On Convergence of Iterative Thresholding Algorithms to Approximate Sparse Solution for Composite Nonconvex Optimization

Yaohua Hu[1], Xinlin Hu[2,3] and Xiaoqi Yang[3*]

[1]School of Mathematical Sciences, Shenzhen University, Shenzhen, 518060, P. R. China.
[2]Shenzhen Audencia Business School, WeBank Institute of Fintech, Shenzhen University, Shenzhen, 518060, P. R. China.
[3*]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong.

*Corresponding author(s). E-mail(s): mayangxq@polyu.edu.hk;
Contributing authors: mayhhu@szu.edu.cn;
thompson-xinlin.hu@connect.polyu.hk;

Abstract

This paper aims to find an approximate true sparse solution of an underdetermined linear system. For this purpose, we propose two types of iterative thresholding algorithms with the continuation technique and the truncation technique respectively. We introduce a notion of limited shrinkage thresholding operator and apply it, together with the restricted isometry property, to show that the proposed algorithms converge to an approximate true sparse solution within a tolerance relevant to the noise level and the limited shrinkage magnitude. Applying the obtained results to nonconvex regularization problems with SCAD, MCP and ℓ_p penalty ($0 \leq p \leq 1$) and utilizing the recovery bound theory, we establish the convergence of their proximal gradient algorithms to an approximate global solution of nonconvex regularization problems. The established results include the existing convergence theory for ℓ_1 or ℓ_0 regularization problems for finding a true sparse solution as special cases. Preliminary numerical results show that our proposed algorithms can find approximate true sparse solutions that are much better than stationary solutions that are found by using the standard proximal gradient algorithm.

Keywords: Nonconvex sparse optimization, Iterative thresholding algorithm, Proximal gradient algorithm, Sparse solution, Global solution.

MSC Classification: 65K05 , 90C26 , 49M37

1 Introduction

Consider the following underdetermined linear system of variable x:

$$b = Ax + \varepsilon, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ $(m \ll n)$ is a linear transformation matrix, $b \in \mathbb{R}^m$ is an observation vector with an unknown noise $\varepsilon \in \mathbb{R}^m$. The sparsity of vector $x \in \mathbb{R}^n$ is defined to be the number of nonzero components of x, denoted as the ℓ_0 quasi-norm $\|x\|_0$. Sparse optimization, which has wide applications in compressive sensing, image science, systems biology, and machine learning, aims to find a true s-sparse solution $\bar{x}$ of (1) satisfying $\|\bar{x}\|_0 = s$, where s is a (small) positive integer. To do this, one popular way is to consider the following composite optimization problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda\varphi(x), \tag{2}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a (sparsity promoting) penalty function and $\lambda > 0$ is a regularization parameter.

When $\varphi(x)$ is the convex ℓ_1 penalty, (2) is called the ℓ_1 regularization problem, also named as Lasso in statistics and basis pursuit in compressive sensing. Benefiting from the convexity property, a great deal of attention has been attracted to explore theoretical properties [1–3] and develop numerical algorithms [4–7] for the ℓ_1 regularization problem. However, it has been revealed by extensive theoretical and empirical studies that the ℓ_1 regularization problem endures significant estimation bias when components of the solution have large magnitude; and the solution induced from the ℓ_1 regularization problem may be much less sparse than the true sparse one; see, e.g., [8–11]. Therefore, there is a great demand for developing alternative sparsity promoting techniques that enjoy nice theoretical property and better numerical performance.

Recently, breakthrough developments have been achieved by virtue of nonconvex regularization methods. Popular nonconvex sparsity promoting penalties include the smoothly clipped absolute deviation (SCAD) [9], minimax concave penalty (MCP) [12], and the ℓ_p penalty with $p \in (0, 1)$ [13]. It is worth noting that all these three types of regularization problems belong to the class of multimodal functions, which have multiple (local) optima, see [14]. It has been shown by several studies that the SCAD and MCP can ameliorate the bias of the ℓ_1 penalty [9, 12], and the ℓ_p regularization, in particular when

$p = \frac{1}{2}$, admits a significantly stronger sparsity promoting capability than the ℓ_1 regularization in the sense that it allows to obtain a more sparse solution from fewer linear measurements than that required by the ℓ_1 regularization; see, e.g., [8, 10, 11]. Motivated by these significant advantages, tremendous efforts have been devoted to the development of theoretical properties and optimization algorithms of the nonconvex regularization problem. It was shown in [1, 10, 15–17] that problem (2) with certain sparse penalties obeys the recovery bound between a true solution of (1) and a global solution of (2) under some regularity conditions. A major challenge of the nonconvex regularization problem is the computational issue as it is intractable to find a global solution of a general nonconvex optimization problem. It is worth noting that second-order optimality conditions of a global solution for nonconvex and nonsmooth penalty problems have been studied in [18, 19].

In sparse optimization, the class of iterative thresholding algorithms (ITA) is one of the most popular and practical numerical algorithms with a simple formulation and a low computational complexity for solving the (convex or nonconvex) regularization problems; see [5, 10, 11, 20, 21, 23] and references therein. The idea of the ITA stems from the first-order optimization algorithms for solving the regularization problem (2). For example, the iterative soft (resp., hard, half) thresholding algorithm can be understood as the proximal gradient algorithm (PGA) for solving the ℓ_1 (resp., ℓ_0, $\ell_{1/2}$) regularization problem; see [5, 11, 20] respectively. However, limited by the difficulty of nonconvexity of problem (2), the convergence theory of its PGA is still far from satisfactory: only convergence to a stationary point of problem (2) is established under the framework of the Kurdyka-Łojasiewicz theory [24, 25].

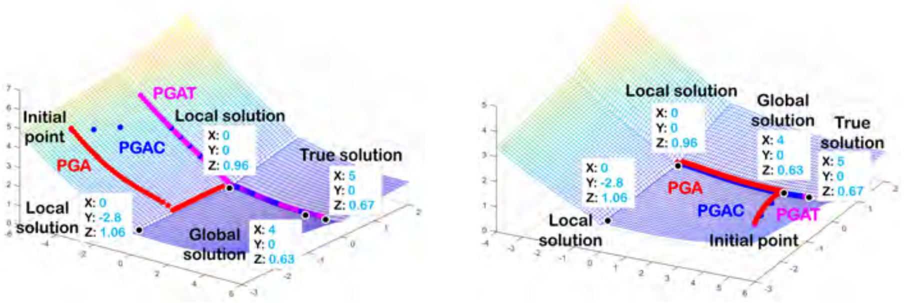
Two acceleration strategies that have been widely applied in the design of ITAs are: the continuation technique and the truncation technique. On one hand, the continuation technique is an easily implemented strategy for speeding up the ITAs by using a decreasing sequence of regularization parameters, which was originally proposed in [6, 26] without theoretical analysis. Xiao and Zhang [7] proved that the iterative soft thresholding algorithm (ISTA) with the continuation technique for the ℓ_1 regularization problem converges to an approximate true sparse solution under the assumption of the restricted isometry property (RIP). Furthermore Jiao et al. [27] showed the convergence of the ITA with the continuation technique for the ℓ_1 and ℓ_0 regularization problems under the assumption of the mutual incoherence property. On the other hand, the truncation technique is widely used to ensure the sparsity structure of the iterates by maintaining its large components and discarding the small ones. One of the most popular algorithms using the truncation technique is the class of iterative hard thresholding algorithms [20, 28, 29], in which their convergence to an approximate true sparse solution was established under the assumption of the RIP. However, to the best of our knowledge, there is still no paper devoted to employing the continuation technique or the truncation technique to accelerate the ITA for solving the nonconvex regularization problems, apart from ℓ_1 and ℓ_0 regularization problems as mentioned above.

Motivated by the results of ℓ_1 and ℓ_0 regularizations in [7, 20, 27], in this paper for a class of nonconvex sparse penalties, we will show that ITAs with the continuation technique or the truncation technique approach an approximate true sparse solution of (1). Furthermore, by virtue of the recovery bounds of problem (2) with SCAD, MCP and ℓ_p penalty in [15, 17], we will show that the ITAs approach an approximate global solution of problem (2). For these purposes, we introduce a notion of limited shrinkage thresholding operator that provides a unified framework of proximal mappings of several nonconvex sparse penalties, including SCAD, MCP, and ℓ_p penalty. Employing the limited shrinkage thresholding operator within the framework of ITA and combining with the continuation technique and the truncation technique, we propose an iterative limited shrinkage thresholding algorithm with continuation (ILSTAC) and the one with truncation (ILSTAT) respectively; see Algorithms 1 and 2. The proposed ILSTAC and ILSTAT are of simple formulation and low storage requirement, and thus extremely efficient for large-scale sparse optimization problems. Under the assumption of the RIP, we show that the output of the ILSTAC approaches an approximate true sparse solution of (1) within a tolerance proportional to the noise level, and that the sequence generated by the ILSTAT converges to an approximate true sparse solution at a geometric rate; see Theorems 1 and 2 respectively. We also obtain the complexities of ILSTAC and ILSTAT.

Since the limited shrinkage thresholding operator includes proximal mappings of SCAD, MCP, and ℓ_p penalty as the special cases, Theorems 1 and 2 are applied to establish the convergence of the PGA with continuation (PGAC) and the PGA with truncation (PGAT) respectively for these nonconvex regularization problems to an approximate true sparse solution of (1), see Theorems 3, 4, 5 and 6 respectively. Moreover, combining these results with the recovery bounds in [15, 17], we present the convergence of the proposed algorithms to an approximate global solution of problem (2) with SCAD, MCP, and ℓ_p penalty, respectively.

We illustrate in Figure 1 the convergence behavior of the PGA, PGAC and PGAT for an $\ell_{1/2}$ regularization problem. In this example, when approaching the local solution $(0, 0)$, a fixed (properly large) regularization parameter in the proximal mapping step in PGA may over-penalize the gradient descent iterate of the least squares at $(0, 0)$, and thus lead PGA to stay at this local solution. In contrast, using a decreasing sequence of regularization parameters in the PGAC is able to avoid over-penalization on variables, and thus help PGAC to escape from the local solution. In this example, the iterates of PGAT will stay on a 1-dimensional subspace after the 1st iterate.

Our preliminary numerical results show that the PGAC and the PGAT have strong sparsity promoting capability and outperform the standard PGA on both accuracy and robustness. In addition, we also compare the numerical performance of PGAs with a branch-and-bound method that was recently proposed in [33].



