A SMOOTHING PROXIMAL GRADIENT ALGORITHM FOR NONSMOOTH CONVEX REGRESSION WITH CARDINALITY PENALTY*

WEI BIAN† AND XIAOJUN CHEN‡

Abstract. In this paper, we focus on the constrained sparse regression problem, where the loss function is convex but nonsmooth, and the penalty term is defined by the cardinality function. Firstly, we give an exact continuous relaxation problem in the sense that both problems have the same optimal solution set. Moreover, we show that a vector is a local minimizer with the lower bound property of the original problem if and only if it is a lifted stationary point of the relaxation problem. Secondly, we propose a *smoothing proximal gradient* (SPG) algorithm for finding a lifted stationary point of the continuous relaxation model. Our algorithm is a novel combination of the classical proximal gradient algorithm and the smoothing method. We prove that the proposed SPG algorithm globally converges to a lifted stationary point of the relaxation problem, has the local convergence rate of $o(k^{-\tau})$ with $\tau \in (0, \frac{1}{2})$ on the objective function value, and identifies the zero entries of the lifted stationary point in finite iterations. Finally, we use three examples to illustrate the validity of the continuous relaxation model and good numerical performance of the SPG algorithm.

Key words. nonsmooth convex regression; cardinality penalty; proximal gradient method; smoothing method; global sequence convergence.

AMS subject classifications. 90C46, 49K35, 90C30, 65K05

1. Introduction.

For a vector $x \in \mathbb{R}^n$, denote its support set by $\mathcal{A}(x) = \{i \in \{1, \ldots, n\} : x_i \neq 0\}$, its cardinality by $|\mathcal{A}(x)|$, and its ℓ_0-norm by $\|x\|_0 = |\mathcal{A}(x)|$. We call $x \in \mathbb{R}^n$ is sparse if $|\mathcal{A}(x)| \ll n$. Sparse optimization problems emerge in many scientific and engineering problems, such as regression [52], imaging decomposition [51], visual coding [44], source separation [10], compressed sensing [12, 22], variable selection [39], etc. Sparse optimization is also the core problem of high-dimensional statistical learning [11, 24]. These problems aim to find the sparse solutions of a system of linear or nonlinear equations. The optimization model with the ℓ_0-norm penalty can improve estimation accuracy by effectively identifying the important predictors, and also enhance its interpretability. However, it is known that the ℓ_0 penalized optimization problems are NP-hard.

Under some conditions on the sensing matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ (such as the RIP and incoherence conditions), Donoho [22], and Candès, Romberg, Tao [12] proved that solving the ℓ_1 minimization can find a sparsest solution satisfying the system of linear equations $\boldsymbol{A}x = b$ with $b \in \mathbb{R}^m$. However, in 2001, Fan and Li [23] pointed out that using the ℓ_1 penalty often results in a biased estimator, and introduced a *smoothly clipped absolute deviation* (SCAD) penalty. Besides SCAD, there are many variant of continuous nonconvex penalties, such as the hard thresholding penalty [56], log-sum penalty [13], bridge ℓ_p $(0 < p < 1)$ penalty [17, 25], capped-ℓ_1 penalty [45, 47, 55] and minimax concave penalty (MCP) [54]. These continuous but nonconvex penalties would bring better sparse solutions than the ℓ_1 penalty in many cases [6, 15, 28, 31]. The estimators obtained by the SCAD, MCP and capped-ℓ_1 penalty functions satisfy the three important properties: unbiasedness, continuity in data and sparsity [23]. Meantime, there are many algorithms for solving these continuous

---

*

Funding: This work was funded in part by the NSF foundation (11871178,61773136) of China and Hong Kong Research Grant Council grant (153000/17P).

†School of Mathematics, Harbin Institute of Technology, Harbin, China; Institute of Advanced Study in Mathematics, Harbin Institute of Technology, Harbin 150001, China (bianweilvse520@163.com).

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (maxjchen@polyu.edu.hk).

nonconvex optimization problems, such as the iterative reweighted algorithm [13, 43, 36], interior point method [7], trust region method [18], cubic method [14], DC (difference of convex) function algorithm [1, 37], iterative thresholding algorithm [8], primal dual active set method [27], etc.

Despite the existing literature on the nonconvex but continuous penalties for replacing the ℓ_0-norm, some important questions still remain. First of all, the relationships between the cardinality penalty problem and its continuous relaxations are not very clear for most cases regarding the minimizers. Apart from the theoretical results for the convex ℓ_1 relaxation under restrictive hypotheses, only a few special cases have been analyzed for the consistency. With a suitable condition on the sensing matrix $\boldsymbol{A}$, the equivalence between ℓ_0 and $\ell_p(0 < p \leq 1)$ problems with constraint $\boldsymbol{A}x = b$ was proved in [25] and then this result was extended to the problem with equality and inequality constraints in [26]. In [19], the authors gave a class of smooth nonconvex penalties to approximate the ℓ_0 penalty in terms of the consistency of global minimizers. In the DC programming framework, an approximation of the ℓ_0 penalty with the consistency of global minimizers was studied in [37]. Recently, Soubies, Blanc-Féraud and Aubert proposed a continuous exact ℓ_0 (CEL0) penalty for the ℓ_2-ℓ_0 problem [51], where the global minimizers of both problems can be the same, and in [50], they verified that the capped-ℓ_1 and SCAD penalties could only guarantee the consistency of global minimizers to the ℓ_2-ℓ_0 problem, while the MCP, truncated-ℓ_p with $0 < p < 1$ and CEL0 penalties could not only own the consistence of global minimizers, but also ensure that its local minimizers are in the set of local minimizers of the ℓ_2-ℓ_0 problem. Next, due to the nonconvexity of the penalties, finding global minimizers of these nonconvex problems is often NP-hard. Most existing work for these continuous nonconvex penalized problems focuses on the stationary points in different sense [1, 7, 8, 14, 18, 31, 35, 36, 46]. Moreover, due to their nonconvexity, only the subsequence convergence to a stationary point can be proved for the proposed algorithms. The K-L (Kurdyka-Łojasiewicz) condition is a popular tool to obtain the algorithmic sequence convergence. In [2], the sequence convergence to a critical point of a class of nonconvex semi-algebra problems is established, where the K-L condition plays the key role. Most recently, the authors in [46] stated that it would be interesting whether the sequence convergence can be established to the DC problem by a given algorithm without the K-L condition on the objective function.

Denote $x^\star$ the true estimator, which is the true solution of the considered (linear or nonlinear) regression problem. Then, the oracle estimator is defined by

$$x^{\texttt{oracle}} \in \arg\min_{x_{\mathcal{A}(x^\star)^c}=\mathbf{0}} f(x), \tag{1.1}$$

where $\mathcal{A}(x^\star)^c$ means the complementary set of $\mathcal{A}(x^\star)$ and $f : \mathbb{R}^n \to [0, \infty)$ is the loss function to evaluate the regression. The oracle estimator can be used as a theoretic benchmark for comparison of computed solutions. We call that the penalized model has the oracle property if it owns a local solution having the same asymptotic distribution as the oracle estimator. The penalized problem with the SCAD, MCP or capped-ℓ_1 penalty owns the oracle property simultaneously [23, 54, 55]. A folded concave penalized problem often has multiple local solutions and the oracle property is established only for one of local solutions [24]. Hence, deriving some appealing properties, such as the optimality, sparsity or statistical properties, of the relevant stationary points is interesting. Ahn, Pang and Xin [1] established some optimality and sparsity properties of the d-stationary points (its definition will be reminded in Section 2) of the continuous relaxation problems. Fan, Xue and Zou [24] proved that as long as there is a reasonable initial estimator, an oracle estimator can be obtained via the one-step local linear approximation algorithm.

In the recent years, algorithmic research on the sparse regression problems with cardinality penalty has received much attention [4, 3, 29, 31, 32]. However, to the best of our knowledge, all the existing results are built up for the problem with a continuously differentiable loss function. The primal dual active set methods are proposed in [29, 31, 32] for the ℓ_2-ℓ_0 problems. Under some regularity conditions, such as the strict complementarity condition [31] or RIP condition on the sensing matrix [29, 32], some variants of the primal dual active set methods were proved to be convergent in finite iterations. The loss functions considered in [4, 3, 40] are continuously differentiable and with Lipschitz continuous gradients.

Our focuses and contributions. In this paper, we consider the following penalized sparse regression problem with cardinality penalty, that is,

$$\min_{x\in\mathcal{X}} \quad \mathcal{F}_{\ell_0}(x) := f(x) + \lambda\|x\|_0, \tag{1.2}$$

where $\mathcal{X} = \{x \in \mathbb{R}^n : l \leq x \leq u\}$, $f : \mathbb{R}^n \to [0, \infty)$ is convex (not necessarily smooth), λ is a positive parameter, and l, $u \in \{\mathbb{R}, \pm\infty\}^n$ with $l \leq 0 \leq u$ and $l < u$.

One application of problem (1.2) comes from the linear regression problem. It is well-known that the least squares estimate with the ℓ_2-ℓ_0 model is not robust for many cases [23]. We need to consider the problem with the outlier-resistant loss function, such as the ℓ_1 loss function given by

$$f(x) = \frac{1}{m}\left\|\boldsymbol{A}x - b\right\|_1, \tag{1.3}$$

or Huber's functions [30], which are convex, but not smooth. Another important application of problem (1.2) comes from the censored regression problem with the nonsmooth convex loss function

$$f(x) = \frac{1}{pm}\sum_{i=1}^{m} \left|\max\{A_i x - c_i, 0\} - b_i\right|^p, \tag{1.4}$$

where $p \in [1, 2]$, $A_i^T \in \mathbb{R}^n$ and $c_i, b_i \in \mathbb{R}$, $i = 1, \ldots, m$. There are some other nonsmooth convex loss functions, for example the negative log-quasi-likelihood function [23] or the check loss function in penalized quantile regression [24, 33]. To the best of our knowledge, only little work has been dedicated to the penalized sparse regression problem (1.2) with a general convex loss function.

For a given parameter $\nu > 0$, let $\Phi(x) = \sum_{i=1}^n \phi(x_i)$ be a continuous relaxation of the ℓ_0 penalty with the capped-ℓ_1 function ϕ given by

$$\phi(t) = \min\{1, |t|/\nu\}. \tag{1.5}$$

We consider the following Lipschitz continuous optimization problem for solving (1.2):

$$\min_{x\in\mathcal{X}} \quad \mathcal{F}(x) := f(x) + \lambda\Phi(x). \tag{1.6}$$

Differently from the previous work [1, 4, 3, 7, 8, 14, 18, 29, 31, 32, 35, 36, 46], this paper considers the original cardinality penalty problem with a continuous convex loss function and uses an exact continuous relaxation problem to solve it. In particular, we focus on problem (1.2) with a continuous convex loss function, which is nonsmooth or whose gradient is not Lipschitz continuous. The main contributions of this paper include the following two aspects. First, we prove that the continuous relaxation problem (1.6) with certain $\nu > 0$ has two advantages: global minimizers of (1.2) and (1.6) are same; any lifted stationary point of (1.6) (its definition will be reminded in Section 2) is a local minimizer of (1.2) with a desired lower bound property. Second, we propose a smoothing proximal gradient (SPG) algorithm with global sequence convergence to a lifted stationary point of

(1.6) without using the K-L condition. Moreover, the SPG algorithm owns a local convergence rate on the objective function value of (1.6) and the finite iterative identification for the zero entries of the limit point.

Notations. We denote $\mathbb{N} = \{0, 1, \dots\}$ and $\mathbb{D}^n = \{d \in \mathbb{R}^n : d_i \in \{1, 2, 3\},\ i = 1, \dots, n\}$. For $x \in \mathbb{R}^n$ and $\delta > 0$, let $\|x\| := \|x\|_2$ and $\mathbb{B}_\delta(x)$ means the open ball centered at x with radius δ. For a nonempty, closed and convex set $\mathcal{X} \subseteq \mathbb{R}^n$, $N_{\mathcal{X}}(x)$ means the normal cone to $\mathcal{X}$ at $x \in \mathcal{X}$. Let $\mathbf{1}_n \in \mathbb{R}^n$ be the all-ones vector and $\boldsymbol{e}_i \in \mathbb{R}^n$ be the ith column of the n dimensional identity matrix. For a locally Lipschitz continuous function $\psi : \mathbb{R}^n \to \mathbb{R}$, we denote $\partial\psi(x)$ the Clarke subgradient [20] of ψ at $x \in \mathbb{R}^n$.

2. An exact continuous relaxation for (1.2).

In this section, we analyze the relationships between (1.2) and (1.6), where the capped-ℓ_1 penalty can let problem (1.6) own the oracle property and then can be seen as one of the best continuous relaxations to the ℓ_0-norm penalty [45].

Assumption 1. *f is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant L_f.*

Assumption 2. *Positive parameter ν in (1.5) satisfies $\nu < \bar{\nu} := \lambda/L_f$.*

If there is no special explanation, we suppose Assumption 1 and Assumption 2 hold throughout the paper, and assume that L_f is large enough such that $L_f \geq \frac{\lambda}{\Gamma}$, where

$$\Gamma := \min\{|l_i|, u_j : l_i \neq 0, u_j \neq 0, i = 1, \dots, n, j = 1, \dots, n\}.$$

When f is defined by the ℓ_1 loss function or the loss function in (1.4) with $p = 1$, we can let $L_f = \max\{\|\boldsymbol{A}\|_\infty, \frac{\lambda}{\Gamma}\}$.

