Group Sparse Optimization for Images Recovery Using Capped Folded Concave Functions *

Lili Pan† and Xiaojun Chen‡

Abstract. This paper considers the image recovery problem by taking group sparsity into account as the prior knowledge. This problem is formulated as a group sparse optimization over the intersection of a polyhedron and a possibly degenerate ellipsoid. It is a convexly constrained optimization problem with a group cardinality objective function. We use a capped folded concave function to approximate the group cardinality function and show that the solution set of the continuous approximation problem and the set of group sparse solutions are same. Moreover, we use a penalty method to replace the constraints in the approximation problem by adding a convex nonsmooth penalty function in the objective function. We show the existence of positive penalty parameters such that the solution sets of the unconstrained penalty problem and the group sparse problem are same. We propose a smoothing penalty algorithm and show that any accumulation point of the sequence generated by the algorithm is a directional stationary point of the continuous approximation problem. Numerical experiments for recovery of group sparse image are presented to illustrate the efficiency of the smoothing penalty algorithm with adaptive capped folded concave functions.

Key words. group sparse recovery, capped folded concave function, exact penalty, smoothing method

AMS subject classifications. 90C46, 90C26

1. Introduction.

In the past decades, sparsity has been emerging as one of significant properties of natural images and used successfully in image recovery. For example, the authors in [13] considered a high-resolution imaging problem of 3D point source image recovery from 2D data. While finding the location and fluxes of the point sources is a large-scale sparse 3D inverse problem, most entries of the recovered 3D variable are zeros. Some images are not sparse themselves but can be represented by sparse linear regression. The basic framework of image recovery is concerned with the recovery of an unknown vector x from an underdetermined system of linear equations $b = Ax + \eta \in \mathbb{R}^r$, where $A \in \mathbb{R}^{r\times n}$ is a measurement matrix and η is the noise term. The sparse image recovery problem is formulated as

$$\begin{array}{ll} \min & \|x\|_0 \\ \text{s.t.} & \|Ax-b\|_2 \le \sigma, \quad Bx \le h, \end{array} \tag{1.1}$$

where $\|x\|_0$ counts the number of nonzero entries of x. The constraint $Bx \le h$ with $B \in \mathbb{R}^{q\times n}$ and $h \in \mathbb{R}^q$ describes some prior knowledge on the true image such as nonnegativity. The positive constant σ is a tolerance for the noise.

Sparse solutions of systems of linear equations for sparse modeling of images have attracted growing interests from theoretical and algorithmic aspects [5, 8, 10, 13, 15, 21, 20]. Since the

*Submitted to the editors DATE.

Funding: The work was supported by Hong Kong Research Grant Council PolyU 153001/18P, the National Natural Science Foundation of China (11771255, 11801325) and Young Innovation Teams of Shandong Province (2019KJI013).

†Department of Applied Mathematics, Shandong University of Technology. panlili1979@163.com

‡Department of Applied Mathematics, The Hong Kong Polytechnic University. xiaojun.chen@polyu.edu.hk

cardinality objective function $\|x\|_0$ is discontinuous, many convex or nonconvex continuous relaxation functions were presented. Compared with convex functions, nonconvex functions would promote better sparse solutions [6, 26], such as SCAD penalty [20], continuous exact L_0 (CEL0) penalty [42], capped L_1 function, [50] and capped L_p function ($0 < p \leq 1$) [38].

It is worth noting that the sparsity of solutions of (1.1) is not structured. However, in many scenarios, the nonzero components of solutions tend to cluster in groups. From the aspect of sparse image recovery, nonzero pixels may only appear in certain regions. Our goal is to explicitly take this predefined group sparse structure into account.

Let variable $\boldsymbol{x}$ be partitioned into m disjoint groups as $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_m^\top)^\top$ with $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$, $i = 1, \cdots, m$ and $\sum_{i=1}^m n_i = n$. The group sparse recovery problem can be formulated as the following group sparse optimization:

$$(P_0) \qquad \begin{aligned} \min \quad & \|\boldsymbol{x}\|_{2,0} \\ \text{s.t.} \quad & \|A\boldsymbol{x} - b\|_2 \leq \sigma, \quad B\boldsymbol{x} \leq h, \end{aligned}$$

where $\|\boldsymbol{x}\|_{2,0} = \#\{i | \, \|\boldsymbol{x}_i\|_2 \neq 0, i = 1, \ldots, m\}$ is the group cardinality function that counts the number of nonzero groups of $\boldsymbol{x}$.

If $n_i = 1$, $i = 1, \cdots, m$, problem (P_0) reduces to (1.1). If $n_i = n_1 < n$, $i = 2, \cdots, m$, then every $\boldsymbol{x}_i^\top$ can be regarded as the ith row of matrix $X \in \mathbb{R}^{m \times n_1}$, denoted as X_i. Problem (P_0) reduces to a row sparse optimization:

$$(1.2) \qquad \begin{aligned} \min \quad & \|X\|_{2,0} \\ \text{s.t.} \quad & \|\mathcal{A}(X) - b\|_2 \leq \sigma, \quad \mathcal{B}(X) \leq h, \end{aligned}$$

where $\|X\|_{2,0}$ is the row cardinality function that counts the number of nonzero rows of X, $\mathcal{A} : \mathbb{R}^{m \times n_1} \to \mathbb{R}^r$ and $\mathcal{B} : \mathbb{R}^{m \times n_1} \to \mathbb{R}^q$ are defined by the trace $\langle \cdot, \cdot \rangle$ of a product of two matrices as $\mathcal{A}(X) = (\langle A_1, X \rangle, \cdots, \langle A_r, X \rangle)^\top$, $\mathcal{B}(X) = (\langle B_1, X \rangle, \cdots, \langle B_q, X \rangle)^\top$. Problem (1.2) is known as a row selection problem, multiple measurement vector problem and simultaneous sparse approximation in various areas [14, 19, 31], and has numerous applications in reconstruction of biomedical signals [37] and joint covariate selection [40].

There are several compelling reasons to consider the group sparsity. In many applications, such as neuroimaging [24], gene expression data [39], bioinformatics [51], the group structure is an important piece of a prior knowledge about the problem. The use of group structure can improve the interpretability of the signals. Moreover, the group sparse optimizations allow to significantly reduce the number of required measurements for perfect recovery in the noiseless case and can be more stable in the presence of noise [28].

Group sparse problems have been extensively studied in the last few decades, see [2, 25, 27, 28, 29, 30, 49]. Many literatures use group $L_{2,1}$ penalty that yields group Lasso model [34, 36, 39, 49]. For example, the group $L_{2,1}$ sparse optimization model in [23] for spherical harmonic representations. In [46], an accelerated proximal method was proposed to solve a regularized $L_{2,1}$ group sparse optimization problem. On the other hand, due to the good performance of nonconvex relaxation for sparse optimizations, some group nonconvex penalties are presented, such as group SCAD [9, 32], group MCP [9, 32] and group $L_{p,q}$ ($0 \leq q \leq 1 \leq p$) [25]. In [3], the results in standard sparse optimization were extended to group sparse

optimization via group cardinality function, which minimizes a continuously differentiable objective function over a set composed of blocks corresponding to the group partition.

In this paper, we consider the following capped folded concave group sparse optimization problem to solve (P_0):

$$(R_\nu) \qquad \begin{aligned} \min \quad & \sum_{i=1}^{m} \phi(\|\boldsymbol{x}_i\|_2) \\ \text{s.t.} \quad & \|A\boldsymbol{x} - b\|_2 \le \sigma, \quad B\boldsymbol{x} \le h, \end{aligned}$$

where function $\phi(\cdot)$: $\mathbb{R}_+ \to \mathbb{R}_+$ is a capped folded concave function that satisfies the following two conditions with a fixed parameter $\nu > 0$:

(i) ϕ is continuous, increasing and concave in $[0, \infty)$ with $\phi(0) = 0$;

(ii) there is a $\nu > 0$ such that ϕ is differentiable in $(0, \nu)$, $\phi'_-(\nu) := \lim_{t\uparrow\nu} \phi'(t) > 0$ and $\phi(t) = 1$ for $t \in [\nu, \infty)$.

Most capped folded concave functions satisfy these two conditions. We list four of them as follows.

(1) Capped L_1: $\phi^{\mathrm{CapL1}}(t) = \min\{1, \frac{t}{\nu}\}$;

(2) Capped L_p: $\phi^{\mathrm{CapLp}}(t) = \min\{1, \frac{t^p}{\nu^p}\}, \quad 0 < p < 1$;

(3) Capped Fraction: $\phi^{\mathrm{CapF}}(t) = \min\{1, \frac{(1+\alpha\nu)t}{\nu(1+\alpha t)}\}$;

(4) Capped Minimax Concave Penalty (MCP):

$$\phi^{\mathrm{C-MCP}}(t) = \min\{1, \frac{2\alpha}{\nu(2\alpha - \nu)}\phi^{\mathrm{MCP}}(t)\}, \quad 0 < \nu < \alpha,$$

$$\text{with } \phi^{\mathrm{MCP}}(t) = \begin{cases} t - \frac{t^2}{2\alpha}, & 0 \le t \le \alpha, \\ \frac{\alpha}{2}, & t > \alpha. \end{cases}$$

To solve (R_ν), we replace its constraints by adding a convex nonsmooth penalty function in its objective function as the following

$$(P_\nu) \qquad \min \ \sum_{i=1}^{m} \phi(\|\boldsymbol{x}_i\|_2) + \lambda\big((\|A\boldsymbol{x} - b\|_2^2 - \sigma^2)_+ + \|(B\boldsymbol{x} - h)_+\|_1\big),$$

where $\lambda > 0$ and $z_+ \in \mathbb{R}^q$ with $(z_+)_i := \max\{0, z_i\}$. Moreover, to study the relation between (P_ν) and (P_0), we consider the corresponding penalty problem

$$\min \ \|\boldsymbol{x}\|_{2,0} + \lambda\big((\|A\boldsymbol{x} - b\|_2^2 - \sigma^2)_+ + \|(B\boldsymbol{x} - h)_+\|_1\big). \tag{1.3}$$

In the following discussion, for simplicity, we denote

$$\Phi(\boldsymbol{x}) = \sum_{i=1}^{m} \phi(\|\boldsymbol{x}_i\|_2)$$

$$F(\boldsymbol{x}) = (\|A\boldsymbol{x} - b\|_2^2 - \sigma^2)_+ + \|(B\boldsymbol{x} - h)_+\|_1, \qquad \Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \|A\boldsymbol{x} - b\|_2 \le \sigma, B\boldsymbol{x} \le h\}.$$

In this paper, we have the following assumption for our theorems.

Assumption We assume matrix A has full row rank, $\mathbf{0} \notin \Omega$ and there exists $\boldsymbol{x}_0 \in \Omega$ such that $\|A\boldsymbol{x}_0 - b\|_2 < \sigma$.

From this Assumption, the optimal value of problem (P_0) is a positive integer, which is denoted as k in this paper.

Using nonconvex continuous functions to approximate the discontinuous function $\|\boldsymbol{x}\|_0$ in optimization problems has been studied in [7, 11, 21, 22, 35, 42, 43, 44]. Some equivalence results of minimal L_0 and L_p norm solutions of linear equalities and inequalities for sufficiently small p have been proved [22]. The relations between minimizers of the L_0 regularized least squares minimization problem and its exact continuous relaxation have been investigated in [42, 43]. However, the relationship between problems (P_0) and (P_ν) regarding optimal solutions are unknown in group structure. Compared with component sparse optimizations, the variables in one group are not separable and the norm $\|\boldsymbol{x}\|_2$ is not differentiable at $\boldsymbol{x} = 0$. Moreover, the functions $\Phi(\cdot)$ and $F(\cdot)$ in (P_ν) are not differentiable and $F(\cdot)$ is not globally Lipschitz continuous in $\mathbb{R}^n$. Group sparse optimization contains component sparse optimization as a special case with more challenges.

Our contributions can be summarized as following:

- We establish the equivalence between problem (P_0) and its capped folded concave relaxation (R_ν) regarding global minimizers.
- We give a lower bound for nonzero group $\|\boldsymbol{x}_i\|$ of directional stationary points of (P_ν) by using directional derivatives. Furthermore, the lower bound is used to establish the relationship between problems (P_0) and (P_ν) regarding global minimizers. These results on relationships between (P_0), (R_ν) and (P_ν) are summarized in Figure 1.
- We propose a smoothing penalty algorithm to solve (R_ν) and show any accumulation point generated by the algorithm is a directional stationary point of (R_ν). It is known that directional stationary points are sharper than lifted stationary points, critical points and C-stationary points for the local optimality [1].

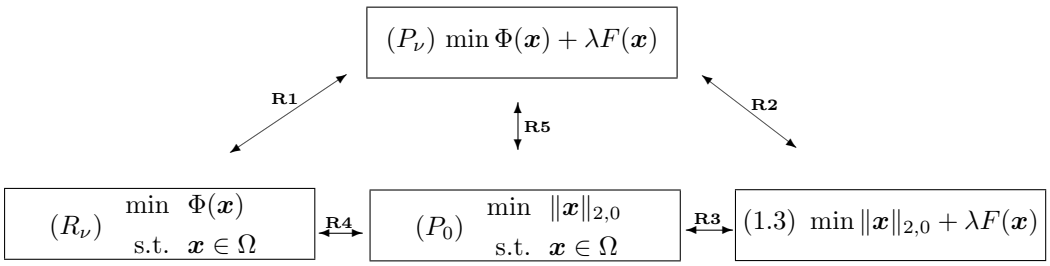

Figure 1. *The relationships of global minimizers between problems (P_0), (R_ν), (1.3) and (P_ν).*

`R1` [Theorem 3.6], `R2` [Theorem 3.4], `R3` [Theorem 3.7], `R4` [Theorem 2.1], `R5` [Theorem 3.3].