(a) Initial point $x^0 := (-5, -2)$ (b) Initial point $x^0 := (5, -2)$

Fig. 1: Consider problem (1) and the regularization problem (2) with $A := \begin{bmatrix} -0.2554, & 0.0778 \\ 0.1084, & -0.1811 \end{bmatrix}$, $b := \begin{bmatrix} -1.2770 \\ 0.5420 \end{bmatrix}$, $\varepsilon = 0$, $\varphi(x) = \ell_{1/2}$ penalty and $\lambda := 0.3$. If starting from $(-5, -2)$, the PGA only converges to a local solution $(0, 0)$, and the PGAC and PGAT first go to the neighborhood of the local solution $(0, 0)$, are then able to escape from this local solution, and finally converge to an approximation of the true sparse solution $(5, 0)$, see (a). If starting from $(5, -2)$, PGA and PGAC will have the same behavoir as starting from $(-5, -2)$. However PGAT will directly converge to an approximation of the true sparse solution $(5, 0)$, see (b). Note that this true sparse solution is also an approximate global solution of the $\ell_{1/2}$ regularization problem.

This paper is organized as follows. In Section 2, we present the notations and preliminary results to be used in this paper. In Section 3, we propose the ILSTAC and the ILSTAT and establish their convergence to an approximate true solution of linear system (1) under the assumption of the RIP. Applications to certain nonconvex regularization problems with SCAD, MCP, and ℓ_p penalty are presented in Section 4. Preliminary numerical results and the conclusion are presented in Sections 5 and 6 respectively.

2 Notation and preliminary results

Let $\mathbb{R}^n$ be an n-dimensional Euclidean space with Euclidean norm $\|x\| := \sqrt{\langle x, x\rangle}$. We use the caligraphic letters $\mathcal{S}$, $\mathcal{I}$, $\mathcal{J}$ to denote the index sets, and use $x_{\mathcal{S}}$ and $A_{\mathcal{S}}$ to denote the subvector of vector x indexed by $\mathcal{S}$ and the submatrix of matrix A with columns indexed by $\mathcal{S}$, respectively. As usual, let $\mathbb{N}$ denote the set of nonnegative integers, $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$ and $\mathbb{R}_{++} := \{x \in \mathbb{R} \mid x > 0\}$. Moreover, we adopt $\mathcal{S}^c$ to denote the complement of $\mathcal{S}$, $\sharp(\cdot)$ to denote the number of elements in an index set, $\mathbb{I}$ represents an identity matrix and $A^\top$ denotes the transpose of matrix A, and $[n] := \{1, 2, \ldots, n\}$.

The support function and the signum function are denoted by supp : $\mathbb{R}^n \to 2^{[n]}$ and sign : $\mathbb{R} \to \mathbb{R}$, respectively; that is,

$$\mathrm{supp}(x) := \{i : x_i \neq 0\} \quad \text{for each } x \in \mathbb{R}^n,$$

and

$$\mathrm{sign}(t) := \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0, \end{cases} \quad \text{for each } t \in \mathbb{R}.$$

The restricted isometry property (RIP) [3] is a well-known regularity condition measuring how close the submatrices are nearly orthonormal restricted on sparse subspaces. The RIP has been widely used for the establishment of oracle property and recovery bound for sparse optimization problems [1–3], and the convergence analysis of sparse optimization algorithms [7, 20, 30]. Many types of random matrices, including Gaussian, Bernoulli, and partial Fourier matrices, have been shown to satisfy the RIP with exponentially high probability [31].

Definition 1 ([3]) Let $A \in \mathbb{R}^{m\times n}$ and $s \in \mathbb{N}$. The s-restricted isometry constant δ_s is defined to be the smallest quantity such that

$$(1-\delta_s)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta_s)\|x\|^2$$

for each $x \in \mathbb{R}^n$ with $\|x\|_0 \leq s$. The matrix A is said to satisfy the s-RIP with δ_s if $\delta_s < 1$.

It is clear by Definition 1 that δ_s is nondecreasing in s, that is $\delta_s \leq \delta_t$ whenever $s \leq t$. The following lemma recalls some properties of the RIP, which will be useful in convergence analysis of our proposed algorithms.

Lemma 1 Suppose that A satisfies s-RIP with $\delta_s < 1$. Let $x \in \mathbb{R}^n$, $\varepsilon \in \mathbb{R}^m$, $\mathcal{I}, \mathcal{J} \subseteq [n]$, and $v \in [0, \frac{1}{1-\delta_s}]$. Then the following assertions are true.

(i) If $\sharp(\mathcal{I} \cup \mathrm{supp}(x)) \leq s$, then $\|((\mathbb{I} - vA^\top A)x)_{\mathcal{I}}\| \leq (1 - v + v\delta_s)\|x\|$.
(ii) If $\sharp(\mathcal{I}) \leq s$, then $\|A_{\mathcal{I}}^\top \varepsilon\| \leq \sqrt{1+\delta_s}\|\varepsilon\|$.
(iii) If $\mathcal{I} \cap \mathcal{J} = \emptyset$ and $\sharp(\mathcal{I} \cup \mathcal{J}) \leq s$, then $\|A_{\mathcal{J}}^\top A_{\mathcal{I}} x_{\mathcal{I}}\| \leq \delta_s \|x_{\mathcal{I}}\|$.

Proof The proof of item (i) with a general v follows an analysis similar to that of [21, Lemma 6.16] with $v = 1$ and is thus omitted. Items (ii) and (iii) are taken from [30, Propositions 3.1 and 3.2] respectively. □

Definition 2 Let $\kappa : \mathbb{R}_{++} \to \mathbb{R}_+$ and $\lambda \in \mathbb{R}_{++}$.

(i) $\mathbb{T}_\lambda : \mathbb{R} \to \mathbb{R}$ is said to be a thresholding operator relative to $\kappa(\lambda)$ if the following thresholding property is satisfied:

$$\mathbb{T}_\lambda(t) = 0 \quad \text{whenever} \quad |t| \leq \kappa(\lambda). \tag{3}$$

(ii) $\mathbb{LT}_\lambda : \mathbb{R} \to \mathbb{R}$ is said to be a limited shrinkage thresholding operator relative to $\kappa(\lambda)$ if (3) and the following limited shrinkage property are satisfied:

$$|\mathbb{LT}_\lambda(t) - t| \le \kappa(\lambda) \quad \text{for each } t \in \mathbb{R}. \tag{4}$$

We use $\mathcal{LT}(\kappa;\ \lambda)$ to denote the family of limited shrinkage thresholding operators.

The limited shrinkage thresholding operator will provide a unified framework of proximal mappings of several nonconvex sparse penalties, including SCAD [9], MCP [12], and ℓ_p penalty $(0 \le p \le 1)$ [13] as special cases. For the sake of simplicity, we adopt the same notation for a separable operator from $\mathbb{R}^n$ to $\mathbb{R}^n$ with each component being the same operator from $\mathbb{R}$ to $\mathbb{R}$, for example $\mathbb{T}_\lambda(x) := (\mathbb{T}_\lambda(x_i))_{i=1}^n$ for each $x \in \mathbb{R}^n$.

3 Iterative limited shrinkage thresholding algorithms

The ITA has simple formulation and low computational complexity, which in general has the following iterative form:

$$x^{k+1} := \mathbb{T}_{v\lambda}(x^k - vA^\top(Ax^k - b)),$$

where $\mathbb{T}_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ is a thresholding operator relative to $\kappa(\lambda)$.

In this section, we will propose two general frameworks of ITAs by using a limited shrinkage thresholding operator and combining with the continuation technique and the truncation technique respectively. We will investigate their convergence to an approximate true sparse solution of (1). The following assumption on the limited shrinkage thresholding operator is made throughout this section.

Assumption 1 Let $\alpha, \beta > 0$ and suppose that $\mathbb{LT}_\lambda \in \mathcal{LT}(\kappa;\ \lambda)$ for $\kappa(\lambda) := \alpha\lambda^\beta$ and all $\lambda > 0$.

It was reported in [10] that the s-sparse solution of linear system (1) is unique under the s-RIP assumption. Throughout this paper, we adopt the following notation:

$$\text{Let } \bar{x} \text{ be an } s\text{-sparse solution of (1) with support } \mathcal{S} := \text{supp}(\bar{x}). \tag{5}$$

3.1 Iterative thresholding algorithms with continuation

Note that the regularization parameter λ plays an important role in the numerical performance of sparse optimization algorithms. According to the recovery bound theory, the regularization parameter λ should be small to guarantee the better recovery; however, the computational mathematics theory and extensive numerical studies show that a too small parameter will result in the ill-posedness of the subproblems and the convergence is faster if the parameter is properly larger. To inherit both advantages in theoretical and numerical aspects, the idea of the continuation technique is using a geometrically decreasing sequence of regularization parameters $\{\lambda_k\}$ starting at a large one in place of a fixed one; see, e.g., [6, 7, 26].

By virtue of a limited shrinkage thresholding operator and inspired by the idea of the continuation technique, we propose the following iterative limited shrinkage thresholding algorithm with the continuation technique (ILSTAC).

Algorithm 1 ILSTAC

```
1: Initialize regularization parameters $\lambda_0 > 0$, $\lambda > 0$ and continuation
   parameter $\gamma \in (0,1)$, initial point $x^0 := 0$, stepsize $v > 0$ and $k = 0$.
2: if $\lambda_k < \lambda$ then
3:     break
4: else
5:     $x^{k+1} := \mathbb{LT}_{v\lambda_k}(x^k - vA^\top(Ax^k - b))$
6:     $\lambda_{k+1} := \gamma\lambda_k$
7:     $k = k + 1$
8: end if
9: Output $x^* := x^k$
```

Remark 1 (i) The step 5 in ILSTAC consists of two steps: the first step is a gradient descent iterate for the least squares of (1), which gradually reduces the error of the linear system (1); and the second one is using a limited shrinkage thresholding operator to gradually transform the outcome of the descent iterate to a sparse subspace.

(ii) When the limited shrinkage thresholding operator $\mathbb{LT}_\lambda$ has a closed-form formula (see examples in Section 4), the ILSTAC inherits the significant advantages of the ITA that is of simple formulation and low computational complexity, and thus is extremely efficient for large-scale sparse optimization problems.

