AIMS Mathematics, 9(12): 33401–33422.
DOI: 10.3934/math.20241594
Received: 22 August 2024
Revised: 04 November 2024
Accepted: 19 November 2024
Published: 25 November 2024

https://www.aimspress.com/journal/Math

Theory article

Smoothing gradient descent algorithm for the composite sparse optimization

Wei Yang, Lili Pan* and Jinhui Wan

Department of Mathematics, Shandong University of Technology, Zibo 255049, China

* **Correspondence:** Email: panlili1979@163.com.

Abstract: Composite sparsity generalizes the standard sparsity that considers the sparsity on a linear transformation of the variables. In this paper, we study the composite sparse optimization problem consisting of minimizing the sum of a nondifferentiable loss function and the ℓ_0 penalty term of a matrix times the coefficient vector. First, we consider an exact continuous relaxation problem with a capped-ℓ_1 penalty that has the same optimal solution as the primal problem. Specifically, we propose the lifted stationary point of the relaxation problem and then establish the equivalence of the original and relaxation problems. Second, we propose a smoothing gradient descent (SGD) algorithm for the continuous relaxation problem, which solves the subproblem inexactly since the objective function is inseparable. We show that if the sequence generated by the SGD algorithm has an accumulation point, then it is a lifted stationary point. At last, we present several computational examples to illustrate the efficiency of the algorithm.

Keywords: nonsmooth convex regression; cardinality penalty; gradient descent; smoothing method
Mathematics Subject Classification: 90C26, 90C30, 90C46, 65K05

1. Introduction

Sparse optimization is a core problem of compressed sensing [1–3], signal and image processing [4–7], and high-dimensional statistical learning [4, 8], etc. Sparsity is usually characterized by the ℓ_0 norm, which is the cardinality of the support set of vector $\mathbf{x} \in \mathbb{R}^n$, denoted by $\|\mathbf{x}\|_0 = |supp(\mathbf{x})| = |\{i \in \{1, 2, \cdots, n\} : \mathbf{x}_i \neq 0\}|$. The penalized formulation of sparse optimization can be expressed as the following cardinality regularized optimization:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) + \lambda\|\mathbf{x}\|_0, \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a loss function depending on the application, $\lambda > 0$ is the penalty parameter.

Compared with sparse optimization, composite sparse optimization problems enforce certain structural constraints instead of pure sparsity on the coefficients, which arise from many important

applications in various fields, such as structural health monitoring [10], fault diagnosis [11], motion planning [12] and impact force identification [13], etc. The most important method is to promote the sparsity of variables through linear transformation [14]. By imposing a regularization matrix $W = (W_1^\top, \ldots, W_p^\top)^\top \in \mathbb{R}^{p\times n}$ on vector $\mathbf{x}$, composite sparse optimization is nicely encapsulated as the following optimization formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \ f(\mathbf{x}) + \lambda\|W\mathbf{x}\|_0. \tag{1.2}$$

A typical choice of function f in problem (1.2) is the ℓ_2 loss function given by $f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - b\|_2^2$, where $A = (A_1^\top, \ldots, A_m^\top)^\top \in \mathbb{R}^{m\times n}$ and $b = (b_1, \ldots, b_m)^\top \in \mathbb{R}^m$, and the ℓ_1 relaxation of the ℓ_0 norm given by $\|W\mathbf{x}\|_1$, which was first defined and summarized as generalized LASSO with the general formulation of W [15]. Unlike traditional LASSO, solving generalized LASSO efficiently on high-dimensional data is very challenging. A few attempts have been made to improve the efficiency of generalized LASSO, but this requires a specific form of the W to work well [14–16], such as the fused LASSO problem [17], the TV regularizer [18] and trending filtering [19].

However, many loss functions of the composite sparse optimization problems cannot be expressed in the form of differentiable functions. The results in [20] showed that the least squares loss function can solve a class of linear regression problems but is not suitable for all types of data. We can choose the outlier that has a strong interference loss function, such as the ℓ_1 function, quantile regression function, or more general Huber class function. On the other hand, J. Fan et al. [20] pointed out that using the ℓ_1 relaxation often results in a biased estimator, various continuous nonconvex relaxation functions for ℓ_0 norm were proposed, such as the smoothly clipped absolute deviation (SCAD) function [20], the hard thresholding function [21], capped-ℓ_1 function [22–26], the transformed ℓ_1 function [27], etc. Here, we are interested in the capped-ℓ_1 function as the relaxation function of the ℓ_0 norm, which is a simple relaxation function that satisfies specific properties. Z. Shen et al. [40] applied locally Lipschitz continuous scaled folded concave functions to the approximate ℓ_0 pseudo-norm. A generic nonsmooth but convex framework was established to gradually approximate the scaled folded concave functions. Numerical experimental results showed the proposed framework and algorithms admitted the exact sparsity-induced capability of the ℓ_0 pseudo-norm. Q. Chen et al. [41] first explored using a class of locally Lipschitz scale folded concave functions to approach the ℓ_0. Then, a convex half-absolute method was proposed to precisely approximate these nonconvex nonsmooth functions. A double iterative algorithm was considered to solve the convex-relaxed composite optimization problems. Both [40] and [41] established a generic nonsmooth convex framework that gradually approximates these scale-folded concave functions based on the Legendre-Fenchel transformation to avoid directly solving a nonconvex optimization problem. However, we use a smoothing function for approximation to achieve the goal of solving the nonconvex optimization problem. The advantages of capped-ℓ_1 function have been explored in various fields. For example, the authors in [28] put forward a capped ℓ_1-norm Sparse Representation method (CSR) for graph clustering. The proposed model learned the optimal graph for clustering by integrating sparse representation and capped ℓ_1-norm loss function. In order to utilize the advantages of the twin extreme learning machine and FDA, Z. Xue et al. [29] first put forward a novel classifier named Fisher-regularized twin extreme learning machine (FTELM). Also considering the instability of the ℓ_2-norm for the outliers, authors introduced the capped ℓ_1-norm into the FTELM model and proposed a more robust capped ℓ_1-norm FTELM (Cℓ_1-FTELM) model. The capped ℓ_1 function was also discussed in the context of sparse group ℓ_0 regularized algorithms

by [30]. It's worth noting that reference [9] gave an exact continuous relaxation problem with capped-ℓ_1 penalty for nonsmooth convex loss function with cardinality penalty in the sense that both problems have the same optimal solution set. Moreover, a smoothing proximal gradient algorithm for finding a lifted stationary point of the continuous relaxation model was proposed. Regarding the solution of relaxation problems, T. Zhang [42] presented a multi-stage convex relaxation scheme for solving problems with non-convex objective functions. However, only parameter estimation performance was analyzed in [42]. Unfortunately, the result in [42] does not directly imply that multi-stage convex relaxation achieves unbiased recovery of the support set. H. Zhou et al. [43] proposed a new unified algorithm based on the local linear approximation (LLA) for maximizing the penalized likelihood for a broad class of concave penalty functions. It did not eliminate the bias issue. Here, we extend the results in [9] to composite sparse optimization and give a smoothing gradient descent algorithm for the continuous relaxation problem. The new algorithm exploits the piecewise linearity of the capped-ℓ_1 penalty term in the relaxation problem. In view of the composite sparsity, if the subproblem in the algorithm does not have a closed solution, then our relaxation problem model analogizes the ℓ_1 penalty model for solving LASSO problems, using the smoothing gradient method to solve the subproblem. We prove that if the sequence generated by the algorithm has an accumulation point, then it is a lifted stationary point of relaxation problem.

In this paper, we consider the following composite sparse regression problem with cardinality penalty,

$$\min_{\mathbf{x}\in\Omega} \quad \mathcal{W}_{\ell_0}(\mathbf{x}) := f(\mathbf{x}) + \lambda\|W\mathbf{x}\|_0, \tag{1.3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex (not necessarily smooth) and bounded from below function, λ is a positive parameter, and $\Omega = \{\mathbf{x} \in \mathbb{R}^n : l \le W\mathbf{x} \le u\}$. For example, the ℓ_1 loss function given by

$$f(\mathbf{x}) = \frac{1}{m}\|A\mathbf{x} - b\|_1, \tag{1.4}$$

or the censored regression problem with

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} |\max\{A_i\mathbf{x}, 0\} - b_i|. \tag{1.5}$$

For a given parameter $\nu > 0$, the continuous relaxation of the ℓ_0 penalty with the capped-ℓ_1 function ϕ is given by

$$\phi(t) = \min\Big\{1, \frac{|t|}{\nu}\Big\}. \tag{1.6}$$

We consider the following continuous optimization problem to solve (1.3):

$$\min_{\mathbf{x}\in\Omega} \quad \mathcal{W}(\mathbf{x}) := f(\mathbf{x}) + \lambda\Phi(W\mathbf{x}), \tag{1.7}$$

where $\Phi(W\mathbf{x}) = \sum_{i=1}^{p} \phi(W_i\mathbf{x})$.

Composite sparse optimization has attracted much attention recently. In [15], a dual path algorithm was proposed for the generalized LASSO problem with any formulation of W. If the composite optimization problem is convex and the W is the general linear map, one feasible approach is to apply the alternating direction method of multipliers (ADMM) [31]. In [31], the author proposed a dual method for the variational problem in the form of $\inf\{f(Av) + g(v)\}$. Z. J. Bai [32] aimed to provide a

coordinate gradient descent method with stepsize chosen by an Armijo-type rule to solve the problem $\min_{\mathbf{x}} f(\mathbf{x}) + c\|L\mathbf{x}\|_1$ and $\min_{\mathbf{x}} \|A\mathbf{x} - b\|_2^2 + c\|L\mathbf{x}\|_1$ efficiently, especially when the problems dimension is large.