2.1. Lifted stationary points of (1.6).

Though ϕ is piecewise linear, problem (1.6) is still a nonconvex optimization problem. It has been proved in [6] that finding a global minimizer of (1.6) is NP-hard in general. Note that ϕ in (1.5) can be reformulated as a DC function, i.e.

$$\phi(t) = \frac{1}{\nu}|t| - \max\{\theta_1(t), \theta_2(t), \theta_3(t)\}$$

with $\theta_1(t) = 0$, $\theta_2(t) = t/\nu - 1$ and $\theta_3(t) = -t/\nu - 1$. For $t \in \mathbb{R}$, denote

$$\mathcal{D}(t) = \{i \in \{1, 2, 3\} : \theta_i(t) = \max\{\theta_1(t), \theta_2(t), \theta_3(t)\}\}. \tag{2.1}$$

Definition 2.1. *[46] We say that $x \in \mathcal{X}$ is a* lifted stationary point *of (1.6) if there exist $d_i \in \mathcal{D}(x_i)$ for $i = 1, \dots, n$, such that*

$$\lambda \sum_{i=1}^n \theta'_{d_i}(x_i)\boldsymbol{e}_i \in \partial f(x) + \frac{\lambda}{\nu}\partial\left(\sum_{i=1}^n |x_i|\right) + N_{\mathcal{X}}(x). \tag{2.2}$$

If (2.2) holds for all $d_i \in \mathcal{D}(x_i)$, $\forall i = 1, \dots, n$, then we call x a d-stationary point [46]. Due to the piecewise linearity of $\max\{\theta_1(t), \theta_2(t), \theta_3(t)\}$, x is a d-stationary point of (1.6) if and only if it is a local minimizer. Recall that $\bar{x}$ is a limiting stationary point [48] of (1.6), if

$$0 \in \bar{\partial}(f + \lambda\Phi)(\bar{x}) + N_{\mathcal{X}}(\bar{x}), \tag{2.3}$$

where "$\bar{\partial}$" indicates the limiting subgradient. And $\bar{x}$ is a Clarke stationary point of (1.6), i $0 \in \partial(f+\lambda\Phi)(\bar{x})+N_{\mathcal{X}}(\bar{x})$. We call $\bar{x} \in \mathcal{X}$ a critical point of (1.6) if it satisfies $0 \in \partial f(\bar{x})+\lambda\partial\Phi(\bar{x})+N_{\mathcal{X}}(\bar{x})$. It holds that

$$\mathcal{S}_d \subseteq \mathcal{S}_{lim} \subseteq \mathcal{S}_{lif} \subseteq \mathcal{S}_{cl} \subseteq \mathcal{S}_{cr},$$

but their inverse may not hold, where $\mathcal{S}_d$, $\mathcal{S}_{lim}$, $\mathcal{S}_{lif}$, $\mathcal{S}_{cl}$ and $\mathcal{S}_{cr}$ denote the d-stationary point set, limiting stationary point set, lifted stationary point set, Clarke stationary point set and critical point set of (1.6), respectively.

A natural question arises why we focus on the lifted stationary points rather than the others. First, the lifted stationary points satisfy a sharper optimal necessary condition than the Clarke and critical stationary points. Second, the d-stationary and limiting stationary points of (1.6) are difficult to be computed. Though Pang, Razaviyayn and Alvarado [46] developed a novel algorithm for computing a d-stationary point of the DC optimization problems, the algorithm in [46] cannot be directly used to solve problem (1.6).

2.2. Characterizations of lifted stationary points of (1.6). With the computable condition on ν defined in Assumption 2, we first verify that the element in $\Pi_{i=1}^n \mathcal{D}(x_i)$ for a lifted stationary point satisfying (2.2) is unique and well-defined.

PROPOSITION 2.2. *If $\bar{x}$ is a lifted stationary point of (1.6), then the vector $d^{\bar{x}} = (d_1^{\bar{x}}, \ldots, d_n^{\bar{x}})^T \in \prod_{i=1}^n \mathcal{D}(\bar{x}_i)$ satisfying (2.2) is unique. In particular, for $i = 1, \ldots, n$,*

$$d_i^{\bar{x}} = \begin{cases} 1 & \text{if } |\bar{x}_i| < \nu, \\ 2 & \text{if } \bar{x}_i \geq \nu, \\ 3 & \text{if } \bar{x}_i \leq -\nu. \end{cases} \tag{2.4}$$

Proof. If $|\bar{x}_i| \neq \nu$, then the statement in this proposition holds naturally. Hence, we only need to consider the case $|\bar{x}_i| = \nu$. When $\bar{x}_i = \nu$, since $\mathcal{D}(\bar{x}_i) = \{1, 2\}$, arguing by contradiction, we assume (2.2) holds with $d_i^{\bar{x}} = 1$. By $\nu < \bar{\nu}$, we have $\bar{x}_i \in (l_i, u_i)$, and by (2.2), there exists $\xi(\bar{x}) \in \partial f(\bar{x})$ such that $0 = \xi_i(\bar{x}) + \lambda/\nu$, which implies $\lambda/\nu = |\xi_i(\bar{x})| \leq L_f$. This leads to a contradiction to $\nu < \lambda/L_f$. Then, (2.4) holds for $\bar{x}_i = \nu$. Similar analysis can be given for the case that $\bar{x}_i = -\nu$, which completes the proof. ∎

For a given $d = (d_1, \ldots, d_n)^T \in \mathbb{D}^n$, we define

$$\Phi^d(x) := \sum_{i=1}^n |x_i|/\nu - \sum_{i=1}^n \theta_{d_i}(x_i), \tag{2.5}$$

which is convex with respect to x. It can be verified that $\Phi(x) = \min_{d \in \mathbb{D}^n} \Phi^d(x)$, $\forall x \in \mathcal{X}$. In particular, for a fixed $\bar{x} \in \mathcal{X}$, $\Phi(\bar{x}) = \Phi^{d^{\bar{x}}}(\bar{x})$ with $d^{\bar{x}}$ defined in (2.4).

Remark 2.1. *Proposition 2.2 implies that $\bar{x}$ is a local minimizer of (1.6) if and only if $\bar{x}$ is a lifted stationary point of (1.6) and $|\bar{x}_i| \neq \nu$, $\forall i = 1, \ldots, n$. Moreover, due to the convexity of $f(x) + \lambda\Phi^d(x)$ and the linearity of $\sum_{i=1}^n \theta_{d_i}(x_i)$ for a fixed $d \in \mathbb{D}^n$, the assertion in Proposition 2.2 implies the following equivalent results:*

$$\bar{x} \text{ is a lifted stationary point of (1.6)} \Leftrightarrow \text{(2.2) holds at } \bar{x} \in \mathcal{X} \text{ with } d = d^{\bar{x}} \text{ defined in (2.4)}$$

$$\Leftrightarrow \bar{x} \in \arg\min\nolimits_{x \in \mathcal{X}} f(x) + \lambda\Phi^{d^{\bar{x}}}(x) \tag{2.6}$$

$$\Leftrightarrow \bar{x} \in \arg\min\nolimits_{x \in \mathcal{X}, d^x = d^{\bar{x}}} f(x) + \lambda\Phi(x), \tag{2.7}$$

where the last equivalence uses $\Phi^{d^{\bar{x}}}(\bar{x}) = \Phi(\bar{x})$ and $\Phi^{d^{\bar{x}}}(x) \geq \Phi(x)$, $\forall x \in \mathbb{R}^n$.

We then show a lower bound property of the lifted stationary points of (1.6).

LEMMA 2.3. *If $\bar{x} \in \mathcal{X}$ is a lifted stationary point of (1.6), then it holds that*

$$\bar{x}_i \in (-\nu, \nu) \quad \Rightarrow \quad \bar{x}_i = 0, \qquad \forall i = 1, \ldots, n. \tag{2.8}$$

Proof. Suppose $\bar{x}$ is a lifted stationary point of (1.6). Assume that $\bar{x}_i \in (-\nu, \nu)\backslash\{0\}$ for some $i \in \{1, \ldots, n\}$. Then, $d_i^{\bar{x}} = 1$ and $\bar{x}_i \in (l_i, u_i)$. By Definition 2.1, there exists $\xi(\bar{x}) \in \partial f(\bar{x})$ such that $\xi_i(\bar{x}) + (\lambda/\nu)\text{sign}(\bar{x}_i) = 0$. Then, $\lambda/\nu = |\xi_i(\bar{x})| \leq \|\xi(\bar{x})\| \leq L_f$, which leads to a contradiction to $\nu < \lambda/L_f$. Thus, for any $i \in \{1, \ldots, n\}$, $\bar{x}_i \in (-\nu, \nu)$ implies $\bar{x}_i = 0$. ◻

Remark 2.2. *On the one hand, if f is not continuously differentiable on $\mathcal{X}_\nu = \{x \in \mathcal{X} : |x_i| = \nu$ for some $i \in \{1, \ldots, n\}\}$, a lifted stationary point of (1.6) is not necessary to be a Clarke stationary point [46]. On the other hand, if f is continuously differentiable on $\mathcal{X}_\nu$, then $\bar{x}$ is a lifted stationary point of (1.6) if and only if it is a limiting stationary point, but is not necessary to be a Clarke stationary point. A counterexample can be provided by setting $f(x) = (x_1 + x_2 - 1)^2$, $l = (0, 0)^T$, $u = (1, 1)^T$, $\lambda = 1$ and $\nu = 0.2$ in (1.6), where $\nu < \bar{\nu} = 0.25$. It follows from Lemma 2.3 that $\mathcal{S}_{\rm cl} = \mathcal{S}_{\rm lif} \bigcup \{(0, 0.2)^T, (0.2, 0)^T\}$, where $\mathcal{S}_{\rm lif} = \{x \in \mathbb{R}^2 : x_1 + x_2 = 1, x_1 \geq 0.2, x_2 \geq 0.2\} \bigcup \{(0, 0)^T, (1, 0)^T, (0, 1)^T\}$.*

2.3. Links between (1.2) and (1.6). The goal of this subsection is to study the links between the ℓ_0 penalized minimization problem (1.2) and its continuous relaxation (1.6). In light of the lower bound characterization of the lifted stationary points of (1.6) given in Lemma 2.3, we show the links between (1.2) and (1.6) by the two following results, where the first result focuses on global minimizers, and the second is on local minimizers.

THEOREM 2.4. *$\bar{x} \in \mathcal{X}$ is a global minimizer of (1.2) if and only if it is a global minimizer of (1.6). Moreover, problems (1.2) and (1.6) have the same optimal value.*

Proof. First, let $\bar{x} \in \mathcal{X}$ be a global minimizer of (1.6), then $\bar{x}$ is a lifted stationary point of (1.6). By (2.8), it gives $\Phi(\bar{x}) = \|\bar{x}\|_0$. Then,

$$f(\bar{x}) + \lambda\|\bar{x}\|_0 = f(\bar{x}) + \lambda\Phi(\bar{x}) \leq f(x) + \lambda\Phi(x) \leq f(x) + \lambda\|x\|_0, \quad \forall x \in \mathcal{X},$$

where the last inequality uses $\Phi(x) \leq \|x\|_0$, $\forall x \in \mathbb{R}^n$. Thus, $\bar{x}$ is a global minimizer of (1.2).

Next, suppose $\bar{x} \in \mathcal{X}$ is a global minimizer of (1.2) but not a global minimizer of (1.6). Then there exists a global minimizer of (1.6) denoted by $\hat{x}$ such that

$$f(\hat{x}) + \lambda\Phi(\hat{x}) < f(\bar{x}) + \lambda\Phi(\bar{x}).$$

From $\Phi(\hat{x}) = \|\hat{x}\|_0$ and $\Phi(\bar{x}) \leq \|\bar{x}\|_0$, we get $f(\hat{x}) + \lambda\|\hat{x}\|_0 < f(\bar{x}) + \lambda\|\bar{x}\|_0$, which leads to a contradiction. Thus, any global minimizer of (1.2) must be a global minimizer of (1.6). Hence, using Lemma 2.3, we ensure that problems (1.2) and (1.6) have the same optimal value. ◻

Theorem 2.4 provides that problems (1.2) and (1.6) have the same global solution set. The following proposition and the subsequent example show that this is not always true for their local minimizers.

PROPOSITION 2.5. *If $\bar{x}$ is a lifted stationary point of (1.6), then it is a local minimizer of (1.2) and the objective functions have the same value at $\bar{x}$, i.e. $\mathcal{F}_{\ell_0}(\bar{x}) = \mathcal{F}(\bar{x})$.*

Proof. Coming back to the definition of $\Phi^{d^{\bar{x}}}$ defined in (2.5) and from the lower bound property of $\bar{x}$ in (2.8), for any $x \in \mathbb{R}^n$, we have

$$\Phi^{d^{\bar{x}}}(x) = \sum_{i=1}^n |x_i|/\nu - \sum_{i=1}^n \theta_{d_i^{\bar{x}}}(x_i) = \sum_{i:|\bar{x}_i| \geq \nu} 1 + \sum_{i:|\bar{x}_i| < \nu} |x_i|/\nu = \|\bar{x}\|_0 + \sum_{i:\bar{x}_i = 0} |x_i|/\nu.$$

Then, there exists $\varrho > 0$ such that $\Phi^{d^{\bar{x}}}(x) \le \|x\|_0$, $\forall x \in \mathbb{B}_\varrho(\bar{x})$. Combining this with $\Phi(x) \le \|x\|_0$ and (2.6) gives

$$f(\bar{x}) + \lambda\|\bar{x}\|_0 \le f(x) + \lambda\|x\|_0, \quad \forall x \in \mathcal{X} \cap \mathbb{B}_\varrho(\bar{x}).$$

Thus, $\bar{x}$ is a local minimizer of (1.2). ◻

Proposition 2.5 states that any lifted stationary point of (1.6) is a local minimizer of (1.2), which implies that any local minimizer of (1.6) is certainly a local minimizer of (1.2). Due to the special structure of the cardinality norm, any minimizer of $\min_{x\in\mathcal{X}} f(x)$ is a local minimizer of (1.2). The following example shows that a lifted stationary point of (1.6) is a local minimizer of (1.2) with the lower bound property in (2.8) and is likely a global minimizer.

Example 2.1. *Let problem (1.2) be in the form of*

$$\min_{0\le x_1,x_2\le 1} \mathcal{F}_{\ell_0}(x_1,x_2) := |x_1 + x_2 - 1| + \lambda\|x\|_0. \tag{2.9}$$

We can easily find that $\mathcal{LM} := \{x \in \mathbb{R}^2 : x_1 + x_2 = 1, 0 \le x_1, x_2 \le 1\} \cup \{(0,0)^T\}$ is the set of local minimizers of (2.9). Moreover, $(0,0)^T$ is the unique global minimizer when $\lambda > 1$, the global minimizers are $\{(0,1)^T,(1,0)^T\}$ when $\lambda < 1$, and the global minimizers are $\{(0,1)^T,(1,0)^T,(0,0)^T\}$ when $\lambda = 1$. Here, $\bar{\nu}$ in Lemma 2.3 can be $\min\{\sqrt{2}\lambda/2, 1\}$. With $\nu < \min\{\sqrt{2}\lambda/2, 1\}$, the lifted stationary points of (1.6) for this example are $\{x \in \mathbb{R}^2 : x_1{+}x_2 = 1, \nu \le x_1, x_2 \le 1\} \bigcup \{(0,0)^T,(1,0)^T,(0,1)^T\}$, which is a proper subset of $\mathcal{LM}$. Specially, if $\sqrt{2}/2 < \lambda \le 1$ and $1/2 < \nu < \min\{\sqrt{2}\lambda/2, 1\}$, the lifted stationary points of (1.6) are $\{(1,0)^T,(0,1)^T,(0,0)^T\}$.