The main theorem of this subsection is as follows, which provides certain parameters setting (relevant to the RIP) in the ILSTAC to guarantee its convergence to an approximate true sparse solution of (1) within a tolerance proportional to a noise level. In addition, the support of the output of the ILSTAC has no false prediction and is exactly a subset of the support of the true sparse solution. Recall that $\bar{x}$ is the true s-sparse solution of (1) with support $\mathcal{S}$; see (5).

Theorem 1 *Suppose that Assumption 1 holds and A satisfies the RIP with*

$$(\sqrt{s}+1)\delta_{s+1} < 1. \tag{6}$$

Let $\alpha, \beta > 0$ be as in Assumption 1,

$$\eta \in \left(0, 1 - (\sqrt{s}+1)\delta_{s+1}\right), \tag{7}$$

set the stepsize $v \le \frac{1}{1-\delta_s}$*, the regularization parameters*

$$\lambda_0 \ge \frac{1}{v}\left(\frac{v\|\bar{x}\|}{\alpha(\sqrt{s}+1)}\right)^{\frac{1}{\beta}} \quad \textit{and} \quad \lambda := \frac{1}{v}\left(\frac{v\sqrt{1+\delta_s}}{\alpha\eta}\|\varepsilon\|\right)^{\frac{1}{\beta}}, \tag{8}$$

and the continuation parameter

$$\gamma \in \left[\left(\frac{(\sqrt{s}+1)v\delta_{s+1}}{1-\eta}+1-v\right)^{\frac{1}{\beta}}, 1\right). \tag{9}$$

Let Algorithm 1 with these parameters output x^**. Then it holds that*

$$\operatorname{supp}(x^*) \subseteq \mathcal{S} \quad \textit{and} \quad \|x^*-\bar{x}\| \le \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\|. \tag{10}$$

Furthermore, if $\min_{i\in\mathcal{S}}|\bar{x}_i| > \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\|$*, then* $\operatorname{supp}(x^*) = \mathcal{S}$*.*

Proof By assumption (6), one checks that $1-(\sqrt{s}+1)\delta_{s+1} > 0$, and thus η in (7) is well-defined. It follows from (7) that $\frac{(\sqrt{s}+1)v\delta_{s+1}}{1-\eta}+1-v<1$. Hence γ in (9) and Algorithm 1 with these parameters are well-defined.

To furniture the proof, we let Algorithm 1 generate the finite sequence $\{x^k\}_{k=0}^K$ and output $x^* = x^K$, and write

$$\rho := \frac{1-\eta}{\delta_{s+1}}, \tag{11}$$

and

$$y^k := x^k - vA^\top(Ax^k-b), \quad \mathcal{S}_k := \operatorname{supp}(x^k) \quad \text{and} \quad r^k := x^k - \bar{x} \tag{12}$$

for each $k = 0,\ldots,K$. By Assumption 1, one has that (3) and (4) are satisfied with $\kappa(\lambda) := \alpha\lambda^\beta$. We shall show by induction that the following inclusion and estimate hold for each $k = 0,\ldots,K$:

$$\mathcal{S}_k \subseteq \mathcal{S} \quad \text{and} \quad \|r^k\| \le \frac{\rho}{v}\kappa(v\lambda_k) = \rho\alpha v^{\beta-1}\lambda_k^\beta \tag{13}$$

(by Assumption 1). By the initial selection that $x^0 := 0$, one has that $\mathcal{S}_0 = \emptyset \subseteq \mathcal{S}$. By definition of ρ in (11) and assumption (7), we obtain by assumption of λ_0 in (8) that

$$\rho\alpha v^{\beta-1}\lambda_0^\beta > (\sqrt{s}+1)\alpha v^{\beta-1}\lambda_0^\beta \ge \|\bar{x}\| = \|r^0\|.$$

It is shown that (13) holds for $k = 0$.

Now suppose that (13) holds for iterate k ($< K$). Then by (12) and (1), we have that

$$y^k = x^k - vA^\top(Ax^k - A\bar{x} - \varepsilon) = x^k - vA^\top A_{\mathcal{S}}r^k_{\mathcal{S}} + vA^\top\varepsilon, \tag{14}$$

where the second equality follows from the hypothesis $\mathcal{S}_k \subseteq \mathcal{S}$ in (13). Fix $i \in \mathcal{S}^c$. It follows from the hypothesis $\mathcal{S}_k \subseteq \mathcal{S}$ in (13) that $x^k_i = 0$, and then (14) is reduced to

$$|y^k_i| \le v|A_i^\top A_{\mathcal{S}}r^k_{\mathcal{S}}| + v|A_i^\top\varepsilon|. \tag{15}$$

Since $\{i\}\cap\mathcal{S} = \emptyset$, we obtain by Lemma 1(iii) and (ii) that

$$|A_i^\top A_{\mathcal{S}}r^k_{\mathcal{S}}| \le \delta_{s+1}\|r^k_{\mathcal{S}}\|{=}\delta_{s+1}\|r^k\| \quad \text{and} \quad |A_i^\top\varepsilon| \le \sqrt{1+\delta_1}\|\varepsilon\| \le \sqrt{1+\delta_s}\|\varepsilon\|$$

(by the nondecreasing property that $\delta_1 \le \delta_s$). This, together with (15), yields that

$$|y^k_i| \le v\delta_{s+1}\|r^k\| + v\sqrt{1+\delta_s}\|\varepsilon\|. \tag{16}$$

By the stopping criterion that $\lambda_k < \lambda$ and the definition of λ in (8), one can get that $\lambda_k \geq \lambda = \frac{1}{v}\left(\frac{v\sqrt{1+\delta_s}}{\alpha\eta}\|\varepsilon\|\right)^{\frac{1}{\beta}}$, that is, $\|\varepsilon\| \leq \frac{\alpha\eta}{v\sqrt{1+\delta_s}}(v\lambda_k)^\beta$. This, together with hypothesis (13), deduces (16) to

$$|y_i^k| \leq v\delta_{s+1}\|r^k\| + v\sqrt{1+\delta_s}\|\varepsilon\| \leq \delta_{s+1}\rho\alpha(v\lambda_k)^\beta + \alpha\eta(v\lambda_k)^\beta = \alpha(v\lambda_k)^\beta, \quad (17)$$

where the equality holds by definition of ρ in (11). Thus (16) is reduced to $|y_i^k| \leq \kappa(v\lambda_k)$ by Assumption 1. Hence it follows from (3) that $x_i^{k+1} = 0$; consequently, $i \in \mathcal{S}_{k+1}^c$. This holds for any $i \in \mathcal{S}^c$, then we get that $\mathcal{S}^c \subseteq \mathcal{S}_{k+1}^c$, and equivalently, $\mathcal{S}_{k+1} \subseteq \mathcal{S}$.

On the other hand, we get by the inclusion $\mathcal{S}_{k+1} \subseteq \mathcal{S}$ that

$$\|x^{k+1} - \bar{x}\| = \|x_{\mathcal{S}}^{k+1} - \bar{x}_{\mathcal{S}}\| \leq \|x_{\mathcal{S}}^{k+1} - y_{\mathcal{S}}^k\| + \|y_{\mathcal{S}}^k - \bar{x}_{\mathcal{S}}\|. \quad (18)$$

By (4) and in view of Algorithm 1 that $x^{k+1} := \mathbb{LT}_{v\lambda_k}(y^k)$, we obtain that

$$\|x_{\mathcal{S}}^{k+1} - y_{\mathcal{S}}^k\| \leq \sqrt{s}\|x^{k+1} - y^k\|_\infty = \sqrt{s}\|\mathbb{LT}_{v\lambda_k}(y^k) - y^k\|_\infty \leq \sqrt{s}\kappa(v\lambda_k), \quad (19)$$

and by (12) and (14) that

$$\|y_{\mathcal{S}}^k - \bar{x}_{\mathcal{S}}\| = \|(x^k - \bar{x} - vA^\top Ar^k)_{\mathcal{S}} + vA_{\mathcal{S}}^\top\varepsilon\| = \|((\mathbb{I} - vA^\top A)r^k)_{\mathcal{S}} + vA_{\mathcal{S}}^\top\varepsilon\|.$$

Then it follows that

$$\|y_{\mathcal{S}}^k - \bar{x}_{\mathcal{S}}\| \leq \|((\mathbb{I} - vA^\top A)r^k)_{\mathcal{S}}\| + v\|A_{\mathcal{S}}^\top\varepsilon\|. \quad (20)$$

It follows from Lemma 1(i) and (ii) that

$$\|((\mathbb{I} - vA^\top A)r^k)_{\mathcal{S}}\| \leq (1 - v + v\delta_s)\|r^k\| \quad \text{and} \quad \|A_{\mathcal{S}}^\top\varepsilon\| \leq \sqrt{1+\delta_s}\|\varepsilon\|,$$

respectively. This, together with (18)-(20), implies that

$$\|x^{k+1} - \bar{x}\| \leq \sqrt{s}\kappa(v\lambda_k) + (1 - v + v\delta_s)\|r^k\| + v\sqrt{1+\delta_s}\|\varepsilon\|. \quad (21)$$

By the fact that $\delta_s \leq \delta_{s+1}$ and by (17), one has that

$$v\delta_s\|r^k\| + v\sqrt{1+\delta_s}\|\varepsilon\| \leq v\delta_{s+1}\|r^k\| + v\sqrt{1+\delta_s}\|\varepsilon\| \leq \kappa(v\lambda_k).$$

Combining this with (13), (21) is reduced to

$$\|x^{k+1} - \bar{x}\| \leq \left(\sqrt{s} + 1 + \frac{\rho}{v}(1-v)\right)\kappa(v\lambda_k) \leq \left(\sqrt{s} + 1 + \frac{\rho}{v}(1-v)\right)\alpha(v\lambda_k)^\beta \quad (22)$$

(by definition $\kappa(\lambda) := \alpha\lambda^\beta$ in Assumption 1). Noting by definition of ρ in (11) that

$$(\sqrt{s} + 1) + \frac{\rho}{v}(1-v) = \frac{\rho}{v}\left(\frac{(\sqrt{s}+1)v\delta_{s+1}}{1-\eta} + 1 - v\right) \leq \frac{\rho}{v}\gamma^\beta$$

(due to definition of γ in (9)), (22) is reduced to

$$\|x^{k+1} - \bar{x}\| \leq \frac{\rho}{v}\alpha(v\gamma\lambda_k)^\beta = \frac{\rho}{v}\kappa(v\lambda_{k+1})$$

(by the continuation rule that $\lambda_{k+1} := \gamma\lambda_k$). This, together with $\mathcal{S}_{k+1} \subseteq \mathcal{S}$, shows that (13) holds for each iterate $k = 0, \ldots, K$. Then we conclude by (13) that $\operatorname{supp}(x^*) \subseteq \mathcal{S}$ and

$$\|x^* - \bar{x}\| \leq \rho\alpha v^{\beta-1}\lambda_K^\beta < \rho\alpha v^{\beta-1}\lambda^\beta = \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\|$$

by definitions of λ and ρ in (8) and (11). Hence (10) is proved.

