A Primal Dual Active Set with Continuation Algorithm for the ℓ^0-Regularized Optimization Problem

Yuling Jiao[*] Bangti Jin[†] Xiliang Lu[‡]

September 3, 2014

Abstract

We develop a primal dual active set with continuation algorithm for solving the ℓ^0-regularized least-squares problem that frequently arises in compressed sensing. The algorithm couples the the primal dual active set method with a continuation strategy on the regularization parameter. At each inner iteration, it first identifies the active set from both primal and dual variables, and then updates the primal variable by solving a (typically small) least-squares problem defined on the active set, from which the dual variable can be updated explicitly. Under certain conditions on the sensing matrix, i.e., mutual incoherence property or restricted isometry property, and the noise level, a finite step global convergence of the overall algorithm is established. Extensive numerical examples are presented to illustrate the efficiency and accuracy of the algorithm and its convergence behavior.

keywords: primal dual active set algorithm, coordinatewise minimizer, continuation strategy, global convergence.

1 Introduction

Over the last ten years, compressed sensing [9, 15] has received a lot of attention amongst engineers, statisticians and mathematicians due to its broad range of potential applications. Mathematically it can be formulated as the following constrained ℓ^0 optimization problem:

$$\begin{aligned} &\min_{x\in\mathbb{R}^p} \|x\|_0, \\ \text{subject to} \quad &\|\Psi x - y\|_2 \le \epsilon, \end{aligned} \tag{1.1}$$

where the sensing matrix $\Psi \in \mathbb{R}^{n\times p}$ with $p \gg n$ has normalized column vectors $\{\psi_i\}$, i.e.,

$$\|\psi_i\| = 1, \quad i = 1, \cdots, p,$$

$\epsilon \ge 0$ is the noise level, and $\|x\|_0$ denotes the number of nonzero components in the vector x. Due to the discrete structure of the term $\|x\|_0$, it is very challenging to develop an efficient algorithm to accurately solve the model (1.1). Hence, approximate methods for the model (1.1), especially greedy heuristics and convex relaxation, are very popular in practice. In greedy algorithms, including orthogonal matching pursuit [34], stagewise orthogonal matching pursuit [17], regularized orthogonal matching pursuit [30], CoSaMP [29], subspace pursuit [13], and greedy gradient pursuit [4, 7] etc., one first identifies the support of the sought-for signal, i.e., the locations of (one or more) nonzero components, iteratively based on the

[*] School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, 430063, P.R. China; School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China. (yulingjiaomath@whu.edu.cn)

[†] Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK (bangti.jin@gmail.com)

[‡] Corresponding author. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China (xllv.math@whu.edu.cn)

current dual variable (correlation), and then updates the components on the support by solving a least-squares problem. There are also several variants of greedy heuristics, e.g., (accelerated) iterative hard thresholding [3, 6] and hard thresholding pursuit [19], which are based on the sum of the current primal and dual variable. In contrast, basis pursuit finds one minimizer of a convex relaxation problem [11, 37], for which a wide variety of convex optimization algorithms can be conveniently applied; see [2, 12, 33, 39] for comprehensive overviews and the references therein.

Besides greedy methods and convex relaxation, the "Lagrange" counterpart of (1.1) (or equivalently, the ℓ^0-regularized minimization problem), which reads

$$\min_{x\in\mathbb{R}^p} J_\lambda(x) = \tfrac{1}{2}\|\Psi x - y\|^2 + \lambda\|x\|_0, \tag{1.2}$$

has been very popular in many applications, e.g., model selection, statistical regression, and image restoration. In the model (1.2), $\lambda > 0$ is a regularization parameter, controlling the sparsity level of the regularized solution. Due to the nonconvexity and discontinuity of the function $\|x\|_0$, the relation between problems (1.1) and (1.2) is not self evident. We shall show that under certain assumptions on the sensing matrix Ψ and the noise level ϵ (and with λ chosen properly), the support of the regularized solution to (1.2) coincides with that of the true signal, cf. Theorem 2.1.

The existence and a characterization of global minimizers to (1.2) were established in [26, 31]. However, it is still very challenging to develop globally convergent algorithms for efficiently solving problem (1.2) in view of its nonconvexity and nonsmoothness. Nonetheless, due to its broad range of applications, several algorithms have been developed to find an approximate solution to problem (1.2), including iterative hard thresholding [5], forward backward splitting [1], penalty decomposition [27] and stochastic continuation [35, 36], to name just a few. Theoretically, these algorithms can at best have a local convergence. Very recently, in [25, 26], based on a coordinatewise characterization of the global minimizers, a novel primal dual active set (PDAS) algorithm was developed to solve problem (1.2). The extensive simulation studies in [26] indicate that when coupled with a continuation technique, the PDAS algorithm merits a global convergence property. The idea of continuation is well established for iterative algorithms with the purpose of "warm starting" and globalizing the convergence. Unsurprisingly, this idea has been extensively pursued in sparsity optimization, especially the ℓ^1 penalty [43, 20, 44, 18]. The popular OMP [34] can be viewed as a continuation in the sparsity level of the solution, where one active set is added at each step; see also [14]. However, to the best of our knowledge, the application of the continuation technique to the PDAS algorithm for the ℓ^0 optimization problem (1.2) and its rigorous convergence analysis are new.

The PDAS algorithm solves the necessary optimality condition of a coordinatewise minimizer to (1.2), cf. (2.2) below, and thus it can at best converge to a coordinatewise minimizer. However, if the support of the coordinatewise minimizer is small and the sensing matrix Ψ satisfies certain mild conditions, then its active set is contained in the support of the true signal, cf. Lemma 2.4. Hence, the support of the minimizer will coincide with that of the true signal if we choose the regularization parameter λ properly (and thus control the size of the active set) during the iteration. This naturally motivates the use of a continuation strategy on the parameter λ. The resulting PDAS continuation (PDASC) algorithm extends the PDAS developed in [26]. In this work, we provide a convergence analysis of the PDASC algorithm under commonly used assumptions on the sensing matrix Ψ for the analysis of existing algorithms, i.e., mutual incoherence property and restricted isometry property. The convergence analysis relies essentially on a novel characterization of the evolution of the active set during the primal-dual active set iterations. To the best of our knowledge, this represents the first work on the global convergence of an algorithm for problem (1.2), without using a knowledge of the exact sparsity level.

The rest of the paper is organized as follows. In Section 2, we describe the problem setting, collect basic estimates, and provide refined properties of a coordinatewise minimizer. In Section 3, we give the complete algorithm, discuss the parameter choices, and provide a global convergence analysis. Finally, in Section 4, several numerical examples are provided to illustrate the efficiency of the algorithm and the convergence theory.

2 Regularized ℓ^0-minimization

In this section, we describe the problem setting, and derive basic estimates, which are essential for the convergence analysis. Further, we give sufficient conditions for a coordinatewise minimizer to be a global minimizer, which allows one to derive equivalence between (1.1) and (1.2), under certain circumstances.

2.1 Problem setting

Suppose that the true signal x^* has T nonzero components with its active set (indices of nonzero components) denoted by A^*, i.e., $T = |A^*|$ and the noisy data y is formed by

$$y = \sum_{i \in A^*} x_i^* \psi_i + \eta.$$

We assume that the noise vector η satisfies $\|\eta\| \leq \epsilon$, with $\epsilon \geq 0$ being the noise level. Further, we let

$$S = \{1, 2, ..., p\} \quad \text{and} \quad I^* = S \backslash A^*.$$

For any index set $A \subseteq S$, we denote by $x_A \in \mathbb{R}^{|A|}$ (respectively $\Psi_A \in \mathbb{R}^{n \times |A|}$) the subvector of x (respectively the submatrix of Ψ) whose indices (respectively column indices) appear in A. Last, we denote by x^o the oracle solution defined by

$$x^o = \Psi_{A^*}^\dagger y, \tag{2.1}$$

where $\Psi_A^\dagger$ denotes the pseudoinverse of the submatrix Ψ_A, i.e., $\Psi_A^\dagger = (\Psi_A^t \Psi_A)^{-1} \Psi_A^t$ if $\Psi_A^t \Psi_A$ is invertible. The oracle solution x^o is the least-squares solution on the true active set A^*.

In compressive sensing, there are two assumptions, i.e., mutual incoherence property (MIP) [16] and restricted isometry property (RIP) [10], on the sensing matrix Ψ that are frequently used for the convergence analysis of sparse recovery algorithms. The MIP assumes that the mutual coherence ν of the sensing matrix Ψ is small, where the mutual coherence ν of Ψ is defined by

$$\nu = \max_{1 \leq i,j \leq p, i \neq j} |\psi_i^t \psi_j|.$$

A sensing matrix Ψ is said to satisfy RIP of level s if there exists a constant $\delta \in (0, 1)$ such that

$$(1 - \delta)\|x\|^2 \leq \|\Psi x\|^2 \leq (1 + \delta)\|x\|^2, \ \forall x \in \mathbb{R}^p \text{ with } \|x\|_0 \leq s,$$

and we denote by δ_s the smallest constant with respect to the sparsity level s. We note that the mutual coherence ν can be easily computed, but the RIP constant δ_s is nontrivial to evaluate (see [38] for some recent results on its computational complexity). Nonetheless, the mutual coherence ν can be used to provide a simple upper bound on the RIP constant δ_s [41, Proposition 21]. However, our assumptions on the mutual coherence ν do not follow from that on the RIP constant δ, or vice versa, and thus we present theoretical results for both conditions.

The next lemma gives basic estimates under the MIP condition.

Lemma 2.1. *Let A and B be disjoint subsets of S. Then*

$$\begin{aligned} \|\Psi_A^t y\|_{\ell^\infty} &\leq \|y\|, \\ \|\Psi_B^t \Psi_A x_A\|_{\ell^\infty} &\leq |A|\nu \|x_A\|_{\ell^\infty}, \\ \|(\Psi_A^t \Psi_A)^{-1} x_A\|_{\ell^\infty} &\leq \frac{\|x_A\|_{\ell^\infty}}{1 - (|A| - 1)\nu} \quad \text{if} \quad (|A| - 1)\nu < 1. \end{aligned}$$

Proof. If $A = \emptyset$, then the estimates are trivial. Hence we will assume A is nonempty. For any $i \in A$,

$$|\psi_i^t y| \le \|\psi_i\| \|y\| \le \|y\|.$$

This shows the first inequality. Next, for any $i \in B$,

$$|\psi_i^t \Psi_A x_A| = |\sum_{j\in A} \psi_i^t \psi_j x_j| \le \sum_{j\in A} |\psi_i^t \psi_j| |x_j| \le |A| \nu \|x_A\|_{\ell^\infty}.$$

This shows the second assertion. To prove the last estimate, we follow the proof strategy of [40, Theorem 3.5], i.e., applying a Neumann series method. First we note that $\Psi_A^t \Psi_A$ has a unit diagonal because all columns of Ψ are normalized. So the off-diagonal part Φ satisfies

$$\Psi_A^* \Psi_A = E_{|A|} + \Phi,$$

where $E_{|A|}$ is an identity matrix. Each column of the matrix Φ lists the inner products between one column of Ψ_A and the remaining $|A| - 1$ columns. By the definition of the mutual coherence ν and the operator norm of a matrix

$$\|\Phi\|_{\ell^\infty,\ell^\infty} = \max_{k\in A} \sum_{j\in A\setminus\{k\}} |\psi_j^t \psi_k| \le (|A| - 1)\nu.$$

Whenever $\|\Phi\|_{\ell^\infty,\ell^\infty} < 1$, the Neumann series $\sum_{k=0}^\infty (-\Phi)^k$ converges to the inverse $(E_{|A|} + \Phi)^{-1}$. Hence, we may compute

$$\begin{aligned} \|(\Psi_A^* \Psi_A)^{-1}\|_{\ell^\infty,\ell^\infty} &= \|(E_{|A|} + \Phi)^{-1}\|_{\ell^\infty,\ell^\infty} = \|\sum_{k=0}^\infty (-\Phi)^k\|_{\ell^\infty,\ell^\infty} \\ &\le \sum_{k=0}^\infty \|\Phi\|_{\ell^\infty,\ell^\infty}^k = \frac{1}{1 - \|\Phi\|_{\ell^\infty,\ell^\infty}} \le \frac{1}{1 - (|A| - 1)\nu}. \end{aligned}$$

The desired estimate now follows immediately. □

The following lemma collects some well known estimates on the RIP constant δ_s; see [29, Propositions 3.1 and 3.2] and [13, Lemma 1] for the proofs.

