Signal Processing 168 (2020) 107369

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Sparse optimization problem with *s*-difference regularization

Yuli Sun, Xiang Tan, Xiao Li, Lin Lei, Gangyao Kuang*

College of Electronic science, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Article history:
Received 29 March 2019
Revised 29 September 2019
Accepted 7 November 2019
Available online 8 November 2019

Keywords:
Sparse optimization
Forward-Backward splitting
Proximal operator
Difference of convex
Truncated function

ABSTRACT

In this paper, a *s*-difference type regularization for sparse recovery problem is proposed, which is the difference of the penalty function $R(\mathbf{x})$ and its corresponding *s*-truncated function $R(\mathbf{x}^s)$. First, we show the equivalent conditions between the ℓ_0 constrained problem and the unconstrained *s*-difference penalty regularized problem. Next, we choose the forward-backward splitting (FBS) method to approximately solve the non-convex regularization function and further derive some closed-form solutions for the proximal mapping of the *s*-difference regularization with some commonly used $R(\mathbf{x})$, which makes the FBS easy and fast. We also show that any cluster point of the sequence generated by the proposed algorithm converges to a stationary point. Numerical experiments demonstrate the efficiency of the proposed *s*-difference regularization in comparison with some other existing penalty functions.

1. Introduction

1.1. Background

In recent years, sparse optimization problems have drawn lots of attentions in many applications such as compressive sensing (CS), machine learning, image processing and medical imaging. Signal and image processing problems are usually expressed as

$$A(\mathbf{x}) + \mathbf{n} = \mathbf{b} \tag{1}$$

where A is the linear or non-linear operator, $\mathbf{b}$ is the observation data, and $\mathbf{n}$ represents the observation noise or error. Since problem (1) is often ill-posed and the error $\mathbf{n}$ is unknown, solving (1) is difficult. To overcome this ill-posed problem, we need to make some constraints to narrow the solution space, such as the prior sparsity of the signals. Then the problem can be formulated as

$$\min_{\mathbf{x}} \phi(\mathbf{x}) + P(\mathbf{x}) \tag{2}$$

where the loss function $\phi(\mathbf{x})$ is the data fidelity term related to (1). For example, the least-square (LS) loss function $\|A(\mathbf{x}) - \mathbf{b}\|_2^2$ or the least-absolute (LA) loss function $\|A(\mathbf{x}) - \mathbf{b}\|_1$; $P(\mathbf{x})$ is the Regularization function to penalize the sparsity of $\mathbf{x}$. Intuitively, $P(\mathbf{x})$ should be selected as the ℓ_0-norm $\|\mathbf{x}\|_0$, which represents the number of nonzero elements in $\mathbf{x}$. However, minimizing the ℓ_0-norm is equivalent to finding the sparsest solution, which is known to be an NP-hard problem. A favorite and popular approach is using the ℓ_1-norm convex approximation, i.e., using $\|\mathbf{x}\|_1$ to replace the ℓ_0 [1]. This ℓ_1 model has been widely used in many different applications, such as radar systems [2,3], communications [4], computed tomography (CT) [5] and magnetic resonant imaging (MRI) [6]. It has been proved that the *s*-sparse signal $\mathbf{x}$ can be recovered by ℓ_1 model under some assumptions of the operator A, such as the restricted isometry property (RIP) of A when the operator is a sensing matrix [1]. However, the ℓ_1-norm regularization tends to underestimate high-amplitude components of $\mathbf{x}$ as it penalizes the amplitude uniformly, unlike the ℓ_0-norm in which all nonzero entries have equal contributions. This may lead to reconstruction failures with the least measurements [7,8], and bring undesirable blocky images [9,10]. It is quite well-known that when it promotes sparsity, the ℓ_1-norm does not provide a performance close to that of the ℓ_0-norm, and lots of theoretical and experimental results in CS and low-rank matrix recovery suggest that better approximations of the ℓ_0-norm and matrix rank give rise to better performances.

Recently, researchers begin to investigate various non-convex regularizations to replace the ℓ_1-norm regularization and gain some better reconstructions. In particular, the ℓ_p(quasi)-norm with $p \in (0,\ 1)$ [11–16], can be regarded as a interpolation between the ℓ_0 and ℓ_1, and a continuation strategy to approximate the ℓ_0 as $p \to 0$. The optimization strategies include half thresholding [14,17–20] and iterative reweighting [11,12,15]. Other non-convex regularizations and algorithms have also been designed to outperform ℓ_1-norm regularization and seek better reconstructions: capped ℓ_1-norm [21–23], transformed ℓ_1-norm [24–26], sorted ℓ_1-norm [27,28], the difference of the ℓ_1 and ℓ_2-norms (ℓ_{1-2}) [29–31],

* Corresponding author.
E-mail address: kuanggangyao@nudt.edu.cn (G. Kuang).

https://doi.org/10.1016/j.sigpro.2019.107369
0165-1684/

the log-sum penalty (LSP) [8], smoothly clipped absolute deviation (SCAD) [32,33], minimax-concave penalty (MCP) [34–36].

On the other hand, there are some approaches which do not approximate the ℓ_0-norm, such as the single best replacement (SBR) algorithm [73] and iterative hard thresholding (IHT) algorithm [37,38], which operate directly on the ℓ_0 regularized cost function or the s-sparse constrained optimization problem. Moreover, there are some acceleration methods for the IHT: accelerated IHT (AIHT) [39], proximal IHT (PIHT) [40], extrapolated proximal IHT (EPIHT) [41] and accelerated proximal IHT [42]. Meanwhile, some researchers transform the ℓ_0-norm problem into an equivalent difference of two convex functions, and then use the difference of convex algorithm (DCA) and the proximal gradient technique to solve the subproblem [43,44].

To address these non-convex regularization problems, many iterative algorithms are investigated by researchers, such as the DCA [45–48] (or Convex-ConCave Procedure (CCCP) [49], or the Multi-Stage (MS) convex relaxation [22]), and its accelerate versions: Boosted Difference of Convex function Algorithms (BDCA) [50] and proximal Difference-of-Convex Algorithm with extrapolation (pDCAe) [51], the alternating direction method of multipliers (ADMM) [52], split Bregman iteration (SBI) [53], General Iterative Shrinkage and Thresholding (GIST) [54], nonmonotone accelerated proximal gradient (nmAPG) [55], which is an extension of the accelerated proximal gradient (APG) [56].

1.2. Contributions

In many applications, the non-convex ℓ_0-norm based regularization has its advantages over the convex ℓ_1-norm, such as image restoration [41,53,57,58], bioluminescence [59], CT [9,10], MRI reconstruction [60,61]. Thus, in this paper, we are interested in the following ℓ_0 constrained problem

$$\min_{\mathbf{x}} \phi(\mathbf{x}) \text{ subject to } \|\mathbf{x}\|_0 \le s \tag{3}$$

where $s \in \{1, 2, \dots, N\}$. This s-sparse problem tries to find the solution minimizing $\phi(\mathbf{x})$ under the constraint that the number of non-zero coefficients below a certain value.

This paper can be viewed as a natural complement and extension of Gotoh et al. framework [43]. First, we rewrite the ℓ_0 constrained problem (3) as difference of two functions, one of which is the convex or non-convex function $R(\mathbf{x})$ and the other is the corresponding s-truncated function $R(\mathbf{x}^s)$, where $\mathbf{x}^s$ is the best s term approximation to $\mathbf{x}$. Then, we consider the unconstrained minimization problem by using this s-difference $R(\mathbf{x}) - R(\mathbf{x}^s)$ type regularization. Second, we propose fast approaches to deal with this non-convex regularization function, which are based on a proximal operator corresponding to $R(\mathbf{x}) - R(\mathbf{x}^s)$. Moreover, we derive some cheap closed-form solutions for the proximal mapping of $R(\mathbf{x}) - R(\mathbf{x}^s)$ with some commonly used $R(\mathbf{x})$, such as $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1 - a\|\mathbf{x}\|_2$, LSP, MCP and so on. Third, we prove the convergence performance of the proposed algorithm, and show that any cluster point of the sequence generated by the proposed algorithm converges to a stationary point. We also show a link between the proposed algorithm with some related regularizations and algorithms. Finally, we evaluate the effectiveness of the proposed algorithm via numerical experiments. The reconstruction results demonstrate that the proposed s-difference penalty function with closed-form solutions is more accurate than the ℓ_1-norm and other non-convex regularization based methods, and faster than the DCA based algorithms.

1.3. Outline and notation

The rest of this paper is structured as follows. In section II, we define the s-difference regularization. In section III, we propose the reconstruction algorithm by using the proximal operator with closed-form solutions. In section IV, we provide some theorems to demonstrate the convergence of the proposed algorithm. In section V, we discuss some related algorithms and extend the proposed regularization to rank-constrained problem. Section VI presents the numerical results. In the end, we provide our conclusion in section VII.

Here, we define our notation. For a vector $\mathbf{x} \in \mathbb{R}^N$, it can be written as $\mathbf{x} = (x_1, x_2, \dots, x_N)$, and its ℓ_p-norm is defined as $\|\mathbf{x}\|_p = (\sum_n |x_n|^p)^{1/p}$. Especially, the ℓ_∞-norm of $\mathbf{x}$ is defined as $\max_n |x_n|$. Given a matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$, the transpose of $\mathbf{A}$ is denoted by $\mathbf{A}^T$, the maximum eigenvalue of $\mathbf{A}^T\mathbf{A}$ is defined as $\|\mathbf{A}\|_2^2$. Some of the arguments in this paper use sub-vectors. The letters Γ, Λ denote sets of indices that enumerate the elements in the vector $\mathbf{x}$. By using this sets as subscripts, $\mathbf{x}_\Gamma$ represents the vector that setting all elements of $\mathbf{x}$ to zero except those in the set Γ. The iteration count is given in square bracket, e.g., $\mathbf{x}^{[k]}$. $\langle \cdot, \cdot \rangle$ denotes the inner product, $\text{sign}(\cdot)$ represents the sign of a quantity with $\text{sign}(0) \in [-1, 1]$. We also use the notation $\mathbb{R}_+ = \{x \in \mathbb{R} : x \ge 0\}$, and if the function f is defined as the composition $f(x) = h(g(x))$, we write $f = h \circ g$.

Given a proper closed function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, the subgradient of h at x is given by

$$\partial h(x) = \{v \in \mathbb{R}^n : h(u) - h(x) - \langle v, u - x\rangle \ge 0, \forall u \in \mathbb{R}^n\} \tag{4}$$

In addition, if $h(x)$ is continuously differentiable, then the subdifferential reduces to the gradient of $h(x)$ denoted by $\nabla h(x)$.

2. Penalty representation for s-sparse problem

In Gotoh et al. work of [43], they expressed the ℓ_0-norm constraint as a difference of convex (DC) function:

$$\|\mathbf{x}\|_0 \le s \Leftrightarrow \|\mathbf{x}\|_1 - \||\mathbf{x}|\|_s = 0 \tag{5}$$

where $s \in \{1, 2, \dots, N\}$ and $\||\mathbf{x}|\|_s$, which named top-$(s, 1)$ norm, denotes the sum of top-s elements in absolute value. This notation is also known as the largest-s norm (or called CVaR norm in [62,63]). Precisely,

$$\||\mathbf{x}|\|_s := \left|x_{\pi_x(1)}\right| + \left|x_{\pi_x(2)}\right| + \cdots + \left|x_{\pi_x(s)}\right| \tag{6}$$

where $x_{\pi_x(i)}$ denotes the element whose absolute value is the i-th largest among the N elements of vector $\mathbf{x} \in \mathbb{R}^N$, i.e., $|x_{\pi_x(1)}| \ge |x_{\pi_x(2)}| \ge \cdots \ge |x_{\pi_x(N)}|$. For convenience of description, we define the set $\Gamma_{\mathbf{x}}^s = \{\pi_x(1), \pi_x(2), \dots, \pi_x(s)\}$, then we have $\Gamma_{\mathbf{x}}^1 \subseteq \Gamma_{\mathbf{x}}^2 \subseteq \cdots \subseteq \Gamma_{\mathbf{x}}^N$. By using $\cdot\backslash\cdot$ as the set difference, we have $\Gamma_{\mathbf{x}}^N \backslash \Gamma_{\mathbf{x}}^s = \{\pi_x(s+1), \pi_x(s+2), \dots, \pi_x(N)\}$.

We define $\mathbf{x}^s$ as the best s term approximation to $\mathbf{x}$, that is, any s-sparse vectors that minimize $\|\mathbf{x} - \mathbf{x}^s\|_2$. By using the definition of $x_{\pi_x(i)}$, we have

$$x_i^s = \begin{cases} x_i, & \text{if} \quad i \in \Gamma_{\mathbf{x}}^s \\ 0, & \text{if} \quad i \in \Gamma_{\mathbf{x}}^N \backslash \Gamma_{\mathbf{x}}^s \end{cases} \tag{7}$$

Then $R(\mathbf{x}^s)$ can be named as a s-truncated function. In this work, we consider a more general s-difference function $R(\mathbf{x}) - R(\mathbf{x}^s)$ instead of $\|\mathbf{x}\|_1$ to replace the ℓ_0-norm constraint, where $R(\mathbf{x})$ can be convex or non-convex, separable or non-separable. Let $P(\mathbf{x}) = R(\mathbf{x}) - R(\mathbf{x}^s)$, $s \in \{1, 2, \dots, N\}$. We defined a class of penalty functions $P, R : \mathbb{R}^N \to \mathbb{R}_+$ as follows (without loss of generality, functions $P(\mathbf{x})$ and $R(\mathbf{x})$ mentioned throughout this paper all satisfy Property 1 when there is no additional illustration).

Properties 1. *The penalty functions $P, R : \mathbb{R}^N \to \mathbb{R}_+$ satisfy the following properties.*

(a) $R(\mathbf{x}) = R(-\mathbf{x})$
(b) $\|\mathbf{x}\|_0 \le s \Leftrightarrow P(\mathbf{x}) = 0$

Table 1
Functions that satisfies Property 1.

Function type	$R(\mathbf{x})$	$P_1(\mathbf{x})$	$P_2(\mathbf{x})$
Convex, Separable	$\|\mathbf{x}\|_1$	$\|\mathbf{x}\|_1$	$\|\mathbf{x}^s\|_1$
	$\|\mathbf{x}\|_2^2$	$\|\mathbf{x}\|_2^2$	$\|\mathbf{x}^s\|_2^2$
Convex, Non-separable	$\|\mathbf{x}\|_2$	$\|\mathbf{x}\|_2$	$\|\mathbf{x}^s\|_2$
	$R(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|_2^2/(2\theta), & \|\mathbf{x}\|_2 \le \theta \\ \|\mathbf{x}\|_2 - \theta/2, & \|\mathbf{x}\|_2 > \theta \end{cases}, \theta > 0$	$R(\mathbf{x})$	$R(\mathbf{x}^s)$
Non-convex, Separable	$R(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$		
	$r_i(x_i) = \log(1 + \|x_i\|/\theta), \theta > 0$	$\|\mathbf{x}\|_1/\theta + (\|\mathbf{x}^s\|_1/\theta - R(\mathbf{x}^s))$	$\|\mathbf{x}^s\|_1/\theta + (\|\mathbf{x}\|_1/\theta - R(\mathbf{x}))$
	$R(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$		
	$r_i(x_i) = \begin{cases} \|x_i\| - x_i^2/(2\theta), & \|x_i\| \le \theta \\ \theta/2, & \|x_i\| > \theta \end{cases}, \theta > 0$	$\|\mathbf{x}\|_1 + (\|\mathbf{x}^s\|_1 - R(\mathbf{x}^s))$	$\|\mathbf{x}^s\|_1 + (\|\mathbf{x}\|_1 - R(\mathbf{x}))$
Non-convex, Non-separable	$\|\mathbf{x}\|_1 - a\|\mathbf{x}\|_2, 0 < a \le 1$	$\|\mathbf{x}\|_1 + a\|\mathbf{x}^s\|_2$	$\|\mathbf{x}^s\|_1 + a\|\mathbf{x}\|_2$
	$\log(1 + \|\mathbf{x}\|_2/\theta), \theta > 0$	$\|\mathbf{x}\|_2/\theta + (\|\mathbf{x}^s\|_2/\theta - R(\mathbf{x}^s))$	$\|\mathbf{x}^s\|_2/\theta + (\|\mathbf{x}\|_2/\theta - R(\mathbf{x}))$
	$\begin{cases} \|\mathbf{x}\|_2 - \|\mathbf{x}\|_2^2/(2\theta), & \|\mathbf{x}\|_2 \le \theta \\ \theta/2, & \|\mathbf{x}\|_2 > \theta \end{cases}, \theta > 0$	$\|\mathbf{x}\|_2 + (\|\mathbf{x}^s\|_2 - R(\mathbf{x}^s))$	$\|\mathbf{x}^s\|_2 + (\|\mathbf{x}\|_2 - R(\mathbf{x}))$

(c) *$P(\mathbf{x})$ is a continuous function which can be decomposed into the difference of two convex (DC) functions, that is, $P(\mathbf{x}) = P_1(\mathbf{x}) - P_2(\mathbf{x})$, where $P_1(\mathbf{x})$ and $P_2(\mathbf{x})$ are convex functions.*

It should be noted that although $P(\mathbf{x})$ is defined as $P(\mathbf{x}) = R(\mathbf{x}) - R(\mathbf{x}^s)$ and it can also be decomposed into DC functions as $P(\mathbf{x}) = P_1(\mathbf{x}) - P_2(\mathbf{x})$, but $P_1(\mathbf{x})$ and $P_2(\mathbf{x})$ are not always equivalent to $R(\mathbf{x})$ and $R(\mathbf{x}^s)$, respectively. They are two different expression forms of the same $P(\mathbf{x})$. In fact, for the convex $R(\mathbf{x})$, we can set $P_1(\mathbf{x}) = R(\mathbf{x})$ and $P_2(\mathbf{x}) = R(\mathbf{x}^s)$ to satisfy the DC decomposition of $P(\mathbf{x})$, which is a special case in the infinite many DC decompositions of $P(\mathbf{x})$; however, for the non-convex $R(\mathbf{x})$, the $P_1(\mathbf{x})$ can never be equal to $R(\mathbf{x})$. This can also be shown in the proof of Proposition 1 and Table 1.

Proposition 1. *The penalty functions listed on Table 1 all satisfy Property 1.*

See appendix A for the Proof of Proposition 1.

Remark 1. For the separable $R(\mathbf{x}) = \sum_{i=1}^N r(x_i)$, and $r(x)$ is continuous, symmetrical and strictly increasing on $\mathbb{R}_+$, if $r(x)$ is convex, then $R(\mathbf{x})$ satisfies Property 1; if $r(x)$ is non-convex, since it can be written as the difference of two convex functions as $r(x) = h(x) - g(x)$, then $R(\mathbf{x})$ also satisfies Property 1.

It is easy to see that the penalty function in Ref [43]. is a special case of $R(\mathbf{x})$=$\|\mathbf{x}\|_1$.