Moreover, suppose that $\min_{i\in\mathcal{S}}|\bar{x}_i| > \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\|$. We prove by contradiction, assuming that $\operatorname{supp}(x^*) \neq \mathcal{S}$. This, together with $\operatorname{supp}(x^*) \subseteq \mathcal{S}$ in (10), indicates that there exists $i \in \mathcal{S}$ such that $x_i^* = 0$. Hence

$$\|x^* - \bar{x}\| \geq |x_i^* - \bar{x}_i| > \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\|,$$

which yields a contradiction with the inequality in (10). Hence $\operatorname{supp}(x^*) = \mathcal{S}$. □

Remark 2 (i) When $\mathcal{S}$ is known, the initial point x^0 in ILSTAC can be chosen as one satisfying $\mathrm{supp}(x^0) \subset \mathcal{S}$ such that the result of Theorem 1 also holds; see the proof of Theorem 1.

(ii) By the continuation technique in Algorithm 1 and regularization parameters setting (8) in Theorem 1, particularly setting $\lambda_0 := \frac{\tau}{v}\left(\frac{v\|\bar{x}\|}{\alpha(\sqrt{s}+1)}\right)^{\frac{1}{\beta}}$ with $\tau \geq 1$, we obtain the complexity K of the ILSTAC to obtain x^*:

$$K := \left\lceil \log_{\gamma^{-1}} \frac{\lambda_0}{\lambda} \right\rceil = \left\lceil \frac{1}{\beta} \log_{\gamma^{-1}} \frac{\eta\tau^\beta\|\bar{x}\|}{(\sqrt{s}+1)\sqrt{1+\delta_s}\|\varepsilon\|} \right\rceil.$$

Adopting a similar proof in [21, Theorem 9.2] and using the concentration inequality in [22, Example 2.11], the following lemma provides a minimal requirement of the sample size for guaranteeing the s-RIP with high probability for a Gaussian matrix.

Lemma 2 Let $A \in \mathbb{R}^{m\times n}$ be a Gaussian matrix with each $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$. Let $0 < \delta, \varepsilon < 1$. Then A satisfies the s-RIP, where $\delta_s < \delta$ with probability at least $1-\varepsilon$ provided

$$m \geq \frac{16}{3}\delta^{-2}\left(s(9 + 2\ln\frac{n}{s}) + 2\ln\frac{2}{\varepsilon}\right).$$

Remark 3 The RIP assumption (6) is critical in Theorem 1 in guaranteeing the convergence of the ILSTAC to an approximate true sparse solution of (1). This remark provides some circumstances, where (6) is fulfilled.

(i) One can obtain by [2, Proposition 4.1] that (6) is satisfied when the following mutual incoherence property (MIP) is satisfied

$$\max_{i\neq j}|\langle A_i, A_j\rangle| \leq \frac{1}{s(\sqrt{s}+1)}.$$

Particularly, when A is column-wise normalized and $v = 1$, following a similar line of analysis, we can obtain the convergence result of Theorem 1 when the assumption (6) is replaced by $\max_{i\neq j}|\langle A_i, A_j\rangle| \leq \frac{1}{2s}$.

(ii) Suppose that $A \in \mathbb{R}^{m\times n}$ is a Gaussian matrix with each $A_{ij} \sim \mathcal{N}(0, \frac{1}{m})$. It follows from Lemma 2 that (6) is satisfied with probability at least $1-\varepsilon$ provided that

$$m \geq \frac{16}{3}(\sqrt{s}+1)^2\left((s+1)(9 + 2\ln\frac{n}{s+1}) + 2\ln\frac{2}{\varepsilon}\right).$$

3.2 Iterative thresholding algorithms with truncation

A truncation operator (also named the hard thresholding operator), denoted as $\mathbb{H}_s$, is to set all but the largest s elements of a vector (in magnitude) to zero [20]. By virtue of the limited shrinkage thresholding operator and the truncation operator, we propose an iterative limited shrinkage thresholding algorithm with the truncation technique (ILSTAT) to approach the true sparse solution of (1).

Algorithm 2 ILSTAT

```
Initialization: regularization parameter $\lambda > 0$, truncation parameter $s \in \mathbb{N}$,
  and random initial point $x^0 \in \mathbb{R}^n$, stepsize $v > 0$ and $k = 0$.
Iteration: for each $k \in \mathbb{N}$, having $x^k$, we determine $x^{k+1}$ via
    $x^{k+1} := \mathbb{H}_s \circ \mathbb{LT}_{v\lambda}(x^k - vA^\top(Ax^k - b)).$    (23)
```

Remark 4 (i) The ILSTAT adopts the truncation operator $\mathbb{H}_s$ to maintain the sparsity level s of the sequence $\{x^k\}$, which is helpful for guaranteeing the convergence to an approximate true solution with the required sparsity of (1); see Theorem 2.

(ii) Since the truncation operator $\mathbb{H}_s$ is very simple to calculate, the ILSTAT inherits the significant advantages of the ITA that is of simple formulation and low computational complexity, and thus is extremely efficient for large-scale sparse optimization problems, when $\mathbb{LT}_\lambda$ has a closed-form formulation.

The main result of this subsection is as follows, in which we establish the convergence of the ILSTAT to an approximate true sparse solution of (1) under the assumption of the RIP.

Theorem 2 *Suppose that Assumption 1 holds and A satisfies the $3s$-RIP. Let $\{x^k\}$ be a sequence generated by Algorithm 2 with stepsize*

$$\frac{1}{2(1-\delta_{3s})} < v \le \frac{1}{1-\delta_{3s}}. \tag{24}$$

Then $\{x^k\}$ converges approximately to $\bar{x}$ at a geometric rate; particularly,

$$\|x^k - \bar{x}\| \le \rho^k \|x^0 - \bar{x}\| + \frac{2}{1-\rho}(v\sqrt{1+\delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)), \tag{25}$$

where $\rho := 2(1 - v + v\delta_{3s}) \in (0, 1)$.

Proof By Assumption 1, one has that (3) and (4) are satisfied with $\kappa(\lambda) := \alpha\lambda^\beta$. To proceed the convergence analysis, we re-write the process (23) of Algorithm 2 into the following three steps:

$$y^k := x^k - vA^\top(Ax^k - b), \quad z^k := \mathbb{LT}_{v\lambda}(y^k), \quad x^{k+1} := \mathbb{H}_s(z^k). \tag{26}$$

Moreover, for the sake of simplicity, we write

$$r^k := x^k - \bar{x}, \quad \mathcal{S}_k := \mathrm{supp}(x^k) \quad \text{and} \quad \mathcal{I}_k = \mathcal{S} \cup \mathcal{S}_k; \tag{27}$$

and then one observes that

$$\sharp(\mathcal{S}) = s, \quad \sharp(\mathcal{S}_k) \le s \quad \text{and} \quad \sharp(\mathcal{I}_k) \le 2s \quad \text{for each } k \in \mathbb{N}. \tag{28}$$

Noting that

$$\mathrm{supp}(\bar{x}) = \mathcal{S} \subseteq \mathcal{I}_{k+1} \quad \text{and} \quad \mathrm{supp}(x^{k+1}) = \mathcal{S}_{k+1} \subseteq \mathcal{I}_{k+1}, \tag{29}$$

by (27) we get that

$$\|r^{k+1}\| = \|x^{k+1}_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\| \le \|x^{k+1}_{\mathcal{I}_{k+1}} - z^k_{\mathcal{I}_{k+1}}\| + \|z^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\|. \tag{30}$$

Noting by (26) that $x^{k+1} = \mathbb{H}_s(z^k) = \arg\min_{\|x\|_0 \le s} \|x - z^k\|$ (cf [20, pp. 266]), and by the fact that $\|\bar{x}\|_0 = s$, we obtain that $\|z^k - x^{k+1}\| \le \|z^k - \bar{x}\|$. This, together with (29), implies that $\|z^k_{\mathcal{I}_{k+1}} - x^{k+1}_{\mathcal{I}_{k+1}}\| \le \|z^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\|$. Consequently, (30) is reduced to

$$\|r^{k+1}\| \le 2\|z^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\| \le 2(\|z^k_{\mathcal{I}_{k+1}} - y^k_{\mathcal{I}_{k+1}}\| + \|y^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\|). \tag{31}$$

Noting by (26) that $z^k = \mathbb{LT}_{v\lambda}(y^k)$, we have by (4) that $\|z^k - y^k\|_\infty \le \kappa(v\lambda)$. Combining this with (28), we achieve that

$$\|y^k_{\mathcal{I}_{k+1}} - z^k_{\mathcal{I}_{k+1}}\| \le \sqrt{\sharp(\mathcal{I}_{k+1})}\|y^k - z^k\|_\infty \le \sqrt{2s}\kappa(v\lambda). \tag{32}$$

On the other hand, we obtain by the first equality of (26) and (1) that

$$y^k = x^k - vA^\top(Ax^k - A\bar{x} - \varepsilon) = x^k - vA^\top Ar^k + vA^\top\varepsilon$$

(due to (27)); and hence it follows that

$$\|y^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\| \le \|((\mathbb{I} - vA^\top A)r^k)_{\mathcal{I}_{k+1}}\| + v\|A^\top_{\mathcal{I}_{k+1}}\varepsilon\|. \tag{33}$$