When f is convex, $\bar{x}$ is a local minimizer of (1.2) if and only if $\bar{x} \in \mathcal{X}$ satisfies

$$0 \in [\partial f(\bar{x}) + N_{\mathcal{X}}(\bar{x})]_i, \quad \forall i \in \mathcal{A}(\bar{x}), \tag{2.10}$$

which is a criterion for the local minimizers of (1.2) [40]. From Lemma 2.3 and Theorem 2.4, we find that the lower bound property in (2.8) holds for any global minimizer of (1.2), but is not true for all of its local minimizers. This inspires us to define a class of strong local minimizers of (1.2) by combining the optimality condition in (2.10) and the lower bound property in (2.8).

Definition 2.6. *We call $\bar{x} \in \mathcal{X}$ a ν-strong local minimizer of (1.2), if there exist $\bar{\xi} \in \partial f(\bar{x})$ and $\bar{\eta} \in N_{\mathcal{X}}(\bar{x})$ such that for any $i \in \mathcal{A}(\bar{x})$, it holds*

$$\bar{\xi}_i + \bar{\eta}_i = 0 \quad \text{and} \quad |\bar{x}_i| \ge \nu.$$

By (2.10), any ν-strong local minimizer of (1.2) is a local minimizer of it. To close this section, we give a result on the relationship between the ν-strong local minimizers of (1.2) and the lifted stationary points of (1.6).

Proposition 2.7. *$\bar{x} \in \mathcal{X}$ is a ν-strong local minimizer of (1.2) if and only if it is a lifted stationary point of (1.6). Moreover, if $\bar{x} \in \mathcal{X}$ is a ν-strong local minimizer of (1.2), then it holds*

$$\mathcal{F}_{\ell_0}(\bar{x}) \le \mathcal{F}_{\ell_0}(x), \quad \forall x \in \mathcal{X} \cap (\bar{x} - \nu e, \bar{x} + \nu e),$$

$$f(\bar{x}) \le f(x), \quad \forall x \in \{x \in \mathcal{X} : \mathcal{A}(x) \subseteq \mathcal{A}(\bar{x})\}, \tag{2.11}$$

$$\bar{x} \text{ is an oracle solution defined in (1.1) if } \mathcal{A}(\bar{x}) = \mathcal{A}(x^\star). \tag{2.12}$$

Proof. From Lemma 2.3, we can easily verify the first statement. By (2.6), we see that if $\bar{x}$ is a lifted stationary point of (1.6), then

$$\mathcal{F}_{\ell_0}(\bar{x}) = f(\bar{x}) + \lambda\|\bar{x}\|_0 = f(\bar{x}) + \lambda\Phi(\bar{x}) = f(\bar{x}) + \lambda\Phi^{d^{\bar{x}}}(\bar{x}) \le f(x) + \lambda\Phi^{d^{\bar{x}}}(x), \quad \forall x \in \mathcal{X}.$$

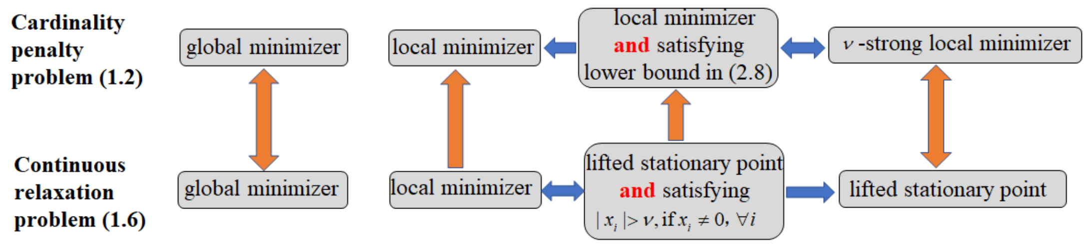


Fig. 2.1: Links between problems (1.2) and (1.6)

Due to Lemma 2.3, we then have $\mathcal{F}_{\ell_0}(\bar{x}) \leq \mathcal{F}_{\ell_0}(x)$, $\forall x \in \mathcal{X} \cap (\bar{x} - \nu \mathbf{1}_n, \bar{x} + \nu \mathbf{1}_n)$, which holds from $\Phi^{d^{\bar{x}}}(x) \leq \|x\|_0$, $\forall x \in (\bar{x} - \nu \mathbf{1}_n, \bar{x} + \nu \mathbf{1}_n)$. Recalling (2.6) again, we obtain $f(\bar{x}) \leq f(x) + \lambda \sum_{i \notin \mathcal{A}(\bar{x})} |x_i|/\nu$, $\forall x \in \mathcal{X}$. If $\mathcal{A}(x) \subseteq \mathcal{A}(\bar{x})$, then $x_i = 0$ for $i \notin \mathcal{A}(\bar{x})$. Hence, (2.11) holds, which immediately implies (2.12). ◻

Remark 2.3. *In [50], the authors gave a unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization, which derives necessary and sufficient conditions on ℓ_0 continuous relaxations such that each (local and global) minimizer of the underlying relaxation is also a minimizer of the ℓ_2-ℓ_0 problem. However, the property that any local minimizer of the relaxation problem with the capped-ℓ_1 penalty is a local minimizer of the ℓ_2-ℓ_0 problem cannot be verified by the results in [50]. In this paper, we prove this property for the capped-ℓ_1 penalty by its lifted stationary points.*

To end this section, we use Fig. 2.1 to give a brief description on the links between problems (1.2) and (1.6) when $\nu < \bar{\nu}$.

3. Numerical Algorithm and its convergence analysis. In this section, we focus on the numerical algorithm for finding a lifted stationary point of (1.6), which is a ν-strong local minimizer of (1.2). The first two subsections briefly introduce some useful preliminary results on smoothing methods and the proximal gradient algorithm, the third subsection presents a new proximal gradient algorithm combined with the smoothing method, and the last two subsections show the convergence of the proposed algorithm for solving (1.6).

3.1. Smoothing approximation method. A well-known method for solving nonsmooth optimization problems is to approximate the original problem by a sequence of smooth problems, which own rich theory and powerful numerical algorithms [42]. For the sake of completeness, we formally define a class of smoothing functions for f in (1.6).

Definition 3.1. *We call $\tilde{f} : \mathbb{R}^n \times [0, \bar{\mu}] \to \mathbb{R}$ with $\bar{\mu} > 0$ a smoothing function of the convex function f in (1.6), if $\tilde{f}(\cdot, \mu)$ is continuously differentiable in $\mathbb{R}^n$ for any fixed $\mu > 0$ and satisfies the following conditions:*

(i) $\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) = f(x)$, $\forall x \in \mathcal{X}$;

(ii) *(convexity)* $\tilde{f}(x, \mu)$ *is convex with respect to* x *in* $\mathcal{X}$ *for any fixed* $\mu > 0$;

(iii) *(gradient consistency)* $\{\lim_{z \to x, \mu \downarrow 0} \nabla_z \tilde{f}(z, \mu)\} \subseteq \partial f(x)$, $\forall x \in \mathcal{X}$;

(iv) *(Lipschitz continuity with respect to μ) there exists a positive constant κ such that*

$$|\tilde{f}(x, \mu_2) - \tilde{f}(x, \mu_1)| \leq \kappa |\mu_1 - \mu_2|, \quad \forall x \in \mathcal{X},\ \mu_1, \mu_2 \in [0, \bar{\mu}];$$

(v) *(Lipschitz continuity with respect to x) there exists a constant $L > 0$ such that for any*

$\mu \in (0, \bar{\mu}]$, $\nabla_x \tilde{f}(\cdot, \mu)$ is Lipschitz continuous on $\mathcal{X}$ with Lipschitz constant $L\mu^{-1}$.

Throughout this paper, we denote $\tilde{f}$ a smoothing function of f in (1.6). When it is clear from the context, the derivative of $\tilde{f}(x, \mu)$ with respect to x is simply denoted as $\nabla \tilde{f}(x, \mu)$. Definition 3.1-(iv) implies

$$|\tilde{f}(x, \mu) - f(x)| \le \kappa\mu, \quad \forall x \in \mathcal{X},\ 0 < \mu \le \bar{\mu}. \tag{3.1}$$

Example 3.1. *Many existing results in [16, 34, 49] give us some theoretical basis for constructing smoothing functions satisfying the conditions in Definition 3.1. A smoothing function of the ℓ_1 loss function in (1.3) can be defined by*

$$\tilde{f}(x, \mu) = \frac{1}{m} \sum_{i=1}^{m} \tilde{\theta}(A_i x - b_i, \mu) \quad \textit{with} \quad \tilde{\theta}(s, \mu) = \begin{cases} |s| & \textit{if } |s| > \mu, \\ \dfrac{s^2}{2\mu} + \dfrac{\mu}{2} & \textit{if } |s| \le \mu. \end{cases} \tag{3.2}$$

For the loss function in (1.4) with $p = 1$, a smoothing function of it can be defined by

$$\tilde{f}(x, \mu) = \frac{1}{m} \sum_{i=1}^{m} \tilde{\theta}(\tilde{\phi}(A_i x, \mu) - b_i, \mu) \quad \textit{with} \quad \tilde{\phi}(s, \mu) = \begin{cases} \max\{s, 0\} & \textit{if } |s| > \mu, \\ \dfrac{(s+\mu)^2}{4\mu} & \textit{if } |s| \le \mu. \end{cases} \tag{3.3}$$

We end this subsection by giving the following notations:

$$\tilde{\mathcal{F}}^d(x, \mu) \triangleq \tilde{f}(x, \mu) + \lambda \Phi^d(x) \quad \text{and} \quad \tilde{\mathcal{F}}(x, \mu) \triangleq \tilde{f}(x, \mu) + \lambda \Phi(x),$$

where $\tilde{f}$ is a smoothing function of f, $\mu > 0$ and $d \in \mathbb{D}^n$. For any fixed $\mu > 0$ and $d \in \mathbb{D}^n$, both $\tilde{\mathcal{F}}^d(x, \mu)$ and $\tilde{\mathcal{F}}(x, \mu)$ are nonsmooth, $\tilde{\mathcal{F}}^d(x, \mu)$ is convex, but $\tilde{\mathcal{F}}(x, \mu)$ is nonconvex. Moreover,

$$\tilde{\mathcal{F}}^d(x, \mu) \ge \tilde{\mathcal{F}}(x, \mu), \quad \forall d \in \mathbb{D}^n,\ x \in \mathcal{X},\ \mu \in (0, \bar{\mu}]. \tag{3.4}$$

3.2. Proximal gradient method. In this subsection, we consider the following constrained convex optimization problem with given smoothing parameter $\mu > 0$ and vector $d \in \mathbb{D}^n$

$$\min_{x \in \mathcal{X}} \quad \tilde{\mathcal{F}}^d(x, \mu). \tag{3.5}$$

It is good news that, for any given vectors $d \in \mathbb{D}^n$, $w \in \mathbb{R}^n$ and a positive number $\tau > 0$, the proximal operator of $\tau\Phi^d$ on $\mathcal{X}$ has a closed form solution, i.e.

$$\hat{x} = \arg\min_{x \in \mathcal{X}} \left\{ \tau \Phi^d(x) + \frac{1}{2}\|x - w\|^2 \right\} \tag{3.6}$$

can be calculated by $\hat{x}_i = \min\{\max\{l_i, y_i\}, u_i\}$ for $i = 1, \ldots, n$, where

$$y_i = \begin{cases} 0 & \text{if } |\bar{w}_i| \le \tau/\nu, \\ \bar{w}_i - \tau/\nu & \text{if } \bar{w}_i > \tau/\nu, \\ \bar{w}_i + \tau/\nu & \text{if } \bar{w}_i < -\tau/\nu, \end{cases} \tag{3.7}$$

with $\bar{w}_i = w_i$ for $d_i = 1$, $\bar{w}_i = w_i + \tau/\nu$ for $d_i = 2$ and $\bar{w}_i = w_i - \tau/\nu$ for $d_i = 3$. Toward this end, we consider an approximation of $\tilde{\mathcal{F}}^d(\cdot,\mu)$ around a given point z, given by

$$Q_{d,\gamma}(x,z,\mu) = \tilde{f}(z,\mu) + \langle x - z, \nabla \tilde{f}(z,\mu)\rangle + \frac{1}{2}\gamma\mu^{-1}\|x - z\|^2 + \lambda\Phi^d(x) \tag{3.8}$$

with a constant $\gamma > 0$. Since $\Phi^d(x)$ is convex with respect to x for any fixed $d \in \mathbb{D}^n$, function $Q_{d,\gamma}(x,z,\mu)$ is a strongly convex function with respect to x for any fixed d, γ, z and μ. Then, minimization problem $\min_{x\in\mathcal{X}} Q_{d,\gamma}(x,z,\mu)$ admits a unique minimizer, denoted by $\hat{x}$, which can be calculated by (3.7) with $\tau = \lambda\gamma^{-1}\mu$ and $w = z - \gamma^{-1}\mu\nabla\tilde{f}(z,\mu)$.

3.3. Smoothing proximal gradient (SPG) algorithm. In this subsection, we propose a new algorithm for finding a lifted stationary point of (1.6). Since the proposed algorithm combines the smoothing method and the proximal gradient algorithm, we call it *Smoothing Proximal Gradient* (SPG) algorithm.

For convenience of further reading, we begin this subsection by emphasizing the following assumptions needed in the convergence analysis of the SPG algorithm.