With the Property 1(b), we consider the following unconstrained minimization problem associated with (3):

$$\min_{\mathbf{x}\in\mathbb{R}^N} \{F(\mathbf{x}) = \phi(\mathbf{x}) + \rho P(\mathbf{x})\} \tag{8}$$

where $\rho > 0$ is the penalty parameter. We make the following assumptions on the above formulation throughout this paper, which are usually used in image processing and many CS fields.

Assumption 1. $\phi(\mathbf{x})$ is continuously differentiable with Lipschitz continuous gradient, i.e., there exists $L > 0$ such that

$$\|\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{y})\|_2 \le L\|\mathbf{x} - \mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N \tag{9}$$

Assumption 2. $F(\mathbf{x})$ is bounded from below.

From (8), we can find that the difference between penalty $P(\mathbf{x})$ and other penalty functions, such as ℓ_1, ℓ_p, ℓ_{1-2} and MCP, is that there is no punishment in model (8) when the sparsity level of $\mathbf{x}$ is under s. This is because that $P(\mathbf{x})$ is equal to zero as $\|\mathbf{x}\|_0 \le s$. Meanwhile, the selection of the regularization parameter ρ has an important influence on the performance of the reconstruction. On the one hand, ρ should be big enough to give a heavy cost for constraint violation: $\|\mathbf{x}\|_0 > s$. On the other hand, if ρ is too big, the reconstruction is mostly over regularized. In light of this, how to choose an appropriate parameter is very difficult and some researchers suggest using adaptive methods to select this parameter during the iterations [74]. The next Theorem ensures that problem (8) is equivalent to the original s-sparse constraint problem (3) as we take the limit of ρ, which can be proved in a similar manner to Theorem 17.1 in [71].

Theorem 1. *Let $\{\rho_t\}$ be an increasing sequence with $\lim_{t\to\infty}\rho_t = \infty$ and suppose that $\mathbf{x}_t$ is an optimal solution of (8) with $\rho = \rho_t$. Then, any limit point $\bar{\mathbf{x}}$ of $\{\mathbf{x}_t\}$ is also optimal to (3).*

See Appendix B for the proof.

In addition to Theorem 1, we have some stricter conclusions for the parameter ρ under some assumptions of $P(\mathbf{x})$ and $\phi(\mathbf{x})$.

Proposition 2. *If $\phi(\mathbf{x})$ is Lipschitz continuous with constant $\beta > 0$, i.e., $\|\phi(\mathbf{x}) - \phi(\mathbf{y})\|_2 \le \beta\|\mathbf{x} - \mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, and $\bar{\mathbf{x}}_\rho$ is an optimal solution of (8) with some ρ. Suppose that there exists a constant $\eta > 0$ such that $R(\mathbf{x}) - R(\mathbf{x} + \mathbf{x}^s - \mathbf{x}^{s+1}) \ge \eta\|\mathbf{x}^{s+1} - \mathbf{x}^s\|_2$ for any $\mathbf{x} \in \mathbb{R}^N$. Then if $\rho > \beta/\eta$, $\bar{\mathbf{x}}_\rho$ is also optimal to (3).*

See Appendix C for the proof.

Remark 2. Suppose that $\phi(\mathbf{x})$ is β-Lipschitz continuous and the regularization is $P(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$. Then if $\rho > \beta$, any optimal solution of (8) is also optimal to (3).

Remark 3. Suppose that $\phi(\mathbf{x})$ is β-Lipschitz continuous. If we choose $R(\mathbf{x})$ as $R(\mathbf{x}) = \|\mathbf{x}\|_1 - a\|\mathbf{x}\|_2, 0 < a \le 1$, then any optimal solution of (8) is also optimal to (3) when $\rho > \frac{\beta}{1 - a/(2\sqrt{s})}$. This can be proved by using that

$$\begin{aligned} \|\mathbf{x}\|_2 - \left\|\mathbf{x} + \mathbf{x}^s - \mathbf{x}^{s+1}\right\|_2 &= \frac{\left\|\mathbf{x}^{s+1} - \mathbf{x}^s\right\|_2^2}{\|\mathbf{x}\|_2 + \|\mathbf{x} + \mathbf{x}^s - \mathbf{x}^{s+1}\|_2} \\ &\le \frac{\left\|\mathbf{x}^{s+1} - \mathbf{x}^s\right\|_2}{2\sqrt{s}} \end{aligned} \tag{10}$$

If we choose $R(\mathbf{x}) = \theta_1\|\mathbf{x}\|_1 - \sum_{i=1}^N \log(1 + |x_i|/\theta_2), \theta_1 > \theta_2 > 0$, then the equivalent condition of ρ is $\rho > \frac{\beta}{\theta_1 - \theta_2}$. Meanwhile, we can obtain similar conclusions for the $R(\mathbf{x})$ which are the difference of $\|\mathbf{x}\|_1$ and MCP (Eq. (A.3)), or SCAD (Eq. (A.4)) function.

The next proposition, which is similar to Theorem 3 in [43] but with wider scope and stricter conclusion, shows another exact penalty parameters ρ requirement for $\phi(\mathbf{x})$ with Lipschitz continuous gradient L.

Proposition 3. *If Assumption 1 is satisfied and $\bar{\mathbf{x}}_\rho$ is an optimal solution of (8) with some ρ. Suppose that there exists a constant $C>0$ such that $\|\bar{\mathbf{x}}_\rho\|_2 \le C$ for any $\rho>0$, and there exists a constant $\eta>0$ such that $R(\mathbf{x})-R(\mathbf{x}+\mathbf{x}^s-\mathbf{x}^{s+1}) \ge \eta\|\mathbf{x}^{s+1}-\mathbf{x}^s\|_2$ for any $\mathbf{x}\in\mathbb{R}^N$, Then if $\rho > \frac{1}{\eta}(\|\nabla\phi(\mathbf{0})\|_2 + (1+\frac{1}{2\sqrt{s+1}})LC)$, $\bar{\mathbf{x}}_\rho$ is also optimal to (3).*

See Appendix D for the proof.

Remark 4. Suppose that $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ and $\|\bar{\mathbf{x}}_\rho\|_2 \le C$. If we choose $R(\mathbf{x})$ as $R(\mathbf{x}) = \|\mathbf{x}\|_1$, $R(\mathbf{x}) = \|\mathbf{x}\|_1 - a\|\mathbf{x}\|_2$ $(0<a\le 1)$ and $R(\mathbf{x}) = \theta_1\|\mathbf{x}\|_1 - \sum_{i=1}^N \log(1+|x_i|/\theta_2)$ $(\theta_1>\theta_2>0)$, then any optimal solution of (8) is also optimal to (3) when $\rho > \|\mathbf{A}^T\mathbf{b}\|_2 + (1+\frac{1}{2\sqrt{s+1}})\|\mathbf{A}\|_2^2 C$, $\rho > \frac{1}{1-a/(2\sqrt{s})}(\|\mathbf{A}^T\mathbf{b}\|_2 + (1+\frac{1}{2\sqrt{s+1}})\|\mathbf{A}\|_2^2 C)$ and $\rho > \frac{1}{\theta_1-\theta_2}(\|\mathbf{A}^T\mathbf{b}\|_2 + (1+\frac{1}{2\sqrt{s+1}})\|\mathbf{A}\|_2^2 C)$, respectively.

Remark 5. Similarly to Theorem 3 in [43] by replacing penalty function $\|\mathbf{x}\|_1 - \||\mathbf{x}|\|_s$ with ordinary function $R(\mathbf{x})-R(\mathbf{x}^s)$, we have the following conclusions without proof. If the conditions in Proposition 3 are satisfied, and supposing that $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$, where $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{N\times N}$ is symmetric and $\mathbf{q} = (q_i) \in \mathbb{R}^N$, then $\bar{\mathbf{x}}_\rho$ is also optimal to (3) when $\rho > \max_i \frac{1}{\eta}\{|q_i| + (\|\mathbf{Q}\mathbf{e}_i\|_2 + \frac{|q_{ii}|}{2\sqrt{s+1}})C\}$, where $\mathbf{e}_i$ denotes the unit vector in the i-th coordinate direction.

3. Forward-backward splitting for the regularization of difference of two functions

In this section, we use the FBS to approximately solve the unconstrained minimization (8). Moreover, we derive closed-form solutions for the proximal mapping of some special s-difference regularizations $P(\mathbf{x})$, and this makes FBS more efficient.

3.1. Forward-backward splitting and proximal operator

Each iteration of forward-backward splitting applies the gradient descent of $\phi(\mathbf{x})$ followed by a proximal operator. That is

$$\mathbf{x}^{[k+1]} = \text{prox}_{\beta\rho P}\left(\mathbf{x}^{[k]} - \beta\nabla\phi\left(\mathbf{x}^{[k]}\right)\right) \tag{11}$$

where $\beta>0$ is the step size, and sometimes this type of FBS is called the proximal gradient (PG) algorithm. The proximal operator is defined as

$$\text{prox}_{\lambda P}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\lambda} + P(\mathbf{x}) \tag{12}$$

with parameter $\lambda>0$.

The Eq. (11) can be broken up into a forward gradient step using the function $\phi(\mathbf{x})$, and a backward step using the function $\rho P(\mathbf{x})$. The proximal operator plays a central role in the analysis and solution of optimization problems. For example, the soft shrinkage operator, which is a proximal operator for ℓ_1-norm regularization, has been widely used in CS and rendered many efficient ℓ_1 algorithms. The proximal operator also has been successfully used with some non-convex regularizations, such as ℓ_p, SCAD, LSP [64], and MCP [52,65]. Normally, the closed-form solution of the proximal operator needs some special properties on $P(\mathbf{x})$, such as convexity or separability (e.g., the ℓ_1-norm, LSP, MCP, and other various separable functions in [66]). Next, we will focus on the solution of (12) with separable and non-separable s-difference $P(\mathbf{x})$.

3.2. Closed-form solution of the proximal operator

Denote $E(\mathbf{x})$ as

$$E(\mathbf{x}) = \frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\lambda} + P(\mathbf{x}) \tag{13}$$

and let $\mathbf{x}^*$ be the optimal solution of (12), i.e., $\mathbf{x}^* = \text{prox}_{\lambda P}(\mathbf{y})$, then we have the following Proposition.

Proposition 4. *Suppose that functions $P(\mathbf{x})$ and $R(\mathbf{x})$ satisfy Property 1, and let $\mathbf{x}^*$ be the optimal solution of (12). Then $\mathbf{x}^* = \mathbf{0}$ if and only if $\mathbf{y} = \mathbf{0}$.*

Proof. Necessary condition: note that $E(\mathbf{x}) \ge 0$ for any $\mathbf{x}$, and when $\mathbf{y}=\mathbf{0}$, we have $E(\mathbf{0})=0$. Thus if $\mathbf{y}=\mathbf{0}$, the optimal solution is $\mathbf{x}^* = \mathbf{0}$. Sufficient condition: assume by contradiction that $\mathbf{y}\neq\mathbf{0}$, then we select an arbitrary non-zero dimension y_j in $\mathbf{y}$, and construct $\tilde{\mathbf{x}} \in \mathbb{R}^N$ as $\tilde{x}_i = \begin{cases} 0, & i\neq j \\ y_j, & i=j \end{cases}$. Then we have

$$E(\mathbf{x}^*) = E(\mathbf{0}) = \frac{1}{2\lambda}\sum_{i=1}^N y_i^2 > \frac{1}{2\lambda}\sum_{i=1,i\neq j}^N y_i^2 = E(\tilde{\mathbf{x}}) \tag{14}$$

This contradicts the optimality of $\mathbf{x}^*$. Thus if $\mathbf{x}^* = \mathbf{0}$, $\mathbf{y}$ must be equal to zero. □

Proposition 5. *Suppose that functions $P(\mathbf{x})$ and $R(\mathbf{x})$ satisfy Property 1, and let $\mathbf{x}^*$ be the optimal solution of (12). Then, for $i\in\{1,2,\dots,N\}$, if $y_i>0$, then we have $x_i^* \ge 0$. If $y_i<0$, then we have $x_i^* \le 0$.*

Proof. We prove it by establishing contradiction. If there exits any $x_i^* < 0$ when $y_i>0$, then we select an arbitrary one and we construct $\tilde{\mathbf{x}} \in \mathbb{R}^N$ as $\tilde{x}_j = \begin{cases} x_j^*, & j\neq i \\ -x_j^*, & j=i \end{cases}$. We have

$$\begin{aligned}\|\tilde{\mathbf{x}}-\mathbf{y}\|_2^2 &= \sum_{j\neq i}\left(\tilde{x}_j - y_j\right)^2 + (\tilde{x}_i - y_i)^2 \\ &< \sum_{j\neq i}\left(x_j^* - y_j\right)^2 + \left(x_i^* - y_i\right)^2 = \|\mathbf{x}^*-\mathbf{y}\|_2^2\end{aligned} \tag{15}$$

The inequality follows from that x_i^* has the opposite sign as y_i and $y_i>0$. Since we have not changed the absolute value of $\tilde{x}_i$ and $R(\mathbf{x}) = R(-\mathbf{x})$, then we have $P(\tilde{\mathbf{x}}) = P(\mathbf{x}^*)$. combining this and (15), we have $E(\tilde{\mathbf{x}}) < E(\mathbf{x}^*)$. This contradicts the optimality of $\mathbf{x}^*$ and proves that $x_i^* \ge 0$ when $y_i>0$. On the other hand, we can prove that $x_i^* \le 0$ when $y_i<0$ by using a similar method. This completes the proof. □

Next, we focus on the closed-form solutions of $\text{prox}_{\lambda P}(\mathbf{y})$ with different types of $R(\mathbf{x})$.

Proposition 6. *Suppose that functions $P(\mathbf{x})$ and $R(\mathbf{x})$ satisfy Property 1, and $R(\mathbf{x})$ is separable, i.e., $R(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$ and each r_i is strictly increasing on $\mathbb{R}_+$. Let $\mathbf{x}^*$ be the optimal solution of (12). Then we have*

$$x_i^* = \begin{cases} y_i, & \text{if } i\in\Gamma_{\mathbf{y}}^s \\ \text{prox}_{\lambda r_{\pi_y(i)}}\left(y_{\pi_y(i)}\right), & \text{if } i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s \end{cases} \tag{16}$$

where $\Gamma_{\mathbf{y}}^s = \{\pi_y(1),\pi_y(2),\dots,\pi_y(s)\}$ and $\pi_y(j)$ is the index of the j-th largest amplitude of $\mathbf{y}$, i.e., $|y_{\pi_y(1)}| \ge |y_{\pi_y(2)}| \ge \cdots \ge |y_{\pi_y(N)}|$.

See Appendix E for the proof. From the proof we can find that if each r_i is convex, then $x_i^* = ((\mathbf{I}_N + \lambda\partial R)^{-1}(\mathbf{y}))_i$ when $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$, where $\mathbf{I}_N$ denotes the identity operator and the mapping $(\mathbf{I}_N+\lambda\partial R)^{-1}$ is called the resolvent of the operator ∂R with parameter λ.

Remark 6. Note that $x_i^* = y_i$ if $i\in\{\pi_y(1),\pi_y(2),\dots,\pi_y(s)\}$ in (16). Suppose that there exits one or more components of y_i, $i\notin\{\pi_y(1),\pi_y(2),\dots,\pi_y(s)\}$ having the same amplitude of $y_{\pi_y(s)}$, i.e., $|y_{\pi_y(s-m)}| = \cdots = |y_{\pi_y(s)}| = \cdots = |y_{\pi_y(s+j)}|$, $m\ge 0$, $j\ge 1$. Then there exit C_{j+m+1}^{m+1} solutions of $\mathbf{x}^*$ as there are C_{j+m+1}^{m+1} arrangements of $y_{\pi_y(s-m)},\dots,y_{\pi_y(s)}$.

Remark 7. If $R(\mathbf{x}) = \|\mathbf{x}\|_1$, i.e., $P(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$, then the solution $\mathbf{x}^*$ of (12) is

$$x_i^* = \begin{cases} y_i, & \text{if } i\in\Gamma_{\mathbf{y}}^s \\ \text{shrink}(y_i,\lambda), & \text{if } i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s \end{cases} \tag{17}$$

where shrink(y_i, λ) denotes the soft shrinkage operator given by

$$\text{shrink}(y_i,\lambda)=\text{sign}(y_i)\max\{|y_i|-\lambda,0\} \tag{18}$$

Remark 8. If $R(\mathbf{x})=\|\mathbf{x}\|_2^2$, i.e., $P(\mathbf{x})=\|\mathbf{x}\|_2^2-\|\mathbf{x}^s\|_2^2$, then the solution $\mathbf{x}^*$ of (12) is

$$x_i^*=\begin{cases} y_i, & \text{if } i\in\Gamma_{\mathbf{y}}^s \\ y_i/(2\lambda+1), & \text{if } i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s \end{cases} \tag{19}$$

Remark 9. If $R(\mathbf{x})$ is the MCP (A.3), that is $r_i(x_i)=\begin{cases}|x_i|-x_i^2/(2\theta), & |x_i|\le\theta\\ \theta/2, & |x_i|>\theta\end{cases}$ $(\theta>0)$, then the solution $\mathbf{x}^*$ of (12) is: under the condition of $\theta>\lambda$, if $i\in\Gamma_{\mathbf{y}}^s$ or $|y_i|>\theta$, then $x_i^*=y_i$; otherwise $x_i^*=\text{sign}(y_i)\max\{\theta(|y_i|-\lambda)/(\theta-\lambda),0\}$. When $\theta\le\lambda$, if $i\in\Gamma_{\mathbf{y}}^s$ or $|y_i|>\theta$, then $x_i^*=y_i$; otherwise $x_i^*=0$. If $R(\mathbf{x})$ is the LSP (A.2), that is $r_i(x_i)=\log(1+|x_i|/\theta),\theta>0$, then the solution $\mathbf{x}^*$ of (12) is: if $i\in\Gamma_{\mathbf{y}}^s$, then $x_i^*=y_i$; otherwise $x_i^*=\text{sign}(y_i)w_i$, and $w_i=\arg\min_{x_i\in\Omega}\{\frac{1}{2\lambda}(x_i-|y_i|)^2+\sum_i\log(1+|x_i|/\theta)\}$, where Ω is a set composed of 3 elements or 1 element. If $(|y_i|-\theta)^2-4(\lambda-|y_i|\theta)\ge 0$, then

$$\Omega=\{0,\max\{\xi_1,0\},\max\{\xi_2,0\}\} \tag{20}$$

where $\xi_1=\frac{1}{2}((|y_i|-\theta)+\sqrt{(|y_i|-\theta)^2-4(\lambda-|y_i|\theta)})$ and $\xi_2=\frac{1}{2}((|y_i|-\theta)-\sqrt{(|y_i|-\theta)^2-4(\lambda-|y_i|\theta)})$. Otherwise, $\Omega=\{0\}$.

Proposition 6 gives the solution of the (12) under the conditions of $R(\mathbf{x})$ with separable and strictly increasing properties. In fact, there are some other commonly used separable and non-convex $R(\mathbf{x})$, which also have the closed-form solutions similar as (16), such as $R(\mathbf{x})=\|\mathbf{x}\|_p^p$ with $p=1/2,2/3$ [14]. However, these $R(\mathbf{x})$ do not satisfy the Property 1(c), so they are not within the scope of this article. Next, we consider two special non-separable cases as the reference for other non-separable regularizations.

