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

Continuous exact relaxation and alternating proximal gradient algorithm for partial sparse and partial group sparse optimization problems

Qingqing Wu, Dingtao Peng, Xian Zhang

Abstract In this paper, we consider a partial sparse and partial group sparse optimization problem, where the loss function is a continuously differentiable function (possibly nonconvex), and the penalty term consists of two parts associated with sparsity and group sparsity. The first part is the ℓ_0 norm of $\mathbf{x}$, the second part is the $\ell_{2,0}$ norm of $\mathbf{y}$, i.e, $\lambda_1\|\mathbf{x}\|_0+\lambda_2\|\mathbf{y}\|_{2,0}$, where $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{n+m}$ is the decision variable. We give a continuous relaxation model of the above original problem, where the two parts of the penalty term are relaxed by Capped-ℓ_1 of $\mathbf{x}$ and group Capped-ℓ_1 of $\mathbf{y}$ respectively. Firstly, we define two kinds of stationary points of the relaxation model. Based on the lower bound property of d-stationary points of the relaxation model, we establish the equivalence of solutions of the original problem and the relaxation model, which provides a theoretical basis for solving the original problem via solving the relaxation problem. Secondly, we propose an alternating proximal gradient (APG) algorithm to solve the relaxation model, and prove that the whole sequence of the APG algorithm converges to a critical point under some mild conditions. Finally, numerical experiments on simulated data and multichannel image as well as comparison with some state-of-art algorithms are presented to illustrate the effectiveness and robustness of the proposed algorithm for partial sparse and partial group sparse optimization problem.

Keywords Partial sparse and partial group sparse optimization problem; continuous exact relaxation; stationary point; alternating proximal gradient algorithm; whole sequence convergence

MSC(2010) 90C26 · 90C46

1 Introduction

In the past decade, sparse optimization problems have attracted great attention in variable selection, image restoration, gene expression, and so on [5, 10, 14, 20, 21, 22, 39, 45]. The basic framework of sparse optimization problem is to seek a sparse solution of an underde-

Dingtao Peng ✉
School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China. E-mail: dingtaopeng@126.com
Qingqing Wu
School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China. E-mail: gs.qqwu21@gzu.edu.cn
Xian Zhang
School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China. E-mail: zhangxian05@163.com

termined linear system. The general sparse optimization problem is as follows:

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

where $f : \mathbb{R}^n \to \mathbb{R}_+$ is a loss function, $\lambda > 0$, $\|\mathbf{x}\|_0 := \sum_{i=1,x_i\neq 0}^{n} |x_i|^0$. A vector $\mathbf{x} \in \mathbb{R}^n$ is said to be sparse if $\|\mathbf{x}\|_0 \ll n$, and the sparsity of vector $\mathbf{x} \in \mathbb{R}^n$ is usually provided by its ℓ_0 norm.

Due to the fact that traditional sparse optimization problems only consider the sparsity of a single item and do not have sufficient ability to handle complex structures such as group sparse structures, Yuan and Lin [39] first use group sparse structures as prior information. Group sparse structure refers to dividing variables into multiple groups, and then considering whether each group as a whole is zero. Let $\mathbf{x} = (\mathbf{x}_{(1)}^\top, \cdots, \mathbf{x}_{(J)}^\top)^\top$ with J disjoint groups, where $\mathbf{x}_{(i)} = (x_{(i)1}, \cdots, x_{(i)n_i})^\top \in \mathbb{R}^{n_i}$, $n_i > 0$ and $\sum_{i=1}^{J} n_i = n$. Then the optimization problem with group sparse structure can be formulated as the following group sparse optimization [24,30,31]:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ F(\mathbf{x}) = f(\mathbf{x}) + \lambda\|\mathbf{x}\|_{2,0},$$

where $\|\mathbf{x}\|_{2,0} := \sharp\{i \mid \|\mathbf{x}_{(i)}\| \neq 0, i = 1, \cdots, J\}$ is called $\ell_{2,0}$ norm that counts the number of nonzero groups of $\mathbf{x}$, in which $\|\mathbf{x}_{(i)}\|$ denotes the ℓ_2 norm of the subvector $\mathbf{x}_{(i)}$. Note that $\|\cdot\|_{2,0}$ is nonconvex, nonsmooth, and even discontinuous, which causes the above problem to be NP-hard. Many researchers consider the relaxation problem of this problem, such as group LASSO model [35], Bayes group LASSO models [9,33], group SCAD model [23,31, 38], group MCP model [31,40] and other models [30,32,42,44,46].

When the data consist of two parts such that the first part has a sparse structure and the second part has a certain group sparse structure, it naturally makes sense for us to investigate the following partial sparse and partial group sparse optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m} \ F(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}) + \lambda_1\|\mathbf{x}\|_0 + \lambda_2\|\mathbf{y}\|_{2,0}, \tag{1.1}$$

where $f(\mathbf{x},\mathbf{y})$ is a loss function which we suppose it to be continuously differentiable but not necessarily convex in this paper. In (1.1), λ_1, $\lambda_2 > 0$, $\|\mathbf{x}\|_0 = \sum_{i=0,x_i\neq 0}^{n} |x_i|^0$ is called ℓ_0 norm of $\mathbf{x}$, $\mathbf{y} = (\mathbf{y}_{(1)}^\top, \cdots, \mathbf{y}_{(J)}^\top)^\top \in \mathbb{R}^m$ with J disjoint groups, and $\|\mathbf{y}\|_{2,0} = \sharp\{j \mid \|\mathbf{y}_{(j)}\| \neq 0, j = 1, \cdots, J\}$ is called $\ell_{2,0}$ norm of $\mathbf{y}$. Specially, if $\mathbf{x}$ and $\mathbf{y}$ are same, problem (1.1) degrades to the following sparse plus group sparse optimization problem [26]:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ F(\mathbf{x}) = f(\mathbf{x}) + \lambda_1\|\mathbf{x}\|_0 + \lambda_2\|\mathbf{x}\|_{2,0}.$$

Since both $\|\cdot\|$ and $\|\cdot\|_{2,0}$ are nonconvex, nonsmooth and discontinuous, problem (1.1) in general is NP-hard. One popular way is to relax ℓ_0 ($\ell_{2,0}$) norm to ℓ_1 ($\ell_{2,1}$) norm which are convex [35,44], but the solution obtained by the relaxation problem is biased and does not satisfy oracle property [16,17]. Therefore, some researchers propose using several classes of folding concave continuous relaxations which are still nonconvex but have some good properties. These nonconvex relaxations includes ℓ_p $(0 < p < 1)$ norm, smoothly clipped absolute deviation (SCAD) penalty [17], minimax concave penalty (MCP) [40], Capped-ℓ_1 penalty [28,41] and their corresponding group structure forms, such as $\ell_{p,q}$, group SCAD and group MCP. The nonconvex relaxations have been widely studied in many works, for example [3,4,11,30,36,37,42]. It has been proved that the solutions obtained by these kinds of nonconvex optimization have some desired properties: unbiasedness, sparsity, continuity and oracle property. Specially, reference [25] has shown that Capped-ℓ_1 relaxation is the tightest difference-of-convex (DC) relaxation for ℓ_0 norm.

In this paper, we consider using Capped-ℓ_1 and group Capped-ℓ_1 to relax ℓ_0 norm and $\ell_{2,0}$ norm in problem (1.1) respectively, that is, we consider the following problem

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m} \quad F(\mathbf{x},\mathbf{y}) := f(\mathbf{x},\mathbf{y}) + \lambda_1 \Phi_1(\mathbf{x}) + \lambda_2 \Phi_2(\mathbf{y}), \tag{1.2}$$

where

$$\Phi_1(\mathbf{x}) := \sum_{i=1}^{n} \varphi_1(|x_i|), \quad \Phi_2(\mathbf{y}) := \sum_{j=1}^{J} \varphi_2(\|\mathbf{y}_{(j)}\|),$$

which are Capped-ℓ_1 regularization and group Capped-ℓ_1 regularization respectively, and

$$\varphi_\upsilon(t) := \min\left\{1, \frac{t}{\alpha_\upsilon}\right\} = \frac{t}{\alpha_\upsilon} - \max\left\{0, \frac{t}{\alpha_\upsilon} - 1\right\} = \begin{cases} \frac{t}{\alpha_\upsilon}, & \text{if } 0 \le t < \alpha_\upsilon, \\ 1, & \text{if } t \ge \alpha_\upsilon, \end{cases}$$

with $\alpha_\upsilon > 0$, $\upsilon = 1, 2$. The penalty function $\varphi_\upsilon : \mathbb{R}_+ \to \mathbb{R}_+$ can be written in the form of DC form as $\varphi_\upsilon(t) := g_\upsilon(t) - h_\upsilon(t)$ with $g_\upsilon(t) = \frac{t}{\alpha_\upsilon}$, $h_\upsilon(t) = \max\{0, \frac{t}{\alpha_\upsilon} - 1\}$. Therefore, problem (1.2) can be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m} F(\mathbf{x},\mathbf{y}) =& f(\mathbf{x},\mathbf{y}) + \lambda_1 \sum_{i=1}^{n} (g_1(|x_i|) - h_1(|x_i|)) \\ &+ \lambda_2 \sum_{j=1}^{J} \big(g_2(\|\mathbf{y}_{(j)}\|) - h_2(\|\mathbf{y}_{(j)}\|)\big). \end{aligned} \tag{1.3}$$

In recent years, many scholars have studied the relaxation models of sparse or group sparse optimization problems. For the sparse optimization problem with the linear least square loss and ℓ_p regularization, the reference [12] established the lower bound property of nonzero entires of local solutions. When the loss function is convex and the constraint set is a box, the reference [3] studied the relationship between the original ℓ_0 regularization problem and the Capped-ℓ_1 relaxation problem. Under certain conditions, the equivalence of global solutions and the inclusion relationship of local solutions between the two problems are proved. The authors also proposed a smoothing proximal gradient algorithm for solving the relaxation problem. The reference [31] considered a class of group sparse optimization problems with nonconvex folding concave continuous relaxations, and researched the first-order and second-order directional stationary points of the problem. The reference [30] considered three kinds of group sparse optimization models with linear inequality constraints and discussed the relationship between stationary points, local solutions and global solutions. The reference [42] considered a class of group sparse optimization models with a general constraint set, and discussed the relationship of local solutions and global solutions between original problem and relaxation problem.

In this paper, inspired by the above works, we study the stationary points of problem (1.2), the equivalence of solutions between problems (1.1) and (1.2), and provide an efficient algorithm for solving problem (1.2).

This paper is organized as follows. In Section 2, we give some preliminaries that will be used in this paper. In Section 3, we define two classes of stationary points for the relaxation model and discuss their characterization, relationship and some properties. In Section 4, we establish the equivalence of solutions between the original problem (1.1) and the relaxation model (1.2). In Section 5, we propose an APG algorithm for problem (1.2) and establish the convergence result of the whole sequence. In Section 6, we test the proposed APG algorithm through rich numerical experiments on recovering the simulated partial sparse and partial group sparse signals and some real images. In Section 7, we make a brief conclusion of this paper.

2 Notations and preliminaries

In this section, we provide some basic notations, and introduce the preliminaries of several kinds of stationary points and subdifferentials.

Notations: For any $n \in \mathbb{N}^+$, $[n] := \{1, \cdots, n\}$. For any $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\nabla f(\mathbf{x}, \mathbf{y}) = \big(\nabla_{\mathbf{x}}^\top f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}}^\top f(\mathbf{x}, \mathbf{y})\big)^\top$, where $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) = \big([\nabla_{\mathbf{y}}^\top f(\mathbf{x}, \mathbf{y})]_{(1)}, \cdots, [\nabla_{\mathbf{y}}^\top f(\mathbf{x}, \mathbf{y})]_{(J)}\big)^\top$, and $[\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]_{(j)} = \big([\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]_{(j)1}, \cdots, [\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]_{(j)m_j}\big)^\top$. For convenience, we define the following index sets

$$
\begin{aligned}
I_1(\mathbf{x}) &:= \{i : |x_i| = 0, \forall i \in [n]\},\\
I_2(\mathbf{x}) &:= \{i : 0 < |x_i| < \alpha_1, \forall i \in [n]\},\\
I_3(\mathbf{x}) &:= \{i : |x_i| = \alpha_1, \forall i \in [n]\},\\
I_4(\mathbf{x}) &:= \{i : |x_i| > \alpha_1, \forall i \in [n]\},\\
J_1(\mathbf{y}) &:= \{j : \|\mathbf{y}_{(j)}\| = 0, \forall j \in [J]\},\\
J_2(\mathbf{y}) &:= \{j : 0 < \|\mathbf{y}_{(j)}\| < \alpha_2, \forall j \in [J]\},\\
J_3(\mathbf{y}) &:= \{j : \|\mathbf{y}_{(j)}\| = \alpha_2, \forall j \in [J]\},\\
J_4(\mathbf{y}) &:= \{j : \|\mathbf{y}_{(j)}\| > \alpha_2, \forall j \in [J]\}.
\end{aligned}
$$

Let $I(\mathbf{x}) := I_2(\mathbf{x}) \cup I_3(\mathbf{x}) \cup I_4(\mathbf{x})$, and $J(\mathbf{y}) := J_2(\mathbf{y}) \cup J_3(\mathbf{y}) \cup J_4(\mathbf{y})$. Denote

$$
\ell(x_i) := |x_i|, \quad \rho_j(\mathbf{y}_{(j)}) := \|\mathbf{y}_{(j)}\|,
$$

then problem (1.3) can be rewritten as follows

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m} F(\mathbf{x}, \mathbf{y}) =& f(\mathbf{x}, \mathbf{y}) + \lambda_1 \sum_{i=1}^{n} (g_1 \circ \ell(x_i) - h_1 \circ \ell(x_i))\\
&+ \lambda_2 \sum_{j=1}^{J} (g_2 \circ \rho_j(\mathbf{y}_{(j)}) - h_2 \circ \rho_j(\mathbf{y}_{(j)})),
\end{aligned} \tag{2.1}
$$

where "$\circ$" denotes the composition of two functions.

Next, we introduce several important concepts to characterize optimal conditions of problem (1.2).

Definition 2.1 *[13, 31] Let $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$, for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$, the directional derivative of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ is defined as*

$$
h'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) := \lim_{t \downarrow 0} \frac{h((\hat{\mathbf{x}}, \hat{\mathbf{y}}) + t(\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) - h(\hat{\mathbf{x}}, \hat{\mathbf{y}})}{t}.
$$

If h is differentiable at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, then $h'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) = \langle \nabla h(\hat{\mathbf{x}}, \hat{\mathbf{y}}), (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle$.