Note by (28) and (29) that $\sharp(\mathcal{I}_{k+1}) \le 2s$ and $\sharp(\mathcal{I}_{k+1} \cup \mathrm{supp}(r^k)) = \sharp(\mathcal{I}_{k+1} \cup \mathcal{S}_k) \le 3s$. Then by the assumption of $3s$-RIP of A, we obtain by Lemma 1(i) and (ii) that

$$\|((\mathbb{I} - vA^\top A)r^k)_{\mathcal{I}_{k+1}}\| \le (1 - v + v\delta_{3s})\|r^k\| \quad \text{and} \quad \|A^\top_{\mathcal{I}_{k+1}}\varepsilon\| \le \sqrt{1 + \delta_{2s}}\|\varepsilon\|,$$

respectively. By the above two inequalities, (33) is reduced to

$$\|y^k_{\mathcal{I}_{k+1}} - \bar{x}_{\mathcal{I}_{k+1}}\| \le (1 - v + v\delta_{3s})\|r^k\| + v\sqrt{1 + \delta_{2s}}\|\varepsilon\|.$$

This, together with (31) and (32), yields that

$$\|r^{k+1}\| \le 2(1 - v + v\delta_{3s})\|r^k\| + 2(v\sqrt{1 + \delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)). \tag{34}$$

Let $\rho := 2(1 - v + v\delta_{3s})$. By assumption (24), we check that $\rho < 1$, and then obtain inductively by (34) that

$$\begin{aligned}\|x^{k+1} - \bar{x}\| &\le \rho\|x^k - \bar{x}\| + 2(v\sqrt{1 + \delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)) \\ &\le \cdots \\ &\le \rho^{k+1}\|x^0 - \bar{x}\| + \tfrac{2}{1-\rho}(v\sqrt{1 + \delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)).\end{aligned}$$

The proof is complete. □

Remark 5 (i) Theorem 2 shows a geometric convergence rate of the ILSTAT to an approximate true sparse solution of (1) within a tolerance. The tolerance in (25) has an additive form of a term on noise level $\mathcal{O}(\|\varepsilon\|)$ and a term on limited shrinkage thresholding operator $\mathcal{O}(\kappa(v\lambda))$.

(ii) As in Assumption 1, $\kappa(\lambda) = \alpha\lambda^\beta$ for some $\alpha, \beta > 0$, which could be small when a small regularization parameter λ is selected. For example, it will be shown later in Lemmas 3 and 4 that Assumption 1 holds with $\kappa(\lambda) := \lambda$ for the SCAD and MCP penalty, and with $\kappa(\lambda) := \alpha_p\lambda^{\frac{1}{2-p}}$ for the ℓ_p penalty (where α_p is given by (45)), respectively. The orders of λ's in these $\kappa(\lambda)$'s are the same as the ones in the corresponding recovery bounds of problem (2) with SCAD/MCP penalty [17, Theorem 1] and with ℓ_p penalty [10, Theorem 9] and [15, Theorem 2], respectively. It will be illustrated in our numerical experiments in Section 5 that the best regularization parameter is about $\lambda = 10^{-4}$.

(iii) By (25), we obtain the complexity of the ILSTAT that

$$\|x^{k^*} - \bar{x}\| \leq \frac{3-\rho}{1-\rho}(v\sqrt{1+\delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)), \tag{35}$$

after at most $k^* := \lceil \log_{\rho^{-1}} \frac{\|x^0 - \bar{x}\|}{v\sqrt{1+\delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda)} \rceil$ iterates. Indeed, we have by definition of k^* that

$$\rho^{k^*}\|x^0 - \bar{x}\| \leq v\sqrt{1+\delta_{2s}}\|\varepsilon\| + \sqrt{2s}\kappa(v\lambda).$$

This, together with (25), implies (35).

4 Proximal gradient algorithms for nonconvex regularization problems

Proximal gradient algorithm (PGA) [4, 10] is one of the most popular and practical algorithms for (convex or nonconvex) composite optimization problem (2), which successively processes the gradient descent operator on the least square and the proximal operator on the penalty function φ:

$$x^{k+1} := \text{Prox}_{v\lambda\varphi}(x^k - vA^\top(Ax^k - b)),$$

where the proximal mapping $\text{Prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\text{Prox}_f(y) := \arg\min_{x\in\mathbb{R}^n} \; f(x) + \frac{1}{2}\|x-y\|^2 \quad \text{for each } y \in \mathbb{R}^n.$$

When the penalty function is separable, i.e.,

$$\varphi(x) := \sum_{i=1}^n \phi(x_i) \quad \text{for each } x \in \mathbb{R}^n, \tag{36}$$

the iteration of the PGA is equivalent to a cycle of one-dimensional proximal optimization subproblems

$$x_i^{k+1} := \text{Prox}_{v\lambda\phi}(x_i^k - vA_i^\top(Ax^k - b)) \quad \text{for } i = 1, \ldots, n, \tag{37}$$

and then $\text{Prox}_{\lambda\varphi}(x) = (\text{Prox}_{\lambda\phi}(x_i))_{i=1}^n$ for each $x \in \mathbb{R}^n$.

Inspired by the idea of Algorithms 1 and 2 (with the proximal mapping $\text{Prox}_{\lambda\varphi}$ in place of the limited shrinkage thresholding operator $\mathbb{LT}_\lambda$), we obtain the PGA with the continuation technique (PGAC) and the PGA with the truncation technique (PGAT) for solving the problem (2) respectively.

The main computational task of PGAC and PGAT is the proximal mapping $\text{Prox}_{\lambda\varphi}$ of the nonconvex penalty function φ. For the popular nonconvex penalty functions including SCAD [9], MCP [12], and ℓ_p penalty with $p \in [0,1]$ [13], the penalty is separable and the one-dimensional proximal mapping (37) has a closed-form formula. Therefore the corresponding algorithms can be efficiently implemented in a parallel and analytical manner and extremely efficient for large-scale sparse optimization problems.

4.1 SCAD and MCP

Let $a > 2$. The SCAD penalty [9] is of separable form (36) with

$$\phi_{\mathrm{SCAD}}(t) := \int_0^{|t|} \min\left\{1, \frac{(a\lambda - \alpha)_+}{(a-1)\lambda}\right\} \mathrm{d}\alpha = \begin{cases} |t|, & |t| < \lambda, \\ \frac{2a|t| - \frac{t^2}{\lambda} - \lambda}{2(a-1)}, & \lambda \le |t| < a\lambda, \\ \frac{(a+1)\lambda}{2}, & |t| > a\lambda, \end{cases} \quad (38)$$

where $t \in \mathbb{R}$. The proximal mapping of the SCAD penalty (38) has a closed-form formula (see [9, Eq. (2.8)]):

$$\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}}(t) = \begin{cases} 0, & |t| \le \lambda, \\ \mathrm{sign}(t)(|t| - \lambda), & \lambda < |t| \le 2\lambda, \\ \dfrac{(a-1)t - \mathrm{sign}(t)a\lambda}{a-2}, & 2\lambda < |t| \le a\lambda, \\ t, & |t| > a\lambda. \end{cases} \quad (39)$$

Let $a > 1$. The MCP penalty [12] is of separable form (36) with

$$\phi_{\mathrm{MCP}}(t) := \int_0^{|t|} \left(1 - \frac{\alpha}{a\lambda}\right)_+ \mathrm{d}\alpha = \begin{cases} |t| - \frac{t^2}{2a\lambda}, & |t| < a\lambda, \\ \frac{1}{2}a\lambda, & |t| \ge a\lambda, \end{cases} \quad (40)$$

where $t \in \mathbb{R}$. The proximal mapping of the MCP penalty (40) has a closed-form formula:

$$\mathrm{Prox}_{\lambda\phi_{\mathrm{MCP}}}(t) = \begin{cases} 0, & |t| \le \lambda, \\ \frac{\mathrm{sign}(t)(|t|-\lambda)}{1-1/a}, & \lambda < |t| \le a\lambda, \\ t, & |t| > a\lambda. \end{cases} \quad (41)$$

The following lemma validates that the proximal mappings of the SCAD penalty and the MCP penalty are limited shrinkage thresholding operators relative to an identity function.

Lemma 3 $\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}} \in \mathcal{LT}(\kappa;\ \lambda)$ and $\mathrm{Prox}_{\lambda\phi_{\mathrm{MCP}}} \in \mathcal{LT}(\kappa;\ \lambda)$ with $\kappa(\lambda) := \lambda$ for each $\lambda \in \mathbb{R}_{++}$.

Proof It directly follows from (39) and (41) that $\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}}$ and $\mathrm{Prox}_{\lambda\phi_{\mathrm{MCP}}}$ satisfy the thresholding property (3) with $\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}}$ in place of $\mathbb{T}_\lambda$, respectively. Moreover, we have by (39) that

$$|\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}}(t) - t| = \begin{Bmatrix} |t|, & |t| \le \lambda, \\ \lambda, & \lambda < |t| \le 2\lambda, \\ \frac{a\lambda - |t|}{a-2}, & 2\lambda < |t| \le a\lambda, \\ 0, & |t| > a\lambda, \end{Bmatrix} \le \lambda,$$

and by (41) that

$$|\mathrm{Prox}_{\lambda\phi_{\mathrm{MCP}}}(t) - t| = \begin{Bmatrix} |t|, & |t| \le \lambda, \\ \frac{a\lambda - |t|}{a-1}, & \lambda < |t| \le a\lambda, \\ 0, & |t| > a\lambda, \end{Bmatrix} \le \lambda.$$

Consequently, one can check the limited shrinkage property (4) with $\mathrm{Prox}_{\lambda\phi_{\mathrm{SCAD}}}$ and $\mathrm{Prox}_{\lambda\phi_{\mathrm{MCP}}}$ in place of $\mathbb{LT}_{\lambda}$, respectively. □

Directly applying Theorems 1 and 2 and Lemma 3, we present in the following theorems the convergence of the PGAC and the PGAT with SCAD proximal mapping (39) or MCP proximal mapping (41) to an approximate true sparse solution of (1) under the assumption of the RIP. Recall that $\bar{x}$ is the true s-sparse solution of (1) with support $\mathcal{S}$.

Theorem 3 *Suppose that A satisfies the s-RIP with (6). Let η be defined in (7), and set*

$$v \le \frac{1}{1-\delta_s}, \quad \lambda_0 \ge \frac{\|\bar{x}\|}{\sqrt{s}+1}, \quad \lambda := \frac{\sqrt{1+\delta_s}}{\eta}\|\varepsilon\|, \quad \gamma \in \left[\frac{(\sqrt{s}+1)v\delta_{s+1}}{1-\eta}+1-v, 1\right).$$

Let the PGAC with these parameters and SCAD proximal mapping (39) or MCP proximal mapping (41) output x^*. Then (10) is satisfied.