- (A1) Assumption 1 and Assumption 2 hold.
- (A2) $\tilde{f}$ is a smoothing function of f defined in Definition 3.1.
- (A3) $\mathcal{F}$ in (1.6) (or $\mathcal{F}_{\ell_0}$ in (1.2)) is level bounded on $\mathcal{X}$[1].

As the feasible region $\mathcal{X}$ is bounded, then assumption (A3) holds naturally. We give some more details on the parameters in these assumptions. Parameter L_f in Assumption 1 is used to define ν such that problems (1.2) and (1.6) have the consistency in Theorem 2.4 and Proposition 2.5. Parameter κ in Definition 3.1 is used in the SPG algorithm, which can be calculated exactly for most smoothing functions [16] and $\kappa = \frac{1}{2}$ for the smoothing functions in (3.2) and (3.3). The value of L in Definition 3.1 is not necessary and we will use a simple line search method to find an acceptable value at each iteration of the SPG algorithm. Upon the above assumptions, we present the SPG algorithm for solving (1.6). See Algorithm 3.1.

At each iteration, this algorithm takes the proximal gradient algorithm for solving (3.5) with fixed μ_k, γ_k and d^k, and uses a simple criterion for updating μ_k. The values of γ_k are chosen independently in Step 1 of each iteration. Step 3 updates the smoothing parameter μ_k by using (3.12), where $\tilde{\mathcal{F}}(x^{k+1},\mu_k)+\kappa\mu_k$ can be seen as an energy function and its monotone non-increasing property will be proved in Lemma 3.3. If the energy function is decreased more than the given scale at the current iteration, then the current smoothing parameter is still acceptable for the next iteration, otherwise we reduce its value by the updating rule in (3.13) for the next iteration. Let

$$\mathcal{N}^s = \{k \in \mathbb{N} : \mu_{k+1} \neq \mu_k\},$$

and denote n_r^s the rth smallest number in $\mathcal{N}^s$. Then, we can obtain following updating method of $\{\mu_k\}$

$$\mu^k = \mu^{n_r^s+1} = \frac{\mu_0}{(n_r^s+1)^\sigma}, \qquad \forall n_r^s + 1 \le k \le n_{r+1}^s, \tag{3.14}$$

which will be used in the proof of Lemma 3.2 and Lemma 3.5.

[1]We call function $\mathcal{F}$ is level bounded on $\mathcal{X}$, if for any $\Gamma > 0$, the level set $\{x \in \mathcal{X} : \mathcal{F}(x) \le \Gamma\}$ is bounded.

Algorithm 3.1 Smoothing Proximal Gradient (SPG) algorithm

Input: Take initial iterates $x^{-1} = x^0 \in \mathcal{X}$ and $\mu_{-1} = \mu_0 \in (0, \bar{\mu}]$. Choose constants $\rho > 1$, $\sigma \in (\frac{1}{2}, 1)$, $\alpha > 0$ and $0 < \underline{\gamma} \le \bar{\gamma}$. Set $k = 0$.

While a termination criterion is not met, **do**

Step 1. Choose $\gamma_k \in [\underline{\gamma}, \bar{\gamma}]$ and let $d^k \triangleq d^{x^k}$, where d^{x^k} is defined in (2.4).

Step 2. 2a) Compute

$$\hat{x}^{k+1} = \arg\min_{x \in \mathcal{X}} Q_{d^k,\gamma_k}(x, x^k, \mu_k). \tag{3.9}$$

2b) If $\hat{x}^{k+1}$ satisfies

$$\tilde{\mathcal{F}}^{d^k}(\hat{x}^{k+1}, \mu_k) \le Q_{d^k,\gamma_k}(\hat{x}^{k+1}, x^k, \mu_k), \tag{3.10}$$

set

$$x^{k+1} = \hat{x}^{k+1} \tag{3.11}$$

and go to **Step 3**. Otherwise, let $\gamma_k = \rho\gamma_k$ and return to 2a).

Step 3: If

$$\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k - \tilde{\mathcal{F}}(x^k, \mu_{k-1}) - \kappa\mu_{k-1} \le -\alpha\mu_k^2, \tag{3.12}$$

set $\mu_{k+1} = \mu_k$, otherwise, set

$$\mu_{k+1} = \frac{\mu_0}{(k+1)^\sigma}. \tag{3.13}$$

Increment k by one and return to **Step 1**.

end while

3.4. Basic convergence analysis of the SPG algorithm. Denote $\{x^k\}$, $\{\gamma_k\}$ and $\{\mu_k\}$ be the sequences generated by the SPG algorithm. In this subsection, we first establish some basic properties of the iterates $\{x^k\}$, $\{\gamma_k\}$ and $\{\mu_k\}$ in Lemma 3.2. Then, by the level boundedness assumption of $\mathcal{F}$ (or $\mathcal{F}_{\ell_0}$) on $\mathcal{X}$, the boundedness of $\{x^k\}$ is obtained in Lemma 3.3. At last, the subsequential convergence of $\{x^k : k \in \mathcal{N}^s\}$ to a lifted stationary point of (1.6) is established in Proposition 3.4.

LEMMA 3.2. *The proposed SPG algorithm is well-defined, and the sequences $\{x^k\}$, $\{\gamma_k\}$ and $\{\mu_k\}$ generated by it own the following properties:*

(i) $\{x^k\} \subseteq \mathcal{X}$ *and* $\{\gamma_k\} \subseteq \left[\underline{\gamma}, \max\{\bar{\gamma}, \rho L\}\right]$*;*

(ii) *there are infinite elements in* $\mathcal{N}^s$ *and* $\lim_{k\to\infty} \mu_k = 0$*.*

Proof. (i). Upon rearranging terms, (3.10) can be rewritten as

$$\tilde{f}(\hat{x}^{k+1}, \mu_k) \le \tilde{f}(x^k, \mu_k) + \langle \nabla \tilde{f}(x^k, \mu_k), \hat{x}^{k+1} - x^k \rangle + \frac{1}{2}\gamma_k\mu_k^{-1}\|\hat{x}^{k+1} - x^k\|^2.$$

Invoking Definition 3.1-(v), (3.10) holds when $\gamma_k \ge L$. Thus the updating of γ_k in Step 2 is at most $\log_\eta(L/\underline{\gamma}) + 1$ times at each iteration. Hence, the SPG algorithm is well-defined and we have that $\gamma_k \le \max\{\bar\gamma, \rho L\}$, $\forall k \in \mathbb{N}$. From (3.11), it is easy to verify that $x^{k+1} \in \mathcal{X}$ by $x^k \in \mathcal{X}$ and $\hat{x}^{k+1} \in \mathcal{X}$.

(ii). Since $\{\mu_k\}$ is non-increasing, to prove (ii), we assume that $\lim_{k\to\infty} \mu_k = \hat\mu > 0$ by contradiction. Then, (3.13) happens finite times at most, which means that there exists $K \in \mathbb{N}$ such that $\mu_k = \hat\mu$, $\forall k \ge K$. Then,

$$\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k - \tilde{\mathcal{F}}(x^k, \mu_{k-1}) - \kappa\mu_{k-1} \le -\alpha\hat\mu^2, \quad \forall k \ge K+1.$$

We obtain from the above inequality that

$$\lim_{k\to\infty} \tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k = -\infty. \tag{3.15}$$

However, by $\{x^k\} \subseteq \mathcal{X}$ and (3.1), we see that

$$\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k \ge \mathcal{F}(x^{k+1}) \ge \min_{x\in\mathcal{X}} \mathcal{F}(x) = \min_{x\in\mathcal{X}} \mathcal{F}_{\ell_0}(x), \quad \forall k \ge K, \tag{3.16}$$

where the last equality follows from Theorem 2.4. Thus, the contradiction between (3.15) and (3.16) implies (ii). ∎

LEMMA 3.3. *For any $k \in \mathbb{N}$, we have*

$$\tilde{\mathcal{F}}(x^{k+1}, \mu_k) - \tilde{\mathcal{F}}(x^k, \mu_k) \le -\frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2, \tag{3.17}$$

which implies $\left\{\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k\right\}$ is non-increasing and $\lim_{k\to\infty} \tilde{\mathcal{F}}(x^{k+1}, \mu_k) = \lim_{k\to\infty} \mathcal{F}(x^k)$. Moreover, there exists $R > 0$ such that $\|x^k\| \le R$, $\forall k \in \mathbb{N}$.

Proof. Since $Q_{d^k,\gamma_k}(x, x^k, \mu_k)$ is strongly convex with modulus $\gamma_k\mu_k^{-1}$, using the definition of $\hat{x}^{k+1}$ in (3.9) and $x^{k+1} = \hat{x}^{k+1}$ when (3.10) holds, we obtain

$$Q_{d^k,\gamma_k}(x^{k+1}, x^k, \mu_k) \le Q_{d^k,\gamma_k}(x, x^k, \mu_k) - \frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x\|^2, \quad \forall x \in \mathcal{X}.$$

By the definition of function Q_{d^k,γ_k} given in (3.8), upon rearranging the terms, we have

$$\begin{aligned}\lambda\Phi^{d^k}(x^{k+1}) \le{}& \lambda\Phi^{d^k}(x) + \langle x - x^{k+1}, \nabla\tilde f(x^k, \mu_k)\rangle \\ &+ \frac{1}{2}\gamma_k\mu_k^{-1}\|x - x^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x\|^2.\end{aligned} \tag{3.18}$$

Moreover, (3.10) can be written as

$$\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) \le \tilde f(x^k, \mu_k) + \langle x^{k+1} - x^k, \nabla\tilde f(x^k, \mu_k)\rangle + \frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2 + \lambda\Phi^{d^k}(x^{k+1}). \tag{3.19}$$

Summing up (3.18) and (3.19), we notice that

$$\begin{aligned}\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) \le{}& \tilde f(x^k, \mu_k) + \lambda\Phi^{d^k}(x) + \langle x - x^k, \nabla\tilde f(x^k, \mu_k)\rangle \\ &+ \frac{1}{2}\gamma_k\mu_k^{-1}\|x - x^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|x^{k+1} - x\|^2, \quad \forall x \in \mathcal{X}.\end{aligned} \tag{3.20}$$

For a fixed $\mu > 0$, the convexity of $\tilde{f}(x, \mu)$ with respect to x invokes

$$\tilde{f}(x^k, \mu_k) + \langle x - x^k, \nabla \tilde{f}(x^k, \mu_k)\rangle \leq \tilde{f}(x, \mu_k), \quad \forall x \in \mathcal{X}. \tag{3.21}$$

Combining (3.20) and (3.21) and recalling the definition of $\tilde{\mathcal{F}}^{d^k}$, one has

$$\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) \leq \tilde{\mathcal{F}}^{d^k}(x, \mu_k) + \frac{1}{2}\gamma_k \mu_k^{-1}\|x - x^k\|^2 - \frac{1}{2}\gamma_k \mu_k^{-1}\|x^{k+1} - x\|^2, \quad \forall x \in \mathcal{X}. \tag{3.22}$$

Letting $x = x^k$ in (3.22) and by $d^k = d^{x^k}$, we obtain

$$\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) + \frac{1}{2}\gamma_k \mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq \tilde{\mathcal{F}}(x^k, \mu_k). \tag{3.23}$$

Thanks to $\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) \geq \tilde{\mathcal{F}}(x^{k+1}, \mu_k)$, (3.23) leads to (3.17).

Since $\tilde{\mathcal{F}}(x^k, \mu_k) \leq \tilde{\mathcal{F}}(x^k, \mu_{k-1}) + \kappa(\mu_{k-1} - \mu_k)$, by (3.17), we obtain

$$\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k + \frac{1}{2}\gamma_k \mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq \tilde{\mathcal{F}}(x^k, \mu_{k-1}) + \kappa\mu_{k-1}, \tag{3.24}$$

which implies the non-increasing property of $\left\{\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k\right\}$. Together this result with (3.16) ensures the existence of $\lim_{k\to\infty} \tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k$. By virtue of $\lim_{k\to\infty} \mu_k = 0$ and Definition 3.1-(i), we get $\lim_{k\to\infty} \tilde{\mathcal{F}}(x^{k+1}, \mu_k) = \lim_{k\to\infty} \mathcal{F}(x^k)$.

Recalling the non-increasing property of $\{\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k\}$ again, we see that

$$\mathcal{F}(x^{k+1}) \leq \tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k \leq \tilde{\mathcal{F}}(x^1, \mu_0) + \kappa\mu_0 < \infty.$$

We then obtain the boundedness of $\{x^k\}$ from $\{x^k\} \subseteq \mathcal{X}$ and the level bounded assumption of $\mathcal{F}$ on $\mathcal{X}$. Observe that

$$\mathcal{F}_{\ell_0}(x) \geq \mathcal{F}(x) = \mathcal{F}_{\ell_0}(x) - \lambda \sum\nolimits_{|x_i|<\nu}(1 - |x_i|/\nu) \geq \mathcal{F}_{\ell_0}(x) - \lambda n, \quad \forall x \in \mathbb{R}^n.$$

Then, it is easy to verify the level boundedness of $\mathcal{F}$ by the level boundedness of $\mathcal{F}_{\ell_0}$ on $\mathcal{X}$. Hence, the same results in Lemma 3.3 hold when $\mathcal{F}_{\ell_0}$ is level bounded on $\mathcal{X}$. ∎

The following proposition shows that there exists a subsequence of $\{x^k\}$ converging to a lifted stationary point of (1.6), which lays a foundation for the sequence convergence of $\{x^k\}$.

PROPOSITION 3.4. *Any accumulation point of $\{x^k : k \in \mathcal{N}^s\}$ is a lifted stationary point of (1.6).*

Proof. When $\mathcal{F}$ (or $\mathcal{F}_{\ell_0}$) is level bounded on $\mathcal{X}$, by Lemma 3.3, $\{x^k\}$ is bounded. Suppose $\bar{x}$ is an accumulation point of $\{x^k\}_{k\in\mathcal{N}^s}$ with the convergence of subsequence $\{x^{k_i}\}_{k_i\in\mathcal{N}^s}$.