By the definition, for any $(\mathbf{x}, \mathbf{y})$, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$, we can get

$$\ell'(\hat{x}_i; x_i - \hat{x}_i) = \begin{cases} |x_i|, & \text{if } i \in I_1(\hat{\mathbf{x}}), \\ \operatorname{sgn}(\hat{x}_i)(x_i - \hat{x}_i), & \text{otherwise}, \end{cases} \tag{2.2}$$

and

$$\rho_j'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) = \begin{cases} \|\mathbf{y}_{(j)}\|, & \text{if } j \in J_1(\hat{\mathbf{y}}), \\ \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|}, & \text{otherwise}, \end{cases} \tag{2.3}$$

where

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

Definition 2.2 *[13] Let $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$ be locally Lipschitz at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$,for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$ near $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, the generalized directional derivative of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is defined as*

$$h^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) := \limsup_{\substack{(\mathbf{x}, \mathbf{y}) \to (\hat{\mathbf{x}}, \hat{\mathbf{y}}) \\ t \downarrow 0}} \frac{h((\mathbf{x}, \mathbf{y}) + t(\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) - h(\mathbf{x}, \mathbf{y})}{t}.$$

As we all know, the existence of the generalized directional derivative does not imply the existence of the directional derivative. But if the directional derivative exists, then

$$h'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) \le h^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})), \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}. \tag{2.4}$$

Next, we introduce several types of definitions of subdifferential.

Definition 2.3 *[34] Let $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$ be a proper convex function, the subdifferential $\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \operatorname{dom} h$ is the set of $\xi \in \partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, called subgradients of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, such that*

$$h(\mathbf{x}, \mathbf{y}) \ge h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) + \langle \xi, (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle, \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}.$$

If $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \notin \operatorname{dom} h$, then $\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \emptyset$.

Definition 2.4 *[13] Let $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$ be a locally Lipschitz function. The Clarke subdifferential of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \operatorname{dom} h$, written $\partial^C h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, is defined as*

$$con\{\xi \in \mathbb{R}^{n+m} | \langle \xi, (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle \le h^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})), \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}\},$$

where "con" represents the convex hull of a set.

The above definition implies that [13, Corollary 2.9.1]

$$h^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) = \max_{\xi \in \partial^C h(\hat{\mathbf{x}}, \hat{\mathbf{y}})} \langle \xi, (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle.$$

It is known that [13, Proposition 2.3.6] if h is convex, then $h^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) = h'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}))$ and $\partial^C h(\mathbf{x}, \mathbf{y}) = \partial h(\mathbf{x}, \mathbf{y})$; if h is continuously differentiable, then $\partial^C h(\mathbf{x}, \mathbf{y}) = \{\nabla h(\mathbf{x}, \mathbf{y})\}$.

Since the penalty terms in (1.2) are known as Capped-ℓ_1 functions, we can gain that the objective function F is nonconvex and lower semicontinuous. We now give the definition of limiting subdifferential.

Definition 2.5 *[34] Let $h : \mathbb{R}^{n+m} \to \mathbb{R}\cup\{\infty\}$ be a proper lower semicontinuous function.*

(i) *The Fréchet subdifferential of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathrm{dom}h$, written $\widehat{\partial} h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, is defined as*

$$\{\xi \in \mathbb{R}^{n+m} | h(\mathbf{x}, \mathbf{y}) \geq h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) + \langle \xi, (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})\rangle + o(\|(\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})\|), \forall\, (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}\},$$

If $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \notin \mathrm{dom}h := \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m} \mid h(\mathbf{x}, \mathbf{y}) < \infty\}$, then $\widehat{\partial} h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \emptyset$.

(ii) *The limiting subdifferential of h at $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathrm{dom}h$, written $\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$, is defined as*

$$\{\xi \in \mathbb{R}^{n+m} | \exists (\mathbf{x}^k, \mathbf{y}^k) \to (\hat{\mathbf{x}}, \hat{\mathbf{y}}), h(\mathbf{x}^k, \mathbf{y}^k) \to h(\hat{\mathbf{x}}, \hat{\mathbf{y}}), \xi^k \in \widehat{\partial} h(\mathbf{x}^k, \mathbf{y}^k) \text{ such that } \xi^k \to \xi\},$$

If $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \notin \mathrm{dom}h$, then $\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \emptyset$.

From [34], it is known that if h is locally Lipschitz, then $\partial^C h(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathrm{cl}(\mathrm{con}(\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}})))$ which is the closed convex hull of $\partial h(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. If h is a convex function, then the Fréchet subdifferential, limit subdifferential and Clarke subdifferential of h at $(\mathbf{x}, \mathbf{y})$ are all consistent with the classical subdifferential of convex function.

3 Directional stationary points and critical points of problem (1.2)

The optimality conditions of optimization problems are often characterized by stationary points. In this section, we give the characterization of the d(irectional)-stationary points and the critical points of problem (1.2), and analyze their properties. Then we investigate the relationship between the two types of stationary points.

Based on the DC expression (1.3) of problem (1.2), we give the definition of critical point of problem (1.2).

Definition 3.1 *[29, 34] [critical point] $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ is called a critical point of problem (1.2), if*

$$\begin{aligned} \mathbf{0} \in \quad & \nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) + \lambda_1 \partial \left(\sum_{i=1}^{n} (g_1 \circ \ell)(\hat{x}_i) \right) - \lambda_1 \partial \left(\sum_{i=1}^{n} (h_1 \circ \ell)(\hat{x}_i) \right) \\ & + \lambda_2 \partial \left(\sum_{j=1}^{J} (g_2 \circ \rho_j)(\hat{\mathbf{y}}_{(j)}) \right) - \lambda_2 \partial \left(\sum_{j=1}^{J} (h_2 \circ \rho_j)(\hat{\mathbf{y}}_{(j)}) \right). \end{aligned}$$

The set of critical points of problem (1.2) is denoted by $\mathrm{crit}F$.

Based on this definition, [34, Proposition 10.5] and [42, Theorem 3.4], we give the characterization of critical point of problem (1.2) as follows.

Theorem 3.2 *Let $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a critical point of problem (1.2), then*

$$\begin{aligned} \mathbf{0} \in \nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) & + \lambda_1 \Big\{ \partial (g_1 \circ \ell)(\hat{x}_1) \times \cdots \times \partial (g_1 \circ \ell)(\hat{x}_n) - \partial (h_1 \circ \ell)(\hat{x}_1) \times \cdots \times \partial (h_1 \circ \ell)(\hat{x}_n) \Big\} \\ & + \lambda_2 \Big\{ \partial (g_2 \circ \rho_1)(\hat{\mathbf{y}}_{(1)}) \times \cdots \times \partial (g_2 \circ \rho_J)(\hat{\mathbf{y}}_{(J)}) - \partial (h_2 \circ \rho_1)(\hat{\mathbf{y}}_{(1)}) \times \cdots \times \partial (h_2 \circ \rho_J)(\hat{\mathbf{y}}_{(J)}) \Big\}, \end{aligned}$$

where

$$\partial(g_1\circ\ell)(\hat{x}_i)=\begin{cases}[-\frac{1}{\alpha_1},\frac{1}{\alpha_1}], & \textit{if } i\in I_1(\hat{\mathbf{x}}),\\ \{\frac{1}{\alpha_1}\mathrm{sgn}(\hat{x}_i)\}, & \textit{otherwise},\end{cases}$$

$$\partial(h_1\circ\ell)(\hat{x}_i)=\begin{cases}0, & \textit{if } i\in I_1(\hat{\mathbf{x}})\cup I_2(\hat{\mathbf{x}}),\\ con\{0,\frac{1}{\alpha_2}\mathrm{sgn}(\hat{x}_i)\}, & \textit{if } i\in I_3(\hat{\mathbf{x}}),\\ \frac{1}{\alpha_2}\mathrm{sgn}(\hat{x}_i), & \textit{if } i\in I_4(\hat{\mathbf{x}}),\end{cases}$$

$$\partial(g_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})=\begin{cases}\frac{1}{\alpha_2}B^{m_j}, & \textit{if } j\in J_1(\hat{\mathbf{y}}),\\ \{\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \textit{otherwise},\end{cases}$$

$$\partial(h_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})=\begin{cases}\mathbf{0}, & \textit{if } j\in J_1(\hat{\mathbf{y}})\cup J_2(\hat{\mathbf{y}}),\\ con\{\mathbf{0},\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \textit{if } j\in J_3(\hat{\mathbf{y}}),\\ \{\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \textit{if } j\in J_4(\hat{\mathbf{y}}),\end{cases}$$

in which "con" denotes the convex hull of a set and B^{m_j} *denotes the closed unit ball in* $\mathbb{R}^{m_j}$*.*

Proof According to [34, Proposition 10.5], we get

$$\partial\left(\sum_{i=1}^{n}(g_1\circ\ell)(\hat{x}_i)\right)=\Big\{\partial(g_1\circ\ell)(\hat{x}_1)\times\cdots\times\partial(g_1\circ\ell)(\hat{x}_n)\Big\}$$

$$\partial\left(\sum_{i=1}^{n}(h_1\circ\ell)(\hat{x}_i)\right)=\Big\{\partial(h_1\circ\ell)(\hat{x}_1)\times\cdots\times\partial(h_1\circ\ell)(\hat{x}_n)\Big\}$$

$$\partial\left(\sum_{j=1}^{J}(g_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})\right)=\Big\{\partial(g_2\circ\rho_1)(\hat{\mathbf{y}}_{(1)})\times\cdots\times\partial(g_2\circ\rho_J)(\hat{\mathbf{y}}_{(J)})\Big\}$$

$$\partial\left(\sum_{j=1}^{J}(h_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})\right)=\Big\{\partial(h_2\circ\rho_1)(\hat{\mathbf{y}}_{(1)})\times\cdots\times\partial(h_2\circ\rho_J)(\hat{\mathbf{y}}_{(J)})\Big\}.$$

By the definition of critical point and the direct calculation, we get that

$$\partial(g_1\circ\ell)(\hat{x}_i)=\begin{cases}[-\frac{1}{\alpha_1},\frac{1}{\alpha_1}], & \text{if } i\in I_1(\hat{\mathbf{x}}),\\ \{\frac{1}{\alpha_1}\mathrm{sgn}(\hat{x}_i)\}, & \text{otherwise},\end{cases}$$

$$\partial(h_1\circ\ell)(\hat{x}_i)=\begin{cases}0, & \text{if } i\in I_1(\hat{\mathbf{x}})\cup I_2(\hat{\mathbf{x}}),\\ con\{0,\frac{1}{\alpha_2}\mathrm{sgn}(\hat{x}_i)\}, & \text{if } i\in I_3(\hat{\mathbf{x}}),\\ \{\frac{1}{\alpha_2}\mathrm{sgn}(\hat{x}_i)\}, & \text{if } i\in I_4(\hat{\mathbf{x}}).\end{cases}$$

Similar to [42, Theorem 3.4], we can get

$$\partial(g_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})=\begin{cases}\frac{1}{\alpha_2}B^{m_j}, & \text{if } j\in J_1(\hat{\mathbf{y}}),\\ \{\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \text{otherwise},\end{cases}$$

$$\partial(h_2\circ\rho_j)(\hat{\mathbf{y}}_{(j)})=\begin{cases}\mathbf{0}, & \text{if } j\in J_1(\hat{\mathbf{y}})\cup J_2(\hat{\mathbf{y}}),\\ con\{\mathbf{0},\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \text{if } j\in J_3(\hat{\mathbf{y}}),\\ \{\frac{\hat{\mathbf{y}}_{(j)}}{\alpha_2\|\hat{\mathbf{y}}_{(j)}\|}\}, & \text{if } j\in J_4(\hat{\mathbf{y}}).\end{cases}$$

The proof is thus finished. □

Now we give the definition of d-stationary point of problem (1.2).

Definition 3.3 *[42] [d-stationary point] $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ is called a d-stationary point of problem (1.2), if*

$$F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) \geq 0, \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}.$$

The following theorem gives the characterization of d-stationary point of problem (1.2).

Theorem 3.4 *Let $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a d-stationary point of problem (1.2), then*

$$\langle \nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}}), (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle + \lambda_1 \left(\sum_{i=1}^{n} (g_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) \right) - \lambda_1 \left(\sum_{i=1}^{n} (h_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) \right)$$
$$+ \lambda_2 \left(\sum_{j=1}^{J} (g_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) \right) - \lambda_2 \left(\sum_{j=1}^{J} (h_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) \right) \geq 0$$

for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$, where

$$\begin{aligned}
(g_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) &= \begin{cases} \frac{|x_i|}{\alpha_1}, & \text{if } i \in I_1(\hat{\mathbf{x}}), \\ \frac{\mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i)}{\alpha_1}, & \text{otherwise}, \end{cases} \\
(h_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) &= \begin{cases} 0, & \text{if } i \in I_1(\hat{\mathbf{x}}) \cup I_2(\hat{\mathbf{x}}), \\ \max\{0, \frac{\mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i)}{\alpha_1}\}, & \text{if } i \in I_3(\hat{\mathbf{x}}), \\ \frac{\mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i)}{\alpha_1}, & \text{if } i \in I_4(\hat{\mathbf{x}}). \end{cases} \\
(g_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) &= \begin{cases} \frac{\|\mathbf{y}_{(j)}\|}{\alpha_2}, & \text{if } j \in J_1(\hat{\mathbf{y}}), \\ \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\alpha_2 \|\hat{\mathbf{y}}_{(j)}\|}, & \text{otherwise}. \end{cases} \\
(h_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) &= \begin{cases} \mathbf{0}, & \text{if } j \in J_1(\hat{\mathbf{y}}) \cup J_2(\hat{\mathbf{y}}), \\ \max\{\mathbf{0}, \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\alpha_2 \|\hat{\mathbf{y}}_{(j)}\|}\}, & \text{if } j \in J_3(\hat{\mathbf{y}}), \\ \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\alpha_2 \|\hat{\mathbf{y}}_{(j)}\|}, & \text{if } j \in J_4(\hat{\mathbf{y}}). \end{cases}
\end{aligned} \tag{3.1}$$

Proof From the definition of d-stationary point, the DC form (1.3) and the analysis similar to [42, Theorem 3.2], we can directly obtain the conclusion. □

The following theorem provides the relationship between d-stationary point and critical point.

