbf{d}_i\rangle.$$

We also have $\langle \mathbf{u}, \mathbf{d}\rangle = \sum\limits_{i\in J} \langle \mathbf{u}_i, \mathbf{d}_i\rangle \leq 0$, which also implies $\mathbf{u}_i = \mathbf{0}, i \in J$. Due to $\|\widehat{\mathbf{x}}\|_{2,0} < k$, $\Gamma(\widehat{\mathbf{x}}) \subseteq J$ and $|J| = k$, it must hold $\bigcup\limits_{J\in\Theta(\widehat{\mathbf{x}})} J = \{1, 2, \cdots, m\}$, and then $N_S^B(\widehat{\mathbf{x}}) = \{\mathbf{0}\}$. □

Next, we give the equivalent characterizations of Clarke tangent cone and Clarke normal cone of the group sparse constraint set S.

Theorem 2. *For any $\widehat{\mathbf{x}} \in S$, the Clarke tangent cone and the Clarke normal cone of the group sparse set S at $\widehat{\mathbf{x}}$ have the following equivalent expressions:*

$$\begin{aligned} T_S^C(\widehat{\mathbf{x}}) &= \{\mathbf{d}\in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\} \\ &= \{\mathbf{d}\in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})\} \\ &= \text{span}\{e_{ij}, i \in \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}; \end{aligned}$$

$$\begin{aligned} N_S^C(\widehat{\mathbf{x}}) &= \{\mathbf{u}\in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})\} \\ &= \text{span}\{e_{ij}, i \notin \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}. \end{aligned}$$

Proof. (i) According to the definition of Clarke tangent cone, we have

$$T_S^C(\widehat{\mathbf{x}}) = \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{ll} \forall\{\mathbf{x}^t\} \subseteq S, & \lim_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}, \forall\{\lambda_t\} \subset \mathbb{R}_+, \lim_{t\to\infty} \lambda_t = 0, \exists\{\mathbf{y}^t\} \subset \mathbb{R}^n, \\ & s.t. \lim_{t\to\infty} \mathbf{y}^t = \mathbf{d}, \|\mathbf{x}^t + \lambda_t \mathbf{y}^t\|_{2,0} \le k, \forall t \in \mathbb{N} \end{array} \right\}.$$

We first prove $T_S^C(\widehat{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$.

To prove $T_S^C(\widehat{\mathbf{x}}) \subseteq \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$, we assume, on the contrary, that there exists $\mathbf{d} \in T_S^C(\widehat{\mathbf{x}})$, but $\Gamma(\mathbf{d}) \nsubseteq \Gamma(\widehat{\mathbf{x}})$. Then there exists $i_0 \in \Gamma(\mathbf{d})$ but $i_0 \notin \Gamma(\widehat{\mathbf{x}})$, which implies that $\widehat{\mathbf{x}}_{i_0} = \mathbf{0}$ but $\mathbf{d}_{i_0} \neq \mathbf{0}$.

Note that $|\Gamma(\widehat{\mathbf{x}})| \le k$. For any $t \in \mathbb{N}$, take $\Gamma_t \subseteq \{1, 2, \cdots, m\} \setminus \{\Gamma(\widehat{\mathbf{x}}) \bigcup \{i_0\}\}$ such that

$$|\Gamma(\widehat{\mathbf{x}})| + |\Gamma_t| = k.$$

Let $\lambda_t = \frac{1}{t^2} \downarrow 0$ and

$$\mathbf{x}_i^t = \begin{cases} \mathbf{x}_i, & i \in \Gamma(\widehat{\mathbf{x}}) \\ \frac{1}{t}\mathbf{1}_{n_i}, & i \in \Gamma_t, \\ \mathbf{0}, & i \in \{1, 2, \cdots, m\} \setminus \{\Gamma_t \bigcup \Gamma(\widehat{\mathbf{x}})\}. \end{cases}$$

where $\mathbf{1}_{n_i}$ is an n_i-dimensional vector of all ones. Then

$$\Gamma(\mathbf{x}^t) = \Gamma(\widehat{\mathbf{x}}) \bigcup \Gamma_t, \quad \|\mathbf{x}^t\|_{2,0} = |\Gamma(\widehat{\mathbf{x}})| + |\Gamma_t| = k,$$

and thus $\{\mathbf{x}^t\} \subseteq S$, $\mathbf{x}_{i_0}^t = \mathbf{0}$, and $\lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}$. For any $\mathbf{y}^t \to \mathbf{d}$, we have

$$\mathbf{x}_i^t + \lambda_t \mathbf{y}_i^t = \begin{cases} \mathbf{x}_i^t + \frac{1}{t^2}\mathbf{y}_i^t \to \widehat{\mathbf{x}}_i, & i \in \Gamma(\widehat{\mathbf{x}}), \\ \frac{1}{t}\mathbf{1}_{n_i} + \frac{1}{t^2}\mathbf{y}_i^t \to \mathbf{0}, & i \in \Gamma_t, \\ \mathbf{0} + \frac{1}{t^2}\mathbf{y}_{i_0}^t \to \mathbf{0}, & i = i_0, \\ \mathbf{0} + \frac{1}{t^2}\mathbf{y}_i^t \to \mathbf{0}, & i \in \{1, 2, \cdots, m\} \setminus \{\Gamma_t \bigcup \Gamma(\widehat{\mathbf{x}}) \bigcup \{i_0\}\}. \end{cases}$$

Since $\mathbf{y}_{i_0}^t \to \mathbf{d}_{i_0} \neq \mathbf{0}$, for any sufficiently large t, we have

$$\|\mathbf{x}^t + \lambda_t \mathbf{y}^t\|_{2,0} \ge |\Gamma(\widehat{\mathbf{x}}) \bigcup \Gamma_t \bigcup \{i_0\}| = k + 1,$$

Therefore, $\mathbf{x}^t + \lambda_t \mathbf{y}^t \notin S$ for any sufficiently large t, which means $\mathbf{d} \notin T_S^C(\widehat{\mathbf{x}})$ according to the definition of $T_S^C(\widehat{\mathbf{x}})$. This contradiction shows that $T_S^C(\widehat{\mathbf{x}}) \subseteq \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$.