In this paper, we use the exact continuous relaxation problem with capped-ℓ_1 function to solve optimization problem (1.3) and present a smoothing gradient descent (SGD) algorithm. Since W is a general linear mapping and $\Phi(W\mathbf{x})$ is an inseparable function, which makes the proximal gradient algorithm unable to be explicitly applied, we approximately solve the subproblem in the SGD algorithm. We prove that if there is an accumulation point, then the accumulation point is a lifted stationary point of (1.7).

Notation. We denote $\mathbb{N} = \{0, 1, \ldots\}$. For $\mathbf{x} \in \mathbb{R}^n$ and $\delta > 0$, let $\|\mathbf{x}\| := \|\mathbf{x}\|_2$ and $\mathbb{B}_\delta(\mathbf{x})$ means the open ball centered at $\mathbf{x}$ with radius δ. For a nonempty, closed, and convex set $\Omega \subseteq \mathbb{R}^n$, $N_\Omega(\mathbf{x})$ means the normal cone to Ω at $\mathbf{x} \in \Omega$. $\sigma_{\min}(W)$ is the minimum singular value of W. We denote $\|\mathbf{x}\|_\infty = \max\{|\mathbf{x}_1|, |\mathbf{x}_2|, \ldots, |\mathbf{x}_n|\}$.

2. An exact continuous relaxation for problem (1.3)

Before starting this section, we first make the following two assumptions:

Assumption 2.1. Function f is Lipschitz continuous on Ω with Lipschitz constant $L_f > 0$ and matrix W has full column rank.

Assumption 2.2. Parameter ν in (1.6) satisfies $0 < \nu < \overline{\nu} := \frac{\lambda\sigma_{\min}(W)}{L_f}$.

We suppose Assumptions 2.1 and 2.2 hold throughout the paper and assume that L_f is large enough such that $L_f \geq \frac{\lambda\sigma_{\min}(W)}{\Gamma}$, where

$$\Gamma := \min\{|l_i|, u_i : l_i \neq 0, u_i \neq 0, i = 1, \ldots, p\}.$$

When f is defined by the ℓ_1 loss function (1.4) or the censored regression loss function (1.5), L_f can be taken as $L_f = \|A\|_\infty$.

2.1. Lifted stationary points of problem (1.7)

We first give the definition of lifted stationary points of (1.7) as that in [33]. Since function ϕ in (1.6) can be rephrased as

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

with $\theta_1(t) = 0$, $\theta_2(t) = \frac{t}{\nu} - 1$ and $\theta_3(t) = -\frac{t}{\nu} - 1$ for $t \in \mathbb{R}$, we denote

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

and

$$\mathbb{D}^p(W\mathbf{x}) := \Pi_{i=1}^p \mathbb{D}(W_i\mathbf{x}). \tag{2.2}$$

Definition 2.1. Point $\mathbf{x} \in \Omega$ is called a lifted stationary point of (1.7) if there exists $d = (d_1, \ldots, d_p)^\top \in \mathbb{D}^p(W\mathbf{x})$ such that

$$\lambda \sum_{i=1}^{p} (\theta'_{d_i}(W_i\mathbf{x}))^\top \in \partial f(\mathbf{x}) + \frac{\lambda}{\nu} \sum_{i=1}^{p} W_i^\top \vartheta^i(\mathbf{x}) + N_\Omega(\mathbf{x}), \tag{2.3}$$

where

$$\vartheta^i(\mathbf{x}) \begin{cases} = 1 & \text{if } W_i\mathbf{x} > 0, \\ \in [-1, 1] & \text{if } |W_i\mathbf{x}| = 0, \\ = -1 & \text{if } W_i\mathbf{x} < 0. \end{cases} \tag{2.4}$$

Under the definition of the range of ν in Assumption 2.2, we first prove that the element in $\mathbb{D}^p(W\mathbf{x})$ for a lifted stationary point satisfying (2.3) is unique.

Proposition 2.2. If $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7), then the vector $d^{W\bar{\mathbf{x}}} = (d_1^{W\bar{\mathbf{x}}}, \ldots, d_p^{W\bar{\mathbf{x}}})^\top \in \mathbb{D}^p(W\bar{\mathbf{x}})$ satisfying (2.3) is unique. In particular, for $i = 1, \ldots, p$,

$$d_i^{W\bar{\mathbf{x}}} = \begin{cases} 1 & \text{if } |W_i\bar{\mathbf{x}}| < \nu, \\ 2 & \text{if } W_i\bar{\mathbf{x}} \geq \nu, \\ 3 & \text{if } W_i\bar{\mathbf{x}} \leq -\nu. \end{cases} \tag{2.5}$$

Proof. If $|W_i\bar{\mathbf{x}}| \neq \nu$, the statement is clearly valid. Hence, we only need to consider the case $|W_i\bar{\mathbf{x}}| = \nu$. When $W_i\bar{\mathbf{x}} = \nu$, since $\mathbb{D}(W_i\bar{\mathbf{x}}) = \{1, 2\}$, arguing by contradiction, we assume (2.3) holds with $d_i^{W\bar{\mathbf{x}}} = 1$, so $\vartheta^i(\bar{\mathbf{x}}) = 1$. By $\nu < \bar{\nu}$, we have $W_i\bar{\mathbf{x}} \in (l_i, u_i)$, and by (2.3), there exists $\xi(\bar{\mathbf{x}}) \in \partial f(\bar{\mathbf{x}})$ such that

$$\mathbf{0} = \xi(\bar{\mathbf{x}}) + \frac{\lambda}{\nu} \sum_{i:|W_i\bar{\mathbf{x}}| \leq \nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}), \tag{2.6}$$

where $\vartheta^i(\bar{\mathbf{x}})$ is defined as (2.4).

It is easy to observe that the following relation holds

$$\left\| \sum_{i:|W_i\bar{\mathbf{x}}| \leq \nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}) \right\|_2 \geq \sigma_{\min}(W). \tag{2.7}$$

In fact, from the definition of the minimum singular value of W,

$$\sigma_{\min}(W) = \min \left\{ \frac{\|W\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \mathbf{x} \neq \mathbf{0} \right\} = \min \left\{ \|W\mathbf{x}\|_2 : \|\mathbf{x}\|_2 = 1 \right\},$$

we have

$$\begin{aligned} \sigma_{\min}(W) &= \min \left\{ \frac{\|W\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \mathbf{x} \neq \mathbf{0} \right\} \\ &\leq \min \left\{ \frac{\|W\mathbf{x}\|_2}{\|\mathbf{x}\|_2} : \|\mathbf{x}\|_2 \geq 1, \|\mathbf{x}\|_\infty \leq 1 \right\} \\ &\leq \min \left\{ \|W\mathbf{x}\|_2 : \|\mathbf{x}\|_2 \geq 1, \|\mathbf{x}\|_\infty \leq 1 \right\} \\ &\leq \min \{ \|W\mathbf{x}\|_2 : \|\mathbf{x}\|_2 = 1 \} \\ &= \sigma_{\min}(W). \end{aligned}$$

Then, we see that

$$\sigma_{\min}(W) = \min\{\|W\mathbf{x}\|_2 : \|\mathbf{x}\|_2 \geq 1, \|\mathbf{x}\|_\infty \leq 1\}.$$

From the definition of $\vartheta^i(\bar{\mathbf{x}})$ (2.4), this yields that

$$\left\| \sum_{i:|W_i\bar{\mathbf{x}}|\leq\nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}) \right\|_2 \geq \sigma_{\min}(W_I) \geq \sigma_{\min}(W),$$

where W_I is the submatrix consisting of the rows in W indexed by $I := \{i : |W_i\bar{\mathbf{x}}| \leq \nu\}$ [34].

Combining (2.6) and (2.7), we have

$$\frac{\lambda\sigma_{\min}(W)}{\nu} \leq \frac{\lambda}{\nu} \left\| \sum_{i:|W_i\bar{\mathbf{x}}|\leq\nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}) \right\| = \|\xi(\bar{\mathbf{x}})\| \leq L_f.$$

This leads to a contradiction to $\nu < \frac{\lambda\sigma_{\min}(W)}{L_f}$. Then, (2.5) holds for $W_i\bar{\mathbf{x}} = \nu$. Similar analysis can be given for the case that $W_i\bar{\mathbf{x}} = -\nu$, which completes the proof. □

For a given $d = (d_1, \ldots, d_p)^\top \in \mathbb{D}^p := \{d \in \mathbb{R}^p : d_i \in \{1, 2, 3\}, i = 1, \cdots, p\}$, we define

$$\Phi^d(W\mathbf{x}) := \sum_{i=1}^{p} \frac{|W_i\mathbf{x}|}{\nu} - \sum_{i=1}^{p} \theta_{d_i}(W_i\mathbf{x}), \tag{2.8}$$

which is convex with respect to $\mathbf{x}$. It can be verified that

$$\Phi(W\mathbf{x}) = \min_{d\in\mathbb{D}^p} \Phi^d(W\mathbf{x}), \ \forall \mathbf{x} \in \Omega.$$

In particular, for a fixed $\bar{\mathbf{x}} \in \Omega$, $\Phi(W\bar{\mathbf{x}}) = \Phi^{d^{W\bar{\mathbf{x}}}}(W\bar{\mathbf{x}})$ and the following relation holds

$$\bar{\mathbf{x}} \text{ is a lifted stationary point of (1.7)} \Leftrightarrow \bar{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\Omega} f(\mathbf{x}) + \lambda\Phi^{d^{W\bar{\mathbf{x}}}}(W\mathbf{x}). \tag{2.9}$$

Next lemma describes a lower bound property.

