Nonconvex Sorted ℓ_1 Minimization for Sparse Approximation

Xiaolin Huang · Lei Shi · Ming Yan

Abstract The ℓ_1 norm is the tight convex relaxation for the ℓ_0 "norm" and has been successfully applied for recovering sparse signals. However, for problems with fewer samples than required for accurate ℓ_1 recovery, one needs to apply nonconvex penalties such as ℓ_p "norm". As one method for solving ℓ_p minimization problems, iteratively reweighted ℓ_1 minimization updates the weight for each component based on the value of the same component at the previous iteration. It assigns large weights on small components in magnitude and small weights on large components in magnitude. The set of the weights is not fixed, and it makes the analysis of this method difficult. In this paper, we consider a weighted ℓ_1 penalty with the set of the weights fixed and the weights are assigned based on the sort of all the components in magnitude. The smallest weight is assigned to the largest component in magnitude. This new penalty is called nonconvex sorted ℓ_1. Then we propose two methods for solving nonconvex sorted ℓ_1 minimization problems: iteratively reweighted ℓ_1 minimization and iterative sorted thresholding, and prove that both methods will converge to a local minimizer of the nonconvex sorted ℓ_1 minimization problems. We also show that both methods are generalizations of iterative

X. Huang
Department of Electrical Engineering, KU Leuven, B-3001 Leuven, Belgium.
E-mail: huangxl06@mails.tsinghua.edu.cn

L. Shi
Shanghai Key Laboratory for Contemporary Applied Mathematics, School of Mathematical Sciences, Fudan University, Shanghai, 200433, P.R. China.
E-mail: leishi@fudan.edu.cn

M. Yan
Department of Mathematics, University of California, Los Angeles, CA 90095, USA.
E-mail: yanm@math.ucla.edu

support detection and iterative hard thresholding respectively. The numerical experiments demonstrate the better performance of assigning weights by sort compared to assigning by value.

Keywords Iteratively Reweighted ℓ_1 Minimization; Iterative Sorted Thresholding; Local Minimizer; Nonconvex Optimization; Sparse Approximation.

Mathematics Subject Classification (2000) 49M37 · 65K10 · 90C26 · 90C52

1 Introduction

Sparse approximation aims at recovering a sparse vector $\mathbf{u} \in \mathbf{R}^n$ from relatively few linear measurements $\mathbf{b} \in \mathbf{R}^m$. More precisely, we determine a sparse vector $\mathbf{u} \in \mathbf{R}^n$ from an underdetermined linear system

$$\mathbf{Au} = \mathbf{b},$$

where $\mathbf{A} \in \mathbf{R}^{m \times n}$ is the measurement matrix with $m < n$. Many problems such as signal and image compression, compressed sensing, and error correcting codes, can be modeled as sparse approximation problems [4,31]. In statistics, sparse approximation is also related to selecting the relevant explanatory variables, which is referred to as model selection [5].

Generally speaking, this underdetermined linear system has infinite solutions and we are interested in the sparest solution only. Finding the sparest vector amounts to solving the ℓ_0 minimization problem

$$\min_{\mathbf{u} \in \mathbf{R}^n} \|\mathbf{u}\|_0, \text{ subject to } \mathbf{Au} = \mathbf{b}, \tag{1.1}$$

where $\|\mathbf{u}\|_0$ is the number of nonzero components in $\mathbf{u}$.

Since the locations of the nonzero components are not available, solving problem (1.1) directly is NP-hard in general [28]. A family of iterative greedy algorithms, including orthogonal matching pursuit [34], CoSaMP [29], subspace pursuit [16], iterative hard thresholding [2], and hard thresholding pursuit [20], have been established to solve problem (1.1) with less computational complexity.

Another alternative is to consider the ℓ_1 minimization problem

$$\min_{\mathbf{u} \in \mathbf{R}^n} \|\mathbf{u}\|_1, \text{ subject to } \mathbf{Au} = \mathbf{b}. \tag{1.2}$$

The convex optimization problem (1.2), commonly known as basis pursuit [12], is an efficient relaxation for problem (1.1). The ℓ_1 minimization often leads to sparse solutions and the mechanism behind its performance has been theoretically analyzed. It is known from the compressed sensing literature that if $\mathbf{A}$ obeys some conditions, such as the restricted isometry property [6], the null space property [15], and the incoherence condition [33], problem (1.1) and its

convex relaxation (1.2) are equivalent. When there is noise in the measurement, the following basis pursuit denoising model is proposed in [32]:

$$\min_{\mathbf{u} \in \mathbf{R}^n} \|\mathbf{u}\|_1 + \alpha \|\mathbf{A}\mathbf{u} - \mathbf{b}\|^2. \tag{1.3}$$

Though ℓ_1 minimization is stable and has a number of theoretical results, it is not able to recover the sparest solutions in many real applications, e.g., computed tomography, where the sufficient conditions for exact recovery are not satisfied. The noncovex optimization is applied to enhance the sparsity of the solutions and improve the recovery performance. The mostly used nonconvex penalty is ℓ_p "norm" with $0 < p < 1$ [8,21,30,37], which connects ℓ_0 and ℓ_1. Other nonconvex penalties in the literature are: Smoothly Clipped Absolute Deviation (SCAD) [19], generalized shrinkage [9,10], etc. For almost all the nonconvex penalties used, the regularization terms penalize components with small magnitudes more than those with large magnitudes and are separable.

In this paper, we introduce a new nonconvex penalty called nonconvex sorted ℓ_1 penalty which penalizes different components in the solution according to their ranks in the absolute value. It is not separable and the set of weights is fixed. The contributions of this paper are summarized below.

- We introduce the nonconvex sorted ℓ_1 minimization to enhance the sparsity and improve recovery performance.
- We build the connection of the nonconvex sorted ℓ_1 to several existing penalties including ℓ_1 and K-sparsity.
- We propose an iteratively reweighted ℓ_1 minimization for solving the problems with nonconvex sorted ℓ_1 terms, and show that it converges in finite steps to a local optimum of the nonconvex sorted ℓ_1 minimization.
- We propose an iterative sorted thresholding method for the unconstrained denoising problem and show that it converges to a local optimum of the corresponding regularization problem. This iterative sorted thresholding is a generalization of iterative soft thresholding and iterative hard thresholding.

1.1 Notation

The following notations are used throughout the paper. For any dimension n, bold lowercase letters are used for vectors and lowercase letters with subscripts denote their components, e.g., $\mathbf{u} = (u_1, \cdots, u_n)^T \in \mathbf{R}^n$. Bold uppercase letters such as $\mathbf{A}$ and $\mathbf{P}$ are used for matrices. $\mathbf{I}_n$ stands for the $n \times n$ identity matrix, and $\mathbf{J}_n$ is the all-ones $n \times n$ matrix. For $\mathbf{u} \in \mathbf{R}^n$, the ℓ_p norm of $\mathbf{u}$ is $\|\mathbf{u}\|_p := \left(\sum_{i=1}^n |u_i|^p\right)^{1/p}$ for $0 < p < \infty$, with the usual extension $\|\mathbf{u}\|_0 := \#\{i : u_i \neq 0\}$ and $\|\mathbf{u}\|_\infty := \max_{1 \leq i \leq n} |u_i|$. Strictly speaking, $\|\cdot\|_0$ is not a real norm and $\|\cdot\|_p$ merely defines a quasi-norm when $0 < p < 1$. For simplicity, let $\|\cdot\|$ stands for $\|\cdot\|_2$. The indicator function $\iota_{\mathcal{C}}$ of the set $\mathcal{C}$ is defined as follows:

$$\iota_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{C}, \\ +\infty, & \text{if } \mathbf{x} \notin \mathcal{C}. \end{cases}$$

The component-wise multiplication of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as follows:

$$(\mathbf{x} \odot \mathbf{y})_i = x_i \times y_i, \quad \forall i \in [1, n].$$

1.2 Organization

The rest of the paper is presented as follows. The previous related works on nonconvex optimization for sparse approximation are given in section 2. We introduce the nonconvex sorted ℓ_1 minimization in section 3, and propose two methods for solving the problem with nonconvex sorted ℓ_1 terms in sections 4 and 5. The convergence results are shown in the corresponding sections. In section 6, we compare the performance of this nonconvex sorted ℓ_1 with other nonconvex approaches on compressed sensing problems for both noise-free and noisy cases. This paper is ended with a short conclusion section.

2 Previous Works

As shown in applications of sparse approximation, nonconvex minimizations have better performance than convex optimizations in enforcing sparsity. The ℓ_p minimization with $0 < p < 1$ is mostly used in literature as a bridge between ℓ_0 and ℓ_1. It was demonstrated that under certain restricted isometry properties, the ℓ_p minimization problem

$$\min_{\mathbf{u} \in \mathbf{R}^n} \|\mathbf{u}\|_p^p, \text{ subject to } \mathbf{A}\mathbf{u} = \mathbf{b} \tag{2.1}$$

recovers sparse vectors from fewer linear measurements than the ℓ_1 minimization (1.2) does [8,21]. Because of the nonconvexity, finding a global minimizer of problem (2.1) or its unconstrained variant is generally NP-hard [23,13].

The first group of methods for solving problem (2.1) is to solve a sequence of convex problems. The iteratively reweighted ℓ_1 minimization (IRL1) [7] is to iteratively solve the following weighted ℓ_1 minimization

$$\mathbf{u}^{l+1} = \arg \min_{\mathbf{u} \in \mathbf{R}^n} \left\{ \sum_{i=1}^n w_i^l |u_i|, \text{ subject to } \mathbf{A}\mathbf{u} = \mathbf{b} \right\}, \tag{2.2}$$

where $w_i^l = (|u_i^l| + \epsilon)^{p-1}$ and $\epsilon > 0$. A parallel approach to IRL1 is iteratively reweighted least squares (IRLS) [11,18], which iteratively solves the following weighted least square problem,

$$\mathbf{u}^{l+1} = \arg \min_{\mathbf{u} \in \mathbf{R}^n} \left\{ \sum_{i=1}^n w_i^l u_i^2, \text{ subject to } \mathbf{A}\mathbf{u} = \mathbf{b} \right\}, \tag{2.3}$$

with $w_i^l = \left((u_i^l)^2 + \epsilon\right)^{p/2-1}$ for $i = 1, \cdots, n$. When $\epsilon \to 0$, the objective functions in (2.2) and (2.3) are approximations to $\|\mathbf{u}\|_p^p$. The ϵ-regularization strategy, which starts with a relatively large ϵ and decreases ϵ at each iteration,

improves the performance of sparse recovery for both methods. Similar iteratively reweighted algorithms have been established in [26, 14] for unconstrained smoothed ℓ_p minimization.