To prove $\{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\} \subseteq T_S^C(\widehat{\mathbf{x}})$, let $\mathbf{d} \in \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$. For any $\{\mathbf{x}^t\} \subseteq S$, with $\lim\limits_{t\to\infty} \mathbf{x}^t = \widehat{\mathbf{x}}$ and any $\{\lambda_t\} \subset \mathbb{R}_+$ with $\lim_{t\to\infty} \lambda_t = 0$, we have

$$\Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}}) \subseteq \Gamma(\mathbf{x}^t). \tag{4}$$

Let $\mathbf{y}^t = \mathbf{x}^t - \widehat{\mathbf{x}} + \mathbf{d}$, then from (4), we get $\Gamma(\mathbf{y}^t) \subseteq \Gamma(\mathbf{x}^t)$ and

$$\|\mathbf{x}^t + \lambda_t \mathbf{y}^t\|_{2,0} = |\Gamma(\mathbf{x}^t + \lambda_t \mathbf{y}^t)| \leq |\Gamma(\mathbf{x}^t)| \leq k.$$

In addition, $\lim\limits_{t\to\infty} \mathbf{y}^t = \lim\limits_{t\to\infty} (\mathbf{x}^t - \widehat{\mathbf{x}} + \mathbf{d}) = \mathbf{d}$. It is easy to know that $\mathbf{d} \in T_S^C(\widehat{\mathbf{x}})$ according to the definition of $T_S^C(\widehat{\mathbf{x}})$. From the arbitrariness of $\mathbf{d} \in \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$, we have

$$\{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\} \subseteq T_S^C(\widehat{\mathbf{x}}).$$

Therefore, we have proved that $T_S^C(\widehat{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$.

Since $\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{d})$ and $\Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})$ for any $\mathbf{d} \in \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\}$, it must hold $\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})$. Hence we get

$$T_S^C(\widehat{\mathbf{x}}) = \{\mathbf{d} \in \mathbb{R}^n : \Gamma(\mathbf{d}) \subseteq \Gamma(\widehat{\mathbf{x}})\} = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})\}. \tag{5}$$

It is easy to prove that $\{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})\} = \text{span}\{e_{ij} : i \in \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}$, then

$$T_S^C(\widehat{\mathbf{x}}) = \text{span}\{e_{ij} : i \in \Gamma(\widehat{\mathbf{x}}), j = 1, \cdots, n_i\}. \tag{6}$$

(ii) According to the definition of Clarke normal cone, we have

$$N_S^C(\widehat{\mathbf{x}}) = [T_S^C(\widehat{\mathbf{x}})]^\circ = \{\mathbf{u} \in \mathbb{R}^n : \langle \mathbf{d}, \mathbf{u} \rangle \leq 0, \forall \mathbf{d} \in T_S^C(\widehat{\mathbf{x}})\}.$$

For any $\mathbf{d} \in T_S^C(\widehat{\mathbf{x}})$ and any $\mathbf{u} \in N_S^C(\widehat{\mathbf{x}})$, we have

$$\langle \mathbf{d}, \mathbf{u} \rangle = \sum_{i \in \Gamma(\widehat{\mathbf{x}})} \langle \mathbf{d}_i, \mathbf{u}_i \rangle + \sum_{i \notin \Gamma(\widehat{\mathbf{x}})} \langle \mathbf{d}_i, \mathbf{u}_i \rangle \leq 0.$$

From (5), $\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\widehat{\mathbf{x}})$, then we get $\sum\limits_{i \notin \Gamma(\widehat{\mathbf{x}})} \langle \mathbf{d}_i, \mathbf{u}_i \rangle = 0$, and thus

$$\langle \mathbf{d}, \mathbf{u} \rangle = \sum_{i \in \Gamma(\widehat{\mathbf{x}})} \langle \mathbf{d}_i, \mathbf{u}_i \rangle \leq 0.$$

which means $\mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})$ due to the arbitrariness of $\mathbf{d}_i \in \mathbb{R}^{n_i}$. Therefore, $N_S^C(\widehat{\mathbf{x}}) = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\widehat{\mathbf{x}})\}$. □

Obviously, the following relationship holds for Boligand tangent cone, Clarke normal cone and the corresponding normal cones of the group sparse set S at any point $\widehat{\mathbf{x}} \in S$:

$$T_S^C(\widehat{\mathbf{x}}) \subseteq T_S^B(\widehat{\mathbf{x}}), \quad N_S^B(\widehat{\mathbf{x}}) \subseteq N_S^C(\widehat{\mathbf{x}}).$$

Remark 1. *In* [14]*, the authors gave the expressions of tangent cone and normal cone to the sparse set* $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq k\}$ *. Theorems 1 and 2 in this paper are the extension of their results.*

In the end of this section, we give an example of the tangent cones of S in $\mathbb{R}^3$.

Example 2. *Consider the group sparse set*

$$S = \{\mathbf{x} = (x_1, (x_2, x_3))^\top \in \mathbb{R}^3 : \|\mathbf{x}\|_{2,0} \leq 1\},$$

where x_1 is the first group, and $(x_2, x_3)^\top$ is the second group. Consider its Bouligand tangent cone and Clarke tangent cone at three points: $\mathbf{x}^1 = (0,(1,1))^\top, \mathbf{x}^2 = (0,(1,0))^\top$ *and* $\mathbf{x}^3 = (1,(0,0))^\top$. *It is easy to get the following statements:* $\Gamma(\mathbf{x}^1) = \{2\}, \Gamma(\mathbf{x}^2) = \{2\}, \Gamma(\mathbf{x}^3) = \{1\}; \quad \Theta(\mathbf{x}^1) =$

$\{2\}, \Theta(\mathbf{x}^2) = \{2\}, \Theta(\mathbf{x}^3) = \{1\};$

$T_S^B(\mathbf{x}^1) = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x}_1 = 0\} = \mathbb{R}^2_{x_2x_3}; \quad T_S^C(\mathbf{x}^1) = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x}_1 = 0\} = \mathbb{R}^2_{x_2x_3};$

$T_S^B(\mathbf{x}^2) = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x}_1 = 0\} = \mathbb{R}^2_{x_2x_3}; \quad T_S^C(\mathbf{x}^2) = \{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x}_1 = 0\} = \mathbb{R}^2_{x_2x_3};$

$T_S^B(\mathbf{x}^3) = \{\mathbf{x} \in \mathbb{R}^3 : x_2 = x_3 = 0\} = \mathbb{R}_{x_1}; \quad T_S^C(\mathbf{x}^3) = \{\mathbf{x} \in \mathbb{R}^3 : x_2 = x_3 = 0\} = \mathbb{R}_{x_1}.$

Therefore, $T_S^B(\mathbf{x}^1) = T_S^C(\mathbf{x}^1) = T_S^B(\mathbf{x}^2) = T_S^C(\mathbf{x}^2) = \mathbb{R}^2_{x_2x_3}$, $T_S^B(\mathbf{x}^3) = T_S^C(\mathbf{x}^3) = \mathbb{R}_{x_1}$.