Theorem 3.5 *Let $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ be a d-stationary point of problem (1.2), then $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a critical point of problem (1.2).*

Proof Sine $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a d-stationary point of problem (1.2), according to inequality (2.4), we have

$$\begin{aligned}
0 &\leq F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) \\
&\leq F^\circ((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) \\
&= \max_{\xi \in \partial^C F(\hat{\mathbf{x}}, \hat{\mathbf{y}})} \langle \xi, (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle .
\end{aligned}$$

Therefore, according to the operational properties of Clarke differential [13, Propostion 2.3.3, Corollary 2.3.3.2], we obtain that

$$\begin{aligned}
&\mathbf{0} \in \partial^C F(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \\
&\subseteq \partial^C f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) + \lambda_1 \partial^C \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) - h_1(|\hat{x}_i|) \right) + \lambda_2 \partial^C \left(\sum_{j=1}^{J} g_2(\|\hat{\mathbf{y}}_{(j)}\|) - h_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) \\
&\subseteq \partial^C f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) + \lambda_1 \partial^C \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) \right) - \lambda_1 \partial^C \left(\sum_{i=1}^{n} h_1(|\hat{x}_i|) \right) \\
&\quad + \lambda_2 \partial^C \left(\sum_{j=1}^{J} g_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) - \lambda_2 \partial^C \left(\sum_{j=1}^{J} h_2(\|\hat{\mathbf{y}}_{(j)}\|) \right).
\end{aligned}$$

Since f is continuously differentiable, g_ν and h_ν ($\nu = 1, 2$) are all convex functions, according to [13, Propostion 2.3.6(b)], we get

$$\begin{aligned}
\partial^C f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) &= \partial f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \{\nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\} \\
\partial^C \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) \right) &= \partial \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) \right) \\
\partial^C \left(\sum_{i=1}^{n} h_1(|\hat{x}_i|) \right) &= \partial \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) \right)
\end{aligned}$$

and

$$\begin{aligned}
\partial^C \left(\sum_{j=1}^{J} g_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) &= \partial \left(\sum_{j=1}^{J} g_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) \\
\partial^C \left(\sum_{j=1}^{J} h_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) &= \partial \left(\sum_{j=1}^{J} h_2(\|\hat{\mathbf{y}}_{(j)}\|) \right),
\end{aligned}$$

then

$$\begin{aligned}
\mathbf{0} \in \nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}}) &+ \lambda_1 \partial \left(\sum_{i=1}^{n} g_1(|\hat{x}_i|) \right) - \lambda_1 \partial \left(\sum_{i=1}^{n} h_1(|\hat{x}_i|) \right) \\
&+ \lambda_2 \partial \left(\sum_{j=1}^{J} g_2(\|\hat{\mathbf{y}}_{(j)}\|) \right) - \lambda_2 \partial \left(\sum_{j=1}^{J} h_2(\|\hat{\mathbf{y}}_{(j)}\|) \right).
\end{aligned}$$

From Definition 3.1, the above inequality implies that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a critical point of problem (1.2). □

Remark 3.6 *From the proof of Lemma 3.5, we have that if* $\mathbf{0} \in \partial^C F(\hat{\mathbf{x}}, \hat{\mathbf{y}})$*, then* $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ *is a critical point of problem (1.2).*

The following lemma characterize the property of gradient of f at the d-stationary point of problem (1.2).

Lemma 3.7 *Let $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ be a d-stationary point of problem (1.2), the following statements hold:*

(i) $|[\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i| = \frac{\lambda_1}{\alpha_1},\ \forall i \in I_2(\hat{\mathbf{x}})$; $\quad [\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i = 0,\ \forall i \in I_4(\hat{\mathbf{x}})$.

(ii) $\|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)}\| = \frac{\lambda_2}{\alpha_2},\ \forall j \in J_2(\hat{\mathbf{y}})$; $\quad \|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)}\| = 0,\ \forall i \in J_4(\hat{\mathbf{y}})$.

Proof (i). From Theorem 3.4, for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$, we have

$$\begin{aligned}
&0 \le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}})) \\
&= \langle \nabla f(\hat{\mathbf{x}}, \hat{\mathbf{y}}), (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}}) \rangle + \lambda_1 \sum_{i=1}^{n} (g_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) - \lambda_1 \sum_{i=1}^{n} (h_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) \\
&\quad + \lambda_2 \sum_{j=1}^{J} (g_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) - \lambda_2 \sum_{j=1}^{J} (h_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}).
\end{aligned} \tag{3.2}$$

According to the arbitrariness of $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$, let $\mathbf{y} = \hat{\mathbf{y}}$, then

$$\begin{aligned}
&0 \le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{0})) \\
&= \sum_{i=1}^{n} [\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i (x_i - \hat{x}_i) + \lambda_1 \sum_{i=1}^{n} (g_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i) - \lambda_1 \sum_{i=1}^{n} (h_1 \circ \ell)'(\hat{x}_i; x_i - \hat{x}_i).
\end{aligned}$$

From (3.1), we have

$$\begin{aligned}
0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{x} - \hat{\mathbf{x}}, \mathbf{0})) \\
&= \sum_{i=1}^{n} [\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i (x_i - \hat{x}_i) + \frac{\lambda_1}{\alpha_1} \Bigg(\sum_{i \in I_1(\hat{\mathbf{x}})} |x_i| + \sum_{i \in [n] \backslash I_1(\hat{\mathbf{x}})} \mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i) \\
&\quad - \sum_{i \in I_3(\hat{\mathbf{x}})} \max\{0, \mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i)\} - \sum_{i \in I_4(\hat{\mathbf{x}})} \mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i) \Bigg) \\
&= \sum_{i=1}^{n} [\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i (x_i - \hat{x}_i) + \frac{\lambda_1}{\alpha_1} \Bigg(\sum_{i \in I_1(\hat{\mathbf{x}})} |x_i| + \sum_{i \in I_2(\hat{\mathbf{x}}) \cup I_3(\hat{\mathbf{x}})} \mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i) \\
&\quad - \sum_{i \in I_3(\hat{\mathbf{x}})} \max\{0, \mathrm{sgn}(\hat{x}_i)(x_i - \hat{x}_i)\} \Bigg).
\end{aligned} \tag{3.3}$$

Let

$$\tilde{x}_i^1 = \begin{cases} \hat{x}_i, & \text{if } i \in [n] \backslash I_2(\hat{\mathbf{x}}), \\ \hat{x}_i - \big([\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i + \frac{\lambda_1}{\alpha_1} \mathrm{sgn}(\hat{x}_i) \big), & \text{if } i \in I_2(\hat{\mathbf{x}}), \end{cases}$$

then from (3.3), we obtain that

$$\begin{aligned}
0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\tilde{\mathbf{x}}^1 - \hat{\mathbf{x}}, \mathbf{0})) \\
&= \sum_{i \in I_2(\hat{\mathbf{x}})} \big([\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i + \frac{\lambda_1}{\alpha_1} \mathrm{sgn}(\hat{x}_i) \big) (\tilde{x}_i^1 - \hat{x}_i), \\
&= - \sum_{i \in I_2(\hat{\mathbf{x}})} \Big([\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i + \frac{\lambda_1}{\alpha_1} \mathrm{sgn}(\hat{x}_i) \Big)^2.
\end{aligned} \tag{3.4}$$

From inequality (3.4), we obtain

$$[\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i + \frac{\lambda_1}{\alpha_1} \mathrm{sgn}(\hat{x}_i) = 0, \quad \forall i \in I_2(\hat{\mathbf{x}}).$$

Thus, $|[\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i| = \frac{\lambda_1}{\alpha_1},\ \forall i \in I_2(\hat{\mathbf{x}})$.

Let

$$\tilde{x}_i^2 = \begin{cases} \hat{x}_i, & \text{if } i \in [n] \backslash I_4(\hat{\mathbf{x}}), \\ \hat{x}_i - [\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i, & \text{if } i \in I_4(\hat{\mathbf{x}}), \end{cases}$$

then from (3.3), we obtain that

$$\begin{aligned} 0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\tilde{\mathbf{x}}^2 - \hat{\mathbf{x}}, \mathbf{0})) \\ &= \sum_{i \in I_4(\hat{\mathbf{x}})} \big([\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i\big)(\tilde{x}_i^2 - \hat{x}_i) = -\sum_{i \in I_4(\hat{\mathbf{x}})} \big([\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i\big)^2, \end{aligned} \tag{3.5}$$

From inequality (3.5), we obtain

$$[\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_i = 0, \quad \forall i \in I_4(\hat{\mathbf{x}}).$$

(ii). The proof is similar to that of (i). By the arbitrariness of $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$ in (3.2), take $\mathbf{x} = \hat{\mathbf{x}}$, then

$$\begin{aligned} 0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})) \\ &= \sum_{j=1}^{J} [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)} (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) + \lambda_2 \sum_{j=1}^{J} (g_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) \\ &\quad - \lambda_2 \sum_{j=1}^{J} (h_2 \circ \rho_j)'(\hat{\mathbf{y}}_{(j)}; \mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}). \end{aligned}$$

From (3.1), we have

$$\begin{aligned} 0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{0}, \mathbf{y} - \hat{\mathbf{y}})) \\ &= \sum_{j=1}^{J} [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) + \frac{\lambda_2}{\alpha_2} \Bigg(\sum_{j \in J_1(\hat{\mathbf{x}})} \|\mathbf{y}_{(j)}\| + \sum_{j \in [m] \backslash J_1(\hat{\mathbf{y}})} \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|} \\ &\quad - \sum_{j \in J_3(\hat{\mathbf{y}})} \max\left\{\mathbf{0}, \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|}\right\} - \sum_{j \in J_4(\hat{\mathbf{y}})} \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|} \Bigg) \qquad (3.6) \\ &= \sum_{j=1}^{J} [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)}) + \frac{\lambda_2}{\alpha_2} \Bigg(\sum_{j \in J_1(\hat{\mathbf{x}})} \|\mathbf{y}_{(j)}\| + \sum_{j \in J_2(\hat{\mathbf{y}}) \cup J_3(\hat{\mathbf{y}})} \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|} \\ &\quad - \sum_{j \in J_3(\hat{\mathbf{y}})} \max\left\{\mathbf{0}, \frac{\hat{\mathbf{y}}_{(j)}^\top (\mathbf{y}_{(j)} - \hat{\mathbf{y}}_{(j)})}{\|\hat{\mathbf{y}}_{(j)}\|}\right\} \Bigg). \end{aligned}$$

Let

$$\tilde{\mathbf{y}}_{(j)}^1 = \begin{cases} \hat{\mathbf{y}}_{(j)}, & \text{if } j \in [m] \backslash J_2(\hat{\mathbf{y}}), \\ \hat{\mathbf{y}}_{(j)} - \left([\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)} + \frac{\lambda_2}{\alpha_2} \frac{\hat{\mathbf{y}}_{(j)}}{\|\hat{\mathbf{y}}_{(j)}\|}\right), & \text{if } j \in J_2(\hat{\mathbf{y}}), \end{cases}$$

then from (3.6), we obtain that

$$\begin{aligned} 0 &\le F'((\hat{\mathbf{x}}, \hat{\mathbf{y}}); (\mathbf{0}, \tilde{\mathbf{y}}^{\mathbf{1}} - \hat{\mathbf{y}})) \\ &= \sum_{j \in J_2(\hat{\mathbf{y}})} \left([\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)} + \frac{\lambda_2}{\alpha_2} \frac{\hat{\mathbf{y}}_{(j)}}{\|\hat{\mathbf{y}}_{(j)}\|}\right)^\top (\tilde{\mathbf{y}}_{(j)}^1 - \hat{\mathbf{y}}_{(j)}) \\ &= -\sum_{j \in J_2(\hat{\mathbf{y}})} \left\| [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})]_{(j)} + \frac{\lambda_2}{\alpha_2} \frac{\hat{\mathbf{y}}_{(j)}}{\|\hat{\mathbf{y}}_{(j)}\|} \right\|^2. \end{aligned} \tag{3.7}$$

From inequality (3.7), we obtain

$$[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)} + \frac{\lambda_2}{\alpha_2}\frac{\hat{\mathbf{y}}_{(j)}}{\|\hat{\mathbf{y}}_{(j)}\|} = 0, \text{ i.e., } [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)} = -\frac{\lambda_2}{\alpha_2}\frac{\hat{\mathbf{y}}_{(j)}}{\|\hat{\mathbf{y}}_{(j)}\|}, \ \forall j \in J_2(\hat{\mathbf{y}}).$$

Take ℓ_2 norm on both sides of the above equality, then we get

$$\|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)}\| = \frac{\lambda_2}{\alpha_2}, \ \forall j \in J_2(\hat{\mathbf{y}}).$$

Let

$$\tilde{\mathbf{y}}^2_{(j)} = \begin{cases} \hat{\mathbf{y}}_{(j)}, & \text{if } j \in [m]\backslash J_4(\hat{\mathbf{y}}), \\ \hat{\mathbf{y}}_{(j)} - [\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)}, & \text{if } j \in J_4(\hat{\mathbf{y}}), \end{cases}$$

then from (3.6), we obtain that

$$\begin{aligned} 0 &\le F'((\hat{\mathbf{x}},\hat{\mathbf{y}});(\mathbf{0},\tilde{\mathbf{y}}^{\mathbf{2}}-\hat{\mathbf{y}})) \\ &= \sum_{j\in J_4(\hat{\mathbf{y}})} \left([\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)}\right)^\top (\tilde{\mathbf{y}}^2_{(j)} - \hat{\mathbf{y}}_{(j)}) \\ &= -\sum_{j\in J_4(\hat{\mathbf{y}})} \left\|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)}\right\|^2. \end{aligned} \tag{3.8}$$

From inequality (3.8), we obtain

$$\|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j)}\| = 0, \ \forall j \in J_4(\hat{\mathbf{y}}).$$

The proof is thus complete. □

The following theorem gives the lower bound property of the d-stationary points of problem (1.2).

Theorem 3.8 *Let* $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ *be a d-stationary point of problem (1.2). Suppose* $\|[\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{I_2(\hat{\mathbf{x}})\cup J_2(\hat{\mathbf{y}})}\| < \min\left\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\right\}$*, then the following statements hold:*

(i) $I_2(\hat{\mathbf{x}}) = \emptyset$*, that is, if* $\hat{x}_i \neq 0$*, then* $|\hat{x}_i| \ge \alpha_1$*;*

(ii) $J_2(\hat{\mathbf{y}}) = \emptyset$*, that is, if* $\hat{\mathbf{y}}_{(j)} \neq \mathbf{0}$*, then* $\|\hat{\mathbf{y}}_{(j)}\| \ge \alpha_2$*.*

Proof (i) Assume, on the contrary, that $I_2(\hat{\mathbf{x}}) \neq \emptyset$. Let $i_0 \in I_2(\hat{\mathbf{x}})$, then from Lemma 3.7, we have

$$\frac{\lambda_1}{\alpha_1} = |[\nabla_{\mathbf{x}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{i_0}| \le \|[\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{I_2(\hat{\mathbf{x}})\cup J_2(\hat{\mathbf{y}})}\| < \frac{\lambda_1}{\alpha_1},$$

which is a contradiction, and implies that $I_2(\hat{\mathbf{x}}) = \emptyset$.