Additionally, iterative thresholding algorithms have been established for unconstrained ℓ_p minimization problems [1, 37]. These algorithms can search a local minimizer starting from any initial point using the so-called shrinkage operator to generate a minimizing sequence such that the objective function is strictly decreasing along the sequence. The iterative thresholding algorithm is very efficient and particularly suitable for large-scale problems. But except for $p = 0, 1/2, 2/3, 1$, one can not give an explicit expression for the shrinkage operator which limits its applications in the ℓ_p minimization [37].

Except the nonconvex ℓ_p, other nonconvex penalties have been proposed. The SCAD penalty in statistics [19], of which the contour map is illustrated in Fig. 1, is defined as $R_{\text{SCAD}}(\mathbf{u}) = \sum_{i=1}^{n} r_{\text{SCAD}}(u_i)$, where

$$r_{\text{SCAD}}(u_i) = \begin{cases} a_1|u_i| & \text{if } |u_i| < a_1, \\ -\frac{a_1|u_i|^2 - 2a_1a_2|u_i| + a_1^3}{2(a_2-a_1)} & \text{if } a_1 \leq |u_i| \leq a_2, \\ \frac{a_1a_2+a_1^2}{2} & \text{if } |u_i| > a_2. \end{cases}$$

The ℓ_{1-2}, i.e., the difference between ℓ_1 and ℓ_2 norms, minimization is proposed in [39]. The difference of convex functions programming can be applied to solve those nonconvex optimization problems [22].

All the penalties mentioned above are separable, i.e., the same penalty function is applied on all the components of the vector, except that ℓ_{1-2} has a ℓ_2 term which couples all the components equally. In addition, there are a group of methods applying different penalty functions on different components based on their ranks in the absolute value. Iterative hard thresholding [2] puts weight 0 on the first K largest components and $+\infty$ on the other components; Iterative support detection [35] puts weight 0 on the first K largest components and 1 on the other components; The two-level ℓ_1 "norm" [25] puts a smaller positive weight on the first K largest components and a larger weight on the other components. In this paper, we generalize all these methods using more than two different weights, and propose iteratively reweighted ℓ_1 minimization and iterative sorted thresholding to solve problems with this generalized penalty.

3 Nonconvex Sorted ℓ_1 Minimization

This section introduces a new nonconvex minimization named *nonconvex sorted* ℓ_1 minimization. Assume $\mathbf{u} \in \mathbf{R}^n$. Let λ be a nondecreasing sequence of nonnegative numbers,

$$0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

with $\lambda_n > 0$. The nonconvex sorted ℓ_1 is defined as

$$R_\lambda(\mathbf{u}) = \lambda_1|\mathbf{u}|_{[1]} + \lambda_2|\mathbf{u}|_{[2]} + \cdots + \lambda_n|\mathbf{u}|_{[n]}, \tag{3.1}$$

where $|\mathbf{u}|_{[1]} \geq |\mathbf{u}|_{[2]} \geq \cdots \geq |\mathbf{u}|_{[n]}$ are the absolute values ranked in decreasing order. It is different from the sorted ℓ_1 norm proposed by Bogdan et.al. in [3], where λ is a nonincreasing sequence of nonnegative numbers, i.e., higher weights are assigned on components with larger absolute values. The contour map of the non-convex sorted ℓ_1 is illustrated in Fig. 1, along with those of ℓ_1 norm, SCAD, and ℓ_p "norm".
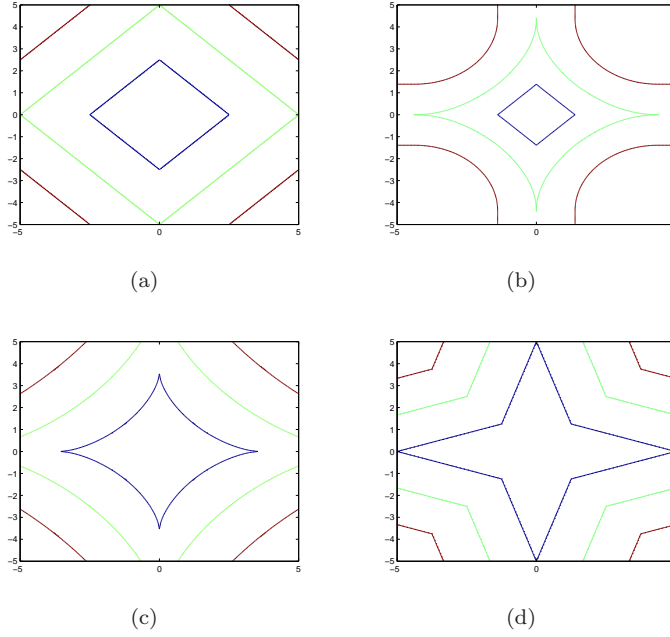

(a)

(b)

(c)

(d)

Fig. 1 The contour maps of several penalties: (a) the ℓ_1 norm; (b) the SCAD penalty with $a_1 = 1.1, a_2 = 3.7$; (c) the ℓ_p-norm with $p = 2/3$; (d) the nonconvex sorted ℓ_1 with $\lambda_1 = 1/3$ and $\lambda_2 = 1$.

First, we establish the connections of this nonconvex sorted ℓ_1 minimization to existing works.

- If $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 1$, it is the ℓ_1 norm of $\mathbf{u}$.
- If $\lambda_1 = \cdots = \lambda_K = 0$ and $\lambda_{K+1} = \cdots = \lambda_n = +\infty$ for some K satisfying $0 < K \leq n$, it is the indicator function for $\{\mathbf{u} : \|\mathbf{u}\|_0 \leq K\}$.
- If $\lambda_1 = \cdots = \lambda_K = 0$ and $\lambda_{K+1} = \cdots = \lambda_n = 1$ for some K satisfying $0 < K < n$, it corresponds to the iterative support detection in [35]. In [35], K can be changed adaptively during the iterations.
- If $\lambda_1 = \cdots = \lambda_K = w_1$ and $\lambda_{K+1} = \cdots = \lambda_n = w_2$ for $0 < w_1 < w_2 < \infty$, it is the two-level ℓ_1 "norm" in [25].
- If $\lambda_i = w_1 + w_2(i-1)$, where $w_1 \geq 0$ and $w_2 > 0$, it is the small magnitude penalized (SMAP) in [40].

In the following of this section, we introduce three equivalent formulations of the nonconvex sorted ℓ_1, which will help us to show the convergence results of the methods for solving problems with nonconvex sorted ℓ_1 terms in the following sections.

We introduce the following three functions with additional variables $\mathbf{P}$, $\mathbf{v}$, and Λ, respectively, and show that the nonconvex sorted ℓ_1 is equivalent to the minimum of those functions by eliminating the new variables with the corresponding constraints.

$$F_1(\mathbf{u}, \mathbf{P}) = \sum_{i=1}^{n} (\mathbf{P}\lambda)_i |u_i|,$$

$$F_2(\mathbf{u}, \mathbf{v}) = \lambda_1 \|\mathbf{u}\|_1 + (\lambda_2 - \lambda_1)\|\mathbf{u} - \mathbf{v}^1\|_1 + (\lambda_3 - \lambda_2)\|\mathbf{u} - \mathbf{v}^1 - \mathbf{v}^2\|_1$$

$$+ \cdots + (\lambda_n - \lambda_{n-1})\|\mathbf{u} - \sum_{j=1}^{n-1} \mathbf{v}^j\|_1,$$

$$F_3(\mathbf{u}, \Lambda) = \lambda_1\|\mathbf{u}\|_1 + (\lambda_2 - \lambda_1)\|\mathbf{u} \odot \Lambda^1\|_1 + (\lambda_3 - \lambda_2)\|\mathbf{u} \odot \Lambda^1 \odot \Lambda^2\|_1$$

$$+ \cdots + (\lambda_n - \lambda_{n-1})\|\mathbf{u} \odot \Lambda^1 \odot \Lambda^2 \odot \cdots \odot \Lambda^{n-1}\|_1,$$

where $\mathbf{v}$ is used to denote the combination of $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{n-1}$ for notational simplicity, and Λ denotes the combination of $\Lambda^1, \Lambda^2, \ldots, \Lambda^{n-1}$.

Theorem 1 *Minimizing $F_1(\mathbf{u}, \mathbf{P})$, $F_2(\mathbf{u}, \mathbf{v})$, and $F_3(\mathbf{u}, \Lambda)$ over $\mathbf{P}$, $\mathbf{v}$, and Λ with their corresponding constraints given below, respectively, eliminates $\mathbf{P}$, $\mathbf{v}$, and Λ and obtains the nonconvex sorted ℓ_1 defined in* (3.1).

$$R_\lambda(\mathbf{u}) = \min_{\mathbf{P} \in \mathcal{P}} F_1(\mathbf{u}, \mathbf{P})$$

$$= \min_{\{\mathbf{v}^j\}_{j=1}^{n-1} : \|\mathbf{v}^j\|_0 \leq 1} F_2(\mathbf{u}, \mathbf{v})$$

$$= \min_{\{\Lambda^j\}_{j=1}^{n-1} : \Lambda_i^j \in \{0,1\}, \sum_i \Lambda_i^j \geq n-1} F_3(\mathbf{u}, \Lambda),$$

where $\mathcal{P}$ is the set of all permutation matrices.

Proof Part 1: Without loss of generality, assume that $|u_1| \geq |u_2| \geq \cdots \geq |u_n|$. We show that $F_1(\mathbf{u}, \mathbf{P}) \geq F_1(\mathbf{u}, \mathbf{I}_n)$ for all $\mathbf{P} \in \mathcal{P}$, which means that $\min_{\mathbf{P}} F_1(\mathbf{u}, \mathbf{P}) = F_1(\mathbf{u}, \mathbf{I}_n) = R_\lambda(\mathbf{u})$.

Given $\mathbf{P} \in \mathcal{P}$. If $(\mathbf{P}\lambda)_1 \neq \lambda_1 = (\mathbf{P}\lambda)_k$ for some $k > 1$, we can exchange the 1st and the kth components of $\mathbf{P}\lambda$ and obtain $\mathbf{P}^1$ such that $(\mathbf{P}^1\lambda)_1 = \lambda_1$, otherwise let $\mathbf{P}^1 = \mathbf{P}$. We have

$$F_1(\mathbf{u}, \mathbf{P}^1) - F_1(\mathbf{u}, \mathbf{P}) = \lambda_1|u_1| + (\mathbf{P}\lambda)_1|u_k| - (\mathbf{P}\lambda)_1|u_1| - \lambda_1|u_k|$$

$$= [\lambda_1 - (\mathbf{P}\lambda)_1](|u_1| - |u_k|) \leq 0.$$

Similarly, for $j > 1$, we can find $\mathbf{P}^j$ such that $(\mathbf{P}^j\lambda)_i = \lambda_i$ for $i = 1, \cdots, j$ and $F_1(\mathbf{u}, \mathbf{P}^j) \leq F_1(\mathbf{u}, \mathbf{P}^{j-1}) \leq \cdots \leq F_1(\mathbf{u}, \mathbf{P})$. When $j = n$, we have $\mathbf{P}^n\lambda = \mathbf{I}_n\lambda$ and $F_1(\mathbf{u}, \mathbf{I}_n) \leq F_1(\mathbf{u}, \mathbf{P})$.