Proposition 7. *If $P(\mathbf{x})=R(\mathbf{x})-R(\mathbf{x}^s)$ with $R(\mathbf{x})=\|\mathbf{x}\|_2$, then the solution $\mathbf{x}^*$ of (12) is that: when $i\in\Gamma_{\mathbf{y}}^s$,*

$$x_i^*=\frac{(\|\mathbf{y}^s\|_2+\lambda)\left(\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+(\|\mathbf{y}^s\|_2+\lambda)^2}-\lambda\right)}{\|\mathbf{y}^s\|_2\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+(\|\mathbf{y}^s\|_2+\lambda)^2}}y_i \tag{21}$$

when $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$,

$$x_i^*=\frac{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+(\|\mathbf{y}^s\|_2+\lambda)^2}-\lambda}{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+(\|\mathbf{y}^s\|_2+\lambda)^2}}y_i \tag{22}$$

See Appendix F for the proof.

Proposition 8. *If $P(\mathbf{x})=R(\mathbf{x})-R(\mathbf{x}^s)$ with $R(\mathbf{x})=\|\mathbf{x}\|_1-a\|\mathbf{x}\|_2$, $0<a\le 1$, then the solution $\mathbf{x}^*$ of (12) is that:*

(1) When $|y_{\pi_y(s+1)}|>\lambda$, for $i\in\Gamma_{\mathbf{y}}^s$,

$$x_i^*=\frac{\|\mathbf{y}^s\|_2-a\lambda}{\|\mathbf{y}^s\|_2}\left(1+\frac{a\lambda}{\sqrt{\|\mathbf{z}-\mathbf{z}^s\|_2^2+(\|\mathbf{y}^s\|_2-a\lambda)^2}}\right)y_i \tag{23}$$

for $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$,

$$x_i^*=\left(1+\frac{a\lambda}{\sqrt{\|\mathbf{z}-\mathbf{z}^s\|_2^2+(\|\mathbf{y}^s\|_2-a\lambda)^2}}\right)z_i \tag{24}$$

where $z_i=y_{\pi_y(1)}$ for $i\in\Gamma_{\mathbf{y}}^s$, and $z_i=\text{shrink}(y_i,\lambda)$ for $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$.

(2) When $|y_{\pi_y(s+1)}|=\lambda$, if $a=1$, $s=1$, $|y_{\pi_y(1)}|=\lambda$, and suppose that there are k components of y_i having the same amplitude of λ, i.e., $|y_{\pi_y(s+1)}|=\cdots=|y_{\pi_y(s+k)}|=\lambda>|y_{\pi_y(s+k+1)}|$. $\mathbf{x}^*$ is an optimal solution of (12) if and only if it satisfies $\|\mathbf{x}^*\|_2=\lambda$, $x_i^*y_i\ge 0$, and $x_i^*=0$ when $i\in\{\pi_y(k+2),\pi_y(k+3),\ldots,\pi_y(N)\}$. In this case, there are infinite many solutions, Eqs. (A.40) and (A.41) are two solution examples. When $|y_{\pi_y(s+1)}|=\lambda$, and any of these conditions of $a=1$, $s=1$, $|y_{\pi_y(1)}|=\lambda$ cannot be satisfied, the solution $\mathbf{x}^*$ is

$$x_i^*=\begin{cases} y_i, & i\in\Gamma_{\mathbf{y}}^s \\ 0, & i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s \end{cases} \tag{25}$$

(3) When $0\le|y_{\pi_y(s+1)}|<\lambda$, the solution $\mathbf{x}^*$ is the same as (25).

We apply the similar proof framework in Ref [29]. for the fast ℓ_{1-2} minimization. See Appendix G for the proof.

Remark 10. When $a=0$, then $R(\mathbf{x})=\|\mathbf{x}\|_1-a\|\mathbf{x}\|_2$ reduces to $R(\mathbf{x})=\|\mathbf{x}\|_1$, and the corresponding solution $\mathbf{x}^*$ of (23,24,25) reduces to (17) as in Remark 7.

From the above Propositions 6–8 and Remarks 6–10, we can find that we have $x_i^*\le y_i$ when $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$ for all these $R(\mathbf{x})=\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2^2$, MCP, LSP, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1-a\|\mathbf{x}\|_2$, which means that it is a shrinkage operator for the bottom-$(N-s)$ elements in absolute value. However, when $i\in\Gamma_{\mathbf{y}}^s$, for the separable $R(\mathbf{x})=\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2^2$, MCP and LSP, we have $x_i^*=y_i$; for the non-separable convex $R(\mathbf{x})=\|\mathbf{x}\|_2$, we have $x_i^*\ge y_i$; for the non-separable non-convex $R(\mathbf{x})=\|\mathbf{x}\|_1-a\|\mathbf{x}\|_2$, we have $x_i^*\le y_i$. Meanwhile, when $\lambda\to\infty$, all these optimal solutions of (12) with different regularizations $P(\mathbf{x})$ degenerate into the hard thresholding operator.

4. Convergence analysis

The purpose of this section is to demonstrate that the sequence of $\{\mathbf{x}^{[k]}\}$ obtained from the FBS of (11) for the minimization problem (8) is convergent.

Theorem 2. *Suppose that functions $P(\mathbf{x})$ and $R(\mathbf{x})$ satisfy Property 1. If Assumption 1 and 2 are satisfied and $\beta<1/L$, let $\{\mathbf{x}^{[k]}\}$ be the sequence generated by the FBS of (11) for minimization problem (8), the following statements hold.*

(1) *The sequence $\{\mathbf{x}^{[k]}\}$ is bounded.*
(2) $\lim_{karrow\infty}\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\|_2=0$.
(3) *Any accumulation point of $\{\mathbf{x}^{[k]}\}$ is a stationary point of $F(\mathbf{x})$.*

Proof.

(1) Rewrite (8) and consider the following inequality

$$\begin{aligned}
&F\left(\mathbf{x}^{[k+1]}\right)-F\left(\mathbf{x}^{[k]}\right)\\
&=\phi\left(\mathbf{x}^{[k+1]}\right)+\rho P\left(\mathbf{x}^{[k+1]}\right)-\phi\left(\mathbf{x}^{[k]}\right)-\rho P\left(\mathbf{x}^{[k]}\right)\\
&\le\left\langle\nabla\phi\left(\mathbf{x}^{[k]}\right),\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\rangle+\tfrac{L}{2}\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2\\
&\quad+\rho P\left(\mathbf{x}^{[k+1]}\right)-\rho P\left(\mathbf{x}^{[k]}\right)\\
&=\rho P\left(\mathbf{x}^{[k+1]}\right)-\rho P\left(\mathbf{x}^{[k]}\right)+\tfrac{L}{2}\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2\\
&\quad+\tfrac{\left\|\mathbf{x}^{[k+1]}-\left(\mathbf{x}^{[k]}-\beta\nabla\phi\left(\mathbf{x}^{[k]}\right)\right)\right\|_2^2}{2\beta}-\tfrac{\left\|\beta\nabla\phi\left(\mathbf{x}^{[k]}\right)\right\|_2^2}{2\beta}\\
&\quad-\tfrac{\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2}{2\beta}\\
&=\rho\left(E\left(\mathbf{x}^{[k+1]}\right)-E\left(\mathbf{x}^{[k]}\right)\right)+\left(\tfrac{L}{2}-\tfrac{1}{2\beta}\right)\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2\\
&\le\left(\tfrac{L}{2}-\tfrac{1}{2\beta}\right)\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2
\end{aligned} \tag{26}$$

where the $E(\mathbf{x})$ in the third equation is the expression (13) with $\mathbf{y}$ replaced by $\mathbf{x}^{[k]}-\beta\nabla\phi(\mathbf{x}^{[k]})$ and set $\lambda=\beta\rho$. The first inequality comes from Assumption 1, and the second inequality is based on the fact that $\mathbf{x}^{[k+1]}$ is

the optimal solution of the $E(\mathbf{x})$. When $\beta < 1/L$, we have $F(\mathbf{x}^{[k]}) \le F(\mathbf{x}^{[0]})$ for all $k \ge 0$. Due to the level-boundedness of $F(\mathbf{x})$ (Assumption 2), then the sequence $\{\mathbf{x}^{[k]}\}$ is bounded.

(2) Summing both sides of (26) from $k = 0$ to ∞, we can obtain

$$\left(\frac{1}{2\beta} - \frac{L}{2}\right)\sum_{k=0}^{+\infty}\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 \le F(\mathbf{0}) - F\left(\mathbf{x}^{[k+1]}\right) < \infty \tag{27}$$

Since $\beta < 1/L$, we can deduce that $\lim_{karrow\infty}\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\|_2 = 0$ from the above relation obviously.

(3) Since the sequence $\{\mathbf{x}^{[k]}\}$ is bounded, there exists a subsequence of $\{\mathbf{x}^{[k]}\}$, denoted as $\{\mathbf{x}^{[k_j]}\}$, converging to an accumulation point $\mathbf{x}^*$. Considering that minimizer $\{\mathbf{x}^{[k_j+1]}\}$ is a critical point of (13) and $P(\mathbf{x})$ can be decomposed into DC functions as $P(\mathbf{x}) = P_1(\mathbf{x}) - P_2(\mathbf{x})$, we have

$$\begin{aligned}\mathbf{0} \in\ & \frac{\mathbf{x}^{[k_j+1]} - \mathbf{x}^{[k]} + \beta\nabla\phi\left(\mathbf{x}^{[k]}\right)}{\beta\rho} \\ & + \partial P_1\left(\mathbf{x}^{[k_j+1]}\right) - \partial P_2\left(\mathbf{x}^{[k_j+1]}\right)\end{aligned} \tag{28}$$

Let $k_j \to \infty$, by using $\|\mathbf{x}^{[k_j+1]} - \mathbf{x}^{[k_j]}\|_2 \to 0$ from the above conclusion and considering the semi-continuity of $\nabla\phi$, ∂P_1 and ∂P_2, we have that $\mathbf{0} \in \nabla\phi(\mathbf{x}^*) + \rho\partial P_1(\mathbf{x}^*) - \rho\partial P_2(\mathbf{x}^*)$. Therefore, $\mathbf{x}^*$ is a critical point of problem (8). This completes the proof.

□

From the proof of Theorem 2, we have that $\lim_{k\to\infty}\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\|_2 = 0$ is a necessary optimality condition of the FBS. Therefore, we can use $\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\|_2$ as a quantity to measure the convergence performance of the sequence $\{\mathbf{x}^{[k]}\}$ to a critical point $\mathbf{x}^*$.

Theorem 3. *If Assumptions 1 and 2 are satisfied and $\beta < 1/L$, let $\{\mathbf{x}^{[k]}\}$ be the sequence generated by the FBS of (11) for minimization problem (8), then for every $K \ge 1$, we have*

$$\min_{0\le k\le K}\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 \le 2\beta\frac{F(\mathbf{0}) - F(\mathbf{x}^*)}{K(1 - L\beta)} \tag{29}$$

Proof. Summing the inequality (26) over $k = 0, \ldots, K$, we can obtain

$$\left(\frac{1}{2\beta} - \frac{L}{2}\right)\sum_{k=0}^{K}\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 \le F(\mathbf{0}) - F\left(\mathbf{x}^{[K+1]}\right) \tag{30}$$

When $\beta < 1/L$, we have that $\{F(\mathbf{x}^{[k]})\}$ is monotonically decreasing, which means that $F(\mathbf{x}^{[K+1]}) \ge F(\mathbf{x}^*)$. Substituting this into (30), we have

$$\begin{aligned}K\min_{0\le k\le K}\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 &\le 2\beta\frac{F(\mathbf{0}) - F\left(\mathbf{x}^{[K+1]}\right)}{(1 - L\beta)} \\ &\le 2\beta\frac{F(\mathbf{0}) - F(\mathbf{x}^*)}{(1 - L\beta)}\end{aligned} \tag{31}$$

This completes the proof. □

In fact, we may have a stricter conclusion for the convergence speed as $F(\mathbf{x}^{[k+1]}) - F(\mathbf{x}^{[k]})$ can be smaller than $(\frac{L}{2} - \frac{1}{2\beta})\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\|_2^2$ in (26).

Proposition 9. *Suppose that functions $P(\mathbf{x})$ and $R(\mathbf{x})$ satisfy Property 1, and $R(\mathbf{x})$ is separable, i.e., $R(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$, and each r_i is strictly increasing on $\mathbb{R}_+$. Let $\{\mathbf{x}^{[k]}\}$ be the sequence generated by the FBS of (11) for minimization (8), then we have*

$$\begin{aligned}F\left(\mathbf{x}^{[k+1]}\right) - F\left(\mathbf{x}^{[k]}\right) \le\ & \left(\tfrac{L}{2} - \tfrac{1}{2\beta}\right)\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 \\ & + \min\left\{-\tfrac{1}{2\beta}\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2 + \rho\Delta_k, 0\right\}\end{aligned} \tag{32}$$

where $\Delta_k = \sum_{i\in\Lambda_{k+1}} r_i(x_i^{[k]}) - \sum_{i\in\Lambda_k} r_i(x_i^{[k]})$, $\Lambda_{k+1} = \Gamma^N_{\mathbf{x}^{[k+1]}}\backslash\Gamma^s_{\mathbf{x}^{[k+1]}}$, *and* $\Lambda_k = \Gamma^N_{\mathbf{x}^{[k]}}\backslash\Gamma^s_{\mathbf{x}^{[k]}}$.

See Appendix H for the proof.

From Proposition 9, we can find that $F(\mathbf{x}^{[k+1]}) - F(\mathbf{x}^{[k]}) \le (\frac{L}{2} - \frac{1}{\beta})\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\|_2^2$ if $\Gamma^s_{\mathbf{x}^{[k+1]}}$ is the same as $\Gamma^s_{\mathbf{x}^{[k]}}$.

5. Extensions

In this section, we discuss some related algorithms for solving (8), show a link between the DC function $P(\mathbf{x})$ with other regularization functions, and simply extend $P(\mathbf{x})$ to rank-constrained problem, which can enlarge the application scope of the proposed algorithm.

5.1. Related algorithms

Here, we discuss some related algorithms. When $\phi(\mathbf{x})$ is convex, it is an intuitive idea that we can use the DCA to solve the minimization (8). Since $P(\mathbf{x})$ can be written as the DC functions, i.e., $P(\mathbf{x}) = P_1(\mathbf{x}) - P_2(\mathbf{x})$, the objective function can be naturally decomposed into

$$F(\mathbf{x}) = \phi(\mathbf{x}) + \rho P(\mathbf{x}) = \{\phi(\mathbf{x}) + \rho P_1(\mathbf{x})\} - \rho P_2(\mathbf{x}) \tag{33}$$

The corresponding DCA solves the minimization problem as

$$\mathbf{x}^{[k+1]} = \arg\min_{\mathbf{x}\in\mathbb{R}^N}\left\{\phi(\mathbf{x}) + \rho P_1(\mathbf{x}) - \rho P_2\left(\mathbf{x}^{[k]}\right) - \rho\left\langle\mathbf{w}^{[k]}, \mathbf{x} - \mathbf{x}^{[k]}\right\rangle\right\} \tag{34}$$

where $\mathbf{w}^{[k]} \in \partial P_2(\mathbf{x}^{[k]})$. Although this problem is convex, it does not necessarily have closed-form solution and the computational cost is very expensive for large-scale problems.

On the other hand, since $\phi(\mathbf{x})$ is continuously differentiable with L-Lipschitz continuous gradient, we can use the Sequential Convex Programming (SCP) [67] to solve problem (8) by updating $\{\mathbf{x}^{[k]}\}$ as

$$\begin{aligned}\mathbf{x}^{[k+1]} = \arg\min_{\mathbf{x}\in\mathbb{R}^N}\Big\{ & \phi\left(\mathbf{x}^{[k]}\right) + \left\langle\nabla\phi\left(\mathbf{x}^{[k]}\right), \mathbf{x} - \mathbf{x}^{[k]}\right\rangle \\ & + \tfrac{L}{2}\left\|\mathbf{x} - \mathbf{x}^{[k]}\right\|_2^2 + \rho P_1(\mathbf{x}) - \rho P_2\left(\mathbf{x}^{[k]}\right) \\ & - \rho\left\langle\mathbf{w}^{[k]}, \mathbf{x} - \mathbf{x}^{[k]}\right\rangle\Big\}\end{aligned} \tag{35}$$

Meanwhile, the SCP can be thought as a variant of DCA with DC decomposition:

$$\begin{aligned}F(\mathbf{x}) = & \left(\rho P_1(\mathbf{x}) + L\|\mathbf{x}\|_2^2/2\right) \\ & - \left(\rho P_2(\mathbf{x}) + L\|\mathbf{x}\|_2^2/2 - \phi(\mathbf{x})\right)\end{aligned} \tag{36}$$

The subproblem can be written as

$$\begin{aligned}\mathbf{x}^{[k+1]} = \arg\min_{\mathbf{x}\in\mathbb{R}^N}\{ & \rho P_1(\mathbf{x}) \\ & + \tfrac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x}^{[k]} - \tfrac{1}{L}\left(\nabla\phi\left(\mathbf{x}^{[k]}\right) - \rho\mathbf{w}^{[k]}\right)\right)\right\|_2^2\Big\}\end{aligned} \tag{37}$$

Due to that the subproblem (37) can be solved by using the proximal operator, Ref [43]. and [44] call this type DCA as proximal DCA (PDCA). For some simple form $P(\mathbf{x})$, subproblem (37) also has closed-form solution. For example, $P(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$ and $P(\mathbf{x}) = \|\mathbf{x}\|_2^2 - \|\mathbf{x}^s\|_2^2$. In the numerical experiment, we will compare the FBS with this PDCA and show that the FBS is more efficient than PDCA in this problem. Meanwhile, as $P(\mathbf{x})$ is a DC function, the FBS reduces to the GIST algorithm proposed in [54].

To improve the performance of the FBS, some acceleration methods can be used in the proximal framework. Such as the Nonmonotone Accelerated proximal gradient (nmAPG) method [55], the extrapolation method in PDCA (pDCAe) [51] and the backtracking line search initialized method with Barzilai-Borwein (BB) rule [68] in GIST [54].