(ii) Assume, on the contrary, that $J_2(\hat{\mathbf{y}}) \neq \emptyset$. Let $j_0 \in J_2(\hat{\mathbf{y}})$, then from Lemma 3.7, we have

$$\frac{\lambda_2}{\alpha_2} = \|[\nabla_{\mathbf{y}} f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{(j_0)}\| \le \|[\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{I_2(\hat{\mathbf{x}})\cup J_2(\hat{\mathbf{y}})}\| < \frac{\lambda_2}{\alpha_2},$$

which is a contradiction, and implies that $J_2(\hat{\mathbf{y}}) = \emptyset$. □

Remark 3.9 *(1) If* $\|\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})\| < \min\left\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\right\}$*, then* $\|[\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})]_{I_2(\hat{\mathbf{x}})\cup J_2(\hat{\mathbf{y}})}\| < \min\left\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\right\}$*.*

(2) If f *is locally Lipschitz at* $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ *with modulus* $L < \min\left\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\right\}$*, then* $\|\nabla f(\hat{\mathbf{x}},\hat{\mathbf{y}})\| < \min\left\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\right\}$*.*

(3) If $f : \mathbb{R}^{n+m} \to \mathbb{R}$ *is convex, then* f *is locally Lipschitz on* $\mathbb{R}^{n+m}$*.*

4 Equivalence of problem (1.1) and problem (1.2)

In this section, we investigate the relationship between the original problem (1.1) and the relaxation problem (1.2) by considering the global solutions and local solutions of them.

Theorem 4.1 *Suppose* $\|\nabla f(\mathbf{x},\mathbf{y})\| < \min\{\frac{\lambda_1}{\alpha_1}, \frac{\lambda_2}{\alpha_2}\}$ *holds on* $\mathbb{R}^{n+m}$*, then the following statements hold.*

(i) *The global optimal solution sets and optimal value of problem (1.1) are same as those of problem (1.2) respectively;*

(ii) *If* $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ *is a local minimizer of problem (1.2), then* $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ *is also a local minimizer of problem (1.1), and the objective function value of problems (1.1) and (1.2) at* $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ *are same.*

Proof (i). (a) Let $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a global optimal solution of problem (1.2), then $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is also a d-stationary point of problem (1.2). From (1.2) and Theorem 3.8, we obtain

$$\varphi_1(|\hat{x}_i|) = \begin{cases} 0, & \text{if } i \in I_1(\hat{\mathbf{x}}) \cup I_2(\hat{\mathbf{x}}), \\ 1, & \text{if } i \in I_3(\hat{\mathbf{x}}) \cup I_4(\hat{\mathbf{x}}), \end{cases} \quad \text{and} \quad \varphi_2(\|\hat{\mathbf{y}}_{(j)}\|) = \begin{cases} 0, & if\ j \in J_1(\hat{\mathbf{y}}) \cup J_2(\hat{\mathbf{y}}), \\ 1, & \text{if } j \in J_3(\hat{\mathbf{y}}) \cup J_4(\hat{\mathbf{y}}), \end{cases}$$

then

$$\Phi_1(\hat{\mathbf{x}}) = \sum_{i \in I_3(\hat{\mathbf{x}}) \cup I_4(\hat{\mathbf{x}})} \varphi_1(|\hat{x}_i|) = \|\hat{\mathbf{x}}\|_0, \quad \Phi_2(\hat{\mathbf{y}}) = \sum_{j \in J_3(\hat{\mathbf{y}}) \cup J_4(\hat{\mathbf{y}})} \varphi_2(\|\hat{\mathbf{y}}_{(j)}\|) = \|\hat{\mathbf{y}}\|_{2,0}. \quad (4.1)$$

For any $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{n+m}$, since $\varphi_\upsilon(t) = \min\left\{1, \frac{t}{\alpha_\upsilon}\right\} \le 1$, $(\upsilon = 1, 2)$, then

$$\begin{aligned} \Phi_1(\mathbf{x}) &= \sum_{i \in [n] \setminus I_1(\mathbf{x})} \varphi_1(|x_i|) \le \sum_{i \in [n] \setminus I_1(\mathbf{x})} 1 = \|\mathbf{x}\|_0, \\ \Phi_2(\mathbf{y}) &= \sum_{j \in [m] \setminus J_1(\hat{\mathbf{y}})} \varphi_2(\|\hat{\mathbf{y}}_{(j)}\|) \le \sum_{j \in [m] \setminus J_1(\hat{\mathbf{y}})} 1 = \|\mathbf{y}\|_{2,0}. \end{aligned}$$

Thus, we have

$$\begin{aligned} f(\hat{\mathbf{x}},\hat{\mathbf{y}}) + \lambda_1\|\hat{\mathbf{x}}\|_0 + \lambda_2\|\hat{\mathbf{y}}\|_{2,0} &= f(\hat{\mathbf{x}},\hat{\mathbf{y}}) + \lambda_1\Phi_1(\hat{\mathbf{x}}) + \lambda_2\Phi_2(\hat{\mathbf{y}}) \\ &\le f(\mathbf{x},\mathbf{y}) + \lambda_1\Phi_1(\mathbf{x}) + \lambda_2\Phi_2(\mathbf{y}) \\ &\le f(\mathbf{x},\mathbf{y}) + \lambda_1\|\mathbf{x}\|_0 + \lambda_2\|\mathbf{y}\|_{2,0}. \end{aligned}$$

Therefore, $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is a global solution of problem (1.1), and (4.1) implies that optimal value of problems (1.1) and (1.2) at $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ are same.

(b) On the other hand, let $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a global minimizer of problem (1.1). Assume, on the contrary, that $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is not a global minimizer of problem (1.2), then

$$\Phi_1(\hat{\mathbf{x}}) \le \|\hat{\mathbf{x}}\|_0 \quad \text{and} \quad \Phi_2(\hat{\mathbf{y}}) \le \|\hat{\mathbf{y}}\|_{2,0}.$$

Let $(\bar{\mathbf{x}},\bar{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a global minimizer of problem (1.2), then

$$f(\bar{\mathbf{x}},\bar{\mathbf{y}}) + \lambda_1\Phi_1(\bar{\mathbf{x}}) + \lambda_2\Phi_2(\bar{\mathbf{y}}) < f(\hat{\mathbf{x}},\hat{\mathbf{y}}) + \lambda_1\Phi_1(\hat{\mathbf{x}}) + \lambda_2\Phi_2(\hat{\mathbf{y}}).$$

What's more, from (i)(a), we know that

$$\Phi_1(\bar{\mathbf{x}}) = \|\bar{\mathbf{x}}\|_0 \quad \text{and} \quad \Phi_2(\bar{\mathbf{y}}) = \|\bar{\mathbf{y}}\|_{2,0}.$$

Thus, we have

$$\begin{aligned} f(\bar{\mathbf{x}},\bar{\mathbf{y}})+\lambda_1\|\bar{\mathbf{x}}\|_0+\lambda_2\|\bar{\mathbf{y}}\|_{2,0} &= f(\bar{\mathbf{x}},\bar{\mathbf{y}})+\lambda_1\Phi_1(\bar{\mathbf{x}})+\lambda_2\Phi_2(\bar{\mathbf{y}}) \\ &< f(\hat{\mathbf{x}},\hat{\mathbf{y}})+\lambda_1\Phi_1(\hat{\mathbf{x}})+\lambda_2\Phi_2(\hat{\mathbf{y}}) \\ &\le f(\hat{\mathbf{x}},\hat{\mathbf{y}})+\lambda_1\|\hat{\mathbf{x}}\|_0+\lambda_2\|\hat{\mathbf{y}}\|_{2,0}, \end{aligned}$$

which contradicts that $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is a global minimizer of problem (1.1). Therefore, $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ must be a global minimizer of problem (1.2).

(ii). Let $(\hat{\mathbf{x}},\hat{\mathbf{y}}) \in \mathbb{R}^{n+m}$ be a local minimizer of problem (1.2), then there exists a neighborhood W of $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that

$$f(\hat{\mathbf{x}},\hat{\mathbf{y}})+\lambda_1\Phi_1(\hat{\mathbf{x}})+\lambda_2\Phi_2(\hat{\mathbf{y}}) \le f(\mathbf{x},\mathbf{y})+\lambda_1\Phi_1(\mathbf{x})+\lambda_2\Phi_2(\mathbf{y}), \quad \forall(\mathbf{x},\mathbf{y})\in W,$$

It is easy to know that $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is also a d-stationary point of problem (1.2). From Theorem 3.8 and (4.1), we have

$$\Phi_1(\hat{\mathbf{x}}) = \|\hat{\mathbf{x}}\|_0 \quad \text{and} \quad \Phi_2(\hat{\mathbf{y}}) = \|\hat{\mathbf{y}}\|_{2,0}. \tag{4.2}$$

Hence, we have

$$\begin{aligned} f(\hat{\mathbf{x}},\hat{\mathbf{y}})+\lambda_1\|\hat{\mathbf{x}}\|_0+\lambda_2\|\hat{\mathbf{y}}\|_{2,0} &= f(\hat{\mathbf{x}},\hat{\mathbf{y}})+\lambda_1\Phi_1(\hat{\mathbf{x}})+\lambda_2\Phi_2(\hat{\mathbf{y}}) \\ &\le f(\mathbf{x},\mathbf{y})+\lambda_1\Phi_1(\mathbf{x})+\lambda_2\Phi_2(\mathbf{y}) \\ &\le f(\mathbf{x},\mathbf{y})+\lambda_1\|\mathbf{x}\|_0+\lambda_2\|\mathbf{y}\|_{2,0}, \quad \forall(\mathbf{x},\mathbf{y})\in W. \end{aligned}$$

Therefore, $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ is a local minimizer of problem (1.1), and (4.2) implies that the objective function value of problems (1.1) and (1.2) at $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ are equal. □

Remark 4.2 *(1) The result in Theorem (4.1) reveals that problems (1.1) and (1.2) have some equivalence, which provides a theoretical basis for solving problem (1.1) via solving problem (1.2).*

(2) From Remark 3.9, we know that the hypothesis of Theorem (4.1) is easy to satisfy.

5 Alternating proximal gradient algorithm for problem (1.2)

In this section, we propose an APG algorithm to solve problem (1.2), and discuss the convergence of the sequence generated by the APG algorithm.

5.1 Scheme of APG algorithm

Noting that the objective function F in (1.2) has two parts of variables, the alternating minimization may be the suitable way to solve problem (1.2), which transforms problem (1.2) into two subproblems.

Take the initial point $(\mathbf{x}^0,\mathbf{y}^0) \in \mathbb{R}^{n+m}$, and let the sequence $\{(\mathbf{x}^{k+1},\mathbf{y}^{k+1})\}_{k\in\mathbb{N}}$ be generated through the following subproblems:

$$\begin{cases} \mathbf{x}^{k+1} \in \arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}^k,\mathbf{y}^k)+\langle \mathbf{x}-\mathbf{x}^k, \nabla_{\mathbf{x}} f(\mathbf{x}^k,\mathbf{y}^k)\rangle + \dfrac{1}{2t_1}\|\mathbf{x}-\mathbf{x}^k\|^2+\lambda_1\Phi_1(\mathbf{x}), & \text{(5.1a)} \\ \mathbf{y}^{k+1} \in \arg\min\limits_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x}^{k+1},\mathbf{y}^k)+\langle \mathbf{y}-\mathbf{y}^k, \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1},\mathbf{y}^k)\rangle + \dfrac{1}{2t_2}\|\mathbf{y}-\mathbf{y}^k\|^2+\lambda_2\Phi_2(\mathbf{y}). & \text{(5.1b)} \end{cases}$$

One can note that the two subproblems in (5.1) are both nonconvex and nonsmooth since $\Phi_1(\mathbf{x})$ and $\Phi_2(\mathbf{y})$ are both nonconvex and nonsmooth. Fortunately, in the following part, we can provide their closed form solutions, which is very important for the efficiency of the APG algorithm.

The subproblem (5.1a) solves $\mathbf{x}$ with the fixed $\mathbf{y}^k$. It can be explicitly reexpressed as the following form:

$$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2t_1}\|\mathbf{x} - (\mathbf{x}^k - t_1\nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k))\|^2 + \lambda_1 \Phi_1(\mathbf{x}) \right\}. \tag{5.2}$$

Denote $\mathbf{v}^k := \mathbf{x}^k - t_1\nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k)$, then (5.2) can be rewritten as

$$\mathbf{x}^{k+1} \in \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2t_1}\|\mathbf{x} - \mathbf{v}^k\|^2 + \lambda_1 \Phi_1(\mathbf{x}) \right\}. \tag{5.3}$$

Note that $\Phi_1(\mathbf{x})$ is separable in the component of $\mathbf{x}$, then problem (5.3) is also separable. That is,

$$\mathbf{x}^{k+1} \in \mathrm{Prox}_{t_1\lambda_1\Phi_1}(\mathbf{v}^k) = \mathrm{Prox}_{t_1\lambda_1\varphi_1}(v_1^k) \times \cdots \times \mathrm{Prox}_{t_1\lambda_1\varphi_1}(v_n^k), \tag{5.4}$$

where the proximal operator $\mathrm{Prox}_{t_1\lambda_1\varphi_1}(\cdot)$ is the optimal solution of the following problem

$$\mathrm{Prox}_{t_1\lambda_1\varphi_1}(v) = \arg\min_{x\in\mathbb{R}} \left\{ \frac{1}{2t_1}(x-v)^2 + \lambda_1\varphi_1(x) \right\},\ \forall v \in \mathbb{R}. \tag{5.5}$$

The solution of (5.5) is known to have the following closed form [3,19,30,43]

$$\begin{aligned}\mathrm{Prox}_{t_1\lambda_1\varphi_1}(v) &= \begin{cases} 0, & |v| \le \frac{\lambda_1 t_1}{\alpha_1}, \\ \mathrm{sgn}(v)(|v| - \frac{\lambda_1 t_1}{\alpha_1}), & \frac{\lambda_1 t_1}{\alpha_1} < |v| < \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}, \\ \mathrm{sgn}(v)(\alpha_1 \pm \frac{\lambda_1 t_1}{2\alpha_1}), & |v| = \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}, \\ v, & |v| > \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}. \end{cases} \\ &= \begin{cases} \mathrm{sgn}(v)(|v| - \frac{\lambda_1 t_1}{\alpha_1})_+, & |v| \le \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}, \\ v, & |v| \ge \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}. \end{cases}\end{aligned} \tag{5.6}$$

which means that $\mathrm{Prox}_{t_1\lambda_1\varphi_1}(v)$ has two values when $|v| = \alpha_1 + \frac{\lambda_1 t_1}{2\alpha_1}$.