Part 2: We show the equivalence by two steps: 1) $F_2(\mathbf{u}, \mathbf{v}) \geq R_\lambda(\mathbf{u})$ for all $\mathbf{v}$'s satisfying $\|\mathbf{v}^j\|_0 \leq 1$ and 2) $F_2(\mathbf{u}, \bar{\mathbf{v}}) = R_\lambda(\mathbf{u})$ for some $\bar{\mathbf{v}}$ satisfying the constraint.

The constraint that $\|\mathbf{v}^j\|_0 \leq 1$ for all j gives us

$$\left\| \mathbf{u} - \sum_{j=1}^{k} \mathbf{v}^j \right\|_1 \geq \sum_{i=k+1}^{n} |\mathbf{u}|_{[i]},$$

for $1 \leq k \leq n-1$. Thus for any $\mathbf{v}$ satisfying the constraint $\|\mathbf{v}^j\|_0 \leq 1$, we have

$$F_2(\mathbf{u}, \mathbf{v}) \geq \lambda_1 \sum_{i=1}^{n} |\mathbf{u}|_{[i]} + (\lambda_2 - \lambda_1) \sum_{i=2}^{n} |\mathbf{u}|_{[i]} + (\lambda_3 - \lambda_2) \sum_{i=3}^{n} |\mathbf{u}|_{[i]}$$
$$+ \cdots + (\lambda_n - \lambda_{n-1}) |\mathbf{u}|_{[n]}$$
$$= \lambda_1 |\mathbf{u}|_{[1]} + \lambda_2 |\mathbf{u}|_{[2]} + \cdots + \lambda_n |\mathbf{u}|_{[n]} = R_\lambda(\mathbf{u}).$$

Let $\bar{\mathbf{v}}^j$ be the vector with all zeros except the component corresponding to the jth largest absolute value of $\mathbf{u}$ having the same value as $\mathbf{u}$. We have

$$\left\| \mathbf{u} - \sum_{j=1}^{k} \mathbf{v}^j \right\|_1 = \sum_{i=k+1}^{n} |\mathbf{u}|_{[i]},$$

for $1 \leq k \leq n-1$, which gives us $F_2(\mathbf{u}, \bar{\mathbf{v}}) = R_\lambda(\mathbf{u})$.

Part 3: The proof is similar to that in Part 2 and we omit it here.   □

Remark 1 The constraint $\Lambda_i^j \in \{0, 1\}$ in $F_3(\mathbf{u}, \Lambda)$ can be relaxed to $\Lambda_i^j \in [0, 1]$ without changing the equivalence result. We will use the exact relaxed constraint $\Lambda_i^j \in [0, 1]$ for Λ.

Remark 2 Define $\mathcal{V} = \{\mathbf{v} : \|\mathbf{v}^j\|_0 \leq 1\}$ and $\mathcal{L} = \{\Lambda : \Lambda_i^j \in [0, 1], \sum_i \Lambda_i^j \geq n-1\}$. Note that the set $\mathcal{P}$ only has finite number of points in $\mathbf{R}^{n \times n}$, thus is not continuous; The set $\mathcal{V}$ is continuous but non-convex; The set $\mathcal{L}$ is continuous and convex.

Remark 2 For each $\mathbf{P} \in \mathcal{P}$, we can construct a $\Lambda \in \mathcal{L}$ such that $F_1(\mathbf{u}, \mathbf{P}) = F_3(\mathbf{u}, \Lambda)$. Each column of $\mathbf{J}_n - \mathbf{P}$ corresponds to one vector Λ^j, i.e., Λ^j is the jth column of $\mathbf{J}_n - \mathbf{P}$. In addition, if $\mathbf{P} \in \mathcal{P}$ is optimal for minimizing $F_1(\mathbf{u}, \mathbf{P})$, the corresponding Λ is also the minimizer of $F_3(\mathbf{u}, \Lambda)$. Thus, we can consider $F_3(\mathbf{u}, \Lambda)$ as a relaxation to $F_1(\mathbf{u}, \mathbf{P})$.

Remark 3 For each $\mathbf{P} \in \mathcal{P}$, we can also construct a $\mathbf{v} \in \mathcal{V}$ such that $F_1(\mathbf{u}, \mathbf{P}) = F_2(\mathbf{u}, \mathbf{v})$. Choose the support of each column of $\mathbf{P}$ as the support for the corresponding $\mathbf{v}^j$, i.e, the supports of the jth column of $\mathbf{P}$ and $\mathbf{v}^j$ are the same. Then $F_2(\mathbf{u}, \mathbf{v}) = F_1(\mathbf{u}, \mathbf{P})$ when $\mathbf{v}^j$ is the projection of $\mathbf{u}$ onto the support of $\mathbf{v}^j$. In addition, if $\mathbf{P}$ is optimal for minimizing $F_1(\mathbf{u}, \mathbf{P})$, then $\mathbf{v}$ is also optimal for minimizing $F_2(\mathbf{u}, \mathbf{v})$. Thus $F_2(\mathbf{u}, \mathbf{v})$ is also a relaxation to $F_1(\mathbf{u}, \mathbf{P})$. It can be shown that if $\mathbf{v}$ is the minimizer for $F_2(\mathbf{u}, \mathbf{v})$, we can find an optimal $\mathbf{P}$ for minimizing $F_1(\mathbf{u}, \mathbf{P})$ from the support of $\{\mathbf{v}^j\}_{j=1}^{n-1}$.

If the ℓ_1 terms in (1.2) and (1.3) are replaced by the nonconvex sorted ℓ_1 terms, the basis pursuit problem (1.2) becomes

$$\underset{\mathbf{u}}{\text{minimize}} \ R_\lambda(\mathbf{u}) + \iota_{\{\mathbf{u}:\mathbf{Au}=\mathbf{b}\}}(\mathbf{u}),$$

and the unconstrained basis pursuit denoising problem (1.3) becomes

$$\underset{\mathbf{u}}{\text{minimize}} \ R_\lambda(\mathbf{u}) + \alpha\|\mathbf{Au} - \mathbf{b}\|^2. \tag{3.2}$$

Combining these two problems together into a more general problem

$$\underset{\mathbf{u}}{\text{minimize}} \ E(\mathbf{u}) := R_\lambda(\mathbf{u}) + L(\mathbf{u}), \tag{3.3}$$

where $L(\mathbf{u})$ can be any convex loss function. Except compressive sensing, this general problem has more applications in geophysics, image processing, sensor networks, and computer vision. The interested reader is referred to [4, 36] for a comprehensive review of these applications. In Theorem 1, we have shown equivalent formulations of $R_\lambda(\mathbf{u})$, which follows that problem (3.3) is equivalent to the following two problems:

$$\underset{\mathbf{u},\mathbf{P}\in\mathcal{P}}{\text{minimize}} \ E_1(\mathbf{u},\mathbf{P}) := F_1(\mathbf{u},\mathbf{P}) + L(\mathbf{u}), \tag{3.4}$$

$$\underset{\mathbf{u},\Lambda\in\mathcal{L}}{\text{minimize}} \ E_3(\mathbf{u},\Lambda) := F_3(\mathbf{u},\Lambda) + L(\mathbf{u}). \tag{3.5}$$

Two lemmas stating sufficient and necessary conditions for $\mathbf{u}^*$ being a local minimizer of $E(\mathbf{u})$ are introduced.

Lemma 1 *If $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$, then for any $\mathbf{P}^*$ minimizing $F_1(\mathbf{u}^*,\mathbf{P})$, $(\mathbf{u}^*,\Lambda^*)$ with Λ^* being constructed from $\mathbf{P}^*$ as in Remark 2 is a local minimizer of $E_3(\mathbf{u},\Lambda)$.*

Proof Since $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$, we can find $\epsilon > 0$ such that for all $\mathbf{u}$ satisfying $\|\mathbf{u}-\mathbf{u}^*\| < \epsilon$, we have $E(\mathbf{u}) \geq E(\mathbf{u}^*)$. For all $(\mathbf{u},\Lambda)$, satisfying $\|(\mathbf{u},\Lambda) - (\mathbf{u}^*,\Lambda^*)\| < \epsilon$, we have $\|\mathbf{u}-\mathbf{u}^*\| < \epsilon$. Thus

$$E_3(\mathbf{u},\Lambda) \geq E(\mathbf{u}) \geq E(\mathbf{u}^*) = E_3(\mathbf{u}^*,\Lambda^*),$$

for all $(\mathbf{u},\Lambda)$ such that $\|(\mathbf{u},\Lambda) - (\mathbf{u}^*,\Lambda^*)\| < \epsilon$. Then $(\mathbf{u}^*,\Lambda^*)$ is a local minimizer of $E_3(\mathbf{u},\Lambda)$. $\square$

Lemma 2 *Given fixed $\mathbf{u}^*$, if for all $\bar{\mathbf{P}} \in \mathcal{P}$ minimizing $E_1(\mathbf{u}^*,\mathbf{P})$, we also have $\mathbf{u}^*$ minimizing $E_1(\mathbf{u},\bar{\mathbf{P}})$, then $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$.*

Proof There exists $\epsilon > 0$ such that when $\|\mathbf{u}-\mathbf{u}^*\| < \epsilon$, we can always find the same $\mathbf{P} \in \mathcal{P}$ such that both $R_\lambda(\mathbf{u}) = F_1(\mathbf{u},\mathbf{P})$ and $R_\lambda(\mathbf{u}^*) = F_1(\mathbf{u}^*,\mathbf{P})$ are satisfied. Therefore, we have $E(\mathbf{u}^*) = R_\lambda(\mathbf{u}^*) + L(\mathbf{u}^*) = F_1(\mathbf{u}^*,\bar{\mathbf{P}}) + L(\mathbf{u}^*) \leq F_1(\mathbf{u},\bar{\mathbf{P}}) + L(\mathbf{u}) = R_\lambda(\mathbf{u}) + L(\mathbf{u}) = E(\mathbf{u})$. $\square$

With these two lemmas, we propose two algorithms for solving problem (3.3) with nonconvex sorted ℓ_1 term in the next two sections.