Figure 1 provides the figures of the above Bouligand tangent cones and Clarke tangent cones.

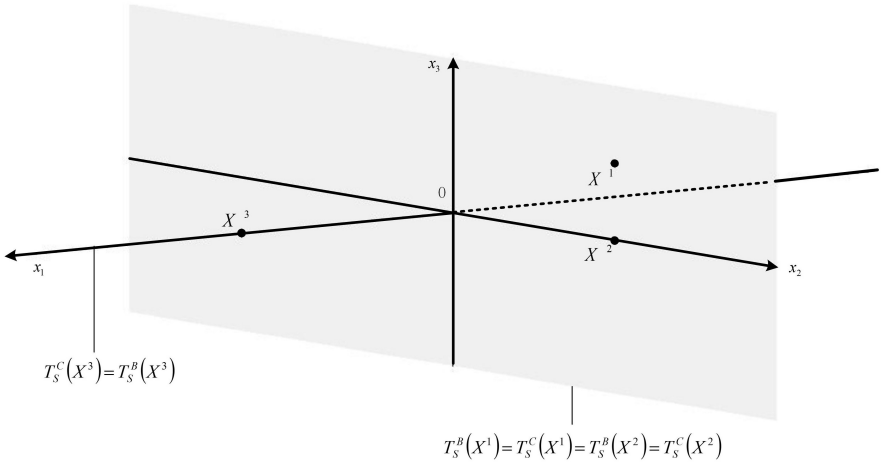

Figure 1. Bouligand tangent cones and Clarke tangent cones of S in $\mathbb{R}^3$, where $S = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_{2,0} \leq 1\}$, $\mathbf{x}^1 = (0,1,1), \mathbf{x}^2 = (0,1,0)$ and $\mathbf{x}^3 = (1,0,0)$.

From example 3.1, we can see that the key of group sparsity is to survey whether each group as a whole is zero instead of checking whether each entry is zero.

4. First-Order Optimality Conditions for Problem (1)

The optimality conditions for optimization problems are usually closely related to their stationary points. In this section, we use Bouligand tangent cones, Clarke tangent cones and their corresponding normal cones to specifically describe the N-stationary points and T-stationary points of Problem (1), then based on the descriptions, we investigate the relationship among the stationary points and the relationship between stationary points and local minimizers.

Definition 2. *$\mathbf{x}^* \in S$ is called an $N^\sharp$-stationary point or $T^\sharp$-stationary point of Problem (1) if it meets the following conditions respectively:*
(i) $N^\sharp$-stationary point: $\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^\sharp(\mathbf{x}^*)$;
(ii) $T^\sharp$-stationary point: $0 = \|\nabla_S^\sharp f(\mathbf{x}^*)\|$;
where $\sharp \in \{B, C\}$ stands for the sense of Bouligand or Clarke, and

$$\nabla_S^\sharp f(\mathbf{x}^*) = \arg\min \left\{ \|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d} \in T_S^\sharp(\mathbf{x}^*) \right\}$$

is the projection gradient on Bouligand tangent cone or Clarke tangent cone.

Next, we will study the link between N^B-stationary point and T^B-stationary point of Problem (1).

Theorem 3. *Suppose $\mathbf{x}^* \in S$, then the following statements hold for Problem (1):*
(i) If $\|\mathbf{x}^*\|_{2,0} = k$, then $\mathbf{x}^*$ is an N^B-stationary point $\Leftrightarrow$ $\mathbf{x}^*$ is a T^B-stationary point;
(ii) If $\|\mathbf{x}^*\|_{2,0} < k$, then $\mathbf{x}^*$ is an N^B-stationary point $\Leftrightarrow \nabla f(\mathbf{x}^*) = \mathbf{0} \Leftrightarrow$ $\mathbf{x}^*$ is a T^B-stationary point.

Proof. (i) Let $\|\mathbf{x}^*\|_{2,0} = k$.

On one hand, suppose $\mathbf{x}^* \in S$ is an N^B-stationary point of Problem (1), then

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*),$$

that is, $-\nabla f(\mathbf{x}^*) \in N_S^B(\mathbf{x}^*)$. By Theorem 2, $N_S^B(\mathbf{x}^*) = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\mathbf{x}^*)\}$, then we have

$$-\nabla f(\mathbf{x}^*) \in \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\mathbf{x}^*)\},$$

i.e.,

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

It is easy to check that the converse is also true. That is, when $\|\mathbf{x}^*\|_{2,0} = k$, it holds that

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*) \Leftrightarrow (\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases} \tag{7}$$

On the other hand, suppose $\mathbf{x}^* \in S$ is a T^B-stationary point of Problem (1), then

$$0 = \|\nabla_S^B f(\mathbf{x}^*)\|.$$

By Theorem 1, $T_S^B(\mathbf{x}^*) = \{\mathbf{d} \in \mathbb{R}^n : \|\mathbf{d}\|_{2,0} \leq k, \|\mathbf{x}^* + \gamma\mathbf{d}\|_{2,0} \leq k, \forall \gamma \in \mathbb{R}\}$. Hence, in the case of $\|\mathbf{x}^*\|_{2,0} = k$, we have

$$\mathbf{d} \in T_S^B(\mathbf{x}^*) \Leftrightarrow \Gamma(\mathbf{d}) \subseteq \Gamma(\mathbf{x}^*).$$