The subproblem (5.1b) solves $\mathbf{y}$ with the fixed $\mathbf{x}^{k+1}$. It can be explicitly reexpressed as

$$\mathbf{y}^{k+1} \in \arg\min_{\mathbf{y}\in\mathbb{R}^m} \left\{ \frac{1}{2t_2}\|\mathbf{y} - (\mathbf{y}^k - t_2\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k))\|^2 + \lambda_2 \Phi_2(\mathbf{y}) \right\}. \tag{5.7}$$

Denote $\mathbf{u}^k := \mathbf{y}^k - t_2\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k)$, then (5.7) can be simplified as

$$\mathbf{y}^{k+1} \in \arg\min_{\mathbf{y}\in\mathbb{R}^m} \left\{ \frac{1}{2t_2}\|\mathbf{y} - \mathbf{u}^k\|^2 + \lambda_2 \Phi_2(\mathbf{y}) \right\}. \tag{5.8}$$

Note that $\Phi_2(\mathbf{y})$ is separable in the group of $\mathbf{y}$, then problem (5.8) is also group separable. That is, the solution of (5.8) have the following closed form

$$\mathbf{y}^{k+1} \in \mathrm{Prox}_{t_2\lambda_2\Phi_2}(\mathbf{u}^k) = [\mathrm{Prox}_{t_2\lambda_2\Phi_2}(\mathbf{u}^k)]_{(1)} \times \cdots \times [\mathrm{Prox}_{t_2\lambda_2\Phi_2}(\mathbf{u}^k)]_{(J)} \tag{5.9}$$

with

$$[\mathrm{Prox}_{t_2\lambda_2\Phi_2}(\mathbf{u})]_{(j)} = \begin{cases} (\|\mathbf{u}_{(j)}\| - \frac{\lambda_2 t_2}{\alpha_2})_+ \frac{\mathbf{u}_{(j)}}{\|\mathbf{u}_{(j)}\|}, & \|\mathbf{u}_{(j)}\| \le \alpha_2 + \frac{\lambda_2 t_2}{2\alpha_2}, \\ \mathbf{u}_{(j)}, & \|\mathbf{u}_{(j)}\| \ge \alpha_2 + \frac{\lambda_2 t_2}{2\alpha_2}, \end{cases}$$

for $j = 1, \cdots, J$, which can be obtained by the similar way to (5.6) or [42].

From (5.4) and (5.9), we give the scheme of the APG algorithm for solving problem (1.2) as below.

Algorithm 1 APG algorithm

- **Initialize:** For given $\alpha_1 > 0, \alpha_2 > 0$, $\lambda_1 > 0, \lambda_2 > 0$, $t_1 > 0, t_2 > 0$, $xtol > 0$, take $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^{n+m}$, and set $k = 0$.
- **Step1.** Compute

$$\begin{cases} \mathbf{x}^{k+1} \in \arg\min\limits_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}^k, \mathbf{y}^k) + \langle \mathbf{x} - \mathbf{x}^k, \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k)\rangle + \frac{1}{2t_1}\|\mathbf{x} - \mathbf{x}^k\|^2 + \lambda_1 \Phi_1(\mathbf{x}), \\ \mathbf{y}^{k+1} \in \arg\min\limits_{\mathbf{y}\in\mathbb{R}^m} f(\mathbf{x}^{k+1}, \mathbf{y}^k) + \langle \mathbf{y} - \mathbf{y}^k, \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k)\rangle + \frac{1}{2t_2}\|\mathbf{y} - \mathbf{y}^k\|^2 + \lambda_2 \Phi_2(\mathbf{y}). \end{cases}$$

  The calculation process is as follows:
  I. Compute $\mathbf{x}^{k+1} \in \mathrm{Prox}_{t_1\lambda_1\Phi_1}(\mathbf{x}^k - t_1\nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k))$ according to (5.4);
  II. Let $\mathbf{u}^k = \mathbf{y}^k - t_2\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k)$, then divide $\mathbf{u}^k$ into J groups according to the given group of $\mathbf{y}$;
  III. Compute $\mathbf{y}^{k+1} \in \mathrm{Prox}_{t_2\lambda_2\Phi_2}(\mathbf{u}^k)$ according to (5.9).
- **Step2.** Let $\mathbf{z}^{k+1} := (\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$, if $\frac{\|\mathbf{z}^{k+1} - \mathbf{z}^k\|}{\max\{1, \|\mathbf{z}^{k+1}\|\}} \le$ xtol, terminate.
  Otherwise, let $k := k + 1$ then return to **Step1**.
- **Output:** $(\mathbf{x}^k, \mathbf{y}^k)$

5.2 Convergence analysis

Before the convergence analysis of the APG algorithm, we give some basic assumptions.

Assumption 5.1 (i) $\inf\{F(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y}) + \lambda_1\Phi_1(\mathbf{x}) + \lambda_2\Phi_2(\mathbf{y})\} > -\infty$.

(ii) $f(\mathbf{x}, \mathbf{y}) \to \infty$ as $\|(\mathbf{x}, \mathbf{y})\| \to \infty$.

(iii) $\nabla_{\mathbf{x}} f(\cdot, \cdot)$ is Lipschitz continuous with modulus L_1, that is

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}^1, \mathbf{y}^1) - \nabla_{\mathbf{x}} f(\mathbf{x}^2, \mathbf{y}^2)\| \le L_1(\|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{y}_1 - \mathbf{y}_2\|), \quad \forall (\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2) \in \mathbb{R}^{n+m}.$$

Meanwhile, for any $\mathbf{x}$, $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ is Lipschitz continuous about $\mathbf{y}$ with modulus L_2.

(v) The parameters satisfy

$$0 < t_1 < \frac{1}{L_1}, \quad 0 < t_2 < \frac{1}{L_2}.$$

It is easy to check that there are many loss functions satisfy Assumption 5.1, for example, ℓ_2 loss and logistic loss.

Next, we investigate the convergence of the proposed APG algorithm under Assumption 5.1.

Lemma 5.2 *Let $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k\in\mathbb{N}}$ be the sequence generated by the APG algorithm. Suppose Assumption 5.1 holds, then*

$$\rho\left(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2\right) \le F(\mathbf{x}^k, \mathbf{y}^k) - F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}), \tag{5.10}$$

where $\rho = \min\left\{\frac{1}{2t_1} - \frac{L_1}{2}, \frac{1}{2t_2} - \frac{L_2}{2}\right\} > 0$, which implies that the sequence $\{F(\mathbf{x}^k, \mathbf{y}^k)\}$ is nonincreasing.

Proof From Step 1 in the APG algorithm, we know that

$$\begin{aligned}\lambda_1 \Phi_1(\mathbf{x}^k) + f(\mathbf{x}^k, \mathbf{y}^k) \geq & f(\mathbf{x}^k, \mathbf{y}^k) + \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle \\ & + \frac{1}{2t_1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \lambda_1 \Phi_1(\mathbf{x}^{k+1}),\end{aligned} \tag{5.11}$$

and that

$$\begin{aligned}\lambda_2 \Phi_2(\mathbf{y}^k) + f(\mathbf{x}^{k+1}, \mathbf{y}^k) \geq & f(\mathbf{x}^{k+1}, \mathbf{y}^k) + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle \\ & + \frac{1}{2t_2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + \lambda_2 \Phi_2(\mathbf{y}^{k+1}).\end{aligned} \tag{5.12}$$

Summing (5.11) and (5.12), we obtain that

$$\begin{aligned}\lambda_1 \Phi_1(\mathbf{x}^k) + \lambda_2 \Phi_2(\mathbf{y}^k) \geq \; & \lambda_1 \Phi_1(\mathbf{x}^{k+1}) + \lambda_2 \Phi_2(\mathbf{y}^{k+1}) \\ & + \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle + \frac{1}{2t_1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ & + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle + \frac{1}{2t_2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2.\end{aligned} \tag{5.13}$$

From the Lipschitz continuity of $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ (Assumption 5.1 (iii)), we can obtain

$$\begin{aligned} f(\mathbf{x}^k, \mathbf{y}^k) &\geq f(\mathbf{x}^{k+1}, \mathbf{y}^k) - \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle - \frac{L_1}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \\ f(\mathbf{x}^{k+1}, \mathbf{y}^k) &\geq f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle - \frac{L_2}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2.\end{aligned}$$

The above two inequalities yield that

$$\begin{aligned} f(\mathbf{x}^k, \mathbf{y}^k) \geq \; & f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle - \frac{L_1}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \\ & - \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{y}^{k+1} - \mathbf{y}^k \right\rangle - \frac{L_2}{2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2.\end{aligned} \tag{5.14}$$

Summing (5.13) and (5.14), we have

$$F(\mathbf{x}^k, \mathbf{y}^k) \geq F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \left(\frac{1}{2t_1} - \frac{L_1}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \left(\frac{1}{2t_2} - \frac{L_2}{2}\right)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2.$$

By Assumption 5.1 (iii), we get $\frac{1}{2t_1} - \frac{L_1}{2} > 0$, $\frac{1}{2t_2} - \frac{L_2}{2} > 0$. Let $\rho = \min\left\{\frac{1}{2t_1} - \frac{L_1}{2}, \frac{1}{2t_2} - \frac{L_2}{2}\right\}$, then we obtain

$$\rho\left(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2\right) \leq F(\mathbf{x}^k, \mathbf{y}^k) - F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}).$$

This completes the proof. □

Theorem 5.3 *Suppose Assumption 5.1 holds. Let* $\{\mathbf{z}^k := (\mathbf{x}^k, \mathbf{y}^k)\}$ *be generated by the APG Algorithm, then the following statements hold.*

(i) $\{\mathbf{z}^k\}$ *is bounded and* $\{F(\mathbf{z}^k)\}$ *is convergent;*

(ii) $\sum_{k=0}^{\infty}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 < \infty$, $\lim\limits_{k\to\infty}\|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0$ *and* $\lim\limits_{k\to\infty}\|\mathbf{y}^{k+1} - \mathbf{y}^k\| = 0$.

Proof (i). From Lemma 5.2 and Assumption 5.1 (i), it follows that $\{F(\mathbf{x}^k, \mathbf{y}^k)\}$ is nonincreasing and F is bounded from below, and hence $\{F(\mathbf{x}^k, \mathbf{y}^k)\}$ is convergent. From $\{(\mathbf{x}^k, \mathbf{y}^k)\} \subset \{(\mathbf{x}, \mathbf{y}) : F(\mathbf{x}, \mathbf{y}) \leq F(\mathbf{x}^0, \mathbf{y}^0)\}$ which is bounded due to Assumption 5.1 (ii) and $\Phi_1(\mathbf{x}) \geq 0$ as well as $\Phi_2(\mathbf{y}) \geq 0$, it follows that $\{\mathbf{x}^k, \mathbf{y}^k\}$ is bounded.

(ii). From (5.10) and (i), we have

$$\rho\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 = \rho(\|(\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2) \leq F(\mathbf{x}^k, \mathbf{y}^k) - F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}).$$

Summing both sides of the above inequality from 0 to N, we get

$$\begin{aligned}\sum_{k=0}^{N} \rho\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 &= \sum_{k=0}^{N} \rho(\|(\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2) \\ &\leq \sum_{k=0}^{N} (F(\mathbf{x}^k, \mathbf{y}^k) - F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})) \\ &= F(\mathbf{x}^0, \mathbf{y}^0) - F(\mathbf{x}^{N+1}, \mathbf{y}^{N+1}),\end{aligned}$$

Letting $N \to \infty$, we obtain

$$\sum_{k=0}^{\infty} \rho\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 = \sum_{k=0}^{\infty} \rho(\|(\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2) \leq F(\mathbf{x}^0, \mathbf{y}^0) - F(\mathbf{x}^*, \mathbf{y}^*) < \infty,$$

Then

$$\lim_{k\to\infty} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 = \lim_{k\to\infty} (\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2) = 0.$$

Thus, $\lim_{k\to\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0$ and $\lim_{k\to\infty} \|\mathbf{y}^{k+1} - \mathbf{y}^k\| = 0$. □

In order to prove a global convergence of the whole sequence $\{(\mathbf{x}^k, \mathbf{y}^k)\}$, we first prove the following results.

Lemma 5.4 *Suppose Assumption 5.1 holds, and* $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k\in\mathbb{N}}$ *is generated by the APG algorithm with the initial point* $(\mathbf{x}^0, \mathbf{y}^0)$. *Let*

$$q_{\mathbf{x}}^{k+1} := \nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k) - \frac{1}{t_1}(\mathbf{x}^{k+1} - \mathbf{x}^k),$$

$$q_{\mathbf{y}}^{k+1} := \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k) - \frac{1}{t_2}(\mathbf{y}^{k+1} - \mathbf{y}^k),$$

then

$$q_{\mathbf{x}}^{k+1} \in \partial_{\mathbf{x}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}), \quad q_{\mathbf{y}}^{k+1} \in \partial_{\mathbf{y}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$$

and

$$\|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\| \leq \alpha\|(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})\|,$$

where $\alpha^2 = \max\left\{2\big(L_1 + \frac{1}{t_1}\big)^2, 2L_1^2 + (L_2 + \frac{1}{t_2})^2\right\}$.

Proof It follows from (5.2) that

$$0 \in \tfrac{1}{t_1}(\mathbf{x}^{k+1} - \mathbf{x}^k) + \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k) + \lambda_1 \partial \Phi_1(\mathbf{x}^{k+1}), \tag{5.15}$$

Adding $\nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$ to both sides of (5.15) and rearranging terms, we obtain

$$\begin{aligned} \nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k) - \frac{1}{t_1}(\mathbf{x}^{k+1} - \mathbf{x}^k) &\in \nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \lambda_1 \partial \Phi_1(\mathbf{x}^{k+1}) \\ &= \partial_{\mathbf{x}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}). \end{aligned} \tag{5.16}$$

Similarly, it follows from (5.7) that

$$0 \in \tfrac{1}{t_2}(\mathbf{y}^{k+1} - \mathbf{y}^k) + \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k) + \lambda_2 \partial \Phi_2(\mathbf{y}^{k+1}). \tag{5.17}$$

Adding $\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$ to both sides of (5.17) and rearranging terms, we obtain

$$\begin{aligned} \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k) - \frac{1}{t_2}(\mathbf{y}^{k+1} - \mathbf{y}^k) &\in \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \lambda_2 \partial \Phi_2(\mathbf{y}^{k+1}) \\ &= \partial_{\mathbf{y}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}). \end{aligned} \tag{5.18}$$

Combining (5.16) and (5.18), we obtain

$$q_{\mathbf{x}}^{k+1} \in \partial_{\mathbf{x}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}), \quad q_{\mathbf{y}}^{k+1} \in \partial_{\mathbf{y}} F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}), \tag{5.19}$$

which then implies $(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1}) \in \partial F(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$.