4 Iteratively Reweighted ℓ_1 Minimization

It is difficult to solve problem (3.3) directly because of the non-convexity of $R_\lambda(\mathbf{u})$. In this section, we apply the equivalence results in the previous section to solve problem (3.3) from minimizing (3.4) and (3.5). There are two variables in $E_1(\mathbf{u}, \mathbf{P})$ (or $E_3(\mathbf{u}, \Lambda)$) and we can apply the alternating minimization procedure to solve this problem because the problem is easy to solve with one of the two variables is fixed. Fix $\mathbf{u}$, the variable $\mathbf{P}$ (or Λ) can be obtained in closed-form; while the problem for $\mathbf{u}$ with $\mathbf{P}$ (or Λ) fixed can be formulated into a weighted ℓ_1 minimization problem which is convex and for which there are many existing solvers. The step of updating $\mathbf{P}$ (or Λ) is just updating the weights in the weighted ℓ_1 minimization, and this procedure is essentially an iteratively reweighted ℓ_1 minimization.

The set of all the weights is fixed and each reweighting is just a permutation of the weights. This is different from the iteratively reweighted ℓ_1 minimization in [7,10] where each weight is updated independently from the corresponding component of the previous result, i.e., no sorting is needed and the set of the weights is not fixed. The proposed iteratively reweighted ℓ_1 minimization for (3.3) is summarized in algorithm 1, whose convergence result is shown below.

Algorithm 1 Iteratively Reweighted ℓ_1 Minimization

Initialize λ, $\mathbf{u}^0$
  for $l = 0, 1, \cdots$ **do**
    Update $\mathbf{P}^l = \arg\min_{\mathbf{P}} F_1(\mathbf{u}^l, \mathbf{P})$ with an optimal $\mathbf{P}$ such that $\mathbf{P}^l$ is different from $\{\mathbf{P}^0,$
$\mathbf{P}^1, \cdots, \mathbf{P}^{l-1}\}$. If there is no optimal $\mathbf{P}$ satisfying this condition, break.
    Update $\mathbf{u}^{l+1} = \arg\min_{\mathbf{u}} E_1(\mathbf{u}, \mathbf{P}^l)$.
  end for

Remark 4 Because $F_1(\mathbf{u}, \mathbf{P}) = \sum_i (\mathbf{P}\lambda)_i |u_i|$, which depends on $\mathbf{P}\lambda$, not $\mathbf{P}$. If there are equivalent components in λ, then different $\mathbf{P}$'s may have the same $\mathbf{P}\lambda$. In the algorithm, we just choose the optimal $\mathbf{P}^l$ for minimizing $F_1(\mathbf{u}^l, \mathbf{P})$ such that $\mathbf{P}^l\lambda$ is different from $\{\mathbf{P}^0\lambda, \mathbf{P}^1\lambda, \cdots, \mathbf{P}^{l-1}\lambda\}$.

Remark 5 The initial $\mathbf{u}^0$ can be chosen as the output of the ℓ_1 minimization problem: $\min_{\mathbf{u}} \|\mathbf{u}\|_1 + L(\mathbf{u})$. When we use alternating minimization procedure on $F_3(\mathbf{u}, \Lambda)$, in order to minimize $F_3(\mathbf{u}^l, \Lambda)$, we can first find an optimal $\mathbf{P}^l$ from minimizing $F_1(\mathbf{u}^l, \mathbf{P})$ and then construct the corresponding optimal Λ^l using Remark 2.

We show that the proposed iteratively reweighted ℓ_1 minimization can obtain a local optimum of problem (3.3) in finite steps.

Theorem 2 *Algorithm 1 will converge in finite steps, and the output $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$. In addition, $(\mathbf{u}^*, \Lambda^*)$ is a local minimizer of $E_3(\mathbf{u}, \Lambda)$.*

Proof Since $\mathbf{P} \in \mathcal{P}$, there are only finite number of $\mathbf{P}$'s (the total number of different $\mathbf{P}$'s is $n!$), and the algorithm will stop in finite steps. Assume that the algorithm stops at step $\bar{l}$. There are two cases:

- Case I: $E_1(\mathbf{u}^{\bar{l}-1}, \mathbf{P}^{\bar{l}-2}) > E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}-1})$. In this case, there is no element in $\{\mathbf{P}^0, \mathbf{P}^1, \cdots, \mathbf{P}^{\bar{l}-2}\}$ being a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$ because if $\mathbf{P}^l$ is a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$ for some integer $l \in [0, \bar{l} - 2]$, we have

$$E(\mathbf{u}^{\bar{l}}) = E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^l) \geq E_1(\mathbf{u}^{l+1}, \mathbf{P}^l)$$
$$\geq E_1(\mathbf{u}^{\bar{l}-1}, \mathbf{P}^{\bar{l}-2}) > E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}-1}) \geq E(\mathbf{u}^{\bar{l}}).$$

  In addition, $\mathbf{P}^{\bar{l}-1}$ is a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$, otherwise, we can find an optimal $\mathbf{P}^{\bar{l}}$ such that $E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}}) = E(\mathbf{u}^{\bar{l}}) < E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}-1})$. Therefore, $\mathbf{P}^{\bar{l}-1}$ is the unique minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$, $\mathbf{u}^{\bar{l}}$ being a local minimizer of $E(\mathbf{u})$ comes from Lemma 2.
- Case II: $E_1(\mathbf{u}^{l_0-1}, \mathbf{P}^{l_0-2}) > E_1(\mathbf{u}^{l_0}, \mathbf{P}^{l_0-1}) = E_1(\mathbf{u}^{l_0+1}, \mathbf{P}^{l_0}) = \cdots = E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}-1})$ for $l_0 \leq \bar{l} - 1$. In the same way, we can show that there is no element in $\{\mathbf{P}^0, \mathbf{P}^1, \cdots, \mathbf{P}^{l_0-2}\}$ being a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$. In addition, $\mathbf{P}^{\bar{l}-1}$ is a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$. For $l \in [l_0 - 1, \bar{l} - 2]$, we have

$$E(\mathbf{u}^{\bar{l}}) = E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{\bar{l}-1}) = E_1(\mathbf{u}^{l+1}, \mathbf{P}^l) \leq E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^l).$$

  For any $l \in [l_0 - 1, \bar{l} - 2]$ such that the equality satisfies, $\mathbf{P}^l$ is a minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$, and $\mathbf{u}^{\bar{l}}$ is also a minimizer of $E_1(\mathbf{u}, \mathbf{P}^l)$. In addition, there is no more minimizer of $F_1(\mathbf{u}^{\bar{l}}, \mathbf{P})$. Then Lemma 2 tells us that $\mathbf{u}^{\bar{l}}$ is a local minimizer.

If $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$, then $(\mathbf{u}^*, \Lambda^*)$ being a local minimizer of $E_3(\mathbf{u}, \Lambda)$ follows from Lemma 1. $\qquad\square$

Notice that the function value may not be strictly decreasing, and the next theorem discusses the case when the function values of two consecutive iterations are the same.

Theorem 3 *If $E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = E_1(\mathbf{u}^l, \mathbf{P}^{l-1})$ for some $l \leq \bar{l}$, $(\mathbf{u}^{l-1}, \mathbf{P}^{l-2})$ is a coordinatewise minimum point of $E_1(\mathbf{u}, \mathbf{P})$, i.e., $\mathbf{u}^{l-1}$ is a minimizer of $E_1(\mathbf{u}, \mathbf{P}^{l-2})$ and $\mathbf{P}^{l-2}$ is a minimizer of $E_1(\mathbf{u}^{l-1}, \mathbf{P})$.*

Proof $E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = E_1(\mathbf{u}^l, \mathbf{P}^{l-1})$, together with the nonincreasing property of the algorithm

$$E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) \geq E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-1}) \geq E_1(\mathbf{u}^l, \mathbf{P}^{l-1}),$$

gives us

$$E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-1}) = E_1(\mathbf{u}^l, \mathbf{P}^{l-1}).$$

Thus

$$E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-1}) = \min_{\mathbf{P}} E_1(\mathbf{u}^{l-1}, \mathbf{P}).$$

In addition, the algorithm gives

$$E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = \min_{\mathbf{u}} E_1(\mathbf{u}, \mathbf{P}^{l-2}).$$

Thus $(\mathbf{u}^{l-1}, \mathbf{P}^{l-2})$ is a coordinatewise minimum point of $E_1(\mathbf{u}, \mathbf{P})$.   □

Remark 6 If there exists $l < \bar{l}$ such that $E_1(\mathbf{u}^{l-1}, \mathbf{P}^{l-2}) = E_1(\mathbf{u}^l, \mathbf{P}^{l-1}) > E_1(\mathbf{u}^{\bar{l}}, \mathbf{P}^{l-1})$, then $\mathbf{u}^{l-1}$ and $\mathbf{u}^l$ may not be local minimizers of $E(\mathbf{u})$.

5 Iterative Sorted Thresholding

In this section, we propose another iterative method for solving problem (3.3). There are several iterative thresholding algorithms for compressive sensing in literature: iterative shrinkage-thresholding for ℓ_1 [17], iterative hard thresholding [2], iterative half thresholding for $\ell_{1/2}$ [37]. In this section, we shall establish a thresholding algorithm for problem (3.3), which generalizes iterative shrinkage-thresholding and iterative hard thresholding.

Assume that $L : \mathbf{R}^n \to \mathbf{R}$ is a smooth, bounded below, convex function of type $C^{1,1}$, i.e., continuously differentiable with Lipschitz continuous gradient:

$$\|\nabla L(\mathbf{u}) - \nabla L(\mathbf{v})\| \le L_L \|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathbf{R}^n$. The iterative sorted thresholding is described in Algorithm 2, and its convergence is shown in Theorem 4.

Algorithm 2 Iteratively Sorted Thresholding

Initialize $\mathbf{u}^0$
for $l = 0, 1, \cdots$ **do**
    Find $\mathbf{u}^{l+1} = \arg\min_{\mathbf{u}} \beta R_\lambda(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - (\mathbf{u}^l - \beta \nabla L(\mathbf{u}^l))\|^2$.
end for

Remark 7 The iteration can be expressed as

$$\mathbf{u}^{l+1} \in \mathbf{prox}_{\beta R_\lambda}(\mathbf{u}^l - \beta \nabla L(\mathbf{u}^l))$$

and the proximal operator can be evaluated in closed form (see Lemma 3).