Accordingly, we have

$$\begin{aligned} \nabla_S^B f(\mathbf{x}^*) &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d} \in T_S^B(\mathbf{x}^*)\} \\ &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \Gamma(\mathbf{d}) \subseteq \Gamma(\mathbf{x}^*)\}. \end{aligned}$$

For $i \notin \Gamma(\mathbf{x}^*)$, $\mathbf{d}_i = \mathbf{0}$, then $\mathbf{0} = (\nabla_S^B f(\mathbf{x}^*))_i$; For $i \in \Gamma(\mathbf{x}^*)$, obviously, $(\nabla_S^B f(\mathbf{x}^*))_i = -(\nabla f(\mathbf{x}^*))_i$. Hence we get

$$(\nabla_S^B f(\mathbf{x}^*))_i = \begin{cases} \mathbf{0}, & i \notin \Gamma(\mathbf{x}^*), \\ -(\nabla f(\mathbf{x}^*))_i, & i \in \Gamma(\mathbf{x}^*), \end{cases}$$

According to $0 = \|\nabla_S^B f(\mathbf{x}^*)\|$, we have

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

It is easy to check that the converse is also true. That is, in the case of $\|\mathbf{x}^*\|_{2,0} = k$, the following equivalence holds

$$0 = \|\nabla_S^B f(\mathbf{x}^*)\| \Leftrightarrow (\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases} \tag{8}$$

Combining (7) with (8), we can conclude that, when $\|\mathbf{x}^*\|_{2,0} = k$, $\mathbf{x}^*$ is an N^B-stationary point of Problem (1) if and only if it is a T^B-stationary point of Problem (1).

(ii) In the case of $\|\mathbf{x}^*\|_{2,0} < k$, we first prove the equivalent relationship between N^B-stationary point of Problem (1) and $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

On one hand, suppose $\mathbf{x}^* \in S$ is an N^B-stationary point of Problem (1), then

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*),$$

that is, $-\nabla f(\mathbf{x}^*) \in N_S^B(\mathbf{x}^*)$. It follows from $N_S^B(\mathbf{x}^*) = \{\mathbf{0}\}$ that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Hence the following implication holds

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*) \Rightarrow \nabla f(\mathbf{x}^*) = \mathbf{0}. \tag{9}$$

On the other hand, suppose $\nabla f(\mathbf{x}^*) = \mathbf{0}$. In the case of $\|\mathbf{x}^*\|_{2,0} < k$, by theorem 1, $N_S^B(\mathbf{x}^*) = \mathbf{0}$. Therefore

$$-\nabla f(\mathbf{x}^*) = \mathbf{0} \in N_S^B(\mathbf{x}^*),$$

i.e., $\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*)$. Hence the following implication holds

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \Rightarrow \mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*). \tag{10}$$

From (9) and (10), we get the following equivalent relationship

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^B(\mathbf{x}^*) \Leftrightarrow \nabla f(\mathbf{x}^*) = \mathbf{0}, \tag{11}$$

that is, in the case of $\|\mathbf{x}^*\|_{2,0} < k$, $\mathbf{x}^*$ is an N^B-stationary point if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

In the following part, we prove the equivalent relationship between T^B-stationary point of Problem (1) and $\nabla f(\mathbf{x}^*) = \mathbf{0}$ in the case of $\|\mathbf{x}^*\|_{2,0} < k$.

Suppose $\mathbf{x}^* \in S$ satisfies $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then by Theorem 1,

$$\begin{aligned}\nabla_S^B f(\mathbf{x}^*) &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d} \in T_S^B(\mathbf{x}^*)\} \\ &= \arg\min\{\|\mathbf{d}\| : \|\mathbf{d}\|_{2,0} \le k, \|\mathbf{x}^* + \gamma\mathbf{d}\|_{2,0} \le k, \forall\gamma \in \mathbb{R}\} \\ &= \mathbf{0}.\end{aligned}$$

That is,

$$\nabla f(\mathbf{x}^*) = \mathbf{0} \Rightarrow 0 = \|\nabla_S^B f(\mathbf{x}^*)\|. \tag{12}$$

Conversely, suppose $\mathbf{x}^*$ is a T^B-stationary point of Problem (1), i.e.,

$$0 = \|\nabla_S^B f(\mathbf{x}^*)\|,$$

then by Theorem 1,

$$\begin{aligned}\mathbf{0} = \nabla_S^B f(\mathbf{x}^*) &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d} \in T_S^B(\mathbf{x}^*)\} \\ &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \|\mathbf{d}\|_{2,0} \le k, \|\mathbf{x}^* + \gamma\mathbf{d}\|_{2,0} \le k, \forall\gamma \in \mathbb{R}\}.\end{aligned}$$

Hence we get that $\|\nabla f(\mathbf{x}^*)\| = \|\mathbf{0} + \nabla f(\mathbf{x}^*)\| \le \|\mathbf{d} + \nabla f(\mathbf{x}^*)\|$ for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $\|\mathbf{d}\|_{2,0} \le k$, $\|\mathbf{x}^* + \gamma\mathbf{d}\|_{2,0} \le k$, $\forall\gamma \in \mathbb{R}$.

For any $i_0 \in \{1, 2, \cdots, m\}$, take $\hat{\mathbf{d}} \in \mathbb{R}^n$ such that $\Gamma(\hat{\mathbf{d}}) = \{i_0\}$ and $\hat{\mathbf{d}}_{i_0} = -(\nabla f(\mathbf{x}^*))_{i_0}$. Following from $|\Gamma(\mathbf{x}^*)| = \|\mathbf{x}^*\|_{2,0} < k$, we have

$$\|\mathbf{x}^* + \gamma\hat{\mathbf{d}}\|_{2,0} = |\Gamma(\mathbf{x}^*) \cup \{i_0\}| \le |\Gamma(\mathbf{x}^*)| + 1 \le k.$$

From $\|\nabla f(\mathbf{x}^*)\| \leq \|\hat{\mathbf{d}} + \nabla f(\mathbf{x}^*)\|$, we obtain $\|(\nabla f(\mathbf{x}^*))_{i_0}\| \leq \|-(\nabla f(\mathbf{x}^*))_{i_0} + (\nabla f(\mathbf{x}^*))_{i_0}\|$, and then

$$(\nabla f(\mathbf{x}^*))_{i_0} = \mathbf{0}.$$

According to the arbitrariness of i_0, we get $\nabla f(\mathbf{x}^*) = \mathbf{0}$. That is,

$$\mathbf{0} = \|\nabla_S^B f(\mathbf{x}^*)\| \Rightarrow \nabla f(\mathbf{x}^*) = \mathbf{0}, \tag{13}$$

Combining (12) with (13), in the case of $\|\mathbf{x}^*\| < k$, the following equivalent relationship holds

$$0 = \|\nabla_S^B f(\mathbf{x}^*)\| \Leftrightarrow \nabla f(\mathbf{x}^*) = \mathbf{0}.$$

The proof is thus finished. □

Furthermore, for Problem (1), its N^C-stationary point and T^C-stationary point have the following equivalent relationship.