Theorem 4 *Suppose that A satisfies the $3s$-RIP. Let $\{x^k\}$ be a sequence generated by the PGAT with SCAD proximal mapping (39) or MCP proximal mapping (41) and stepsize (24). Then $\{x^k\}$ converges approximately to $\bar{x}$ at a geometric rate; particularly,*

$$\|x^k - \bar{x}\| \le \rho^k \|x^0 - \bar{x}\| + \frac{2v}{1-\rho}(\sqrt{1+\delta_{2s}}\|\varepsilon\| + \sqrt{2s}\lambda),$$

where $\rho := 2(1 - v + v\delta_{3s}) \in (0, 1)$.

Combining Theorems 3 and 4 with the recovery bound results [17], we obtain the convergence of the PGAC and the PGAT with SCAD proximal mapping (39) or MCP proximal mapping (41) to an approximation global solution of the corresponding problem (2) in Corollaries 1 and 2 below respectively, when the noise ε in (1) is a Gaussian one or a sub-Gaussian one. For notation simplicity, let $\varphi_{\mathrm{S/M}}$ and $x_{\mathrm{S/M}}$ denote the penalty function and a global solution of problem (2) when φ is the φ_{SCAD} or φ_{MCP} penalty.

Corollary 1 *Suppose that assumptions in Theorem 3 are satisfied. Suppose that ε in (1) is a Gaussian noise or a sub-Gaussian noise and the following restricted invertibility condition related to SCAD/MCP penalty is satisfied with $\eta > 1$:*

$$\min\left\{\frac{\sqrt{s}\|A^\top Ax\|_\infty}{m\|x\|} : \varphi_{\mathrm{S/M}}(x_{\mathcal{S}^c}) \le \eta\, \varphi_{\mathrm{S/M}}(x_{\mathcal{S}})\right\} > 0. \tag{42}$$

Then it holds with probability $1 - \exp(-\frac{3-2\sqrt{2}}{2}m)$ that

$$\|x^* - x_{S/M}\| \le \mathcal{O}(\sqrt{s}\lambda).$$

Proof By assumption of the restricted invertibility condition (42), it follows from the recovery bound result of problem (2) with SCAD/MCP penalty [17, Theorem 1] that

$$\|x_{\mathrm{S/M}} - \bar{x}\| \leq \mathcal{O}(\sqrt{s}\lambda) \tag{43}$$

with probability $1 - \exp(-\frac{3-2\sqrt{2}}{2}m)$. On the other hand, Theorem 3 shows that the PGAC with SCAD/MCP proximal mapping outputs x^* that satisfies

$$\|x^* - \bar{x}\| \leq \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\| = \frac{(1-\eta)}{\delta_{s+1}}\lambda = \mathcal{O}(\sqrt{s}\lambda).$$

This, together with (43), implies that

$$\|x^* - x_{\mathrm{S/M}}\| \leq \|x^* - \bar{x}\| + \|x_{\mathrm{S/M}} - \bar{x}\| \leq \mathcal{O}(\sqrt{s}\lambda)$$

with probability $1 - \exp(-\frac{3-2\sqrt{2}}{2}m)$. The proof is complete. □

Corollary 2 *Suppose that assumptions in Theorem 4 and Corollary 1 are satisfied. Then it holds with probability* $1 - \exp(-\frac{3-2\sqrt{2}}{2}m)$ *that*

$$\|x^k - x_{\mathrm{S/M}}\| \leq \rho^k\|x^0 - \bar{x}\| + \mathcal{O}(\|\varepsilon\|) + \mathcal{O}(\sqrt{s}\lambda),$$

where $\rho := 2(1 - v + v\delta_{3s}) \in (0,1)$.

4.2 ℓ_p penalty

For $0 \leq p \leq 1$, the ℓ_p penalty [13] is of separable form (36) with

$$\phi_{\ell_p}(t) := |t|^p \quad \text{for each } t \in \mathbb{R}, \tag{44}$$

where we adopt the convenience that $0^0 = 0$. Write

$$\alpha_p := \begin{cases} 1, & p = 1, \\ (2-p)(2-2p)^{-\frac{1-p}{2-p}}, & p \in [0,1). \end{cases} \tag{45}$$

The proximal mapping of the ℓ_p penalty (44) has a solution formulated as (see [32, Theorem 1])

$$\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t) = \begin{cases} 0, & |t| \leq \alpha_p\lambda^{\frac{1}{2-p}}, \\ t^*, & |t| > \alpha_p\lambda^{\frac{1}{2-p}}, \end{cases} \tag{46}$$

where t^* is the unique (nonzero) solution of the following problem:

$$\min_{x\in\mathbb{R}} h(x) := \lambda|x|^p + \frac{1}{2}(x-t)^2. \tag{47}$$

The following lemma validates that the proximal mapping of the ℓ_p penalty is a limited shrinkage thresholding operator.

Lemma 4 $\mathrm{Prox}_{\lambda\phi_{\ell_p}} \in \mathcal{LT}(\kappa;\ \lambda)$ with $\kappa(\lambda) := \alpha_p\lambda^{\frac{1}{2-p}}$ for each $\lambda \in \mathbb{R}_{++}$.

Proof It directly follows from (46) that $\mathrm{Prox}_{\lambda\phi_{\ell_p}}$ satisfies the thresholding property (3) with $\mathrm{Prox}_{\lambda\phi_{\ell_p}}$ in place of $\mathbb{T}_\lambda$ and $\kappa(\lambda) := \alpha_p\lambda^{\frac{1}{2-p}}$. Then it remains to show that

$$|\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t) - t| \le \alpha_p\lambda^{\frac{1}{2-p}} \quad \text{for each } t \in \mathbb{R}. \tag{48}$$

When $|t| \le \alpha_p\lambda^{\frac{1}{2-p}}$, one has by (46) that $|\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t)-t| = |t| \le \alpha_p\lambda^{\frac{1}{2-p}}$. Below we consider the case when $|t| > \alpha_p\lambda^{\frac{1}{2-p}}$. In this case, we get by (46) that $\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t) = t^*$ is the unique (nonzero) solution of problem (47); consequently, the optimality conditions of (47) says that

$$h'(t^*) = 0 \quad \text{and} \quad h''(t^*) \ge 0.$$

By definition of $h(\cdot)$ in (47), it follows that $h'(t^*) = \lambda p|t^*|^{p-1}\mathrm{sign}(t^*) + t^* - t = 0$; and consequently,

$$|\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t) - t| = |t^* - t| = \lambda p|t^*|^{p-1}. \tag{49}$$

Moreover, one has that $h''(t^*) = \lambda p(p-1)|t^*|^{p-2} + 1 \ge 0$; hence, $|t^*|^{p-1} \le (\lambda p(1-p))^{-\frac{1-p}{2-p}}$. Therefore, (49) is reduced to

$$|\mathrm{Prox}_{\lambda\phi_{\ell_p}}(t) - t| \le (\lambda p)^{\frac{1}{2-p}}(1-p)^{-\frac{1-p}{2-p}}. \tag{50}$$

Below we will claim that

$$(\lambda p)^{\frac{1}{2-p}}(1-p)^{-\frac{1-p}{2-p}} \le \alpha_p\lambda^{\frac{1}{2-p}} \quad \text{for each } p \in [0,1]. \tag{51}$$

Granting this, (50) is reduced to (48), as desired.

To show (51), we define $f : [0,1] \to \mathbb{R}_+$ by

$$f(p) := \frac{(\lambda p)^{\frac{1}{2-p}}(1-p)^{-\frac{1-p}{2-p}}}{\alpha_p\lambda^{\frac{1}{2-p}}} = \frac{2}{2-p}\left(\frac{p}{2}\right)^{\frac{1}{2-p}} \quad \text{for each } p \in [0,1]$$

(by (45)). By the elementary calculus, one can check that $f(0) = 0$, $f(1) = 1$ and $f'(p) > 0$ for each $p \in (0,1)$. Then it follows that $f(p) \le 1$ for each $p \in [0,1]$, and thus (51) is shown to hold. The proof is complete. □

Directly applying Theorems 1 and 2 and Lemma 4, we present the convergence of the PGAC and the PGAT with ℓ_p proximal mapping (46) to an approximate true sparse solution of (1) under the assumption of the RIP.

Theorem 5 *Suppose that assumptions in Theorem 1 holds with $\alpha := \alpha_p$ and $\beta := \frac{1}{2-p}$. Let the PGAC with the given parameters and ℓ_p proximal mapping (46) output x^*. Then (10) is satisfied.*

Theorem 6 *Suppose that A satisfies the $3s$-RIP. Let $\{x^k\}$ be a sequence generated by the PGAT with ℓ_p proximal mapping (46) and stepsize (24). Then $\{x^k\}$ converges approximately to $\bar{x}$ at a geometric rate; particularly,*

$$\|x^k - \bar{x}\| \le \rho^k\|x^0 - \bar{x}\| + \frac{2}{1-\rho}(v\sqrt{1+\delta_{2s}}\|\varepsilon\| + \alpha_p\sqrt{2s}(v\lambda)^{\frac{1}{2-p}}),$$

where $\rho := 2(1 - v + v\delta_{3s}) \in (0,1)$.

Combining Theorems 5 and 6 with the recovery bound result for the ℓ_p regularization problem [15], we obtain the convergence of the PGAC and the PGAT with the ℓ_p proximal mapping (46) to an approximation global solution of the corresponding problem (2) in Corollaries 3 and 4 below respectively, when the noise ε in (1) is a Gaussian one or a sub-Gaussian one.