Lemma 3 *The proximal operator of R_λ can be evaluated as*

$$\mathbf{prox}_{R_\lambda}(\mathbf{x}) = \max(|\mathbf{x}| - \mathbf{P}\lambda, \mathbf{0}) \odot \text{sign}(\mathbf{x}),$$

for any $\mathbf{P} \in \mathcal{P}$ such that $R_\lambda(\mathbf{x}) = \sum_{i=1}^{n} (\mathbf{P}\lambda)_i |x_i|$. Here $\max$ and sign are both component-wise.

Proof Evaluating the proximal operator is equivalent to solving the problem

$$\underset{\mathbf{u}}{\text{minimize}}\ R_\lambda(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2, \tag{5.1}$$

which is the same as the following problem with additional variable $\mathbf{P}$:

$$\underset{\mathbf{u},\mathbf{P}}{\text{minimize}}\ \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|u_i| + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2. \tag{5.2}$$

To solve problem (5.2), we eliminate the variable $\mathbf{u}$ from the objective function.

$$
\begin{aligned}
&\min_{\mathbf{u}} \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|u_i| + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \\
&= \min_{\mathbf{u}} \sum_{i=1}^{n}\lambda_i|(\mathbf{P}^T\mathbf{u})_i| + \frac{1}{2}\sum_{i=1}^{n}((\mathbf{P}^T\mathbf{u})_i - (\mathbf{P}^T\mathbf{x})_i)^2 \\
&= \sum_{i=1}^{n} f_{\lambda_i}((\mathbf{P}^T\mathbf{x})_i),
\end{aligned}
$$

where $f_{\lambda_i}(t)$ is a function defined on $\mathbf{R}$ as the following,

$$f_{\lambda_i}(t) = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| < \lambda_i, \\ \lambda_i|t| - \frac{1}{2}\lambda_i^2, & \text{if } |t| \geq \lambda_i. \end{cases}$$

The optimal $\mathbf{u}$ is chosen to be

$$
\begin{aligned}
\mathbf{u} = \mathbf{P}(\mathbf{P}^T\mathbf{u}) &= \mathbf{P}[\max(|\mathbf{P}^T\mathbf{x}| - \lambda, \mathbf{0}) \odot \text{sign}(\mathbf{P}^T\mathbf{x})] \\
&= \max(|\mathbf{x}| - \mathbf{P}\lambda, \mathbf{0}) \odot \text{sign}(\mathbf{x}).
\end{aligned}
$$

Next, we shall find an optimal $\mathbf{P}$ for problem (5.2). If $\lambda_i < \lambda_j$, the function $f_{\lambda_i} - f_{\lambda_j}$ can be expressed as

$$f_{\lambda_i}(t) - f_{\lambda_j}(t) = \begin{cases} 0, & \text{if } |t| < \lambda_i, \\ -\frac{1}{2}(|t| - \lambda_i)^2, & \text{if } \lambda_i \leq |t| < \lambda_j, \\ -\frac{1}{2}(\lambda_j - \lambda_i)(2|t| - \lambda_i - \lambda_j), & \text{if } |t| \geq \lambda_j. \end{cases}$$

Therefore, the even function $f_{\lambda_i}(t) - f_{\lambda_j}(t)$ is decreasing when $t \geq 0$, i.e.,

$$f_{\lambda_i}(x_i) - f_{\lambda_j}(x_i) \geq f_{\lambda_i}(x_j) - f_{\lambda_j}(x_j) \tag{5.3}$$

if $|x_i| \leq |x_j|$. Rearranging (5.3) gives us

$$f_{\lambda_i}(x_i) + f_{\lambda_j}(x_j) \geq f_{\lambda_i}(x_j) + f_{\lambda_j}(x_i). \tag{5.4}$$

Applying the same technique in part I of the proof for Theorem 1, we show that $|\mathbf{P}^T\mathbf{x}|$ is in decreasing order for optimal $\mathbf{P}$'s. The definition of R_λ gives us that $R_\lambda(\mathbf{x}) = \sum_{i=1}^{n}\lambda_i|(\mathbf{P}^T\mathbf{x})_i| = \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|x_i|$ for optimal $\mathbf{P}$'s. $\qquad\square$

Remark 8 The proximal operator can be multi-valued. For different optimal **P**, the output can be different. For example, when $\lambda = (0, 1)$ and $\mathbf{x} = (1, 1)$, then both $(1, 0)$ and $(0, 1)$ are optimal for problem (5.1).

Before proving the convergence, we state and prove two lemmas.

Lemma 4 *If there exists* $\mathbf{u}^*$ *such that*

$$\mathbf{u}^* \in \mathbf{prox}_{R_\lambda}(\mathbf{w}^*), \tag{5.5}$$

then

- $|w_i^*| \geq |w_j^*|$ *if* $|u_i^*| > |u_j^*|$;
- $|w_i^*| = |w_j^*|$ *if* $|u_i^*| = |u_j^*| > 0$;
- *In addition, if* $|\mathbf{u}^*|$ *is in decreasing order, i.e.,* $|u_1^*| \geq |u_2^*| \geq \cdots \geq |u_{I-1}^*| > |u_I^*| = \cdots = |u_n^*| = 0$, *we have* $|w_1^*| \geq |w_2^*| \geq \cdots \geq |w_I^*|$, *and* $|w_i^*| \geq |w_j^*|$ *for all* $i < I$ *and* $j \geq I$.

Proof Assumption (5.5) ensures the existence of $\mathbf{P}^* \in \{\mathbf{P} \in \mathcal{P} : R_\lambda(\mathbf{w}^*) = \sum_{k=1}^n (\mathbf{P}\lambda)_k |w_k^*|\}$ such that

$$|\mathbf{u}^*| = \max(|\mathbf{w}^*| - \mathbf{P}^*\lambda, \mathbf{0}). \tag{5.6}$$

Part I: If $|u_i^*| > |u_j^*|$, (5.6) shows that

$$|w_i^*| - (\mathbf{P}^*\lambda)_i = |u_i^*| > |u_j^*| \geq |w_j^*| - (\mathbf{P}^*\lambda)_j. \tag{5.7}$$

Proof by contradiction: Assume $|w_i^*| < |w_j^*|$, then $(\mathbf{P}^*\lambda)_i \geq (\mathbf{P}^*\lambda)_j$. Therefore,

$$|w_i^*| - (\mathbf{P}^*\lambda)_i < |w_j^*| - (\mathbf{P}^*\lambda)_j,$$

which is a contradiction to (5.7). Thus $|w_i^*| \geq |w_j^*|$.

Part II: If $|u_i^*| = |u_j^*| > 0$, (5.6) shows that

$$|w_i^*| - (\mathbf{P}^*\lambda)_i = |u_i^*| = |u_j^*| = |w_j^*| - (\mathbf{P}^*\lambda)_j. \tag{5.8}$$

Proof by contradiction: Without loss of generality, assume $|w_i^*| < |w_j^*|$, we have $(\mathbf{P}^*\lambda)_i \geq (\mathbf{P}^*\lambda)_j$. Therefore,

$$|w_i^*| - (\mathbf{P}^*\lambda)_i < |w_j^*| - (\mathbf{P}^*\lambda)_j,$$

which is a contradiction to (5.8). Thus $|w_i^*| = |w_j^*|$.

Part III: Proof by contradiction: Assume $|w_i^*| < |w_{i+1}^*|$ for $i < I$, we have $(\mathbf{P}^*\lambda)_i \geq (\mathbf{P}^*\lambda)_{i+1}$. Then

$$|u_i^*| = |w_i^*| - (\mathbf{P}^*\lambda)_i < |w_{i+1}^*| - (\mathbf{P}^*\lambda)_{i+1} \leq |u_{i+1}^*|, \tag{5.9}$$

which a contradiction to the assumption that $|\mathbf{u}^*|$ is in decreasing order. Thus $|w_i^*| \geq |w_{i+1}^*|$ for all $i < I$.

In addition if $i < I$ and $j \geq I$, then $|u_i^*| > |u_j^*| = 0$, Part I gives us that $|w_i^*| \geq |w_j^*|$. $\qquad\square$

Lemma 5 *If there exists* $\mathbf{u}^*$ *such that*

$$\mathbf{u}^* \in \mathbf{prox}_{\beta R_\lambda}(\mathbf{u}^* - \beta \nabla L(\mathbf{u}^*)), \qquad (5.10)$$

then $\mathbf{u}^*$ *is a local minimizer of* $E(\mathbf{u})$.

Proof Without loss of generality, we assume that $|\mathbf{u}^*|$ is in decreasing order. If $|\mathbf{u}^*|$ is not in decreasing order, we can find $\mathbf{P} \in \mathcal{P}$ such that $|\mathbf{P}^T\mathbf{u}^*|$ is in decreasing order and reformulate function L accordingly. When $L(\mathbf{u}) = \alpha\|\mathbf{A}\mathbf{u} - \mathbf{b}\|^2$, we can just rearrange the columns of $\mathbf{A}$.

Let $\mathbf{w}^* = \mathbf{u}^* - \beta\nabla L(\mathbf{u}^*)$ and define two subsets of $\mathcal{P}$ as follows:

$$\mathcal{P}^* := \{\mathbf{P} \in \mathcal{P} : R_\lambda(\mathbf{w}^*) = \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|w_i^*|\},$$

$$\hat{\mathcal{P}} := \{\mathbf{P} \in \mathcal{P} : R_\lambda(\mathbf{u}^*) = \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|u_i^*|\}.$$

Assumption (5.10) ensures the existence of $\mathbf{P}^* \in \mathcal{P}^*$ such that

$$\mathbf{u}^* = \max(|\mathbf{w}^*| - \beta\mathbf{P}^*\lambda, \mathbf{0}) \odot \mathrm{sign}(\mathbf{w}^*). \qquad (5.11)$$

First of all, we show that $\mathbf{P}^* \in \hat{\mathcal{P}}$, i.e., $(\mathbf{P}^*\lambda)_i \leq (\mathbf{P}^*\lambda)_j$ if $|u_i^*| > |u_j^*|$. If $|u_i^*| > |u_j^*|$, (5.11) shows that

$$|w_i^*| - \beta(\mathbf{P}^*\lambda)_i = |u_i^*| > |u_j^*| \geq |w_j^*| - \beta(\mathbf{P}^*\lambda)_j. \qquad (5.12)$$

Proof by contradiction: Assume $(\mathbf{P}^*\lambda)_i > (\mathbf{P}^*\lambda)_j$, then $|w_i^*| \leq |w_j^*|$ because $\mathbf{P}^* \in \mathcal{P}^*$. Therefore,

$$|w_i^*| - \beta(\mathbf{P}^*\lambda)_i < |w_j^*| - \beta(\mathbf{P}^*\lambda)_j,$$

which is a contradiction to (5.12). Thus $\mathbf{P}^* \in \hat{\mathcal{P}}$.