Since (3.12) fails for $k_i \in \mathcal{N}^s$, by rearranging (3.24), we obtain that $\gamma_{k_i}\mu_{k_i}^{-1}\|x^{k_i+1} - x^{k_i}\|^2 \leq 2\alpha\mu_{k_i}^2$, which gives $\|x^{k_i+1} - x^{k_i}\| \leq \sqrt{2\alpha\gamma_{k_i}^{-1}\mu_{k_i}^3}$. Thus, $\gamma_{k_i}\mu_{k_i}^{-1}\|x^{k_i+1} - x^{k_i}\| \leq \sqrt{2\alpha\gamma_{k_i}\mu_{k_i}}$, which together with $\lim_{i\to\infty} \mu_{k_i} = 0$ and $\{\gamma_{k_i}\} \subseteq \left[\underline{\gamma}, \max\{\bar{\gamma}, \rho L\}\right]$ implies

$$\lim_{i\to\infty} \gamma_{k_i}\mu_{k_i}^{-1}\|x^{k_i+1} - x^{k_i}\| = 0 \quad \text{and} \quad \lim_{i\to\infty} x^{k_i+1} = \bar{x}. \tag{3.25}$$

Recalling $x^{k_i+1} = \hat{x}^{k_i+1}$ defined in (3.9) and by its first order necessary optimality condition, we have

$$\langle \nabla \tilde{f}(x^{k_i}, \mu_{k_i}) + \gamma_{k_i}\mu_{k_i}^{-1}(x^{k_i+1} - x^{k_i}) + \lambda\zeta^{k_i}, x - x^{k_i+1}\rangle \geq 0, \quad \forall \zeta^{k_i} \in \partial\Phi^{d^{k_i}}(x^{k_i+1}), \, x \in \mathcal{X}. \tag{3.26}$$

Since the elements in $\{d^{k_i} : i \in \mathbb{N}\}$ are finite and $\lim_{i\to\infty} x^{k_i+1} = \bar{x}$, there exists a subsequence of $\{k_i\}$, denoted as $\{k_{i_j}\}$, and $\bar{d} \in \mathcal{D}(\bar{x})$ such that $d^{k_{i_j}} = \bar{d}$, $\forall j \in \mathbb{N}$. By the upper semicontinuity of $\partial\Phi^{\bar{d}}$ and $\lim_{j\to\infty} x^{k_{i_j}+1} = \bar{x}$, it gives

$$\{\lim_{j\to\infty} \zeta^{k_{i_j}} : \zeta^{k_{i_j}} \in \partial\Phi^{d^{k_{i_j}}}(x^{k_{i_j}+1})\} \subseteq \partial\Phi^{\bar{d}}(\bar{x}). \tag{3.27}$$

Along with the subsequence $\{k_{i_j}\}$ and letting $j \to \infty$ in (3.26), from Definition 3.1-(iii), (3.25) and (3.27), we obtain that there exist $\bar{\xi} \in \partial f(\bar{x})$ and $\bar{\zeta}^{\bar{d}} \in \partial\Phi^{\bar{d}}(\bar{x})$ such that

$$\langle \bar{\xi} + \lambda\bar{\zeta}^{\bar{d}}, x - \bar{x}\rangle \geq 0, \quad \forall x \in \mathcal{X}. \tag{3.28}$$

By $\bar{d} \in \mathcal{D}(\bar{x})$, the definition of $\Phi^{\bar{d}}$ in (2.5) and the convexity of $\mathcal{X}$, (3.28) implies that $\bar{x}$ is a lifted stationary point of (1.6). ◻

Remark 3.1. *The convexity of Φ^d plays an important role in the analysis of the SPG algorithm. It is easy to check that all the results in subsection 3.4 are true when the penalty can be described by the* min *of a class of simple convex functions whose proximal operators can be calculated effectively.*

3.5. Global sequence convergence of the SPG algorithm for problem (1.6). It is interesting that the proposed SPG algorithm for this kind of nonconvex nonsmooth optimization problem owns the global sequence convergence without the K-L condition or error bound condition on the objective function, while the special structure of the continuous relaxation for $\|x\|_0$ and the updating rule for μ_k are the key points. Throughout this subsection, the analysis uses the same assumptions in subsection 3.4.

We begin this subsection by giving some preliminary analysis, which are Lemma 3.5, Lemma 3.6 and Proposition 3.7. Based on these results, we present the two main results for the SPG algorithm: the sequence convergence of $\{x^k\}$ in Theorem 3.8; the local convergence rate of $\{\mathcal{F}(x^k)\}$ and the finite-iteration identification of $\mathcal{A}(x^k)$ in Theorem 3.9.

Lemma 3.5. *The following statements hold:*

(i) $\sum_{k=0}^{\infty} \gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq 2\left(\mathcal{F}(x^0, \mu_{-1}) + \kappa\mu_{-1} - \min_{\mathcal{X}} \mathcal{F}\right)$;

(ii) $\sum_{k=0}^{\infty} \mu_k^2 \leq \Lambda$ *with* $\Lambda = \frac{1}{\alpha}\left(\tilde{\mathcal{F}}(x^0, \mu_{-1}) + \kappa\mu_{-1} - \min_{x\in\mathcal{X}} \mathcal{F}(x)\right) + \frac{2\mu_0^2\sigma}{2\sigma-1} < \infty$;

(iii) $\mathcal{A}(x^{k+1}) \subseteq \mathcal{A}(x^k)$.

Proof. (i). Recalling (3.24), for all $k \in \mathbb{N}$, we obtain

$$\gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq 2\left(\tilde{\mathcal{F}}(x^k, \mu_{k-1}) + \kappa\mu_{k-1} - \tilde{\mathcal{F}}(x^{k+1}, \mu_k) - \kappa\mu_k\right). \tag{3.29}$$

Summing up the above inequality over $k = 0, \ldots, K$, it gives

$$\sum_{k=0}^{K} \gamma_k\mu_k^{-1}\|x^{k+1} - x^k\|^2 \leq 2\left(\tilde{\mathcal{F}}(x^0, \mu_{-1}) + \kappa\mu_{-1} - \tilde{\mathcal{F}}(x^{K+1}, \mu_K) - \kappa\mu_K\right). \tag{3.30}$$

By letting K in (3.30) tend to infinity and along with (3.16), we obtain (i).

(ii). From (3.14), we have

$$\sum_{k\in\mathcal{N}^s}\mu_k^2=\sum_{r=1}^{\infty}\mu_0^2\frac{1}{(n_r^s+1)^{2\sigma}}\le\sum_{k=1}^{\infty}\frac{\mu_0^2}{k^{2\sigma}}\le\frac{2\mu_0^2\sigma}{2\sigma-1}, \tag{3.31}$$

where n_r^s is the rth smallest element in $\mathcal{N}^s$.

When $k\notin\mathcal{N}^s$, (3.12) gives $\alpha\mu_k^2\le\tilde{\mathcal{F}}(x^k,\mu_{k-1})+\kappa\mu_{k-1}-\tilde{\mathcal{F}}(x^{k+1},\mu_k)-\kappa\mu_k$, which together with the non-increasing property of $\left\{\tilde{\mathcal{F}}(x^{k+1},\mu_k)+\kappa\mu_k\right\}$ and (3.16) implies

$$\sum_{k\notin\mathcal{N}^s}\mu_k^2\le\frac{1}{\alpha}\left(\tilde{\mathcal{F}}(x^0,\mu_{-1})+\kappa\mu_{-1}-\min_{\mathcal{X}}\mathcal{F}\right). \tag{3.32}$$

Combining (3.31) and (3.32), we finish the proof for the estimation in item (ii).

(iii). We only need to prove that if $x_i^k=0$, then $x_i^{k+1}=0$. If $x_i^k=0$, we get $d_i^k=1$. From (3.7) and $\nu<\lambda/L_f$, we have

$$\left|x_i^k-\gamma_k^{-1}\mu_k\nabla_i\tilde{f}(x^k,\mu_k)\right|\le\gamma_k^{-1}\mu_k\left\|\nabla\tilde{f}(x^k,\mu_k)\right\|\le(\lambda\gamma_k^{-1}\mu_k)/\nu.$$

By (3.7), we obtain $x_i^{k+1}=0$, which completes the proof of this statement. ∎

For $\{x^k\}$, denote

$$\mathcal{N}_1=\{k\in\mathbb{N}:\text{ there exists } i\in\{1,\dots,n\}\text{ such that } 0<|x_i^k|<\nu\}. \tag{3.33}$$

Next lemma gives some estimation on $\{x^k\}$ and $\{\mu_k\}$ when k is sufficiently large.

LEMMA 3.6. *There exists $K\in\mathbb{N}$ such that for all $k\ge K$, it holds*

(i) $\left\|\nabla\tilde{f}(x^k,\mu_k)\right\|<\frac{1}{2}\left(\lambda/\nu+L_f\right)$;

(ii) $\left\|x^{k+1}-x^k\right\|\le 3(\lambda/\nu)\sqrt{n}\gamma^{-1}\mu_k$;

(iii) *for any $k\in\mathcal{N}_1$, either* $\left\|x^{k+1}\right\|_0\le\left\|x^k\right\|_0-1$ *or* $\left\|x^{k+1}-x^k\right\|\ge\frac{1}{2}\left(\lambda/\nu-L_f\right)\gamma_k^{-1}\mu_k$;

(iv) $\sum_{k\in\mathcal{N}_1,k\ge K}\left\|x^{k+1}-x^k\right\|<\infty$ *and* $\sum_{k\in\mathcal{N}_1,k\ge K}\mu_k<\infty$.

Proof. (i). We argue it by contradiction. Suppose there is a subsequence of $\{x^k\}$, denoted by $\{x^{k_i}\}$, such that

$$\left\|\nabla\tilde{f}(x^{k_i},\mu_{k_i})\right\|\ge\frac{1}{2}(\lambda/\nu+L_f)>L_f,\quad\forall i\in\mathbb{N}. \tag{3.34}$$

Since $\{x^{k_i}\}$ is bounded, which is proved in Lemma 3.3, there exists a subsequence of $\{x^{k_i}\}$ (also denoted by $\{x^{k_i}\}$ for simplicity) and $\bar{x}\in\mathcal{X}$ such that $\lim_{i\to\infty}x^{k_i}=\bar{x}$. Due to $\lim_{i\to\infty}\mu_{k_i}=0$, the property of $\tilde{f}$ in Definition 3.1-(iii), λ/ν and (3.34) imply the existence of $\bar{\xi}\in\partial f(\bar{x})$ such that $\|\bar{\xi}\|>L_f$, which leads to a contradiction to the definition of L_f given in Assumption 1. Hence, we establish result (i) in this lemma.

(ii). For any $i\in\{1,2,\dots,n\}$, by (3.7) and $L_f<\lambda/\nu$, we have

$$\left|x_i^{k+1}-x_i^k\right|\le 2(\lambda/\nu)\gamma_k^{-1}\mu_k+\gamma_k^{-1}\mu_k\left|\nabla_i\tilde{f}(x^k,\mu_k)\right|\le 3(\lambda/\nu)\gamma_k^{-1}\mu_k,$$

which completes the proof for item (ii).

(iii). Denote $w^k = x^k - \gamma_k^{-1}\mu_k\nabla\tilde{f}(x_k,\mu_k)$. For a fixed $k \in \mathcal{N}_1$ and $k \geq K$, there exists j such that $0 < |x_j^k| < \nu$. Then, $d_j^k = 1$ by (2.4). Next, we will prove that either $x_j^{k+1} = 0$ or $\left|x_j^{k+1} - x_j^k\right| \geq \frac{1}{2}\left(\lambda/\nu - L_f\right)\gamma_k^{-1}\mu_k$. We split the proof into three cases.

Case 1. If $|w_j^k| \leq (\lambda/\nu)\gamma_k^{-1}\mu_k$, by (3.7), we get $x_j^{k+1} = 0$, which together with $\mathcal{A}(x^{k+1}) \subseteq \mathcal{A}(x^k)$ implies $\|x^{k+1}\|_0 \leq \|x^k\|_0 - 1$.

Case 2. If $w_j^k > (\lambda/\nu)\gamma_k^{-1}\mu_k$, by (3.7) and result (i) of this lemma, we obtain that

$$\left|x_j^{k+1} - x_j^k\right| \geq (\lambda/\nu)\gamma_k^{-1}\mu_k - \left|\gamma_k^{-1}\mu_k\nabla_i\tilde{f}(x^k,\mu_k)\right| \geq \frac{1}{2}\left(\lambda/\nu - L_f\right)\gamma_k^{-1}\mu_k,$$

which implies

$$\left\|x^{k+1} - x^k\right\| \geq \frac{1}{2}\left(\lambda/\nu - L_f\right)\gamma_k^{-1}\mu_k. \tag{3.35}$$

Case 3. If $w_j^k < -(\lambda/\nu)\gamma_k^{-1}\mu_k$, similar to the analysis in Case 1, we see that (3.35) holds. Thus, we complete the proof of statement (iii).

(iv). We introduce the notations $\mathcal{N}_{11} = \{k \in \mathcal{N}_1 : k \geq K, \left\|x^{k+1}\right\|_0 \leq \left\|x^k\right\|_0 - 1\}$ and $\mathcal{N}_{12} = \{k : k \geq K, k \in \mathcal{N}_1 \backslash \mathcal{N}_{11}\}$. By Lemma 3.5-(iii), $\mathcal{N}_{11}$ has at most n elements. From result (iii) of this lemma, we have $\gamma_k\mu_k^{-1}\left\|x^{k+1} - x^k\right\| \geq \frac{1}{2}(\lambda/\nu - L_f)$, $\forall k \in \mathcal{N}_{12}$. Then, we have

$$\frac{1}{2}(\lambda/\nu - L_f)\sum_{k\in\mathcal{N}_{12}}\left\|x^{k+1} - x^k\right\| \leq \sum_{k\in\mathcal{N}_{12}}\gamma_k\mu_k^{-1}\left\|x^{k+1} - x^k\right\|^2 \leq 2\left(\tilde{\mathcal{F}}(x^0,\mu_{-1}) + \kappa\mu_{-1} - \min_{\mathcal{X}}\mathcal{F}\right), \tag{3.36}$$

where the second inequality follows from Lemma 3.5-(i). (3.36) implies $\sum_{k\in\mathcal{N}_{12}}\left\|x^{k+1} - x^k\right\| < \infty$, which together with the finiteness of the elements in $\mathcal{N}_{11}$ gives $\sum_{k\in\mathcal{N}_1,k\geq K}\left\|x^{k+1} - x^k\right\| < \infty$. Moreover,

$$\sum_{k\in\mathcal{N}_{12}}\gamma_k\mu_k^{-1}\left\|x^{k+1} - x^k\right\|^2 = \sum_{k\in\mathcal{N}_{12}}\left(\gamma_k\mu_k^{-1}\left\|x^{k+1} - x^k\right\|\right)^2\gamma_k^{-1}\mu_k \geq \frac{1}{4}\left(\lambda/\nu - L_f\right)^2\sum_{k\in\mathcal{N}_{12}}\gamma_k^{-1}\mu_k,$$

which together with the second inequality of (3.36) and Lemma 3.2-(i) implies $\sum_{k\in\mathcal{N}_{12}}\mu_k < \infty$. By $\sum_{k\in\mathcal{N}_{11}}\mu_k \leq n\mu_0$, we conclude that $\sum_{k\in\mathcal{N}_1,k\geq K}\mu_k < \infty$. ∎

The next proposition explores that all accumulation points of $\{x^k\}$ own a common support set and a unified lower bound, which provides the main technical support for the forthcoming Theorem 3.8.