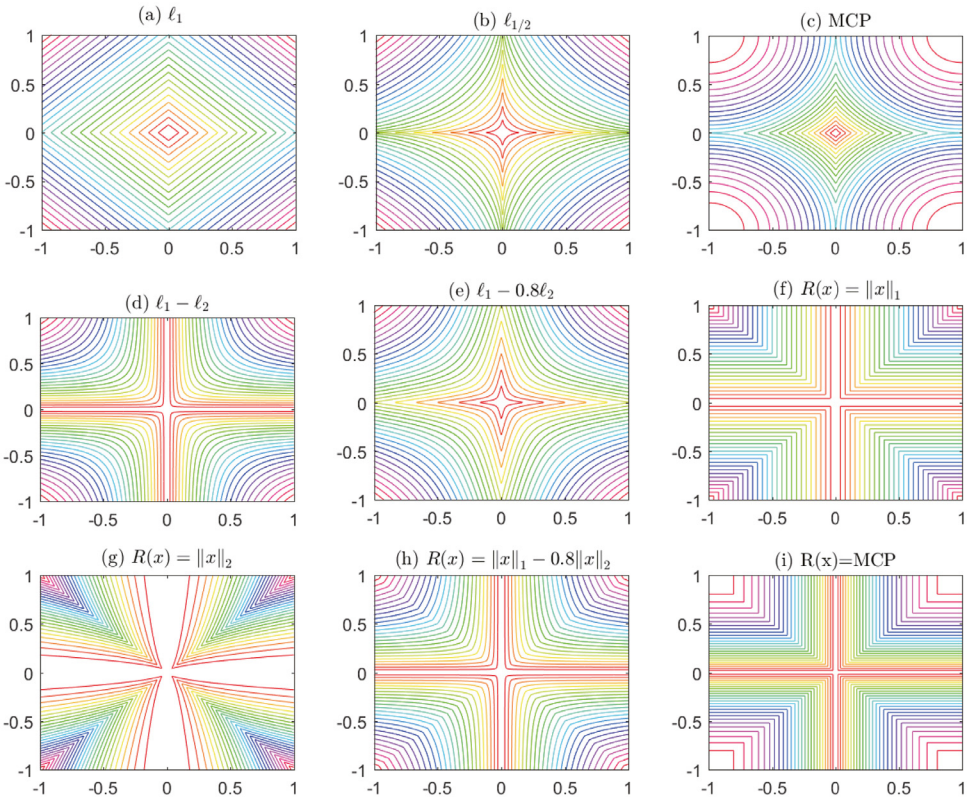


Fig. 1. Level curves of different metrics.

5.2. Comparing with other regularization

From the previous discussion, we have illustrated that the DC function $P(\mathbf{x})$ can replace the ℓ_0-norm constraint. And in Theorem 1 and Proposition 2, we have proved that the unconstrained problem (8) is equal to the original sparsity constrained problem (3) when we select proper parameter ρ. On the other hand, in the minimization problem (8), $P(\mathbf{x})$ can also be considered as a regularization function. Then, we can investigate its performance from the aspect of sparsity metric. Fig. 1 shows the contours of various regularizers for comparing.

From Fig. 1, we can find that the level curves of $R(\mathbf{x}) - R(\mathbf{x}^s)$ approach the x and y axes as the values get small, hence promoting sparsity. Inspire by Sidky et al. work of [69] and Rahimi et al. work of [70], where they using toy examples to illustrate the advantages of ℓ_p and ℓ_1/ℓ_2, respectively, we also use a similar example to show that with some special data sets (**A, b**), the $R(\mathbf{x}) - R(\mathbf{x}^s)$ tends to select a sparser solution.

Example 1. Let $N = 6$ and define

$$\mathbf{A} := \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 & 1 & 0 \\ 0.5 & 0.5 & 3 & 0 & 0 & -1 \end{bmatrix}, \mathbf{b} := \begin{bmatrix} 0 \\ 0 \\ 15 \\ 20 \\ 40 \end{bmatrix}$$

It is straightforward that any general solution of $\mathbf{Ax} = \mathbf{b}$ has the form of $\mathbf{x} = (t, t, t, 15 - 3t, 20 - 4t, 4t - 40)^T$ for a scalar $t \in \mathbb{R}$. The sparest solution occurs at $t = 0$ for the sparsity of $\mathbf{x}$ being 3, and some local solutions include $t = 5$ for sparsity being 4 and $t = 10$ for sparsity being 5. We plot the various regularization functions with respect to t in Fig. 2, including ℓ_1, ℓ_p $(p = 1/2)$, ℓ_{1-2}, ℓ_1/ℓ_2, MCP $(\theta = 15)$ of (A.3) and the proposed $R(\mathbf{x}) - R(\mathbf{x}^s)$ with $R(\mathbf{x}) = \|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1/\|\mathbf{x}\|_2$, MCP, and $s = 3$.

From Fig. 2, we can find that all these regularized functions are not differentiable at the values of $t = 0, 5$, and 10, where the corresponding sparsities of $\mathbf{x}$ are all smaller than 6. However, only the ℓ_1/ℓ_2 and the s-difference $R(\mathbf{x}) - R(\mathbf{x}^s)$ can find the sparsest vector $\mathbf{x}$ at $t = 0$ as a global minimum, where the other functions find $t = 5$ as the minimum and lead to the sparsity of $\mathbf{x}$ being 4.

5.3. Extending to rank-constrained problem

Similar to Gotoh et al. [43], the penalty function $P(\mathbf{x}) = R(\mathbf{x}) - R(\mathbf{x}^s)$ can also be extended to rank-constrained problem based on the connection between the ℓ_0-norm on $\mathbb{R}^N$ and the rank function for a matrix. The rank-constrained minimization problem can be formulated as

$$\min_{\mathbf{W}} \phi(\mathbf{W}) \quad \text{subject to} \quad \text{rank}(\mathbf{W}) \le s, \mathbf{W} \in \mathbb{R}^{M\times N} \tag{38}$$

where s is a non-negative integer with $s \le q = \min\{M, N\}$. As the rank of a matrix is equal to the number of its nonzero singular values, i.e., $\text{rank}(\mathbf{W}) = \|\sigma(\mathbf{W})\|_0$, where $\sigma(\mathbf{W})$ represents the singular value vector of $\mathbf{W}$ and $\sigma_i(\mathbf{W})$ is the i-th largest term, then we can construct the penalty functions $P, R : \mathbb{R}_+^q \to \mathbb{R}_+$, $P(\sigma(\mathbf{W})) = R(\sigma(\mathbf{W})) - R(\sigma^s(\mathbf{W}))$ that satisfy Property 1 (b) and (c), where $\sigma_i^s(\mathbf{W}) = \sigma_i(\mathbf{W})$ for $i \in \{1, 2, \dots, s\}$ and $\sigma_i^s(\mathbf{W}) = 0$ for else. Replacing the rank constraint with the DC penalty function $P(\sigma(\mathbf{W}))$ and considering the unconstrained problem:

$$\min_{\mathbf{W}} \phi(\mathbf{W}) + \rho P(\sigma(\mathbf{W})) \tag{39}$$

then we can use the FBS, DCA or ADMM algorithms to solve this rank-constrained problem.

6. Numerical experiments

In this section, simulations are performed to demonstrate the proposed conclusions and evaluate the performance of the *s*-

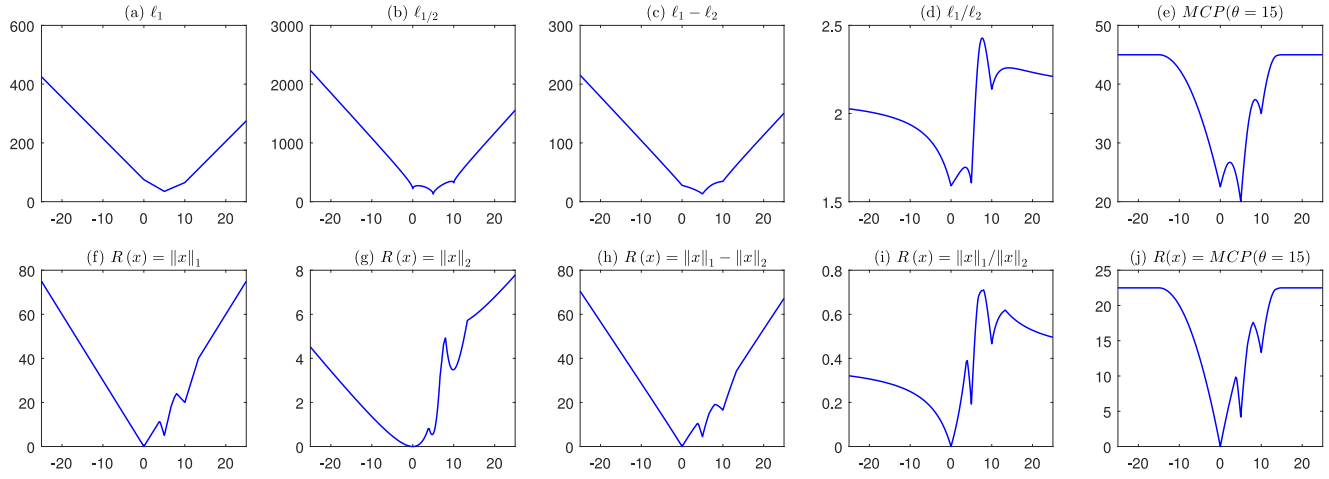


Fig. 2. The objective functions of a toy example. For the top row, from the left to right, the five columns are functions of $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_{0.5}$, $\|\mathbf{x}\|_1-\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1/\|\mathbf{x}\|_2$, MCP of (A.3) with $\theta=15$, respectively. While for the bottom row, from the left to right, the five columns are functions of $R(\mathbf{x})-R(\mathbf{x}^s)$ with $R(\mathbf{x})=\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1-\|\mathbf{x}\|_2$, $\|\mathbf{x}\|_1/\|\mathbf{x}\|_2$, MCP, respectively.

difference regularization. We mainly apply seven methods in comparison with the proposed algorithm: (1) the ℓ_1-norm regularization based ℓ_1-ADMM [72]; (2) the semismooth Newton augmented Lagrangian (SSNAL) method [75] for LASSO problem (http://www.math.nus.edu.sg/~mattohkc/SuiteLasso.html); (3) the iterative p-shrinkage (IPS) algorithm [17] with $p=-1$, which uses the p-shrinkage mapping as $S_{p,\lambda}(x_i)=\text{sign}(x_i)\max\{|x_i|-\lambda^{2-p}|x_i|^{p-1},0\}$; (4) the generalized minimax-concave penalty (GMC) with $\gamma=0.8$ [35], which uses the proximal algorithm to find the global minimizer (https://codeocean.com/2017/06/21/gmc-sparse-regularization/); (5) the ℓ_p-norm ($p=1/2$) regularization based half thresholding [14]; (6) the ℓ_0-norm regularization based accelerate IHT (AIHT) [39]; (7) the difference of the ℓ_1 and ℓ_2-norms (ℓ_{1-2}) regularization based ℓ_{1-2}-DCA [31]. We choose the representative $R(\mathbf{x})$ as $R(\mathbf{x})=\|\mathbf{x}\|_1, \|\mathbf{x}\|_2, \|\mathbf{x}\|_1-\|\mathbf{x}\|_2$ for comparing. Meanwhile, among all these penalties of comparisons, the GMC, ℓ_{1-2} and the proposed s-difference penalties all belong to the type of difference of convex functions. However, the solutions of these penalties based optimization problems are different: the GMC based problem is minimized by finding the saddle-point with proximal algorithms comprising simple computations in [35], the ℓ_{1-2} based problem is approximately solved by using the DCA framework in [31], while the s-difference penalty regularized problem is approximately solved by using the iterative FBS. In addition, although the GMC is nonconvex, the convexity of the total objective function is maintained, which means that it allows the leveraging of globally convergent. This convexity preservation is also the most attractive aspect of GMC. All experiments are performed in MATLAB 2015b running on ASUS laptop with Intel (R) Core (TM) i7-8550U CPU, 8 GB of RAM and 64-bit Windows 10 operating system.

We focus on the following least squares problem:

$$\min_{\mathbf{x}\in\mathbb{R}^N}\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2+\rho P(\mathbf{x}) \tag{40}$$

and conduct experiments on simulated vector signals.

We test two types of matrices $\mathbf{A}$: the random Gaussian matrix with i.i.d. standard Gaussian entries and being normalized that each column has unit norm, and the random partial DCT matrix which is formed by randomly selecting rows from the full DCT matrix. For the original sparse vector $\bar{\mathbf{x}}$, we generate it with random index set and draw non-zero elements with standard normal distribution. The observation is $\mathbf{b}=\mathbf{A}\bar{\mathbf{x}}+\mathbf{n}$, where $\mathbf{n}$ is zeros for the noiseless test, and Gaussian noise for the contaminated measurements. The initial value for all the methods is an approximated solution of the ℓ_1 minimization using ADMM after N iterations. The maximum number of iterations for all these methods is $5N$ except for DCA, whose maximum number of internal iterations is $5N$ and the maximum number of external iterations is 20. The stopping condition is set to be $\frac{\|\mathbf{x}^{[k]}-\mathbf{x}^{[k-1]}\|_2}{\max\{\|\mathbf{x}^{[k]}\|_2,1\}}<10^{-5}$.

In the first study, we look at the success rates with 100 random instances under the noise-free condition, in which we set the size of matrices $\mathbf{A}$ as 64×256 and let the sparsity level of $\bar{\mathbf{x}}$ being 1, 2, 4, 6, ..., 40. We vary the regularization parameter ρ from 10^{-6} to 10 (with 30 logarithmically equally spaced) and set $\beta=10\rho$ for each method, and then select the best one as the result. Here we consider a recovery $\mathbf{x}^*$ as successful if the relative error of recovery (Rel.Err) satisfies $\|\mathbf{x}^*-\bar{\mathbf{x}}\|_2/\|\bar{\mathbf{x}}\|_2\le10^{-3}$. In addition, we set sparsity parameter s to be the ground truth s_{truth} for the proposed s-difference $P(\mathbf{x})$. Fig. 3 plots the success rates of the comparing methods for both the Gaussian matrix and the partial DCT matrix. From this, we can find that the s-difference regularization with $R(\mathbf{x})=\|\mathbf{x}\|_1$ has the best performance for both Gaussian matrix and partial DCT matrix. The $R(\mathbf{x})=\|\mathbf{x}\|_1-\|\mathbf{x}\|_2$ is comparable to ℓ_{1-2}-DCA, followed by $R(\mathbf{x})=\|\mathbf{x}\|_2$, half thresholding and GMC, which outperform the SSNAL, IPS and ℓ_1-ADMM.

In the second study, we focus on the recovery quantity of these methods under different sizes of matrix. For the noiseless case, we set $\rho=10^{-1}$ for s-difference regularization based FBS and $\rho=10^{-6}$ for the ADMM and other types of methods. We set $\beta=10\rho$, and consider $(M,N,s_{truth})=(256i,1024i,48i)$ for $i=1,2,\ldots,8$. Here, we also set the sparsity threshold parameter to be s_{truth} for the AIHT and s-difference $P(\mathbf{x})$. For each triple (M, N, s_{truth}), we generate 30 random realizations. Tables 2 and 3 list the mean and standard deviation of Rel.Err for Gaussian matrix and partial DCT matrix, respectively. We also test these methods in the presence of Gaussian noise as $\mathbf{n}=0.01*randn(M,1)$. We set $\rho=1$ for s-difference regularization based FBS and $\rho=10^{-3}$ for the ADMM and other types of methods, and consider $(M,N,s_{truth})=(256i,1024i,48i)$ for $i=1,2,3,4$. The recovery performance is listed in Tables 4 and 5 for comparing. From Tables 2–5, we can find that the s-difference $P(\mathbf{x})$ with the ground truth sparsity threshold parameter can provide a quite competitive or slightly superior performance comparing with AIHT and other methods under the noise-free conditions. However, under the condition of noise, AIHT performance decreases rapidly, while the s-difference $P(\mathbf{x})$ is still able to provide a relatively best result, especially the $P(\mathbf{x})$ with $R(\mathbf{x})=\|\mathbf{x}\|_1-\|\mathbf{x}\|_2$. From the first and second studies, we can find that the s-difference $P(\mathbf{x})$ can obtain a

Table 2
Mean and standard deviation of Rel.Err for different methods with Gaussian matrix under noiseless condition.

M	N	s_{true}	ℓ_1-ADMM	SSNAL	IPS	GMC	ℓ_{1-2}-DCA	Half thresholding	AIHT	s-difference (ℓ_1)	s-difference (ℓ_{1-2})	s-difference (ℓ_2)
256	1024	48	1.212e-04	1.613e-05	1.877e-05	1.670e-05	2.390e-05	2.361e-05	**1.436e-05**	1.546e-05	1.548e-05	1.558e-05
			(4.427e-08)	(1.162e-11)	(1.780e-10)	(6.475e-11)	(2.073e-11)	(4.974e-10)	(3.670e-11)	(3.675e-11)	(1.809e-10)	(1.772e-10)
512	2048	96	9.836e-05	1.615e-05	1.607e-05	1.529e-05	2.601e-05	2.079e-05	1.319e-05	1.323e-05	**1.281e-05**	1.290e-05
			(1.774e-08)	(5.634e-12)	(3.291e-12)	(2.990e-12)	(1.224e-11)	(7.632e-12)	(2.544e-12)	(2.581e-12)	(2.307e-12)	(2.341e-12)
768	3072	144	1.014e-04	1.662e-05	1.640e-05	1.461e-05	2.483e-05	2.125e-05	1.351e-05	1.372e-05	**1.317e-05**	1.325e-05
			(1.346e-08)	(4.768e-12)	(2.322e-12)	(2.302e-12)	(4.595e-12)	(7.547e-12)	(2.001e-12)	(2.123e-12)	(1.892e-12)	(1.910e-12)
1024	4096	192	1.205e-04	1.562e-05	1.562e-05	1.433e-05	2.529e-05	2.093e-05	1.284e-05	1.444e-12	**1.254e-05**	1.260e-05
			(2.263e-08)	(3.440e-12)	(1.834e-12)	(2.268e-12)	(8.490e-12)	(8.213e-12)	(1.444e-12)	(1.462e-12)	(1.447e-12)	(1.454e-12)
1280	5120	240	1.297e-04	1.642e-05	1.610e-05	1.507e-05	2.533e-05	2.113e-05	**1.288e-05**	1.316e-05	1.292e-05	1.297e-05
			(1.422e-08)	(3.132e-12)	(1.839e-12)	(2.379e-12)	(9.386e-12)	(7.449e-12)	(1.455e-12)	(1.533e-12)	(1.477e-12)	(1.486e-12)
1536	6144	288	1.155e-04	1.576e-05	1.574e-05	1.449e-05	2.468e-05	2.108e-05	1.294e-05	1.282e-05	**1.256e-05**	1.261e-05
			(1.397e-08)	(1.992e-12)	(1.219e-12)	(2.208e-12)	(7.396e-12)	(8.514e-12)	(9.803e-13)	(9.757e-13)	(9.163e-13)	(9.234e-13)
1792	7168	336	1.069e-04	1.540e-05	1.538e-05	1.436e-05	2.527e-05	2.101e-05	1.265e-05	1.270e-05	**1.229e-05**	1.233e-05
			(1.089e-08)	(1.600e-12)	(1.062e-12)	(2.017e-12)	(9.967e-12)	(6.653e-12)	(9.040e-13)	(9.084e-13)	(8.588e-13)	(8.629e-13)
2018	8192	384	1.456e-04	1.573e-05	1.564e-05	1.442e-05	2.539e-05	2.067e-05	1.287e-05	1.279e-05	**1.251e-05**	1.255e-05
			(1.810e-08)	(1.405e-12)	(1.029e-12)	(1.986e-12)	(4.754e-12)	(7.831e-12)	(7.885e-13)	(7.812e-13)	(7.730e-13)	(7.783e-13)

Table 3
Mean and standard deviation of Rel.Err for different methods with partial DCT matrix under noiseless condition.