Notation. For a vector $x \in \mathbb{R}^n$, we denote L_2 norm by $\|x\|$, L_1 norm by $\|x\|_1$ and L_0 norm by $\|x\|_0 = \sum_{i=1}^n |x_i|^0$ where $|x_i|^0 = \begin{cases} 1, & x_i \neq 0, \\ 0, & x_i = 0. \end{cases}$ Letters in bold font denote that they are partitioned in the same way as $\boldsymbol{x}$. The group support set of $\boldsymbol{x}$ is denoted by

$$\Gamma(\boldsymbol{x}) = \{i \,|\, \|\boldsymbol{x}_i\| \neq 0, i = 1, \ldots, m\} = \Gamma_1(\boldsymbol{x}) \cup \Gamma_2(\boldsymbol{x}),$$

$$\Gamma_1(\boldsymbol{x}) = \{i \,|\, \|\boldsymbol{x}_i\| < \nu,\, i \in \Gamma(\boldsymbol{x})\} \quad \text{and} \quad \Gamma_2(\boldsymbol{x}) = \{i \,|\, \|\boldsymbol{x}_i\| \geq \nu,\, i \in \Gamma(\boldsymbol{x})\}.$$

For a fixed subset $\Gamma \subset \{1, \ldots, m\}$, let $\boldsymbol{x}_\Gamma$ be an n-dimensional vector with $(\boldsymbol{x}_\Gamma)_i = 0$, for $i \notin \Gamma$ and $(\boldsymbol{x}_\Gamma)_i = \boldsymbol{x}_i$ for $i \in \Gamma$. Let $(B^j)^\top$ $(j = 1, \cdots, q)$ be the jth row of matrix B. The distance from $\boldsymbol{x}$ to a closed set $\mathbb{S} \subseteq \mathbb{R}^n$ is defined by $\mathrm{dist}(\boldsymbol{x}, \mathbb{S}) = \inf\{\|\boldsymbol{x} - \boldsymbol{y}\| : \boldsymbol{y} \in \mathbb{S}\}$.

The paper is organized as follows. The link between problems (P_0) and (R_ν) and link between problems (P_0) and (P_ν) are studied in Section 2 and Section 3, respectively. The smoothing penalty algorithm using adaptive capped folded concave functions for (R_ν) is proposed in Section 4. In Section 5, numerical performance of the smoothing penalty algorithm is illustrated through randomly generated examples and image recovery examples. Section 6 gives conclusions.

2. Link between problems (P_0) and (R_ν).

From the conditions of $\phi(\cdot)$, we see that

$$|t|^0 - \phi(|t|) > 0 \text{ if } |t| \in (0, \nu), \quad \phi(|t|) = |t|^0 \text{ if } |t| \in \{0\} \cup [\nu, \infty) \text{ and } \int_0^\infty |t|^0 - \phi(|t|) \in (0, \nu].$$

In this section, we show that there is $\widehat{\nu} > 0$, such that problems (P_0) and (R_ν) have the same global optimal solutions for any $\nu \in (0, \widehat{\nu})$.

To show the existence of $\widehat{\nu}$, we define a positive constant $\overline{\nu}$ based on the global minimum of problem (P_0). For an integer s with $0 \leq s \leq m$, denote $Q_s := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_{2,0} \leq s\}$ and $\mathrm{dist}(\Omega, Q_s) := \inf\{\mathrm{dist}(\boldsymbol{x}, Q_s) : \boldsymbol{x} \in \Omega\}$. Recall that the global minimum of (P_0) is a positive integer k. Then the feasible set Ω of (P_0) does not have a vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_{2,0} < k$, which means $\mathrm{dist}(\Omega, Q_{k-K}) > 0$ for all $K = 1, \cdots, k$. Define

$$\overline{\nu} = \min\left\{\frac{1}{K}\mathrm{dist}(\Omega, Q_{k-K}) : K = 1, \cdots, k\right\}. \tag{2.1}$$

In the following, we show the equivalence of global optimality of problems (P_0) and (R_ν).

Theorem 2.1. *For any capped folded concave function ϕ satisfying $0 < \nu < \overline{\nu}$ and $\phi^{\mathrm{CapL1}}(t) \leq \phi(t) < |t|^0$ for $t \in (0, \nu)$, problems (P_0) and (R_ν) have same global minimizers and same optimal value.*

Proof. (i) Let $\boldsymbol{x}^* \in \mathbb{R}^n$ with $\|\boldsymbol{x}^*\|_{2,0} = k$ be a global minimizer of (P_0). We prove $\boldsymbol{x}^*$ is also a global minimizer of (R_ν) for any $0 < \nu < \overline{\nu}$. Since the global optimality of (P_0) yields $\|\boldsymbol{x}\|_{2,0} \geq k$ for $\boldsymbol{x} \in \Omega$, we show the conclusion by two cases.

Case 1. $\|\boldsymbol{x}\|_{2,0} = k$. It is easy to see that for any $i \in \Gamma(\boldsymbol{x})$,

$$\|\boldsymbol{x}_i\| \geq \min\{\|\boldsymbol{x}_j\| > 0 : j = 1, \cdots, m\} = \mathrm{dist}(\boldsymbol{x}, Q_{k-1}) \geq \mathrm{dist}(\Omega, Q_{k-1}) \geq \overline{\nu},$$

where the last inequality comes from (2.1). By $0 < \nu < \overline{\nu}$, we obtain $\|\boldsymbol{x}_i\| > \nu$ for all $i \in \Gamma(\boldsymbol{x})$, which means that $\Phi(\boldsymbol{x}) = k = \Phi(\boldsymbol{x}^*)$.

Case 2. $\|\boldsymbol{x}\|_{2,0} = r > k$. If $|\Gamma_2(\boldsymbol{x})| = r' \geq k$, from $\phi(t) > 0$ for $t > 0$ and $r > k$, we have $\Phi(\boldsymbol{x}) > k$. Now assume $r' < k$, without loss of generality, assume $\|\boldsymbol{x}_1\|, \cdots, \|\boldsymbol{x}_{r-r'}\| \in (0, \nu)$. Since $r' < k$, we know from (2.1) that $\frac{1}{k-r'}\mathrm{dist}(\Omega, Q_{r'}) \geq \overline{\nu}$. Together with

$$\|\boldsymbol{x}_1\| + \cdots + \|\boldsymbol{x}_{r-r'}\| \geq \sqrt{\|\boldsymbol{x}_1\|^2 + \cdots + \|\boldsymbol{x}_{r-r'}\|^2} \geq \mathrm{dist}(\boldsymbol{x}, Q_{r'}) \geq \mathrm{dist}(\Omega, Q_{r'}),$$

we get

$$\begin{aligned}\Phi(\boldsymbol{x}) &= \phi(\|\boldsymbol{x}_1\|) + \cdots + \phi(\|\boldsymbol{x}_{r-r'}\|) + \cdots + \phi(\|\boldsymbol{x}_r\|) \\ &\geq \phi^{\mathrm{CapL1}}(\|\boldsymbol{x}_1\|) + \cdots + \phi^{\mathrm{CapL1}}(\|\boldsymbol{x}_{r-r'}\|) + r' \\ &= \frac{1}{\nu}(\|\boldsymbol{x}_1\| \cdots + \|\boldsymbol{x}_{r-r'}\|) + r' \\ &\geq \frac{1}{\nu}\mathrm{dist}(\Omega, Q_{r'}) + r' \geq \frac{1}{\nu}(k-r')\overline{\nu} + r' \\ &> \frac{1}{\overline{\nu}}(k-r')\overline{\nu} + r' = k,\end{aligned} \tag{2.2}$$

where the first inequality comes from $\phi(t) \geq \phi^{\mathrm{CapL1}}(t)$ for $t \in (0, \nu)$ and the last inequality comes from $0 < \nu < \overline{\nu}$. The above two cases imply that $\Phi(\boldsymbol{x}) \geq k = \Phi(\boldsymbol{x}^*)$ for all $\boldsymbol{x} \in \Omega$. Hence $\boldsymbol{x}^*$ is also a global minimizer of (R_ν). Moreover, we have $\|\boldsymbol{x}^*\|_0 = \Phi(\boldsymbol{x}^*)$ for each global minimizer $\boldsymbol{x}^*$ of (R_ν).

(ii) Let $\boldsymbol{x}^*$ be a global minimizer of (R_ν) with $0 < \nu < \overline{\nu}$. Assume on the contrary $\boldsymbol{x}^*$ is not a solution of (P_0). Let $\widetilde{\boldsymbol{x}}$ be a global minimizer of (P_0), that is, $\|\widetilde{\boldsymbol{x}}\|_{2,0} = k$. By $\phi^{\mathrm{CapL1}}(t) \leq \phi(t) \leq |t|^0$, we have $\Phi(\widetilde{\boldsymbol{x}}) \leq \|\widetilde{\boldsymbol{x}}\|_{2,0}$. Using similar ways in the proof for Case 2 above, we will obtain $\Phi(\boldsymbol{x}^*) > k = \|\widetilde{\boldsymbol{x}}\|_{2,0} \geq \Phi(\widetilde{\boldsymbol{x}})$ for any $0 < \nu < \overline{\nu}$. This contradicts the global optimality of $\boldsymbol{x}^*$ for (R_ν). Hence $\boldsymbol{x}^*$ is a global minimizer of (P_0).

Therefore, whenever $0 < \nu < \overline{\nu}$, (P_0) and (R_ν) have the same global minimizers and optimal values. ∎

Remark 2.2. We can easily see that the four capped folded concave penalty functions given in Section 1 satisfy the conditions of Theorem 2.1. Hence problems (P_0) and (R_ν) with any one of the four functions have same global minimizers and same optimal values, whenever $0 < \nu < \overline{\nu}$.

For simplicity, in the theoretical results of this paper, we will use $\nu \in (0, \overline{\nu})$ such that problems (P_0) and (R_ν) have same global minimizers and same optimal value.

3. Links between problems (P_0), (1.3) and (P_ν).

We first characterize the d(directional)-stationary point of (P_ν), which can be used to study the relationship of global optimal solutions of problems (P_0), (1.3) and (P_ν).

3.1. d-stationary point of (P_ν).

Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous and directionally differentiable at point $x \in \mathbb{R}^n$. The directional derivative of f along a vector $w \in \mathbb{R}^n$ at x is defined by

$$f'(x; w) := \lim_{\tau \downarrow 0} \frac{f(x + \tau w) - f(x)}{\tau}.$$

If f is differentiable at x, then $f'(x; w) = \langle \nabla f(x), w \rangle$. Next, we consider the directional derivative of L_2 norm and $F(\boldsymbol{x})$. Denote L_2 norm as $l(x) := \|x\|$. By simple computation, we have

$$l'(x^*; x - x^*) = \begin{cases} \|x\|, & \|x^*\| = 0, \\ \frac{\langle x^*, x - x^* \rangle}{\|x^*\|}, & \text{otherwise.} \end{cases} \tag{3.1}$$

Then by [41, Exercise 8.31], the directional derivative of F at $\boldsymbol{x}^*$ has the form

$$F'(\boldsymbol{x}^*;\boldsymbol{x}-\boldsymbol{x}^*)=\Delta^a+\Delta^b, \tag{3.2}$$

where

$$\Delta^a:=\begin{cases}0, & \text{if } \|A\boldsymbol{x}^*-b\|^2<\sigma^2,\\ \max\{0,\langle 2A^\top(A\boldsymbol{x}^*-b),\boldsymbol{x}-\boldsymbol{x}^*\rangle\}, & \text{if } \|A\boldsymbol{x}^*-b\|^2=\sigma^2,\\ \langle 2A^\top(A\boldsymbol{x}^*-b),\boldsymbol{x}-\boldsymbol{x}^*\rangle, & \text{otherwise}\end{cases}$$

and $\Delta^b=\sum_{j=1}^q\Delta_j^b$ with

$$\Delta_j^b:=\begin{cases}0, & \text{if } \langle B^j,\boldsymbol{x}^*\rangle<h_j,\\ \max\{0,\langle B^j,\boldsymbol{x}-\boldsymbol{x}^*\rangle\}, & \text{if } \langle B^j,\boldsymbol{x}^*\rangle=h_j,\\ \langle B^j,\boldsymbol{x}-\boldsymbol{x}^*\rangle, & \text{otherwise.}\end{cases}$$

Definition 3.1. *We say that $\boldsymbol{x}^*\in\mathbb{R}^n$ is a d-stationary point of (P_ν) if*

$$\Phi'(\boldsymbol{x}^*;\boldsymbol{x}-\boldsymbol{x}^*)+\lambda F'(\boldsymbol{x}^*;\boldsymbol{x}-\boldsymbol{x}^*)\geq 0,\quad \forall \boldsymbol{x}\in\mathbb{R}^n.$$

From (3.2), $\boldsymbol{x}^*\in\mathbb{R}^n$ is a d-stationary point of (P_ν) if

$$\Phi'(\boldsymbol{x}^*;\boldsymbol{x}-\boldsymbol{x}^*)+\lambda\Delta^a+\lambda\Delta^b\geq 0,\quad \forall \boldsymbol{x}\in\mathbb{R}^n. \tag{3.3}$$

It is known that [16, Lemma 2.1] there exists a $\beta>0$ such that for all $\boldsymbol{x}\in\mathbb{R}^n$, we have

$$\operatorname{dist}(\boldsymbol{x},\Omega)\leq\beta[(\|A\boldsymbol{x}-b\|_2^2-\sigma^2)_+ +\|(B\boldsymbol{x}-h)_+\|_1]=\beta F(\boldsymbol{x}). \tag{3.4}$$

Let $L:\mathbb{R}^n\to\mathbb{R}$ be defined as

$$L(\boldsymbol{x})=2\|A\|_F\|A\boldsymbol{x}-b\|+\sqrt{q}\|B\|_F.$$

Since A has nonzero entries and $\phi'_-(\nu)\to\infty$ as $\nu\to 0$, for any $\Upsilon>\sigma$ and $\lambda>0$, there are $\hat{\boldsymbol{x}}\in\mathbb{R}^n$ and a sufficiently small $\nu>0$ such that $\|A\hat{\boldsymbol{x}}-b\|\geq\Upsilon$ and $\phi'_-(\nu)>\lambda L(\hat{\boldsymbol{x}})$. In the rest of this paper, we choose Υ, ν, λ and $\hat{\boldsymbol{x}}\in\mathbb{R}^n$ satisfying

$$\|A\hat{\boldsymbol{x}}-b\|\geq\Upsilon,\quad \lambda>\frac{\beta}{\bar{\nu}}\quad\text{and}\quad \phi'_-(\nu)>\lambda L(\hat{\boldsymbol{x}}). \tag{3.5}$$

At the end of this section, we show how to choose these parameters by Example 3.8.