Theorem 4. *For Problem* (1)*, let* $\mathbf{x}^* \in S$*, then* $\mathbf{x}^*$ *is an* N^C*-stationary point if and only if it is a* T^C*-stationary point.*

Proof. On one hand, by Theorem 2, $N_S^C(\mathbf{x}^*) = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u}_i = \mathbf{0}, i \in \Gamma(\mathbf{x}^*)\}$. Then we have the following equivalences:

$$\begin{aligned}
&\mathbf{x}^* \text{ is an } N^C\text{-stationary point of Problem (1)} \\
&\Leftrightarrow \mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^C(\mathbf{x}^*) \\
&\Leftrightarrow -\nabla f(\mathbf{x}^*) \in N_S^C(\mathbf{x}^*) \\
&\Leftrightarrow (\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}
\end{aligned} \tag{14}$$

On the other hand, by Theorem 2, $T_S^C(\mathbf{x}^*) = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\}$. Then according to the definition of $\nabla_S^C f(\mathbf{x}^*)$, we have that

$$\begin{aligned}
\nabla_S^C f(\mathbf{x}^*) &= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d} \in T_S^C(\mathbf{x}^*)\} \\
&= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\| : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\} \\
&= \arg\min\{\|\mathbf{d} + \nabla f(\mathbf{x}^*)\|^2 : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\} \\
&= \arg\min\left\{\left(\sum_{i\in\Gamma(\mathbf{x}^*)} + \sum_{i\notin\Gamma(\mathbf{x}^*)}\right)\|\mathbf{d}_i + (\nabla f(\mathbf{x}^*))_i\|^2 : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\right\} \\
&= \arg\min\left\{\sum_{i\in\Gamma(\mathbf{x}^*)}\|\mathbf{d}_i + (\nabla f(\mathbf{x}^*))_i\|^2 + \sum_{i\notin\Gamma(\mathbf{x}^*)}\|(\nabla f(\mathbf{x}^*))_i\|^2 : \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\right\} \\
&= \arg\min\left\{\sum_{i\in\Gamma(\mathbf{x}^*)}\|\mathbf{d}_i + (\nabla f(\mathbf{x}^*))_i\|^2 : \mathbf{d}_i \in \mathbb{R}^{n_i}, i \in \Gamma(\mathbf{x}^*); \mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*)\right\}.
\end{aligned}$$

Thus by directly computing, $\nabla_S^C f(\mathbf{x}^*)$ satisfies

$$(\nabla_S^C f(\mathbf{x}^*))_i = \begin{cases} -(\nabla f(\mathbf{x}^*))_i, & i \in \Gamma(\mathbf{x}^*), \\ \mathbf{0}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

Therefore, the following equivalent relationships hold:

$$\begin{aligned} &\mathbf{x}^* \text{ is a } T^C\text{-stationary point of Problem (1)} \\ &\Leftrightarrow \mathbf{0} = \nabla_S^C f(\mathbf{x}^*) \\ &\Leftrightarrow -(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases} \end{aligned} \tag{15}$$

Combine (14) and (15), then we get the following equivalent relationships:

$$\mathbf{0} \in \nabla f(\mathbf{x}^*) + N_S^C(\mathbf{x}^*) \Leftrightarrow (\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*), \end{cases} \Leftrightarrow \|\nabla_S^C f(\mathbf{x}^*)\| = 0.$$

The proof is thus complete. □

Next, we investigate the relationship among the four types of stationary points of Problem (1).

Theorem 5. *Let $\mathbf{x}^* \in S$, then the following statements hold for Problem (1):*
(i) If $\mathbf{x}^*$ is an N^B-stationary point, then it must be an N^C-stationary point;
(ii) If $\mathbf{x}^*$ is a T^B-stationary point, then it must be a T^C-stationary point.

Proof. (i) Let $\mathbf{x}^*$ is an N^B-stationary point of Problem (1). There are two cases: $\|\mathbf{x}^*\|_{2,0} = k$ and $\|\mathbf{x}^*\|_{2,0} < k$.

Case 1: $\|\mathbf{x}^*\|_{2,0} = k$. In this case, by (7), x^* is an N^B-stationary point if and only if

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*), \end{cases}$$

which, by (14), is equivalent to that $\mathbf{x}^*$ is an N^C-stationary point of Problem (1). Thus we obtain that N^B-stationary point and N^C-stationary point are equivalent in the case of $\|\mathbf{x}^*\|_{2,0} = k$.

Case 2: $\|\mathbf{x}^*\|_{2,0} < k$. By (11), in this case, $\mathbf{x}^*$ is an N^B-stationary point of Problem (1) if and only if

$$\nabla f(\mathbf{x}^*) = \mathbf{0}.$$

By (14), $\mathbf{x}^*$ is an N^C-stationary point of Problem (1) if and only if

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

Clearly, in the case of $\|\mathbf{x}^*\|_{2,0} < k$, if $\mathbf{x}^*$ is an N^B-stationary point of Problem (1), it must be an N^C-stationary point (the converse is not true). That is,

$$N^B-\text{stationary point} \Rightarrow \text{N}^\text{C}-\text{stationary point}. \tag{16}$$

(ii) According to Theorems 3 and 4, the N^B-stationary point of Problem (1) is equivalent to its T^B-stationary point, and the N^C-stationary point of Problem (1) is equivalent to its T^C-stationary point, this is,

$$N^B-\text{stationary point} \Leftrightarrow T^B-\text{stationary point};$$

$$N^C-\text{stationary point} \Leftrightarrow T^C-\text{stationary point}.$$

Moreover, from (16),

$$N^B-\text{stationary point} \Rightarrow N^C-\text{stationary point}.$$

Therefore,

$$T^B-\text{stationary point} \Rightarrow T^C-\text{stationary point}.$$

The proof is finished. □

To have a clear presentation, based on the proofs of Theorems 3 and 4, we use Table 1 to display the characterizations of the four types of stationary points of Problem (1).