Lemma 2.2. *Let A and B be disjoint subsets of S. Then*

$$\begin{aligned} \|\Psi_A^t \Psi_A x_A\| &\gtreqless (1 \mp \delta_{|A|}) \|x_A\|, \\ \|(\Psi_A^t \Psi_A)^{-1} x_A\| &\gtreqless \frac{1}{1 \pm \delta_{|A|}} \|x_A\|, \\ \|\Psi_A^t \Psi_B\| &\le \delta_{|A|+|B|}, \\ \|\Psi_A^\dagger y\| &\le \frac{1}{\sqrt{1 - \delta_{|A|}}} \|y\|, \\ \delta_s &\le \delta_{s'},\ \text{if } s < s'. \end{aligned}$$

The next lemma gives some crucial estimates for one-step primal dual active set iteration on the active set A. These estimates provide upper bounds on the dual variable $d = \Psi^t(y - \Psi x)$ and the error $\bar{x}_A = x_A - x_A^*$ on the active set A. They will play an essential role for subsequent analysis, including the convergence of the PDASC algorithm.

Lemma 2.3. *For any set $A \subseteq S$ with $|A| \le T$, let $B = A^* \setminus A$ and $I = S \setminus A$, and consider the following primal dual iteration on A*

$$x_A = \Psi_A^\dagger y, \quad x_I = 0, \quad d = \Psi^t(y - \Psi x).$$

Then the quantities $\bar{x}_A \equiv x_A - x_A^*$ and d satisfy the following estimates.

(a) If $\nu < 1/(T-1)$, then $d_A = 0$ and

$$
\begin{aligned}
\|\bar{x}_A\|_{\ell^\infty} &\leq \frac{1}{1-(|A|-1)\nu}\left(|B|\nu\|x_B^*\|_{\ell^\infty} + \epsilon\right), \\
|d_j| &\geq |x_j^*| - \|x_B^*\|_{\ell^\infty}(|B|-1)\nu - \epsilon - |A|\nu\|\bar{x}_A\|_{\ell^\infty}, \quad \forall j \in B, \\
|d_j| &\leq |B|\nu\|x_B^*\|_{\ell^\infty} + \epsilon + |A|\nu\|\bar{x}_A\|_{\ell^\infty}, \quad \forall j \in I^* \cap I.
\end{aligned}
$$

(b) If the RIP of sparsity level $s := \max\{|A|+|B|, T+1\}$ is satisfied, then $d_A = 0$ and

$$
\begin{aligned}
\|\bar{x}_A\| &\leq \frac{\delta_{|A|+|B|}}{1-\delta_{|A|}}\|x_B^*\| + \frac{1}{\sqrt{1-\delta_{|A|}}}\epsilon, \\
|d_j| &\geq |x_j^*| - \delta_{|B|}\|x_B^*\| - \epsilon - \delta_{|A|+1}\|\bar{x}_A\|, \quad \forall j \in B, \\
|d_j| &\leq \delta_{|B|+1}\|x_B^*\| + \epsilon + \delta_{|A|+1}\|\bar{x}_A\|, \quad \forall j \in I^* \cap I,
\end{aligned}
$$

Proof. We show only the estimates under the RIP condition and using Lemma 2.2, and that for the MIP condition follow similarly from Lemma 2.1. If $A = \emptyset$, then all the estimates clearly hold. In the case $A \neq \emptyset$, then by the assumption, $\Psi_A^t \Psi_A$ is invertible. By the definition of the update x_A and the data y we deduce that

$$d_A = \Psi_A^t(y - \Psi_A x_A) = 0,$$

and

$$
\begin{aligned}
\bar{x}_A &= (\Psi_A^t\Psi_A)^{-1}\Psi_A^t(\Psi_{A^*}x_{A^*}^* + \eta - \Psi_A x_A^*) \\
&= (\Psi_A^t\Psi_A)^{-1}\Psi_A^t(\Psi_B x_B^* + \eta).
\end{aligned}
$$

Consequently, by Lemma 2.2 and the triangle inequality, there holds

$$
\begin{aligned}
\|\bar{x}_A\| &\leq \frac{1}{1-\delta_{|A|}}\|\Psi_A^t\Psi_B x_B^*\| + \|\Psi_A^\dagger \eta\| \\
&\leq \frac{1}{1-\delta_{|A|}}\delta_{|A|+|B|}\|x_B^*\| + \frac{1}{\sqrt{1-\delta_{|A|}}}\epsilon.
\end{aligned}
$$

Next, it follows from the definition of the dual variable d, i.e.,

$$d_j = \psi_j^t(y - \Psi_A x_A) = \psi_j^t(\Psi_B x_B^* + \eta - \Psi_A \bar{x}_A),$$

Lemma 2.2, and the assumption $\psi_j^t\psi_j = 1$ that for any $j \in B$, there holds

$$
\begin{aligned}
|d_j| &= |\psi_j^t\psi_j x_j^* + \psi_j^t(\Psi_{B\setminus\{j\}}x_{B\setminus\{j\}}^* + \eta - \Psi_A\bar{x}_A)| \\
&\geq |x_j^*| - (|\psi_j^t\Psi_{B\setminus\{j\}}x_{B\setminus\{j\}}^*| + |\psi_j^t\eta| + |\psi_j^t\Psi_A\bar{x}_A|) \\
&\geq |x_j^*| - \delta_{|B|}\|x_B^*\| - \epsilon - \delta_{|A|+1}\|\bar{x}_A\|.
\end{aligned}
$$

Similarly, for any $j \in I^* \cap I$, there holds

$$|d_j| \leq \delta_{|B|+1}\|x_B^*\| + \epsilon + \delta_{|A|+1}\|\bar{x}_A\|.$$

This completes the proof of the lemma. □

2.2 Coordinatewise minimizer

Due to the nonconvexity and discontinuity of the function $\|x\|_0$, the classical theory [24] on the existence of a Lagrange multiplier cannot be applied directly to show the equivalence between the constrained problem (1.1) and the Lagrange counterpart (1.2). Nonetheless, both formulations aim at recovering the true sparse signal x^*, and thus we expect that they are closely related to each other. We shall establish below that under certain circumstances (with the regularization parameter λ properly chosen) the oracle solution x^o is the only global minimizer of problem (1.2), and as a consequence, we derive directly the equivalence between problems (1.1) and (1.2).

To this end, we first characterize minimizers of problem (1.2). Since the cost function $J_\lambda(x)$ is nonconvex and discontinuous, instead of a global minimizer, we study its coordinatewise minimizers, which was introduced in [42]. A vector $x = (x_1, x_2, \ldots, x_p)^t \in \mathbb{R}^p$ is called a coordinatewise minimizer to $J_\lambda(x)$ if it is the minimum along each coordinate direction, i.e.,

$$x_i \in \arg\min_{t\in\mathbb{R}} J_\lambda(x_1, ..., x_{i-1}, t, x_{i+1}, ..., x_p).$$

The necessary and sufficient condition for a coordinatewise minimizer x to $J_\lambda(x)$ is given by [25, 26]:

$$x_i \in S_\lambda^{\ell^0}(x_i + d_i) \quad \forall i \in S, \tag{2.2}$$

where $d = \Psi^t(y - \Psi x)$ denotes the dual variable, and $S_\lambda^{\ell^0}$ is the hard thresholding operator defined by

$$S_\lambda^{\ell^0}(v) \begin{cases} = 0, & |v| < \sqrt{2\lambda}, \\ \in \{0, \mathrm{sgn}(v)\sqrt{2\lambda}\}, & |v| = \sqrt{2\lambda}, \\ = v, & |v| > \sqrt{2\lambda}. \end{cases} \tag{2.3}$$

Using the operator $S_\lambda^{\ell^0}$, the condition (2.2) can be equivalently written as

$$\begin{cases} |x_i + d_i| > \sqrt{2\lambda} \Rightarrow d_i = 0, \\ |x_i + d_i| < \sqrt{2\lambda} \Rightarrow x_i = 0, \\ |x_i + d_i| = \sqrt{2\lambda} \Rightarrow x_i = 0 \text{ or } d_i = 0. \end{cases}$$

We note that these conditions as the optimality condition for the stationary point to (1.2) are well known in the literature (see e.g., [5]). In [5], it forms the basis of the iterative hard thresholding algorithm. Consequently, with the active set $A = \{i : x_i \neq 0\}$, there holds

$$\min_{i\in A} |x_i| \geq \sqrt{2\lambda} \geq \|d\|_{\ell^\infty}. \tag{2.4}$$

It is known that any coordinatewise minimizer x for problem (1.2) is a local minimizer [26, Theorem 3.1 (i)] [31, Section 3.1]. To further analyze the coordinatewise minimizer, we need the following assumption on the noise level ϵ:

Assumption 2.1. *The noise level ϵ is small in the sense $\epsilon \leq \beta \min_{i\in A^*} |x_i^*|$, for some $0 \leq \beta < 1/2$.*

The next lemma gives an interesting characterization of the active set of the coordinatewise minimizer.

Lemma 2.4. *Let Assumption 2.1 hold, and x be a coordinatewise minimizer with a support A and $|A| \leq T$. If either (a) $\nu < (1-2\beta)/(3T-1)$ or (b) $\delta \triangleq \delta_{2T} \leq (1-2\beta)/(2\sqrt{T}+1)$ holds, then $A \subseteq A^*$.*

Proof. Let $I = S \setminus A$. Since x is a coordinatewise minimizer, it follows from (2.2) that

$$x_A = \Psi_A^\dagger y, \quad x_I = 0, \quad d = \Psi^t(y - \Psi x).$$

We shall prove the assertions by means of contradiction. Assume the contrary, i.e., $A \not\subseteq A^*$. We let $B = A^* \setminus A$, which is nonempty by assumption, and denote by $i_A \in \{i \in I : |x_i^*| = \|x_B^*\|_{\ell^\infty}\}$. Then $i_A \in B$. Further by (2.4), there holds

$$|d_{i_A}| \leq \|d\|_{\ell^\infty} \leq \min_{i\in A} |x_i| \leq \min_{i\in A\setminus A^*} |x_i| \leq \|\bar{x}_A\|_{\ell^\infty} \leq \|\bar{x}_A\|. \tag{2.5}$$

Now we discuss the two cases separately.
Case (a). By Lemma 2.1, $\epsilon \le \beta \min_{i\in A^*} |x_i^*| \le \beta \|x_B^*\|_{\ell^\infty}$ from Assumption 2.1 and the choice of the index i_A, we have

$$\begin{aligned}
\|\bar{x}_A\|_{\ell^\infty} &\le \frac{1}{1-(|A|-1)\nu}\left(|B|\nu\|x_B^*\|_{\ell^\infty}+\epsilon\right)\\
&\le \frac{1}{1-(|A|-1)\nu}(|B|\nu+\beta)\|x_B^*\|_{\ell^\infty},\\
|d_{i_A}| &\ge |x_{i_A}^*| - \|x_B^*\|_{\ell^\infty}(|B|-1)\nu - \epsilon - |A|\nu\|\bar{x}_A\|_{\ell^\infty}\\
&\ge \|x_B^*\|_{\ell^\infty}\left(1-(|B|-1)\nu-\beta-|A|\nu\frac{1}{1-(|A|-1)\nu}(|B|\nu+\beta)\right).
\end{aligned}$$

Consequently, we deduce

$$\begin{aligned}
|d_{i_A}| - \|\bar{x}_A\|_{\ell^\infty} &\ge \frac{\|x_B^*\|_{\ell^\infty}}{1-(|A|-1)\nu}\left[1-(|A|+2|B|)\nu-(|A|+|B|)\nu^2+2\nu+\nu^2-\beta(\nu+2)\right]\\
&\ge \frac{\|x_B^*\|_{\ell^\infty}}{1-(|A|-1)\nu}\left[1-3T\nu+\nu-2\beta+\nu(1-\beta-2T\nu)\right]\\
&\ge \frac{\|x_B^*\|_{\ell^\infty}}{1-(|A|-1)\nu}\left[1-(3T-1)\nu-2\beta\right]>0,
\end{aligned}$$

under assumption (a) $\nu < (1-2\beta)/(3T-1)$. This leads to a contradiction to (2.5).
Case (b). By assumption, $|A|+|B| \le 2T$ and by Lemma 2.2, there hold

$$\begin{aligned}
\|\bar{x}_A\| &\le \frac{\delta}{1-\delta}\|x_B^*\| + \frac{1}{\sqrt{1-\delta}}\epsilon\\
&\le \frac{\delta}{1-\delta}\|x_B^*\| + \frac{1}{1-\delta}\epsilon,\\
|d_{i_A}| &\ge |x_{i_A}^*| - \delta\|x_B^*\| - \epsilon - \delta\|\bar{x}_A\|\\
&\ge |x_{i_A}^*| - \frac{\delta}{1-\delta}\|x_B^*\| - \frac{1}{1-\delta}\epsilon.
\end{aligned}$$

Consequently, with the assumption on ϵ and $\delta < \frac{1-2\beta}{2\sqrt{T}+1}$, we get

$$\begin{aligned}
|d_{i_A}| - \|\bar{x}_A\| &\ge |x_{i_A}^*| - \frac{2\delta}{1-\delta}\|x_B^*\| - \frac{2}{1-\delta}\epsilon\\
&\ge |x_{i_A}^*|\left(1-\frac{2\sqrt{T}\delta+2\beta}{1-\delta}\right)>0,
\end{aligned}$$

which is also a contradiction to (2.5). This completes the proof of the lemma. □

From Lemma 2.4, if the support size of the active set A of a coordinatewise minimizer x can be controlled, then we may obtain information of the true active set A^*: it is a superset of A. However, a local minimizer generally does not yield such information; see following result. Hence the coordinatewise minimizer is more informative. The proof can be found also in [31], but we include a short sketch here for completeness.