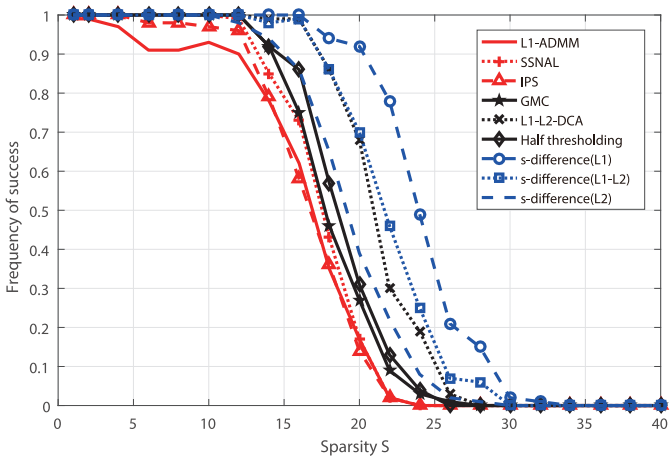
M	N	s_{true}	ℓ_1-ADMM	SSNAL	IPS	GMC	ℓ_{1-2}-DCA	Half thresholding	AIHT	s-difference (ℓ_1)	s-difference (ℓ_{1-2})	s-difference (ℓ_2)
256	1024	48	6.766e-05	5.792e-06	5.119e-06	4.693e-06	2.521e-05	7.681e-06	4.208e-06	3.132e-06	**2.969e-06**	3.113e-06
			(2.360e-09)	(1.345e-12)	(4.761e-13)	(3.474e-13)	(7.891e-12)	(4.678e-12)	(3.970e-13)	(2.459e-13)	(2.743e-13)	(2.958e-13)
512	2048	96	9.139e-05	5.770e-06	5.466e-06	4.682e-06	2.298e-05	7.936e-06	4.509e-06	3.437e-06	**3.200e-06**	3.310e-06
			(1.507e-08)	(8.264e-13)	(3.370e-12)	(2.995e-12)	(2.539e-11)	(5.357e-12)	(2.496e-12)	(2.129e-12)	(1.456e-12)	(1.528e-12)
768	3072	144	7.828e-05	5.845e-06	5.265e-06	4.546e-06	2.418e-05	7.860e-06	4.343e-06	3.254e-06	**3.075e-06**	3.165e-06
			(4.769e-09)	(4.563e-13)	(2.411e-13)	(2.368e-13)	(1.930e-11)	(8.569e-12)	(2.047e-13)	(1.973e-13)	(1.447e-13)	(1.513e-13)
1024	4096	192	1.088e-04	5.567e-06	5.049e-06	4.881e-06	2.189e-05	7.700e-06	4.156e-06	3.065e-06	**2.935e-06**	3.008e-06
			(1.350e-08)	(2.997e-13)	(1.758e-13)	(1.630e-13)	(2.626e-11)	(8.614e-12)	(1.495e-13)	(1.119e-13)	(1.052e-13)	(1.095e-13)
1280	5120	240	1.201e-04	5.745e-06	5.175e-06	4.621e-06	2.398e-05	7.868e-06	4.267e-06	3.284e-06	**3.018e-06**	3.086e-06
			(1.605e-08)	(1.215e-13)	(6.515e-14)	(5.992e-14)	(1.856e-11)	(9.326e-12)	(5.541e-14)	(5.338e-14)	(4.026e-14)	(4.148e-14)
1536	6144	288	1.085e-04	5.793e-06	5.351e-06	5.075e-06	2.418e-05	6.895e-06	4.409e-06	3.265e-06	**3.117e-06**	3.181e-06
			(1.192e-08)	(2.182e-13)	(4.024e-13)	(3.065e-13)	(1.533e-11)	(9.014e-12)	(2.849e-13)	(1.976e-13))	(1.550e-13)	(1.613e-13))
1792	7168	336	1.441e-04	5.632e-06	5.117e-06	4.605e-06	2.224e-05	7.843e-06	4.212e-06	3.178e-06	**2.973e-06**	3.030e-06
			(1.177e-08)	(1.092e-13)	(6.588e-14)	(4.996e-14)	(2.069e-11)	(6.979e-12)	(5.368e-14)	(3.995e-14)	(3.654e-14)	(3.779e-14)
2018	8192	384	8.597e-05	5.641e-06	5.131e-06	4.639e-06	2.286e-05	7.654e-06	4.217e-06	3.188e-06	**2.969e-06**	3.023e-06
			(8.211e-09)	(1.210e-13)	(6.572e-14)	(5.519e-14)	(1.592e-11)	(5.937e-12)	(5.711e-14)	(5.042e-14)	(4.083e-14)	(4.188e-14)

Table 4
Mean and standard deviation of Rel.Err for different methods with Gaussian matrix under Gaussian noise.

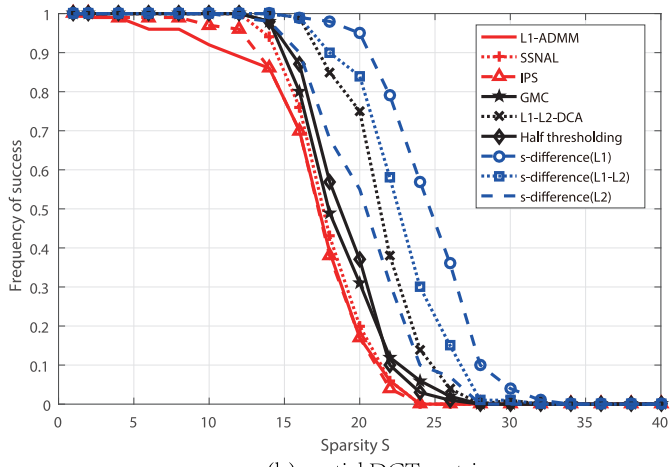
M	N	s_{true}	ℓ_1-ADMM	SSNAL	IPS	GMC	ℓ_{1-2}-DCA	Half thresholding	AIHT	s-difference (ℓ_1)	s-difference (ℓ_{1-2})	s-difference (ℓ_2)
256	1024	48	1.236e-01	1.222e-01	1.217e-01	1.035e-01	1.050e-01	7.381e-02	2.123e-01	6.192e-02	**5.857e-02**	5.887e-02
			(4.490e-04)	(4.419e-04)	(4.331e-04)	(2.065e-04)	(2.965e-04)	(1.717e-04))	(1.539e-03)	(3.765e-04)	(2.500e-04)	(2.369e-04)
512	2048	96	1.234e-01	1.218e-01	1.210e-01	1.049e-01	1.081e-01	9.214e-02	2.154e-01	6.185e-02	**5.958e-02**	5.977e-02
			(3.272e-04)	(3.261e-04)	(3.245e-04)	(2.126e-04)	(2.027e-04)	(7.262e-05)	(7.431e-04)	(1.705e-04)	(1.431e-04)	(1.438e-04)
768	3072	144	1.243e-01	1.224e-01	1.219e-01	1.087e-01	1.118e-01	1.020e-01	2.202e-01	6.296e-02	6.148e-02	**6.120e-02**
			(1.266e-04)	(1.289e-04)	(1.275e-04)	(1.155e-04)	(9.412e-05)	(9.371e-05)	(3.419e-04))	(9.727e-05)	(7.771e-05))	(7.661e-05)
1024	4096	192	1.228e-01	1.208e-01	1.202e-01	1.052e-01	1.097e-01	1.101e-01	2.182e-01	6.188e-02	**5.901e-02**	5.989e-02
			(6.416e-05)	(6.635e-05)	(6.516e-05)	(9.954e-05)	(5.421e-05)	(4.824e-05)	(2.430e-04)	(9.485e-05)	(7.434e-05)	(7.842e-05)

Table 5
Mean and standard deviation of Rel.Err for different methods with partial DCT matrix under Gaussian noise.

M	N	s_{true}	ℓ_1-ADMM	SSNAL	IPS	GMC	ℓ_{1-2}-DCA	Half thresholding	AIHT	s-difference (ℓ_1)	s-difference (ℓ_{1-2})	s-difference (ℓ_2)
256	1024	48	7.361e-02	7.352e-02	7.348e-02	6.837e-02	6.273e-02	4.303e-02	1.777e-01	4.005e-02	**3.105e-02**	3.116e-02
			(1.460e-04)	(1.358e-04)	(1.231e-04)	(9.501e-05)	(8.720e-05)	(4.242e-05)	(9.091e-04)	(5.439e-05)	(3.367e-05)	(4.104e-05)
512	2048	96	7.941e-02	7.934e-02	8.033e-02	7.529e-02	7.140e-02	5.271e-02	1.857e-01	3.951e-02	**3.245e-02**	3.273e-02
			(9.035e-05)	(9.012e-05)	(8.977e-05)	(7.255e-05)	(6.952e-05)	(3.233e-05)	(5.822e-04)	(5.122e-05)	(2.568e-05)	(2.506e-05)
768	3072	144	7.428e-02	7.288e-02	7.125e-02	7.039e-02	6.760e-02	6.078e-02	1.728e-01	4.299e-02	**3.032e-02**	3.062e-02
			(5.154e-05)	(5.123e-05)	(5.045e-05)	(4.880e-05)	(4.318e-05)	(1.717e-05)	(1.917e-04))	(3.023e-05)	(3.032e-02))	(9.071e-06)
1024	4096	192	7.480e-02	7.369e-02	7.357e-02	7.163e-02	6.888e-02	6.608e-02	1.758e-01	4.289e-02	2.997e-02	**2.984e-02**
			(3.288e-05)	(3.144e-05)	(3.055e-05)	(2.835e-05)	(2.619e-05)	(1.337e-05)	(1.537e-04)	(9.025e-06)	(6.900e-06)	(6.853e-06)



(a) Gaussian matrix



(b) partial DCT matrix

Fig. 3. Success rates versus sparsity for compared methods: (a) Gaussian matrix, (b) partial DCT matrix.

better recovery than other compared methods. However, we can also find that it is hard to determine theoretically which one is the best for all application scenarios. This is maybe because that all these $P(\mathbf{x})$ are non-convex regularizations. Their similarity is that they are equal to zero when the sparsity level of $\mathbf{x}$ is under s, and the difference between these $P(\mathbf{x})$ is that they use different ways to punish the signals $\mathbf{x}$ that are not sparse enough, which leads to different shrinkage operators as shown in section III.

In the third study, we focus on the accuracy and efficiency of the methods under fixed matrix $\mathbf{A}$ and sparsity level as $(M, N, s_{truth}) = (256, 1024, 48)$. To illustrate the benefit of the closed-form solutions of proposed s-difference regularization, we selectively analyse the performance of DCA, PDCA and FBS under the condition of the same regularization $P(\mathbf{x}) = \|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$. The DCA solves the minimization problem (40) by using (34), which is

$$\mathbf{x}^{[k+1]} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \rho\|\mathbf{x}\|_1 - \rho\left\langle \mathbf{w}^{[k]}, \mathbf{x}\right\rangle \right\} \tag{41}$$

where $\mathbf{w}^{[k]} \in \partial\|\mathbf{x}^{s[k]}\|_1$. This problem can be solved by ADMM as

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{v}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \rho\|\mathbf{v}\|_1 - \rho\left\langle \mathbf{w}^{[k]}, \mathbf{x}\right\rangle \right\} \\ &subject\ to \quad \mathbf{x}-\mathbf{v}=0 \end{aligned} \tag{42}$$

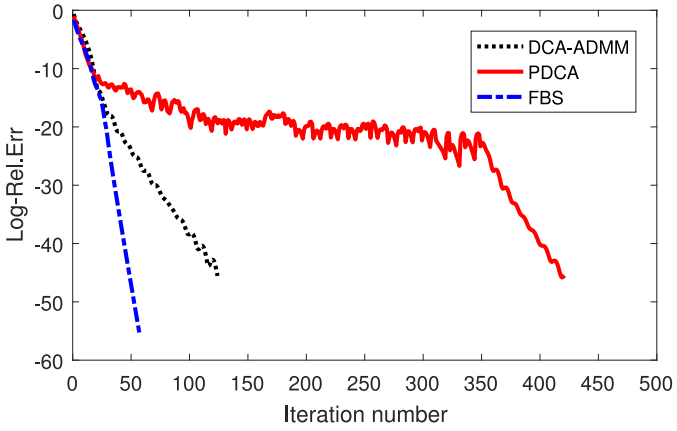


Fig. 4. Convergence performance of DCA-ADMM, PDCA and FBS for solving the s-difference $\|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$ regularization problem.

We denote this method as DCA-ADMM for short. The PDCA solve the minimization problem (40) by using (37), that is

$$\begin{aligned} \mathbf{x}^{[k+1]} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} \{&\rho\|\mathbf{x}\|_1 \\ &+\tfrac{L}{2}\left\|\mathbf{x}-\left(\mathbf{x}^{[k]}-\tfrac{1}{L}\left(\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^{[k]}-\mathbf{b}\right)-\rho\mathbf{w}^{[k]}\right)\right)\right\|_2^2\Big\} \end{aligned} \tag{43}$$

and it can be solved by using the soft shrinkage operator (18). We denote this method as PDCA for short. The FBS solves the problem by using closed-form solution (17) in Remark 7.

Fig. 4 shows the convergence performance of three methods under noise-free condition with partial DCT matrix, which is measured by the Log-Rel.Err (defined as $10\log_{10}$(Rel.Err)) versus iteration numbers. Table 6 lists the mean of relative error, iteration number and computational time (in seconds) under the noise-free and Gaussian noise conditions as $\mathbf{n} = 0.01 * randn(M, 1)$. From Fig. 4 and Table 6, it is clear that the FBS with closed-form method leads to less error and converges faster than the DCA type methods.

From the definition of s-difference $P(\mathbf{x})$, it is easy to understand that the parameter s plays an important role in the algorithm. Here we focus on the problem of how to select the proper parameter s. We consider the fixed matrix $\mathbf{A}$ and s_{truth} as $(M, N, s_{truth}) = (256, 1024, 48)$. Fig. 5 shows the performance of s-difference $P(\mathbf{x}) = (\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2) - (\|\mathbf{x}^s\|_1 - \|\mathbf{x}^s\|_2)$ under different s from 1 to 1000. In addition to using the FBS with closed-form solution as Proposition 8, we also consider the approximate DCA-ADMM using the similar solution of (42) but set $\mathbf{w}^{[k]} \in \partial(\|\mathbf{x}^{[k]}\|_2 + \|\mathbf{x}^{s[k]}\|_1 - \|\mathbf{x}^{s[k]}\|_2)$. This method is not a true DCA as the decomposition is not a convex function. However, this DCA-ADMM still works well as shown in Fig. 5. From Fig. 5, we can find that once the parameter s is less than the true sparsity s_{truth}, the performance of FBS with closed-form will drop sharply. However, the DCA-ADMM is almost unaffected. This is probably because that the FBS solves the problem as the hard thresholding way when $|y_{\pi_y(s+1)}|$ is smaller than λ in Proposition 8, whereas the DCA-ADMM make full use of the non-convex $P(\mathbf{x})$ and bring better results than the ℓ_1-norm methods. According to this deduction, designing an adaptive penalty parameter for FBS is quite necessary, which also is our future work. The good performance of DCA-ADMM also shows the superiority of this s-difference regularization from another point of view.

From Fig. 5, we also have a suggestion that if we already have a preliminary range of judgements about sparsity based on prior knowledge, i.e., $s_{truth} \in (s_{\max}, s_{\min})$, then we suggest that s decreases from the $s_{\max}$, but no less than $s_{\min}$, or just set s be equal to $s_{\max}$ when the range of sparsity is not very large. Here, we also introduce an adjustment strategy to estimate

Table 6
Mean of relative error, iteration number and computational time (sec.) under the noise-free and Gaussian noise conditions.

Methods	Noiseless condition Gaussian matrix		Noiseless condition partial DCT matrix		Noisy condition Gaussian matrix		Noisy condition partial DCT matrix	
	Rel.Err	Iter/Time	Rel.Err	Iter/Time	Rel.Err	Iter/Time	Rel.Err	Iter/Time
ℓ_1-ADMM	1.098E-04		1.357E-04		1.198E-01		7.485E-02	
$\|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$ (DCA-ADMM)	2.298E-05	178/0.05	2.501E-05	170/0.05	7.182E-02	302/0.08	4.430E-02	511/0.12
$\|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$ (PDCA)	3.735E-05	530/0.13	4.063E-05	460/0.12	1.179E-01	5120/1.46	1.005E-01	3559/1.08
$\|\mathbf{x}\|_1 - \|\mathbf{x}^s\|_1$ (FBS)	1.368E-05	126/0.04	3.059E-06	65/0.03	6.190E-02	195/0.06	4.264E-02	108/0.05

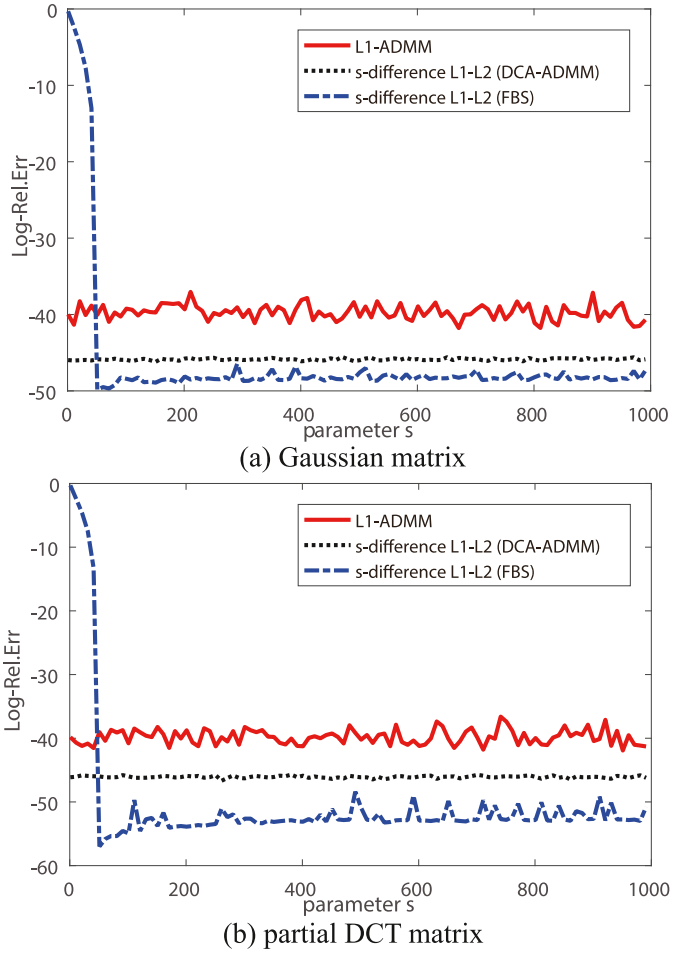


Fig. 5. Recovery performance of DCA-ADMM and FBS for solving the s-difference regularization problem with different parameter s : (a) Gaussian matrix, (b) partial DCT matrix.

the parameter s when we don't know the prior sparsity range: set $s^{[k+1]} = size(find(|\mathbf{x}^{[k]}| \geq \min\{|x^{[k-1]}_{\pi_x(s^{[k-1]})}|, \varepsilon\}))$, where constant $\varepsilon > 0$ is given. Some experiments show that this adjustment strategy can often find the approximate true sparsity level s_{truth}, which means that it maybe can be used to estimate the sparsity of the unknown signal.