Table 1. The characterizations of T^B-, N^B-, T^C-, N^C- stationary point for Problem (1).

Stationary Point	$\|\mathbf{x}^*\|_{2,0} = k$	$\|\mathbf{x}^*\|_{2,0} < k$
T^B-stationary point	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$
N^B-stationary point	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$
T^C-stationary point	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$
N^C-stationary point	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$	$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*) \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*) \end{cases}$

In the end of this section, we discuss the relationship between the local minimizers of Problem (1) and its stationary points.

Theorem 6. *Let $\mathbf{x}^* \in S$ be a local minimizer of Problem (1), then the following two statements hold:*
(i) $\mathbf{x}^*$ is an N^B-stationary point and hence an N^C-stationary point;
(ii) $\mathbf{x}^*$ is a T^B-stationary point and hence a T^C-stationary point.

Proof. Since $\mathbf{x}^*$ is a local minimizer of Problem (1), for sufficiently small $\alpha > 0$, it holds that

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \alpha e_{ij}), \quad \forall i \in J \supseteq \Gamma(x^*), |J| = k;\ j = 1, \cdots, n_i,$$

and then

$$0 \in \arg\min\{h_{ij}(\alpha) \triangleq f(\mathbf{x}^* + \alpha e_{ij}) : \alpha \geq 0\}, \quad \forall i \in J \supseteq \Gamma(x^*), |J| = k;\ j = 1, \cdots, n_i.$$

Due to $\mathbf{x}^* \in S$, there are two cases: $\|\mathbf{x}^*\|_{2,0} < k$ and $\|\mathbf{x}^*\|_{2,0} = k$.
Case 1: $\|\mathbf{x}^*\|_{2,0} < k$. In this case, $\bigcup_{J \supseteq \Gamma(x^*), |J|=k} J = \{1, \cdots, m\}$, then

$$0 \in \arg\min\{h_{ij}(\alpha) = f(\mathbf{x}^* + \alpha e_{ij}) : \alpha \geq 0\}, \quad \forall i = 1, \cdots, m;\ \forall j = 1, \cdots, n_i.$$

By the optimality conditions for the above problems, we have

$$(\nabla f(\mathbf{x}^*))_{ij} = h'_{ij}(0) = 0, \quad \forall i = 1, \cdots, m;\ \forall j = 1, \cdots, n_i.$$

That is, $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
Case 2: $\|\mathbf{x}^*\|_{2,0} = k$. In this case,

$$0 \in \arg\min\{h_{ij}(\alpha) = f(\mathbf{x}^* + \alpha e_{ij}) : \alpha \geq 0\}, \quad \forall i \in \Gamma(\mathbf{x}^*);\ \forall j = 1, \cdots, n_i.$$

It can be derived that $(\nabla f(\mathbf{x}^*))_{ij} = 0, \forall i \in \Gamma(\mathbf{x}^*), \forall j = 1, \cdots, n_i$. That is, $(\nabla f(\mathbf{x}^*))_i = \mathbf{0}, \forall i \in \Gamma(\mathbf{x}^*)$.

Combining the above two cases with (7) and (11), we know that $\mathbf{x}^*$ is an N^B-stationary point of Problem (1). From Theorem 5, $\mathbf{x}^*$ is also an N^C-stationary point of Problem (1).

(ii) From (i), $\mathbf{x}^*$ is both N^B-stationary point and N^C-stationary point. According to Theorems 3 and 4, $\mathbf{x}^*$ is both T^B-stationary point and T^C-stationary point. The proof is complete. □

As a summary of this section, we conclude the relationship among local minimizers and the four stationary points of Problem (1) as follows:

$$\begin{array}{lcll} \text{local minimizer} & \Rightarrow & N^B\text{-stationary point} & \Leftrightarrow \; T^B\text{-stationary point} \\ & & \Downarrow & \Downarrow \\ & & N^C\text{-stationary point} & \Leftrightarrow \; T^C\text{-stationary point.} \end{array}$$

5. Second-Order Optimality Conditions for Problem (1)

In this section, we provide some second-order necessary or sufficient optimality conditions for Problem (1) by use of Clarke tangent cone.

Theorem 7 (**Second-order necessary condition**). *Let $\mathbf{x}^* \in S$ be a local minimizer of Problem (1), then for any $\mathbf{d} \in T_S^C(\mathbf{x}^*)$, it must hold that $\mathbf{d}^\top \nabla f(\mathbf{x}^*) = 0$ and*

$$\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} \geq 0,$$

where $\nabla^2 f(\mathbf{x}^*)$ is the Hessian matrix of f at $\mathbf{x}^*$.

Proof. Since $\mathbf{x}^* \in S$ is a local minimizer of Problem (1), by Theorem 6, $\mathbf{x}^*$ is also an N^C-stationary point. By (14),

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

According to (5), for any $\mathbf{d} \in T_S^C(\mathbf{x}^*)$,

$$\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*).$$

Thus, for any $\mathbf{d} \in T_S^C(\mathbf{x}^*)$, it holds

$$\mathbf{d}^\top \nabla f(\mathbf{x}^*) = 0. \tag{17}$$

In addition, since $\mathbf{x}^*$ is a local minimizer of Problem (1), for sufficiently small $\alpha > 0$ and any $\mathbf{d} \in T_S^C(\mathbf{x}^*)$, we have

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \alpha\mathbf{d}). \tag{18}$$

By Taylor's Theorem,

$$f(\mathbf{x}^* + \alpha\mathbf{d}) = f(\mathbf{x}^*) + \alpha\mathbf{d}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2). \tag{19}$$

Combine (17)–(19), then

$$\begin{aligned} f(\mathbf{x}^*) \leq & f(\mathbf{x}^*) + \alpha\mathbf{d}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2) \\ = & f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2). \end{aligned}$$

Hence,

$$0 \leq \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2),$$

which implies $\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} \geq 0, \forall \mathbf{d} \in T_S^C(\mathbf{x}^*)$. The desired result is derived. □

Finally, we give a second-order sufficient condition for the optimality of Problem (1).