Proposition 2.1. *for any given index set $A \subseteq S$, the solution x to the least-squares problem $\min_{\mathrm{supp}(x)\subseteq A}\|\Psi x - y\|$ is a local minimizer.*

Proof. Let $\tau = \min\{|x_i| : x_i \neq 0\}$. Then for any small perturbation h in the sense $\|h\|_{\ell^\infty} < \tau$, we have $x_i \neq 0 \to x_i + h_i \neq 0$. Now we show that x is a local minimizer. To see this, we consider two cases, i.e., $\mathrm{supp}(h) \subseteq A$ and $\mathrm{supp}(h) \nsubseteq A$. First consider the case $\mathrm{supp}(h) \subseteq A$. By the definition of x, and $\|x\|_0 \leq \|x+h\|_0$, we deduce

$$\begin{aligned} J_\lambda(x+h) &= \tfrac{1}{2}\|\Psi(x+h)-y\|^2 + \lambda\|x+h\|_0 \\ &\geq \tfrac{1}{2}\|\Psi x - y\|^2 + \lambda\|x\|_0 = J_\lambda(x). \end{aligned}$$

Alternatively, if $\mathrm{supp}(h) \nsubseteq A$, then $\|x+h\|_0 \geq \|x\|_0 + 1$. Since

$$\lim_{\|h\|\to 0} \|\Psi(x+h)-y\| = \|\Psi x - y\|,$$

we again have $J_\lambda(x+h) > J_\lambda(x)$ for sufficiently small h. This completes the proof of the proposition. □

Now we can study global minimizers to problem (1.2). For any $\lambda > 0$, there exists a global minimizer x_λ to problem (1.2) [26]. Further, the following monotonicity relation holds [23][22, Section 3.2].

Lemma 2.5. *For $\lambda_1 > \lambda_2 > 0$, there holds $\|x_{\lambda_1}\|_0 \leq \|x_{\lambda_2}\|_0$.*

If the noise level ϵ is sufficiently small, and the parameter λ is properly chosen, the oracle solution x^o is the only global minimizer to $J_\lambda(x)$, cf. Theorem 2.1, which in particular implies the equivalence between the two formulations (1.1) and (1.2); see Remark 2.1 below.

Theorem 2.1. *Let Assumption 2.1 hold.*

(a) Suppose $\nu < (1-2\beta)/(3T-1)$ and $\beta \leq (1-2(T-1)\nu)/(T+3)$, and let

$$\xi = \frac{1-2(T-1)\nu - 2\beta - \beta^2}{2T} \min_{i\in A^*} |x_i^*|^2.$$

Then for any $\lambda \in (\epsilon^2/2, \xi)$, the oracle solution x^o, cf. (2.1), is the only global minimizer to $J_\lambda(x)$.

(b) Suppose $\delta \triangleq \delta_{2T} \leq (1-2\beta)/(2\sqrt{T}+1)$ and $\beta \leq (1-2\delta-\delta^2)/4$, and let

$$\xi = \left[\frac{1}{2}(1-\delta) - \frac{\delta^2}{1-\delta} - \frac{\beta}{\sqrt{1-\delta}} - \frac{1}{2}\beta^2\right] \min_{i\in A^*} |x_i^*|^2.$$

Then for any $\lambda \in (\epsilon^2/2, \xi)$, the oracle solution x^o, cf. (2.1), is the only global minimizer to $J_\lambda(x)$.

Proof. Let x be a global minimizer to problem (1.2), and its support be A. It suffices to show $A = A^*$. If $|A| \geq T+1$, then by the choice of λ, we deduce

$$J_\lambda(x) \geq \lambda(T+1) > \lambda T + \tfrac{1}{2}\epsilon^2 \geq J_\lambda(x^o),$$

which contradicts the minimizing property of x. Hence, $|A| \leq T$. Since a global minimizer is always a coordinatewise minimizer, by Lemma 2.4, we deduce $A \subseteq A^*$. If $A \neq A^*$, then $B = A^* \backslash A$ is nonempty. By the global minimizing property of x, there holds $x = \Psi_A^\dagger y$. Using the notation $\bar{x}_A$ from Lemma 2.3, we have

$$J_\lambda(x) = \tfrac{1}{2}\|\Psi_B x_B^* + \eta - \Psi_A \bar{x}_A\|^2 + \lambda|A|. \tag{2.6}$$

Now we consider the cases of the MIP and RIP separately.
Case (a) MIP: Let $i_A \in \{i \in I : |x_i^*| = \|x_B^*\|_{\ell^\infty}\}$, then $i_A \in B$ and $|x_{i_A}^*| = \|x_B^*\|_{\ell^\infty}$. Hence, by Lemmas 2.3 and 2.1, there holds

$$\begin{aligned} &\tfrac{1}{2}\|\psi_{i_A} x_{i_A}^* + \Psi_{B\backslash\{i_A\}} x_{B\backslash\{i_A\}}^* + \eta - \Psi_A \bar{x}_A\|^2 \\ &\quad \geq \tfrac{1}{2}|x_{i_A}^*|^2 - |x_{i_A}^*| \left(|\langle \psi_{i_A}, \Psi_{B\backslash\{i_A\}} x_{B\backslash\{i_A\}}^*\rangle| + |\langle \psi_{i_A}, \eta\rangle| + |\langle \psi_{i_A}, \Psi_A \bar{x}_A\rangle|\right) \\ &\quad \geq \tfrac{1}{2}|x_{i_A}^*|^2 - |x_{i_A}^*| \left((|B|-1)\nu|x_{i_A}^*| + \epsilon + \frac{|A|\nu}{1-(|A|-1)\nu}(|B|\nu|x_{i_A}^*| + \epsilon)\right). \end{aligned}$$

Now with $\epsilon < \beta \min_{i\in A^*} |x_i^*| \leq \beta |x_{i_A}^*|$ from Assumption 2.1, we deduce

$$\begin{aligned}
J_\lambda(x) \geq & |x_{i_A}^*|^2 \left(\frac{1}{2} - \left((|B|-1)\nu + \beta + \frac{|A|\nu}{1-(|A|-1)\nu}(|B|\nu+\beta)\right)\right) + \lambda|A| \\
= & |x_{i_A}^*|^2 \left(\frac{1}{2} - (T-1)\nu - \beta\right) + |x_{i_A}^*|^2 |A|\nu \left(1 - \frac{|B|\nu+\beta}{1-(|A|-1)\nu}\right) + \lambda|A| \\
\geq & |x_{i_A}^*|^2 \left(\frac{1}{2} - (T-1)\nu - \beta\right),
\end{aligned}$$

where the last inequality follows from $(|A|+|B|-1)\nu + \beta < 1$. By Assumption 2.1, there holds $\epsilon^2/2 \leq \beta^2/2 \min_{i\in A^*} |x_i^*|^2$. Now by the assumption $\beta \leq (1-2(T-1)\nu)/(T+3)$, we deduce $(T+1)\beta^2 + 2\beta < (T+3)\beta \leq 1-2(T-1)\nu$, and hence $T\beta^2 < 1-2(T-1)\nu - 2\beta - \beta^2$. Together with the definition of ξ, this implies $\xi > \epsilon^2/2$. Further, by the choice of the parameter λ, i.e., $\lambda \in (\epsilon^2/2, \xi)$, there holds

$$J_\lambda(x) - J_\lambda(x^o) \geq \left[\frac{1}{2} - (T-1)\nu - \beta - \frac{1}{2}\beta^2\right] \min_{i\in A^*} |x_i^*|^2 - \lambda T > 0,$$

which contradicts the optimality of x.
Case (b) RIP: It follows from (2.6) that

$$\begin{aligned}
J_\lambda(x) &\geq \tfrac{1}{2}\|\Psi_B x_B^*\|^2 - |\langle \eta, \Psi_B x_B^* \rangle| - |\langle x_B^*, \Psi_B^t \Psi_A \bar{x}_A \rangle| + \lambda|A| \\
&\geq \|\Psi_B x_B^*\| (\tfrac{1}{2}\|\Psi_B x_B^*\| - \epsilon) - \|x_B^*\| \delta \|\bar{x}_A\| + \lambda|A|.
\end{aligned}$$

By Assumption 2.1 and the assumptions on β and δ, we deduce $\sqrt{1-\delta}\|x_B^*\| \geq \epsilon$. Now in view of the monotonicity of the function $t(t/2-\epsilon)$ for $t \geq \epsilon$, and the inequality $\|\Psi_B x_B^*\| \geq \sqrt{1-\delta}\|x_B^*\|$ from the definition of the RIP constant δ, we have

$$\|\Psi_B x_B^*\| (\tfrac{1}{2}\|\Psi_B x_B^*\| - \epsilon) \geq \sqrt{1-\delta}\|x_B^*\| (\tfrac{1}{2}\sqrt{1-\delta}\|x_B^*\| - \epsilon).$$

Thus by Lemma 2.3, we deduce

$$\begin{aligned}
J_\lambda(x) &\geq \frac{1-\delta}{2}\|x_B^*\|^2 - \epsilon\sqrt{1-\delta}\|x_B^*\| - \|x_B^*\| \left(\frac{\delta^2}{1-\delta}\|x_B^*\| + \frac{\delta}{\sqrt{1-\delta}}\epsilon\right) + \lambda|A| \\
&= \frac{1-\delta}{2}\|x_B^*\|^2 - \frac{1}{\sqrt{1-\delta}}\epsilon\|x_B^*\| - \|x_B^*\|^2 \frac{\delta^2}{1-\delta} + \lambda|A| \\
&\geq \|x_B^*\|^2 \left[\frac{1-\delta}{2} - \frac{\delta^2}{1-\delta} - \frac{\beta}{\sqrt{1-\delta}}\right] + \lambda|A|,
\end{aligned}$$

where the last line follows from $\epsilon < \beta\|x_B^*\|$, in view of Assumption 2.1. Appealing again to Assumption 2.1, $\epsilon^2/2 \leq \beta^2 \min_{i\in A^*} |x_i^*|^2/2 \leq \beta^2\|x_B^*\|^2/2$. Next it follows from the assumption $\beta \leq (1-\delta-\delta^2)/4$ that the inequality

$$\begin{aligned}
\beta^2 + \frac{\beta}{\sqrt{1-\delta}} \leq \frac{\beta^2+\beta}{\sqrt{1-\delta}} \leq \frac{2\beta}{1-\delta} \\
\leq \frac{1-2\delta-\delta^2}{2(1-\delta)} = \frac{1-\delta}{2} - \frac{\delta^2}{1-\delta}
\end{aligned}$$

holds. This together with the definition of ξ yields $\xi > \epsilon^2/2$. Further, the choice of $\lambda \in (\epsilon^2/2, \xi)$ implies

$$J_\lambda(x) - J_\lambda(x^o) \geq \|x_B^*\|^2 \left[\frac{1-\delta}{2} - \frac{\delta^2}{1-\delta} - \frac{\beta}{\sqrt{1-\delta}} - \frac{1}{2}\beta^2\right] - \lambda|B| > 0,$$

which again leads to a contradiction. This completes the proof of the theorem. □

Proposition 2.2. *Let the conditions in Theorem 2.1 hold. Then the oracle solution x^o is a minimizer of* (1.1)*. Moreover, the support to any solution of problem* (1.1) *is A^*.*

Proof. First we observe that there exists a solution $\bar{x}$ to problem (1.1) with $|\text{supp}(\bar{x})| \leq T$, upon noting that the true solution x^* satisfies $\|\Psi x^* - y\| \leq \epsilon$ and $\|x^*\|_0 \leq T$. Clearly, for any minimizer $\bar{x}$ to problem (1.1) with support $|A| \leq T$, then $\Psi_A^\dagger y$ is also a minimizer with $\|\Psi\Psi_A^\dagger y - y\| \leq \|\Psi\bar{x} - y\|$. Now if there is a minimizer $\bar{x}$ with $A \neq A^*$, by repeating the arguments in the proof of Theorem 2.1, we deduce

$$\tfrac{1}{2}\|\Psi\Psi_A^\dagger y - y\|^2 + \lambda\|\Psi_A^\dagger y\|_0 = J_\lambda(\Psi_A^\dagger y) > J_\lambda(x^o) = \tfrac{1}{2}\epsilon^2 + \lambda T \Rightarrow \|\Psi\bar{x} - y\| > \epsilon,$$

which leads a contradiction to the assumption that $\bar{x}$ is a minimizer to problem (1.1). Hence, any minimizer of (1.1) has a support A^*, and thus the oracle solution x^o is a minimizer. □

Remark 2.1. *Due to the nonconvex structure of problem* (1.1)*, the equivalence between problem* (1.1) *and its "Lagrange" version* (1.2) *is generally not clear. However under certain assumptions, their equivalence can be obtained, cf. Theorem 2.1 and Proposition 2.2. Further, we note that very recently, the equivalence between* (1.2) *and the following constrained sparsity problem*

$$\min \|\Psi x - y\| \quad \textit{subject to} \quad \|x\|_0 \leq T$$

was discussed in [32].