Corollary 3 *Suppose that assumptions in Theorem 5 are satisfied. Suppose that ε in (1) is a Gaussian noise or a sub-Gaussian noise and the following p-restricted eigenvalue condition is satisfied with $\eta > 1$:*

$$\min\left\{\frac{\|Ax\|}{\sqrt{m}\|x\|} : \|x_{\mathcal{S}^c}\|_p^p \le \eta\,\|x_{\mathcal{S}}\|_p^p\right\} > 0. \tag{52}$$

Let x_{ℓ_p} be a global solution of problem (2) with ℓ_p penalty. Then it holds with probability $1-\exp(-m)-(\sqrt{\pi\log n})^{-1}$ that

$$\|x^* - x_{\ell_p}\| \le \mathcal{O}(\sqrt{s}\lambda^{\frac{1}{2-p}}).$$

Proof By assumption of the p-restricted eigenvalue condition (52), it follows from the recovery bound result of the ℓ_p regularization problem [15, Theorem 2] that

$$\|x_{\ell_p} - \bar{x}\| \le \mathcal{O}(\sqrt{s}\lambda^{\frac{1}{2-p}}) \tag{53}$$

with probability $1-\exp(-m)-(\sqrt{\pi\log n})^{-1}$. On the other hand, Theorem 5 shows that the PGAC with ℓ_p proximal mapping (46) outputs x^* that satisfies

$$\|x^* - \bar{x}\| \le \frac{(1-\eta)\sqrt{1+\delta_s}}{\eta\delta_{s+1}}\|\varepsilon\| = \frac{(1-\eta)\alpha_p}{\delta_{s+1}v^{\frac{1-p}{2-p}}}\lambda^{\frac{1}{2-p}} = \mathcal{O}(\sqrt{s}\lambda^{\frac{1}{2-p}}).$$

This, together with (53), implies that

$$\|x^* - x_{\ell_p}\| \le \|x^* - \bar{x}\| + \|x_{\ell_p} - \bar{x}\| \le \mathcal{O}(\sqrt{s}\lambda^{\frac{1}{2-p}}).$$

The proof is complete. □

Corollary 4 *Suppose that assumptions in Theorem 6 and Corollary 3 are satisfied. Then it holds with probability $1-\exp(-m)-(\sqrt{\pi\log n})^{-1}$ that*

$$\|x^k - x_{\ell_p}\| \le \rho^k\|x^0 - \bar{x}\| + \mathcal{O}(\|\varepsilon\|) + \mathcal{O}(\sqrt{s}\lambda^{\frac{1}{2-p}}),$$

where $\rho := 2(1-v+v\delta_{3s}) \in (0,1)$.

Remark 6 It was shown in [27, Theorem 2] that the ISTA and the IHTA with the continuation technique for solving the ℓ_1 and ℓ_0 regularization problems converge to an approximate true sparse solution of (1) under the assumption of the mutual incoherence property (MIP). Theorem 5 extends and improves [27, Theorem 2] in several aspects:

- Theorem 5 considers a unified framework of the PGAC for solving the ℓ_p regularization problem with $p \in [0,1]$, which covers the ISTA (when $p=1$) and the IHTA (when $p=0$) with the continuation technique in [27] as special cases.

- Jiao et al. [27] considered problem (1) with A being column normalized and the ISTA and the IHTA with stepsize $v = 1$, while Theorem 5 considers problem (1) with A being a general matrix and the PGAC with a general stepsize $0 < v < \frac{1}{1-\delta_s}$.
- Even for the special cases of the ISTA and the IHTA with the continuation technique, Theorem 5 improves [27, Theorem 2] in the sense that our convergence result is established under the assumption of the RIP, which is weaker than the MIP assumed in [27, Theorem 2]; see [2, Proposition 4.1].

5 Numerical experiments

In this section, we carry out experiments to illustrate the numerical performance of the PGAC and the PGAT for problem (2) with SCAD, MCP and ℓ_p penalty ($p = 0, \frac{1}{2}, 1$) respectively, and to compare it with the standard PGA. All numerical experiments are implemented in R (4.0.0) and executed on a personal desktop (Intel Core Duo i7-8550, 1.80 GHz, 8.00 GB of RAM).

In numerical experiments, the simulated data are generated via the standard process of compressive sensing. In details, we randomly generate an i.i.d. Gaussian ensemble $A \in \mathbb{R}^{m\times n}$ satisfying $AA^\top = I$, and a true $\bar{s}$-sparse solution $\bar{x} \in \mathbb{R}^n$ via randomly picking $\bar{s}$ nonzero entries with an i.i.d. Gaussian ensemble. Then the observation b is generated via

$$b = A\bar{x} + \sigma\varepsilon_1,$$

where $\sigma \in \mathbb{R}$ and $\varepsilon_1 \sim \mathcal{N}(0, 1)$ is a standard Gaussian noise. In the numerical experiments, the problem size is set as $m = 256$ and $n = 1024$ and the noise level $\sigma = 0.1\%$.

The parameters in nonconvex penalties are set as: $a = 16$ in SCAD (38) and MCP (40) and $p = 0, \frac{1}{2}, 1$ in ℓ_p penalty (44). In the implementation, we select the initial point $x^0 := 0$, the stepsize $v := 1$, the maximum number of iterations as 500; the regularization parameters are set as $\lambda_0 := \frac{\|\bar{x}\|}{\sqrt{\bar{s}}+1}$ and $\lambda := 10^{-4}$ in PGAC and $\lambda \in [10^{-4}, 1]$ via cross validation in PGA and PGAT. The performance of the algorithms is evaluated via two major criteria:

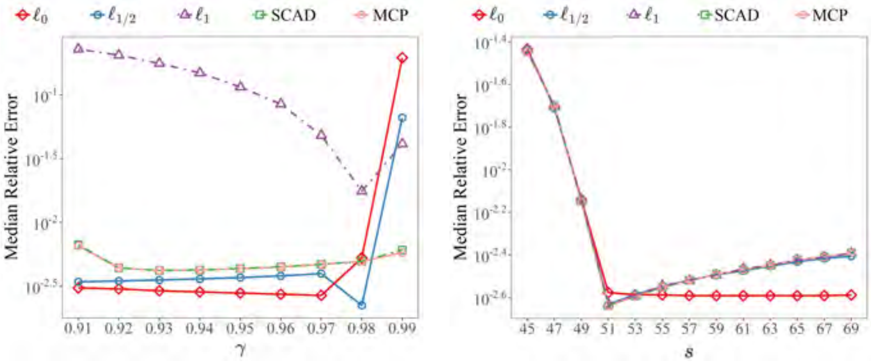
- (Accuracy) The relative error (RE): RE $:= \frac{\|x-\bar{x}\|}{\|\bar{x}\|}$.
- (Stability) The successful recovery rate: the ratio of successful recovery with RE $< 10^{-2}$.

The stopping criteria of the algorithms are listed as follows.

- PGAC: the number of iterations is greater than 500, or $\lambda^k < \lambda$.
- PGAT and PGA: the number of iterations is greater than 500 or $\|x^k - x^{k-1}\| \le 10^{-6}$.

The first experiment aims to show the numerical performance of the PGAC with different continuation parameter γ and the PGAT with different truncation parameter s. In this experiment, the sparsity of the true solution is set as

$\bar{s} = 51$. Figure 2(a) and (b) plot the average RE of the solution generated by the PGAC with γ varying from $(0.9, 1)$ and the one of the PGAT with s varying from $(45, 70)$ in 200 random trials, respectively. It is demonstrated from Figure 2(a) that the PGAC with nonconvex penalty obtains an accurate estimation when $\gamma \in [0.91, 0.98]$, while the PGAC with ℓ_1 penalty only approaches an accurate estimation when $\gamma = 0.98$. It is indicated from Figure 2(b) that the PGAT with these five penalties have similar performance: the PGAT cannot achieve an accurate estimation when $s < \bar{s}$ and approaches an accurate estimation when $s \geq \bar{s}$ slightly (within 20%). Therefore, in the following two numerical experiments, let the continuation parameter in PGAC $\gamma = 0.98$ and the truncation parameter in PGAT $s = \bar{s}$ as default.



(a) Continuation parameter γ in PGAC

(b) Truncation parameter s in PGAT

Fig. 2: Numerical results of PGAC and PGAT with different parameters.

The second experiment aims to compare the convergence behavior of the PGAC and the PGAT with the standard PGA. In this experiment, the sparsity ratio of the true solution is set as $\bar{s}/n := 2\%$. Figure 3 plots the average RE of the PGAs starting from 100 different and random initial points along the number of iterations in a random trials. It is displayed from Figure 3 that the PGAC and the PGAT converge faster and achieve a more accurate solution than the standard PGA. This validates the accelerating capability and the convergence to an approximate true solution of the continuation technique and the truncation technique in PGA for sparse optimization.

The third experiment aims to compare the stability of the PGAC and the PGAT with the standard PGA and with different sparse penalties. Figure 4 plots the successful recovery rates of PGA, PGAC and PGAT within 200 random trials at each sparsity level. It is indicated from Figure 4 that (i) the PGAC and the PGAT can achieve a higher successful recovery rate than the standard PGA; and (ii) the PGAC with $\ell_{1/2}$ penalty outperforms other penalties, while the PGAT with different sparse penalties share comparable stability.

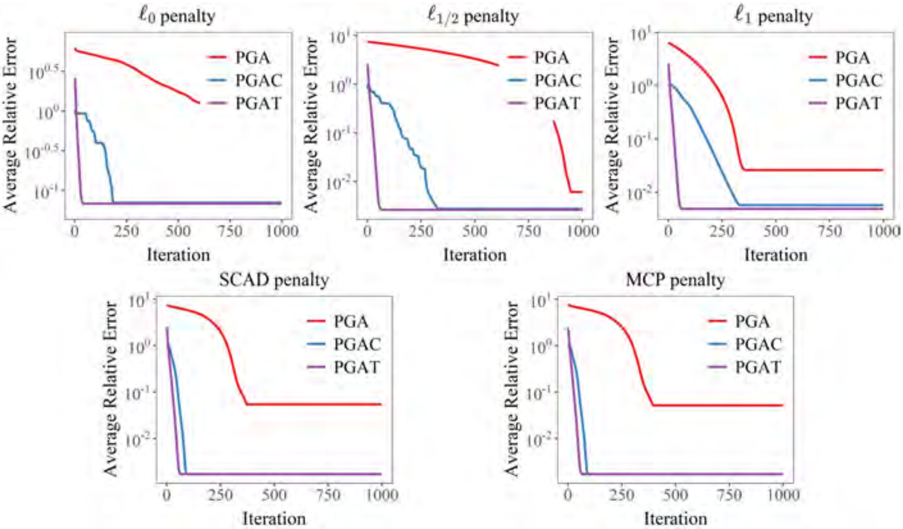


Fig. 3: Convergence behavior of PGAs starting from different initial points.