Theorem 8 (Second-order sufficient condition). *Let $\mathbf{x}^* \in S$ be an N^C-stationary point of Problem (1), if for any $\mathbf{d} \in T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\}$, it holds $\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} > 0$, then the following two statements hold:*
(i) $\mathbf{x}^* \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)}$ is a strictly local minimizer of Problem (1);
(ii) $\mathbf{x}^*$ satisfies the second-order growth condition, that is, there are $\omega > 0$ and $\delta > 0$ such that for any $\mathbf{x} \in B(\mathbf{x}^*, \delta) \bigcap \mathbb{R}^n_{\Gamma(\mathbf{x}^*)}$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \omega \|\mathbf{x} - \mathbf{x}^*\|^2.$$

where $\mathbb{R}^n_{\Gamma(\mathbf{x}^*)} = \text{span}\{e_{ij}, i \in \Gamma(\mathbf{x}^*), j = 1, \cdots, n_i\}$.

Proof. (i) Since $\mathbf{x}^*$ is an N^C-stationary point of Problem (1), from Theorem 2, we have

$$(\nabla f(\mathbf{x}^*))_i \begin{cases} = \mathbf{0}, & i \in \Gamma(\mathbf{x}^*), \\ \in \mathbb{R}^{n_i}, & i \notin \Gamma(\mathbf{x}^*). \end{cases}$$

For any $\mathbf{d} \in T_S^C(\mathbf{x}^*)$, by (5),

$$\mathbf{d}_i = \mathbf{0}, i \notin \Gamma(\mathbf{x}^*).$$

Then for any $\mathbf{d} \in T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\}$, it holds

$$\mathbf{d}^\top \nabla f(\mathbf{x}^*) = 0.$$

By Taylor's Theorem, for any sufficiently small $\alpha > 0$,

$$\begin{aligned} f(\mathbf{x}^* + \alpha \mathbf{d}) =& f(\mathbf{x}^*) + \alpha \mathbf{d}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2) \\ =& f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2). \end{aligned}$$

Since $\mathbf{d}^T \nabla^2 f(\mathbf{x}^*)\mathbf{d} > 0, \forall \mathbf{d} \in T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\}$, then for any sufficiently small $\alpha > 0$,

$$f(\mathbf{x}^* + \alpha \mathbf{d}) = f(\mathbf{x}^*) + \frac{1}{2}\alpha^2 \mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} + o(\alpha^2) > f(\mathbf{x}^*).$$

Therefore, $\mathbf{x}^*$ is a strictly local minimizer of Problem (1).

(ii) Assume, on the contrary, that the second-order growth condition does not hold at $\mathbf{x}^*$, then there is a sequence $\{\mathbf{x}^t\}_{t \in \mathbb{N}} \subset \mathbb{R}^n_{\Gamma(\mathbf{x}^*)}$ such that $\{\mathbf{x}^t\}_{t \in \mathbb{N}} \to \mathbf{x}^*$ but

$$f(\mathbf{x}^t) < f(\mathbf{x}^*) + \frac{1}{t}\|\mathbf{x}^t - \mathbf{x}^*\|^2.$$

Let $\mathbf{z}^t = \frac{\mathbf{x}^t - \mathbf{x}^*}{\|\mathbf{x}^t - \mathbf{x}^*\|}$, then $\|\mathbf{z}^t\| = 1$. Since $\{\frac{\mathbf{x}^t - \mathbf{x}^*}{\|\mathbf{x}^t - \mathbf{x}^*\|}\}_{t \in \mathbb{N}}$ is bounded, without loss of generality, suppose $\mathbf{z}^t \to \mathbf{z}$, then $\|\mathbf{z}\| = 1$.

It follows $\mathbf{x}^t \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)}$ that $\Gamma(\mathbf{x}^t) \subseteq \Gamma(\mathbf{x}^*)$. Due to $\lim_{t \to \infty} \mathbf{x}^t = \mathbf{x}^*$, we have $\Gamma(\mathbf{x}^*) \subseteq \Gamma(\mathbf{x}^t)$, then

$$\Gamma(\mathbf{x}^t) = \Gamma(\mathbf{x}^*)$$

for any sufficiently large t. From $\mathbf{z}^t = \frac{\mathbf{x}^t - \mathbf{x}^*}{\|\mathbf{x}^t - \mathbf{x}^*\|}$, we get

$$\Gamma(\mathbf{z}^t) \subseteq \Gamma(\mathbf{x}^t) = \Gamma(\mathbf{x}^*) \text{ and } \mathbf{z}^t \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)} \setminus \{\mathbf{0}\}.$$

Moreover, from $\lim_{t\to\infty} \mathbf{z}^t = \mathbf{z}$, it follows

$$\Gamma(\mathbf{z}) \subseteq \Gamma(\mathbf{z}^t) \subseteq \Gamma(\mathbf{x}^*) \text{ and } \mathbf{z} \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)} \setminus \{\mathbf{0}\}$$

for any sufficiently large t. According to (6), it holds

$$\mathbb{R}^n_{\Gamma(\mathbf{x}^*)} = \text{span}\{e_{ij}, i \in \Gamma(\mathbf{x}^*), j = 1, \cdots, n_i\} = T_S^C(\mathbf{x}^*).$$

Hence, for any $\mathbf{z}^t \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)} \setminus \{\mathbf{0}\}$, we have $\mathbf{z}^t \in T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\}$, which together with (5) yields

$$(\mathbf{z}^t)^\top \nabla f(\mathbf{x}^*) = 0.$$

By Taylor's Theorem,

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) = (\mathbf{x}^t - \mathbf{x}^*)^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}^t - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{x}^*)(\mathbf{x}^t - \mathbf{x}^*) + o(\|\mathbf{x}^t - \mathbf{x}^*\|^2).$$