Lemma 3.2. *Let $\boldsymbol{x}^*\in\mathbb{R}^n$ be a d-stationary point of (P_ν) satisfying $\|A\boldsymbol{x}^*-b\|\leq\Upsilon$ and $\phi'_-(\nu)>\lambda L(\hat{\boldsymbol{x}})$, then either $\|\boldsymbol{x}_i^*\|\geq\nu$ or $\|\boldsymbol{x}_i^*\|=0$ for $i=1,\cdots,m$.*

Proof. To prove this Lemma, we only need to show $\Gamma_1(\boldsymbol{x}^*)=\emptyset$. Assume on contradiction that $\Gamma_1(\boldsymbol{x}^*)\neq\emptyset$. From (3.3), we have the following inequality for any $\boldsymbol{x}\in\mathbb{R}^n$ satisfying $\boldsymbol{x}_j=\boldsymbol{x}_j^*$ for all $j\notin\Gamma_1(\boldsymbol{x}^*)$ and for which $\exists\, i\in\Gamma_1(\boldsymbol{x}^*)$ such that $\boldsymbol{x}_i\neq\boldsymbol{x}_i^*$,

$$\sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\phi'(\|\boldsymbol{x}_i^*\|)\frac{\langle\boldsymbol{x}_i^*,\boldsymbol{x}_i-\boldsymbol{x}_i^*\rangle}{\|\boldsymbol{x}_i^*\|}+\lambda\Delta^a+\lambda\Delta^b\geq 0. \tag{3.6}$$

Notice that $\Delta^a \leq \sum_{i\in\Gamma_1(\boldsymbol{x}^*)} \|2A_i^\top(A\boldsymbol{x}^*-b)\|\|\boldsymbol{x}_i-\boldsymbol{x}_i^*\|$ and $\Delta^b \leq \sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\sum_{j=1}^q \|B^j\|\|\boldsymbol{x}_i-\boldsymbol{x}_i^*\|$. By (3.6) and letting $\boldsymbol{x}_i=\boldsymbol{x}_i^*-\epsilon\boldsymbol{x}_i^*$ $(\epsilon>0)$, $i\in\Gamma_1(\boldsymbol{x}^*)$, we have

$$\begin{aligned}\sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\phi'(\|\boldsymbol{x}_i^*\|)\|\boldsymbol{x}_i^*\| \leq \lambda \sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\Big(2\|A_i^\top(A\boldsymbol{x}^*-b)\|+\sum_{j=1}^q\|B^j\|\Big)\|\boldsymbol{x}_i^*\| \\ \leq \lambda\sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\Big(2\|A_i^\top(A\boldsymbol{x}^*-b)\|+\sqrt{q}\|B\|\Big)\|\boldsymbol{x}_i^*\| \leq \lambda L(\hat{\boldsymbol{x}})\sum_{i\in\Gamma_1(\boldsymbol{x}^*)}\|\boldsymbol{x}_i^*\|.\end{aligned}$$

By $\phi'(\|\boldsymbol{x}_i^*\|)\geq\phi'_-(\nu)$ for all $i\in\Gamma_1(\boldsymbol{x}^*)$, we have $\phi'_-(\nu)\leq\lambda L(\hat{\boldsymbol{x}})$. This contradicts the condition of $\phi'_-(\nu)>\lambda L(\hat{\boldsymbol{x}})$. The proof is completed. ∎

3.2. Link between problems (P_0) and (P_ν). Utilizing Lemma 3.2, we obtain the following relationship between the global optimality of problems (P_0) and (P_ν). We recall the choice of λ in (3.5) for the following theorem.

Theorem 3.3. *Let $\lambda>\frac{\beta}{\overline{\nu}}$ and $\phi'_-(\nu)>\lambda L(\hat{\boldsymbol{x}})$.*

(i) If $\boldsymbol{x}^*\in\mathbb{R}^n$ is a global minimizer of (P_ν) with $\|A\boldsymbol{x}^*-b\|\leq\Upsilon$, then $\boldsymbol{x}^*$ is a global minimizer of (P_0).

(ii) If $\boldsymbol{x}^*\in\mathbb{R}^n$ is a global minimizer of (P_0) and (P_ν) has a global minimizer $\widetilde{\boldsymbol{x}}$ with $\|A\widetilde{\boldsymbol{x}}-b\|\leq\Upsilon$, then $\boldsymbol{x}^*$ is a global minimizer of (P_ν).

Proof. (i) Since $\boldsymbol{x}^*\in\mathbb{R}^n$ is a global minimizer of (P_ν) and the objective function is locally Lipschitz continuous, $\boldsymbol{x}^*$ is a d-stationary point of (P_ν). From $\|A\boldsymbol{x}^*-b\|\leq\|A\hat{\boldsymbol{x}}-b\|$ and Lemma 3.2, $\Phi(\boldsymbol{x}^*)=\|\boldsymbol{x}^*\|_{2,0}$. Assume now that $\boldsymbol{x}^*$ is not a global minimizer of (P_0) and $\boldsymbol{x}'$ with $\|\boldsymbol{x}'\|_{2,0}=k$ is a global minimizer of (P_0).

Then, we distinguish two cases.

- $\boldsymbol{x}^*\in\Omega$. Then $\|\boldsymbol{x}'\|_{2,0}<\|\boldsymbol{x}^*\|_{2,0}$ by the assumption. From $F(\boldsymbol{x}^*)=0$, we have

$$\Phi(\boldsymbol{x}')+\lambda F(\boldsymbol{x}')\leq\|\boldsymbol{x}'\|_{2,0}+\lambda F(\boldsymbol{x}')=\|\boldsymbol{x}'\|_{2,0}<\|\boldsymbol{x}^*\|_{2,0}\leq\Phi(\boldsymbol{x}^*)+\lambda F(\boldsymbol{x}^*),$$

  which contradicts the global optimality of $\boldsymbol{x}^*$ for (P_ν).
- $\boldsymbol{x}^*\notin\Omega$. Then we consider two cases with $F(\boldsymbol{x}^*)>0$.
  - If $\|\boldsymbol{x}^*\|_{2,0}\geq k$, it holds that

$$\Phi(\boldsymbol{x}')+\lambda F(\boldsymbol{x}')\leq\|\boldsymbol{x}'\|_{2,0}+\lambda F(\boldsymbol{x}')=k<\|\boldsymbol{x}^*\|_{2,0}+\lambda F(\boldsymbol{x}^*)=\Phi(\boldsymbol{x}^*)+\lambda F(\boldsymbol{x}^*)$$

    which contradicts the global optimality of $\boldsymbol{x}^*$ for (P_ν).
  - If $\|\boldsymbol{x}^*\|_{2,0}=k'<k$, then as $\boldsymbol{x}'\in\Omega$ we have $\|A\boldsymbol{x}'-b\|^2\leq\sigma^2<\Upsilon^2$ and Lemma 3.2 implies $\|\boldsymbol{x}'\|_{2,0}=\Phi(\boldsymbol{x}')$. From the definition of $\overline{\nu}$, $k\geq k-k'\geq 1$ and (3.4)-(3.5), we have

$$\overline{\nu}\leq\frac{1}{k-k'}\mathrm{dist}(\Omega,Q_{k'})\leq\frac{1}{k-k'}\mathrm{dist}(\Omega,\boldsymbol{x}^*)\leq\frac{\beta F(\boldsymbol{x}^*)}{k-k'}.$$

This together with $\lambda > \frac{\beta}{\overline{\nu}}$ gives

$$\begin{aligned}\Phi(\boldsymbol{x}^*) + \lambda F(\boldsymbol{x}^*) &= k' + \lambda F(\boldsymbol{x}^*)\\ &> k' + \frac{\beta}{\overline{\nu}} F(\boldsymbol{x}^*) \geq k' + \frac{\beta}{\overline{\nu}} \frac{(k-k')\overline{\nu}}{\beta} = k = \|\boldsymbol{x}'\|_{2,0} + \lambda F(\boldsymbol{x}')\\ &\geq \Phi(\boldsymbol{x}') + \lambda F(\boldsymbol{x}'),\end{aligned}$$

which contradicts the global optimality of $\boldsymbol{x}^*$ for (P_ν).
This shows that $\boldsymbol{x}^*$ is a global minimizer of (P_0).

(ii) Suppose that $\boldsymbol{x}^*$ is a global minimizer of (P_0) but not a global minimizer of (P_ν). Since $\widetilde{\boldsymbol{x}}$ is a global minimizer of (P_ν) with $\|A\widetilde{\boldsymbol{x}} - b\| \leq \Upsilon$, from Lemma 3.2 and (i), we have $\Phi(\widetilde{\boldsymbol{x}}) = \|\widetilde{\boldsymbol{x}}\|_{2,0}$ and $\widetilde{\boldsymbol{x}} \in \Omega$. Using this, we conclude that

$$\|\widetilde{\boldsymbol{x}}\|_{2,0} \leq \|\widetilde{\boldsymbol{x}}\|_{2,0} + \lambda F(\widetilde{\boldsymbol{x}}) = \Phi(\widetilde{\boldsymbol{x}}) + \lambda F(\widetilde{\boldsymbol{x}}) < \Phi(\boldsymbol{x}^*) + \lambda F(\boldsymbol{x}^*) = \Phi(\boldsymbol{x}^*) \leq \|\boldsymbol{x}^*\|_{2,0},$$

which leads to a contradiction with the global optimality of $\boldsymbol{x}^*$ for (P_0). Hence $\boldsymbol{x}^*$ is a global minimizer of problem (P_ν) and the proof is completed. ∎

3.3. Link between problems (1.3) and (P_ν). In [7], the authors showed the relationship between (1.3) and (P_ν) regarding global minimizers in the case where $m = n$, Ω is a box feasible set, ϕ is the capped L_1 penalty function and F is a Lipschitz continuous function. Although the objective functions here are not globally Lipschitz continuous, by Lemma 3.2 and similar method as in [7], we can prove problems (1.3) and (P_ν) have the same optimal solutions if the parameters λ and ν satisfy (3.5). For completeness, we state it as following.

Theorem 3.4. *Suppose that $\phi'_-(\nu) > \lambda L(\hat{\boldsymbol{x}})$. If $\boldsymbol{x}^*$ is a global minimizer of problem (P_ν) and satisfies $\|A\boldsymbol{x}^* - b\| \leq \Upsilon$, then $\boldsymbol{x}^*$ is a global minimizer of problem (1.3). Conversely, if problem (P_ν) has a global minimizer $\widetilde{\boldsymbol{x}}$ satisfying $\|A\widetilde{\boldsymbol{x}} - b\| \leq \Upsilon$ and $\boldsymbol{x}^*$ is a global minimizer of problem (1.3), then $\boldsymbol{x}^*$ is a global minimizer of problem (P_ν).*

Proof. Firstly, if $\boldsymbol{x}^*$ is a global minimizer of problems (P_ν), by Lemma 3.2, we have $\Phi(\boldsymbol{x}^*) = \|\boldsymbol{x}^*\|_{2,0}$. Therefore, it holds that

$$\|\boldsymbol{x}^*\|_{2,0} + \lambda F(\boldsymbol{x}^*) = \Phi(\boldsymbol{x}^*) + \lambda F(\boldsymbol{x}^*) \leq \Phi(\boldsymbol{x}) + \lambda F(\boldsymbol{x}) \leq \|\boldsymbol{x}\|_{2,0} + \lambda F(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$

which means $\boldsymbol{x}^*$ is a global minimizer of problem (1.3).

Secondly, assume $\boldsymbol{x}^*$ is a global minimizer of problem (1.3). Then from from $\Phi(\boldsymbol{x}^*) \leq \|\boldsymbol{x}^*\|_{2,0}$, we have

$$\Phi(\boldsymbol{x}^*) + \lambda F(\boldsymbol{x}^*) \leq \|\boldsymbol{x}^*\|_{2,0} + \lambda F(\boldsymbol{x}^*) \leq \|\boldsymbol{x}\|_{2,0} + \lambda F(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

From Lemma 3.2 and the assumption of this theorem, the global minimizer $\widetilde{\boldsymbol{x}}$ of problem (P_ν) satisfies $\|\widetilde{\boldsymbol{x}}\|_{2,0} = \Phi(\widetilde{\boldsymbol{x}})$. Hence the above inequalities imply

$$\Phi(\widetilde{\boldsymbol{x}}) + \lambda F(\widetilde{\boldsymbol{x}}) \leq \Phi(\boldsymbol{x}^*) + \lambda F(\boldsymbol{x}^*) \leq \|\boldsymbol{x}^*\|_{2,0} + \lambda F(\boldsymbol{x}^*) \leq \|\widetilde{\boldsymbol{x}}\|_{2,0} + \lambda F(\widetilde{\boldsymbol{x}}) = \Phi(\widetilde{\boldsymbol{x}}) + \lambda F(\widetilde{\boldsymbol{x}}).$$

This shows $\boldsymbol{x}^*$ is a global minimizer of problem(P_ν). ∎

Remark 3.5. From Theorem 3.4, we can claim that problem (1.3) has a unique solution if and only if (P_ν) has a unique solution. Moreover, (1.3) and (P_ν) have the same unique solution if one of the two problems has a unique solution.