PROPOSITION 3.7. *Denote $\bar{\mathcal{X}} = \{\bar{x} \in \mathcal{X} : \bar{x}$ is an accumulation point of $\{x^k\}\}$, then there exists $\mathcal{A}(\bar{\mathcal{X}}) \subseteq \{1,2,\ldots,n\}$ such that for any $\bar{x} \in \bar{\mathcal{X}}$, it holds that*

$$|\bar{x}_i| \geq \nu \text{ for any } i \in \mathcal{A}(\bar{\mathcal{X}}) \text{ and } \bar{x}_i = 0 \text{ for any } i \notin \mathcal{A}(\bar{\mathcal{X}}).$$

Proof. We first prove the following result:

$$\text{for any } \bar{x} \in \bar{\mathcal{X}} \text{ and any } i \in \{1,\ldots,n\}, \text{ either } \bar{x}_i = 0 \text{ or } |\bar{x}_i| \geq \nu. \tag{3.37}$$

If (3.37) does not hold, there exists $\hat{x} \in \bar{\mathcal{X}}$ with the convergence sequence $\{x^{k_j}\}$ and $\iota \in \{1,\ldots,n\}$ such that $0 < |\hat{x}_\iota| < \nu$. In what follows, without loss of generality, we suppose $\hat{x}_\iota > 0$.

Since any accumulation point of $\{x^k\}_{k\in\mathcal{N}^s}$ is an accumulation point of $\{x^k\}$, there exists $\bar{x} \in \bar{\mathcal{X}}$ and a subsequence of $\{x^k\}$, denoted by $\{x^{t_j}\}$, such that $\lim_{j\to\infty} x^{t_j} = \bar{x}$. By taking subsequences of $\{x^{k_j}\}$ and $\{x^{t_j}\}$ if necessary, we assume for the simplicity of notation that $k_j < t_j < k_{j+1}$, $\forall j \in \mathbb{N}$. Combining Proposition 2.5, Lemma 2.3 and Proposition 3.4, either $\bar{x}_\iota = 0$ or $|\bar{x}_\iota| \geq \nu$.

Let $\varepsilon = \min\left\{\frac{\nu-\hat{x}_\iota}{2}, \frac{\hat{x}_\iota}{4}\right\} > 0$. If $\bar{x}_\iota = 0$, there exists $J \in \mathbb{N}$ such that

$$|x_\iota^{k_j} - \hat{x}_\iota| \leq \varepsilon \quad \text{and} \quad |x_\iota^{t_j}| \leq \varepsilon, \quad \forall j \geq J,$$

which implies

$$\frac{3}{4}\hat{x}_\iota \leq \hat{x}_\iota - \varepsilon \leq x_\iota^{k_j} \leq \varepsilon + \hat{x}_\iota \leq \frac{\nu + \hat{x}_\iota}{2} < \nu \quad \text{and} \quad -\frac{1}{4}\hat{x}_\iota \leq x_\iota^{t_j} \leq \frac{1}{4}\hat{x}_\iota, \quad \forall j \geq J. \tag{3.38}$$

Then, $x_\iota^{k_j} - x_\iota^{t_j} \geq \frac{1}{2}\hat{x}_\iota$, $\forall j \geq J$. Thus,

$$\sum_{j=J}^{\infty} \left|x_\iota^{t_j} - x_\iota^{k_j}\right| = +\infty. \tag{3.39}$$

If there exists $r \geq J$, such that $x_\iota^{t_r} = 0$, Lemma 3.5-(iii) gives $x_\iota^{k_{j+1}} = 0$, $\forall j \geq r$, which leads to a contradiction to the first inequality in (3.38). Thus, (3.38) gives $0 < |x_\iota^{k_j}| < \nu$ and $0 < |x_\iota^{t_j}| < \nu$, which implies $\{x^{k_j}, x^{t_j} : j \geq J\} \subseteq \mathcal{N}_1$ with $\mathcal{N}_1$ defined in (3.33). Together this with Lemma 3.6-(ii), (iv) and $\lim_{k\to\infty} \mu_k = 0$, there exists $J_1 \geq J$ such that

$$\sum_{j=J_1}^{\infty} \left|x_\iota^{t_j} - x_\iota^{k_j}\right| \leq \sum_{k\in\mathcal{N}_1, k\geq K} \left\|x^{k+1} - x^k\right\| < \infty,$$

which leads to a contradiction to (3.39). Likewise, we can obtain a similar contradiction when $|\bar{x}_\iota| \geq \nu$. Therefore, the above analysis ensures the validity of statement (3.37). Together (3.37) with $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$, we complete the proof of this proposition. ∎

We next prove the global sequence convergence of iterates $\{x^k\}$.

THEOREM 3.8. *The iterates $\{x^k\}$ generated by the SPG algorithm is globally convergent to a lifted stationary point of (1.6), i.e. there exists a lifted stationary point $\bar{x}$ of (1.6) such that $\lim_{k\to\infty} x^k = \bar{x}$.*

Proof. Let K be a positive integer such that the estimations in Lemma 3.6 hold and $\bar{x}$ be an accumulation point of $\{x^k\}_{k\in\mathcal{N}^s}$. Suppose $\{x^{k_j}\}$ is a subsequence of $\{x^k\}$ such that

$$\lim_{j\to\infty} x^{k_j} = \bar{x}. \tag{3.40}$$

By Proposition 3.4, $\bar{x}$ is a lifted stationary point of (1.6).

From Lemma 2.3, for any $i \in \{1, \ldots, n\}$, either $\bar{x}_i = 0$ or $|\bar{x}_i| \geq \nu$. Denote

$$\mathcal{N}(\bar{x}) = \{k \in \mathbb{N} : \, d_i^k \in \mathcal{D}(\bar{x}_i), \, \forall i = 1, \ldots, n\},$$

where $\mathcal{D}(\bar{x}_i)$ is defined in (2.1). We then evaluate $\|x^{k+1} - \bar{x}\|^2$ by considering two cases.

Case 1: In this case, we consider the iteration for $k \in \mathcal{N}(\bar{x})$, which implies that $\tilde{\mathcal{F}}^{d^k}(\bar{x}, \mu_k) = \tilde{\mathcal{F}}(\bar{x}, \mu_k)$. Letting $x = \bar{x}$ in (3.22), we have

$$\tilde{\mathcal{F}}^{d^k}(x^{k+1}, \mu_k) - \tilde{\mathcal{F}}(\bar{x}, \mu_k) \leq \frac{1}{2}\gamma_k\mu_k^{-1}\left\|x^k - \bar{x}\right\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\left\|x^{k+1} - \bar{x}\right\|^2,$$

combining which with (3.1) and (3.4), we obtain

$$2\gamma_k^{-1}\mu_k\left(\tilde{\mathcal{F}}(x^{k+1},\mu_k)+\kappa\mu_k-\mathcal{F}(\bar{x})\right)\leq \left\|x^k-\bar{x}\right\|^2-\left\|x^{k+1}-\bar{x}\right\|^2+4\kappa\gamma_k^{-1}\mu_k^2. \tag{3.41}$$

Due to the non-increasing property of $\left\{\tilde{\mathcal{F}}(x^{k+1},\mu_k)+\kappa\mu_k\right\}$ and $\lim_{k\to\infty}\tilde{\mathcal{F}}(x^{k+1},\mu_k)+\kappa\mu_k=\mathcal{F}(\bar{x})$, we obtain

$$\left\|x^{k+1}-\bar{x}\right\|^2\leq\left\|x^k-\bar{x}\right\|^2+4\kappa\gamma_k^{-1}\mu_k^2,\quad \forall k\in\mathcal{N}(\bar{x}). \tag{3.42}$$

Case 2: In this case, we consider the iteration for $k\notin\mathcal{N}(\bar{x})$. From Proposition 3.7, there exists $K_1\geq K$ such that for any $k\geq K_1$, it holds

$$\left|x_i^k\right|<\nu/2 \text{ for } i\notin\mathcal{A}(\bar{\mathcal{X}}) \text{ and } \left|x_i^k\right|\geq\nu/2 \text{ for } i\in\mathcal{A}(\bar{\mathcal{X}}),$$

where $\mathcal{A}(\bar{\mathcal{X}})$ is defined in Proposition 3.7.

Hence, for $k\notin\mathcal{N}(\bar{x})$ and $k\geq K_1$, there exists $\iota^k\in\mathcal{A}(\bar{\mathcal{X}})$ such that $\nu/2\leq\left|x_{\iota^k}^k\right|<\nu$, which means that $k\in\mathcal{N}_1$ with $\mathcal{N}_1$ defined in (3.33). Then,

$$\begin{aligned}\left\|x^{k+1}-\bar{x}\right\|^2&=\left\|x^k-\bar{x}\right\|^2+\left\|x^{k+1}-x^k\right\|^2+2\langle x^{k+1}-x^k,x^k-\bar{x}\rangle\\&\leq\left\|x^k-\bar{x}\right\|^2+c_1\mu_k^2+4R\left\|x^{k+1}-x^k\right\|,\quad\forall k\notin\mathcal{N}(\bar{x}),\end{aligned} \tag{3.43}$$

where $c_1=9(\lambda/\nu)^2n\underline{\gamma}^{-2}$ follows from Lemma 3.6-(ii), and R comes from Lemma 3.3.

By (3.42) and (3.43), for any $t\geq K_1$ and $s\in\mathbb{N}$, we have

$$\left\|x^{t+s+1}-\bar{x}\right\|^2\leq\left\|x^t-\bar{x}\right\|^2+c_2\sum_{k=t}^{t+s}\mu_k^2+4R\sum_{\substack{k\,=\,t,\\ k\,\notin\,\mathcal{N}(\bar{x})}}^{t+s}\left\|x^{k+1}-x^k\right\|, \tag{3.44}$$

where $c_2=\max\{4\kappa\underline{\gamma}^{-1},c_1\}$.

Fix an $\epsilon>0$. There exists $K_2\geq K_1$ such that when $k_j\geq K_2$, it holds that

$$\left\|x^{k_j}-\bar{x}\right\|^2\leq\epsilon^2/3,\qquad\sum_{k=k_j}^{\infty}\mu_k^2\leq\epsilon^2/3c_2,\qquad\sum_{\substack{k\,=\,k_j,\\ k\,\notin\,\mathcal{N}(\bar{x})}}^{\infty}\left\|x^{k+1}-x^k\right\|\leq\epsilon^2/12R, \tag{3.45}$$

where the first inequality follows from (3.40), the second inequality follows from Lemma 3.5-(ii), and the third inequality follows from and $\{k:k\geq K_1,k\notin\mathcal{N}(\bar{x})\}\subseteq\mathcal{N}_1$ and Lemma 3.6-(iv).

Letting $t=k_j$ in (3.44) with $k_j\geq K_2$, from (3.45), we obtain $\|x^k-\bar{x}\|\leq\epsilon$, $\forall k\geq K_3$, where $K_3=\min\{k_j:k_j\geq K_2\}$. Due to the arbitrariness of $\epsilon>0$, we get $\lim_{k\to\infty}x^k=\bar{x}$. ◻

The lower bound property is used to prove the estimation in Lemma 3.6-(iii), which is the key point to guarantee the global sequence convergence of $\{x^k\}$. Without this lower bound property, due to the nonconvexity of the objective function in (1.6), it is almost impossible to propose a global sequence convergence algorithm without the regularity conditions. Among the existing penalties, only capped-ℓ_1 penalty can be expressed by the min of a class of simple convex functions and make the stationary points of the corresponding minimization problem own a unified lower bound. This is the main motivation of this paper on studying the cardinality penalty problem by the capped-ℓ_1

relaxation. Moreover, from the proof of Theorem 3.9, we find that the descent criterion and the updating method for μ_k are also important to guarantee the global sequence convergence of $\{x^k\}$, since it needs that $\Sigma_{k=1}^{\infty}\mu_k^2 < +\infty$.

The limit point of $\{x^k\}$ is most likely different with different initial iterates x^0 and μ_0. The zero vector is a trivial ν-strong local minimizer of (1.2), which is not we want. By property (iii) of Lemma 3.5, our theoretical results hold for any initial iterate $x^0 \in \mathcal{X}$. To find interesting ν-strong local minimizers, we chose x^0 without zero component in the numerical experiments. How to choose an initial point such that the accumulation point of $\{x^k\}$ is a global minimizer (or an oracle solution) of (1.2) is an interesting work. To the best of our knowledge, it is still an open problem. [24, Theorem 1] gave some discussion on this topic for the linear approximation algorithm to solve the sparsity problem with SCAD penalty. Similar results can be expected for the SPG algorithm.

The following theorem gives a local convergence rate of the SPG algorithm on the objective function values of problem (1.6) and the finite iteration convergence of $\{x^k\}$ in a subspace.

THEOREM 3.9. *There exist $c > 0$ and $K \in \mathbb{N}$ such that, for $k \geq K$, we have*

$$\mathcal{F}(x^{k+1}) - \mathcal{F}(\bar{x}) \leq ck^{-(1-\sigma)} \quad \text{and} \quad \left\| x^k_{\mathcal{A}(\bar{x})^c} - \bar{x}_{\mathcal{A}(\bar{x})^c} \right\| = 0, \tag{3.46}$$

where $\bar{x}$ is the limit of $\{x^k\}$.