3 Primal-dual active set algorithm with continuation

In this section, we present the primal-dual active set with continuation (PDASC) algorithm, and establish its finite step convergence property.

3.1 The PDASC algorithm

The PDASC algorithm combines the strengthes of the PDAS algorithm [26] and the continuation technique. The complete procedure is described in Algorithm 1. The PDAS algorithm (the inner loop of Algorithm 1) first determines the active set A from the primal variable x and dual variable d, then updates the primal variable x by solving a least-squares problem on the active set A, and finally updates the dual variable d explicitly. It is well known that for convex optimization problems the PDAS algorithm can be interpreted as the semismooth Newton method [24]. Thus the algorithm merits a local superlinear convergence, and it reaches convergence with a good initial guess. In contrast, the continuation technique on the regularization parameter λ allows one to control the size of the active set A, and thus the active set of the coordinatewise minimizer lies within the true active set A^*, under appropriate assumptions. For example, for the choice of the parameter $\lambda_0 \geq \|\Psi^t y\|_{\ell^\infty}^2/2$, $x(\lambda_0) = 0$ is the unique global minimizer to the function J_{λ_0}, and its active set A is empty.

In the algorithm, there are a number of free parameters: the starting value λ_0 for the parameter λ, the decreasing factor $\rho \in (0, 1)$ (for λ), and the maximum number J_{max} of iterations for the inner PDAS loop. Further, one needs to set the stopping criteria at lines 6 and 10. Below we discuss their choices.

The choice of the initial value λ_0 is not important. For any choice $\lambda_0 \geq \|\Psi^t y\|_{\ell^\infty}^2/2$, $x = 0$ is the unique global minimizer, and $A = \emptyset$. Both the decreasing factor ρ and the iteration number J_{max} affect the accuracy and efficiency of the algorithm: Larger ρ and J_{max} values make the algorithm have better exact support recovery probability but at the expense of more computing time. Numerically, ρ is determined by the number of grid points for the parameter λ. Specifically, given an initial value $\lambda_0 \geq \|\Psi^t y\|_{\ell^\infty}^2/2$ and a small constant λ_{min}, e.g., 1e-15λ_0, the interval $[\lambda_{min}, \lambda_0]$ is divided into N equally distributed subintervals in the logarithmic scale. A large N value implies a large decreasing factor ρ. The choice $J_{max} = 1$ generally works well, which is also covered in the convergence theory in Theorems 3.1 and 3.2 below.

Algorithm 1 Primal dual active set with continuation (PDASC) algorithm

```
Set $\lambda_0 \geq \frac{1}{2}\|\Psi^t y\|^2_{\ell^\infty}$, $A(\lambda_0) = \emptyset$, $x(\lambda_0) = 0$ and $d(\lambda_0) = \Psi^t y$, $\rho \in (0, 1)$, $J_{max} \in \mathbb{N}$.
for $k = 1, 2, ...$ do
    Let $\lambda_k = \rho\lambda_{k-1}$, $A_0 = A(\lambda_{k-1})$, $(x^0, d^0) = (x(\lambda_{k-1}), d(\lambda_{k-1}))$.
    for $j = 1, 2, ..., J_{max}$ do
        Compute the active and inactive sets $A_j$ and $I_j$:
            $A_j = \left\{ i : |x_i^{j-1} + d_i^{j-1}| > \sqrt{2\lambda_k} \right\} \quad \text{and} \quad I_j = A_j^c.$
        Check stopping criterion $A_j = A_{j-1}$.
        Update the primal and dual variables $x^j$ and $d^j$ respectively by
            $\begin{cases} x^j_{I_j} = 0, \\ \Psi^t_{A_j}\Psi_{A_j}x^j_{A_j} = \Psi^t_{A_j}y, \\ d^j = \Psi^t(\Psi x^j - y). \end{cases}$
    end for
    Set $\widetilde{j} = \min(J_{max}, j)$, and $A(\lambda_k) = \left\{ i : |x_i^{\widetilde{j}} + d_i^{\widetilde{j}}| > \sqrt{2\lambda_k} \right\}$ and $(x(\lambda_k), d(\lambda_k)) = (x^{\widetilde{j}}, d^{\widetilde{j}})$.
    Check stopping criterion: $\|\Psi x(\lambda_k) - y\| \leq \epsilon$.
end for
```

The stopping criterion for each λ-problem in Algorithm 1 is either $A_j = A_{j-1}$ or $j = J_{max}$, instead of the standard criterion $A_j = A_{j-1}$ for active set type algorithms [24]. The condition $j = J_{max}$ is very important for nonconvex problems. This is motivated by the following empirical observation: When the true signal x^* does not have a strong decay property, e.g., 0-1 signal, the inner PDAS loop (for each λ-problem) may never reach the condition $A_j = A_{j-1}$ within finite steps; see the example below.

Example 3.1. *In this example, we illustrate the convergence of the PDAS algorithm. Let* $-1 < \mu < 0$, $A^* = \{1, 2\}$, *and*

$$\Psi_1 = \frac{1}{\sqrt{1+\mu^2}}(1, \mu, 0, ..., 0)^t, \quad \Psi_2 = \frac{1}{\sqrt{1+\mu^2}}(\mu, 1, 0, ..., 0)^t, \quad x_1^* = x_2^* = 1.$$

In the absence of data noise η, *the data* y *is given by*

$$y = \frac{1}{\sqrt{1+\mu^2}}(1+\mu, 1+\mu, 0, ..., 0)^t.$$

Now we let $\sqrt{2\lambda} \in (\frac{(1+\mu)^2}{1+\mu^2}, \frac{(1-\mu^2)^2}{(1+\mu^2)^2})$, *the initial guess* $A_1 = \{1\}$. *Then direct computation yields*

$$\begin{aligned} x^1 &= \frac{1}{1+\mu^2}((1+\mu)^2, 0)^t, \\ y - \Psi x^1 &= \frac{1-\mu^2}{(\sqrt{1+\mu^2})^3}(-\mu, 1, 0, ..., 0)^t, \\ d^1 &= \frac{1}{(1+\mu^2)^2}(0, (1-\mu^2)^2)^t. \end{aligned}$$

Hence $d_2^1 > \sqrt{2\lambda} > x_1^1$, *and* $A_2 = \{2\}$. *Similarly, we have* $A_3 = \{1\} = A_1$, *which implies that the algorithm simply alternates between the two sets* $\{1\}$ *and* $\{2\}$ *and will never reach the stopping condition* $A_k = A_{k+1}$.

The stopping criterion at line 10 of Algorithm 1 is a discrete analogue of the discrepancy principle. This principle is well established in the inverse problem community for selecting an appropriate regularization parameter [22]. The rationale behind the rule is that one cannot expect the reconstruction to be more accurate than the data accuracy, i.e., the discrepancy ϵ. In the PDASC algorithm, if the active set is always contained in the true active set A^* throughout the iteration, then the discrepancy principle can always be satisfied for some λ_k, and the solution $x(\lambda_k)$ resembles closely the oracle solution x^o.

3.2 Convergence analysis

Now we discuss the convergence of Algorithm 1. First we note that for a fixed λ value, since there are only a finite number of active sets, the PDAS algorithm is asymptotically periodic, and one naturally expects a finite-step convergence, if the desired convergence does occur. However, the convergence of the iteration itself does not follow directly, cf. Example 3.1. We shall discuss the cases of the MIP and RIP conditions separately. The general proof strategy is as follows. It essentially relies on the precise characterization of the evolution of the active set during the iteration, especially a certain monotonicity relation of the active set $A(\lambda_k)$ (via the continuation technique). In particular, we introduce two auxiliary sets G_{λ,s_1} and G_{λ,s_2}, cf. (3.1) below, to precisely characterize the evolution of the active set A during the PDASC iteration.

First we consider the MIP case. We begin with an elementary observation: under the assumption $\nu < (1-2\beta)/(2T-1)$ of the mutual coherence ν, there holds $(2T-1)\nu + 2\beta < 1$.

Lemma 3.1. *If $\nu < (1-2\beta)/(2T-1)$, then for any $\rho \in (((2T-1)\nu+2\beta)^2, 1)$ there exist $s_1, s_2 \in (1/(1-T\nu+\nu-\beta), 1/(T\nu+\beta))$, $s_1 > s_2$, such that $s_2 = 1 + (T\nu - \nu + \beta)s_1$ and $\rho = s_2^2/s_1^2$.*

Proof. By the assumption $v < (1-2\beta)/(2T-1)$, $T\nu + \beta < 1 - T\nu + \nu - \beta$. Hence for any $s_1 \in (1/(1-T\nu+\nu-\beta), 1/(T\nu+\beta))$, there holds

$$s_1 > 1 + (T\nu - \nu + \beta)s_1 \quad \text{and} \quad 1 + (T\nu - \nu + \beta)s_1 > \frac{1}{1 - T\nu + \nu - \beta},$$

i.e.,

$$\frac{1}{T\nu + \beta} > s_1 > 1 + (T\nu - \nu + \beta)s_1 > \frac{1}{1 - T\nu + \nu - \beta}.$$

Upon letting $s_2 = 1 + (T\nu - \nu + \beta)s_1$, we deduce

$$\frac{1}{T\nu + \beta} > s_1 > s_2 > \frac{1}{1 - T\nu + \nu - \beta}.$$

Now the monotonicity of the function $f(s_1) = s_2/s_1$ over the interval $(1/(1-T\nu+\nu-\beta), 1/(T\nu+\beta))$, and the identities

$$\frac{1 + (T\nu - \nu + \beta)/(T\nu + \beta)}{1/(T\nu + \beta)} = (2T-1)\nu + 2\beta,$$
$$\frac{1 + (T\nu - \nu + \beta)/(1 - T\nu + \nu - \beta)}{1/(1 - T\nu + \nu - \beta)} = 1,$$

imply that there exists an s_1 in the internal such that $s_2/s_1 = \sqrt{\rho}$ for any $\rho \in (((2T-1)\nu + 2\beta)^2, 1)$. □

Next for any $\lambda > 0$ and $s > 0$, we denote by

$$G_{\lambda,s} \triangleq \left\{ i : |x_i^*| \geq \sqrt{2\lambda}s \right\}. \tag{3.1}$$

The set $G_{\lambda,s}$ characterizes the true sparse signal x^* (via level sets). The lemma below provides an important monotonicity relation on the active set A_k during the iteration, which is essential for showing the finite step convergence of the algorithm in Theorem 3.1 below.