From (5.19) and Assumption 5.1 (iii), we have

$$\begin{aligned} \|q_{\mathbf{x}}^{k+1}\| &= \|\nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k) - \frac{1}{t_1}(\mathbf{x}^{k+1} - \mathbf{x}^k)\| \\ &\le \|\nabla_{\mathbf{x}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^k, \mathbf{y}^k)\| + \frac{1}{t_1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \\ &\le \left(L_1 + \frac{1}{t_1}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\| + L_1\|\mathbf{y}^{k+1} - \mathbf{y}^k\|. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \|q_{\mathbf{y}}^{k+1}\| &= \|\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k) - \frac{1}{t_2}(\mathbf{y}^{k+1} - \mathbf{y}^k)\| \\ &\le \|\nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{k+1}, \mathbf{y}^k)\| + \frac{1}{t_2}\|\mathbf{y}^{k+1} - \mathbf{y}^k\| \\ &\le \left(L_2 + \frac{1}{t_2}\right)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|. \end{aligned}$$

then

$$\begin{aligned} \|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\|^2 &= \|q_{\mathbf{x}}^{k+1}\|^2 + \|q_{\mathbf{y}}^{k+1}\|^2 \\ &\le 2\left(L_1 + \frac{1}{t_1}\right)^2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \left(2L_1^2 + \big(L_2 + \frac{1}{t_2}\big)^2\right)\|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2. \end{aligned}$$

Let $\alpha^2 = \max\{2\big(L_1 + \frac{1}{t_1}\big)^2, 2L_1^2 + (L_2 + \frac{1}{t_2})^2\}$, then we have

$$\|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\|^2 \le \alpha^2(\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2).$$

As a consequence, we get

$$\|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\| \le \alpha\|(\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{y}^{k+1} - \mathbf{y}^k)\|.$$

The proof is thus complete. □

To analyze the convergence of the generated sequence of the APG algorithm, we discuss some properties of the limit point sets of the sequence at first. For convenience, we denote $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$ and $F(\mathbf{z}^k) = F(\mathbf{x}^k, \mathbf{y}^k)$. The set of all limit points of $\{\mathbf{z}^k\}$ is denoted by $\Gamma(\mathbf{z}^0)$, i.e.,

$$\Gamma(\mathbf{z}^0) = \left\{ \bar{\mathbf{z}} = (\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathbb{R}^{n+m} | \ \exists \{k_j\} \in \mathbb{N},\ s.t.\ \mathbf{z}^{k_j} \to \bar{\mathbf{z}}, j \to \infty \right\}.$$

Theorem 5.5 *Suppose Assumption 5.1 holds. Let $\{\mathbf{z}^k\}$ be generated by the APG algorithm with the initial point $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0)$, then the following statements hold.*

(i) *$\Gamma(\mathbf{z}^0)$ is a nonempty and compact set, and the objective value of F is finite and constant on $\Gamma(\mathbf{z}^0)$.*

(ii) *$\Gamma(\mathbf{z}^0) \subset \mathrm{crit} F$.*

(iii) *$\lim\limits_{k\to\infty} \mathrm{dist}(\mathbf{z}^k, \Gamma(\mathbf{z}^0)) = 0$.*

Proof (i). From the boundedness of $\{\mathbf{z}^k\}$, it follows that $\Gamma(\mathbf{z}^0)$ is nonempty. Note that $\Gamma(\mathbf{z}^0)$ can be represented as an intersection of compact sets, i.e.,

$$\Gamma(\mathbf{z}^0) = \cap_{s\in\mathbb{N}} \overline{\cup_{k\geq s}\{\mathbf{z}^k\}}.$$

Since the intersection of bounded closed sets is still bounded and closed, $\Gamma(\mathbf{z}^0)$ is also a compact set. For any $\bar{\mathbf{z}} = (\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \Gamma(\mathbf{z}^0)$, $\exists\{k_j\} \subset \mathbb{N}$ such that

$$\lim_{j\to\infty} \mathbf{z}^{k_j} = \bar{\mathbf{z}}.$$

By the continuity of F, we have

$$\lim_{j\to\infty} F(\mathbf{z}^{k_j}) = F(\bar{\mathbf{z}}).$$

From Theorem 5.3 (i), we have $F(\mathbf{z}^k) \to F^*(k \to \infty)$. Then, for arbitrary subsequence $F(\mathbf{z}^{k_j})$, it holds

$$\lim_{j\to\infty} F(\mathbf{z}^{k_j}) = F(\bar{\mathbf{z}}) = F^*. \tag{5.20}$$

That is, the value of F on $\Gamma(\mathbf{z}^0))$ is a constant.

(ii) From Theorem 5.3 (ii), we have

$$\lim_{j\to\infty} \|\mathbf{x}^{k_j+1} - \mathbf{x}^{k_j}\| = 0,\ \lim_{j\to\infty} \|\mathbf{y}^{k_j+1} - \mathbf{y}^{k_j}\| = 0,$$

then

$$\lim_{j\to\infty} \mathbf{x}^{k_j+1} = \lim_{j\to\infty} \mathbf{x}^{k_j} = \bar{\mathbf{x}},\ \lim_{j\to\infty} \mathbf{y}^{k_j+1} = \lim_{j\to\infty} \mathbf{y}^{k_j} = \bar{\mathbf{y}}.$$

From Lemma 5.4, we have

$$\|(q_{\mathbf{x}}^{k_j+1}, q_{\mathbf{y}}^{k_j+1})\| \leq \alpha \|(\mathbf{x}^{k_j+1} - \mathbf{x}^{k_j}, \mathbf{y}^{k_j+1} - \mathbf{y}^{k_j})\|.$$

Let $j \to \infty$, then

$$\lim_{j\to\infty} \|(q_{\mathbf{x}}^{k_j+1}, q_{\mathbf{y}}^{k_j+1})\| = 0, \quad \text{i.e.,} \quad (q_{\mathbf{x}}^{k_j+1}, q_{\mathbf{y}}^{k_j+1}) \to (0,0),\ for\ j \to \infty.$$

From Lemma 5.4, we know $(q_{\mathbf{x}}^{k_j+1}, q_{\mathbf{y}}^{k_j+1}) \in \partial F(\mathbf{x}^{k_j+1}, \mathbf{y}^{k_j+1}) \subset \partial^C F(\mathbf{x}^{k_j+1}, \mathbf{y}^{k_j+1})$. Further, by the closedness of the mapping $\partial^C F(\cdot)$ [13, Propostion 2.1.5(b)], we obtain

$$(0,0) \in \partial^C F(\bar{\mathbf{x}}, \bar{\mathbf{y}}).$$

From Remark 3.6, this implies that $\bar{\mathbf{z}} = (\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a critical point of problem (1.2), and $\Gamma(\mathbf{z}^0) \subset \mathrm{crit} F$.

(iii) This conclusion follows from the definition of $\Gamma(\mathbf{z}^0)$. □

In order to give the global convergence of the whole sequence $\{(\mathbf{x}^k, \mathbf{y}^k)\}$, we first introduce the Kurdyka-Łojasiewicz (KŁ) property of F. The KŁ property was used to analyze smooth problems, then Bolte, Daniilidis and Lewis [6] used KŁ property to analyze nonsmooth problems. Since then, lots of researchers have done much research on this basis, for example, [1,2,7,27]. Now, it is well-known that the KŁ property have played the important roles in the convergence analysis of proximal algorithms. Let's recall the KŁ property.

Let $\eta \in (0, +\infty]$, we denote by Ψ_η the class of all concave and continuous functions $\psi : [0, \eta) \to [0, \infty)$ such that

(i) $\psi(0) = 0$;

(ii) ψ is continuously differentiable on $(0, \eta)$;

(iii) $\psi'(s) > 0$ for all $s \in (0, \eta)$.

Definition 5.6 *[2, 7, 27] [KŁ property] Let $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function.*

(i) *h is said to have the KŁ property at $\bar{\mathbf{w}} \in \mathrm{dom}\partial h := \{\mathbf{w} \in \mathbb{R}^{n+m} | \partial h(\mathbf{w}) \neq \emptyset\}$, if there exist $\eta \in (0, +\infty]$, a neighborhood Ω of $\bar{\mathbf{w}}$ and a function $\psi \in \Psi_\eta$, such that for all*

$$\mathbf{w} \in \Omega \cap [h(\bar{\mathbf{w}}) < h(\mathbf{w}) < h(\bar{\mathbf{w}}) + \eta],$$

the following inequality holds

$$\psi'(h(\mathbf{w}) - h(\bar{\mathbf{w}}))\mathrm{dist}(0, \partial h(\mathbf{w})) \geq 1.$$

(ii) *If h satisfies the KŁ property at each point of* $\mathrm{dom}\partial h$*, then h is called a KŁ function.*

Lemma 5.7 *[7, 27] [Uniformized KŁ property] Let Ω be a compact set and $h : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{\infty\}$ be a proper lower semicontinuous function. Assume that h is constant on Ω and satisfies the KŁ property at each point of Ω. Then there exist $\epsilon > 0, \eta > 0$ and $\psi \in \Psi_\eta$ such that for all $\bar{\mathbf{w}}$ in Ω and all*

$$\mathbf{w} \in \{\mathbf{w} \in \mathbb{R}^{n+m} : \mathrm{dist}(\mathbf{w}, \Omega) < \epsilon\} \cap [h(\bar{\mathbf{w}}) < h(\mathbf{w}) < h(\bar{\mathbf{w}}) + \eta],$$

one has,

$$\psi'(h(\mathbf{w}) - h(\bar{\mathbf{w}}))\mathrm{dist}(0, \partial h(\mathbf{w})) \geq 1.$$

The KŁ functions have a wide range including semi-algebraic, subanalytic and log-exp and so on [7]. It is easy to check that our objective function F in (1.2) meets the KŁ property.

Now we can give the global convergence of the whole sequence $\{(\mathbf{x}^k, \mathbf{y}^k)\}$ under the condition of KŁ function.

Theorem 5.8 *Suppose Assumption 5.1 holds and F is a KŁ function. Let $\{\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)\}$ be generated by the APG algrithom. Then the following statements hold.*

(i) *$\sum_{k=0}^{\infty} \|\mathbf{z}^{k+1} - \mathbf{z}^k\| < \infty$;*

(ii) *The sequence $\{\mathbf{z}^k\}_{k \in N}$ converges to a critical point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ of problem (1.2).*

Proof (i). Firstly, we suppose that $F(\mathbf{z}^k) \neq F(\bar{\mathbf{z}})$ for all $k \in \mathbb{N}$; Otherwise, the algorithm will terminate.

On the one hand, it follows from (5.20) that $\lim_{k\to\infty} F(\mathbf{z}^k) = F^* = F(\bar{\mathbf{z}})$. Then, for any $\eta > 0$, there exists $k_0 > 0$, such that for any $k > k_0$, it holds

$$F(\bar{\mathbf{z}}) < F(\mathbf{z}^k) < F(\bar{\mathbf{z}}) + \eta,$$

that is,

$$\mathbf{z}^k \in [F(\bar{\mathbf{z}}) < F(\mathbf{z}) < F(\bar{\mathbf{z}}) + \eta], \;\; \forall k > k_0.$$

On the other hand, by Theorem 5.5 (iii), we have $\lim_{k\to\infty} \mathrm{dist}(\mathbf{z}^k, \Gamma(\mathbf{z}^0)) = 0$. Therefore, for any $\epsilon > 0$, there exists $k_1 > 0$, such that for any $k > k_1$, it holds

$$\mathrm{dist}(\mathbf{z}^k, \Gamma(\mathbf{z}^0)) < \epsilon.$$

Let $k_2 = \max\{k_0, k_1\}$, then for any $k > k_2$, we have

$$\mathbf{z}^k \in \left\{\mathbf{z} \mid \mathrm{dist}(\mathbf{z}, \Gamma(\mathbf{z}^0)) < \epsilon\right\} \cap \left[F(\bar{\mathbf{z}}) < F(\mathbf{z}) < F(\bar{\mathbf{z}}) + \eta\right], \;\; \forall k > k_2.$$

Since the value of F on $\Gamma(\mathbf{z}^0)$ is a constant, by the uniformized KŁ property (Lemma 5.7), there exists $\psi \in \Psi_\eta$, such that

$$\psi'(F(\mathbf{z}^k) - F(\bar{\mathbf{z}}))\mathrm{dist}(0, \partial F(\mathbf{z}^k)) \geq 1. \tag{5.21}$$

By Lemma 5.4, we have $(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1}) \in \partial F(\mathbf{z}^{k+1})$ and $\|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\| \leq \alpha\|(\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{y}^{k+1} - \mathbf{y}^k)\|$, then

$$\mathrm{dist}(0, \partial F(\mathbf{z}^k)) \leq \|(q_{\mathbf{x}}^{k+1}, q_{\mathbf{y}}^{k+1})\| \leq \alpha\|(\mathbf{x}^{k+1} - \mathbf{x}^k, \mathbf{y}^{k+1} - \mathbf{y}^k)\| = \alpha\|\mathbf{z}^{k+1} - \mathbf{z}^k\|.$$

Substitute the above inequality into (5.21), then we obtain

$$\psi'(F(\mathbf{z}^k) - F(\bar{\mathbf{z}})) \geq \frac{1}{\mathrm{dist}(0, \partial F(\mathbf{z}^k))} \geq \frac{1}{\alpha\|\mathbf{z}^{k+1} - \mathbf{z}^k\|}.$$

Since ψ is a concave function, we have

$$\psi(F(\mathbf{z}^{k+1}) - F(\bar{\mathbf{z}})) \leq \psi(F(\mathbf{z}^k) - F(\bar{\mathbf{z}})) + \psi'(F(\mathbf{z}^k) - F(\bar{\mathbf{z}}))(F(\mathbf{z}^{k+1}) - F(\mathbf{z}^k)).$$

Due to the above two inequalituies and the sufficient descending property of function F given by Lemma 5.2, we get

$$\begin{aligned}
&\psi(F(\mathbf{z}^k) - F(\bar{\mathbf{z}})) - \psi(F(\mathbf{z}^{k+1}) - F(\bar{\mathbf{z}})) \\
\geq\ &\psi'(F(\mathbf{z}^k) - F(\bar{\mathbf{z}}))(F(\mathbf{z}^k) - F(\mathbf{z}^{k+1})) \\
\geq\ &\frac{\rho\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2}{\alpha\|\mathbf{z}^k - \mathbf{z}^{k-1}\|}.
\end{aligned}$$

Denote by $C = \alpha/\rho$ and $\triangle_k = \psi(F(\mathbf{z}^k) - F(\bar{\mathbf{z}}))$, then $\triangle_k$ is monotonically non-increasing with respect to k, and $\bar{\triangle} = \lim_{k\to\infty} \triangle_k$ makes sense. Therefore, the above inequality can be rewritten as

$$\triangle_k - \triangle_{k+1} \geq \frac{\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2}{C\|\mathbf{z}^k - \mathbf{z}^{k-1}\|}.$$

By the inequality $4ab \leq (a+b)^2$, then

$$\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2 \leq C\|\mathbf{z}^k-\mathbf{z}^{k-1}\|(\triangle_k - \triangle_{k+1}) \leq \left(\frac{\|\mathbf{z}^k-\mathbf{z}^{k-1}\| + C(\triangle_k - \triangle_{k+1})}{2}\right)^2,$$

hence,

$$2\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \leq \|\mathbf{z}^k-\mathbf{z}^{k-1}\| + C(\triangle_k - \triangle_{k+1}).$$

Summing the left and right sides of the above inequality respecting to k, we obtain

$$\begin{aligned} 2\sum_{k=k_2+1}^{K}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| &\leq C(\triangle_{k_2+1} - \triangle_{K+1}) + \sum_{k=k_2+1}^{K}\|\mathbf{z}^k-\mathbf{z}^{k-1}\| \\ &= C(\triangle_{k_2+1} - \triangle_{K+1}) + \|\mathbf{z}^{k_2+1}-\mathbf{z}^{k_2}\| \\ &\quad -\|\mathbf{z}^{K+1}-\mathbf{z}^K\| + \sum_{k=k_2+1}^{K}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|, \end{aligned}$$

then

$$\sum_{k=k_2+1}^{K}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \leq C(\triangle_{k_2+1} - \triangle_{K+1}) + \|\mathbf{z}^{k_2+1}-\mathbf{z}^{k_2}\| - \|\mathbf{z}^{K+1}-\mathbf{z}^K\|.$$

Letting $K \to \infty$, we obtain

$$\sum_{k=k_2+1}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \leq C(\triangle_{k_2+1} - \bar{\triangle}) + \|\mathbf{z}^{k_2+1}-\mathbf{z}^{k_2}\| < \infty.$$

Therefore,

$$\sum_{k=0}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| = \sum_{k=0}^{k_2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| + \sum_{k=k_2+1}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| < \infty.$$

(ii) For any $p > q \geq k_2$, we have

$$\|\mathbf{z}^p - \mathbf{z}^q\| = \|\textstyle\sum_{k=q}^{p-1}(\mathbf{z}^{k+1}-\mathbf{z}^k)\| \leq \sum_{k=q}^{p-1}\|\mathbf{z}^{k+1}-\mathbf{z}^k\| < \sum_{k=q}^{\infty}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|.$$

Then $\|\mathbf{z}^p - \mathbf{z}^q\| \to 0$ as $q \to \infty$, which indicates that $\{\mathbf{z}^k\}$ is a Cauchy sequence, and hence is a convergent sequence. It then follows from Theorem 5.5 that the limit point $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ of $\{\mathbf{z}^k\}$ is a critical point of problem (1.2). The proof is thus complete. □

6 Numerical experiments

In this section, we conduct numerical experiments on the relaxation problem (1.2) to test the APG algorithm.