Proof. Denote $\epsilon = \min\{\nu, \min\{|\bar{x}_i| - \nu : |\bar{x}_i| > \nu, \, i = 1, \ldots, n\}\}$. From Theorem 3.8, there exists $K_1 \in \mathbb{N}$ such that $\|x^k - \bar{x}\| < \epsilon$, $\forall k \geq K_1$. Then, $k \in \mathcal{N}(\bar{x})$, $\forall k \geq K_1$.

From the proof of Theorem 3.8, (3.41) holds for any $k \geq K_1$. Summing up (3.41) for $k = K_1, K_1 + 1, \ldots, K_1 + t$, we have

$$\begin{aligned} &2t \max\{\bar{\gamma}, \rho L\}^{-1} \mu_{K_1+t} \left(\tilde{\mathcal{F}}(x^{K_1+t+1}, \mu_{K_1+t}) + \kappa\mu_{K_1+t} - \mathcal{F}(\bar{x}) \right) \\ \leq &\|x^{K_1} - \bar{x}\|^2 - \|x^{K_1+t+1} - \bar{x}\|^2 + 4\kappa \sum_{k=K_1}^{K_1+t} \gamma_k^{-1}\mu_k^2, \end{aligned} \tag{3.47}$$

where we use $\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k \geq \mathcal{F}(\bar{x})$, $\{\gamma_k\} \subseteq \left[\underline{\gamma}, \max\{\bar{\gamma}, \rho L\}\right]$, and the non-increasing property of $\{\mu_k\}$ and $\left\{\tilde{\mathcal{F}}(x^{k+1}, \mu_k) + \kappa\mu_k\right\}$.

We first consider the right hand side of (3.47). We observe that $4\kappa \sum_{k=K_1}^{K_1+t} \gamma_k^{-1}\mu_k^2 \leq 4\kappa\underline{\gamma}^{-1}\Lambda$, where Λ is defined in Lemma 3.5-(ii). Then,

$$\left\|x^{K_1} - \bar{x}\right\|^2 - \left\|x^{K_1+t+1} - \bar{x}\right\|^2 + 4\kappa \sum_{k=K_1}^{K_1+t} \gamma_k^{-1}\mu_k^2 \leq 4R^2 + 4\kappa\underline{\gamma}^{-1}\Lambda, \quad \forall t \in \mathbb{N}, \tag{3.48}$$

with R defined in Lemma 3.3.

By $\mu_{K_1+t} \geq \mu_0(K_1+t)^{-\sigma}$ and $\tilde{\mathcal{F}}(x^{K_1+t+1}, \mu_{K_1+t}) + \kappa\mu_{K_1+t} \geq \mathcal{F}(x^{K_1+t+1})$, $\forall t \in \mathbb{N}$, we observe from (3.47) and (3.48) that

$$\mathcal{F}(x^{K_1+t+1}) - \mathcal{F}(\bar{x}) \leq \left(\frac{(4R^2 + 4\kappa\underline{\gamma}^{-1}\Lambda)\max\{\bar{\gamma}, \rho L\}}{2\mu_0} \right) \left(\frac{(K_1+t)^\sigma}{t} \right).$$

Therefore, letting $c = (4R^2 + 4\kappa\underline{\gamma}^{-1}\Lambda)\max\{\bar{\gamma}, \rho L\}/\mu_0$, we obtain

$$\mathcal{F}(x^{k+1}) - \mathcal{F}(\bar{x}) \leq \frac{c}{2}\left(\frac{k^\sigma}{k - K_1}\right) \leq ck^{-(1-\sigma)}, \ \forall k \geq 2K_1.$$

To prove the second statement in (3.46), we argue it by contradiction. If there is no $K \in \mathbb{N}$ such that $x_i^k = 0$ for all $i \in \mathcal{A}(\bar{x})^c$ and $k \geq K$, then there is a subsequence of $\{x^k\}$, denoted by $\{x^{k_j}\}$, and $\hat{i} \in \mathcal{A}(\bar{x})^c$ such that $|x_{\hat{i}}^{k_j}| \neq 0$. Since $\mathcal{A}(x^{k+1}) \subseteq \mathcal{A}(x^k)$ and $\lim_{k\to\infty} x^k = \bar{x}$, the above assumption implies that there exists $K_1 \in \mathbb{N}$ such that $0 < |x_{\hat{i}}^k| < \nu$, $\forall k \geq K_1$. Thus, for all $k \geq K_1$, it gives $k \in \mathcal{N}_1$ with $\mathcal{N}_1$ given in (3.33). Recalling Lemma 3.6-(iv), we get $\sum_{k=K_1}^{\infty} \mu_k < \infty$. However, due to $\mu_k \geq \mu_0 k^{-\sigma}$ with $\sigma < 1$, we have $\sum_{k=K_1}^{\infty} \mu_k = \infty$, which leads to a contradiction. Therefore, the second statement in (3.46) holds. □

Following the proof of Theorem 3.9, the local convergence rate of $\mathcal{F}(x^k) - \mathcal{F}(\bar{x})$ is $\mathrm{O}(\frac{1}{k\mu_k})$. Moreover, thanks to the lower bound property, the SPG algorithm owns the finite iteration identification on the support set of the limit point of $\{x^k\}$, which inspires us that the local convergence rate can be improved when f satisfies some proper conditions. For example, when f is strongly convex with modulus $\delta > 0$, then the local convergence rate can be exponential; when f satisfies the K-L inequality on $\mathcal{X}$ with exponent $\alpha \in [0, 1)$, then $\{x^k\}$ is convergent finitely if $\alpha = 0$, linearly if $\alpha \in (0, \frac{1}{2}]$ and sublinearly if $\alpha \in (\frac{1}{2}, 1)$.

4. Numerical experiments. To verify and illustrate the performance of the continuous relaxation (1.6) and the SPG algorithm, we use a test example and generate two examples randomly with normal distribution. All experiments are performed in MATLAB 2016a on a Lenovo PC (3.00GHz, 2.00GB of RAM). In the following examples, the stopping criterion is set as

$$\text{number of iterations} \leq \texttt{Maxiter} \qquad \text{or} \qquad \mu_k \leq \epsilon. \tag{4.1}$$

Denote $\bar{x}$ the output of iterate x^k, `Iter` the number of running iterations and `Time` the CPU time of the SPG algorithm by the criterion in (4.1). Examples 4.1 and 4.2 are for the under-determined linear regression problems. Moreover, Example 4.1 is a typical under-determined linear regression problem, which shows that the proposed method in this paper can find a global solution with certain sparsity. The aim of Example 4.2 is to solve a random generated under-determined sparse linear regression problem, while Example 4.3 is to solve a over-determined censored regression problem.

Example 4.1. (A test example) We consider the problem in Example 2.1 to verify the validity of the theoretical results and the efficiency of SPG algorithm. Problem (2.9) is an example of problem (1.2) with the ℓ_1 loss function given in (1.3), where $m = 1$, $\boldsymbol{A} = (1\ 1)$ and $b = 1$.

Let the smoothing function of f be defined by (3.2). Some fixed parameters in the SPG algorithm are given as follows:

$$\underline{\gamma} = \bar{\gamma} = \sqrt{2},\ \alpha = 1,\ \sigma = 0.8,\ \rho = 1.1,\ \texttt{Maxiter} = 10^4,\ \epsilon = 10^{-3},\ \kappa = 1/2,\ L_f = \sqrt{2}.$$

Let $\mathcal{LM}$, $\nu - \mathcal{LM}$ and $\mathcal{GM}$ denote the sets of local minimizers, ν-strong local minimizers and global minimizers of (2.9), respectively. When $\nu < \lambda/L_f$,

$$\nu - \mathcal{LM} = \{x : x_1 + x_2 = 1, \nu \leq x_1, x_2 \leq 1\} \cup \{(1,0)^T, (0,1)^T, (0,0)^T\}.$$

Set $\mu_0 = 0.1$ and $x^0 = (1, 0.8)^T$. The other parameters and the numerical results are listed in Table 4.1, where the global minimizers are same for the cases in one line. For problem (2.9), many different values of λ and the corresponding ν if $\nu < \lambda/L_f$ are given in Table 4.1, which shows that $\bar{x}$ is always a ν-strong local minimizer and sometimes a global minimizer of (2.9). In particular, when $\lambda = 0.7$, $\nu = 0.4$ and $x^0 = (1, 0.8)^T$, the SPG algorithm finds a global solution of (2.9). Moreover, we consider the influence of the values of ν on the SPG algorithm for solving (2.9) in

Table 4.1. When $\lambda = 1$, $\bar{\nu}$ as defined in Assumption 2 is 0.7071. From Table 4.1, we find that the SPG algorithm finds different ν-strong local minimizer for different values of ν satisfying $\nu < \bar{\nu}$. And it is interesting that when $\nu \geq 0.5$, the SPG algorithm converges to a global minimizer. We notice that when ν is a lower bound for the global minimizers, it holds that

$$\mathcal{GM} \subseteq \nu_1\text{-}\mathcal{LM} \subseteq \nu_2\text{-}\mathcal{LM},\ \forall \nu_2 \leq \nu_1 \leq \nu. \tag{4.2}$$

Hence when ν is a lower bound for the global minimizers, the larger ν is likely to let the SPG algorithm converge to a global minimizer with higher possibility.

The updating rule for μ_k in the SPG algorithm is to ensure its global sequence convergence. How to improve the local convergence rate with the guarantee of global sequence convergence is an interesting work for further research.

λ	$\mathcal{GM}$	ν	Iter	$\bar{x}$
0.7/0.8/0.9	$(1,0)$, $(0,1)$	0.4/0.5/0.6	18/19/10	$(1,0)/(1,0)/(1,0)$
1/1/1	$(1,0)$, $(0,1)$, $(0,0)$	0.7/0.5/0.3	21/11/5	$(0,0)/(1,0)/(0.6,0.4)$
1.1/1.2/1.3	$(0,0)$	0.7/0.9/1	18/17/16	$(0,0)/(0,0)/(0,0)$

Table 4.1: Numerical results of the SPG algorithm for problem (2.9) with different λ and ν

Using the same parameters and initial point, the IRL1 and IRTight algorithms in [43] may generate

$$x^k = \arg\min_{0\leq x_1,x_2\leq 1} |x_1 + x_2 - 1| \tag{4.3}$$

for $k \geq 0$ with $x^k \equiv (\alpha, \beta) > 0$ and $\alpha+\beta = 1$. Obviously, x^k is not a global minimizer of (2.9). Hence almost surely the reweighted algorithms in [43] cannot find a global minimizer (2.9). In fact, at any point $x^k > 0$, the derivative of $\|x^k\|_0$ is $(0,0)^T$ and $x^{k+1} = (\alpha, \beta) > 0$ with $\alpha + \beta = 1$ is an optimal solution of the subproblem $\min_{0\leq x_1,x_2\leq 1} |x_1 + x_2 - 1|$ in the algorithms. Hence the SPG algorithm has better performance than the algorithms in [43] for solving the nonsmooth optimization problem with cardinality penalty (1.2).

Example 4.2. (Linear regression problem) Linear regression problem is the most representative problem in sparse regression, which has been widely used in information theory [12], image restoration[5, 10, 41], signal precessing [10, 41] and variable selection [23, 24] problems. Least square function is the most frequently used loss function due to its convexity and differentiability [24, 29, 31, 32, 54]. However, the ℓ_1 loss function often owns the stronger outlier-resistant property than the least square loss function [23]. So, in this example, we consider the following cardinality penalty problem with ℓ_1 loss function:

$$\min_{0\leq x\leq 10\mathbf{1}_n} \quad \mathcal{F}_{\ell_0}(x) := \frac{1}{m}\|\boldsymbol{A}x - b\|_1 + \lambda\|x\|_0, \tag{4.4}$$

where $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$ with $m < n$.

Generating data and setting parameters. For positive integers m, n and s, we generate the original signal $x^\star$ with $\|x^\star\|_0 = s$, sensing matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and observation $b \in \mathbb{R}^m$ as follows:

```
index=randperm(n); index=index(1:s); x⋆=zeros(n,1); B=randn(n,m);
```

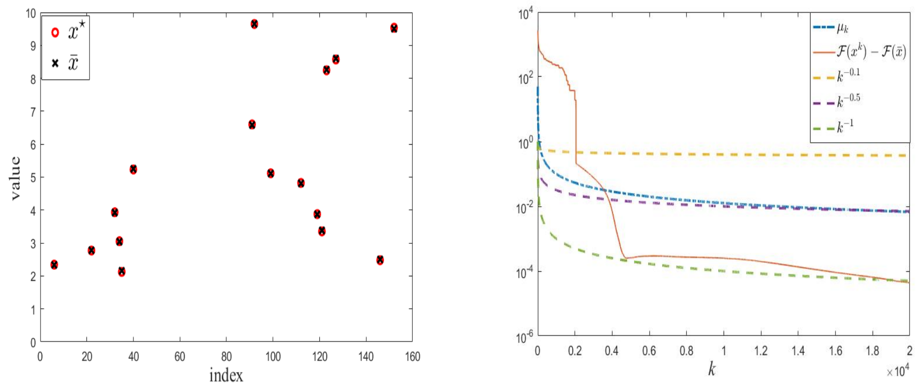
$x^\star$(index)=unifrnd(2,10,[s,1]); A=orth(B)'; b = A*$x^\star$+ 0.01*randn(size(b)).

In the proposed SPG algorithm, we use the smoothing function of f in (3.2) and set the parameters as below

$$\underline{\gamma} = \bar{\gamma} = 1,\ \alpha = 1,\ \mu_0 = 50, \rho = 1.1, \sigma = 0.9, \texttt{Maxiter} = 10^4,\ \kappa = 1/2.$$

It is not hard to show that all assumptions in sections 2 and 3 hold. Thus, the sequence $\{x^k\}$ of the SPG algorithm should be convergent to a ν-strong local minimizer of (4.4).