Since $\frac{(\mathbf{x}^t - \mathbf{x}^*)^\top}{\|\mathbf{x}^t - \mathbf{x}^*\|} \nabla f(\mathbf{x}^*) = (\mathbf{z}^t)^\top \nabla f(\mathbf{x}^*) = 0$, we have

$$\begin{aligned}
\frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|^2} &= \frac{1}{\|\mathbf{x}^t - \mathbf{x}^*\|^2}\Big((\mathbf{x}^t - \mathbf{x}^*)^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}^t - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{x}^*)(\mathbf{x}^t - \mathbf{x}^*) \\
&\quad + o(\|\mathbf{x}^t - \mathbf{x}^*\|^2)\Big) \\
&= \frac{\frac{1}{2}(\mathbf{x}^t - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{x}^*)(\mathbf{x}^t - \mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|^2} + o(1) \\
&= \frac{1}{2}(\mathbf{z}^t)^\top \nabla^2 f(\mathbf{x}^*)\mathbf{z}^t + o(1).
\end{aligned}$$

Under the assumption that $f(\mathbf{x}^t) < f(\mathbf{x}^*) + \frac{1}{t}\|\mathbf{x}^t - \mathbf{x}^*\|^2$, we obtain

$$\frac{1}{t} > \frac{f(\mathbf{x}^t) - f(\mathbf{x}^*)}{\|\mathbf{x}^t - \mathbf{x}^*\|^2} = \frac{1}{2}(\mathbf{z}^t)^\top \nabla^2 f(\mathbf{x}^*)\mathbf{z}^t + o(1).$$

Letting $t \to \infty$, we get

$$\mathbf{z}^T \nabla^2 f(\mathbf{x}^*)\mathbf{z} \leq 0, \quad \text{where } \mathbf{z} \in \mathbb{R}^n_{\Gamma(\mathbf{x}^*)} \setminus \{\mathbf{0}\} = T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\},$$

which contradicts the condition that $\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{d} > 0$ holds for any $\mathbf{d} \in T_S^C(\mathbf{x}^*) \setminus \{\mathbf{0}\}$. Therefore, the second-order growth condition must hold at $\mathbf{x}^*$. □

6. Concluding Remarks

In this paper, the first-order optimality conditions are built for group sparsity constrained optimization problems by use of Bouligand tangent cone, Clarke tangent and their corresponding normal cones, and the relationship among the local minimizers and the four types of stationary points of Problem (1) is investigated. Furthermore, the second-order sufficient and second-order necessary optimality conditions for group sparsity constrained optimization problems are provided. The results show that N^C-stationary points of Problem (1) may be strictly local minimizers, and even can fulfill the second-order growth condition under some mild conditions. The results provide the theoretical basis for analyz-

ing or solving the group sparsity constrained optimization problems. In the future, we will use the optimality conditions to design algorithms for solving the problems.

Author Contributions: Methodology, D.P.; Project administration, D.P.; Supervision, D.P.; Writing original draft, W.W.; Writing review and editing, D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by NSFC (11861020), the Growth Project of Education Department of Guizhou Province for Young Talents in Science and Technology ([2018]121), the Foundation for Selected Excellent Project of Guizhou Province for High-level Talents Back from Overseas ([2018]03), and the Science and Technology Planning Project of Guizhou Province ([2018]5781).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
2. Huang, J.; Breheny, P.; Ma, S. A selective review of group selection in high-dimensional models. *Stat. Sci.* **2012**, *27*, 481–499. [CrossRef] [PubMed]
3. Huang, J.; Ma, S.; Xue, H.; Zhang, C.H. A group bridge approach for variable selection. *Biometrika* **2009**, *96*, 339–355. [CrossRef] [PubMed]
4. Meier, L.; van de Geer, S.; Bühlmann, P. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 53–71. [CrossRef]
5. Yang, Y.; Zou, H. A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.* **2015**, *25*, 1129–1141. [CrossRef]
6. Beck, A.; Hallak, N. Optimization involving group sparsity terms. *Math. Program.* **2018**, *178*, 39–67. [CrossRef]
7. Hu, Y.; Li, C.; Meng, K.; Qin, J.; Yang, X. Group sparse optimization via $\ell_{p,q}$ regularization. *J. Mach. Learn. Res.* **2017**, *18*, 1–52.
8. Jiao, Y.; Jin, B.; Lu, X. Group sparse recovery via the $\ell_0(\ell_2)$ penalty: theory and algorithm. *IEEE Trans. Signal Process.* **2017**, *65*, 998–1012. [CrossRef]
9. Huang, J.; Zhang, T. The benefit of group sparsity. *Ann. Stat.* **2010**, *38*, 1978–2004. [CrossRef]
10. Agarwal, A.; Negahban, S.; Wainwright, M.J. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Int. Conf. Neural Inf. Process. Syst.* **2010**, *23*, 37–45.
11. Attouch, H.; Bolte, J.; Svaiter, B.F. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.* **2013**, *137*, 91–129. [CrossRef]
12. Beck, A.; Eldar, Y. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM J. Optim.* **2013**, *23*, 1480–1509. [CrossRef]
13. Calamai, P.H.; More, J.J. Projection gradient methods for linearly constrained problems. *Math. Program.* **1987**, *39*, 93–116. [CrossRef]
14. Pan, L.L.; Xiu, N.H.; Zhou, S.L. On Solutions of Sparsity Constrained Optimization. *J. Oper. Res. Soc. China* **2015**, *3*, 421–439. [CrossRef]
15. Chen, X. J.; Pan, L.L.; Xiu, N.H. Solution sets of three sparse optimization problems for multivariate regression. *Appl. Comput. Harmon. A* **2020**, revised.
16. Bian, W.; Chen, X.J. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM J. Numer. Anal.* **2020**, *58*, 858–883. [CrossRef]
17. Peng, D.T.; Chen, X.J. Computation of second-order directional stationary points for group sparse optimization. *Optim. Methods Softw.* **2020**, *35*, 348–376. [CrossRef]
18. Pan, L.L.; Chen, X.J. Group sparse optimization for images recovery using capped folded concave functions. *SIAM J. Imaging Sci.* **2021**. Available online: https://www.polyu.edu.hk/ama/staff/xjchen/Re_gsparseAugust.pdf (accessed on 5 November 2020).
19. Rockafellar, R.T.; Wets, R.J. *Variational Analysis*; Springer: Berlin/Heidelberg, Germany, 2009.