Lemma 2.3. If $\bar{\mathbf{x}} \in \Omega$ is a lifted stationary point of (1.7), then it holds that

$$W_i\bar{\mathbf{x}} \in (-\nu, \nu) \ \Rightarrow \ W_i\bar{\mathbf{x}} = 0, \quad \forall i = 1, \ldots, p. \tag{2.10}$$

Proof. Suppose that $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7). Now we assume that $W_i\bar{\mathbf{x}} \in (-\nu, \nu) \setminus \{0\}$ for some $i \in 1, \ldots, p$. So from (2.5) and Assumption 2.1, we have $d_i^{W\bar{\mathbf{x}}} = 1$ and $W_i\bar{\mathbf{x}} \in (l_i, u_i)$. By (2.3), there exists $\xi(\bar{\mathbf{x}}) \in \partial f(\bar{\mathbf{x}})$. We have

$$\mathbf{0} = \xi(\bar{\mathbf{x}}) + \frac{\lambda}{\nu} \sum_{i:|W_i\bar{\mathbf{x}}|<\nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}).$$

Through the analysis in the proof of Proposition 2.2, combining (2.6) and (2.7), we have

$$\frac{\lambda\sigma_{\min}(W)}{\nu} \leq \frac{\lambda}{\nu} \left\| \sum_{i:|W_i\bar{\mathbf{x}}|<\nu} W_i^\top \vartheta^i(\bar{\mathbf{x}}) \right\| = \|\xi(\bar{\mathbf{x}})\| \leq L_f,$$

which leads to a contradiction to $\nu < \frac{\lambda\sigma_{\min}(W)}{L_f}$. Consequently, $W_i\bar{\mathbf{x}} \in (-\nu, \nu)$ implies $W_i\bar{\mathbf{x}} = 0$ for $i \in 1, \ldots, p$ and the proof is completed. □

2.2. The relationship between problems (1.3) *and* (1.7)

This subsection discusses the relationship between the global minimizers and local minimizers of (1.3) and (1.7). First, Theorem 2.4 discusses the relationship between the local minimizers of (1.3) and (1.7). Second, Theorem 2.5 states that (1.3) and (1.7) have the same global minimizers. We use the lower bound property mentioned in Lemma 2.3 to prove Theorems 2.4 and 2.5.

Theorem 2.4. If $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7), then it is a local minimizer of (1.3) and the objective functions have the same value at $\bar{\mathbf{x}}$, i.e., $f(\bar{\mathbf{x}}) + \lambda\Phi(W\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \lambda\|W\bar{\mathbf{x}}\|_0$.

Proof. Combining the lower bound property of $W\bar{\mathbf{x}}$ in (2.10) and the definition of $\Phi^{d^{W\bar{\mathbf{x}}}}$ defined in (2.8), for any $\mathbf{x} \in \mathbb{R}^n$, we have

$$\begin{aligned}\Phi^{d^{W\bar{\mathbf{x}}}}(W\mathbf{x}) :&= \sum_{i=1}^{p} \frac{|W_i\mathbf{x}|}{\nu} - \sum_{i=1}^{p} \theta_{d_i^{W\bar{\mathbf{x}}}}(W_i\mathbf{x}) \\ &= \sum_{i:|W_i\bar{\mathbf{x}}|\geq\nu} 1 + \sum_{i:|W_i\bar{\mathbf{x}}|<\nu} \frac{|W_i\mathbf{x}|}{\nu} \\ &= \|W\bar{\mathbf{x}}\|_0 + \sum_{i:W_i\bar{\mathbf{x}}=0} \frac{|W_i\mathbf{x}|}{\nu}.\end{aligned}$$

Then

$$\Phi^{d^{W\bar{\mathbf{x}}}}(W\mathbf{x}) \leq \|W\mathbf{x}\|_0, \ \forall \mathbf{x} \in \mathbb{B}_\varrho(\bar{\mathbf{x}}), \ \varrho > 0. \tag{2.11}$$

Combining this with $\Phi(W\bar{\mathbf{x}}) = \|W\bar{\mathbf{x}}\|_0$ and (2.9), we have

$$f(\bar{\mathbf{x}}) + \lambda\|W\bar{\mathbf{x}}\|_0 \leq f(\mathbf{x}) + \lambda\|W\mathbf{x}\|_0, \quad \forall \mathbf{x} \in \Omega \cap \mathbb{B}_\varrho(\bar{\mathbf{x}}).$$

Thus, $\bar{\mathbf{x}}$ is a local minimizer of (1.7). □

Theorem 2.4 indicates that any lifted stationary point of (1.7) is a local minimizer of (1.3), which means that any local minimizer of (1.7) is also certainly a local minimizer of (1.3).

Theorem 2.5. If $\bar{\mathbf{x}} \in \Omega$ is a global minimizer of (1.3) if and only if it is a global minimizer of (1.7). Moreover, problems (1.3) and (1.7) have the same optimal value.

Proof. On the one hand, let $\bar{\mathbf{x}} \in \Omega$ be a global minimizer of (1.7), and according to Definition 2.1, then we can obtain that $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7). By (2.10), from $W_i\bar{\mathbf{x}} \in (-\nu, \nu)$, then $W_i\bar{\mathbf{x}} = 0$, so it gives $\Phi(W\bar{\mathbf{x}}) = \|W\bar{\mathbf{x}}\|_0$. We have

$$f(\bar{\mathbf{x}}) + \lambda\|W\bar{\mathbf{x}}\|_0 = f(\bar{\mathbf{x}}) + \lambda\Phi(W\bar{\mathbf{x}}) \leq \ f(\mathbf{x}) + \lambda\Phi(W\mathbf{x}) \leq f(\mathbf{x}) + \lambda\|W\mathbf{x}\|_0, \forall \mathbf{x} \in \Omega,$$

where the last inequality uses $\Phi(W\mathbf{x}) \leq \|W\mathbf{x}\|_0$. Therefore, $\bar{\mathbf{x}}$ is a global minimizer of (1.3).

On the other hand, let $\bar{\mathbf{x}} \in \Omega$ be a global minimizer of (1.3). Assume on the contrary $\bar{\mathbf{x}}$ is not a solution of (1.7). Let $\hat{\mathbf{x}}$ be a global minimizer of (1.7), we obtain

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

From

$$\Phi(W\hat{\mathbf{x}}) = \|W\hat{\mathbf{x}}\|_0 \ \text{ and } \ \Phi(W\bar{\mathbf{x}}) \leq \|W\bar{\mathbf{x}}\|_0,$$

we have

$$f(\hat{\mathbf{x}}) + \lambda\|W\hat{\mathbf{x}}\|_0 < f(\bar{\mathbf{x}}) + \lambda\|W\bar{\mathbf{x}}\|_0.$$

This contradicts the global optimality of $\bar{\mathbf{x}}$ for (1.3). Hence $\bar{\mathbf{x}}$ is a global minimizer of (1.7). Therefore, (1.3) and (1.7) have the same global minimizers and optimal values. □

When f is convex, $\bar{\mathbf{x}}$ is a local minimizer of (1.3) if and only if $\bar{\mathbf{x}} \in \Omega$ satisfies

$$\mathbf{0} \in \partial f(\bar{\mathbf{x}}) + N_\Omega(\bar{\mathbf{x}}). \tag{2.12}$$

which is often used as a criterion for the local minimizers of problem (1.3).

Definition 2.6. We call $\bar{\mathbf{x}} \in \Omega$ a ν-strong local minimizer of (1.3), if there exists $\overline{\xi} \in \partial f(\bar{\mathbf{x}})$ and $\overline{\eta} \in N_\Omega(\bar{\mathbf{x}})$ such that for any $i \in supp(W\bar{\mathbf{x}})$, it holds

$$\mathbf{0} = \overline{\xi} + \overline{\eta} \ \text{ and } \ |W_i\bar{\mathbf{x}}| \geq \nu.$$

By (2.12), any ν-strong local minimizer of (1.3) is a local minimizer of it. Below we provide a result on the relationship between the ν-strong local minimizers of (1.3) and the lifted stationary points of (1.7).

Proposition 2.7. $\bar{\mathbf{x}} \in \Omega$ is a ν-strong local minimizer of (1.3) if and only if it is a lifted stationary point of (1.7).