Generate A, b and $x^\star$ with $m = 80$, $n = 160$ and $s = 16$, set $\lambda = 18.8$ in (4.4) and $\epsilon = 10^{-3}$ in the stopping criterion (4.1). We calculate that $L_f = 10.6168$ and define $\nu = 1.77$, $x^0 = 1.97 * \mathrm{ones}(n, 1)$. The numerical results are shown in Fig. 4.1. Fig. 4.1(a) plots $x^\star$ and $\bar{x}$. From Fig 4.1(a), we see that the output of x^k is very close to the original generated signal and satisfies the lower bound property in (2.8). Fig. 4.1(b) exhibits the convergence of μ^k and $\mathcal{F}(x^k) - \mathcal{F}(\bar{x})$.



(a) (b)

Fig. 4.1: Numerical results of the SPG algorithm for Example 4.2

Example 4.3. (Censored regression problem) A typical class of censored regression problem is the linear regression model with left-censoring (or right-censoring) at zero, i.e.

$$\max\{A_i x - c_i, 0\} \approx b_i, \quad i = 1, \ldots, m,$$

where A_i, b_i and c_i are defined as in (1.4). This class of problems have wide applications in wireless communication [38], machine learning [21], variable selection[23, 53], economics [9], etc. To solve it, the loss function is often defined by (1.4), which is nonsmooth for any $p \in [1, 2]$. So the censored regression problem is a typical class of sparse regression problems with nonsmooth convex loss functions [53]. Different from the case considered in Example 4.2, we let $m \gg n$ in this example, which comes from the stochastic optimization models in the portfolio management.

In this example, we let $l = \mathbf{0}$ and $u = \mathbf{1}_n$ in (1.2), and define the loss function f by (1.4) with $c_i = 0$, $i = 1, \ldots, m$ and $p = 1$. The aim of this model is to find a sparse signal $x^\star \in [\mathbf{0}, \mathbf{1}_n]$ for the nonlinear system $\max\{\boldsymbol{A}x^\star, 0\} \approx b$ with some unobservable noise, where $\boldsymbol{A} = (A_1^T, \ldots, A_m^T)^T$ and $b = (b_1, \ldots, b_m)^T$. We use the relative error (rel-err), sparsity regression rate (spa-rat) and successful rate (suc-rat) to judge the performance of the continuous relaxation model for (1.2) and the proposed SPG algorithm. Here, the relative error (`rel-err`) and sparsity regression rate (`spa-rat`) of $\bar{x}$ with respect to $x^\star$ are defined by

$$\texttt{rel-err} := \frac{\|\bar{x}-x^\star\|}{\|\bar{x}\|}, \quad \texttt{spa-rat} := \frac{|\mathcal{A}(x^\star)\cap\mathcal{A}(\bar{x})|}{\max\{|\mathcal{A}(\bar{x})|,|\mathcal{A}(x^\star)|\}},$$

where $|\Xi|$ means the cardinality of set Π with finite elements. The running regression test is regarded as a successful one, if the relative error is smaller then 10^{-2} and $\mathcal{A}(\bar{x})=\mathcal{A}(x^*)$.

For the given positive integers m, n and s, the data are generated by

```
index = randperm(n); index = index(1:s); x⋆=zeros(n,1);
x⋆(index)=unifrnd(0,0.9,[s,1]); x⋆=sign(x⋆)*(abs(x⋆)+0.1)
A=randn(m,n); b =max{A*x⋆+0.01*randn(size(b)),0},
```

which let $x^\star$ satisfies $|x_i^\star|\geq 0.1$, $\forall i\in\mathcal{A}(x^\star)$.

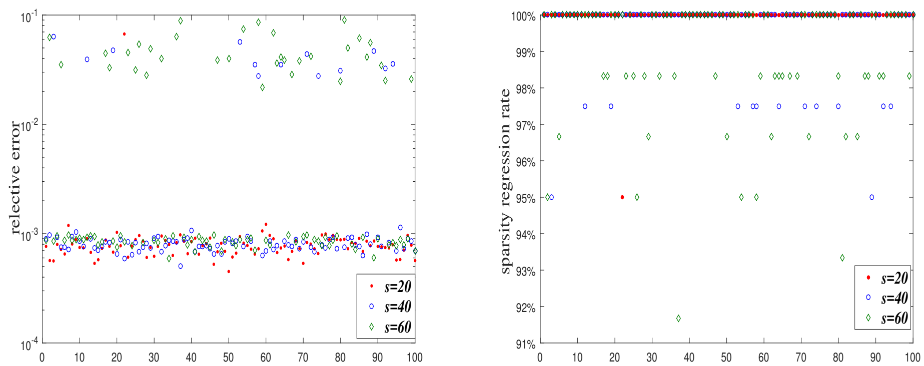
We use the smoothing function of f in (3.3). Let $L_f=\|\boldsymbol{A}\|_\infty$ and set $\nu=\min\{\lambda/L_f,1\}$. Set $x^0=0.1*\mathrm{ones(n,1)}$, $\mu_0=1$ and $\epsilon=0.01$. Let the other parameters in the SPG algorithm be the same as in Example 4.2.

For each group of given numbers m, n and s, we generate the codes with 100 independent trials, and the results displayed in Table 4.2 are the average values for these 100 independent tests. For each test, regarding the lower bound of the true solution $x^\star$, we run the SPG algorithm for problem (1.2) with $\lambda:=\delta L_f$ for $\delta\in[0.001:0.001:0.1]$, and report the result with the smallest `rel-err` for this test. From the displayed results in Table 4.2, we see that the the proposed SPG algorithm can find the true solution with high possibility, and all the sparsity regression rates are more than 90%. In particular, when $m=2000$ and $n=400$, the SPG algorithm can identify almost all the locations of $\mathcal{A}(x^\star)$ when the sparsity levels of $x^\star$ are 10%, 20% and 30%. Correctly identifying the zero and nonzero locations of the true solution is the most important thing in solving the variable selection and classification problems. When $m=1000$ and $n=200$, the values of relative error and sparsity regression rates by the 100 tests are plotted in Fig. 4.2 for $s=20$, 40 and 60, respectively.

m	n	s	Time	Iter	rel-err	$\mathcal{A}(\bar{x})$	spa-rat	suc-rat
1000	200	20	0.612	166	173e-3	19.99	100%	99%
1000	200	40	0.659	178	5.73e-3	39.96	99.7%	89%
1000	200	60	0.708	204	9.12e-3	59.94	92.7%	69%
2000	400	40	2.079	181	1.96e-3	40	100%	96%
2000	400	80	2.686	217	6.93e-3	79.89	99.7%	83%
2000	400	120	3.658	291	9.34e-3	119.91	99.3%	64%

Table 4.2: Average numerical results of the SPG algorithm for the censored regression problem

5. Conclusions. Problem (1.2) includes a class of constrained optimization problems with the objective function defined by the sum of a nonsmooth convex function and a cardinality function. Using the capped-ℓ_1 penalty, we propose a continuous relaxation (1.6) of problem (1.2). We prove that the sets of global minimizers of problems (1.2) and (1.6) are same, and local minimizers of (1.6) are local minimizers of (1.2) with the lower bound property. Moreover, $\bar{x}$ is a local minimizer of (1.2) satisfying a desired lower bound property if and only if it is a lifted stationary point of the continuous relaxation problem (1.6). Though problem (1.6) is a nonsmooth and nonconvex



(a) (b)

Fig. 4.2: The values of relative error and sparsity regression rate for the 100 tests with $m = 1000$ and $n = 200$

optimization problem, its piecewise linear penalty offers us the opportunity to solve it efficiently. Following this idea, we propose the SPG algorithm based on the smoothing method and the proximal gradient algorithm to solve problem (1.6), which can find a "good" local minimizer of (1.2) that satisfies the desired lower bound. The proposed algorithm is simple, whose subproblem has a closed form solution, and can be run efficiently. We prove the global sequence convergence without using the K-L condition. Another interesting result is that the local convergence rate of the SPG algorithm on the objective function value is $o(k^{-\tau})$ with $\tau \in (0, \frac{1}{2})$ and the zero entries of a lifted stationary point of (1.6) can be identified in finite iterations.

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. 1637–1655.
[2] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems*, Math. Program., 137 (2013), pp. 961–978.
[3] A. Beck and N. Hallak, *Proximal mapping for symmetric penalty and sparsity*, SIAM J. Optim., 28 (2018), pp. 496–527.
[4] A. Beck and N. Hallak, *Optimization problems involving group sparsity terms*, Math. Program., 178 (2019), pp. 39–67.
[5] W. Bian and X. Chen, *Linearly constrained non-Lipschitz optimization for image restoration*, SIAM J. Imaging Sci., 8 (2015), pp. 2294–2322.
[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, X. Chen, and Y. Ye, *Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization*, Math. Program., 149 (2015), pp. 301–327.
[8] T. Blumensath and M. E. Davies, *Iterative thresholding for sparse approximations*, J. Fourier Anal. Appl., 14 (2008), pp. 629–654.
[9] R. Blundell and J. L. Powell, *Censored regression quantiles with endogenous regressors*, J. Econometrics, 141 (2007), pp. 65–83.
[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] P. Buhlmann, M. Kalisch, and L. Meier, *High-dimensional statistics with a view toward applications in biology*, Ann. Rev. Stat. Appl., 1 (2014), pp. 255–278.
[12] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inf. Theory, 52 (2006), pp. 489–509.
[13] E. J. Candès, M. B. Walkin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.
[14] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245–295.
[15] R. Chartrand and V. Staneva, *Restricted isometry properties and nonconvex compressive sensing*, Inverse Probl., 24 (2008), pp. 1–14.
[16] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99.
[17] X. Chen, D. Ge, Z. Wang, and Y. Ye, *Complexity of unconstrained ℓ_2-ℓ_p minimization*, Math. Progrm., 143 (2014), pp. 371–383.
[18] X. Chen, L. Niu, and Y. Yuan, *Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization*, SIAM J. Optim., 23 (2013), pp. 1528–1552.
[19] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, *A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization*, SIAM J. Imaging Sci., 6 (2013), pp. 563–591.
[20] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM, New York, 1990.
[21] C. Cortes and Vapnik, *Support-vector networks*, Mach. Learn., 20 (1995), pp. 273–297.
[22] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inf. Theory, 52 (2006), pp. 1289–1306.
[23] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 9 (2001), pp. 1348–1360.
[24] J. Fan, L. Xue, and H. Zou, *Strong oracle optimization of folded concave penalized estimation*, Ann. Statist., 42 (2014), pp. 819–849.
[25] S. Foucart and M.-J. Lai, *Sparsest solutions of underdetermined linear system via ℓ_q-minimization for* $0 < q \leq 1$, Appl. Comput. Harmon. Anal., 26 (2009), pp. 395–407.
[26] G. M. Fung and O. L. Mangasarian, *Equivalence of minimal ℓ_0- and ℓ_p- norm solutions of linear equations, inequalities and linear programs for sufficiently small* p, J. Optim. Theory Appl., 153 (2011), pp. 1–10.
[27] D. Ghilli and K. Kunisch, *On monotone and primal-dual active set schemes for ℓ_p-type problems, $p \in (0, 1]$*, Comput. Optim. Appl., 72 (2019), pp. 45–85.
[28] 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.
[29] J. Huang, Y. Jiao, B. Jin, J. Liu, X. Lu, and C. Yang, *A unified primal dual active set algorithm for nonconvex sparse recovery.*
[30] P. J. Huber, *Robust Estimation*, Wiley, 1981.
[31] K. Ito and K. Karl, *A variational approach to sparsity optimization based on lagrange multiplier theory*, Inverse Probl., 30 (2014), 015001 (23 pages).
[32] Y. Jiao, B. Jin, and X. Lu, *A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem*, Appl. Comput. Harmon. Anal., 39 (2015), pp. 400–426.
[33] R. Koenker, *Quantile Regression*, Cambridge Univ. Press, 2005.
[34] J. Kreimer and R. Y. Rubinstein, *Nondifferentiable optimization via smooth approximation: general analytical approach*, Ann. Oper. Res., 39 (1992), pp. 97–119.
[35] M. Lai and J. Wang, *An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of under-determined linear systems*, SIAM J. Optim., 21 (2011), pp. 82–101.
[36] M. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization*, SIAM J. Numer. Anal., 51 (2013), pp. 927–957.
[37] H. A. Le Thi, T. Pham Dinh, H. M. Le, and X. T. Vo, *DC approximation approaches for sparse optimization*, Eur. J. Oper. Res., 244 (2015), pp. 26–46.
[38] Y. Liu, S. Ma, Y. Dai, and S. Zhang, *A smoothing SQP framework for a class of composite l_q minimization over polyhedron*, Math. Program., 158 (2016), pp. 467–500.
[39] Y. Liu and Y. Wu, *Variable selection via a combination of the ℓ_0 and ℓ_1 penalties*, J. Comput. Graph. Statist., 16 (2007), pp. 4036–4048.
[40] Z. Lu, *Iterative hard thresholding methods for ℓ_0 regularized convex programming*, Math. Program., 147 (2014), pp. 125–154.
[41] M. Nikolova and M. K. Ng, *Analysis of half-quadratic minimization methods for signal and image recovery*, SIAM J. Sci. Comput., 27 (2005), pp. 937–966.

[42] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 2006.

[43] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, *On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision*, SIAM J. Imaging Sci., 8 (2015), pp. 331–372.

[44] B. A. Olshausen and D. J. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature, 381 (1996), pp. 607–609.

[45] C. S. Ong and L. T. H. An, *Learning sparse classifiers with difference of convex functions algorithms*, Optim. Method Softw., 28 (2013), pp. 830–854.

[46] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth DC programs*, Math. Oper. Res., 42 (2017), pp. 95–118.

[47] D. Peleg and R. Meir, *A bilinear formulation for vector sparsity optimization*, Signal Process., 88 (2008), pp. 375–389.

[48] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, USA, 1972.

[49] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, Germany, 1998.

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

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

[52] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. B, 58 (1996), pp. 267–288.

[53] L. Wang, Y. Wu, and L. Runze, *Quantile regression for analyzing heterogeneity in ultra-high dimension*, Ann. Stat., 107 (2012), pp. 214–222.

[54] C. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Stat., 38 (2010), pp. 894–942.

[55] T. Zhang, *Multi-stage convex relaxation for feature selection*, Bernoulli, 19 (2013), pp. 2277–2293.

[56] Z. Zhang, Y. Fan, and J. Lv, *High dimensional thresholded regression and shrinkage effect*, J. R. Stat. Soc. Ser. B. Sta. Methodol., 76 (2014), pp. 627–649.