Lemma 3.2. *Let Assumption 2.1 hold, $\nu < (1-2\beta)/(2T-1)$, $\rho \in (((2T-1)\nu+2\beta)^2, 1)$, and s_1 and s_2 be defined in Lemma 3.1. If $G_{\lambda,s_1} \subseteq A_k \subseteq A^*$, then $G_{\lambda,s_2} \subseteq A_{k+1} \subseteq A^*$.*

Proof. Let $A = A_k$, $B = A^* \backslash A$. By Lemma 2.3, we have

$$\begin{aligned}
&|x_i| \geq |x_i^*| - \|\bar{x}_A\|_{\ell^\infty} \geq |x_i^*| - \frac{|B|\nu\|x_B^*\|_{\ell^\infty} + \epsilon}{1-(|A|-1)\nu}, \quad \forall i \in A,\\
&|d_j| \leq |B|\nu\left(1 + \frac{|A|\nu}{1-(|A|-1)\nu}\right)\|x_B^*\|_{\ell^\infty} + \epsilon\left(1+\frac{|A|\nu}{1-(|A|-1)\nu}\right), \quad \forall j \in I^*,\\
&|d_i| \geq |x_i^*| + \nu\|x_B^*\|_{\ell^\infty} - |B|\nu\left(1+\frac{|A|\nu}{1-(|A|-1)\nu}\right)\|x_B^*\|_{\ell^\infty} - \epsilon\left(1+\frac{|A|\nu}{1-(|A|-1)\nu}\right), \quad \forall i \in B.
\end{aligned}$$

Using the fact $\epsilon \leq \beta \min_{i\in A^*} |x_i^*| \leq \beta\|x_B^*\|_{\ell^\infty}$ from Assumption 2.1 and the trivial inequality $\frac{|B|\nu+\beta}{1-T\nu+\nu+|B|\nu} \leq \frac{T\nu+\beta}{1+\nu}$, we arrive at

$$\begin{aligned}
|B|\nu\left(1+\frac{|A|\nu}{1-(|A|-1)\nu}\right)&\|x_B^*\|_{\ell^\infty} + \epsilon\left(1+\frac{|A|\nu}{1-(|A|-1)\nu}\right)\\
&\leq \frac{|B|\nu+\beta}{1-T\nu+\nu+|B|\nu}(1+\nu)\|x_B^*\|_{\ell^\infty} \leq (T\nu+\beta)\|x_B^*\|_{\ell^\infty}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&|d_j| \leq (T\nu+\beta)\|x_B^*\|_{\ell^\infty}, \quad \forall j \in I^*,\\
&|d_i| \geq |x_i^*| - (T\nu-\nu+\beta)\|x_B^*\|_{\ell^\infty}, \quad \forall i \in B.
\end{aligned}$$

It follows from the assumption $G_{\lambda,s_1} \subseteq A = A_k$ that $\|x_B^*\|_{\ell^\infty} < s_1\sqrt{2\lambda}$. Then for all $j \in I^*$, we have

$$|d_j| < s_1(T\nu+\beta)\sqrt{2\lambda} < \sqrt{2\lambda},$$

i.e., $j \in I_{k+1}$. This shows $A_{k+1} \subseteq A^*$. For any $i \in I \cap G_{\lambda,s_2}$, we have

$$|d_i| > s_2\sqrt{2\lambda} - (T\nu-\nu+\beta)s_1\sqrt{2\lambda} \geq \sqrt{2\lambda},$$

This implies $i \in A_{k+1}$ by (2.4). It remains to show that for any $i \in A \cap G_{\lambda,s_2}$, $i \in A_{k+1}$. Clearly, if $A = \emptyset$, the assertion holds. Otherwise

$$\begin{aligned}
|x_i| &\geq |x_i^*| - \frac{|B|\nu+\beta}{1-(|A|-1)\nu}\|x_B^*\|_{\ell^\infty}\\
&> s_2\sqrt{2\lambda} - (T\nu-\nu+\beta)s_1\sqrt{2\lambda} \geq \sqrt{2\lambda},
\end{aligned}$$

where the last line follows from the elementary inequality

$$\frac{|B|\nu+\beta}{1-(|A|-1)\nu} \leq T\nu-\nu+\beta.$$

This together with (2.4) also implies $i \in A_{k+1}$. This concludes the proof of the lemma. □

Now we can state the convergence result.

Theorem 3.1. *Let Assumption 2.1 hold, and $\nu < (1-2\beta)/(2T-1)$. Then for any $\rho \in (((2T-1)\nu+2\beta)^2, 1)$, Algorithm 1 converges in finite steps.*

Proof. For each λ_k-problem, we denote by $A_{k,0}$ and $A_{k,\diamond}$ the active set for the initial guess and the last inner step (i.e., $A(\lambda_k)$ in Algorithm 1), respectively. Now with s_1 and s_2 from Lemma 3.1, there holds

$G_{\lambda,s_1} \subset G_{\lambda,s_2}$, and using Lemma 3.2, for any index k before the stopping criterion at line 10 of Algorithm 1 is reached, there hold

$$G_{\lambda_k,s_1} \subseteq A_{k,0} \quad \text{and} \quad G_{\lambda_k,s_2} \subseteq A_{k,\diamond}. \tag{3.2}$$

Note that for $k = 0$, $G_{\lambda_0,s_2} = \emptyset$ and thus the assertion holds. To see this, it suffices to check $\|x^*\|_{\ell^\infty} < s_2\|\Psi^t y\|_{\ell^\infty}$. By Lemma 2.1 and the inequality $s_2 > 1/(1 - T\nu + \nu - \beta)$ we obtain that

$$\begin{aligned} \|\Psi^t y\|_{\ell^\infty} &\geq \|\Psi^t_{A^*}\Psi_{A^*}x^*_{A^*}\|_{\ell^\infty} - \|\Psi^t\eta\|_{\ell^\infty} \\ &\geq (1 - (T-1)\nu)\|x^*\|_{\ell^\infty} - \epsilon > \|x^*\|_{\ell^\infty}/s_2. \end{aligned}$$

Now for $k > 0$, it follows by mathematical induction and the relation $A_{k,\diamond} = A_{k+1,0}$. It follows from (3.2) that during the iteration, the active set $A_{k,\diamond}$ always lies in A^*. Further, for k sufficiently large, by Lemma 2.5, the stopping criterion at line 10 must be reached and thus the algorithm terminates; otherwise

$$A^* \subseteq G_{\lambda_k,s_1},$$

then the stopping criterion at line 10 is satisfied, which leads to a contradiction. □

Next we turn to the convergence of Algorithm 1 under the RIP condition. Let $1 - (2\sqrt{T} + 1)\delta > 2\beta$, an argument analogous to Lemma 3.1 implies that for any $\sqrt{\rho} \in ((2\delta\sqrt{T} + 2\beta)/(1-\delta), 1)$ there exist s_1 and s_2 such that

$$\frac{1-\delta}{\delta\sqrt{T}+\beta} > s_1 > s_2 > \frac{1-\delta}{1-\delta-\delta\sqrt{T}-\beta}, \quad s_2 = 1 + \frac{\delta\sqrt{T}+\beta}{1-\delta}s_1, \quad \frac{s_2}{s_1} = \sqrt{\rho}. \tag{3.3}$$

The next result is an analogue of Lemma 3.2.

Lemma 3.3. *Let Assumption 2.1 hold, $\delta \triangleq \delta_{T+1} \leq (1-2\beta)/(2\sqrt{T}+1)$, and $\sqrt{\rho} \in ((2\delta\sqrt{T} + 2\beta)/(1-\delta), 1)$. Let s_1 and s_2 are defined by (3.3). If $G_{\lambda,s_1} \subseteq A_k \subseteq A^*$, then $G_{\lambda,s_2} \subseteq A_{k+1} \subseteq A^*$.*

Proof. Let $A = A_k$, $B = A^*\backslash A$. Using the notation in Lemma 2.3, we have

$$\begin{aligned} |x_i| &\geq |x^*_i| - \|\bar{x}_A\| \geq |x^*_i| - \frac{\delta\|x^*_B\| + \epsilon}{1-\delta}, \quad \forall i \in A, \\ |d_j| &\leq \delta\|x^*_B\| + \epsilon + \delta\|\bar{x}_A\| \leq \frac{\delta\|x^*_B\| + \epsilon}{1-\delta}, \quad \forall j \in I^*, \\ |d_i| &\geq |x^*_i| - \delta\|x^*_B\| - \epsilon - \delta\|\bar{x}_A\| \geq |x^*_i| - \frac{\delta\|x^*_B\| + \epsilon}{1-\delta}, \quad \forall i \in B. \end{aligned}$$

By the assumption $G_{\lambda,s_1} \subseteq A_k$, we have $\|x^*_B\|_{\ell^\infty} < s_1\sqrt{2\lambda}$. Now using the relation $s_1 < (1-\delta)/(\delta\sqrt{T}+\beta)$ and Assumption 2.1, we deduce

$$\frac{\delta\|x^*_B\| + \epsilon}{1-\delta} \leq \frac{\delta\sqrt{T}+\beta}{1-\delta}\|x^*_B\|_{\ell^\infty} < \sqrt{2\lambda}.$$

Thus for $j \in I^*$, $|d_i| < \sqrt{2\lambda}$, i.e., $A_{k+1} \subset A^*$. Similarly, using the relations $s_2 = 1 + s_1(\delta\sqrt{T}+\beta)/(1-\delta)$ and $s_1 > (1-\delta)/(1-\delta-\delta\sqrt{T}-\beta)$, we arrive at that for any $i \in G_{\lambda,s_2}$, there holds

$$|x^*_i| - \frac{\delta\|x^*_B\| + \epsilon}{1-\delta} > s_2\sqrt{2\lambda} - \frac{\delta\sqrt{T}+\beta}{1-\delta}s_1\sqrt{2\lambda} = \sqrt{2\lambda}.$$

This implies that for $i \in G_{\lambda,s_2} \cap A$, $|x_i| > \sqrt{2\lambda}$, and for $i \in G_{\lambda,s_2} \cap I$, $|d_i| > \sqrt{2\lambda}$. Consequently, (2.4) yields the desired relation $(G_{\lambda,s_2} \cap A) \subseteq A_{k+1}$, and this concludes the proof of the lemma. □

Now we can state the convergence of Algorithm 1 under the RIP assumption. The proof is similar to that for Theorem 3.1, and hence omitted.

Theorem 3.2. *Let Assumption 2.1 hold, and $\delta \triangleq \delta_{T+1} \leq (1-2\beta)/(2\sqrt{T}+1)$. Then for any $\sqrt{\rho} \in \Big((2\delta\sqrt{T}+2\beta)/(1-\delta), 1\Big)$, Algorithm 1 converges in finite steps.*

Remark 3.1. *Theorems 3.1 and 3.2 indicate that under designated assumptions, Algorithm 1 converges in finite steps, and the active set $A(\lambda_k)$ remains a subset of the true active set A^*.*

Corollary 3.1. *Let the assumptions in Theorem 2.1 hold. Then Algorithm 1 terminates at the oracle solution x^o.*

Proof. First, we note the monotonicity relation $A(\lambda_k) \subset A^*$ before the stopping criterion at line 10 of Algorithm 1 is reached. For any $A \subsetneq A^*$, let $x = \Psi_A^\dagger y$. Then by the argument in the proof of Theorem 2.1, we have

$$J_\lambda(x) = \tfrac{1}{2}\|\Psi x - y\|^2 + \lambda|A| > \tfrac{1}{2}\epsilon^2 + \lambda T \Rightarrow \|\Psi x - y\| > \epsilon,$$

which implies that the stopping criterion at line 10 in Algorithm 1 cannot be satisfied until the oracle solution x^o is reached. □

3.3 Connections with other algorithms

Now we discuss the connections of Algorithm 1 with three existing greedy methods, i.e., orthogonal matching pursuit (OMP), iterative hard thresholding (IHT) and hard thresholding pursuit (HTP).

Connection with the OMP. To prove the convergence of Algorithm 1, we require either the MIP condition ($\nu < (1-2\beta)/(2T-1)$) or the RIP condition ($\delta_{T+1} \leq (1-2\beta)/(2\sqrt{T}+1)$) on the sensing matrix Ψ. These assumptions have been used to analyze the OMP before: the MIP appeared in [8] and the RIP appeared in [21]. Further, for the OMP, the MIP assumption is fairly sharp, but the RIP assumption can be improved [45, 28]. Our convergence analysis under these assumptions, unsurprisingly, follows the same line of thought as that for the OMP, in that we require the active set $A(\lambda_k)$ always lies in the true active set A^* throughout the iteration. However, we note that this requirement is unnecessary for implementing the PDASC algorithm, since the active set $A(\lambda_k)$ can move inside and outside the true active set A^* during the iteration. The numerical examples in Section 4 below confirm this observation. This makes the PDASC much more flexible than the OMP.

Connection with the IHT and HTP. The IHT due to Blumensath and Davies [6] also defines the active set by both primal and dual variables, but with a projection step, i.e., hard thresholding, in place of the least-squares step (cf. Step 7) in the PDASC algorithm. The HTP due to Foucart [19] can be viewed a primal-dual active set method in the T-version, i.e., at each iteration, the active set is chosen by the first T components based on primal and dual variables. This is equivalent to a variable regularization parameter λ, where $\sqrt{2\lambda}$ is set to the T-th component of $|x^k + d^k|$ at each iteration. The convergence of IHT and HTP were provided under RIP condition, with RIP constants being $\delta_{3T} \leq 1/\sqrt{32}$ [6] and $\delta_{3T} \leq 1/\sqrt{3}$ [19], respectively. These results are stronger than our convergence result based on the RIP, i.e., Theorem 3.2, but the former require an a priori knowledge of the exact sparsity level T. In addition, the IHT has also been applied to the Lagrange formulation (1.2) [5], but the convergence seems unknown.

4 Numerical tests and discussions

In this section we present numerical examples to illustrate the efficiency and accuracy, and the convergence behavior of the proposed PDASC algorithm. The sensing matrix Ψ is of size $n \times p$, the true solution x^* is a T-sparse signal with an active set A^*. The dynamical range R of the true signal x^* is defined by $R = M/m$, with $M = \max\{|x_i^* : i \in A^*|\}$ and $m = \min\{|x_i^*| : i \in A^*\}$. The data y is generated by

$$y = \Psi x^* + \eta,$$

where η denotes the measurement noise, with each entry η_i following the Gaussian distribution $N(0,\sigma^2)$ with mean zero and standard deviation σ. The exact noise level ϵ is given by $\epsilon = \|\eta\|_2$.

In Algorithm 1, we always take $\lambda_0 = \frac{1}{2}\|\Psi^t y\|^2_{\ell^\infty}$, and λ_{min} = 1e-15λ_0. The choice of the number of grid points N and the maximum number J_{max} of inner iterations will be specified later.