3.4. Link between (R_ν) and (P_ν) and link between (P_0) and (1.3). The authors of [16] showed the link between (R_ν) and (P_ν) with $m = n$ and the authors of [17] showed the link between (P_0) and (1.3) with $n_i = n_1$, $i = 1, \ldots, m$ regarding global minimizers under certain conditions, respectively. The following Theorem 3.6 and Theorem 3.7 extend these results in [16] and [17] to the group cases, respectively. The proof is similar, and we omit it.

Theorem 3.6. *Suppose that ϕ is globally Lipschitz continuous on $[0, \nu]$, then there exists a $\lambda^* > 0$ such that any global minimizer of (R_ν) is a global minimizer of (P_ν) whenever $\lambda \geq \lambda^*$; moreover, if $\boldsymbol{x}^*$ is a global minimizer of (P_ν) for some $\lambda > \lambda^*$, then $\boldsymbol{x}^*$ is a global minimizer of (R_ν).*

Theorem 3.7. *There exists a $\lambda^* > 0$ such that any global minimizer of (P_0) is a global minimizer of (1.3) whenever $\lambda \geq \lambda^*$; moreover, if $\boldsymbol{x}^*$ is a global minimizer of (1.3) for some $\lambda \geq \lambda^*$, then $\boldsymbol{x}^*$ is a global minimizer of (P_0).*

To end this section, we use the following example to illustrate the choice of these parameters in (3.5) for the links between problems (P_0), (R_ν), (1.3) and (P_ν).

Example 3.8 Let $A = \frac{1}{\sqrt{2}}[1, 1]$, $b = \frac{1}{\sqrt{2}}$, $\sigma \in (0, \frac{1}{\sqrt{2}})$ and

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^2 : \|A\boldsymbol{x} - b\|_2 \leq \sigma\} = \{\boldsymbol{x} \in \mathbb{R}^2 : \frac{1}{\sqrt{2}}|x_1 + x_2 - 1| \leq \sigma\}.$$

The feasible set Ω is from Example 1.1 in [18].

The solution set of problem (P_0) is

$$X^* := \{(0, t)^\top : t \in [1 - \sqrt{2}\sigma, 1 + \sqrt{2}\sigma]\} \cup \{(t, 0)^\top : t \in [1 - \sqrt{2}\sigma, 1 + \sqrt{2}\sigma]\}.$$

The solution set of problem (R_ν) with capped L_1 function ϕ is X^* for any $0 < \nu < \overline{\nu}$, where $\overline{\nu} = \text{dist}(0, \Omega) = \frac{1-\sqrt{2}\sigma}{\sqrt{2}}$.

To consider the link between problem (P_0) and the two penalty problems (P_ν) and (1.3), we need parameters β, Υ and λ. By [16, Lemma 2.2], we can set $\beta = \frac{1}{\sigma}\|A^\top(AA^\top)^{-1}\| = \frac{1}{\sigma}$ such that for any $\boldsymbol{x} \in \mathbb{R}^2$, $\text{dist}(\boldsymbol{x}, \Omega) \leq \beta F(\boldsymbol{x})$, where $F(\boldsymbol{x}) = [\|A\boldsymbol{x} - b\|^2 - \sigma^2]_+ = [\frac{1}{2}(x_1 + x_2 - 1)^2 - \sigma^2]_+$.

We choose $\hat{\boldsymbol{x}} = \boldsymbol{0}$. Then $\Upsilon = \|b\| > \sigma$. Moreover for $\lambda > \frac{\beta}{\bar{\nu}} = \frac{1}{\sigma\bar{\nu}}$ and $\nu < \frac{1}{\sqrt{2}\lambda} < \frac{\sigma\bar{\nu}}{\sqrt{2}} < \frac{\bar{\nu}}{2}$, we have $\lambda L(\boldsymbol{0}) = \lambda\sqrt{2} < \phi'_-(\nu) = \frac{1}{\nu}$. Hence all inequalities in (3.5) holds.

The solution set of problem (1.3) is X^* if $\lambda > \frac{\sqrt{2}}{\sigma(1-\sqrt{2}\sigma)}$.

The solution set of problem (P_ν) with capped L_1 function ϕ is X^* if $\lambda > \frac{\sqrt{2}}{\sigma(1-\sqrt{2}\sigma)}$ and $\nu < \frac{1}{\sqrt{2}\lambda}$.

Moreover, we can use other three capped functions in Section 1 with ν satisfying $\lambda L(\boldsymbol{0}) = \lambda\sqrt{2} < \phi'_-(\nu)$. From the concavity of $\phi(t)$ on $[0, \nu]$, we have $\phi'_-(\nu) \leq \frac{\phi(\nu)-\phi(0)}{\nu-0} = \frac{1}{\nu}$. Table 1 gives the value of $\phi'_-(\nu)$.

	capped L_1	capped L_p	capped MCP	capped fraction
$\phi'_-(\nu)$	$\frac{1}{\nu}$	$\frac{p}{\nu}$	$\frac{2(\alpha-\nu)}{\nu(2\alpha-\nu)}$	$\frac{1}{\nu(1+\alpha\nu)}$

Table 1

$\phi'_-(\nu) = \lim_{t\uparrow\nu} \phi'(t)$

274 **4. Algorithm.** We propose a smoothing penalty method to solve problem (R_ν). The
275 following smoothing function is used to approximate t_+:

276 $$h_\mu(t) := \max_{0\le s\le 1}\left\{ts - \frac{\mu}{2}s^2\right\} = \begin{cases} t - \frac{\mu}{2}, & t \ge \mu \\ \frac{t^2}{2\mu}, & 0 < t < \mu \\ 0, & t \le 0 \end{cases}$$

277 where $\mu > 0$ and $h'_\mu(t) = \min\left\{\max\left\{\frac{t}{\mu}, 0\right\}, 1\right\}$. The smoothing function of $F(\boldsymbol{x})$ is

278 $$F_\mu(\boldsymbol{x}) := h_\mu(\|A\boldsymbol{x} - b\|^2 - \sigma^2) + \sum_{i=1}^{q} h_\mu([B\boldsymbol{x} - h]_i).$$

279 Since $0 \le t_+ - h_\mu(t) \le \frac{\mu}{2}$, we have that for $\boldsymbol{x} \in \mathbb{R}^n$

280 $$0 \le F_\mu(\boldsymbol{x}) \le F(\boldsymbol{x}) \le F_\mu(\boldsymbol{x}) + \frac{q+1}{2}\mu.$$

281 The smoothing penalty method for solving problem (R_ν) is presented as follows.

Algorithm 4.1 Smoothing penalty algorithm for problem (R_ν)

Choose $\boldsymbol{x}^{feas} \in \Omega$, $\boldsymbol{x}^0 \in \mathbb{R}^n$, $\lambda_0 > 0$, $\mu_0 > 0$, $\epsilon_0 > 0$, $\rho > 1$, and $\theta \in (0, 1)$ arbitrarily. Set $k = 0$ and $\boldsymbol{x}^{0,0} = \boldsymbol{x}^0$.

(1) If $F_{\mu_k}(\boldsymbol{x}^{k,0}) > F_{\mu_k}(\boldsymbol{x}^{feas})$, set $\boldsymbol{x}^{k,0} = \boldsymbol{x}^{feas}$. Use $\boldsymbol{x}^{k,0}$ as an initial point to find an approximate solution $\boldsymbol{x}^k$ of $\min\{G_{\lambda_k,\mu_k}(\boldsymbol{x}) := \Phi(\boldsymbol{x}) + \lambda_k F_{\mu_k}(\boldsymbol{x})\}$ such that

$$\max\left\{0, -\min_{x\in\mathbb{R}^n}\{\Phi'(\boldsymbol{x}^k, \boldsymbol{x} - \boldsymbol{x}^k) + \lambda_k\langle\nabla F_{\mu_k}(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k\rangle\}\right\} \le \epsilon_k. \tag{4.1}$$

(2) Set $\lambda^{k+1} = \rho\lambda_k$, $\mu^{k+1} = \theta\mu_k$, $\epsilon^{k+1} = \theta\epsilon_k$, and $\boldsymbol{x}^{k+1,0} = \boldsymbol{x}^k$.

(3) Set $k \leftarrow k + 1$ and go to step (1).

end

282 Algorithm 4.1 is motivated by the smoothing penalty method for minimizing $\|x\|_p^p$ over
283 Ω where $0 < p < 1$ in [16]. In Step 1, $G_{\lambda_k,\mu_k}(\boldsymbol{x})$ is a smoothing approximation of the
284 objective function $\Phi(\boldsymbol{x}) + \lambda_k F(\boldsymbol{x})$ of problem (P_ν) with $\lambda = \lambda_k$. Compared the objective
285 function in [16] and the objective function in (P_ν), there are three differences. Firstly, the
286 function $\|x\|_p^p$ in [16] has bounded level sets, while our $\Phi(\cdot)$ does not. Secondly, $\|x\|_p^p$ in [16] is
287 differentiable on $\mathbb{R}$ except points involving zero components, while our $\Phi(\cdot)$ is not differentiable

at points that have components with values of zero or ν. Moreover, our $\Phi(\cdot)$ is a composite function of ϕ and L_2 norm, and L_2 norm is not differentiable at original point. At last, $\|x\|_p^p$ is not Lipschitz continuous at zero, while our $\phi(\cdot)$ is Lipschitz continuous and directionally differentiable. Hence, the stop condition (4.1) for solving the subproblem is different and it can generate d-stationary points of (R_ν). It is known that d-stationary points are sharper than critical points and C-stationary points for local optimality [1]. In view of these differences, for completeness we give the convergence analysis for Algorithm 4.1 with Capped L_1 function in the following. For simplicity, in the subsequent arguments, ϕ refers to ϕ^{CapL1}. The convergence of Algorithm 4.1 with other capped folded concave functions can be derived similarly.

We will use the directional derivative of Φ at $\boldsymbol{x}^*$, which is

$$\Phi'(\boldsymbol{x}^*, \boldsymbol{x}-\boldsymbol{x}^*) = \sum_{i=1}^{m} \phi'(\|\boldsymbol{x}_i^*\|, \boldsymbol{x}_i - \boldsymbol{x}_i^*),$$

$$\phi'(\|\boldsymbol{x}_i^*\|, \boldsymbol{x}_i - \boldsymbol{x}_i^*) = \begin{cases} \frac{\|\boldsymbol{x}_i\|}{\nu}, & \|\boldsymbol{x}_i^*\| = 0, \\ \frac{\langle \boldsymbol{x}_i^*, \boldsymbol{x}_i - \boldsymbol{x}_i^* \rangle}{\nu \|\boldsymbol{x}_i^*\|}, & \|\boldsymbol{x}_i^*\| \in (0, \nu), \\ \max\{0, \frac{\langle \boldsymbol{x}_i^*, \boldsymbol{x}_i - \boldsymbol{x}_i^* \rangle}{\nu \|\boldsymbol{x}_i^*\|}\}, & \|\boldsymbol{x}_i^*\| = \nu, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 4.1. *Let $\{\boldsymbol{x}^k\}$ be generated by Algorithm 4.1 with Capped L_1 function. Then any accumulation point $\boldsymbol{x}^*$ of $\{\boldsymbol{x}^k\}$ is a d-stationary point of (R_ν), that is,*

$$\boldsymbol{x}^* \in \Omega \quad \text{and} \quad \Phi(\boldsymbol{x}^*, \boldsymbol{x}-\boldsymbol{x}^*) \geq 0, \quad \forall \boldsymbol{x} \in \Omega.$$

Proof. Let $\{\boldsymbol{x}^k\}_{\mathcal{K}}$ be a subsequence of $\{\boldsymbol{x}^k\}$ such that $\boldsymbol{x}^k \to \boldsymbol{x}^*$ as $k \to \infty, k \in \mathcal{K}$.

First we prove $\boldsymbol{x}^*$ is a feasible point of problem (R_ν). From $G_{\lambda_k,\mu_k}(\boldsymbol{x}^k) \leq G_{\lambda_k,\mu_k}(\boldsymbol{x}^{feas})$, we have

$$\begin{aligned} &(\|A\boldsymbol{x}^k - b\|^2 - \sigma^2)_+ + \|(B\boldsymbol{x}^k - h)_+\|_1 \\ \leq & F_{\mu_k}(\boldsymbol{x}^k) + \frac{q+1}{2}\mu_k = \frac{1}{\lambda_k} G_{\lambda_k,\mu_k}(\boldsymbol{x}^k) + \frac{q+1}{2}\mu_k \\ \leq & \frac{1}{\lambda_k} G_{\lambda_k,\mu_k}(\boldsymbol{x}^{feas}) + \frac{q+1}{2}\mu_k = \frac{1}{\lambda_k}\Phi(\boldsymbol{x}^{feas}) + \frac{q+1}{2}\mu_k. \end{aligned}$$

Taking limits as $k \to \infty$, $k \in \mathcal{K}$, we have $(\|A\boldsymbol{x}^* - b\|^2 - \sigma^2)_+ + \|(B\boldsymbol{x}^* - h)_+\|_1 \leq 0$. Hence $\boldsymbol{x}^* \in \Omega$.

Now we prove that $\boldsymbol{x}^*$ a d-stationary point of problem (R_ν). Denote $I^* = \{i : (B\boldsymbol{x}^* - h)_i = 0\}$. Then $(B\boldsymbol{x}^* - h)_i < 0$ and $(B\boldsymbol{x}^k - h)_i < 0$ for sufficiently large k if $i \notin I^*$. Using this, we have $w_i^k := h'_{\mu_k}((B\boldsymbol{x}^k - h)_i) = 0$ for $i \notin I^*$, when k is sufficiently large.