Proof. First, by (2.9), we see that if $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7), then

$$\mathcal{W}_{\ell_0}(\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \lambda\|W\bar{\mathbf{x}}\|_0 = f(\bar{\mathbf{x}}) + \lambda\Phi(W\bar{\mathbf{x}}) = f(\bar{\mathbf{x}}) + \lambda\Phi^{d^{W\bar{\mathbf{x}}}}(W\bar{\mathbf{x}}) \leq f(\mathbf{x}) + \lambda\Phi^{d^{W\bar{\mathbf{x}}}}(W\mathbf{x}), \quad \forall \mathbf{x} \in \Omega.$$

Combining the Lemma 2.3 and $\Phi^{d^{W\bar{\mathbf{x}}}}(W\mathbf{x}) \leq \|W\mathbf{x}\|_0, \ \forall \mathbf{x} \in \mathbb{B}_\varrho(\bar{\mathbf{x}}), \ \varrho > 0$ in (2.11), then we have

$$\mathcal{W}_{\ell_0}(\bar{\mathbf{x}}) \leq \mathcal{W}_{\ell_0}(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega \cap \mathbb{B}_\varrho(\bar{\mathbf{x}}),$$

so $\bar{\mathbf{x}}$ is a ν-strong local minimizer of (1.3).
Next, because $\bar{\mathbf{x}}$ is a ν-strong local minimizer of (1.3), it is also a local minimizer of (1.3), suppose $\bar{\mathbf{x}}$ is a local minimizer of (1.3) but not a local minimizer of (1.7). Then there exists a local minimizer of (1.7) denoted by $\hat{\mathbf{x}}$, combining (2.9), $\Phi^{d^{W\hat{\mathbf{x}}}}(W\bar{\mathbf{x}}) \leq \|W\bar{\mathbf{x}}\|_0, \ \forall \bar{\mathbf{x}} \in \mathbb{B}_\varrho(\hat{\mathbf{x}})$ in (2.11) and $\Phi(W\hat{\mathbf{x}}) = \|W\hat{\mathbf{x}}\|_0$, we have

$$f(\hat{\mathbf{x}})+\lambda\|W\hat{\mathbf{x}}\|_0 = f(\hat{\mathbf{x}})+\lambda\Phi(W\hat{\mathbf{x}}) = f(\hat{\mathbf{x}})+\lambda\Phi^{d^{W\hat{\mathbf{x}}}}(W\hat{\mathbf{x}}) \leq f(\bar{\mathbf{x}})+\lambda\Phi^{d^{W\hat{\mathbf{x}}}}(W\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}})+\lambda\|W\bar{\mathbf{x}}\|_0, \ \forall \bar{\mathbf{x}} \in \mathbb{B}_\varrho(\hat{\mathbf{x}}),$$

which leads to a contradiction. Thus, the local minimizer of (1.3) is the local minimizer of (1.7), that is to say, the ν-strong local minimizer of (1.3) is the lifted stationary point of (1.7). □

We use Table 1 to clearly demonstrate the relationship between (1.3) and (1.7).

Table 1. Link between problems (1.3) and (1.7).

Continuous relaxation problem (1.7)		**Cardinality penalty problem** (1.3)
global minimizer	$\Longleftrightarrow$	global minimizer
local minimizer	$\Longrightarrow$	local minimizer
$\Updownarrow$		$\Uparrow$
lifted stationary point	$\Longrightarrow$	local minimizer and satisfying lower bound in (2.10)
		$\Updownarrow$
	$\Longleftrightarrow$	$\nu-$ strong local minimizer

3. Smoothing gradient descent algorithm (SGD)

The main content of this section is to find the lifted stationary point of (1.7). Due to the existence of matrix W, we cannot express the explicit solution using the proximal gradient method. We first approximate f by a smoothing function and propose some preliminary theories on the smoothing methods; the second section proposes our algorithm; and the third section conducts a convergence analysis on the proposed algorithm for solving (1.7).

3.1. Smoothing approximation method

Throughout this paper, we approximate the loss function f by a smoothing function $\tilde{f}$ in (1.7). When it is clear from the context, the derivative of $\tilde{f}(\mathbf{x},\mu)$ with respect to $\mathbf{x}$ is simply denoted as $\nabla\tilde{f}(\mathbf{x},\mu)$. We denote

$$\widetilde{\mathcal{W}}^d(\mathbf{x},\mu) := \tilde{f}(\mathbf{x},\mu) + \lambda\Phi^d(W\mathbf{x}),\ \ \widetilde{\mathcal{W}}(\mathbf{x},\mu) := \tilde{f}(\mathbf{x},\mu) + \lambda\Phi(W\mathbf{x}),$$

where smoothing parameter $\mu > 0$ and $d \in \mathbb{D}^p$. For any fixed $\mu > 0$ and $d \in \mathbb{D}^p$, $\widetilde{\mathcal{W}}^d(\mathbf{x},\mu)$ is nonsmooth and convex, and $\widetilde{\mathcal{W}}(\mathbf{x},\mu)$ is nonsmooth and nonconvex. Due to

$$\Phi(W\mathbf{x}) = \min_{d\in\mathbb{D}^p} \Phi^d(W\mathbf{x}),\ \forall \mathbf{x} \in \Omega,$$

we obtain

$$\widetilde{\mathcal{W}}^d(\mathbf{x},\mu) \geq \widetilde{\mathcal{W}}(\mathbf{x},\mu),\ \forall d \in \mathbb{D}^p,\ \mathbf{x} \in \Omega,\ \mu \in (0,\bar{\mu}]. \tag{3.1}$$

The following definition describes some theories about the smoothing function $\tilde{f}$, which is frequently used in the proof of convergence analysis.

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 (1.7), if $\tilde{f}(\cdot,\mu)$ is continuously differentiable in $\mathbb{R}^n$ for any fixed $\mu > 0$ and satisfies the following conditions:
(i) $\lim_{\mathbf{z}\to\mathbf{x},\mu\downarrow 0} \tilde{f}(\mathbf{z},\mu) = f(\mathbf{x}), \forall \mathbf{x} \in \Omega$;

(ii) (convexity) $\tilde{f}(\mathbf{x},\mu)$ is convex with respect to $\mathbf{x}$ in Ω for any fixed $\mu > 0$;
(iii) (gradient consistency) $\{\lim_{\mathbf{z}\to\mathbf{x},\mu\downarrow 0} \nabla_{\mathbf{z}}\tilde{f}(\mathbf{z},\mu)\} \subseteq \partial f(\mathbf{x}), \forall \mathbf{x} \in \Omega$;
(iv) (Lipschitz continuity with respect to μ) there exists a positive constant κ such that

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

(v) (Lipschitz continuity with respect to $\mathbf{x}$) there exists a constant $L > 0$ such that for any $\mu \in (0,\bar{\mu}]$, $\nabla_{\mathbf{x}}\tilde{f}(\cdot,\mu)$ is Lipschitz continuous on Ω with Lipschitz constant $L\mu^{-1}$.

By virtue of Definition 3.1-(iv), we obtain that

$$|\tilde{f}(\mathbf{x},\mu) - f(\mathbf{x})| \leq \kappa\mu, \forall \mathbf{x} \in \Omega, 0 < \mu \leq \bar{\mu}. \tag{3.2}$$

Next, we aim to solve the following problem with $\mu > 0$ and vector $d \in \mathbb{D}^p$

$$\min_{\mathbf{x}\in\Omega} \quad \widetilde{\mathcal{W}}^d(\mathbf{x},\mu) = \tilde{f}(\mathbf{x},\mu) + \lambda\Phi^d(W\mathbf{x}), \tag{3.3}$$

by introducing an approximation of $\widetilde{\mathcal{W}}^d(\cdot,\mu)$ around a given point $\mathbf{z}$ as follows:

$$G_{d,\gamma}(\mathbf{x},\mathbf{z},\mu) = \tilde{f}(\mathbf{z},\mu) + \langle \mathbf{x}-\mathbf{z}, \nabla\tilde{f}(\mathbf{z},\mu)\rangle + \frac{1}{2}\gamma\mu^{-1}\|\mathbf{x}-\mathbf{z}\|^2 + \lambda\Phi^d(W\mathbf{x}) \tag{3.4}$$

with a constant $\gamma > 0$. $\Phi^d(W\mathbf{x})$ is convex with respect to $\mathbf{x}$ for any fixed $d \in \mathbb{D}^p$, function $G_{d,\gamma}(\mathbf{x},\mathbf{z},\mu)$ is a strongly convex function with respect to $\mathbf{x}$ for any fixed $d, \gamma, \mathbf{z}$ and μ. Then, we solve the following problem

$$\min_{\mathbf{x}\in\Omega} G_{d,\gamma}(\mathbf{x},\mathbf{z},\mu)$$

to find the approximate solution of (3.3).

3.2. Smoothing gradient descent algorithm

In this subsection, we propose a new algorithm (see Algorithm 1) for finding a lifted stationary point of (1.7). Specially, since W is a general linear mapping and $\Phi(W\mathbf{x})$ is an inseparable function, which makes the smoothing proximal gradient algorithm [9] cannot explicitly solve a subproblem. The proposed algorithm combines the smoothing method and the smoothing gradient descent algorithm, so we call it the smoothing gradient descent (SGD) algorithm. We use the SGD algorithm to obtain approximate solutions of the subproblem. Let

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

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

$$\mu_k = \mu_{p_r^s+1} = \frac{\mu_0}{(p_r^s+1)^\sigma}, \forall p_r^s + 1 \leq k \leq p_{r+1}^s, \tag{3.5}$$

which will be used in the proof of Lemmas 3.2 and 3.4.

Algorithm 1 Smoothing Gradient Descent (SGD) algorithm

Require: Take $\mathbf{x}^{-1} = \mathbf{x}^0 \in \Omega$ and $\mu_{-1} = \mu_0 \in (0, \bar{\mu}]$. Give parameters $\rho > 1$, $\sigma > \frac{1}{2}$, $\alpha > 0$ and $0 < \underline{\gamma} < \bar{\gamma}$. Set $k = 0$.

while a termination criterion is not met **do**

Step 1. Choose $\gamma_k \in [\underline{\gamma}, \bar{\gamma}]$ and let $d^k \triangleq d^{W\mathbf{x}^k}$, where $d^{W\mathbf{x}^k}$ is defined in (2.5).

Step 2.