All the numerical experiments are implemented in MATLAB R2018b and on a Lenovo PC (Intel(R) Core(TM) i5-9500, 3.00GHz, 8.00GB of RAM).

6.1 Simulated Data

In this simulation experiment part, the APG algorithm is applied to solve the following model.

Example 6.1 *We consider the least square loss* $f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{c}\|^2$*, that is*

$$\min_{\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m} \quad \frac{1}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{c}\|^2 + \lambda_1\Phi_1(\mathbf{x}) + \lambda_2\Phi_2(\mathbf{y}), \tag{6.1}$$

where $A \in \mathbb{R}^{p\times n}$*,* $B \in \mathbb{R}^{p\times m}$*,*

$$\Phi_1(\mathbf{x}) := \sum_{i=1}^{n} \varphi_1(|x_i|), \quad \Phi_2(\mathbf{y}) := \sum_{j=1}^{J} \varphi_2(\|\mathbf{y}_{(j)}\|)$$

and φ_i $(i = 1, 2)$ *is defined in (1.2).*

For this model, the data are generated as follows. We first use MATLAB codes `randn(p,n)` and `randn(p,m)` to randomly generate the i.i.d. Gaussian matrices $A \in \mathbb{R}^{p\times n}$ and $B \in \mathbb{R}^{p\times m}$. Then we generate a sparse solution $\mathbf{xo} \in \mathbb{R}^n$ and a group sparse solution $\mathbf{yo} \in \mathbb{R}^m$ as the real solution. Let kkx be the number of non-zero entries of $\mathbf{xo}$, then the sparsity level of $\mathbf{xo}$ is kkx/n. Meanwhile, $\mathbf{yo} \in \mathbb{R}^m$ is randomly divided into J groups. The kky non-zero groups are randomly selected from these J groups, and the remaining $J - kky$ groups are all set to be zero vectors, so the group sparsity level of $\mathbf{yo}$ is kky/J.

For the given positive integers p, n, m, J, kkx, kky, the real solution $\mathbf{zo} = (\mathbf{xo}, \mathbf{yo})$ are generated by the following codes:

```
xo =zeros(n,1); Indx =randperm(n); xo(Indx(1:kkx)) =randn(kkx,1);
avgsize =floor(m/J); idy =[ ]; gidy =[gidy; j*ones(avgsize,1)], j =1:J;
qqy =randperm(J); suppy =sort(qqy(1:kky)); yo =zeros(m,1);
idy =find(gidy ==suppy(k)), yo(idy) =randn(avgsize,1), k =1:kky;
zo =[xo;yo];
```

The observed data $\mathbf{c} \in \mathbb{R}^p$ is generated by

$$\texttt{c = A*xo + B*yo + }\sigma\texttt{*randn(p,1)},$$

where σ is the standard deviation of additive Gaussian noise.

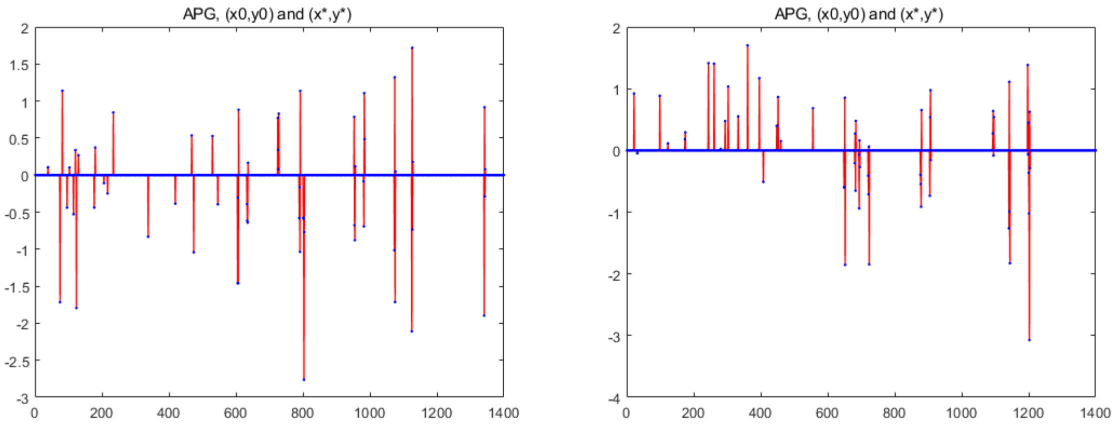
The parameters and initial values in the APG algorithm are given as follows: $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) = \mathbf{0}_{n+m}$, $t_1 = 0.7$, $t_2 = 0.9$. In each iteration of the APG algorithm, we sort $\mathrm{Ex} = \{|x_i^k|\}_{i\in[n]}$ and $\mathrm{Ey} = \{\|[\mathbf{y}^k]_{(j)}\|\}_{j\in[J]}$ in ascending order, we take $\mathrm{crix} = \mathrm{Ex}_{n-kkx}$, $\mathrm{criy} = \mathrm{Ey}_{(m-kky)}$, $\alpha_1 = 1.2 * \mathrm{crix}$, $\alpha_2 = 1.8 * \mathrm{criy}$, $\lambda_1 = \mathrm{crix} * \alpha_1/t_1$, and $\lambda_2 = \mathrm{criy} * \alpha_2/t_2$. Let $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*) \in \mathbb{R}^{n+m}$ denote the solution produced by the APG algorithm.

In this example, for each set of given numbers $\{p, n, m, J = m/4, kkx, kky, \sigma\}$, we run 100 instances and use three indicators to evaluate the experimental effect of the proposed APG algorithm: average relative error (Rel-err $:= \frac{\|\mathbf{z}^* - \mathbf{zo}\|}{\max\{1, \|\mathbf{zo}\|\}}$), average CPU time and successful rate (Suc-rat) where Rel-err $< 10^{-2}$ is regarded success. Set $xtol = 10^{-4}$. The experimental results are shown in Table 1, where we consider two cases: noiseless $\sigma = 0$ and noised $\sigma = 10^{-3}$.

In Figure 1, the scatter plots of real and numerical solutions for $p = 2000$, $n = 3000$, $m = 4000$, $kkx = 50$, $kky = 20$, $J = 1000$ are displayed.

Table 1: Average numerical results of the APG algorithm

Problem						$\sigma = 0$			$\sigma = 10^{-3}$		
p	n	m	kkx	kky	J	Time	Rel-err	Suc-rat	Time	Rel-err	Suc-rat
800	1200	1600	5	5	400	0.08	1.58e − 4	100%	0.22	1.69e − 3	100%
800	1200	1600	40	20	400	0.11	2.88e − 4	100%	0.25	1.59e − 3	100%
800	1200	1600	80	40	400	0.19	6.19e − 4	100%	0.55	1.65e − 3	100%
1000	1500	2000	5	5	500	0.12	1.36e − 4	100%	0.12	2.05e − 3	100%
1000	1500	2000	100	50	500	0.31	5.74e − 4	100%	0.53	1.73e − 3	100%
2000	3000	4000	5	5	1000	0.49	1.39e − 4	100%	0.43	2.25e − 3	100%
2000	3000	4000	50	20	1000	0.54	1.79e − 4	100%	1.41	1.55e − 3	100%
2000	3000	4000	200	100	1000	1.14	5.77e − 4	100%	1.19	1.63e − 3	100%



(a) $\sigma = 0$ (b) $\sigma = 10^{-3}$

Figure 1. Visual numerical results

From Table 1 and Figure 1, we can see that the proposed APG algorithm can quickly obtain the true solution with high success rate.

Next, we compare our APG algorithm with several state-of-art algorithms: PGM-GSO algorithm [22] for solving ℓ_2-$\ell_{p,q}$ model: $\min \|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_{p,q}^q$ $(p \geq 1, 0 \leq q \leq 1)$, IRLS-th algrithm [18] for solving $\ell_{2,q}$ model: $\min \|A\mathbf{x}-\mathbf{b}\|_2^2+\lambda\|\mathbf{x}\|_{2,q}^q$ $(0 < q < 1)$, GCD algorithm [8] for solving group MCP model, and SPGl1 algrithm [15] for solving group lasso model: $\min \|\mathbf{x}\|_{2,1}$ $s.t.$ $\|A\mathbf{x}-\mathbf{b}\|_2 \leq \delta$. One can refer to the references for their implementation details. In order for these algorithms to be used to solve problem (6.1), we group all partial sparse and partial group sparse data into groups as follows:

```
gidxy=[]; Jx=floor(n/avgsize); gidxy=[gidxy;i*ones(avgsize,1)], i=1:(Jx+J);
```

The above grouping way is applied to PGM-GSO, GCD, SPGl1 and IRLS-th. We run 100 times for each instant and record the average CPU time and the average relative error, as shown in Table 2 and Table 3.

Table 2: Comparison of five algorithms for problem (6.1) with $\sigma = 0$

Problem					APG		PGM-GSO		SPGl1		GCD		IRLS-th	
p	n	m	kkx	kky	Time	Rel-err	Time	Rel-err	Time	Rel-err	Time	Rel-err	Time	Rel-err
400	600	800	5	5	0.03	3.59e − 4	0.04	2.05e − 4	0.01	3.18e − 4	1.17	3.41e − 2	0.13	1.99e − 3
400	600	800	25	15	0.05	8.15e − 4	0.08	7.16e − 4	0.06	3.50e − 4	3.27	4.21e − 2	0.42	3.87e − 3
800	1200	1600	5	5	0.14	3.10e − 4	0.16	1.25e − 4	0.03	2.87e − 4	3.38	3.91e − 2	0.59	6.28e − 4
800	1200	1600	20	10	0.17	3.48e − 4	0.18	1.92e − 4	0.05	3.91e − 4	4.10	4.10e − 2	0.75	1.17e − 2
800	1200	1600	40	20	0.24	5.65e − 4	0.24	4.87e − 4	0.11	6.03e − 4	6.79	2.97e − 2	1.04	1.73e − 2
1000	1500	2000	5	5	0.25	2.54e − 4	0.27	1.35e − 4	0.06	2.90e − 4	4.06	1.53e − 2	1.01	4.06e − 4
1000	1500	2000	60	30	0.40	7.49e − 4	0.42	5.91e − 4	0.26	7.59e − 4	13.07	2.89e − 2	2.06	2.61e − 2
2000	3000	4000	5	5	0.82	3.00e − 4	1.32	1.09e − 4	0.18	1.58e − 4	12.96	1.85e − 2	4.94	5.58e − 4
2000	3000	4000	50	20	0.95	3.69e − 4	1.46	1.95e − 4	0.34	1.48e − 4	26.84	2.20e − 2	7.08	7.62e − 4
2000	3000	4000	100	50	1.34	5.35e − 4	1.81	4.67e − 4	0.83	3.00e − 4	43.37	1.68e − 2	9.76	2.79e − 3
4000	6000	8000	5	5	2.92	2.57e − 4	8.63	9.79e − 5	0.75	2.01e − 4	50.47	8.99e − 17	30.52	3.54e − 4
4000	6000	8000	60	30	3.42	3.17e − 4	8.89	1.40e − 4	1.40	1.02e − 4	93.57	6.65e − 3	43.74	3.02e − 3
4000	6000	8000	200	100	5.40	6.73e − 4	10.57	4.12e − 4	3.41	2.26e − 4	144.43	1.15e − 2	83.90	1.29e − 2
6000	9000	12000	5	5	6.61	2.43e − 4	93.44	8.95e − 5	1.47	5.10e − 5	166.46	1.12e − 16	283.06	2.94e − 4
6000	9000	12000	80	40	7.54	2.41e − 4	104.10	1.77e − 4	2.81	4.40e − 5	186.57	6.03e − 3	456.53	2.82e − 3
6000	9000	12000	300	150	10.90	5.20e − 4	99.21	4.35e − 4	6.17	1.70e − 4	312.56	7.78e − 3	544.52	1.21e − 2