From (4.1), we have

$$\Phi'(\boldsymbol{x}^k, \boldsymbol{x}-\boldsymbol{x}^k) + \lambda_k F'_{\mu_k}(\boldsymbol{x}^k, \boldsymbol{x}-\boldsymbol{x}^k) \geq -\epsilon_k, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{4.2}$$

Combining the expression of directional derivative, there exist $\zeta^k := \big((\zeta_1^k)^\top, \cdots, (\zeta_m^k)^\top\big)^\top$ with $\zeta_i^k \in \partial\phi(\|\boldsymbol{x}_i^k\|)$, $i = 1, \cdots, m$ such that

$$\big\langle \zeta^k + 2\lambda_k h'_{\mu_k}(\|A\boldsymbol{x}^k - b\|^2 - \sigma^2)A^\top(A\boldsymbol{x}^k - b) + \lambda_k \sum_{j\in I^*} w_j^k B^j, \boldsymbol{x} - \boldsymbol{x}^k\big\rangle \geq -\epsilon_k,\ \boldsymbol{x} \in \mathbb{R}^n. \tag{4.3}$$

The global Lipschitz continuity yields that $\{\|\zeta_i^k\|\}$, $i = 1, \cdots, m$ are bounded. Then let $\{\zeta^k\}_{\mathcal{K}} \to \zeta^* = \big((\zeta_1^*)^\top, \cdots, (\zeta_m^*)^\top\big)^\top \in \partial\Phi(\|\boldsymbol{x}^*\|)$.

We consider two case: $\|A\boldsymbol{x}^* - b\|^2 < \sigma^2$ and $\|A\boldsymbol{x}^* - b\|^2 = \sigma^2$.

Case 1. Suppose that $\|A\boldsymbol{x}^* - b\|^2 < \sigma^2$. Then for sufficiently large k, we have $\|A\boldsymbol{x}^k - b\|^2 < \sigma^2$ and $h'_{\mu_k}(\|A\boldsymbol{x}^k - b\|^2 - \sigma^2) = 0$. Hence, (4.3) reduces to

$$\big\langle \zeta^k + \lambda_k \sum_{j\in I^*} w_j^k B^j, \boldsymbol{x} - \boldsymbol{x}^k\big\rangle \geq -\epsilon_k,\ \boldsymbol{x} \in \mathbb{R}^n.$$

By passing to the limit on the above inequality, making use of $\epsilon_k \to 0$, the closedness of the conical hull of the finite set $\{B^j : j \in I^*\}$, there exist y_j, $j \in I^*$ such that for any $\boldsymbol{x} \in \mathbb{R}^n$, it holds

$$\begin{aligned} 0 \leq & \big\langle \zeta^* + \sum_{j\in I^*} y_j B^j, \boldsymbol{x} - \boldsymbol{x}^*\big\rangle \leq \max_{\zeta_i^* \in \partial\phi(\|x_i^*\|)} \langle \zeta_i^*, \boldsymbol{x}_i - \boldsymbol{x}_i^*\rangle + \bar{y} \sum_{j\in I^*} \max\{0, \langle B^j, \boldsymbol{x} - \boldsymbol{x}^*\rangle\} \\ = & \max_{\zeta_i^* \in \partial\phi(\|x_i^*\|)} \langle \zeta_i^*, \boldsymbol{x}_i - \boldsymbol{x}_i^*\rangle + \bar{y} \sum_{j\in I^*} \max_{i\in I_2^j(\boldsymbol{x}^*)} \xi_i'(\langle B^j, \boldsymbol{x}^*\rangle - h_j)\langle B^j, \boldsymbol{x} - \boldsymbol{x}^*\rangle \\ = & \Phi'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) + \bar{y}F'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*), \end{aligned}$$

where $\bar{y} = \max\{|y_j| : j \in I^*\}$.

Case 2. Suppose that $\|A\boldsymbol{x}^* - b\|^2 = \sigma^2$. Denote $t_k := \lambda_k h'_{\mu_k}(\|A\boldsymbol{x}^k - b\|^2 - \sigma^2) \geq 0$. We claim $\{t_k\}$ is bounded. Assume on the contrary that $\{t_k\}$ is unbounded and $\{t_k\} \to \infty$. From (4.3), we have

$$\big\langle \frac{\zeta^k}{t_k} + A^\top(A\boldsymbol{x}^k - b) + \sum_{j\in I^*} \frac{w_j^k}{t_k} B^j, \boldsymbol{x} - \boldsymbol{x}^k\big\rangle \geq -\epsilon_k,\ \boldsymbol{x} \in \mathbb{R}^n.$$

Passing to the limit in the above equality and using the boundedness of $\{\zeta^k\}$ as well as the closedness of finitely generated cones, we find

$$0 \in A^\top(A\boldsymbol{x}^* - b) + \mathcal{N}_{B\cdot\leq h}(\boldsymbol{x}^*).$$

This means that $\boldsymbol{x}^*$ is an optimal solution of the problem $\{\min \frac{1}{2}\|A\boldsymbol{x} - b\|^2 \ \text{s.t.}\ B\boldsymbol{x} \leq h\}$. Since $\|A\boldsymbol{x}^* - b\| = \sigma$, this contradicts our assumption that there is $\boldsymbol{x}^0 \in \Omega$ with $\|A\boldsymbol{x}^0 - b\| < \sigma$. Thus $\{t_k\}$ is a nonnegative bounded sequence. Let $\{t_k\} \to t^* \geq 0$. Taking limits on both sides of (4.3), invoking the closedness of finitely generated cones, we can see that there exists

$\zeta^* \in \partial\Phi(\|\boldsymbol{x}^*\|)$ such that

$$\begin{aligned} 0 \leq & \langle \zeta^*, \boldsymbol{x} - \boldsymbol{x}^* \rangle + \langle t^* A^\top (A\boldsymbol{x}^k - b) + \sum_{j \in I^*} y_j B^j, \boldsymbol{x} - \boldsymbol{x}^* \rangle \\ & \leq \Phi'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) + \bar{y} F'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*), \quad \boldsymbol{x} \in \mathbb{R}^n, \end{aligned}$$

where $\bar{y} := \max\{|t^*|, |y_j| : j \in I^*\}$. Therefore, we have $\Phi'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) + \bar{y} F'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$.

From the expression of directional derivative, it holds that for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$0 \leq \Phi'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) + \bar{y} F'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) = \max_{\boldsymbol{v}_1 \in \partial\Phi(\boldsymbol{x}^*)} \langle \boldsymbol{v}_1, \boldsymbol{x} - \boldsymbol{x}^* \rangle + \bar{y} \max_{\boldsymbol{v}_2 \in \partial F(\boldsymbol{x}^*)} \langle \boldsymbol{v}_2, \boldsymbol{x} - \boldsymbol{x}^* \rangle. \tag{4.4}$$

Notice that for $\boldsymbol{x}^* \in \Omega$, $\partial F(\boldsymbol{x}^*) = \vartheta A^\top (A\boldsymbol{x}^* - b) + \sum_{j=1}^l \mu_j' B^j$ with

$$\vartheta \begin{cases} \in [0,1], & \|A\boldsymbol{x}^* - b\|^2 = \sigma^2, \\ = 0, & \|A\boldsymbol{x}^* - b\|^2 < \sigma^2 \end{cases} \quad \text{and} \quad \mu_j' \begin{cases} \in [0,1], & \langle B^j, \boldsymbol{x} \rangle = h_j, \\ = 0, & \langle B^j, \boldsymbol{x} \rangle < h_j. \end{cases} \tag{4.5}$$

From [41, Theorem 6.42], we have $\partial F(\boldsymbol{x}^*) \subseteq \mathcal{N}_\Omega(\boldsymbol{x}^*)$. Hence $\langle \boldsymbol{v}_2, \boldsymbol{x} - \boldsymbol{x}^* \rangle \leq 0$ for any $\boldsymbol{v}_2 \in \partial F(\boldsymbol{x}^*)$ and $\boldsymbol{x} \in \Omega$. Together with (4.4), it yields that $\Phi'(\boldsymbol{x}^*, \boldsymbol{x} - \boldsymbol{x}^*) = \max_{\boldsymbol{v} \in \partial\Phi(\boldsymbol{x}^*)} \langle \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$ for any $\boldsymbol{x} \in \Omega$ and $\boldsymbol{x}^*$ is a d-stationary point of (R_ν). ∎

5. Numerical simulations.

In our numerical simulations, we use the nonmonotone proximal gradient (NPG) method [16, 47] to solve the subproblem in Step (1) of Algorithm 4.1. The subproblem is an unconstrained optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} G_{\lambda,\mu}(\boldsymbol{x}) := \Phi(\boldsymbol{x}) + \lambda F_\mu(\boldsymbol{x}). \tag{5.1}$$

The NPG method for solving (5.1) is presented as follows.

Algorithm 5.1 NPG method for (5.1)

Let $\boldsymbol{x}^0 \in \Omega$ be given. Choose $L_{\max} \geq L_{\min} > 0$, $\kappa > 1$, $c > 0$ and an integer $M \geq 0$ arbitrarily. Set $k = 0$.

(1) Choose $L_k^0 \in [L_{\min}, L_{\max}]$ arbitrarily. Set $L_k = L_k^0$.

(1a) Solve the subproblem

$$\boldsymbol{u} \in \operatorname{Argmin}_{\boldsymbol{x}} \left\{ \lambda \langle \nabla F_\mu(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k \rangle + \frac{L_k}{2} \|\boldsymbol{x} - \boldsymbol{x}^k\|^2 + \Phi(\boldsymbol{x}) \right\}. \tag{5.2}$$

(1b) If $G_{\lambda,\mu}(\boldsymbol{u}) \leq \max_{[k-M]_+ \leq i \leq k} G_{\lambda,\mu}(\boldsymbol{u})(\boldsymbol{x}^i) - \frac{c}{2}\|\boldsymbol{u} - \boldsymbol{x}^k\|^2$, then go to step (2).

(1c) Set $L_k \leftarrow \kappa L_k$ and go to step (1a).

(2) Set $\boldsymbol{x}^{k+1} \leftarrow \boldsymbol{u}$, $\bar{L}_k \leftarrow L_k$, $k \leftarrow k+1$ and go to step (1).

end

We can find $\boldsymbol{u}$ in (5.2) by solving m optimization problems in $\mathbb{R}^{n_i}$ for each group in parallel. Moreover, (5.2) can be written as a proximal operator

$$\boldsymbol{u} \in \operatorname{Argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2} \left\| \boldsymbol{x} - \boldsymbol{x}^k + \frac{\lambda}{L_k} \nabla F_{\lambda,\mu}(\boldsymbol{x}^k) \right\|^2 + \frac{1}{L_k} \Phi(\boldsymbol{x}) \right\},$$

which has a closed form solution. See Appendix A.

For the NPG method to solve the unconstrained subproblem (5.1) at $\lambda = \lambda_k$ and $\mu = \mu_k$, we set $L_{\min} = 1$, $L_{\max} = 10^8$, $\kappa = 2$, $c = 10^{-4}$, $L_0^0 = 1$, and for any $l \geq 1$,

$$L_l^0 := \min\left\{\max\left\{\frac{[\boldsymbol{x}^{k,l} - \boldsymbol{x}^{k,l-1}]^\top[\nabla F_{\lambda_k,\mu_k}(\boldsymbol{x}^{k,l}) - \nabla F_{\lambda_k,\mu_k}(\boldsymbol{x}^{k,l-1})]}{\|\boldsymbol{x}^{k,l} - \boldsymbol{x}^{k,l-1}\|^2}, L_{\min}\right\}, L_{\max}\right\}.$$

The NPG method is terminated when

$$\|\boldsymbol{x}^{k,l} - \boldsymbol{x}^{k,l-1}\|_\infty \leq \sqrt{\epsilon_k} \quad \text{and} \quad \frac{|\nabla F_{\lambda_k,\mu_k}(\boldsymbol{x}^{k,l}) - \nabla F_{\lambda_k,\mu_k}(\boldsymbol{x}^{k,l-1})|}{\max\{1, |\nabla F_{\lambda_k,\mu_k}(\boldsymbol{x}^{k,l})|\}} \leq \min\{\epsilon_k^2, 10^{-4}\}.$$

Algorithm 4.1 is terminated when

$$\max\{(\|A\boldsymbol{x}^k - b)\|^2 - \sigma^2)_+, 0.01\epsilon_k\} \leq 10^{-6}$$

and ϵ_{k+1} in Step (2) of Algorithm 4.1 is updated as $\max\{\theta\epsilon_k, 10^{-6}\}$ in our numerical implementation.

All codes are written in MATLAB, and the experiments were performed in MATLAB R2017a on a laptop computer with 2.6GHz CPU and 8GB RAM. Our testing problems in this section are from recovering a group sparse solution and images from an underdetermined linear system with noisy measurements, which are formulated as optimization problem (R_ν) without the constraint $B\boldsymbol{x} \leq h$.

5.1. Random data. For problem (R_ν), we set $\nu = 0.02$ and $n_i = 8$. Parameters in Algorithm 4.1 and Algorithm 5.1 are set as $\boldsymbol{x}^0 = \mathbf{1}_n$, the vector of all ones, $\lambda_0 = 40$, $\mu_0 = \epsilon_0 = 1$, $\rho = 2$, $\theta = \frac{1}{\rho}$, $M = 3$ and $\boldsymbol{x}^{feas} = A^\dagger b$.

We compare Algorithm 4.1 with SPGl1 [45] (http://www.cs.ubc.ca/.mpf/SPGl1/) for solving group lasso model $\min \|\boldsymbol{x}\|_{2,1}$ s.t. $\|A\boldsymbol{x} - b\| \leq \sigma$, and FISTA [4] for solving $\min \frac{1}{2}\|A\boldsymbol{x} - b\|^2 + \lambda\|\boldsymbol{x}\|_{2,1}$ (https://github.com/tiepvupsu/FISTA), where $\|\boldsymbol{x}\|_{2,1} = \sum_{i=1}^m \|\boldsymbol{x}_i\|$. We use warm restart strategy for FISTA: start from $\lambda = 5$, decrease it by half in every iteration and use the result as the initial point in the next iteration until $\|A\boldsymbol{x} - b\|^2 < 1$ at $\lambda = 0.01$. Then we use FISTA with fixed $\lambda = 0.01$ for the rest of iterations. The data is generated as [45], where Kg is the number of nonzero groups in the signal. In Table 2, we report the number of nonzero groups (nnz) in the approximate solution $\boldsymbol{x}$ obtained by the algorithms and the CPU time in seconds. One can observe that Algorithm 4.1 produces sparser solutions than the group SPGl1 method and FISTA. Among the four capped folded concave functions given in Section 1, capped L_1 outperforms the other three functions in terms of CPU time for this test. Since the group lasso model $\min \|\boldsymbol{x}\|_{2,1}$ s.t. $\|A\boldsymbol{x} - b\| \leq \sigma$ is closely related to problem (P_0), we compare Algorithm 4.1 with SPGl1 only in the following image recovery test problems.