7. Conclusion

In this paper, we propose a new s-difference type penalty function for the sparse optimization problem, which is the difference of the normal convex or non-convex penalty function and its corresponding s-truncated function. To approximately solve this non-convex regularization problem, we use the FBS method based on the proximal operator, which has some cheap closed-form solutions for commonly used $R(\mathbf{x})$, such as ℓ_1, ℓ_2, ℓ_{1-2} and so on. The convergence and effectiveness of the proposed algorithm are proved and demonstrated by the theoretical proof and numerical experiments, respectively. In addition, we have observed that the DCA with s-difference regularization gives better recovery results than the FBS using close-form solutions when the parameter s is less than the true sparsity, which motivates us to find an adaptive strategy for the penalty and sparsity parameters in the future. Meanwhile, how to choose the appropriate s-difference $P(\mathbf{x})$ among these different regularization functions for specific application scenarios is also the unsolved problem that we need to consider.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled "Sparse Optimization Problem with s-difference Regularization".

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (61701508). The authors would like to thank the editors and anonymous reviewers for their careful reading of an earlier version of this article and constructive suggestions that improved the presentation of this work.

Appendix A. Proof of Proposition 1

To prove the Proposition 1, we use the following Lemma:

Lemma 1. If $R : \mathbb{R}^N \to \mathbb{R}$ is convex, then for any $s \in \{1, 2, \ldots, N\}$, $R(\mathbf{x}^s)$ is also convex.

Proof. let $\mathbf{v} = diag\{v_1, v_2, \ldots, v_N\}$, since $R(\mathbf{x})$ is convex, then $R(\mathbf{vx})$ is convex. Then the $R(\mathbf{x}^s)$ can be written as a pointwise maximum of convex functions:

$$R(\mathbf{x}^s) = \max_{\mathbf{v}}\{R(\mathbf{vx}) : v_i \in \{0, 1\}, \|\mathbf{v}\|_1 = s\} \tag{A.1}$$

Then we have that $R(\mathbf{x}^s)$ is convex.

(1) For the convex and separable functions $R(\mathbf{x}) = \|\mathbf{x}\|_p^p (p \geq 1)$, such as $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2^2$, and the convex and non-separable functions $R(\mathbf{x}) = \|\mathbf{x}\|_p, (p > 1)$, such as $R(\mathbf{x}) = \|\mathbf{x}\|_2$, it is obviously that they satisfy Property 1(a) and (b). Then by using Lemma 1, it completes the Property 1(c) by setting $P_1(\mathbf{x}) = R(\mathbf{x})$ and $P_2(\mathbf{x}) = R(\mathbf{x}^s)$.

(2) For the non-convex and separable functions $R(\mathbf{x}) = \sum_{i=1}^N r_i(x_i)$, where $r_i(x_i)$ are equations (A.2), (A.3) and (A.4) corresponding to LSP, MCP and SCAD, respectively.

$$r_i(x_i) = \log(1 + |x_i|/\theta), \theta > 0 \tag{A.2}$$

$$r_i(x_i) = \begin{cases} |x_i| - x_i^2/(2\theta), & |x_i| \leq \theta \\ \theta/2, & |x_i| > \theta \end{cases}, \theta > 0 \tag{A.3}$$

$$r_i(x_i)=\begin{cases}|x_i|, & |x_i|<1\\ \frac{2\theta|x_i|-x_i^2-1}{2(\theta-1)}, & 1\le|x_i|<\theta \quad ,\theta>2\\ (\theta+1)/2, & |x_i|\ge\theta\end{cases} \tag{A.4}$$

Property 1(a) and (b) are obvious. Then we need to give the DC formulations for $P(\mathbf{x})$. Take the LSP as an example, we have that

$$\begin{aligned}&\tfrac{\|\mathbf{x}^s\|_1}{\theta}-R(\mathbf{x}^s)\\&=\max_{\mathbf{v}}\left\{\sum_{i=1}^{N}\tfrac{|v_ix_i|}{\theta}-\log\left(1+\tfrac{|v_ix_i|}{\theta}\right):v_i\in\{0,1\},\|\mathbf{v}\|_1=s\right\}\end{aligned} \tag{A.5}$$

which means that $\|\mathbf{x}^s\|_1/\theta-R(\mathbf{x}^s)$ is convex as $|v_ix_i|/\theta-\log(1+|v_ix_i|/\theta)$ is convex. Then we can rewrite $P(\mathbf{x})$ as

$$\begin{aligned}P(\mathbf{x})&=R(\mathbf{x})-R(\mathbf{x}^s)\\&=\underbrace{\{\|\mathbf{x}\|_1/\theta+(\|\mathbf{x}^s\|_1/\theta-R(\mathbf{x}^s))\}}_{P_1(\mathbf{x})}\\&\quad-\underbrace{\{\|\mathbf{x}^s\|_1/\theta+(\|\mathbf{x}\|_1/\theta-R(\mathbf{x}))\}}_{P_2(\mathbf{x})}\end{aligned} \tag{A.6}$$

where $P_1(\mathbf{x})$ and $P_2(\mathbf{x})$ are two convex functions. For MCP and SCAD, we can obtain similar formulations in the same way.

(3) For the non-convex and non-separable functions, when $R(\mathbf{x})=\|\mathbf{x}\|_1-a\|\mathbf{x}\|_2, 0<a\le1$, we have $R(\mathbf{x})=R(-\mathbf{x})$. When $\|\mathbf{x}\|_0\le s$, it is easy to see that $P(\mathbf{x})=0$. When $P(\mathbf{x})=0$, we have $\|\mathbf{x}\|_0\le s$; otherwise $\|\mathbf{x}\|_0>s$, then $\|\mathbf{x}\|_2^2\le\|\mathbf{x}^s\|_2^2+(\|\mathbf{x}\|_1-\|\mathbf{x}^s\|_1)^2<(\|\mathbf{x}^s\|_2+\|\mathbf{x}\|_1-\|\mathbf{x}^s\|_1)^2$, then we have $\|\mathbf{x}\|_2-\|\mathbf{x}^s\|_2<\|\mathbf{x}\|_1-\|\mathbf{x}^s\|_1$, which means that $P(\mathbf{x})=R(\mathbf{x})-R(\mathbf{x}^s)=\|\mathbf{x}\|_1-\|\mathbf{x}^s\|_1-a(\|\mathbf{x}\|_2-\|\mathbf{x}^s\|_2)>0$, and this is contradiction to $P(\mathbf{x})=0$. Meanwhile, $P(\mathbf{x})$ can be formulated as

$$\begin{aligned}P(\mathbf{x})&=R(\mathbf{x})-R(\mathbf{x}^s)\\&=\underbrace{\{\|\mathbf{x}\|_1+a\|\mathbf{x}^s\|_2\}}_{P_1(\mathbf{x})}-\underbrace{\{\|\mathbf{x}^s\|_1+a\|\mathbf{x}\|_2\}}_{P_2(\mathbf{x})}\end{aligned} \tag{A.7}$$

when $R(\mathbf{x})$ is the non-separable LSP denoted as $R(\mathbf{x})=\log(1+\|\mathbf{x}\|_2/\theta), \theta>0$, Property 1(a) and (b) are obvious. Note that $\|\mathbf{x}^s\|_2/\theta-R(\mathbf{x}^s)$ can be thought as a composition function $h\circ g$, where $h(x)=|x|/\theta-\log(1+|x|/\theta)$ and $g(\mathbf{x})=\|\mathbf{x}^s\|_2$, by using the above deduction, we have that $\|\mathbf{x}^s\|_2/\theta-R(\mathbf{x}^s)$ is convex. Then $P(\mathbf{x})$ can be rewritten as

$$\begin{aligned}P(\mathbf{x})&=R(\mathbf{x})-R(\mathbf{x}^s)\\&=\underbrace{\{\|\mathbf{x}\|_2/\theta+(\|\mathbf{x}^s\|_2/\theta-R(\mathbf{x}^s))\}}_{P_1(\mathbf{x})}\\&\quad-\underbrace{\{\|\mathbf{x}^s\|_2/\theta+(\|\mathbf{x}\|_2/\theta-R(\mathbf{x}))\}}_{P_2(\mathbf{x})}\end{aligned} \tag{A.8}$$

For the non-separable type MCP and SCAD, we can obtain similar formulations in the same way. □

Appendix B. Proof of Theorem 1

Proof. This theorem can be proved in a similar manner to Theorem 17.1 in [71]. Let $\hat{\mathbf{x}}$ be an optimal solution of (3), that is,

$$\phi(\hat{\mathbf{x}})\le\phi(\mathbf{x}) \quad \text{for all} \quad \mathbf{x} \quad \text{with} \quad \|\mathbf{x}\|_0\le s \tag{A.9}$$

Since $\mathbf{x}_t$ minimizes (8) at $\rho=\rho_t$, we have that

$$\phi(\mathbf{x}_t)+\rho_tP(\mathbf{x}_t)\le\phi(\hat{\mathbf{x}})+\rho_tP(\hat{\mathbf{x}})=\phi(\hat{\mathbf{x}}) \tag{A.10}$$

By rearranging this expression, we have

$$R(\mathbf{x}_t)-R(\mathbf{x}_t^s)\le\frac{1}{\rho_t}\left(\phi(\hat{\mathbf{x}})-\phi(\mathbf{x}_t)\right) \tag{A.11}$$

Suppose that $\bar{\mathbf{x}}$ is a limit point of $\{\mathbf{x}_t\}$, then there exits an infinite subsequence $\mathcal{T}$ such that $\lim_{t\in\mathcal{T}}\mathbf{x}_t=\bar{\mathbf{x}}$. By taking the limit as $t\to\infty, t\in\mathcal{T}$, on both side of (A.11)

$$0\le R(\bar{\mathbf{x}})-R(\bar{\mathbf{x}}^s)\le\lim_{t\in\mathcal{T}}\frac{1}{\rho_t}\left(\phi(\hat{\mathbf{x}})-\phi(\mathbf{x}_t)\right)=0 \tag{A.12}$$

Therefore, we have that $R(\bar{\mathbf{x}})-R(\bar{\mathbf{x}}^s)=0$, which means that $\bar{\mathbf{x}}$ is feasible to (3). Moreover, by taking the limit as $t\to\infty$ for $t\in\mathcal{T}$ on (A.10), we have that

$$\phi(\bar{\mathbf{x}})\le\phi(\bar{\mathbf{x}})+\lim_{t\in\mathcal{T}}\rho_tP(\mathbf{x}_t)\le\phi(\hat{\mathbf{x}}) \tag{A.13}$$

Since $\bar{\mathbf{x}}$ is feasible to (3) and $\hat{\mathbf{x}}$ is an optimal solution of (3), then $\bar{\mathbf{x}}$ is also optimal to (3). □

Appendix C. Proof of Proposition 2

Proof. For simplicity, we use $\bar{\mathbf{x}}$ instead of $\bar{\mathbf{x}}_\rho$ for an optimal solution of (8) with some ρ. First, we proof that $\|\bar{\mathbf{x}}\|_0\le s$. If $\|\bar{\mathbf{x}}\|_0>s$, which means that $\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\|_2>0$. We construct a vector $\tilde{\mathbf{x}}$ as $\tilde{\mathbf{x}}=\bar{\mathbf{x}}+\bar{\mathbf{x}}^s-\bar{\mathbf{x}}^{(s+1)}$, easily we have that $\tilde{\mathbf{x}}^s=\bar{\mathbf{x}}^s$. When $\rho>\beta/\eta$, we have

$$\begin{aligned}F(\bar{\mathbf{x}})-F(\tilde{\mathbf{x}})&=\phi(\bar{\mathbf{x}})+\rho(R(\bar{\mathbf{x}})-R(\bar{\mathbf{x}}^s))-\phi(\tilde{\mathbf{x}})-\rho(R(\tilde{\mathbf{x}})-R(\tilde{\mathbf{x}}^s))\\&=\phi(\bar{\mathbf{x}})-\phi(\tilde{\mathbf{x}})+\rho(R(\bar{\mathbf{x}})-R(\tilde{\mathbf{x}}))\\&\ge-\beta\|\bar{\mathbf{x}}-\tilde{\mathbf{x}}\|_2+\rho\eta\|\bar{\mathbf{x}}-\tilde{\mathbf{x}}\|_2\\&=(\rho\eta-\beta)\left\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\|_2>0\end{aligned} \tag{A.14}$$

This contradicts the optimality of $\bar{\mathbf{x}}$. Then we have that $\|\bar{\mathbf{x}}\|_0$ satisfies the s-sparse constraint of (3). Let $\hat{\mathbf{x}}$ be an optimal solution of (3), then we have

$$\begin{aligned}\phi(\bar{\mathbf{x}})-\phi(\hat{\mathbf{x}})&=F(\bar{\mathbf{x}})-\rho P(\bar{\mathbf{x}})-F(\hat{\mathbf{x}})+\rho P(\hat{\mathbf{x}})\\&=F(\bar{\mathbf{x}})-F(\hat{\mathbf{x}})\le0\end{aligned} \tag{A.15}$$

The inequality comes from that $\bar{\mathbf{x}}$ is the optimal solution of (8). This means that $\bar{\mathbf{x}}$ is also optimal to (3). □

Appendix D. Proof of Proposition 3

Proof. Similar to the previous proof of Proposition 2, we use $\bar{\mathbf{x}}$ instead of $\bar{\mathbf{x}}_\rho$ for an optimal solution of (8) with some ρ. Assume by contradiction that $\|\bar{\mathbf{x}}\|_0>s$, which means that $\|\mathbf{x}^{s+1}-\mathbf{x}^s\|_2>0$. By constructing $\tilde{\mathbf{x}}=\bar{\mathbf{x}}+\bar{\mathbf{x}}^s-\bar{\mathbf{x}}^{(s+1)}$, we have

$$\begin{aligned}F(\bar{\mathbf{x}})-F(\tilde{\mathbf{x}})&=\phi(\bar{\mathbf{x}})+\rho(R(\bar{\mathbf{x}})-R(\bar{\mathbf{x}}^s))-\phi(\tilde{\mathbf{x}})-\rho(R(\tilde{\mathbf{x}})-R(\tilde{\mathbf{x}}^s))\\&=\phi(\bar{\mathbf{x}})-\phi(\tilde{\mathbf{x}})+\rho(R(\bar{\mathbf{x}})-R(\tilde{\mathbf{x}}))\\&\ge\left\langle\nabla\phi(\bar{\mathbf{x}}),\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\rangle-\tfrac{L}{2}\left\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\|_2^2\\&\quad+\rho\eta\left\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\|_2\\&\ge\left\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\|_2\left(\rho\eta-\|\nabla\phi(\bar{\mathbf{x}})\|_2-\tfrac{LC}{2\sqrt{s+1}}\right)\\&\ge\left\|\bar{\mathbf{x}}^{(s+1)}-\bar{\mathbf{x}}^s\right\|_2\left(\rho\eta-\|\nabla\phi(\mathbf{0})\|_2-\left(1+\tfrac{1}{2\sqrt{s+1}}\right)LC\right)\\&>0\end{aligned} \tag{A.16}$$

The first inequality uses Assumption 1 that

$$\phi(\mathbf{y})\le\phi(\mathbf{x})+\langle\nabla\phi(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2,\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^N \tag{A.17}$$

The third inequality follows from that

$$\begin{aligned}\|\nabla\phi(\bar{\mathbf{x}})\|_2&\le\|\nabla\phi(\mathbf{0})\|_2+\|\nabla\phi(\bar{\mathbf{x}})-\nabla\phi(\mathbf{0})\|_2\\&\le\|\nabla\phi(\mathbf{0})\|_2+LC\end{aligned} \tag{A.18}$$

(A.16) contradicts the optimality of $\tilde{\mathbf{x}}$, then we have that $\|\tilde{\mathbf{x}}\|_0$ satisfies the s-sparse constraint of (3). Then we can prove that $\tilde{\mathbf{x}}$ is also optimal to (3) similar as the previous proof of Proposition 2. □

Appendix E. Proof of Proposition 6

Proof. Suppose that $\mathbf{x}^*$ is the optimal solution of (12). First, we prove that if $|y_i|>|y_j|$.we have $|x_i^*|\geq|x_j^*|$; otherwise $|x_i^*|<|x_j^*|$, then we construct $\tilde{\mathbf{x}}\in\mathbb{R}^N$ as $\tilde{x}_i^*=\text{sign}(y_i)|x_j^*|$ and $\tilde{x}_j^*=\text{sign}(y_j)|x_i^*|$. Whether $i,j\in\Gamma_{\mathbf{y}}^s$ or $i,j\notin\Gamma_{\mathbf{y}}^s$ or $i\in\Gamma_{\mathbf{y}}^s, j\notin\Gamma_{\mathbf{y}}^s$, we always have that $R(\tilde{\mathbf{x}})=R(\mathbf{x}^*)$ and $R(\tilde{\mathbf{x}}^s)=R(\mathbf{x}^{*s})$. As $\|\tilde{\mathbf{x}}-\mathbf{y}\|_2^2<\|\mathbf{x}^*-\mathbf{y}\|_2^2$, then we can obtain $E(\tilde{\mathbf{x}})<E(\mathbf{x}^*)$. However, this contradicts the optimality of $\mathbf{x}^*$.

Next, we prove that $|x^*_{\pi_y(s+1)}|\leq|y_{\pi_y(s)}|$. To prove this, we need to prove that $|x^*_{\pi_y(j)}|\leq|y_{\pi_y(s)}|$ for all $j\in\{s+1,s+2,\ldots,N\}$. We can do this one by one, i.e., we look at $x^*_{\pi_y(N)}$ first. Easily, we have $|x^*_{\pi_y(N)}|\leq|y_{\pi_y(s)}|$; otherwise we construct $\tilde{x}_{\pi_y(N)}=\text{sign}(y_{\pi_y(N)})|y_{\pi_y(s)}|$, as r_i is strictly increasing on $\mathbb{R}_+$ and symmetrical, thus we have the contradiction $E(\tilde{\mathbf{x}})<E(\mathbf{x}^*)$, then we have $|x^*_{\pi_y(N)}|\leq|y_{\pi_y(s)}|$. By using this deduction, we can prove that $|x^*_{\pi_y(N-1)}|\leq|y_{\pi_y(s)}|$ in a similar way. At last, we have $|x^*_{\pi_y(s+1)}|\leq|y_{\pi_y(s)}|$.