Step 7 of Algorithm 1 requires solving a linear system, which is the most expensive piece of the algorithm. A direct linear solver can be expensive for large-scale problems or even infeasible when the sensing matrix Ψ is given only implicitly. Hence, in practice one may employ iterative solvers for symmetric positive definite systems, e.g., (preconditioned) conjugate gradient (CG) method. In the following numerical examples, when the matrix Ψ is the partial DCT or composition of partial FFT with an inverse wavelet transform (for 1D signal or 2D MRI image), we employ the CG method to solve the resulting linear systems. In practice, only a few CG steps are needed, in view of the good conditioning of the linear systems and a good initial guess from the continuation strategy.

All the computations were performed on a dual core desktop with 3.40 GHz and 8 GB RAM using `MATLAB` version 2013b. The `MATLAB` package `PDASCl0` for reproducing all the numerical results can be found at http://www0.cs.ucl.ac.uk/staff/b.jin/companioncode.html.

4.1 The behavior of the PDASC algorithm

First we study the influence of the free parameters in the PDASC algorithm on the exact support recovery probability. To this end, we fix Ψ to be a 500 × 1000 random Gaussian matrix, and σ = 1e-2. All the results are computed based on 100 independent realizations of the problem setup. To this end, we consider the following three settings:

(a) $J_{max} = 5$, and varying N; see Fig. 1(a).

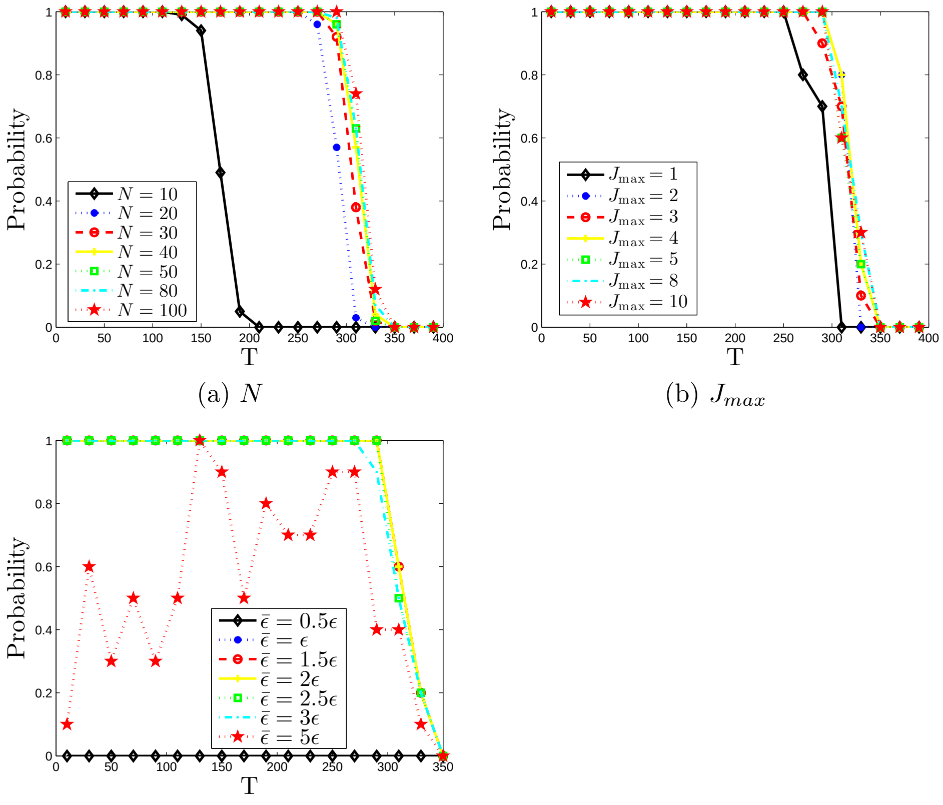
(b) $N = 100$, and varying J_{max}; see Fig. 1(b).

(c) $N = 100$, $J_{max} = 5$, and an approximate noise level $\bar{\epsilon}$; see Fig. 1(c).

We observe that the influence of the parameters N and J_{max} is very mild on the exact support recovery probability. In particular, a reasonably small value for these parameters (e.g. $N = 50$, $J_{max} = 1$) is sufficient for accurately recovering the exact active set A^*. Unsurprisingly, a very small value of N can degrade the accuracy of support recovery greatly, due to insufficient resolution of the solution path. In practice, the exact noise level ϵ is not always available, and often only an approximate estimate $\bar{\epsilon}$ is provided. The use of the estimate $\bar{\epsilon}$ in place of the exact one ϵ in Algorithm 1 may sacrifice the exact recovery probability. Hence it is important to study the sensitivity of Algorithm 1 with respect to the variation of the parameter ϵ. We observe from Fig. 1(c) that the use of the estimate $\bar{\epsilon}$ does not affect the recovery probability much, unless it is grossly erroneous. The case of an overly underestimated noise level is especially dangerous, which may render the reconstruction completely useless due to an insufficient amount of regularization.

To gain further insight into the PDASC algorithm, in Fig. 2, we show the evolution of the active set (for simplicity let $A_k = A(\lambda_k)$) . It is observed that the active set A_k can generally move both "inside" and "outside" of the true active set A^*. This observation is valid for random Gaussian, random Bernoulli and partial DCT sensing matrices. This behavior is in sharp contrast to the OMP, where the size of the active set is monotonically increasing during the iteration, by its construction. The flexible change in the active set might be essential for the efficiency of the PDASC algorithm.

For each λ_k-problem, with $x(\lambda_{k-1})$ $(x(\lambda_0) = 0)$ as the initial guess, the PDASC generally reaches convergence within a few iterations, typically two or three, cf. Fig. 3, which is observed for random Gaussian, random Bernoulli and partial DCT sensing matrices. This is attributed to the local superlinear convergence of the PDAS algorithm. Hence, when coupled with the continuation strategy, the overall PDASC procedure is very efficient.



(a) N

(b) J_{max}

(c) ϵ

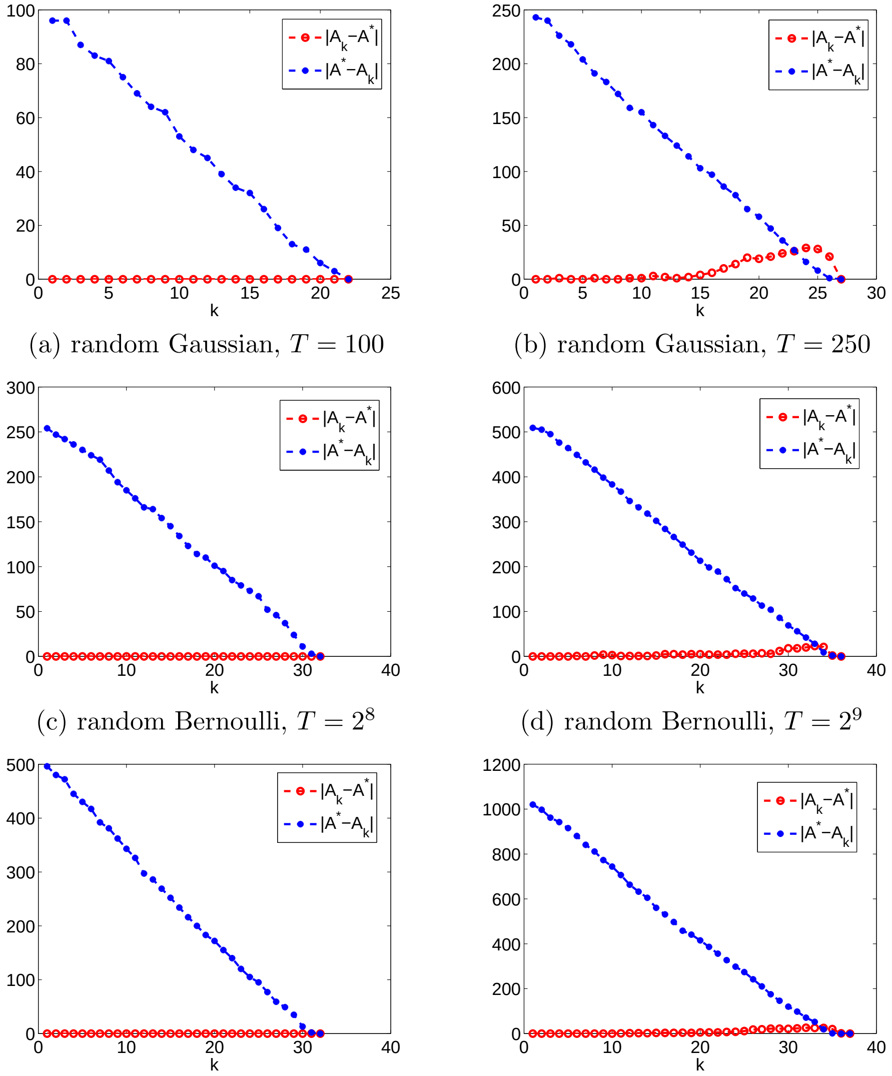
Figure 1: The influence of the algorithmic parameters (N, J_{max} and ϵ) on the exact recovery probability.

4.2 Comparison with existing algorithms

In this part, we compare Algorithm 1 with five state-of-the-art algorithms in the compressive sensing literature, including orthogonal matching pursuit (OMP) [34], greedy gradient pursuit (GreedyGP) [4], accelerated iterative hard thresholding (AIHT) [3], hard thresholding pursuit (HTP) [19], compressive sampling matching pursuit (CoSaMP) [29].

First, we consider the exact support recovery probability, i.e., the percentage of the reconstructions whose support agrees with the true active set A^*. To this end, we fix the sensing matrix Ψ as a 500×1000 random Gaussian matrix, $\sigma =$ 1e-3, $(N, J_{max}) = (100, 5)$ or $(50, 1)$, and all results are computed from 100 independent realizations of the problem setup. Since a different dynamical range R may give different results, we take $R = 1$, 10, 1e3, 1e5 as four exemplary values. The numerical results are summarized in Fig. 4. We observe that when the dynamical range R is not very small, the proposed PDASC algorithm with $(N, J_{max}) = (100, 5)$ has a better exact support recovery probability, and that with the choice $(N, J_{max}) = (50, 1)$ is also largely comparable with other algorithms.

To further illustrate the accuracy and efficiency of the proposed PDASC algorithm, we compare it with other greedy methods in terms of CPU time and reconstruction error. To this end, we fix $\sigma =$ 1e-2, $(N, J_{max}) = (100, 5)$ or $(50, 1)$. The numerical results for random Gaussian, random Bernoulli and partial DCT sensing matrices with different parameter tuples (R, n, p, T) are shown in Tables 4.1-4.3, respectively. The results in the tables are computed from 10 independent realizations of the problem setup. It is observed that the PDASC algorithm yields reconstructions that are comparable with that by other methods, e.g., HTP and AIHT, but usually with less computing time. Further, we observe that it scales well with the problem size. By increasing the maximum number of inner iterations J_{max} and