Data			FISTA		G-SPGl1		L_1		MCP		$L_{1/2}$		Fraction	
r	m	Kg	nnz	Time	nnz	Time	nnz	Time	nnz	Time	nnz	Time	nnz	Time
720	768	38	633	12.82	561	7.37	554	1.37	554	5.69	554	3.20	554	3.14
900	960	48	770	20.96	717	9.67	699	1.90	700	9.90	697	4.53	697	4.49
1080	1152	57	933	31.72	849	11.49	837	2.48	836	13.79	836	6.26	836	6.17
1260	1344	67	1091	45.52	1045	15.64	964	2.88	962	10.82	961	8.07	961	8.01
1440	1536	76	1218	64.66	1117	22.82	1099	3.60	1106	11.81	1096	7.02	1096	6.97
1620	1728	86	1419	81.73	1263	22.34	1255	3.79	1253	15.77	1252	13.05	1252	12.92
1800	1920	96	1524	106.64	1448	25.01	1389	5.34	1390	23.62	1385	11.16	1385	11.26

Table 2

Comparing Algorithm 4.1 with four capped folded concave functions ($\alpha = 20$ for MCP and Fraction) and Group SPGl1, FISTA with $\lambda = 0.01$.

5.2. Multichannel Image Reconstruction. We consider recovering four 2D images from compressive and noisy measurement by Group SPGl1 and Algorithm 4.1. We set parameters $\mu_0 = 1$ and $\epsilon_0 = 1$ for the four images. Other parameters are set according to different images. We use the output iterates obtained by group SPGl1 as the initial point of Algorithm 4.1. In our tables, the PSNR is defined by PSNR $= 10 \cdot \log \frac{V^2}{MSE}$, where V and MSE are the maximum absolute value and the mean squared error of the reconstruction, respectively. Parameters in group SPGl1 are default.

The first example is a multichannel image recovery problem (denoted as Image 1) taken from [29, 45]. We adopt the same method as in [29] to process the image. The observational data b is generated by $b = A\boldsymbol{x}_{orig} + \eta$, where A is a random Gaussian matrix (without correlation within each group), $\boldsymbol{x}_{orig}$ is the target coefficient with a group sparse structure and η is Gaussian noise with the noise level $Var(\eta)$. The parameters are set as $r = 1152$, $m = 2304$, $n_i = 3$ $(i = 1, \cdots, m)$, $Kg = 152$. Parameters in Algorithm 4.1 and Algorithm 5.1 are set as $\lambda_0 = 40$, $M = 3$, $\rho = 5$, $\theta = \frac{1}{\rho}$, $\nu = 0.02$ for Capped L_1, $\nu = 0.01$ for other three functions. The PSNR and CPU time are reported in Table 3 and the recovered images with noise level $Var(\eta) = 10^{-3}$ are presented in Figure 2.

	Group SPGl1		L_1		MCP, $\alpha = 4$		$L_{1/2}$		Fraction, $\alpha = 4$	
Var(η)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)
10^{-1}	19.2400	0.2581	19.6033	1.1254	19.5734	1.0860	19.5734	1.2206	19.5734	1.2337
10^{-2}	23.8819	1.1281	48.2727	1.8495	48.2611	1.8204	48.2611	1.9159	48.2611	1.9206
10^{-3}	23.8675	1.1066	64.2713	1.9797	64.2079	2.0062	64.3523	2.3888	64.2079	2.1527
10^{-4}	23.6657	3.4767	88.3341	4.9455	87.2615	4.7036	87.2615	4.8911	87.2615	4.7603
10^{-5}	23.4267	4.2950	103.6857	6.0991	106.8684	5.6479	106.8684	5.8563	106.8684	5.8547
10^{-6}	24.0451	3.9983	105.3054	5.5021	107.7318	5.3069	107.7318	5.4371	107.7318	5.3849

Table 3

Comparing Algorithm 4.1 with four capped folded concave functions and Group SPGl1 using Image 1.

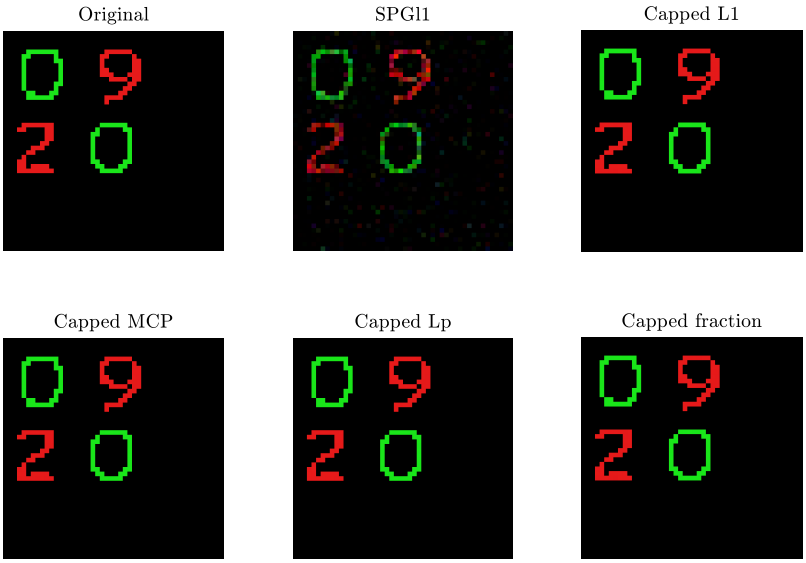

Figure 2. *Recovery results of Image 1.*

The second color image with three-channels (denoted as Image 2) is downloaded from http://www.med.harvard.edu/AANLIB/cases/case12/mr2-tc2/007.html. We preprocess the image at first, including resizing the image into 96×96 and setting the pixels less than 0.3 as zero. We then take the same method as the first example to produce the group sparse target signal, which has 9216 groups with 1363 nonzero groups (each of group size 3). The sampling matrix A is a random Gaussian matrix with a size 6912×27648. Parameters in Algorithm 4.1 and Algorithm 5.1 are set as $\rho = 3$, $M = 5$, $\lambda_0 = 55$, $\theta = \frac{1}{\rho}$, $\nu = 0.001$ for Capped L_1, $\nu = 0.03$ for other three functions. The PSNR and CPU time are reported in Table 4 and the recovered images with noise level $Var(\eta) = 10^{-3}$ are presented in Figure 3.

	Group SPGl1		L_1		MCP, $\alpha = 10$		$L_{1/2}$		Fraction, $\alpha = 10$	
Var(η)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)
10^{-1}	21.4192	4.3258	21.7508	15.5052	21.6872	15.5435	21.7846	16.0082	21.6872	16.2580
10^{-2}	23.9901	23.4516	35.2165	44.4575	35.9986	47.3201	35.9986	48.6960	35.9986	48.5604
10^{-3}	24.3623	49.7471	52.0845	81.3036	52.1006	75.9230	52.1006	77.7247	52.1006	77.4327
10^{-4}	24.4011	66.4492	71.5541	108.9791	71.3558	105.9179	71.3558	108.6231	71.3558	106.9660
10^{-5}	24.3855	126.2611	90.5679	174.5781	94.7775	173.9237	94.7775	177.3438	94.7775	175.7780
10^{-6}	24.4086	195.3681	101.6114	245.3174	99.4979	243.8206	97.0763	241.1279	99.4979	246.7084

Table 4

Comparing Algorithm 4.1 with four capped folded concave functions and Group SPGl1 using Image 2.

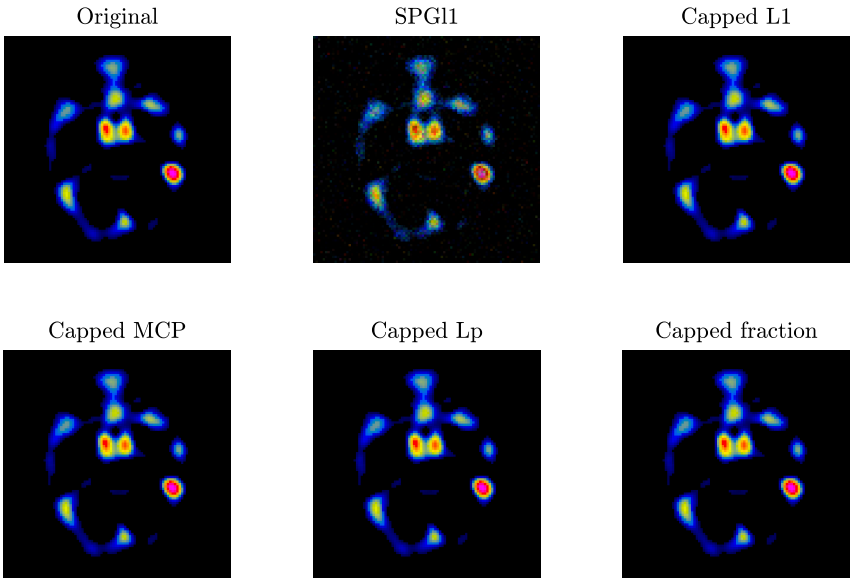

Figure 3. *Recovery results of Image 2.*

The third example is a multichannel MRI recovery problem (denoted as Image 3) taken from [29, 45]. The sampling matrix A is the composition of a partial FFT with an inverse wavelet transform, with a size 3771×12288, where we have used 6 levels of Daubechies 1 wavelet. The three channels for each wavelet expansion are organized into one group, and the underlying image has 4096 groups (each of group size 3) with 724 nonzero groups. The data is formed as $b = A\boldsymbol{x}_{orig} + \eta$, where $\boldsymbol{x}_{orig}$ is the target coefficient with a group sparse structure and η is the Gaussian noise. The recovered image is then obtained by applying the inverse wavelet transform to the estimated coefficient. Parameters of Algorithm 4.1 and Algorithm 5.1 are set as $\rho = 3$, $M = 5$, $\lambda_0 = 40$, $\theta = \frac{1}{\rho}$. We set $\nu = 0.001$ for Capped L_1 and $\nu = 0.01$ for other three functions. The results are reported in Table 5 and the recovered images with noise level $Var(\eta) = 10^{-3}$ are presented in Figure 4.

	Group SPGl1		L_1		MCP, $\alpha = 10$		$L_{1/2}$		Fraction, $\alpha = 10$	
Var(η)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)
10^{-1}	17.9925	1.5646	18.1336	1.4649	18.1336	1.4655	18.1336	1.4708	18.1336	1.4560
10^{-2}	23.9676	4.3351	39.0348	4.2087	38.9066	4.2006	38.9066	4.2438	38.9066	4.2089
10^{-3}	24.9022	11.7074	57.7403	11.4018	58.4163	11.3901	58.4163	11.4976	58.4163	11.5707
10^{-4}	24.9071	19.4300	78.1145	18.5822	76.5253	18.3975	76.5253	18.3766	76.5253	18.1372
10^{-5}	24.8083	34.8007	103.1244	33.8207	96.2496	33.7517	96.2496	33.6962	96.2496	33.7564
10^{-6}	24.7521	28.1514	106.5624	26.9065	103.5421	26.9485	103.5421	27.0806	103.5421	26.9169

Table 5

Comparing Algorithm 4.1 with four capped folded concave functions and Group SPGl1 using Image 3.

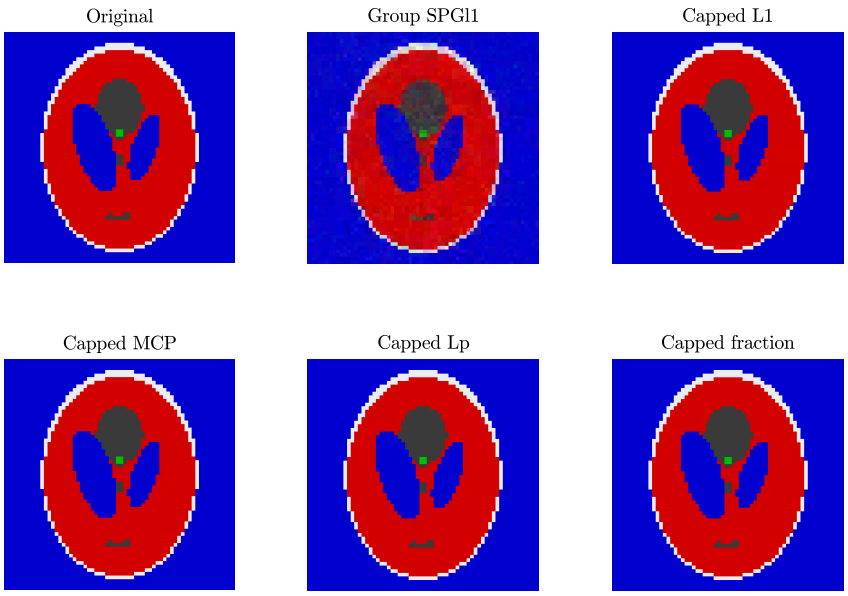

Figure 4. *Recovery results of Image 3.*

The fourth 256×256 grayscale image (denoted as Image 4) is download from http://www.med.harvard.edu/AANLIB/cases/caseNN1/mr1-dg1/015.htm. We resize the image into 128×128 and set the pixels less than 0.3 as zero to reduce the group sparsity. Then we partition the image into 4096 grids with size of 2×2. The pixels in the same grid are organized into one group and reordered into a vector.. All the groups are reordered into a vector and the target coefficient with group sparsity structure is obtained. The sampling matrix A is a random Gaussian matrix with a size 4096×16384. Parameters of penalty methods for three-dimensional image, where $\nu = 0.001$, $M = 8$, $\lambda_0 = 40$, $\theta = \frac{1}{\rho}$ are the same. We set $\rho = 1.5$ for Capped L_1, $\rho = 1.2$ for Capped MCP, $\rho = 2$ for Capped L_p and Capped fraction. The results of PSNR and CPU time are reported in Table 6 and the recovered images with noise level $Var(\eta) = 10^{-3}$ are presented in Figure 5.