Rewrite $E(\mathbf{x})$ as

$$E(\mathbf{x})=\sum_{j=1}^{s}\tfrac{1}{2\lambda}\left(x_{\pi_y(j)}-y_{\pi_y(j)}\right)^2+\sum_{j=s+1}^{N}\left(\tfrac{1}{2\lambda}\left(x_{\pi_y(j)}-y_{\pi_y(j)}\right)^2+r_{\pi_y(j)}\left(x_{\pi_y(j)}\right)\right) \tag{A.19}$$

As $|x^*_{\pi_y(s+1)}|\leq|y_{\pi_y(s)}|$, we have that $x^*_{\pi_y(j)}=y_{\pi_y(j)}$, $j=1,2,\ldots,s$ and $x^*_{\pi_y(j)}=\text{prox}_{\lambda r_{\pi_y(j)}}(y_{\pi_y(j)})$, $j=s+1,s+2,\ldots,N$. Moreover, if each r_i is convex, we have that $x^*_{\pi_y(j)}=(1+\lambda\partial r_{\pi_y(j)})^{-1}(y_{\pi_y(j)})$ for $j=s+1,s+2,\ldots,N$. This completes the proof. □

Appendix F. Proof of Proposition 7

Proof. First, we show that when $R(\mathbf{x})=\|\mathbf{x}\|_2$, we also have if $|y_i|>|y_j|$. We have $|x_i^*|\geq|x_j^*|$. Otherwise, we can always construct a $\tilde{\mathbf{x}}\in\mathbb{R}^N$, which swaps the absolute value of x_i^* and x_j^* as the same way in the proof of Proposition 6, then we can obtain a smaller objective value. As proved in Proposition 4, $\mathbf{x}^*=\mathbf{0}$ if and only if $\mathbf{y}=\mathbf{0}$. Then, we only need to consider the case $\mathbf{y}\neq\mathbf{0}$.

(1) If $|y_{\pi_y(s)}|\neq|y_{\pi_y(s+1)}|$, then we have

$$\{\pi_x(1),\pi_x(2),\ldots,\pi_x(s)\}=\{\pi_y(1),\pi_y(2),\ldots,\pi_y(s)\} \tag{A.20}$$

Easily, we have that if $y_{\pi_y(s+1)}=0$, then $\mathbf{x}^*=\mathbf{y}$ and $E(\mathbf{x}^*)=0$.

When $y_{\pi_y(s+1)}\neq0$, the first-order optimality condition of minimizing $E(\mathbf{x})$ is that

$$\begin{cases}\left(1+\frac{\lambda}{\|\mathbf{x}\|_2}-\frac{\lambda}{\|\mathbf{x}^s\|_2}\right)x_i=y_i, & i\in\Gamma_{\mathbf{y}}^s\\ \left(1+\frac{\lambda}{\|\mathbf{x}\|_2}\right)x_i=y_i, & i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s\end{cases} \tag{A.21}$$

By using Proposition 5, we have that $1+\frac{\lambda}{\|\mathbf{x}\|_2}-\frac{\lambda}{\|\mathbf{x}^s\|_2}\geq0$ in (A.21). Using (A.21), we have

$$\begin{cases}\left(1+\frac{\lambda}{\|\mathbf{x}\|_2}\right)\|\mathbf{x}^s\|_2=\|\mathbf{y}^s\|_2+\lambda\\ \|\mathbf{x}\|_2=\frac{\lambda\|\mathbf{x}-\mathbf{x}^s\|_2}{\|\mathbf{y}-\mathbf{y}^s\|_2-\|\mathbf{x}-\mathbf{x}^s\|_2}\end{cases} \tag{A.22}$$

Substitute one equation of (A.22) into another, we have

$$\|\mathbf{x}^s\|_2=\frac{\|\mathbf{y}^s\|_2+\lambda}{\|\mathbf{y}-\mathbf{y}^s\|_2}\|\mathbf{x}-\mathbf{x}^s\|_2 \tag{A.23}$$

By using the equation $\|\mathbf{x}\|_2=\sqrt{\|\mathbf{x}^s\|_2^2+\|\mathbf{x}-\mathbf{x}^s\|_2^2}$, we have

$$\|\mathbf{x}\|_2=\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}-\lambda \tag{A.24}$$

$$\|\mathbf{x}^s\|_2=\left(\|\mathbf{y}^s\|_2+\lambda\right)\frac{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}-\lambda}{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}} \tag{A.25}$$

$$\|\mathbf{x}-\mathbf{x}^s\|_2=\|\mathbf{y}-\mathbf{y}^s\|_2\frac{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}-\lambda}{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}} \tag{A.26}$$

Substitute these into (A.21), then we have

$$x_i^*=\begin{cases}\frac{\left(\|\mathbf{y}^s\|_2+\lambda\right)\left(\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}-\lambda\right)}{\|\mathbf{y}^s\|_2\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}}y_i, & i\in\Gamma_{\mathbf{y}}^s\\ \frac{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}-\lambda}{\sqrt{\|\mathbf{y}-\mathbf{y}^s\|_2^2+\left(\|\mathbf{y}^s\|_2+\lambda\right)^2}}y_i, & i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s\end{cases} \tag{A.27}$$

(2) If $|y_{\pi_y(s)}|=|y_{\pi_y(s+1)}|$, then we have a similar conclusion as Remark 6.

From the above deduction, we have the expression of $\mathbf{x}^*$ in (21) and (22) when $R(\mathbf{x})=\|\mathbf{x}\|_2$. This completes the proof.

□

Appendix G. Proof of Proposition 8

Proof. Similar to the previous proof of Proposition 6, we have that

$$|x_i^*|\geq\left|x_j^*\right| \quad \text{if} \quad |y_i|>\left|y_j\right| \tag{A.28}$$

As proved in Proposition 4, $\mathbf{x}^*=\mathbf{0}$ if and only if $\mathbf{y}=\mathbf{0}$. Then, we just consider the condition of $\mathbf{y}\neq\mathbf{0}$. Firstly, we suppose that $|y_{\pi_y(s)}|\neq|y_{\pi_y(s+1)}|$, then we have $\{\pi_x(1),\pi_x(2),\ldots,\pi_x(s)\}=\{\pi_y(1),\pi_y(2),\ldots,\pi_y(s)\}$.

The first-order optimality condition of minimizing $E(\mathbf{x})$ is that

$$\left(1-\frac{a\lambda}{\|\mathbf{x}\|_2}+\frac{a\lambda}{\|\mathbf{x}^s\|_2}\right)x_i=y_i, i\in\Gamma_{\mathbf{y}}^s \tag{A.29}$$

$$\left(1-\frac{a\lambda}{\|\mathbf{x}\|_2}\right)x_i=y_i-\lambda q_i, i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s \tag{A.30}$$

where $\mathbf{q}\in\partial\|\mathbf{x}-\mathbf{x}^s\|_1$ is a subgradient.

(1) First case, when $|y_{\pi_y(s+1)}|>\lambda$. Easily we have $1-\frac{a\lambda}{\|\mathbf{x}^*\|_2}>0$ by using Proposition 5: $x_i^*\begin{cases}\geq0, & \text{if } y_i>0\\ \leq0, & \text{if } y_i<0\end{cases}$. When $y_{\pi_y(s+1)}>\lambda$, then $y_{\pi_y(s+1)}-\lambda q>0$, so we have $1-\frac{a\lambda}{\|\mathbf{x}^*\|_2}>0$; when $y_{\pi_y(s+1)}<-\lambda$, then $y_{\pi_y(s+1)}-\lambda q<0$, and we also have $1-\frac{a\lambda}{\|\mathbf{x}^*\|_2}>0$.

For $i\in\Gamma_{\mathbf{y}}^N\backslash\Gamma_{\mathbf{y}}^s$, if $|y_i|\leq\lambda$, then $x_i^*=0$. Otherwise, for this i, if $0<y_i\leq\lambda$, then $x_i^*>0$ based on Proposition 5. As $1-\frac{a\lambda}{\|\mathbf{x}^*\|_2}>0$, the left-hand side (LHS) of (A.30) is positive, while the right-hand side (RHS) of (A.30) nonpositive; if $-\lambda\leq y_i<0$, then $x_i^*<0$ based

on Proposition 5. The LHS of (A.30) is negative, while the RHS of (A.30) is nonnegative; if $y_i = 0$, we have $x_i^* = 0$ based on (A.28).

For $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$, if any $|y_i| > \lambda$, then we have $x_i^* \neq 0$ based on (A.30). For this i, we construct a vector $\mathbf{z} \in \mathbb{R}^N$ as

$$z_i = \begin{cases} \text{shrink}(y_i, \lambda), & i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s \\ y_{\pi_y(1)}, & i \in \Gamma_{\mathbf{y}}^s \end{cases} \tag{A.31}$$

For $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$, we have $(1 - \frac{a\lambda}{\|\mathbf{x}\|_2})x_i = z_i$, then we can obtain

$$\left(1 - \frac{a\lambda}{\|\mathbf{x}\|_2}\right)\|\mathbf{x} - \mathbf{x}^s\|_2 = \|\mathbf{z} - \mathbf{z}^s\|_2 \tag{A.32}$$

For $i \in \Gamma_{\mathbf{y}}^s$, we have

$$\left(1 - \frac{a\lambda}{\|\mathbf{x}\|_2} + \frac{a\lambda}{\|\mathbf{x}^s\|_2}\right)\|\mathbf{x}^s\|_2 = \|\mathbf{y}^s\|_2 \tag{A.33}$$

Substitute (A.32) into (A.33), we have

$$\|\mathbf{x}^s\|_2 = \frac{\|\mathbf{y}^s\|_2 - a\lambda}{\|\mathbf{z} - \mathbf{z}^s\|_2}\|\mathbf{x} - \mathbf{x}^s\|_2 \tag{A.34}$$

By using the equation $\|\mathbf{x}\|_2 = \sqrt{\|\mathbf{x}^s\|_2^2 + \|\mathbf{x} - \mathbf{x}^s\|_2^2}$, we have

$$\|\mathbf{x} - \mathbf{x}^s\|_2 = \|\mathbf{z} - \mathbf{z}^s\|_2 + \frac{a\lambda\|\mathbf{z} - \mathbf{z}^s\|_2}{\sqrt{\|\mathbf{z} - \mathbf{z}^s\|_2^2 + (\|\mathbf{y}^s\|_2 - a\lambda)^2}} \tag{A.35}$$

$$\|\mathbf{x}^s\|_2 = (\|\mathbf{y}^s\|_2 - a\lambda)\left(1 + \frac{a\lambda}{\sqrt{\|\mathbf{z} - \mathbf{z}^s\|_2^2 + (\|\mathbf{y}^s\|_2 - a\lambda)^2}}\right) \tag{A.36}$$

$$\|\mathbf{x}\|_2 = \|\mathbf{z} - \mathbf{z}^s\|_2\sqrt{1 + \frac{(\|\mathbf{y}^s\|_2 - a\lambda)^2}{\|\mathbf{z} - \mathbf{z}^s\|_2^2}} + a\lambda \tag{A.37}$$

Substitute these into (A.29) and (A.30), then we have: for $i \in \Gamma_{\mathbf{y}}^s$,

$$x_i^* = \frac{\|\mathbf{y}^s\|_2 - a\lambda}{\|\mathbf{y}^s\|_2}\left(1 + \frac{a\lambda}{\sqrt{\|\mathbf{z} - \mathbf{z}^s\|_2^2 + (\|\mathbf{y}^s\|_2 - a\lambda)^2}}\right)y_i \tag{A.38}$$

for $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$,

$$x_i^* = \left(1 + \frac{a\lambda}{\sqrt{\|\mathbf{z} - \mathbf{z}^s\|_2^2 + (\|\mathbf{y}^s\|_2 - a\lambda)^2}}\right)z_i \tag{A.39}$$

(2) Second case, if $|y_{\pi_y(s+1)}| = \lambda$, for $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$, suppose that there are k components of y_i having the same amplitude of λ, i.e., $|y_{\pi_y(s+1)}| = \cdots = |y_{\pi_y(s+k)}| = \lambda > |y_{\pi_y(s+k+1)}|$.

For $i \in \{\pi_y(s+k+1), \pi_y(s+k+2), \ldots, \pi_y(N)\}$, we have $x_i^* = 0$. Otherwise, for this i, if $0 < y_i < \lambda$, then $x_i^* > 0$ based on Proposition 5. Easily, we have $y_i - \lambda q_i < 0$ and $1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2} < 0$ from (A.30). Meanwhile, as $|y_{\pi_y(s+1)}| = \lambda$, we have $|x^*_{\pi_y(s+1)}| \geq |x_i^*| > 0$, then $y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)} = 0$, and this contradicts to the equation $(1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2})x^*_{\pi_y(s+1)} = y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)}$ in (A.30). If $-\lambda < y_i < 0$, then $x_i^* < 0$ based on Proposition 5, we have $y_i - \lambda q_i > 0$ and $1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2} < 0$ from (A.30). However, as $y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)} = 0$, this also contradicts to the equation $(1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2})x^*_{\pi_y(s+1)} = y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)}$. If $y_i = 0$, we have $x_i^* = 0$ based on (A.28). Then we obtain that $x_i^* = 0$ for $i \in \{\pi_y(s+k+1), \pi_y(s+k+2), \ldots, \pi_y(N)\}$.

For $i \in \{\pi_y(s+1), \pi_y(s+2), \ldots, \pi_y(s+k)\}$, if there exits $x_i^* \neq 0$, for this i we have $y_i - \lambda q_i = 0$, then we obtain $1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2} = 0$ and $\|\mathbf{x}^*\|_2 = a\lambda$. Substitute this into (A.29), we have $\|\mathbf{y}^s\|_2 = a\lambda$. As $|y_{\pi_y(s+1)}| = \lambda$, then we have that there exits $x_i^* \neq 0$ if and only if the conditions of $a = 1$, $s = 1$, $|y_{\pi_y(1)}| = \lambda$ and $\|\mathbf{x}^*\|_2 = \lambda$ are all satisfied. In this case, there are infinite many solutions, and all these $\mathbf{x}^*$ should satisfy $\|\mathbf{x}^*\|_2 = \lambda$, $x_i^* y_i \geq 0$ and $x_i^* = 0$ when $i \in \{\pi_y(k+2), \pi_y(k+3), \ldots, \pi_y(N)\}$. For example,

$$x_i^* = \begin{cases} \text{sign}\left(y_{\pi_y(1)}\right)\lambda, & i = \pi_y(1) \\ 0, & i \in \{\pi_y(2), \pi_y(3), \ldots, \pi_y(N)\} \end{cases} \tag{A.40}$$

or

$$x_i^* = \begin{cases} \frac{\text{sign}\left(y_{\pi_y(i)}\right)\lambda}{(k+1)}, & i \in \{\pi_y(1), \pi_y(2), \ldots, \pi_y(k+1)\} \\ 0, & i \in \{\pi_y(k+2), \pi_y(k+3), \ldots, \pi_y(N)\} \end{cases} \tag{A.41}$$

When any of these conditions of $a = 1$, $s = 1$, $|y_{\pi_y(1)}| = \lambda$ cannot be satisfied, we have $x_i^* = 0$ for $i \in \{\pi_y(s+1), \pi_y(s+2), \ldots, \pi_y(s+k)\}$. Then we have $\mathbf{x}^* = \mathbf{x}^{*s}$. Substitute this into (A.29), we have $x_i^* = y_i$ for $i \in \Gamma_{\mathbf{y}}^s$. Then the solution $\mathbf{x}^*$ can be expressed as

$$x_i^* = \begin{cases} y_i, & i \in \Gamma_{\mathbf{y}}^s \\ 0, & i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s \end{cases} \tag{A.42}$$

(3) Third case, if $0 < |y_{\pi_y(s+1)}| < \lambda$, for $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$, suppose that there are k components of y_i having the same amplitude with $y_{\pi_y(s+1)}$, i.e., $|y_{\pi_y(s+1)}| = \cdots = |y_{\pi_y(s+k)}| > |y_{\pi_y(s+k+1)}|$.

For $i \in \{\pi_y(s+k+1), \pi_y(s+k+2), \ldots, \pi_y(N)\}$, we have $x_i^* = 0$. Otherwise, for this i, as $|y_{\pi_y(s+1)}| > |y_i|$, we have $|x^*_{\pi_y(s+1)}| \geq |x_i^*| > 0$ based on (A.28). Then we obtain $1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2} < 0$ from (A.30), and we have $(1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2})|x^*_{\pi_y(s+1)}| \leq (1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2})|x_i^*|$, which means that $|y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)}| \geq |y_i - \lambda q_i|$ through (A.30). Since $|x^*_{\pi_y(s+1)}| \geq |x_i^*| \neq 0$, then we can obtain $q_{\pi_y(s+1)} = \text{sign}(y_{\pi_y(s+1)})$ based on Proposition 5. If $y_i \neq 0$, then we have $q_i = \text{sign}(y_i)$ and obtain $|\text{sign}(y_{\pi_y(s+1)})| \cdot ||y_{\pi_y(s+1)}| - \lambda| \geq |\text{sign}(y_i)| \cdot ||y_i| - \lambda|$, which means that $\lambda - |y_{\pi_y(s+1)}| \geq \lambda - |y_i|$. However, this contradicts $|y_{\pi_y(s+1)}| > |y_i|$. If $y_i = 0$, we have $|y_{\pi_y(s+1)} - \lambda q_{\pi_y(s+1)}| \geq \lambda$, then we can obtain $||y_{\pi_y(s+1)}| - \lambda| \geq \lambda$, which contradicts $0 < |y_{\pi_y(s+1)}| < \lambda$. Then, we obtain that $x_i^* = 0$ for $i \in \{\pi_y(s+k+1), \pi_y(s+k+2), \ldots, \pi_y(N)\}$.