Table 3: Comparison of five algorithms for problem (6.1) with $\sigma = 0.01$

Problem					APG		PGM-GSO		SPGl1		GCD		IRLS-th	
p	n	m	kkx	kky	Time	Rel-err	Time	Rel-err	Time	Rel-err	Time	Rel-err	Time	Rel-err
400	600	800	5	5	0.06	3.69e − 2	0.05	2.02e − 2	0.03	4.17e − 2	1.06	1.99e − 2	0.26	6.53e − 2
400	600	800	25	15	0.07	3.72e − 2	0.08	2.40e − 2	0.02	1.00e − 1	1.83	3.03e − 2	0.37	4.55e − 2
800	1200	1600	5	5	1.57	3.27e − 2	0.18	1.87e − 2	0.04	2.95e − 2	2.03	1.95e − 2	0.85	6.25e − 2
800	1200	1600	20	10	1.62	3.66e − 2	0.20	2.03e − 2	0.04	3.74e − 2	3.50	2.43e − 2	1.14	4.93e − 2
800	1200	1600	40	20	0.28	3.43e − 2	0.30	1.18e − 2	0.05	3.21e − 2	2.94	1.19e − 2	2.08	6.89e − 2
1000	1500	2000	5	5	0.26	4.29e − 2	0.28	1.84e − 2	0.05	3.05e − 2	4.69	2.37e − 2	2.03	8.72e − 2
1000	1500	2000	60	30	2.99	3.92e − 2	0.50	2.45e − 2	0.12	6.09e − 2	12.79	2.83e − 2	2.75	4.18e − 2
2000	3000	4000	5	5	0.89	2.75e − 2	1.38	1.28e − 2	0.18	3.00e − 2	12.43	1.44e − 2	9.79	9.93e − 2
2000	3000	4000	50	20	10.06	3.14e − 2	1.57	2.06e − 2	0.23	3.78e − 2	26.96	2.51e − 2	11.14	5.13e − 2
2000	3000	4000	100	50	10.02	3.25e − 2	2.01	2.29e − 2	0.41	5.29e − 2	39.29	2.87e − 2	14.35	4.41e − 2
4000	6000	8000	5	5	2.94	3.45e − 2	8.23	1.61e − 2	0.62	2.83e − 2	43.36	1.64e − 2	49.48	1.11e − 1
4000	6000	8000	60	30	36.14	2.68e − 2	8.59	1.84e − 2	0.85	3.55e − 2	89.79	2.25e − 2	64.84	5.54e − 2
4000	6000	8000	200	100	35.95	2.17e − 2	9.89	2.17e − 2	1.49	5.08e − 2	124.66	2.75e − 2	86.07	5.25e − 2
6000	9000	12000	5	5	7.68	3.96e − 2	297.53	2.16e − 2	3.31	3.93e − 2	73.93	1.95e − 2	3810.94	1.66e − 1
6000	9000	12000	80	40	78.77	2.90e − 2	555.80	1.97e − 2	13.59	3.74e − 2	178.94	2.71e − 2	4951.63	6.65e − 2
6000	9000	12000	300	150	80.17	2.89e − 2	967.80	2.10e − 2	10.82	4.77e − 2	319.92	2.83e − 2	12533.64	4.79e − 2

From Tables 2 and 3, we can observe that in the absence of noise, the average relative errors of APG are similar to SPGl1 and PGM-GSO, but smaller than IRLS-th and GCD in most cases; Meanwhile, the average CPU time of APG is less than PGM-GSO, GCD and

IRLS-th but more than SPGl1. In the presence of noise, the average relative errors of APG are similar to the other four algorithms; Meanwhile, the average CPU time of APG is more than SPGl1, but less than SPGl1, GCD and IRLS-th; It is funny that, for the small scale instances, the average CPU time of APG is more than PGM-GSO but for the large scale instances, the average CPU time of APG is less than PGM-GSO. The results indicate that our APG algorithm is competitive with the four state-of-art algorithms in solving problem (6.1).

6.2 Multichannel image reconstruction

In this section, we consider recovering three-channel images from compressive and noisy measurement. In our experiments, the PSNR (peak signal to noise ratio) is defined by

$$\mathrm{PSNR} = 10 \cdot \log \frac{\mathrm{V}^2}{\mathrm{MSE}},$$

in which V and MSE= $\frac{\|\mathbf{z}-\mathbf{zo}\|^2}{n+m}$ (mean squared error) are the maximum absolute value and the mean squared error of the reconstruction respectively.

The example is taken from [24,30,26,42]. The observed data $\mathbf{c}$ is generated by $\mathbf{c} = \mathbf{Ax} + \mathbf{By} + \sigma * \texttt{randn(p,1)}$, where $\mathbf{A}$, $\mathbf{B}$ are random Gaussian matrices, σ is a positive scalar, $\mathbf{x}$ with sparse struture and $\mathbf{y}$ with group sparse structure are the target coefficients. For this experiment: $n = 48*48*1, m = 48*48*2, p = m/2, J = m/4, kkx = 152, kky = 172$. We still compare experimental results among APG, PGM-GSO, SPGl1, GCD and IRLS-th. The PSNR and CPU time are presented in Table 4, while the original image and the recovered images for $\sigma = 0.1$ are presented in Figure 2.

Table 4: Numerical results for the three-channel image

σ	algorithm	APG	PGM-GSO	SPGl1	GCD	IRLS-th
$\sigma = 0$	CPU time(s)	1.67	3.97	7.70	33.83	17.06
	PSNR	80.11	72.97	61.57	33.27	37.74
$\sigma = 1\mathrm{e}-3$	CPU time(s)	3.11	3.72	1.10	33.56	20.90
	PSNR	64.55	60.16	43.21	60.71	37.75
$\sigma = 1\mathrm{e}-2$	CPU time(s)	5.67	4.38	0.45	29.72	26.77
	PSNR	39.04	37.29	34.68	38.02	33.85
$\sigma = 1\mathrm{e}-1$	CPU time(s)	6.61	5.03	0.38	14.44	77.55
	PSNR	29.29	23.50	24.05	26.72	21.18

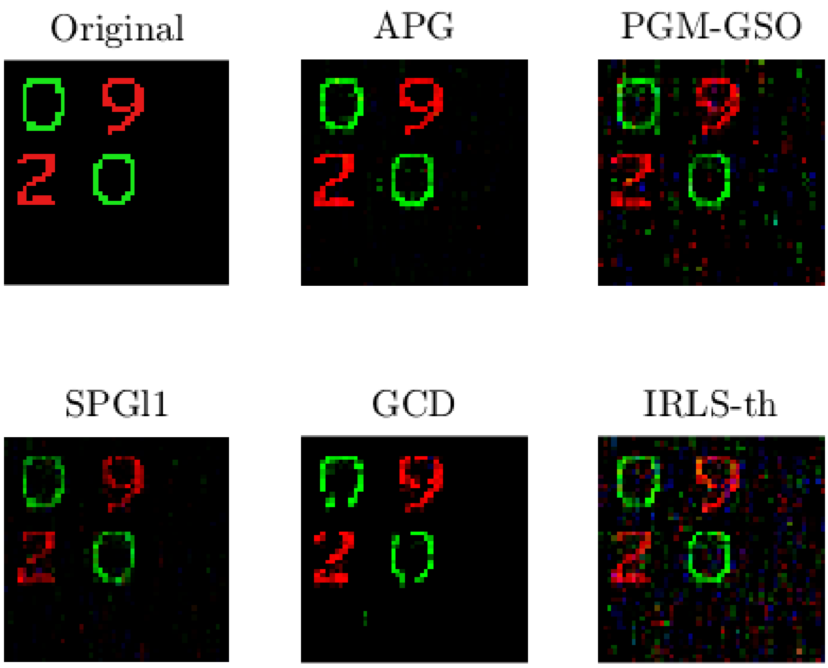


Figure 2. Original image and recovered images by five algorithms for $\sigma = 0.1$

From Table 4 and Figure 2, we can see that APG performs better than PGM-GSO, SPGl1, GCD and IRLS-th in restoring the PSNR value of the image. Although APG is not superior to SPGl1 and PGM-GSO in CPU time, it takes less time than GCD and IRLS-th. The results indicate that our model and APG algorithm are also competitive with the four state-of-art algorithms in multichannel image reconstruction.

7 Conclusion

In this paper, we initially studied the partial sparse and partial group sparse optimization problem. Firstly, we give the Capped-ℓ_1 relaxation and group Capped-ℓ_1 relaxation problem of the original problem. Secondly, we introduced d-stationary point and critical point for the relaxation problem, and prove that any d-stationary point is a critical point. Under some mild assumptions, we gave the lower bound properties of d-stationary points of the relaxation problem, based on which, we proved the equivalence of the original problem and the relaxation problem. This result provides a theoretical basis for solving the original problem via solving the relaxation problem. Then, we proposed an APG algorithm for the relaxation problem, and proved that the whole sequence generated by the APG algorithm converges to a critical point of the relaxation problem. Finally, the rich numerical experiments show that the partial sparse and partial group sparse model and the APG algorithm have good performance and some practical value.

Acknowledgements This work is supported by the National Natural Science Foundation of China (12261020), the Guizhou Provincial Science and Technology Program (ZK[2021]009), the Foundation for Selected Excellent Project of Guizhou Province for High-level Talents Back from Overseas ([2018]03), and the Research Foundation for Postgraduates of Guizhou Province (YJSCXJH[2020]085).

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming, 2009, 116(1-2): 5-16.
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Mathematics of Operations Research, 2010, 35(2): 438-457.
3. Bian, W., Chen, X.: A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. SIAM Journal on Numerical Analysis, 2020, 58(1): 858-883.
4. Bian, W., Chen, X.: Optimality and complexity for constrained optimization problems with nonconvex regularization. Mathematics of Operations Research, 2017, 42(4): 1063-1084.
5. Blumensath, T.: Compressed sensing with nonlinear observations and related nonlinear optimization problems. IEEE Transactions on Information Theory, 2013, 59(6): 3466-3474.
6. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization, 2007, 17(4): 1205-1223.
7. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming, 2014, 146(1-2): 459-494.
8. Breheny, P., Huang, J.: Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing, 2015, 25(2): 173-187.
9. Chandran, M.: Analysis of Bayesian group-Lasso in regression models. University of Florida, 2011.
10. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. IEEE Signal Processing Letters, 2007, 14(10): 707-710.
11. Chen, X., Pan, L.L., Xiu, N.: Solution sets of three sparse optimization problems for multivariate regression. Journal of Global Optimization, 2022, https://doi.org/10.1007/s10898-021-01124-w.
12. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. SIAM Journal on Scientific Computing, 2010, 32(5): 2832-2852.
13. Clarke, F.H.: Optimization and Nonsmooth Analysis. SIAM Journal on Control and Optimization, 1990.
14. Elad, M., Figueiredo, M.A.T., Ma, Y.: On the role of sparse and redundant representations in image processing. Proceedings of the IEEE, 2010, 98(6): 972-982.
15. Van den Berg, E., Friedlander, M.P.: Probing the pareto frontier for basis pursuit solutions. SIAM Journal on Scientific Computing, 2009, 31(2): 890-912.
16. Fan, J., Li, R.: Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the International Congress of Mathematicians, 2006, 3: 595-622.
17. Fan, J., Li, R.: Variable selection via nonconvave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.
18. Feng, X., Yan, S., Wu, C.: The $\ell_{2,q}$ regularized group sparse optimization: lower bound theory, recovery bound and algorithms. Applied and Computational Harmonic Analysis, 2020, 49(2): 381-414.
19. Gong, P., Zhang, C., Lu, Z., Huang, J., Ye, J.: A general iterative shrinkage and thresholding algorithm for nonconvex regularized optimization problems. Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13), **28(2)**, 37-45 (2013)
20. Huang, J., Ma, S., Xie, H., Zhang, C.H.: A group bridge approach for variable selection. Biometrika, 2009, 96(2): 339-355.
21. Huang, J., Zhang, T.: The benefit of group sparsity. The Annals of Statistics, 2010, 38(4): 1978-2004.
22. Hu, Y., Li, C., Meng, K.: Group sparse optimization via $\ell_{p,q}$ regularization. Journal of Machine Learning Research, 2017, 18(30): 1-52.
23. Jiang, D.: Concave selection in generalized linear models. University of Iowa, 2012.
24. Jiao, Y., Jin, B., Lu, X.: Group sparse recovery via the $\ell_0(\ell_2)$ penalty: theory and algorithm. IEEE Transactions on Signal Processing, 2017, 65(4): 998-1012.

25. Le Thi, H.A., Pham Dinh, T., Le, H.M., Vo, X.T.: DC approximation approaches for sparse optimization. European Journal of Operational Research, 2015, 244(1): 26-46.
26. Li, W., Bian, W., Toh, K.C.: DC algorithms for a class of sparse group ℓ_0 regularized optimization problems. SIAM Journal on Optimization, 2022, 32(3): 1614-1641.
27. Nikolova, M., Tan, P.: Alternating structure-adapted proximal gradient descent for nonconvex nonsmooth block-regularized problems. SIAM Journal on Optimization, 2019, 29(3): 2053-2078.
28. Ong, C.S., An, L.T.H.: Learning sparse classifiers with difference of convex functions algorithms. Optimization Methods and Software, 2013, 28(4): 830-854.
29. Pang, J.S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Mathematics of Operations Research, 2017, 42(1): 95-118.
30. Pan, L., Chen, X.: Group sparse optimization for images recovery using capped folded concave functions. SIAM Journal on Imaging Sciences, 2021, 14(1): 1-25.
31. Peng, D., Chen, X.: Computation of second-order directional stationary points for group sparse optimization. Optimization Methods and Software, 2020, 35(2): 348-376.
32. Phan, D.N., Le Thi, H.A.: Group variable selection via $\ell_{p,0}$ regularization and application to optimal scoring. Neural Networks, 2019, 118: 220-234.
33. Raman, S., Fuchs, T.J., Wild, P.J.: The Bayesian group-Lasso for analyzing contingency tables. Proceedings of the 26th Annual International Conference on Machine Learning, 2009, 881-888.
34. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer Science & Business Media, 2009.
35. Simon, N., Friedman, J., Hastie T., Tibshirani R.: A sparse-group Lasso. Journal of computational and graphical statistics, 2013, 22(2): 231-245.
36. Soubies, E., Blanc-Féraud, L., Aubert, G.: A continous exact ℓ_0 penalty (Capped-ℓ_0) for least squares regularized problem. SIAM Journal on Imaging Sciences, 2015, 8(3): 1574-1606.
37. Soubies, E., Blanc-Féraud, L., Aubert, G.: A unified view of exact continuous penalties for $\ell_2 - \ell_0$ minimization. SIAM Journal on Optimization, 2017, 27(3): 2034-2060.
38. Wang, L., Chen, G., Li, H.: Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics, 2007, 23(12): 1486-1494.
39. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.
40. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 2010, 38(2): 894-942.
41. Zhang, T.: Analysis of multi-stage convex relaxation for sparse regularization. Journal of Machine Learning Research, 2010, 11(35): 1081-1107.
42. Zhang, X., Peng, D.: Solving constrained nonsmooth group sparse optimization via group Capped-ℓ_1 relaxation and group smoothing proximal gradient algorithm. Computational Optimimization and Applications, 2022, 83(3): 801-844.
43. Zhang, X., Peng, D., Su, Y.: A singular value shrinkage thresholding algorithm for folded concave penalized low-rank matrix optimization problems. Journal of Global Optimization, 2023, https://doi.org/10.1007/s10898-023-01322-8.
44. Zhang, Y., Zhang, N., Sun, D.: An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems. Mathematical Programming, 2020, 179(1): 223-263.
45. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics, 2009, 37(6A): 3468-3497.
46. Zhou, Y., Han, J., Yuan, X.: Inverse sparse group Lasso model for robust object tracking. IEEE Transactions on Multimedia, 2017, 19(8): 1798-1810.