	Group SPGl1		L_1		MCP, $\alpha = 10$		$L_{1/2}$		Fraction, $\alpha = 10$.	
Var(η)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)	PSNR	Time(s)
10^{-1}	23.5054	2.0878	23.8205	1.5634	23.8205	2.5109	23.8205	1.4602	23.8205	1.4539
10^{-2}	25.4102	7.8816	29.4150	9.2259	29.8325	29.4280	29.8325	12.0694	29.8325	11.2612
10^{-3}	25.4730	17.6577	39.7290	59.7255	39.6593	61.1911	39.6593	63.5210	39.6593	59.6978
10^{-4}	25.6481	29.3783	41.0732	69.9874	40.4339	70.0647	40.4339	62.6833	40.4339	58.7915
10^{-5}	25.4717	31.4714	40.3723	58.1390	40.5878	66.8393	40.5878	60.5256	40.5878	57.3971
10^{-6}	25.4778	39.1183	41.7962	67.0524	40.6657	65.4038	40.6657	74.7206	40.6657	70.9018

Table 6

Comparing Algorithm 4.1 with four capped folded concave functions and Group SPGl1 using Image 4.

To see the influence of parameter ν to PSNR, we fix the noise level $Var(\eta) = 10^{-3}$ and use the same parameter for Table 6 to present PSNR with different ν for Image 4 in Table 7. The results show that decreasing ν can increase PSNR.

ν	0.002	0.004	0.006	0.008	0.01	0.012	0.014	0.016	0.018	0.02
L_1	38.5384	38.5384	35.8403	35.6003	36.5972	33.2918	34.7058	34.2895	33.4616	31.8339
MCP	39.6427	39.6427	39.6427	38.7634	38.5288	38.8392	38.9495	39.1603	38.5645	37.3792
$L_{1/2}$	39.6427	39.6427	39.6427	38.7634	38.5288	38.8392	38.9495	39.1603	38.5645	37.3792
Fraction	39.6427	39.6427	39.6427	38.7634	38.5288	38.8392	38.9495	39.1603	38.5645	37.3792

Table 7

PSNR of Image 4 by Algorithm 4.1 with different values of parameter ν.

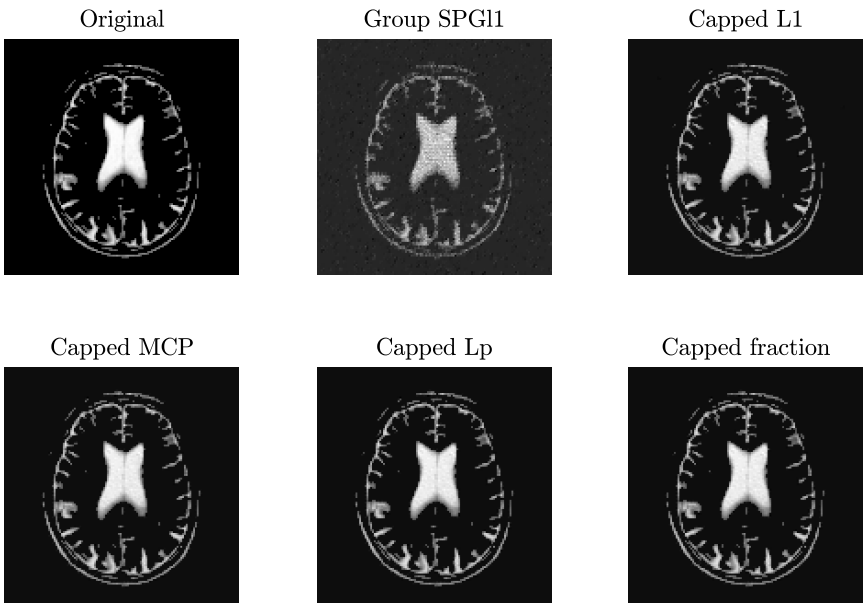

Figure 5. *Recovery results of Image 4.*

To summarize the numerical experiments, we give the variation of PSNR and CPU time of the SPGl1 and Algorithm 4.1 with respect to the noise level in Figure 6, where we present the average of the results of the four capped folded concave penalty functions.

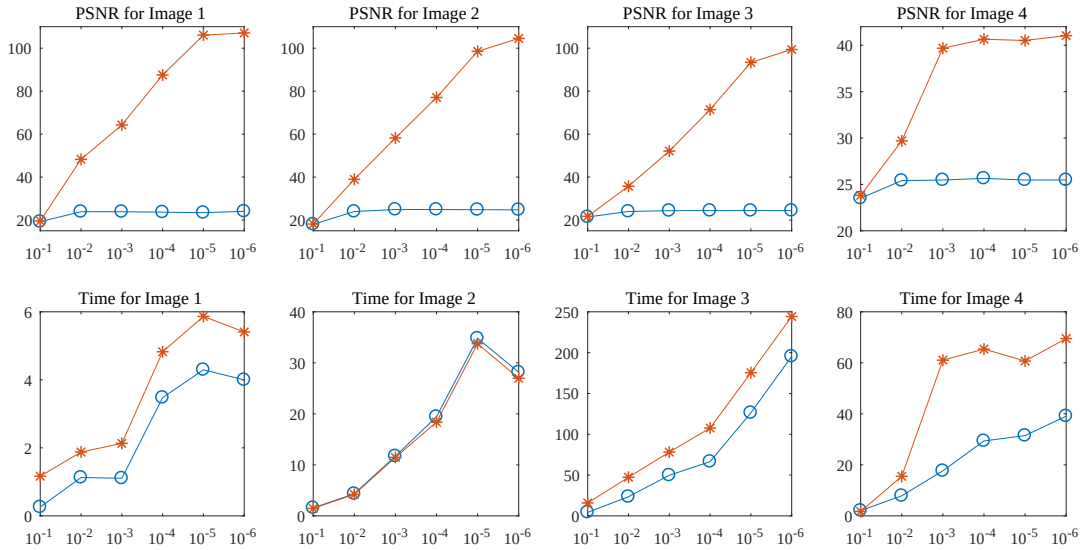

Figure 6. *Comparing Algorithm 4.1 with Group SPGl1 on PSNR and CPU time (specified in second) for Images 1-4. The red star line stands for Algorithm 4.1 and blue circle line stands for group SPGl1.*

Our numerical results show that Algorithm 4.1 with capped folded concave functions can significantly improve PSNR values obtained by group SPGl1 method as the noise level decrease.

6. Conclusions. In this paper, we consider constrained group sparse optimization (P_0) for image recovery problem. We study its continuous relaxation problem (R_ν), its exact penalty problem (1.3) and its continuous relaxation penalty problem (P_ν). We establish the links between the four problems regarding global minimizers. Moreover, we propose a smoothing penalty algorithm (Algorithm 4.1) to solve problem (R_ν) and show any accumulation point generated by Algorithm 4.1 is a d-stationary point of (R_ν). The numerical experiments show that Algorithm 4.1 with the four capped folded concave functions can achieve higher quality solutions.

Appendix A. Proximal operator. Given a function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the proximal operator of $x \in \mathbb{R}^n$ with respect to h is defined as

$$\mathrm{prox}_h(x) = \mathrm{Argmin}\{\frac{1}{2}\|z - x\|^2 + h(z) : z \in \mathbb{R}^n\}.$$

When h is not convex, the proximal operator may return multiple minimizers and should therefore be multivalued. In this section, we list the proximal operators of the four capped folded concave functions given in Section 1. We first observe

$$\psi(\|\boldsymbol{z}\|)\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \in \mathrm{Argmin}\{\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 + \lambda\phi(\|\boldsymbol{x}\|) : \boldsymbol{x} \in \mathbb{R}^n\}, \quad \boldsymbol{z} \in \mathbb{R}^n, \tag{A.1}$$

where $\psi(z) \in \text{Argmin}\{\frac{1}{2}(u-z)^2 + \lambda\phi(u) : u \in \mathbb{R}_+\}$. Indeed, it is from

$$\begin{aligned}&\frac{1}{2}\left\|\psi(\|\boldsymbol{z}\|)\frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} - \boldsymbol{z}\right\|^2 + \lambda\phi\left(\psi(\|\boldsymbol{z}\|)\right) = \frac{1}{2}\left(\psi(\|\boldsymbol{z}\|) - \|\boldsymbol{z}\|\right)^2 + \lambda\phi\left(\psi(\|\boldsymbol{z}\|)\right)\\ \leq&\frac{1}{2}\left(\|\boldsymbol{x}\| - \|\boldsymbol{z}\|\right)^2 + \lambda\phi(\|\boldsymbol{x}\|) \leq \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{z}\right\|^2 + \lambda\phi(\|\boldsymbol{x}\|),\ \forall \boldsymbol{x} \in \mathbb{R}^n.\end{aligned}$$

A.1. Proximal operator of capped L_1 penalty function. Let

$$\boldsymbol{u}^{\text{CapL1}}(\boldsymbol{z}) \in \text{Argmin}_{\boldsymbol{u}\in\mathbb{R}^n}\{f^{\text{CapL1}}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{z}\|^2 + \lambda\sum_{i=1}^m \phi^{\text{CapL1}}(\|\boldsymbol{u}_i\|)\}, \tag{A.2}$$

where $\boldsymbol{z} \in \mathbb{R}^n$ and $\lambda > 0$. Since problem (A.2) is separable on group level, we compute every $\boldsymbol{u}_i^{\text{CapL1}}(\boldsymbol{z})$, $i = 1, \cdots, m$. From (A.1), we only need to compute the following problem:

$$u^{\text{CapL1}}(z) = \text{argmin}\{f^{\text{CapL1}}(u) := \frac{1}{2}(u-z)^2 + \lambda\phi^{\text{CapL1}}(u) : u \in \mathbb{R}_+\}.$$

Note that $f^{\text{CapL1}}(u)$ can be written as

$$f^{\text{CapL1}}(u) = \begin{cases} \frac{1}{2}(u-z)^2 + \frac{\lambda}{\nu}u, & 0 \leq u < \nu,\\ \frac{1}{2}(u-z)^2 + \lambda, & u \geq \nu.\end{cases}$$

When $0 \leq u < \nu$, the minimizer of $f^{\text{CapL1}}(u)$ is $u^* = z - \frac{\lambda}{\nu}$. By $0 \leq u^* < \nu$, we obtain $\frac{\lambda}{\nu} \leq z < \nu + \frac{\lambda}{\nu}$. Therefore, we have $u^{\text{CapL1}}(z) = 0$ if $z < \frac{\lambda}{\nu}$ and $u^{\text{CapL1}}(z) = z - \frac{\lambda}{\nu}$ if $\frac{\lambda}{\nu} \leq z < \nu + \frac{\lambda}{\nu}$. This means $u^{\text{CapL1}}(z) = (z - \frac{\lambda}{\nu})_+$ when $z \leq \nu + \frac{\lambda}{\nu}$.

We compare the values of $f^{\text{CapL1}}(u^*)$ and $f^{\text{CapL1}}(z)$ and find the minimizer

$$\boldsymbol{u}_i^{\text{CapL1}}(\boldsymbol{z}_i) = \begin{cases} (\|\boldsymbol{z}_i\| - \frac{\lambda}{\nu})_+\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & \|\boldsymbol{z}_i\| \leq \nu + \frac{\lambda}{2\nu},\\ \boldsymbol{z}_i, & \|\boldsymbol{z}_i\| > \nu + \frac{\lambda}{2\nu}, \quad i = 1, \ldots, m.\end{cases}$$

A.2. Proximal operator of capped MCP penalty function. Let

$$\boldsymbol{u}^{\text{C-MCP}}(\boldsymbol{z}) \in \text{argmin}_{\boldsymbol{u}\in\mathbb{R}^n}\{f^{\text{C-MCP}}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{z}\|^2 + \lambda\sum_{i=1}^m \phi^{\text{C-MCP}}(\|\boldsymbol{u}_i\|)\},$$

where $\boldsymbol{z} \in \mathbb{R}^n$ and $\lambda > 0$. The corresponding single variable minimization problem is

$$u^{\text{C-MCP}}(\boldsymbol{z}) \in \text{argmin}_{u\in\mathbb{R}}\{f^{\text{C-MCP}}(u) := \frac{1}{2}(u-z)^2 + \lambda\phi^{\text{C-MCP}}(u)\}.$$

For simplicity, denote $\nu^M := \frac{2\alpha}{2\alpha\nu - \nu^2}$. The function $f^{\text{C-MCP}}(u)$ can be expressed as

$$f^{\text{C-MCP}}(u) = \begin{cases} \frac{1}{2}(u-z)^2 + \frac{\lambda}{\nu^M}(u - \frac{u^2}{2\alpha}), & 0 \leq u < \nu,\\ \frac{1}{2}(u-z)^2 + \lambda, & u \geq \nu.\end{cases}$$

When $0 \le u < \nu$, let $\frac{d}{du} f^{\mathrm{C-MCP}}(u) = u - z + \frac{\lambda}{\nu^M}(1 - \frac{u}{\alpha}) = 0$. We have $u^*(z) = \frac{z - \frac{\lambda}{\nu^M}}{1 - \frac{\lambda}{\alpha\nu^M}}$. Then the minimizer of $f^{\mathrm{C-MCP}}(u)$ in $[0, \nu]$ is

$$u_1^*(z) = \min\left\{ \left(\frac{z - \frac{\lambda}{\nu^M}}{1 - \frac{\lambda}{\alpha\nu^M}} \right)_+, \nu \right\},$$

and the minimizer of $f^{\mathrm{C-MCP}}(u)$ in $[\nu, \infty)$ is $u_2^*(z) = \max\{z, \nu\}$. Hence we have

$$\boldsymbol{u}_i^{\mathrm{C-MCP}}(\boldsymbol{z}_i) = \begin{cases} u_1^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & f^{\mathrm{C-MCP}}(u_1^*(\|\boldsymbol{z}_i\|)) \le f^{\mathrm{C-MCP}}(u_2^*(\|\boldsymbol{z}_i\|)), \\ u_2^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & \text{otherwise}, \end{cases} \qquad i = 1, \cdots, m.$$

A.3. Proximal operator of capped L_p penalty function.