According to Lemma 2, in order to show that $\mathbf{u}^*$ is a local minimizer of $E(\mathbf{u})$, we need to show that $\mathbf{u}^*$ minimizes $E_1(\mathbf{u}, \mathbf{P})$, where $\mathbf{P} \in \hat{\mathcal{P}}$ is optimal for $\min_{\mathbf{P}} E_1(\mathbf{u}^*, \mathbf{P})$. Since $E_1(\mathbf{u}^*, \mathbf{P}) = \sum_{i=1}^{n}(\mathbf{P}\lambda)_i|u_i^*|$ for any $\mathbf{P} \in \hat{\mathcal{P}}$, we shall show that

$$\mathbf{u}^* = \max(|\mathbf{w}^*| - \beta\mathbf{P}\lambda, \mathbf{0}) \odot \mathrm{sign}(\mathbf{w}^*) \qquad (5.13)$$

for all $\mathbf{P} \in \hat{\mathcal{P}}$.

Comparing (5.11) and (5.13), in order to prove the second one, we need to show

- $(\mathbf{P}\lambda)_i = (\mathbf{P}^*\lambda)_i$ for $i < I$ and $\mathbf{P} \in \hat{\mathcal{P}}$,
- $|w_j^*| - \beta(\mathbf{P}\lambda)_j \leq 0$ for $j \geq I$ and $\mathbf{P} \in \hat{\mathcal{P}}$,

where $I = \min\{i : |u_i^*| = 0\}$.

Firstly, we consider the case when $i < I$. Combining the results in Lemma 4 and $\mathbf{P}^* \in \hat{\mathcal{P}}$, we have

$$|u_1^*| \geq |u_2^*| \geq \cdots \geq |u_{I-1}^*| > |u_j^*| = 0,$$
$$|w_1^*| \geq |w_2^*| \geq \cdots \geq |w_{I-1}^*| \geq |w_j^*|,$$
$$(\mathbf{P}^*\lambda)_1 \leq (\mathbf{P}^*\lambda)_2 \leq \cdots \leq (\mathbf{P}^*\lambda)_{I-1} \leq (\mathbf{P}^*\lambda)_j,$$

for $j \geq I$. For any $\mathbf{P} \in \hat{\mathcal{P}}$ and $i < I \leq j$, $|u_i^*| > |u_j^*| = 0$ implies $(\mathbf{P}\lambda)_i \leq (\mathbf{P}\lambda)_j$. Therefore, the set $\{(\mathbf{P}\lambda)_i\}_{i=1}^{I-1}$ is the same for all $\mathbf{P} \in \hat{\mathcal{P}}$. We shall show that $(\mathbf{P}\lambda)_{i-1} \leq (\mathbf{P}\lambda)_i$ for all $i < I$ and $\mathbf{P} \in \hat{\mathcal{P}}$.

Proof by contradiction: Assume that i_0 is the first index such that $(\mathbf{P}\lambda)_{i_0} > (\mathbf{P}\lambda)_{i_0+1}$ for some $\mathbf{P} \in \hat{\mathcal{P}}$. Then there exists $i_1 < I$ such that $(\mathbf{P}\lambda)_{i_0} > (\mathbf{P}\lambda)_{i_1}$ and $|u_{i_0}^*| > |u_{i_1}^*|$, which contradicts $\mathbf{P} \in \hat{\mathcal{P}}$.

Secondly, we consider the case when $j \geq I$. $\mathbf{P}^* \in \mathcal{P}^*$ implies $\max_{j \geq I} |w_j^*| \leq \min_{j \geq I} (\beta \mathbf{P}^* \lambda)_j$. Thus for any $\mathbf{P} \in \hat{\mathcal{P}}$, we have $|w_j^*| \leq (\beta \mathbf{P}\lambda)_j$ for all $j \geq I$.   □

Theorem 4 *If $\beta < 1/L_L$, the iterative sorted thresholding converges to a local optimum of* (3.3).

Proof From the iterations, we have

$$E(\mathbf{u}^{l+1}) + \left(\frac{1}{2\beta} - \frac{L_L}{2}\right) \|\mathbf{u}^{l+1} - \mathbf{u}^l\|_2^2$$

$$\leq R_\lambda(\mathbf{u}^{l+1}) + L(\mathbf{u}^l) + (\mathbf{u}^{l+1} - \mathbf{u}^l)^T \nabla L(\mathbf{u}^l) + \frac{L_L}{2} \|\mathbf{u}^{l+1} - \mathbf{u}^l\|_2^2$$

$$+ \left(\frac{1}{2\beta} - \frac{L_L}{2}\right) \|\mathbf{u}^{l+1} - \mathbf{u}^l\|_2^2 \tag{5.14}$$

$$= R_\lambda(\mathbf{u}^{l+1}) + L(\mathbf{u}^l) + (\mathbf{u}^{l+1} - \mathbf{u}^l)^T \nabla L(\mathbf{u}^l) + \frac{1}{2\beta} \|\mathbf{u}^{l+1} - \mathbf{u}^l\|_2^2$$

$$= R_\lambda(\mathbf{u}^{l+1}) + L(\mathbf{u}^l) + \frac{1}{2\beta} \|\mathbf{u}^{l+1} - \mathbf{u}^l + \beta \nabla L(\mathbf{u}^l)\|_2^2 - \frac{\beta}{2} \|\nabla L(\mathbf{u}^l)\|_2^2$$

$$\leq R_\lambda(\mathbf{u}^l) + L(\mathbf{u}^l) + \frac{1}{2\beta} \|\mathbf{u}^l - \mathbf{u}^l + \beta \nabla L(\mathbf{u}^l)\|_2^2 - \frac{\beta}{2} \|\nabla L(\mathbf{u}^l)\|_2^2 \tag{5.15}$$

$$= R_\lambda(\mathbf{u}^l) + L(\mathbf{u}^l) = E(\mathbf{u}^l).$$

Here, (5.14) comes from the assumption that $L \in C^{1,1}$, and (5.15) holds because $\mathbf{u}^{l+1}$ is a minimizer of $\beta R_\lambda(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - (\mathbf{u}^l - \beta\nabla L(\mathbf{u}^l))\|_2^2$. Therefore $E(\mathbf{u}^l)$ is decreasing when $\beta < 1/L_L$. In addition, $E(\mathbf{u}^l)$ is bounded below. Then $E(\mathbf{u})$ converges to $E(\bar{\mathbf{u}})$, where $\bar{\mathbf{u}}$ is a fixed point, i.e., $\bar{\mathbf{u}} \in \mathbf{prox}_{\beta R_\lambda}(\bar{\mathbf{u}} - \beta\nabla L(\bar{\mathbf{u}}))$. Lemma 5 states that $\bar{\mathbf{u}}$ is a local minimizer of $E(\mathbf{u})$. Furthermore, all subsequence limit points of $\mathbf{u}^l$ are fixed points.   □

6 Numerical Experiment

We, in this section, illustrate the performance of the proposed nonconvex sorted ℓ_1 minimization. All the experiments are done in Matlab R2013a with Core i7-3.40 GHz and 16.0G RAM. As discussed previously, if we only give two different weights, i.e., $\lambda_1 = \cdots = \lambda_K = \omega_1$ and $\lambda_{K+1} = \cdots = \lambda_n = 1$, the nonconvex sorted ℓ_1 minimization becomes the two-level ℓ_1 minimization (2level) [25]. Furthermore, if one sets $\omega_1 = 0$, it reduces to the iterative support detection (ISD) [35] with fixed number of supports. The most general case is the multiple-level ℓ_1 minimization (mlevel) and we in this paper use the following strategy:

$$\lambda_i = \begin{cases} 1, & i > K, \\ e^{-r(K-i)/K}, & i \le K, \end{cases}$$

where r controls the rate of decreasing λ_i from 1 to 0. The aim of this experimental section is to test different weight setting methods. The comparative method is the iterative reweighted ℓ_1 minimization (IRL1) in [7], which is also nonconvex but the weights are set according to value not the sort.

The tuning parameters in nonconvex sorted ℓ_1 minimization, such as K, w_1, and r, are related to the sparsity and nonconvexity. Generally, with the decreasing of w_1 and the increasing of K (as long as $K < n/2$), the nonconvexity of R_λ increases. In practice, it is better to start from the ℓ_1 minimization, of which the result is denoted by $\mathbf{u}^0$, and increase the nonconvexity during the iterations. For 2level and mlevel, we fix K to be $\lfloor \|\mathbf{u}^0\|_0/3 \rfloor$, i.e., the largest integer smaller than one third of $\#\{i : u_i^0 \ne 0\}$, but change ω_1 for 2level, and r for mlevel as follows:

$$\omega_1^0 = 0.5 \quad \text{and} \quad \omega_1^l = \max\{0.1, 0.9\omega_1^{l-1}\}, \forall l \ge 1;$$
$$r^l = \min\{10, 0.15l\}, \forall l \ge 1.$$

For ISD, $w_1 = 0$ is fixed and we enhance the sparsity by increasing K following the strategy used in [35] [38]. For IRL1, the weight is set as

$$\lambda_i^l = \frac{1}{|x_i| + \max\{0.5^{l-1}, 0.5^8\}},$$

which is also related to the iteration count.

We first illustrate the phase transition diagram via nonconvex sorted ℓ_1 minimization. In this experiment, 100-dimensional s-sparse signals $\mathbf{u}$, i.e., $\mathbf{u}$ has only s nonzero components, and matrix $\mathbf{A}$ are randomly generated. The non-zero components of $\mathbf{u}$ and the elements of $\mathbf{A}$ come from the standard Gaussian distribution. Then we let $\mathbf{b} = \mathbf{A}\mathbf{u}$ and use Algorithm 1 to recover $\mathbf{u}$ from $\mathbf{b}$ and $\mathbf{A}$. Algorithm 1 involves a series of weighted ℓ_1 minimization, for which we apply YALL1 [38]. ISD and IRL1 also can be solved by YALL1. Those reweighted methods iteratively minimize the weighted ℓ_1 penalty and the maximum iteration is controlled by $l_{\max}$. Notice that in Theorem 1, the convergence of iteratively reweighted ℓ_1 minimization for nonconvex sorted

ℓ_1 minimization has been proved, but the maximal number of iterations is related to the elements in $\mathcal{P}$ and could be very large. Hence, in practice, we set a maximum iteration and the result is still satisfactory. IRL1 takes more time than other methods on the weighted ℓ_1 minimization problems because the dynamic range of the weights can be very high for IRL1 when the nonzero components are Gaussian distributed. Hence we set $l_{\max} = 3$ for IRL1 and $l_{\max} = 10$ for others. One will observe that even with this setting, the computational time of IRL1 is more than other methods.