2a) Compute

$$\hat{\mathbf{x}}^{k+1} = \arg\min_{\mathbf{x}\in\Omega} G_{d^k,\gamma^k}(\mathbf{x}, \mathbf{x}^k, \mu_k). \tag{3.6}$$

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

$$\widetilde{\mathcal{W}}^{d^k}(\hat{\mathbf{x}}^{k+1}, \mu_k) \le G_{d^k,\gamma^k}(\hat{\mathbf{x}}^{k+1}, \mathbf{x}^k, \mu_k). \tag{3.7}$$

Set

$$\mathbf{x}^{k+1} = \hat{\mathbf{x}}^{k+1} \tag{3.8}$$

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

Step 3. If

$$\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k - \widetilde{\mathcal{W}}(\mathbf{x}^k, \mu_{k-1}) - \kappa\mu_{k-1} \le -\alpha\mu_k^2, \tag{3.9}$$

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

$$\mu_{k+1} = \frac{\mu_0}{(k+1)^\sigma}. \tag{3.10}$$

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

end while

3.3. Convergence analysis

Lemma 3.2. The proposed SGD algorithm is well-defined, and the sequences $\{\mathbf{x}^k\}$, $\{\gamma^k\}$ and $\{\mu_k\}$ generated by it own the following properties:
(i) $\{\mathbf{x}^k\} \subseteq \Omega$ and $\{\gamma_k\} \subseteq [\underline{\gamma}, \max\{\overline{\gamma}, \rho L\}]$;
(ii) there are infinite elements in $\mathcal{P}^s$ and $\lim_{k\to\infty} \mu_k = 0$.

Proof. (i) By organizing (3.7), we can obtain

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

According to Definition 3.1-(v), (3.7) holds when $\gamma_k \ge L$. Thus the updating of γ_k in Step 2 is at most $\log_\rho(\frac{L}{\underline{\gamma}}) + 1$ times at each iteration. Hence, the SGD algorithm is well-defined, and we have that

$\gamma_k \leq \max\{\overline{\gamma}, \rho L\}, \forall k \in \mathbb{N}$. From (3.8), it is easy to verify that $\mathbf{x}^{k+1} \in \Omega$ by $\mathbf{x}^k \in \Omega$ and $\hat{\mathbf{x}}^{k+1} \in \Omega$.
(ii) Since $\{\mu_k\}$ is non-increasing, to prove (ii), we assume that $\lim_{k\to\infty} \mu_k = \hat{\mu} > 0$ by contradiction. If $\{\mu_k\}$ converges to a non-zero value, then the iteration of (3.10) is finite, which means that there exists $\mathrm{K} \in \mathbb{N}$ such that $\mu_k = \hat{\mu}, \forall k \geq \mathrm{K}$. Substituting $\hat{\mu}$ into (3.9), we obtain

$$\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k - \widetilde{\mathcal{W}}(\mathbf{x}^k, \mu_{k-1}) - \kappa\mu_{k-1} \leq -\alpha\hat{\mu}^2, \forall k \geq \mathrm{K}+1.$$

By the above inequality, we have

$$\lim_{k\to\infty} \widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k = -\infty. \tag{3.11}$$

However, by $\{\mathbf{x}^k\} \subseteq \Omega$, (3.2) and Theorem 2.5, then

$$\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k \geq \mathcal{W}(\mathbf{x}^{k+1}) \geq \min_{\mathbf{x}\in\Omega} \mathcal{W}(\mathbf{x}) = \min_{\mathbf{x}\in\Omega} \mathcal{W}_{\ell_0}(\mathbf{x}), \quad \forall k \geq \mathrm{K}, \tag{3.12}$$

(3.11) and (3.12) are contradictory; (ii) holds. □

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

$$\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) - \widetilde{\mathcal{W}}(\mathbf{x}^k, \mu_k) \leq -\frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \tag{3.13}$$

which implies $\{\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k\}$ is non-increasing and $\lim_{k\to\infty} \widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) = \lim_{k\to\infty} \mathcal{W}(\mathbf{x}^k)$.

Proof. Since $G_{d^k,\gamma_k}(\mathbf{x}, \mathbf{x}^k, \mu_k)$ is strongly convex with modulus $\gamma_k\mu_k^{-1}$, we have

$$\begin{aligned} G_{d^k,\gamma_k}(\mathbf{x}, \mathbf{x}^k, \mu_k) \geq G_{d^k,\gamma_k}(\hat{\mathbf{x}}^{k+1}, \mathbf{x}^k, \mu_k) + \langle \nabla G_{d^k,\gamma_k}(\hat{\mathbf{x}}^{k+1}, \mathbf{x}^k, \mu_k), \mathbf{x} - \hat{\mathbf{x}}^{k+1} \rangle \\ + \frac{1}{2}\gamma_k\mu_k^{-1}\|\hat{\mathbf{x}}^{k+1} - \mathbf{x}\|^2, \end{aligned} \tag{3.14}$$

using the definition of $\hat{\mathbf{x}}^{k+1}$ in (3.6) and $\mathbf{x}^{k+1} = \hat{\mathbf{x}}^{k+1}$ when (3.7) holds, we obtain

$$G_{d^k,\gamma_k}(\mathbf{x}^{k+1}, \mathbf{x}^k, \mu_k) \leq G_{d^k,\gamma_k}(\mathbf{x}, \mathbf{x}^k, \mu_k) - \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2, \forall \mathbf{x} \in \Omega.$$

By the definition of function G_{d^k,γ_k} given in (3.4), organizing the inequalities above, we have

$$\begin{aligned} \lambda\Phi^{d^k}(W\mathbf{x}^{k+1}) \leq & \lambda\Phi^{d^k}(W\mathbf{x}) + \langle \mathbf{x} - \mathbf{x}^{k+1}, \nabla\tilde{f}(\mathbf{x}^k, \mu_k) \rangle \\ & + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x} - \mathbf{x}^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\ & - \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}\|^2. \end{aligned} \tag{3.15}$$

Moreover, (3.7) can be written as

$$\begin{aligned} \widetilde{\mathcal{W}}^{d^k}(\mathbf{x}^{k+1}, \mu_k) \leq & \tilde{f}(\mathbf{x}^k, \mu_k) + \langle \mathbf{x}^{k+1} - \mathbf{x}^k, \nabla\tilde{f}(\mathbf{x}^k, \mu_k) \rangle \\ & + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \lambda\Phi^{d^k}(W\mathbf{x}^{k+1}). \end{aligned} \tag{3.16}$$

Summing up (3.15) and (3.16), we obtain that

$$\begin{aligned}\widetilde{\mathcal{W}}^{d^k}(\mathbf{x}^{k+1},\mu_k) \leq & \tilde{f}(\mathbf{x}^k,\mu_k) + \lambda\Phi^{d^k}(W\mathbf{x}) + \langle \mathbf{x}-\mathbf{x}^k, \nabla\tilde{f}(\mathbf{x}^k,\mu_k)\rangle \\ & + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}-\mathbf{x}^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}\|^2, \forall \mathbf{x}\in\Omega.\end{aligned} \tag{3.17}$$

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

$$\tilde{f}(\mathbf{x}^k,\mu_k) + \langle \mathbf{x}-\mathbf{x}^k, \nabla\tilde{f}(\mathbf{x}^k,\mu_k)\rangle \leq \tilde{f}(\mathbf{x},\mu_k), \forall \mathbf{x}\in\Omega. \tag{3.18}$$

Combining (3.17) and (3.18) and utilizing the definition of $\widetilde{\mathcal{W}}^{d^k}$, one has

$$\begin{aligned}\widetilde{\mathcal{W}}^{d^k}(\mathbf{x}^{k+1},\mu_k) \leq \widetilde{\mathcal{W}}^{d^k}(\mathbf{x},\mu_k) + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}-\mathbf{x}^k\|^2 - \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}\|^2, \\ \forall \mathbf{x}\in\Omega.\end{aligned} \tag{3.19}$$

Letting $\mathbf{x} = \mathbf{x}^k$ in (3.19) and by $d^k = d^{W\mathbf{x}^k}$, we obtain

$$\widetilde{\mathcal{W}}^{d^k}(\mathbf{x}^{k+1},\mu_k) + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq \widetilde{\mathcal{W}}(\mathbf{x}^k,\mu_k). \tag{3.20}$$

Because $\widetilde{\mathcal{W}}^{d^k}(\mathbf{x}^{k+1},\mu_k) \geq \widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k)$, (3.20) leads to (3.13). Due to Definition 3.1-(iv), we have

$$\tilde{f}(\mathbf{x}^k,\mu_{k-1}) \geq \tilde{f}(\mathbf{x}^k,\mu_k) - \kappa(\mu_{k-1}-\mu_k),$$

then, it follows that

$$\widetilde{\mathcal{W}}(\mathbf{x}^k,\mu_k) \leq \widetilde{\mathcal{W}}(\mathbf{x}^k,\mu_{k-1}) + \kappa(\mu_{k-1}-\mu_k),$$

by (3.13), we obtain

$$\widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k) + \kappa\mu_k + \frac{1}{2}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq \widetilde{\mathcal{W}}(\mathbf{x}^k,\mu_{k-1}) + \kappa\mu_{k-1}, \tag{3.21}$$

(3.21) implies the non-increasing property of $\{\widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k) + \kappa\mu_k\}$. This result and (3.12) ensure the existence of $\lim_{k\to\infty}\widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k) + \kappa\mu_k$. By virtue of $\lim_{k\to\infty}\mu_k = 0$ and Definition 3.1-(i), we obtain

$$\lim_{k\to\infty}\widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k) = \lim_{k\to\infty}\mathcal{W}(\mathbf{x}^k).$$