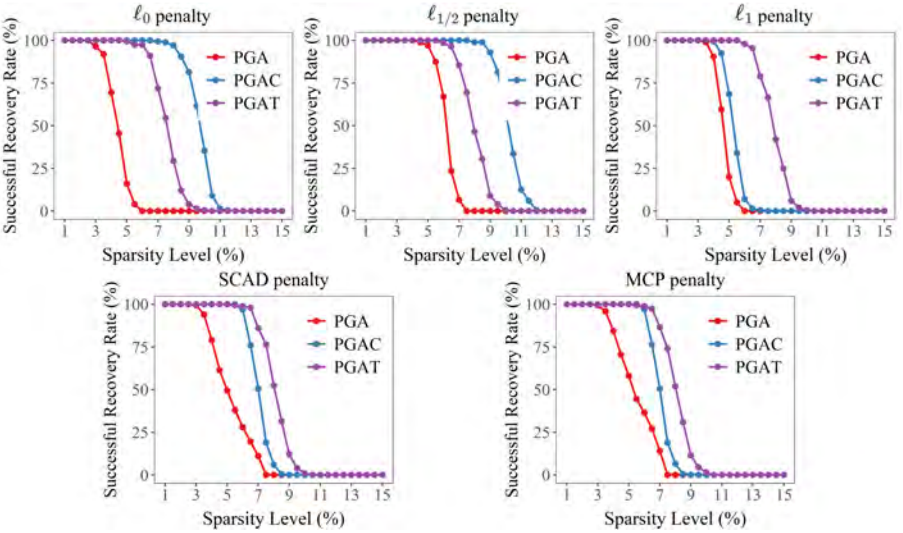
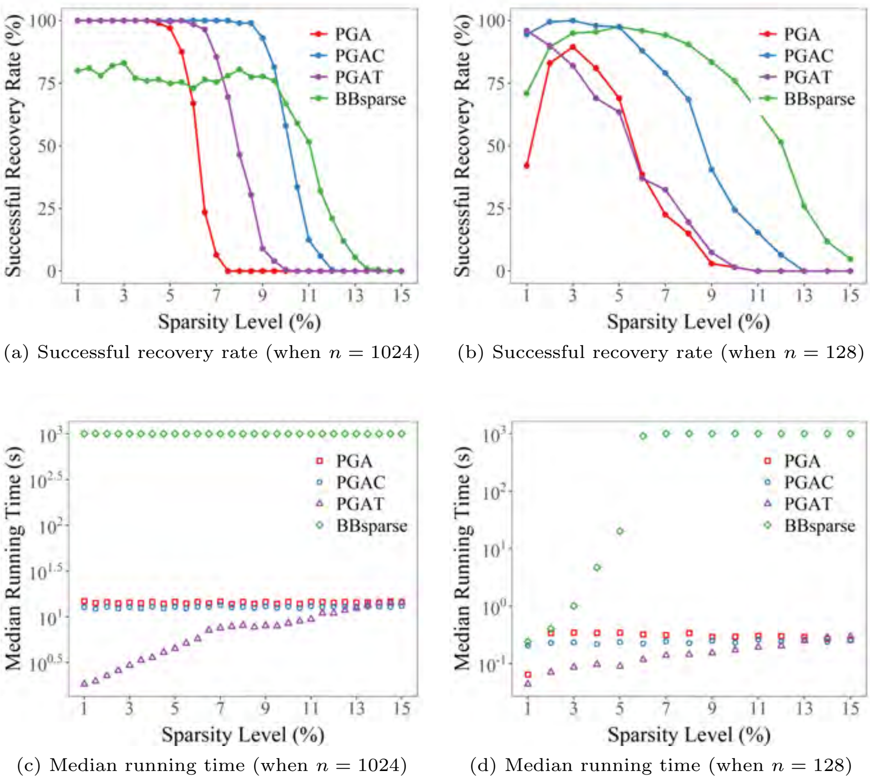


Fig. 4: Successful recovery rates of PGAs.

We also compare the PGAs (for problem (2) with $\ell_{1/2}$ penalty) with a recently proposed branch-and-bound method (BBsparse in short) for problem (2) with ℓ_0 penalty[1] [33], in which an additional stopping criterion is the maximum running time set as 1000 seconds. Figure 5 presents the successful

[1] The code is available at https://github.com/ramzi-benmhenni/BBsparse.

recovery rates and median running times of PGAs and BBsparse for two different situations of variable dimensions: $n = 1024$ and $n = 128$, for which the sample size is $m = 256$ and $m = 32$, respectively. Figures 5(a) and 5(b) demonstrate that PGAs outperform BBsparse with high successful recovery rates for both high and low dimensional problems when the sparsity level is relatively small, while BBsparse outperforms PGAs when the sparsity level is relatively large. In Figures 5(c) and 5(d), it is observed that the PGAs are almost 100 times faster than the BBsparse for both high and low dimensional problems.



(a) Successful recovery rate (when $n = 1024$) (b) Successful recovery rate (when $n = 128$)

(c) Median running time (when $n = 1024$) (d) Median running time (when $n = 128$)

Fig. 5: Numerical results of PGAs and BBsparse.

In conclusion, the numerical experiments show that the PGAC and the PGAT for nonconvex regularization problems (2) have the strong sparsity promoting capability and outperforms the standard PGA on both accuracy and robustness, benefiting from the nonconvex sparse penalty and the continuation or truncation technique.

6 Conclusion

In this paper, we proposed two frameworks of ITAs by employing the limited shrinkage thresholding operator with the continuation technique and the truncation technique respectively, and established their convergence to an approximate true sparse solution of linear system (1) under the assumption of the RIP. Moreover, applying to nonconvex regularization problems (2) with SCAD, MCP and ℓ_p penalty ($0 \le p \le 1$), we obtained the convergence of the PGA with the continuation technique or the truncation technique to an approximate true sparse solution of (1), and an approximate global solution of (2) by virtue of their recovery bound theory. Preliminary numerical results show that the continuation technique and the truncation technique can speed up the convergence and improve the stability of the algorithm, and particularly, are able to escape from the local solution to converge to the true sparse solution.

Acknowledgements. The authors are grateful to the editor and the anonymous reviewers for their valuable comments and suggestions toward the improvement of this paper. Yaohua Hu's work is supported in part by the National Natural Science Foundation of China (12222112, 12071306, 32170655), Project of Educational Commission of Guangdong Province (2023ZDZX1017), and Shenzhen Science and Technology Program (RCJC20221008092753082). Xiaoqi Yang's work was supported in part by the Research Grants Council of Hong Kong (PolyU 15218219).

References

[1] Bickel, P. J., Ritov, Y. & Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732 (2009).

[2] Cai, T. T., Xu, G. W. & Zhang, J. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory* **55** (7), 3388–3397 (2009).

[3] Candès, E. & Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory* **51**, 4203–4215 (2005).

[4] Beck, A. & Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** (1), 183–202 (2009).

[5] Daubechies, I., Defrise, M. & De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57**, 1413–1457 (2004).

[6] Hale, E. T., Yin, W. & Zhang, Y. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization* **19** (3), 1107–1130 (2008).

[7] Xiao, L. & Zhang, T. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization* **23** (2), 1062–1091 (2013).

[8] Chartrand, R. & Staneva, V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems* **24**, 1–14 (2008).

[9] Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** (456), 1348–1360 (2001).

[10] Hu, Y., Li, C., Meng, K., Qin, J. & Yang, X. Group sparse optimization via $\ell_{p,q}$ regularization. *Journal of Machine Learning Research* **18** (30), 1–52 (2017).

[11] Xu, Z., Chang, X., Xu, F. & Zhang, H. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems* **23**, 1013–1027 (2012).

[12] Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** (2), 894–942 (2010).

[13] Chen, X., Xu, F. & Ye, Y. Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. *SIAM Journal on Scientific Computing* **32** (5), 2832–2852 (2010).

[14] Locatelli, M. & Schoen, F. *Global Optimization: Theory, Algorithms, and Applications* (Mathematical Programming Society and Society for Industrial and Applied Mathematics, Philadelphia, 2013).

[15] Li, X., Hu, Y., Li, C., Yang, X. & Jiang, T. Sparse estimation via lower-order penalty optimization methods in high-dimensional linear regression. *Journal of Global Optimization* (2022).

[16] Liu, H., Yao, T., Li, R. & Ye, Y. Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical Programming* **166**, 207–240 (2017).

[17] Zhang, C.-H. & Zhang, T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27** (4), 576 – 593 (2012).

[18] Yang, X. Second-order global optimality conditions for convex composite optimization. *Mathematical Programming* **81**, 327–347 (1998).

[19] Yang, X. & Zhou, Y. Second-order analysis of penalty function. *Journal of Optimization Theory and Applications* **146** (2), 445–461 (2010).

[20] Blumensath, T. & Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis* **27** (3), 265–274 (2009).

[21] Foucart, S. & Rauhut, H. *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013).

[22] Wainwright, Martin J. *High-dimensional statistics: A non-asymptotic viewpoint* (Cambridge University Press, Cambridge, 2019).

[23] Lu, Z. Iterative reweighted minimization methods for ℓ_p regularized unconstrained nonlinear programming. *Mathematical Programming* **147** (1), 277–307 (2014).

[24] Attouch, H., Bolte, J. & Svaiter, B. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming* **137**, 91–129 (2013).

[25] Boţ, R. I., Csetnek, E. R. & Nguyen, D.-K. A proximal minimization algorithm for structured nonconvex and nonsmooth problems. *SIAM Journal on Optimization* **29** (2), 1300–1328 (2019).

[26] Wen, Z., Yin, W., Goldfarb, D. & Zhang, Y. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing* **32** (4), 1832–1857 (2010).

[27] Jiao, Y., Jin, B. & Lu, X. Iterative soft/hard thresholding with homotopy continuation for sparse recovery. *IEEE Signal Processing Letters* **24** (6), 784–788 (2017).

[28] Blumensath, T. & Davies, M. E. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing* **4** (2), 298–309 (2010).

[29] Zhao, Y.-B. Optimal k-thresholding algorithms for sparse optimization problems. *SIAM Journal on Optimization* **30** (1), 31–55 (2020).

[30] Needell, D. & Tropp, J. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic*

Analysis **26** (3), 301–321 (2009).

[31] Blanchard, J. D., Cartis, C. & Tanner, J. Compressed sensing: How sharp is the restricted isometry property? *SIAM Review* **53** (1), 105–125 (2011).

[32] Marjanovic, G. & Solo, V. On ℓ_p optimization and matrix completion. *IEEE Transactions on Signal Processing* **60** (11), 5714–5724 (2012).

[33] Mhenni, R. B., Bourguignon, S. & Ninin, J. Global optimization for sparse solution of least squares problems. *Optimization Methods and Software* **37** (5), 1740–1769 (2022).