For $i \in \{\pi_y(s+1), \pi_y(s+2), \ldots, \pi_y(s+k)\}$, if there exits $x_i^* \neq 0$, then we have $1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2} < 0$ as the signs of $y_i - \lambda q_i$ and x_i^* are opposite. For $i \in \Gamma_{\mathbf{y}}^s$, from (A.29), we have

$$\|\mathbf{x}^{*s}\|_2 = (\|\mathbf{y}^s\|_2 - a\lambda)\left(1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2}\right) \tag{A.43}$$

If $\|\mathbf{y}^s\|_2 \geq a\lambda$, we have $\|\mathbf{x}^{*s}\|_2 \leq 0$, which contradicts $x_i^* \neq 0$. So, when $\|\mathbf{y}^s\|_2 \geq a\lambda$, we have $x_i^* = 0$, and then the solution $\mathbf{x}^*$ is

$$x_i^* = \begin{cases} y_i, & i \in \Gamma_{\mathbf{y}}^s \\ 0, & i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s \end{cases} \tag{A.44}$$

If $\|\mathbf{y}^s\|_2 < a\lambda$, for $i \in \{\pi_y(s+1), \pi_y(s+2), \ldots, \pi_y(s+k)\}$, suppose there are c components of $x_i^* \neq 0$ and $c \leq k$. From (A.30), we have $\|\mathbf{x}^* - \mathbf{x}^{*s}\|_2 = \sqrt{c}(|y_{\pi(s+1)}| - \lambda)/(1 - \frac{a\lambda}{\|\mathbf{x}^*\|_2})$. Substitute this into $\|\mathbf{x}^{*s}\|_2 = (\|\mathbf{y}^s\|_2 - a\lambda)/(1 - \frac{a\lambda}{\|\mathbf{x}\|_2})$ from (A.29), we have

$$\|\mathbf{x}^*\|_2 = a\lambda - \sqrt{(\|\mathbf{y}^s\|_2 - a\lambda)^2 + c\left(\left|y_{\pi(s+1)}\right| - \lambda\right)^2} \tag{A.45}$$

Reconsidering the expression of $E(\mathbf{x})$, and using the first-order optimality condition, we have

$$\begin{aligned}
E(\mathbf{x}^*) &= \tfrac{\|\mathbf{x}^*\|_2^2+\|\mathbf{y}\|_2^2}{2\lambda} - \left\langle \mathbf{x}^*, \tfrac{\mathbf{y}}{\lambda}\right\rangle \\
&\quad + \|\mathbf{x}^*\|_1 - a\|\mathbf{x}^*\|_2 - \|\mathbf{x}^{*S}\|_1 + a\|\mathbf{x}^{*S}\|_2 \\
&= \tfrac{\|\mathbf{x}^*\|_2^2+\|\mathbf{y}\|_2^2}{2\lambda} - \left\langle \mathbf{x}^{*S}, \left(\tfrac{1}{\lambda} - \tfrac{a}{\|\mathbf{x}^*\|_2} + \tfrac{a}{\|\mathbf{x}^{*S}\|_2}\right)\mathbf{x}^{*S}\right\rangle \\
&\quad - \left\langle \mathbf{x}^* - \mathbf{x}^{*S}, q + \left(\tfrac{1}{\lambda} - \tfrac{a}{\|\mathbf{x}^*\|_2}\right)\left(\mathbf{x}^* - \mathbf{x}^{*S}\right)\right\rangle \\
&\quad + \|\mathbf{x}^*\|_1 - a\|\mathbf{x}^*\|_2 - \|\mathbf{x}^{*S}\|_1 + a\|\mathbf{x}^{*S}\|_2 \\
&= \tfrac{\|\mathbf{x}^*\|_2^2+\|\mathbf{y}\|_2^2}{2\lambda} - \tfrac{\|\mathbf{x}^{*S}\|_2^2}{\lambda} + \tfrac{a\|\mathbf{x}^{*S}\|_2^2}{\|\mathbf{x}^*\|_2} - a\|\mathbf{x}^{*S}\|_2 \\
&\quad - \|\mathbf{x}^* - \mathbf{x}^{*S}\|_1 - \tfrac{\|\mathbf{x}^*-\mathbf{x}^{*S}\|_2^2}{\lambda} + \tfrac{a\|\mathbf{x}^*-\mathbf{x}^{*S}\|_2^2}{\|\mathbf{x}^*\|_2} \\
&\quad + \|\mathbf{x}^*\|_1 - a\|\mathbf{x}^*\|_2 - \|\mathbf{x}^{*S}\|_1 + a\|\mathbf{x}^{*S}\|_2 \\
&= -\tfrac{\|\mathbf{x}^*\|_2^2}{2\lambda} + \tfrac{\|\mathbf{y}\|_2^2}{2\lambda}
\end{aligned} \tag{A.46}$$

Then we have $E(\mathbf{x}^*) < E(\mathbf{0})$, and we need to find the $\mathbf{x}^*$ with the largest norm among all $\mathbf{x}^*$ that satisfying (A.29) and (A.30). From this, we have that c should be zero to make the largest $\|\mathbf{x}^*\|$ in (A.45). So, when $\|\mathbf{y}^s\|_2 < a\lambda$, we have the solution $\mathbf{x}^*$ the same as (A.44).

(4) Fourth case, if $y_{\pi(s+1)} = 0$, for $i \in \Gamma_{\mathbf{y}}^N \backslash \Gamma_{\mathbf{y}}^s$, we have $x_i^* = 0$. Otherwise, we can construct a vector $\tilde{\mathbf{x}} \in \mathbb{R}^N$, which is equal to $\mathbf{x}^*$ except setting these corresponding $\tilde{x}_i$ to be zero. Then we can obtain a smaller objective value, which contradicts the optimality of $\mathbf{x}^*$. For $i \in \Gamma_{\mathbf{y}}^s$, we have $x_i^* = y_i$. Then the solution $\mathbf{x}^*$ can be expressed as (A.44).

Once again, if there exits one or more components of y_i, $i \notin \Gamma_{\mathbf{y}}^s$ having the same amplitude of $y_{\pi_y(s)}$, then we have a similar conclusion as Remark 6. This completes the proof.

□

Appendix H. Proof of Proposition 9

Proof. Let $\mathbf{x}^{[k+1]}$ be the optimal solution of (13) with $\mathbf{y} = \mathbf{x}^{[k]} - \beta\nabla\phi(\mathbf{x}^{[k]})$, then we have

$$\begin{aligned}
E\left(\mathbf{x}^{[k+1]}\right) - E\left(\mathbf{x}^{[k]}\right) &= \tfrac{\left\|\mathbf{x}^{[k+1]}-\mathbf{y}\right\|_2^2}{2\lambda} + R\left(\mathbf{x}^{[k+1]}\right) - R\left(\mathbf{x}^{[k+1]^S}\right) \\
&\quad - \tfrac{\left\|\mathbf{x}^{[k]}-\mathbf{y}\right\|_2^2}{2\lambda} - R\left(\mathbf{x}^{[k]}\right) + R\left(\mathbf{x}^{[k]^S}\right) \\
&= -\tfrac{\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2}{2\lambda} + \tfrac{\left\langle \mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}, \mathbf{x}^{[k+1]}-\mathbf{y}\right\rangle}{\lambda} \\
&\quad + R\left(\mathbf{x}^{[k+1]}\right) - R\left(\mathbf{x}^{[k+1]^S}\right) - R\left(\mathbf{x}^{[k]}\right) + R\left(\mathbf{x}^{[k]^S}\right) \\
&= -\tfrac{\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2}{2\lambda} + \sum_{i\in\Lambda_{k+1}} \left(x_i^{[k]} - x_i^{[k+1]}\right)\left(\partial r_i\left(x_i^{[k+1]}\right)\right) \\
&\quad + \sum_{i\in\Lambda_{k+1}} r_i\left(x_i^{[k+1]}\right) - \sum_{i\in\Lambda_k} r_i\left(x_i^{[k]}\right) \\
&\le -\tfrac{\left\|\mathbf{x}^{[k+1]}-\mathbf{x}^{[k]}\right\|_2^2}{2\lambda} + \sum_{i\in\Lambda_{k+1}} r_i\left(x_i^{[k]}\right) - \sum_{i\in\Lambda_k} r_i\left(x_i^{[k]}\right)
\end{aligned} \tag{A.47}$$

The third equation comes from Proposition 6, and the last inequality is based on the property of subgradient. Then we have

$$E\left(\mathbf{x}^{[k+1]}\right) - E\left(\mathbf{x}^{[k]}\right) \le \min\left\{-\frac{\left\|\mathbf{x}^{[k+1]} - \mathbf{x}^{[k]}\right\|_2^2}{2\lambda} + \Delta_k, 0\right\} \tag{A.48}$$

where $\Delta_k = \sum_{i\in\Lambda_{k+1}} r_i(x_i^{[k]}) - \sum_{i\in\Lambda_k} r_i(x_i^{[k]})$, $\Lambda_{k+1} = \Gamma_{\mathbf{x}^{[k+1]}}^N \backslash \Gamma_{\mathbf{x}^{[k+1]}}^s$, and $\Lambda_k = \Gamma_{\mathbf{x}^{[k]}}^N \backslash \Gamma_{\mathbf{x}^{[k]}}^s$. Substituting this into (26), then we have (32). This completes the proof. □

References

[1] E.J. Candes, The restricted isometry property and its implications for compressed sensing, Comptes rendus - Mathematique 346 (9–10) (2008) 589–592.
[2] V.M. Patel, G.R. Easley, D.M. HealyJr, et al., Compressed synthetic aperture radar, IEEE J. Sel. Top. Signal Process. 4 (2) (2010) 244–254.
[3] J. Yang, J. Thompson, X. Huang, et al., Random-frequency SAR imaging based on compressed sensing, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 983–994.
[4] C.R. Berger, Z. Wang, J. Huang, et al., Application of compressive sensing to sparse channel estimation, IEEE Commun. Mag. 48 (11) (2010) 164–174.
[5] Z. Chen, X. Jin, L. Li, et al., A limited-angle CT reconstruction method based on anisotropic TV minimization, Phys. Med. Biol. 58 (7) (2013) 2119.
[6] M. Lustig, D. Donoho, J.M. Pauly, Sparse MRI: the application of compressed sensing for rapid MR imaging, Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med. 58 (6) (2007) 1182–1195.
[7] R. Chartrand, V. Staneva, Restricted isometry properties and nonconvex compressive sensing, Inverse. Probl. 24 (3) (2008) 035020.
[8] E.J. Candes, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted l1 minimization, J. Fourier Anal. Applica. 14 (5–6) (2008) 877–905.
[9] Y. Sun, J. Tao, Few views image reconstruction using alternating direction method viaâ norm minimization, Int. J. Imag. Syst. Technol. 24 (3) (2014) 215–223.
[10] S. Yu-Li, T. Jin-Xu, Image reconstruction from few views by â 0-norm optimization, Chin. Phys. B 23 (7) (2014) 078703.
[11] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Process. Lett. 14 (10) (2007) 707–710.
[12] R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing//2008 IEEE international conference on acoustics, in: Speech and Signal Processing, IEEE, 2008, pp. 3869–3872.
[13] D. Krishnan, R. Fergus, Fast image deconvolution using hyper-laplacian priors, Adv Neural Inf Process Syst (2009) 1033–1041.
[14] Z. Xu, X. Chang, F. Xu, et al., $l_{1/2}$ Regularization: a thresholding representation theory and a fast solver, IEEE Trans. Neural Netw. Learn. Syst. 23 (7) (2012) 1013–1027.
[15] M.J. Lai, Y. Xu, W. Yin, Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization, SIAM J. Numer. Anal. 51 (2) (2013) 927–957.
[16] J.K. Pant, W.S. Lu, A. Antoniou, New improved algorithms for compressive sensing based on ℓ_p norm, IEEE Trans. Circuits Syst. II Express Briefs 61 (3) (2014) 198–202.
[17] J. Woodworth, R. Chartrand, Compressed sensing recovery via nonconvex shrinkage penalties, Inverse Probl. 32 (7) (2016) 075004.
[18] L. Wu, Z. Sun, D.H. Li, A barzilaiâ borwein-like iterative half thresholding algorithm for the $l_{1/2}$ regularized problem, J. Sci. Comput. 67 (2) (2016) 581–601.
[19] X. Fengmin, W. Shanhe, A hybrid simulated annealing thresholding algorithm for compressed sensing, Signal Process. 93 (6) (2013) 1577–1585.
[20] C. Miao, H. Yu, A general-thresholding solution for $l_p(0 < p < 1)$ regularized CT reconstruction, IEEE Trans. Image Process. 24 (12) (2015) 5455–5468.
[21] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, J. Mach. Learn. Res. 11 (2010) 1081–1107.
[22] T. Zhang, Multi-stage convex relaxation for feature selection, Bernoulli 19 (5B) (2013) 2277–2293.
[23] Y. Lou, P. Yin, J. Xin, Point source super-resolution via non-convex l_1 based methods, J. Sci. Comput. 68 (3) (2016) 1082–1100.
[24] S. Z. Zhang, J. Xin, Minimization of transformed l_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing, Math. Program. 169 (1) (2018) 307–336.
[25] T. Dinh, J. Xin, Convergence of a relaxed variable splitting method for learning sparse neural networks via ℓ_1, ℓ_0, and transformed-ℓ_1 penalties, 2018, arXiv:1812.05719.
[26] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, Ann. Stat. 37 (6A) (2009) 3498–3528.
[27] M. Bogdan, E.V.D. Berg, W. Su, et al., Statistical estimation and testing via the sorted l1 norm, Statistics (2013).
[28] X. Zeng, M.A.T. Figueiredo, Decreasing weighted sorted ℓ_1 regularization, IEEE Signal Process. Lett. 21 (10) (2014) 1240–1244.
[29] Y. Lou, M. Yan, Fast l1â l2 minimization via a proximal operator, J. Sci. Comput. 74 (2) (2018) 767–785.
[30] Y. Lou, P. Yin, Q. He, J. Xin, Computing sparse representation in a highly coherent dictionary based on difference of l1 and l2, J. Sci. Comput. 64 (1) (2015) 178–196.
[31] P. Yin, Y. Lou, Q. He, et al., Minimization of l1-2 for compressed sensing, SIAM J. Sci. Comput. 37 (1) (2015) A536–A563.
[32] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J Am Stat Assoc 96 (456) (2001) 1348–1360.
[33] A. Mehranian, H.S. Rad, A. Rahmim, et al., Smoothly clipped absolute deviation (SCAD) regularization for compressed sensing MRI using an augmented lagrangian scheme, Magn. Reson. Imag. 31 (8) (2013) 1399–1411.
[34] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Stat. 38 (2) (2010) 894–942.
[35] I. Selesnick, Sparse regularization via convex analysis, IEEE Trans. Signal Process. 65 (17) (2017) 4481–4494.
[36] Y. Sun, H. Chen, J. Tao, Sparse signal recovery via minimax-concave penalty and ℓ_1-norm loss function, IET Signal Proc. 12 (9) (2018) 1091–1098.

[37] T. Blumensath, M.E. Davies, Iterative thresholding for sparse approximations, J. Fourier Anal. Appl. 14 (2008) 629–654.
[38] T. Blumensath, M.E. Davies, Iterative hard thresholding for compressed sensing, Appl. Comput. Harmon. Anal. 27 (3) (2009) 265–274.
[39] T. Blumensath, Accelerated iterative hard thresholding, Signal Process. 92 (3) (2012) 752–756.
[40] Z. Lu, Iterative hard thresholding methods for ℓ_0 regularized convex cone programming, Math. Program. 147 (1–2) (2014) 125–154.
[41] C. Bao, B. Dong, L. Hou, et al., Image restoration by minimizing zero norm of wavelet frame coefficients, Inverse Probl. 32 (11) (2016) 115004.
[42] X. Zhang, X. Zhang, An accelerated proximal iterative hard thresholding method for ℓ_0 minimization, 2017, arXiv:1709.01668.
[43] J. Gotoh, A. Takeda, K. Tono, DC Formulations and algorithms for sparse optimization problems, Math. Program. (2018) 1–36.
[44] K. Tono, A. Takeda, J. Gotoh, Efficient DC algorithm for constrained sparse optimization, 2017, arXiv:1701.08498.
[45] P.D. Tao, L.T.H. An, Convex analysis approach to dc programming: theory, algorithms and applications, Acta Mathematica Vietnamica 22 (1) (1997) 289–355.
[46] M. Ahn, J.S. Pang, J. Xin, Difference-of-convex learning: directional stationarity, optimality, and sparsity, SIAM J. Optim. 27 (3) (2017) 1637–1665.
[47] P. Yin, J. Xin, Iterative ℓ_1 minimization for non-convex compressed sensing, Journal of Computational Mathematics 35 (4) (2017) 439–451.
[48] T. Liu, T.K. Pong, A. Takeda, A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems, Math. Program. (2017) 1–29.
[49] A.L. Yuille, A. Rangarajan, The concave-convex procedure, Neural Comput. 15 (4) (2003) 915–936.
[50] F.J.A. Artacho, R.M.T. Fleming, P.T. Vuong, Accelerating the DC algorithm for smooth functions, Math. Program. 169 (1) (2018) 95–118.
[51] B. Wen, X. Chen, T.K. Pong, A proximal difference-of-convex algorithm with extrapolation, Comput. Optim. Appl. 69 (2) (2018) 297–324.
[52] F. Wen, L. Pei, Y. Yang, et al., Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization, IEEE Trans. Comput. Imag. 3 (4) (2017) 566–579.
[53] J. Zhang, C. Zhao, D. Zhao, et al., Image compressive sensing recovery using adaptively learned sparsifying basis via l0 minimization, Signal Process. 103 (2014) 114–126.
[54] P. Gong, C. Zhang, Z. Lu, et al., A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems, Int. Conf. Mach. Learn. (2013) 37–45.
[55] H. Li, Z. Lin, Accelerated proximal gradient methods for nonconvex programming, Adv. Neural Inf. Process. Syst. (2015) 379–387.
[56] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imag. Sci. 2 (1) (2009) 183–202.
[57] B. Dong, Y. Zhang, An efficient algorithm for â 0 minimization in wavelet frame based image restoration, J Sci Comput 54 (2–3) (2013) 350–368.
[58] Y. Zhang, B. Dong, Z. Lu, Â â minimization for wavelet frame based image restoration, Math. Comput. 82 (282) (2013) 995–1015.
[59] X. Zhang, Y. Lu, T. Chan, A novel sparsity reconstruction method from poisson data for 3d bioluminescence tomography, J. Sci. Comput. 50 (3) (2012) 519–535.
[60] J. Trzasko, A. Manduca, E. Borisch, Sparse MRI reconstruction via multiscale l0-continuation, in: 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, IEEE, 2007, pp. 176–180.
[61] J. Trzasko, A. Manduca, Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization, IEEE Trans. Med. Imag. 28 (1) (2009) 106–121.
[62] K. Pavlikov, S. Uryasev, CVAr norm and applications in optimization, Optim. Lett. 8 (7) (2014) 1999–2020.
[63] J. Gotoh, S. Uryasev, Two pairs of families of polyhedral norms versus ℓ_p-norms: proximity and applications in optimization, Math. Program. 156 (1–2) (2016) 391–431.
[64] Y. Sun, H. Chen, J. Tao, et al., Computed tomography image reconstruction from few views via log-norm total variation minimization, Digit. Signal Process. 88 (2019) 172–181.
[65] F. Wen, P. Liu, Y. Liu, et al., Robust sparse recovery in impulsive noise via $\ell_p - \ell_1$ optimization, IEEE Trans. Signal Process. 65 (1) (2017) 105–118.
[66] P.L. Combettes, J.C. Pesquet, Proximal splitting methods in signal processing[m], in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, New York, NY, 2011, pp. 185–212.
[67] Z. Lu, Sequential convex programming methods for a class of structured nonlinear programming, 2012, arXiv:1210.3039.
[68] J. Barzilai, J.M. Borwein, Two-point step size gradient methods, IMA J. Num. Anal. 8 (1) (1988) 141–148, doi:10.1093/imanum/8.1.141.
[69] E.Y. Sidky, R. Chartrand, X. Pan, Image reconstruction from few views by non–convex optimization, in: 2007 IEEE Nuclear Science Symposium Conference Record, volume 5, IEEE, 2007, pp. 3526–3530.
[70] Y. Rahimi, C. Wang, H. Dong, et al., A scale invariant approach for sparse signal recovery, 2018, arXiv:1812.08852.
[71] J. Nocedal, S.J. Wright, Numerical Optimization, 2nd, Springer, Berlin, 2006.
[72] S. Boyd, N. Parikh, E. Chu, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. TrendsÂ®Mach. Learn. 3 (1) (2011) 1–122.
[73] C. Soussen, J. Idier, D. Brie, et al., From bernoulliâ gaussian deconvolution to sparse signal restoration, IEEE Trans. Signal Process. 59 (10) (2011) 4572–4584.
[74] D. Lazzaro, E.L. Piccolomini, F. ZAMA, A nonconvex penalization algorithm with automatic choice of the regularization parameter in sparse imaging, Inverse Probl. (2019). In press https://doi.org/10.1088/1361-6420/ab1c6b.
[75] X. Li, D. Sun, K.C. Toh, A highly efficient semismooth newton augmented lagrangian method for solving lasso problems, SIAM J. Optim. 28 (1) (2018) 433–458.