After recovering the signal, we calculate the ℓ_∞ distance between $\mathbf{u}$ and the recovered signal. If the distance is smaller than 10^{-3}, we claim the signal is successfully recovered. In the phase transition diagram, we test 100 trials for each s and m and then show the results in Fig.2 for ISD, 2level, mlevel, and IRL1. For a clear comparison, we in Fig.2(d) plot 50% successful recovery lines for all these four methods.
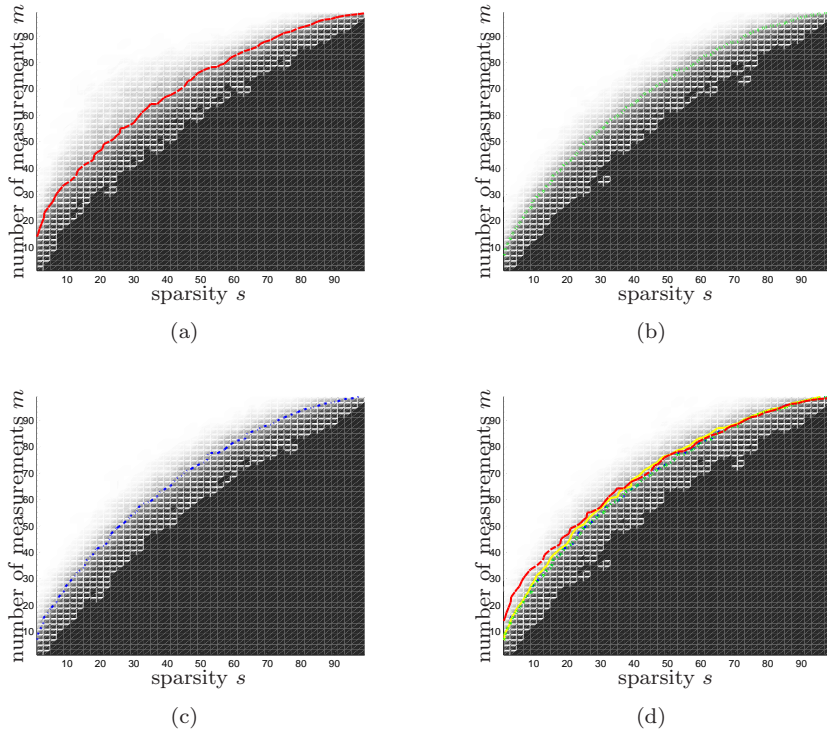
Fig. 2 Phase transition diagram for the considered algorithms: (a) ISD; (b) 2level; (c) mlevel; (d) IRL1. The grey level stands for the successful recovery percentage: white means 100% recovery and black means 0% recovery, the 50% recovery is also displayed by the green solid line. The red, blue, green and yellow lines show the 50% successful recovery by ISD, 2level, mlevel, and IRL1, respectively. In (d), the lines are shown together for comparison.

We also repeat the experiment used in [7], where a recovery problem with $m = 100$ and $n = 256$ is considered. $\mathbf{u}$ is a s-sparse signal and the recovery performance of different s values is evaluated. In Fig.3(a) and Fig.3(b), the recovery percentage of different methods and the computational time are shown, respectively. From both Fig.2 and Fig.3, one can find that compared with setting weights by value, setting weights according to the sort can enhance the sparse recovery performance.
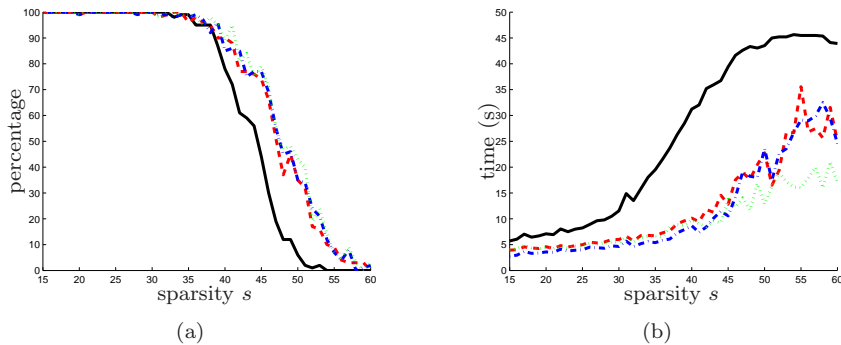

Fig. 3 Performance on signal recovery of ISD (red dashed line), 2level (blue dot-dashed line), mlevel (green dotted line), and IRL1 (black solid line): (a) recovery percentage; (b) average computational time.

Besides the above noise-free experiments, the algorithms are also tested on real-life electrocardiography (ECG) data. The ECG data come from the National Metrology Institute of Germany, which is online available in the PhysioNet [24] [27]. This data set has 15 signal channels and each channel contains 38400 data points. Notice that ECG signal is not sparse in the time domain and is sparse on the orthogonal Daubechies wavelets (db 10), of which the matrix is denoted by $\mathbf{\Psi}$. Then we start from the first 1024 data, denoted by $\mathbf{u}$ and randomly generate one Gaussian matrix $\mathbf{A} \in \mathbf{R}^{m \times n}$, where $n = 1024$ and m varies from 64 to 1024. Since $\mathbf{\Psi u}$ is sparse, we can apply the considered algorithms to recover the signal from $\mathbf{b} = \mathbf{A\Psi u}$. We calculate the mean of squared error between the recovered and the original signals, then we move to the next 1024 data.

In this experiment, ISD, 2level, mlevel, and IRL1 are evaluated. For 2level and mlevel, on the one hand, we can use Algorithm 1 and apply YALL1 to solve the involved weighted ℓ_1 minimization problems. On the other hand, we can also use the iterative sorted thresholding, i.e., Algorithm 2, to solve the unconstrained problems and evaluate their performance. Not like Algorithm 1, Algorithm 2 does not calculate $\mathbf{u}^0$ and hence we set K heuristically as $\lfloor n/5 \rfloor$.

For $m = 128, 256, 512$ and different algorithms, the mean squared error (MSE) and the corresponding mean computational time are reported in Tables 1 and 2, where the best ones in the view of MSE are underlined. From

the results, one can see that solving nonconvex sorted ℓ_1 minimization by Algorithm 1 provides accurate signals. When the signal length m increases, the iterative sorted thresholding becomes attractive due to its computational effectiveness, because only thresholding and matrix multiplication operations are involved. Compared with weighting by value, weighting by sort shows better performance on this experiment. Especially, when the compression ratio is high, the advantage of mlevel and 2level is significant. In that case, it is worthy designing a flexible and suitable weighting strategy. While, for the low compression ratio situation, we suggest ISD or 2level/mlevel solved by Algorithm 2.

7 Conclusion

The nonconvex sorted ℓ_1 minimization is proposed to enhance the sparse signal recovery. In this penalty, the set of the weights is fixed and the weights are assigned based on the ranks of the corresponding components among all the components in magnitude. Iteratively reweighted ℓ_1 minimization and iterative sorted thresholding are proposed to solve nonconvex sorted ℓ_1 minimization problems. Both methods are generalizations of existing methods. Iteratively reweighted ℓ_1 minimization is a generalization of iterative support detection, and iterative sorted thresholding is a generalization of iterative hard thresholding. Both methods are shown to converge to a local optimum of the corresponding sorted ℓ_1 minimization problem. The numerical experiments demonstrate the better performance of assigning weighted by sort.

References

1. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. Journal of Fourier Analysis and Applications **14**(5-6), 629–654 (2008) 5
2. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis **27**(3), 265 – 274 (2009) 2, 5, 12
3. Bogdan, M., van den Berg, E., Su, W., Candes, E.: Statistical estimation and testing via the sorted L1 norm. ArXiv e-prints (2013) 6
4. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review **51**(1), 34–81 (2009) 2, 9
5. Bühlmann, P., Van De Geer, S.: Statistics for high-dimensional data: methods, theory and applications. Springer (2011) 2
6. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. Information Theory, IEEE Transactions on **52**(2), 489–509 (2006) 2
7. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. Journal of Fourier Analysis and Applications **14**(5-6), 877–905 (2008) 4, 10, 17, 18
8. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. Signal Processing Letters, IEEE **14**(10), 707–710 (2007) 3, 4

Table 1 Mean Squared Error and Computational Time on ECG Data Sets

No.	m	IRL1	ISD	two-level		multi-level	
				Alg.1	Alg.2	Alg.1	Alg.2
1	128	4.805	3.429	3.493	4.147	<u>3.290</u>	4.006
		0.23 s	0.73 s	1.22 s	1.64 s	1.98 s	1.80 s
	256	1.791	1.295	1.302	1.393	<u>1.283</u>	1.394
		0.35 s	0.73 s	1.15 s	1.23 s	2.38 s	1.28 s
	512	0.935	0.836	<u>0.830</u>	0.857	<u>0.830</u>	0.853
		1.09 s	2.09 s	3.33 s	1.31 s	3.45 s	1.33 s
2	128	6.205	3.189	3.072	6.207	<u>2.673</u>	6.033
		0.24 s	0.55 s	0.98 s	1.76 s	1.90 s	1.85 s
	256	1.407	0.925	0.922	1.131	<u>0.903</u>	1.130
		0.31 s	0.52 s	0.93 s	1.37 s	1.89 s	1.41 s
	512	0.673	0.596	0.594	0.610	<u>0.587</u>	0.609
		1.09 s	1.18 s	2.11 s	1.63 s	2.11 s	1.63 s
3	128	6.188	3.897	3.692	6.482	<u>3.191</u>	6.281
		0.35 s	0.57 s	1.08 s	2.11 s	1.95 s	1.29 s
	256	1.732	1.085	1.128	1.412	<u>1.083</u>	1.418
		0.50 s	0.71 s	1.20 s	1.92 s	2.37 s	1.39 s
	512	0.756	0.678	0.682	0.688	<u>0.676</u>	0.686
		1.24 s	1.38 s	2.37 s	1.58 s	3.01 s	1.42 s
4	128	5.598	4.939	<u>4.689</u>	5.335	4.828	5.303
		0.43 s	0.60 s	1.11 s	1.01 s	1.69 s	1.24 s
	256	1.367	0.952	0.972	1.096	<u>0.949</u>	1.103
		0.39 s	0.64 s	0.84 s	0.96 s	1.76 s	0.90 s
	512	0.707	0.631	0.629	0.646	<u>0.624</u>	0.643
		1.20 s	1.95 s	3.11 s	1.19 s	4.14 s	1.20 s
5	128	4.598	3.099	3.027	3.288	<u>2.821</u>	3.151
		0.28 s	0.65 s	1.21 s	1.16 s	2.08 s	1.54 s
	256	1.628	1.120	1.133	1.311	<u>1.111</u>	1.306
		0.32 s	0.77 s	1.33 s	1.20 s	2.39 s	1.20 s
	512	0.787	0.704	<u>0.700</u>	0.720	<u>0.700</u>	0.718
		1.05 s	1.36 s	2.39 s	1.21 s	4.04 s	1.23 s
6	128	5.780	3.229	2.985	4.626	<u>2.770</u>	3.757
		0.25 s	0.59 s	1.05 s	1.27 s	1.38 s	1.27 s
	256	1.243	0.791	0.821	0.988	<u>0.787</u>	0.995
		0.32 s	0.51 s	0.86 s	0.89 s	2.00 s	0.95 s
	512	0.559	0.504	0.506	0.516	<u>0.500</u>	0.515
		0.87 s	1.12 s	1.92 s	1.00 s	3.15 s	1.02 s
7	128	5.376	3.666	3.715	4.889	<u>3.390</u>	4.712
		0.18 s	0.60 s	0.91 s	1.22 s	1.55 s	1.20 s
	256	1.181	<u>0.685</u>	0.730	1.047	0.707	1.054
		0.20 s	0.42 s	0.68 s	1.00 s	1.32 s	1.09 s
	512	0.410	0.395	0.401	<u>0.391</u>	0.396	0.392
		0.42 s	0.49 s	1.84 s	0.95 s	1.39 s	0.95 s