The proof is completed. □

Lemma 3.4. The following statements hold:
(i) $\sum_{k=0}^{\infty}\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq 2(\mathcal{W}(\mathbf{x}^0,\mu_{-1}) + \kappa\mu_{-1} - \min_\Omega\mathcal{W})$;
(ii) $\sum_{k=0}^{\infty}\mu_k^2 \leq \Lambda$ with $\Lambda = \frac{1}{\alpha}(\widetilde{\mathcal{W}}(\mathbf{x}^0,\mu_{-1}) + \kappa\mu_{-1} - \min_{\mathbf{x}\in\Omega}\mathcal{W}(\mathbf{x})) + \frac{2\mu_0^2\sigma}{2\sigma-1} < \infty$;

Proof. (i) Recalling (3.21), for all $k\in\mathbb{N}$, we obtain

$$\gamma_k\mu_k^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2 \leq 2(\widetilde{\mathcal{W}}(\mathbf{x}^k,\mu_{k-1}) + \kappa\mu_{k-1} - \widetilde{\mathcal{W}}(\mathbf{x}^{k+1},\mu_k) - \kappa\mu_k). \tag{3.22}$$

Now adding up the above inequality over $k = 0, \ldots, \mathrm{K}$, it gives

$$\sum_{k=0}^{\mathrm{K}} \gamma_k \mu_k^{-1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \leq 2(\widetilde{\mathcal{W}}(\mathbf{x}^0, \mu_{-1}) + \kappa\mu_{-1} - \widetilde{\mathcal{W}}(\mathbf{x}^{\mathrm{K}+1}, \mu_{\mathrm{K}}) - \kappa\mu_{\mathrm{K}}). \tag{3.23}$$

By letting K in (3.23) tend to infinity and along with (3.12), we obtain (i).

(ii) From (3.5), we have

$$\sum_{k\in\mathcal{P}^s} \mu_k^2 = \sum_{r=1}^{\infty} \frac{\mu_0^2}{(p_r^s + 1)^{2\sigma}} \leq \sum_{k=1}^{\infty} \frac{\mu_0^2}{k^{2\sigma}} \leq \frac{2\mu_0^2\sigma}{2\sigma - 1}, \tag{3.24}$$

where p_r^s is the rth smallest element in $\mathcal{P}^s$. When $k \notin \mathcal{P}^s$, (3.9) gives

$$\alpha\mu_k^2 \leq \widetilde{\mathcal{W}}(\mathbf{x}^k, \mu_{k-1}) + \kappa\mu_{k-1} - \widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) - \kappa\mu_k,$$

which together with the non-increasing property of $\{\widetilde{\mathcal{W}}(\mathbf{x}^{k+1}, \mu_k) + \kappa\mu_k\}$ and (3.12) implies

$$\sum_{k\notin\mathcal{P}^s} \mu_k^2 \leq \frac{1}{\alpha}(\widetilde{\mathcal{W}}(\mathbf{x}^0, \mu_{-1}) + \kappa\mu_{-1} - \min_{\Omega} \mathcal{W}). \tag{3.25}$$

Combining (3.24) and (3.25), the proof of (ii) is completed. □

Theorem 3.5. If there is an accumulation point in $\{\mathbf{x}^k : k \in \mathcal{P}^s\}$, then the accumulation point is a lifted stationary point of (1.7).

Proof. Since (3.9) fails for $k \in \mathcal{P}^s$, by rearranging (3.21), we obtain that

$$\gamma_k \mu_k^{-1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \leq 2\alpha\mu_k^2,$$

which gives

$$\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leq \sqrt{2\alpha\gamma_k^{-1}\mu_k^3}.$$

Thus,

$$\gamma_k \mu_k^{-1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| \leq \sqrt{2\alpha\gamma_k\mu_k},$$

which together with $\lim_{k\to\infty} \mu_k = 0$ and $\{\gamma_k\} \subseteq [\underline{\gamma}, \max\{\bar{\gamma}, \rho L\}]$ implies

$$\lim_{k\to\infty} \gamma_k \mu_k^{-1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0 \ \text{ and } \lim_{k\to\infty} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| = 0. \tag{3.26}$$

Let $\bar{\mathbf{x}}$ be an accumulation point of $\{\mathbf{x}^k\}_{k\in\mathcal{P}^s}$, (3.26) indicates that $\{\mathbf{x}^k\}$ exists a subsequence $\{\mathbf{x}^{k_t}\}_{k_t\in\mathcal{P}^s}$ converges to $\bar{\mathbf{x}}$. Similar analysis can be given for the case that $k_t \in \mathcal{P}^s$ implies

$$\lim_{t\to\infty} \gamma_{k_t} \mu_{k_t}^{-1} \|\mathbf{x}^{k_t+1} - \mathbf{x}^{k_t}\| = 0 \ \text{ and } \lim_{t\to\infty} \mathbf{x}^{k_t+1} = \bar{\mathbf{x}}. \tag{3.27}$$

Recalling $\mathbf{x}^{k_t+1} = \hat{\mathbf{x}}^{k_t+1}$ defined in (3.6) and by its first-order necessary optimality condition, we have

$$\begin{aligned} &\langle \nabla \tilde{f}(\mathbf{x}^{k_t}, \mu_{k_t}) + \gamma_{k_t} \mu_{k_t}^{-1}(\mathbf{x}^{k_t+1} - \mathbf{x}^{k_t}) + \lambda\zeta^{k_t}, \mathbf{x} - \mathbf{x}^{k_t+1}\rangle \geq 0, \\ &\forall \zeta^{k_t} \in \partial\Phi^{d^{k_t}}(W\mathbf{x}^{k_t+1}), \mathbf{x} \in \Omega. \end{aligned} \tag{3.28}$$

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

$$\left\{\lim_{j\to\infty} \zeta^{k_{t_j}} : \zeta^{k_{t_j}} \in \partial\Phi^{d^{k_{t_j}}}(W\mathbf{x}^{k_{t_j}+1})\right\} \subseteq \partial\Phi^{\bar{d}}(W\bar{\mathbf{x}}). \tag{3.29}$$

Along with the subsequence $\{k_{t_j}\}$ and letting $j \to \infty$ in (3.28), from Definition 3.1-(iii), (3.27) and (3.29), we obtain that there exist $\bar{\xi} \in \partial f(\bar{\mathbf{x}})$ and $\bar{\zeta}^{\bar{d}} \in \partial\Phi^{\bar{d}}(W\bar{\mathbf{x}})$ such that

$$\langle \bar{\xi} + \lambda\bar{\zeta}^{\bar{d}}, \mathbf{x} - \bar{\mathbf{x}}\rangle \geq 0, \forall \mathbf{x} \in \Omega. \tag{3.30}$$

By $\bar{d} \in \mathbb{D}^p(W\bar{\mathbf{x}})$, thanks to the convexity of $f + \lambda\Phi^{\bar{d}}$, (3.30) implies

$$f(\mathbf{x}) + \lambda\Phi^{\bar{d}}(W\mathbf{x}) - f(\bar{\mathbf{x}}) - \lambda\Phi^{\bar{d}}(W\bar{\mathbf{x}}) \geq \langle \bar{\xi} + \lambda\bar{\zeta}^{\bar{d}}, \mathbf{x} - \bar{\mathbf{x}}\rangle \geq 0, \forall \mathbf{x} \in \Omega,$$

which implies that $\bar{\mathbf{x}}$ is a lifted stationary point of (1.7). □

4. Numerical experiments

The purpose of this part is to test and verify the theoretical results and the properties of the SGD algorithm by the numerical experiments. We present Examples 4.1 and 4.2, which are respectively an under-determined linear regression problem and an over-determined censored regression problem. Especially, the process of solving subproblem (3.6) is very similar to the algorithm process of solving the LASSO problem.

All experiments are performed in MATLAB 2016a on a Lenovo PC with an Intel(R) Core(TM) i5-8250U CPU @1.60GHz 1801 Mhz and 8GB RAM. In the following examples, stopping criterion is set as

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

We stop the proposed algorithm if the number of iterations exceeds **Maxiter** or the smoothing parameter is less than ε. Denote $\bar{\mathbf{x}}$ the output of iterate $\mathbf{x}^k$. Set the fixed parameter $\alpha = 1$ throughout the numerical experiments.