(a) random Gaussian, $T = 100$

(b) random Gaussian, $T = 250$

(c) random Bernoulli, $T = 2^8$

(d) random Bernoulli, $T = 2^9$

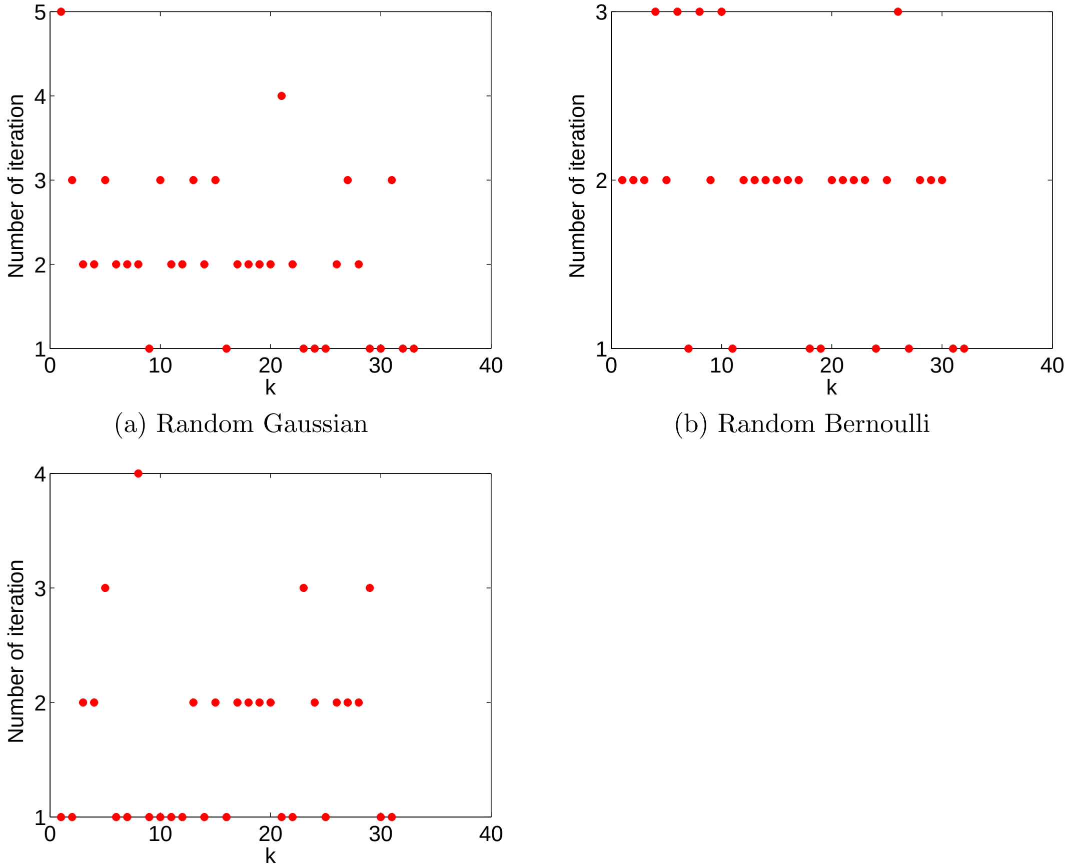
(e) partial DCT, $T = 2^9$

(f) partial DCT, $T = 2^{10}$

Figure 2: Numerical results for random Gaussian (top row, $R = 100$, $n = 500$, $p = 1000$, $\sigma =$ 1e-3), random Bernoulli (middle row, $R = 1000$, $n = 2^{10}$, $p = 2^{12}$, $\sigma =$ 1e-3) and partial DCT (bottom row, $R = 1000$, $n = 2^{11}$, $p = 2^{13}$, $\sigma =$ 1e-3) sensing matrix. The parameters N and J_{max} are set to $N = 50$ and $J_{max} = 1$, respectively.

the continuation steps N, one can improve the reconstruction accuracy slightly, but the enhancement is small. This indicates that with the "good" initial guess provided by the continuation strategy, one inner iteration is sufficient to achieve the desired accuracy, due to the fast local convergence of the PDASC algorithm, and thus it is also very efficient for large-scale problems.

Lastly, we consider one-dimensional signals and two-dimensional images. In this case the explicit form of the sensing matrix Ψ may be not available, and we employ the CG method for the least-squares step at line 7 of Algorithm 1. The most natural initial guess for the CG method for the λ_k-problem is the solution $x(\lambda_{k-1})$ (projected on the current active set), and the stopping criterion for the CG method is as follows: either the number of CG iterations is greater than a given (small) integer or the residual is smaller than a given tolerance. With the continuation strategy, a few (often one or two) CG iterations at the inner loop



(a) Random Gaussian

(b) Random Bernoulli

(c) Partial DCT

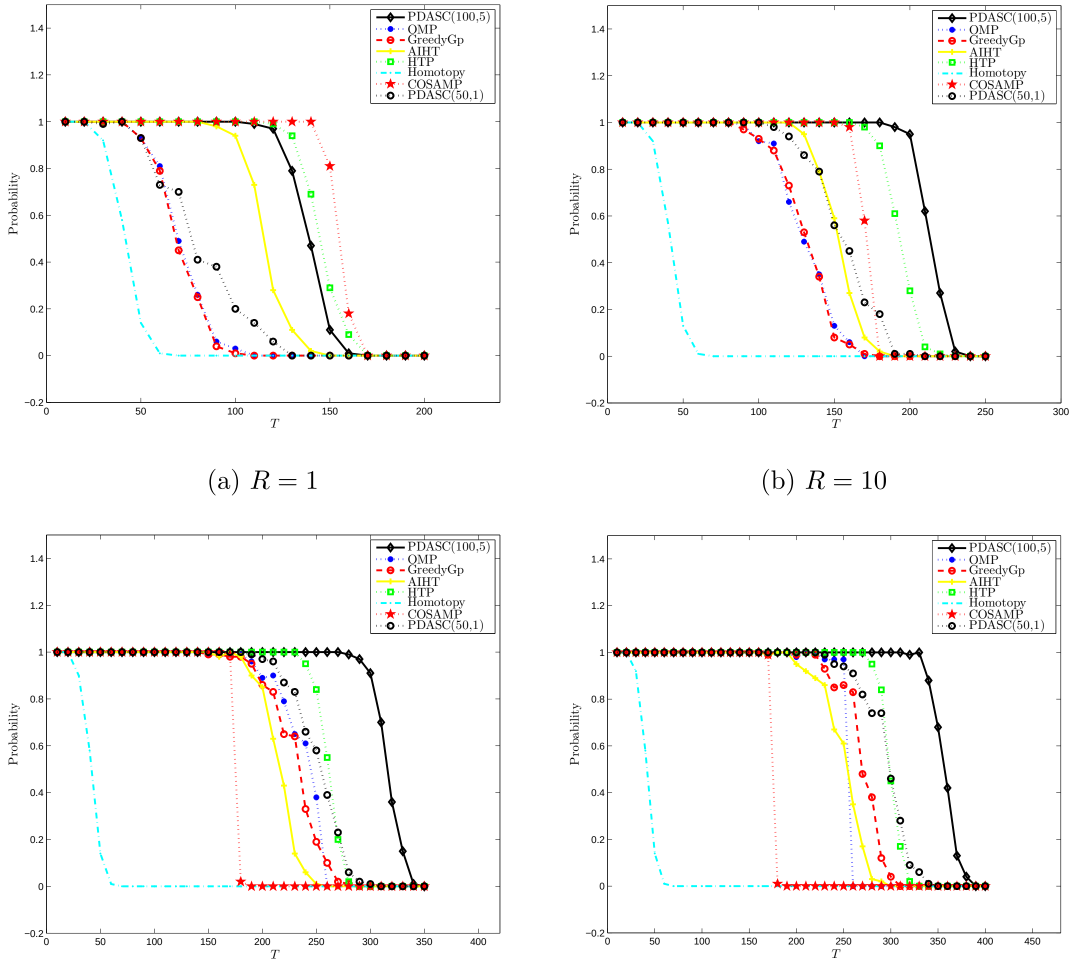
Figure 3: The number of iterations of PDASC at each λ_k for random Gaussian (top left with $R = 1000$, $n = 500$, $p = 1000$, $T = 200$, σ = 1e-3), random Bernoulli (top right with $R = 1000$, $n = 2^{10}$, $p = 2^{12}$, $T = 2^8$, σ = 1e-3) and partial DCT (bottom with $R = 1000$, $n = 2^{11}$, $p = 2^{13}$, $T = 2^8$, σ = 1e-3) sensing matrix. The parameters N and J_{max} are set to $N = 50$ and $J_{max} = 5$, respectively.

of Algorithm 1 suffice the desired accuracy. Numerically we find that the choice $(N, J_{max}) = (50, 1)$ and one CG iteration for the least-squares problem works well in practice, and thus we present the results below only for this particular choice. This surprising "superconvergence" phenomenon awaits theoretical justification.

For the one-dimensional signal, the sampling matrix Ψ is of size 665×1024, and it consists of applying a partial FFT and an inverse wavelet transform (with two level of Daubechies 1 wavelet), and the signal under wavelet transformation has 247 nonzero entries and σ = 1e-4, $N = 50$, $J_{\max} = 1$. The results are shown in Fig. 5 and Table 4.4. The reconstructions by all the methods, except the AIHT and CoSaMP, are visually very appealing and in excellent agreement with the exact solution. The reconstructions by the AIHT and CoSaMP suffer from pronounced oscillations. This is further confirmed by the PSNR values which is defined as

$$PSNR = 10 \cdot \log \frac{V^2}{MSE}$$

where V is the maximum absolute value of the reconstruction and the true solution, and MSE is the mean squared error of the reconstruction, cf. Table 4.4. One finds that except the CoSaMP, all other methods can yield almost identical reconstructions within similar computational efforts.



(a) $R = 1$

(b) $R = 10$

(c) $R = 10^3$

(d) $R = 10^5$

Figure 4: The exact support recovery probability for four different dynamical ranges: $R = 1,\ 10,\ 10^3$, and 10^5.

For the two-dimensional MRI image, the sampling matrix Ψ amounts to a partial FFT and an inverse wavelet transform, and it has a size 29729×262144. The image under wavelet transformation (with four level of Daubechies 1 wavelet) has 8450 nonzero entries and σ = 1e-3, $N = 50$, and $J_{\max} = 1$. The numerical results are shown in Fig. 6 and Table 4.5. Since OMP is too costly for this example (more than one hour), we do not present the corresponding numerical result. All other methods produce similar results. Therefore proposed PDASC algorithm is competitive with state-of-the-art algorithms, and feasible for large-scale problems with implicit operators.

5 Conclusion

We have developed an efficient and accurate primal-dual active set with continuation algorithm for the ℓ^0 penalized least-squares problem arising in compressive sensing. It combines the fast local convergence of the active set technique and the globalizing property of the continuation technique. The global finite step convergence of the algorithm was established under the mutual incoherence property or restricted isometry

Table 4.1: Numerical results (CPU time and errors) for medium-scale problems, with random Gaussian sensing matrix Ψ, of size $p = 10000,\ 15000,\ 20000,\ 25000,\ 30000$, $n = \lfloor p/4 \rfloor$, $T = \lfloor n/3 \rfloor$. The dynamical range R is $R = 1000$, and the noise variance σ is $\sigma =$ 1e-2.

p	method	time(s)	Relative ℓ^2 error	Absolute ℓ^∞ error
10000	PDASC(50,1)	1.46	4.49e-5	3.81e-2
	PDASC(100,5)	2.82	4.49e-5	3.81e-2
	OMP	15.1	4.49e-5	3.81e-2
	GreedyGP	16.5	8.30e-5	1.32e-1
	AIHT	4.23	4.49e-5	3.81e-2
	HTP	1.54	4.49e-5	3.81e-2
	CoSaMP	7.98	8.87e-2	6.34e+1
15000	PDASC(50,1)	3.44	4.61e-5	4.56e-2
	PDASC(100,5)	6.63	4.61e-5	4.56e-2
	OMP	51.3	4.61e-5	4.55e-2
	GreedyGP	54.8	7.17e-5	1.24e-1
	AIHT	9.33	4.61e-5	4.56e-2
	HTP	3.88	4.61e-5	4.56e-2
	CoSaMP	22.6	8.66e-2	6.16e+1
20000	PDASC(50,1)	6.47	4.56e-5	4.34e-2
	PDASC(100,5)	12.3	4.56e-5	4.34e-2
	OMP	119	4.56e-5	4.34e-2
	GreedyGP	127	6.29e-5	1.12e-1
	AIHT	16.2	4.56e-5	4.34e-2
	HTP	7.12	4.56e-5	4.34e-2
	CoSaMP	50.1	8.68e-1	6.59e+1
25000	PDASC(50,1)	11.1	4.55e-5	4.61e-2
	PDASC(100,5)	20.3	4.55e-5	4.61e-2
	OMP	230	4.55e-5	4.61e-2
	GreedyGP	245	5.87e-5	1.10e-1
	AIHT	25.2	4.55e-5	4.61e-2
	HTP	12.0	4.55e-5	4.61e-2
	CoSaMP	90.9	9.29e-1	6.71e+1
30000	PDASC(50,1)	17.5	4.53e-5	4.53e-2
	PDASC(100,5)	31.8	4.53e-5	4.53e-2
	OMP	399	4.53e-5	4.53e-2
	GreedyGP	430	5.62e-5	1.07e-1
	AIHT	36.9	4.53e-5	4.53e-2
	HTP	18.6	4.53e-5	4.53e-2
	CoSaMP	153	9.32e-2	7.61e+1

property on the sensing matrix. Our extensive numerical results indicate that the proposed algorithm is competitive with state-of-the-art algorithms in terms of efficiency, accuracy and exact recovery probability, without a knowledge of the exact sparsity level.

Our numerical experiment indicates that the conjugate gradient method is very effective for solving the least-squares problems arising in the PDAS iterations. A complete analysis of the excellent convergence behavior is of significant interest. Further, the primal dual active set with continuation algorithm extends naturally to other popular nonconvex sparse models, e.g., bridge penalty, smoothly clipped absolute deviation and minmax concave penalty. However, the convergence analysis of the algorithm for these penalties remains unclear.