9. Chartrand, R.: Generalized shrinkage and penalty functions. In: IEEE Global Conference on Signal and Information Processing (2013) 3

10. Chartrand, R.: Shrinkage mappings and their induced penalty functions. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014, pp. 1026–1029 (2014) 3, 10

11. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE international conference on, pp. 3869–3872. IEEE (2008) 4

12. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing **20**(1), 33–61 (1998) 2

Table 2 Mean Squared Error and Computational Time on ECG Data Sets (cont.)

No.	m	IRL1	ISD	two-level Alg.1	Alg.2	multi-level Alg.1	Alg.2
8	128	7.397	3.993	4.302	5.903	<u>3.652</u>	5.612
		0.28 s	0.65 s	1.23 s	1.70 s	1.97 s	1.04 s
	256	1.219	<u>0.596</u>	0.652	0.810	0.613	0.890
		0.37 s	0.84 s	1.29 s	1.79 s	2.32 s	1.38 s
	512	0.414	0.379	0.388	0.386	<u>0.370</u>	0.386
		0.75 s	0.72 s	2.25 s	1.51 s	2.12 s	1.08 s
9	128	8.459	4.267	4.509	6.982	<u>3.859</u>	8.727
		0.21 s	0.49 s	0.96 s	1.60 s	1.67 s	1.67 s
	256	1.329	<u>0.685</u>	0.726	0.832	0.698	0.853
		0.31 s	0.74 s	1.20 s	1.89 s	2.03 s	1.91 s
	512	0.429	0.416	0.417	0.407	<u>0.400</u>	0.408
		0.49 s	0.63 s	0.79 s	1.52 s	1.31 s	1.55 s
10	128	4.685	3.390	3.287	5.151	<u>3.212</u>	3.888
		0.25 s	0.86 s	1.53 s	1.57 s	2.38 s	1.56 s
	256	1.154	0.635	0.654	0.665	<u>0.633</u>	0.657
		0.37 s	0.72 s	1.28 s	1.81 s	2.36 s	1.57 s
	512	0.427	0.386	0.392	0.399	<u>0.381</u>	0.398
		0.60 s	0.72 s	1.23 s	1.14 s	2.30 s	1.14 s
11	128	3.522	2.448	2.064	3.455	<u>1.937</u>	3.281
		0.20 s	0.50 s	0.95 s	1.06 s	1.52 s	1.12 s
	256	0.930	<u>0.588</u>	0.615	0.728	0.583	0.729
		0.28 s	0.56 s	1.28 s	0.77 s	1.87 s	0.79 s
	512	0.402	0.353	0.359	0.370	<u>0.351</u>	0.358
		0.75 s	1.12 s	2.02 s	0.70 s	3.94 s	0.69 s
12	128	3.159	<u>1.776</u>	1.921	2.471	1.820	2.386
		0.22 s	0.53 s	0.91 s	0.94 s	1.48 s	0.95 s
	256	0.775	0.502	0.503	0.576	<u>0.491</u>	0.577
		0.28 s	0.44 s	0.79 s	0.56 s	1.65 s	0.57 s
	512	0.348	0.308	0.310	0.322	<u>0.306</u>	0.321
		0.78 s	1.38 s	2.17 s	0.58 s	3.23 s	0.59 s
13	128	2.316	1.614	1.448	2.281	<u>1.417</u>	2.222
		0.25 s	0.58 s	0.95 s	0.84 s	1.64 s	0.90 s
	256	0.514	<u>0.304</u>	0.320	0.424	0.310	0.427
		0.37 s	0.67 s	1.25 s	0.72 s	1.94 s	0.70 s
	512	0.159	0.158	0.160	0.168	<u>0.156</u>	0.187
		0.60 s	0.55 s	0.95 s	0.54 s	1.62 s	0.55 s
14	128	4.169	2.558	2.344	2.621	<u>2.223</u>	3.417
		0.21 s	0.44 s	0.81 s	1.20 s	1.35 s	1.25 s
	256	0.958	0.568	0.588	0.647	<u>0.566</u>	0.750
		0.27 s	0.45 s	0.76 s	0.73 s	1.51 s	0.75 s
	512	0.410	0.355	0.359	0.377	<u>0.353</u>	0.377
		0.70 s	0.87 s	1.59 s	0.70 s	3.26 s	0.71 s
15	128	2.836	<u>1.754</u>	1.899	2.488	1.791	2.291
		0.20 s	0.40 s	0.73 s	1.01 s	1.33 s	1.07 s
	256	0.630	0.342	0.359	0.413	<u>0.337</u>	0.411
		0.27 s	0.52 s	0.83 s	0.61 s	1.54 s	0.61 s
	512	0.175	<u>0.174</u>	0.178	0.201	<u>0.174</u>	0.178
		0.48 s	0.43 s	0.79 s	0.52 s	1.32 s	0.52 s

13. Chen, X., Ge, D., Wang, Z., Ye, Y.: Complexity of unconstrained $l_2 - l_p$ minimization. Mathematical Programming **143**(1-2), 371–383 (2014) 4
14. Chen, X., Zhou, W.: Convergence of the reweighted ℓ_1 minimization algorithm for $\ell_2-\ell_p$ minimization. Computational Optimization and Applications (2013) 5
15. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. Journal of the American Mathematical Society **22**(1), 211–231 (2009) 2
16. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. Information Theory, IEEE Transactions on **55**(5), 2230–2249 (2009) 2
17. 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**(11), 1413–1457 (2004) 12
18. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics **63**(1), 1–38 (2010) 4
19. 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) 3, 5
20. Foucart, S.: Hard thresholding pursuit: An algorithm for compressive sensing. SIAM Journal on Numerical Analysis **49**(6), 2543–2563 (2011) 2
21. Foucart, S., Lai, M.J.: Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Applied and Computational Harmonic Analysis **26**(3), 395–407 (2009) 3, 4
22. Gasso, G., Rakotomamonjy, A., Canu, S.: Recovering sparse signals with a certain family of nonconvex penalties and DC programming. Signal Processing, IEEE Transactions on **57**(12), 4686–4698 (2009) 5
23. Ge, D., Jiang, X., Ye, Y.: A note on the complexity of l_p minimization. Mathematical Programming **129**(2), 285–299 (2011) 4
24. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000) 18
25. Huang, X., Liu, Y., Shi, L., Van Huffel, S., Suykens, J.A.K.: Two-level ℓ_1 minimization for compressed sensing. Signal Processing **108**, 459–475 (2015) 5, 6, 16
26. Lai, M.J., Xu, Y., Yin, W.: Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. SIAM Journal on Numerical Analysis **51**(2), 927–957 (2013) 5
27. Moody, G.B., Mark, R.G., Goldberger, A.L.: Physionet: a web-based resource for the study of physiologic signals. IEEE Engineering in Medicine and Biology Magazine **20**(3), 70–75 (2001) 18
28. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM Journal on Computing **24**(2), 227–234 (1995) 2
29. Needell, D., Tropp, J.A.: Cosamp: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis **26**(3), 301–321 (2009) 2
30. Saab, R., Yılmaz, Ö.: Sparse recovery by non-convex optimization–instance optimality. Applied and Computational Harmonic Analysis **29**(1), 30–48 (2010) 3
31. Starck, J.L., Murtagh, F., Fadili, J.M.: Sparse image and signal processing: wavelets, curvelets, morphological diversity. Cambridge University Press (2010) 2
32. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B **58**, 267–288 (1996) 3
33. Tropp, J.A.: Greed is good: Algorithmic results for sparse approximation. Information Theory, IEEE Transactions on **50**(10), 2231–2242 (2004) 2
34. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. Information Theory, IEEE Transactions on **53**(12), 4655–4666 (2007) 2
35. Wang, Y., Yin, W.: Sparse signal reconstruction via iterative support detection. SIAM Journal on Imaging Sciences **3**(3), 462–491 (2010) 5, 6, 16, 17

36. Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., Yan, S.: Sparse representation for computer vision and pattern recognition. Proceedings of the IEEE **98**(6), 1031–1044 (2010). DOI 10.1109/JPROC.2010.2044470 9

37. Xu, Z., Chang, X., Xu, F., Zhang, H.: $\ell_{1/2}$ regularization: A thresholding representation theory and a fast solver. Neural Networks and Learning Systems, IEEE Transactions on **23**(7), 1013–1027 (2012) 3, 5, 12

38. Yang, J., Zhang, Y.: Alternating direction algorithms for ℓ_1-problems in compressive sensing. SIAM Journal on Scientific Computing **33**(1), 250–278 (2011) 17

39. Yin, P., Lou, Y., He, Q., Xin, J.: Minimizaiton of $\ell_1 - \ell_2$ for compressed sensing. CAM-report 14-01, UCLA (2014) 5

40. Zeng, X., Figueiredo, M.: Decreasing weighted sorted ℓ_1 regularization. Signal Processing Letters, IEEE **21**(10), 1240–1244 (2014) 6