Example 4.1. (Linear regression problem) Linear regression problems have been widely used in information theory [1], signal processing [35,36] and image restoration [6,36]. As pointed out in [20], ℓ_1 loss function is nonsmooth, but more robust and has stronger capability of outlier-resistance than the least squares loss function in the linear regression problems. Then we consider the following ℓ_0 regularized linear regression problem with ℓ_1 loss function:

$$\min_{\mathbf{x}\in\Omega} \quad \mathcal{W}_{\ell_0}(\mathbf{x}) := \frac{1}{m}\|A\mathbf{x} - b\|_1 + \lambda\|W\mathbf{x}\|_0, \tag{4.2}$$

where $A \in \mathbb{R}^{m\times n}$ with $m = n$, $b \in \mathbb{R}^m$. A smoothing function of the ℓ_1 loss function can be defined by

$$\tilde{f}(\mathbf{x},\mu) = \frac{1}{m}\sum_{i=1}^{m} \tilde{\theta}(A_i\mathbf{x} - b_i, \mu) \;\; with \;\; \tilde{\theta}(s,\mu) = \begin{cases} |s| & if\ |s| > \mu, \\ \frac{s^2}{2\mu} + \frac{\mu}{2} & if\ |s| \leq \mu. \end{cases} \tag{4.3}$$

Denote s the ℓ_0 norm of true solution $\mathbf{x}^*$, i.e., $\|W\mathbf{x}^*\|_0 = s$. For the given positive integers m, n and s, the data are generated by

$$\texttt{W=randn(p,n)}; \texttt{B=randn(n,m)}; \texttt{A=orth(B)'}; \texttt{b=A*x}^*\texttt{+0.01*randn(m,1)}.$$

In the algorithm, we set the parameters as below: $\underline{\gamma} = \overline{\gamma} = 1$, $\mu_0 = 3.533$, $\texttt{Maxiter} = 10^3$, $\nu = 35.6014$, $\sigma = 3.0003$, $\rho = 1.0001$, $\kappa = \frac{1}{2}$. Generate A, b and $\mathbf{x}^*$ with $m = n = 45$, $p = 45$ and $s = 2$, set $\lambda = 10^{-3}$ in (4.2) and $\varepsilon = 10^{-3}$ in the stopping criterion (4.1). We set $\mathbf{x}_0$=$ones(n, 1)$. Figure 1 shows the numerical results. Figure 1 plots $\mathbf{x}^*$ and $\overline{\mathbf{x}}$, where $\mathbf{x}^*$ and $\overline{\mathbf{x}}$ denote the original signal (which can also be expressed as true solution) and the output of iterate $\mathbf{x}^k$ from the SGD algorithm. From Figure 1, we can see that the output of $\mathbf{x}^k$ is very close to the original generated signal.

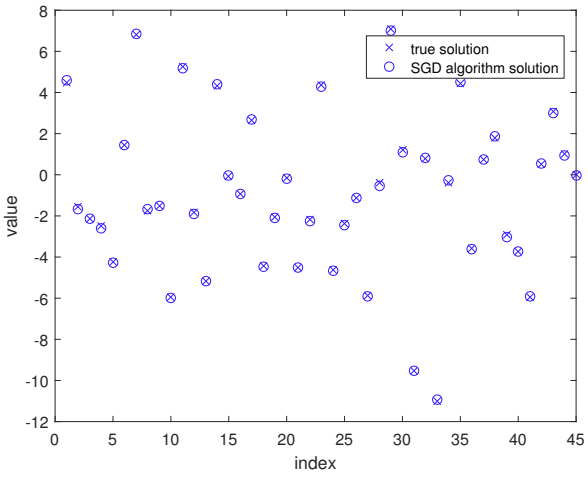

Figure 1. Digital experiment of the SGD algorithm in Example 4.1 under the first form of W.

Now we use another form of matrix W to solve Example 4.1:

$$W = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ -3 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -3 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -3 & 1 \end{pmatrix}_{p\times n}.$$

Set $\underline{\gamma} = \overline{\gamma} = 1$, $\mu_0 = 3.533$, $\texttt{Maxiter} = 10^3$, $\nu = 36$, $\sigma = 7$, $\rho = 1.0001$ and $\kappa = \frac{1}{2}$. We randomly generate the data as follows:

$$\texttt{B=randn(n,m)}; \texttt{A=orth(B)'}; \texttt{b=A*x}^*\texttt{+0.01*randn(m,1)}.$$

We run numerical experiments with (m, n, p, s)=(45,45,45,2). Set $\lambda = 10^{-3}$ in (4.2) and $\varepsilon = 10^{-3}$ in the stopping criterion (4.1). We define $\mathbf{x}_0$=$randn(n, 1)$. From Figure 2, we can see that the output of $\mathbf{x}^k$ obtained by the SGD algorithm is also close to the true solution $\mathbf{x}^*$.

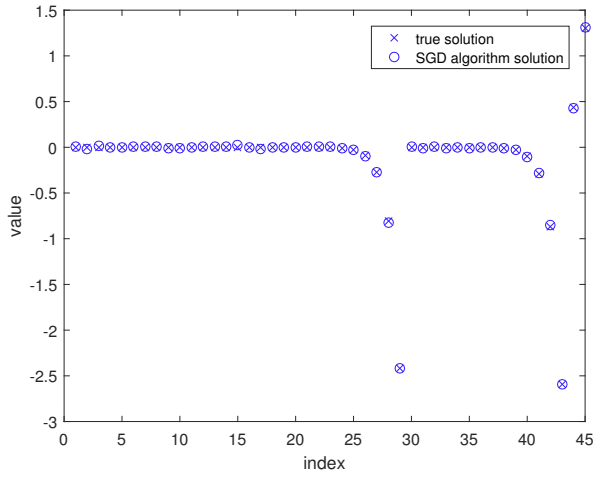

Figure 2. Digital experiment of SGD algorithm in Example 4.1 under the second form of W.

The last special case of W is the penalty matrix in 1-dimensional Fused LASSO [39]:

$$W = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}_{p\times n}.$$

Set $v = 40,\ p = 45,\ n = 46$ and $\mathbf{x}_0$=$ones(n, 1)$. The remaining parameter values are the same as the previous situation (see Figure 3).

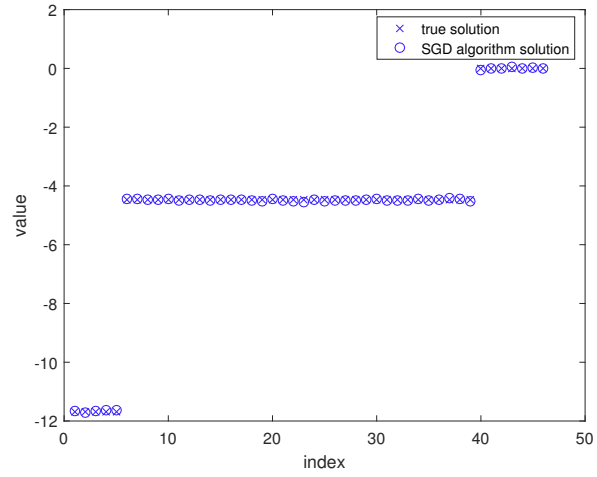

Figure 3. Digital experiment of SGD algorithm in Example 4.1 under the third form of W.

From Figures 1–3, we can see that the output of $\mathbf{x}^k$ obtained by the SGD algorithm is close to the true solution $\mathbf{x}^*$.

Example 4.2. (Censored regression problem) The application of censored regression problems has been studied in machine learning [37], economics [38], biomedical, and technical systems. The following censored regression problem is also a typical class of composite sparse optimization problems with nonsmooth convex loss functions. Now we consider the following ℓ_0 regularized censored regression problem:

$$\min_{\mathbf{x}\in\Omega} \quad \mathcal{W}_{\ell_0}(\mathbf{x}) := \frac{1}{m}\|\max\{A\mathbf{x},0\} - b\|_1 + \lambda\|W\mathbf{x}\|_0,$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. For the loss function in (1.5), a smoothing function of it can be defined by

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

Set $\varepsilon = 10^{-2}$, $\nu = 16.0009$, $\lambda = 10^{-3}$, $\mu_0 = 10.8999$, $\sigma = 4.0003$, $\kappa = \frac{1}{2}$, $\rho = 1.2006$ and $\mathbf{x}_0$=*randn*$(n,1)$. In this example, we run numerical experiments with (m,n,p,s)=(40,40,40,2), we randomly generate the problem data as follows:

A=randn(m,n); **W= randn(p,n)**; **b = max(A*x*+0.01*randn(m,1),0)**.

The computational results of $\mathbf{x}^*$ and $\mathbf{x}$ are shown in Figure 4.

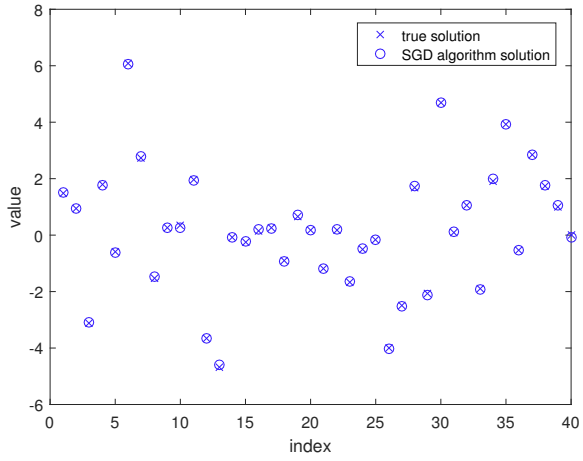

Figure 4. Numerical results of the SGD algorithm for Example 4.2.

We use the following form of W to solve Example 4.2:

$$W = \begin{pmatrix} 0 & -2 & 0 & 0 & \cdots & 0 & 0 \\ -2 & 1 & -3 & 0 & \cdots & 0 & 0 \\ 0 & -3 & 1 & -4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -m \\ 0 & 0 & 0 & 0 & \cdots & -m & 1 \end{pmatrix}_{p\times n}.$$

Set $\underline{\gamma} = \overline{\gamma} = 3$, $\mu_0 = 2$, **Maxiter** $= 10^3$, $\nu = 1.2200$, $\sigma = 2.6915$, $\rho = 1.0001$ and $\kappa = 0.711$. For the given positive integers m, n and s, the data are generated by

A=randn(m,n); **b=A*x*****+0.01*randn(m,1)**.

We run numerical experiments with (m, n, p, s)=(40,40,40,2). Set $\lambda = 10^{-3}$ in (4.2), $\mathbf{x}_0$=*ones*$(n, 1)$ and $\varepsilon = 10^{-3}$. From Figures 4 and 5, it can be seen that the output of $\mathbf{x}^k$ is very close to the true solution.