Let

$$\boldsymbol{u}^{\mathrm{CapLp}}(\boldsymbol{z}) \in \mathrm{argmin}_{\boldsymbol{u}\in\mathbb{R}^n}\{f^{\mathrm{CapLp}}(\boldsymbol{u}) := \|\boldsymbol{u} - \boldsymbol{z}\|^2 + \lambda \sum_{i=1}^m \phi^{\mathrm{CapLp}}(\|\boldsymbol{u}_i\|)\},$$

where $\boldsymbol{z} \in \mathbb{R}^n$, $p = \frac{1}{2}$ and $\lambda > 0$. Denote $\tilde{u}^{\mathrm{CapLp}}(z) \in \mathrm{Argmin}\{\frac{1}{2}(u - z)^2 + \frac{\lambda}{\nu^{\frac{1}{2}}} u^{\frac{1}{2}} : u \ge 0\}$, which was given in [48] that

$$\tilde{u}^{\mathrm{CapLp}}(z) = \frac{2}{3} z \big(1 + \cos(\frac{2\pi}{3} - \frac{2}{3}\varphi_{\lambda,\nu}(z))\big),$$

where $\varphi_{\lambda,\nu}(z) = \arccos\left(\frac{\lambda}{4\nu^{\frac{1}{2}}}(\frac{z}{3})^{-\frac{3}{2}}\right)$. The minimizer of $f^{\mathrm{CapLp}}(u)$ in $[0, \nu]$ is

$$u_1^*(z) = \min\left\{ \left(\tilde{u}^{\mathrm{CapLp}}(z) \right)_+, \nu \right\},$$

and the minimizer of $f^{\mathrm{CapLp}}(u)$ in $[\nu, \infty)$ is $u_2^*(z) = \max\{z, \nu\}$. Hence we have

$$\boldsymbol{u}_i^{\mathrm{CapLp}}(\boldsymbol{z}_i) = \begin{cases} u_1^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & f^{\mathrm{CapLp}}(u_1^*(\|\boldsymbol{z}_i\|)) \le f^{\mathrm{CapLp}}(u_2^*(\|\boldsymbol{z}_i\|)), \\ u_2^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & \text{otherwise}, \end{cases} \qquad i = 1, \cdots, m.$$

A.4. Proximal operator of capped fraction penalty function.

Let

$$\boldsymbol{u}^{\mathrm{CapF}}(\boldsymbol{z}) \in \mathrm{argmin}_{\boldsymbol{u}\in\mathbb{R}^n}\{f^{\mathrm{CapF}}(\boldsymbol{u}) := \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{z}\|^2 + \lambda \sum_{i=1}^m \phi^{\mathrm{CapF}}(\|\boldsymbol{u}_i\|)\},$$

where $\boldsymbol{z} \in \mathbb{R}^n$ and $\lambda > 0$. Denote $\nu^F := \frac{1+\alpha\nu}{\alpha\nu}$ and

$$\tilde{u}^{\mathrm{CapF}}(z) \in \mathrm{Argmin}\{\frac{1}{2}(u - z)^2 + \frac{\lambda}{\nu^F}\frac{\alpha u}{1 + \alpha u} : u \ge 0\}.$$

It was given in [33] that

$$\tilde{u}^{\mathrm{CapF}}(z) = \begin{cases} \mathrm{sgn}(t)\frac{\frac{1+\alpha t}{3}(1+2\cos(\frac{\varphi(t)}{3} - \frac{\pi}{3})) - 1}{\alpha}, & |z| > t, \\ 0, & |z| \le t, \end{cases}$$

where $\varphi(t) = \arccos(\frac{27\lambda\alpha^2}{2\nu^F(1+\alpha|z|)^3} - 1)$, $t = \begin{cases} t_1^*, & \lambda \le \frac{\nu^F}{2\alpha^2}, \\ t_2^*, & \lambda > \frac{\nu^F}{2\alpha^2}, \end{cases}$ with $t_1^* := \frac{\lambda\alpha}{\nu^F}$ and $t_2^* := \sqrt{\frac{2\lambda}{\nu^F}} - \frac{1}{2\alpha}$.

The minimizer of $f^{\mathrm{CapF}}(u)$ in $[0, \nu]$ is

$$u_1^*(z) = \min\left\{\left(\tilde{u}^{\mathrm{CapF}}(z)\right)_+, \nu\right\},$$

and the minimizer of $f^{\mathrm{CapF}}(u)$ in $[\nu, \infty)$ is $u_2^*(z) = \max\{z, \nu\}$. Hence we have

$$\boldsymbol{u}_i^{\mathrm{CapF}}(\boldsymbol{z}_i) = \begin{cases} u_1^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & f^{\mathrm{CapF}}(u_1^*(\|\boldsymbol{z}_i\|)) \le f^{\mathrm{CapF}}(u_2^*(\|\boldsymbol{z}_i\|)), \\ u_2^*(\|\boldsymbol{z}_i\|)\frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|}, & \text{otherwise}, \end{cases} \qquad i = 1, \cdots, m.$$

REFERENCES

[1] M. Ahn, J. S. Pang, and J. Xin, *Difference-of-convex learning: directional stationarity, optimality, and sparsity*, SIAM J. Optim., 27 (2017), pp. 1637–1665.

[2] M. E. Ahsen and M. Vidyasagar, *Error bounds for compressed sensing algorithms with group sparsity: a unified approach*, Appl. Comput. Harmon. Anal., 43 (2017), pp. 212–232.

[3] A. Beck and N. Hallak, *Optimization problems involving group sparsity terms*, Math. Program., (2017), pp. 1–29.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[5] D. Bertsimas, A. King, and R. Mazumder, *Best subset selection via a modern optimization lens*, Ann. Statist, 44 (2016), pp. 813–852.

[6] W. Bian and X. Chen, *Optimality and complexity for constrained optimization problems with nonconvex regularization*, Math. Oper. Res., 42 (2017), pp. 1063–1084.

[7] W. Bian and X. Chen, *A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty*, SIAM J. Numer. Anal., 58(2020), pp. 858-883.

[8] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau, *Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance*, IEEE Trans. Signal Process., 64 (2015), pp. 1405–1419.

[9] P. Breheny and J. Huang, *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors*, Stat. Comput., 25 (2015), pp. 173–187.

[10] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.

[11] M. Carlsson, *On convex envelopes and regularization of non-convex functionals without moving global minima*, J. Optim. Theory Appl., 183 (2019), pp. 66-84.

[12] A. Chambolle and Ch. Dossal, *On the convergence of the iterates of the "Fast iterative shrinkage/thresholding algorithm"*, J. Optim. Theory Appl., 166 (2015), pp. 968-982.

[13] W. Chao, R. Chan, M. Nikolova, R. Plemmons, and S. Prasad, *Nonconvex optimization for 3D point source localization using a rotating point spread function*, SIAM J. Imaging Sci., 12 (2018), pp. 259–286.

[14] J. Chen and X. Huo, *Theoretical results on sparse representations of multiple-measurement vectors*, IEEE Trans. Signal Process., 54 (2006), pp. 4634–4643.

[15] X. Chen, D. Ge, Z. Wang, and Y. Ye, *Complexity of unconstrained l_2-l_p minimization*, Math. Program., 143 (2014), pp. 371–383.

[16] X. Chen, Z. Lu, and T. K. Pong, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492.

[17] X. Chen, L. Pan and N. Xiu, *Solution sets of three sparse optimization problems for multivariate regression*, revision 2019.

[18] X. Chen and R.Womersley, *Spherical designs and nonconvex minimization for recovery of sparse signals on the sphere*, SIAM J. Imaging Sci., 11(2018), pp. 1390–1415.

[19] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, *Block-sparse signals: uncertainty relations and efficient recovery*, IEEE Trans. Signal Process., 58 (2010), pp. 3042–3054.

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

[21] S. Foucart and M.-J. Lai, *Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q < 1$*, Appl. Comput. Harmon. Anal., 26 (2009), pp. 395–407.

[22] G. Fung and O. Mangasarian, *Equivalence of minimal l_0-and l_p-norm solutions of linear equalities, inequalities and linear programs for sufficiently small p*, J. Optim. Theory Appl., 151 (2011), pp. 1–10.

[23] Q. T. L. Gia, I. H. Sloan, R. S. Womersley, and Y. G. Wang, *Isotropic sparse regularization for spherical harmonic representations of random fields on the sphere*, Appl. Comput. Harmon. Anal., 49(2019), pp. 257-278.

[24] A. Gramfort and M. Kowalski, *Improving M/EEG source localization with an inter-condition sparse prior*, in IEEE International Symposium on Biomedical Imaging (ISBI), 2009, pp. 141–144.

[25] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang, *Group sparse optimization via $l_{p,q}$ regularization*, J. Mach. Learn. Res., 18 (2017), pp. 1–52.

[26] J. Huang, J. L. Horowitz, and S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Ann. Statist., 36 (2008), pp. 587–613.

[27] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, *A group bridge approach for variable selection*, Biometrika, 96 (2009), pp. 339–355.

[28] J. Huang and T. Zhang, *The benefit of group sparsity*, Ann. Statist., 38 (2010), pp. 1978–2004.

[29] Y. Jiao, B. Jin, and X. Lu, *Group sparse recovery via the $l_0(l_2)$ penalty: theory and algorithm*, IEEE Trans. Signal Process., 65 (2017), pp. 998–1012.

[30] A. Juditsky, F. Kilinc Karzan, A. Nemirovski, and B. Polyak, *Accuracy guaranties for ℓ_1 recovery of block-sparse signals*, Ann. Statist., 40 (2012), pp. 3077–3107.

[31] K. Lee, Y. Bresler, and M. Junge, *Subspace methods for joint sparse recovery*, IEEE Trans. Inform. Theory., 58 (2012), pp. 3613–3641.

[32] S. Lee, M. Oh, and Y. Kim, *Sparse optimization for nonconvex group penalized estimation*, J. Stat. Comput. Sim., 86 (2016), pp. 597–610.

[33] H. Li, Q. Zhang, A. Cui, and J. Peng, *Minimization of fraction function penalty in compressed sensing*, arXiv preprint arXiv:1705.06048, (2017).

[34] X. Liao, H. Li, and L. Carin, *Generalized alternating projection for weighted-$\ell_{2,1}$ minimization with applications to model-based compressive sensing*, SIAM J. Imaging Sci., 7 (2014), pp. 797–823.

[35] Y. Liu, S. Bi, and S. Pan, *Equivalent Lipschitz surrogates for zero-norm and rank optimization problems*, J. Glob. Optim., 72 (2018), pp. 679-704.

[36] X. Lv, G. Bi, and C. Wan, *The group lasso for stable recovery of block-sparse signal representations*, IEEE Trans. Signal Process., 59 (2011), pp. 1371–1382.

[37] P. R. Muduli and A. Mukherjee, *A subspace projection-based joint sparse recovery method for structured biomedical signals*, IEEE Trans. Instrum. Measurem., 66 (2017), pp. 234–242.

[38] M. Nikolova, M. K. Ng, and C. P. Tam, *Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction*, IEEE Trans. Image Process., 19 (2010), pp. 3073–3088.

[39] G. Obozinski, L. Jacob, and J. P. Vert, *Group lasso with overlaps: the latent group lasso approach*, Research report. Available at https://hal.inria.fr/inria-00628498, (2011).

[40] G. Obozinski, B. Taskar, and M. I. Jordan, *Joint covariate selection and joint subspace selection for multiple classification problems*, Stat. Comput., 20 (2010), pp. 231–252.

[41] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, 3rd Edition, Springer-Verlag, Berlin, 2009.

[42] E. Soubies, L. Blanc-Féraud, and G. Aubert, *A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem*, SIAM J. Imaging Sci., 8 (2015), pp. 1607–1639.

[43] E. Soubies, L. Blanc-Fraud, and G. Aubert, *A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization*, SIAM J. Optim., 27 (2017), pp. 2034–2060.

[44] H. A. Le. Thi, H. M. Le, and T. P. Dinh, *Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm*, Mach. Learn., 101 (2015), pp. 163–186.

[45] E. Van Den Berg and M. P. Friedlander, *Probing the pareto frontier for basis pursuit solutions*, SIAM J. Sci. Comput., 31 (2008), pp. 890–912.
[46] S. Villa, L. Rosasco, S. Mosci, and A. Verri, *Proximal methods for the latent group lasso penalty*, Comput. Optim. Appl., 58 (2014), pp. 381–407.
[47] S. J. Wright, R. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[48] Z. Xu, X. Chang, F. Xu, and H. Zhang, *$l_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE T Neur. Net. Lear., 23 (2012), pp. 1013–1027.
[49] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. B, 68 (2006), pp. 49–67.
[50] T. Zhang, *Analysis of multi-stage convex relaxation for sparse regularization*, J. Mach. Learn Res., 11 (2010), pp. 1081–1107.
[51] H. Zhou, M. E. Sehl, J. S. Sinsheimer, and K. Lange, *Association screening of common and rare genetic variants by penalized regression*, Bioinformatics, 26 (2010), pp. 2375–2382.