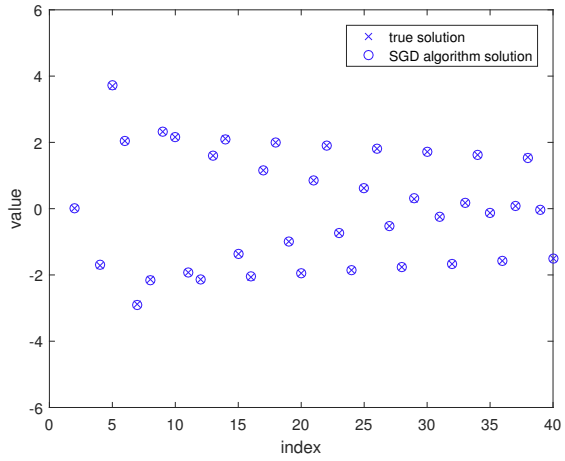

Figure 5. Numerical results of the SGD algorithm for Example 4.2.

5. Conclusions

We have intensively studied the composite sparse optimization problem consisting of the sum of a nonsmooth convex function and the ℓ_0 penalty term of a matrix times the coefficient vector. Considering the original cardinality penalty problem and an exact continuous relaxation problem with capped-ℓ_1 penalty, we have proved several novel and interesting results: the consistency between global minimizers of the relaxation problem and the original problem, and local minimizers of relaxation problems are local minimizers of the original problem. We propose the SGD algorithm based on the smoothing method and the smoothing gradient descent algorithm. Then SGD algorithm has been investigated from both a theoretical and an algorithmic point of view. So we prove that if the sequence generated by the algorithm has an accumulation point, then it is a lifted stationary point of relaxation problem. This well explains why the algorithm is expected to enjoy an appealing performance from the theoretical perspective, which is testified by the numerical experiments. Our initial numerical results confirm the predicted underlying theoretical results.

Author contributions

Wei Yang: Conceptualization, writing-original draft, validation, software; Lili Pan: Conceptualization, supervision, funding acquisition, validation, software; Jinhui Wan: Software, methodology, data curation. All authors have read and approved the final version of the manuscript for publication.

Acknowledgments

The work of the authors was supported by the National Natural Science Foundation of China grants 12271309.

Conflict of interest

The authors declare no conflicts of interest.

References

1. E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE T. Inform. Theory*, **52** (2006), 489–509.
2. D. L. Donoho, Compressed sensing, *IEEE T. Inform. Theory*, **52** (2006), 1289–1306.
3. F. Facchinei, Minimization of SC1 functions and the Maratos effect, *Oper. Res. Lett.*, **17** (1995), 131–137. https://doi.org/10.1016/0167-6377(94)00059-F
4. M. Elad, *Sparse and redundant representations: From theory to applications in signal and image processing*, Springer Science & Business Media, 2010.
5. M. Elad, M. A. Figueiredo, Y. Ma, On the role of sparse and redundant representations in image processing, *P. IEEE*, **98** (2010), 972–982. https://doi.org/10.1109/JPROC.2009.2037655
6. W. Bian, X. Chen, Linearly constrained non-Lipschitz optimization for image restoration, *SIAM J. Imaging Sci.*, **8** (2015), 2294–2322. https://doi.org/10.1137/140985639
7. X. Chen, M. K. Ng, C. Zhang, Non-Lipschitz ℓ_p-regularization and box constrained model for image restoration, *IEEE T. Image Process.*, **21** (2012), 4709–4721. https://doi.org/10.1109/TIP.2012.2214051
8. J. Fan, L. Xue, H. Zou, Strong oracle optimality of folded concave penalized estimation, *Ann. Stat.*, **42** (2014), 819. https://doi.org/10.1214/13-AOS1198
9. W. Bian, X. Chen, A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty, *SIAM J. Numer. Anal.*, **58** (2020), 858–883. https://doi.org/10.1137/18M1186009
10. X. Li, Z. Yang, X. Chen, Quantitative damage detection and sparse sensor array optimization of carbon fiber reinforced resin composite laminates for wind turbine blade structural health monitoring, *Sensors*, **14** (2014), 7312–7331. https://doi.org/10.3390/s140407312
11. W. Huang, Q. Fu, H. Dou, Z. Dong, *Resonance-based sparse signal decomposition based on genetic optimization and its application to composite fault diagnosis of rolling bearings*, In: ASME International Mechanical Engineering Congress and Exposition (Vol. 57403, p. V04BT04A054). American Society of Mechanical Engineers, 2015. https://doi.org/10.1115/IMECE2015-50874
12. L. L. Beyer, N. Balabanska, E. Tal, S. Karaman, *Multi-modal motion planning using composite pose graph optimization*, In: 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, 9981–9987.

13. R. Zhou, Y. Wang, B. Qiao, W. Zhu, J. Liu, X. Chen, Impact force identification on composite panels using fully overlapping group sparsity based on Lp-norm regularization, *Struct. Health Monit.*, **23** (2024), 137–161.

14. J. Liu, L. Yuan, J. Ye, *Guaranteed sparse recovery under linear transformation*, In: International Conference on Machine Learning, 2013, 91–99.

15. R. J. Tibshirani, *The solution path of the generalized lasso*, Stanford University, 2011.

16. B. Xin, Y. Kawahara, Y. Wang, W. Gao, *Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease*, In: Proceedings of the AAAI Conference on Artificial Intelligence, 2014. https://doi.org/10.1145/2847421

17. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, *J. R. Stat. Soc. B*, **67** (2005), 91–108 https://doi.org/10.1111/j.1467-9868.2005.00490.x

18. L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, *Physica D*, **60** (1992), 259–268.

19. S. J. Kim, K. Koh, S. Boyd, D. Gorinevsky, ℓ_1 trend filtering, *SIAM Rev.*, **51** (2009), 339–360. https://doi.org/10.1137/070690274

20. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, **96** (2001), 1348–1360. https://doi.org/10.1198/016214501753382273

21. Z. Zheng, Y. Fan, J. Lv, High dimensional thresholded regression and shrinkage effect, *J. R. Stat. Soc. B*, **76** (2014), 627–649. https://doi.org/10.1111/rssb.12037

22. D. Peleg, R. Meir, A bilinear formulation for vector sparsity optimization, *Signal Process.*, **88** (2008), 375–389.

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

24. T. Zhang, Multi-stage convex relaxation for feature selection, *Bernoulli*, **19** (2013), 2277–2293. https://doi.org/10.3150/12-BEJ452

25. W. Jiang, F. Nie, H. Huang, *Robust dictionary learning with capped ℓ_1-norm*, In: Twenty-fourth international joint conference on artificial intelligence, 2015.

26. L. Pan, X. Chen, Group sparse optimization for images recovery using capped folded concave functions, *SIAM J. Imaging Sci.*, **14** (2021), 1–25.

27. M. Nikolova, Local strong homogeneity of a regularized estimator, *SIAM J. Appl. Math.*, **61** (2000), 633–658.

28. M. Chen, Q. Wang, S. Chen, X. Li, Capped ℓ_1-norm sparse representation method for graph clustering, *IEEE Access*, **7** (2019), 54464–54471.

29. Z. Xue, L. Cai, Robust fisher-regularized twin extreme learning machine with capped ℓ_1-norm for classification, *Axioms*, **12** (2023), 717.

30. E. Soubies, L. Blanc-Féraud, G. Aubert, A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization, *SIAM J. Optimiz.*, **27** (2017), 2034–2060. https://doi.org/10.1137/16m1059333

31. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. Appl.*, **2** (1976), 17–40. https://doi.org/10.1016/0898-1221(76)90003-1
32. Z. J. Bai, M. K. Ng, L. Qi, A coordinate gradient descent method for nonsmooth nonseparable minimization, *Numer. Math.-Theory Me.*, **2** (2009), 377–402. https://doi.org/10.4208/nmtma.2009.m9002s
33. J. S. Pang, M. Razaviyayn, A. Alvarado, Computing B-stationary points of nonsmooth DC programs, *Math. Oper. Res.*, **42** (2017), 95–118. https://doi.org/10.1287/MOOR.2016.0795
34. H. Lütkepohl, *Handbook of matrices*, John Wiley & Sons, 1997.
35. A. M. Bruckstein, D. L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Rev.*, **51** (2009), 34–81.
36. M. Nikolova, M. K. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, *SIAM J. Sci. Comput.*, **27** (2005), 937–966.
37. C. Cortes, V. Vapnik, *Support-vector networks*, Machine learning, **20** (1995), 273–297. https://doi.org/10.1023/A:1022627411411
38. R. Blundell, J. L. Powell, Censored regression quantiles with endogenous regressors, *J. Econometrics*, **141** (2007), 65–83.
39. T. B. Arnold, R. J. Tibshirani, Efficient implementations of the generalized lasso dual path algorithm, *J. Comput. Graph. Stat.*, **25** (2016), 1–27.
40. Z. Shen, Q. Chen, F. Yang, A convex relaxation framework consisting of a primal-dual alternative algorithm for solving ℓ_0 sparsity-induced optimization problems with application to signal recovery based image restoration, *J. Comput. Appl. Math.*, **421** (2023), 114878.
41. Q. Chen, Z. Shen, A two-metric variable scaled forward-backward algorithm for ℓ_0 optimization problem and its applications, *Numer. Algorithms*, **97** (2024), 191–221.
42. T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, *J. Mach. Learn. Res.*, **11** (2010).
43. H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models, *Anna. Stat.*, **36** (2008), 1509.

AIMS Press