Table 4.2: Numerical results (CPU time and errors) for medium-scale problems, with random Bernoulli sensing matrix Ψ, of size $p = 10000,\ 15000,\ 20000,\ 25000,\ 30000$, $n = \lfloor p/4 \rfloor$, $T = \lfloor n/4 \rfloor$. The dynamical range R is $R = 10$, and the noise variance σ is $\sigma =$ 1e-2.

p	method	time(s)	Relative ℓ^2 error	Absolute ℓ^∞ error
10000	PDASC(50,1)	0.64	2.46e-3	3.82e-2
	PDASC(100,5)	1.21	2.45e-3	3.82e-2
	OMP	10.3	2.46e-3	3.83e-2
	GreedyGP	12.8	2.40e-2	1.07e+0
	AIHT	3.50	2.45e-3	3.82e-2
	HTP	0.95	2.45e-3	3.82e-2
	CoSaMP	4.93	5.00e-3	7.57e-2
15000	PDASC(50,1)	1.49	2.53e-3	4.13e-2
	PDASC(100,5)	2.85	2.52e-3	4.11e-2
	OMP	34.6	2.52e-3	4.10e-2
	GreedyGP	43.7	2.20e-2	1.07e+0
	AIHT	8.38	2.52e-3	4.11e-2
	HTP	2.35	2.52e-3	4.11e-2
	CoSaMP	13.9	5.01e-3	7.75e-2
20000	PDASC(50,1)	2.63	2.52e-3	3.96e-2
	PDASC(100,5)	4.93	2.51e-3	3.97e-2
	OMP	78.4	2.52e-3	3.99e-2
	GreedyGP	96.4	2.41e-2	1.09e+0
	AIHT	13.6	2.51e-3	3.97e-2
	HTP	4.37	2.51e-3	3.97e-2
	CoSaMP	29.0	5.10e-3	8.07e-2
25000	PDASC(50,1)	4.37	2.50e-3	3.99e-2
	PDASC(100,5)	8.30	2.49e-3	3.99e-2
	OMP	157	2.49e-3	3.99e-2
	GreedyGP	191	2.31e-2	1.08e+0
	AIHT	20.1	2.49e-3	3.99e-2
	HTP	7.36	2.49e-3	3.99e-2
	CoSaMP	54.4	5.10e-3	7.98e-2
30000	PDASC(50,1)	7.06	2.48e-3	4.19e-2
	PDASC(100,5)	12.3	2.48e-3	4.18e-2
	OMP	265	2.48e-3	4.18e-2
	GreedyGP	325	2.09e-2	1.08e+0
	AIHT	26.6	2.48e-3	4.18e-2
	HTP	10.5	2.48e-3	4.18e-2
	CoSaMP	85.6	5.00e-3	9.44e-2

Acknowledgements

The authors are grateful to the anonymous referees for their constructive comments, which have led to an improvement in the quality of the paper. The research of Y. Jiao is partially supported by National Science Foundation of China No. 61179039, No. 11301545 and National Key Basic Research Development Program (973 Program) of China (Grant No. 2011CB707100), the work of B. Jin is supported by NSF Grant DMS-1319052, and that of X. Lu is partially supported by National Science Foundation of China No. 11101316, No. 91230108 and No. 11471253.

Table 4.3: Numerical results (CPU time and errors) for large-scale problems, with partial DCT sensing matrix Ψ, of size $p = 2^{13}$, 2^{14}, 2^{15}, 2^{16}, 2^{17}, $n = \lfloor p/4 \rfloor$, $T = \lfloor n/3 \rfloor$. The dynamical range R is $R = 100$, and the noise variance σ is $\sigma =$ 1e-2.

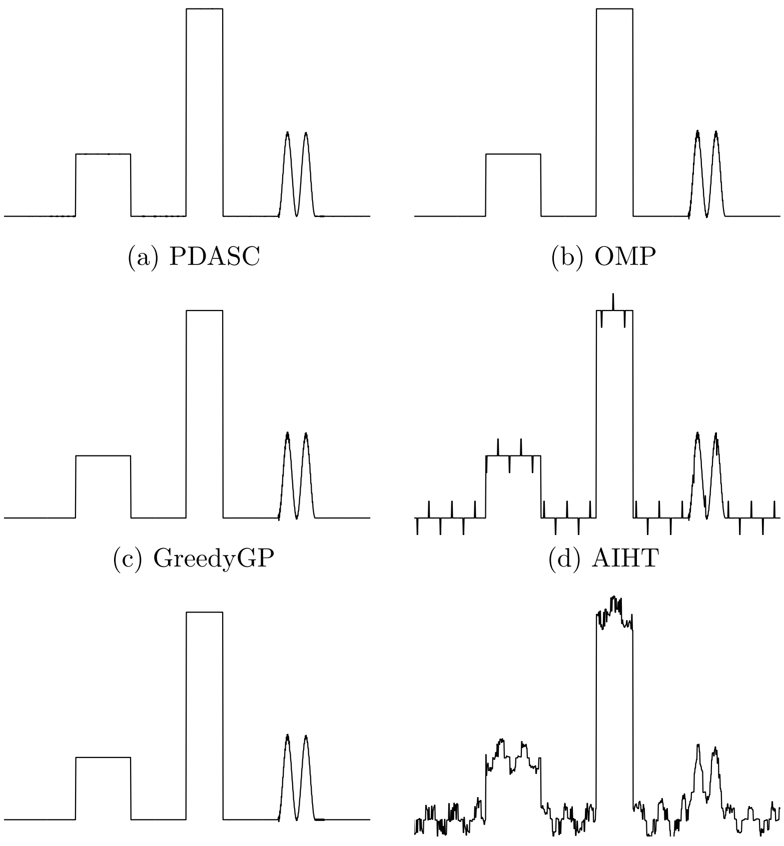
p	method	time (s)	Relative ℓ^2 error	Absolute ℓ^∞ error
$p = 2^{13}$	PDASC(50,1)	0.21	7.09e-4	7.93e-2
	PDASC(100,5)	0.39	7.08e-4	7.89e-2
	OMP	2.26	7.08e-4	7.93e-2
	GreedyGP	0.74	9.99e-4	1.68e-1
	AIHT	0.30	7.08e-4	7.93e-2
	HTP	0.32	7.08e-4	7.87e-2
	CoSaMP	0.50	3.63e-1	3.24e+1
2^{14}	PDASC(50,1)	0.35	6.97e-4	8.60e-2
	PDASC(100,5)	0.66	6.95e-4	8.52e-2
	OMP	11.3	6.95e-4	8.49e-2
	GreedyGP	2.52	9.01e-4	2.49e-1
	AIHT	0.48	6.95e-4	8.50e-2
	HTP	0.52	6.96e-4	8.57e-2
	CoSaMP	0.85	3.87e-1	3.63e+1
2^{15}	PDASC(50,1)	0.64	7.16e-4	8.20e-2
	PDASC(100,5)	1.20	7.83e-4	4.58e-1
	OMP	66.8	9.96e-4	1.01e+0
	GreedyGP	9.50	7.98e-4	1.90e-1
	AIHT	0.98	7.14e-4	8.23e-2
	HTP	0.97	7.15e-4	8.24e-2
	CoSaMP	1.53	3.79e-1	3.71e+1
2^{16}	PDASC(50,1)	1.23	7.43e-4	2.75e-1
	PDASC(100,5)	2.30	7.37e-4	1.81e-1
	OMP	423	1.11e-3	1.03e+0
	GreedyGP	34.2	8.39e-4	5.20e-1
	AIHT	1.87	7.07e-4	8.80e-2
	HTP	2.12	7.08e-4	8.78e-2
	CoSaMP	2.77	3.87e-1	3.93e+1
2^{17}	PDASC(50,1)	3.04	7.47e-4	3.73e-1
	PDASC(100,5)	6.29	7.43e-4	1.95e-2
	OMP	3.17e+3	1.11e-3	1.03e+0
	GreedyGP	200	7.96e-4	5.99e-1
	AIHT	4.76	7.13e-4	9.86e-2
	HTP	4.90	7.14e-4	9.88e-2
	CoSaMP	7.67	3.91e-1	4.25e+1

References

[1] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2, Ser. A):91–129, 2013.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with Sparsity-Inducing Penalties. *Found. Trend. Mach. Learn.*, 4(1):1–106, 2012.

[3] T. Blumensath. Accelerated iterative hard thresholding. *Sig. Proc.*, 92(3):752–756, 2012.



(a) PDASC (b) OMP

(c) GreedyGP (d) AIHT

(e) HTP (f) CoSaMP

Figure 5: Reconstruction results of the one-dimension signal.

[4] T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Trans. Signal Proc.*, 56(6):2370–2382, 2008.

[5] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.*, 14(5-6):629–654, 2008.

[6] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.

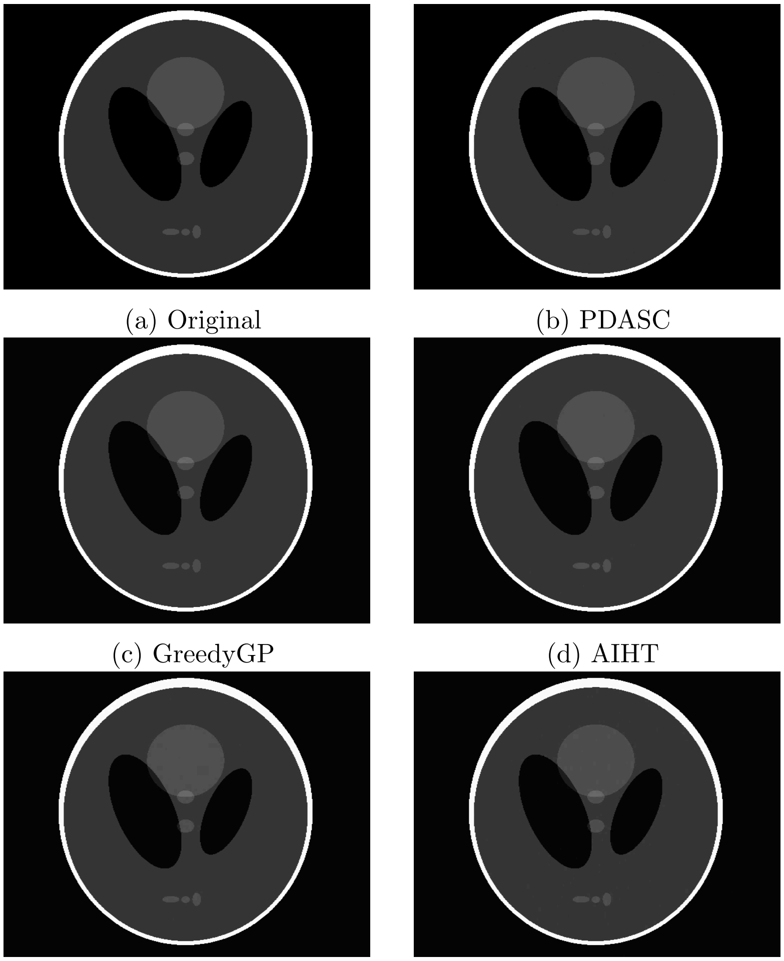
[7] T. Blumensath and M. E. Davies. Stagewise weak gradient pursuits. *IEEE Trans. Signal Process.*, 57(11):4333–4346, 2009.

[8] T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inform. Theory*, 57(7):4680–4688, 2011.

[9] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[10] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.



(a) Original (b) PDASC

(c) GreedyGP (d) AIHT

(e) HTP (f) CoSaMP

Figure 6: Reconstruction results of the two-dimensional Logan-Shepp image.

[12] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, Berlin, 2011.

[13] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, 55(5):2230–2249, 2009.

[14] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran. Sparsity adaptive matching pursuit algorithm for practical compressed sensing. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 581–587. IEEE, 2008.

[15] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[16] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7):2845–2862, 2001.

[17] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 58(2):1094–1121, 2012.

Table 4.4: One-dimensional signal: $n = 665$, $p = 1024$, $T = 247$, $\sigma = 1\text{e-}4$.

method	CPU time (s)	PSNR
PDASC	0.49	53
OMP	1.45	49
GreedyGp	0.76	49
AIHT	0.58	34
HTP	0.40	51
CoSaMP	0.91	26

Table 4.5: Two-dimensional image: $n = 29729$, $p = 262144$, $T = 8450$, $\sigma = 1\text{e-}3$.

method	CPU time (s)	PSNR
PDASC	22.33	62
GreedyGP	448.7	63
AIHT	25.77	62
HTP	20.51	58
CoSaMP	48.74	56

[18] Q. Fan, Y. Jiao, and X. Lu. A primal dual active set algorithm with continuation for compressed sensing. preprint, arXiv:1312.7039, 2013.

[19] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6):2543–2563, 2011.

[20] E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for l_1-minimization: methodology and convergence. *SIAM J. Optim.*, 19(3):1107–1130, 2008.

[21] S. Huang and J. Zhu. Recovery of sparse signals using OMP and its variants: convergence analysis based on RIP. *Inverse Problems*, 27(3):035003, 14 pp., 2011.

[22] K. Ito and B. Jin. *Inverse Problems: Tikhonov Theory and Algorithms.* World Scientific, Singapore, 2014.

[23] K. Ito, B. Jin, and T. Takeuchi. A regularization parameter for nonsmooth Tikhonov regularization. *SIAM J. Sci. Comput.*, 33(3):1415–1438, 2011.

[24] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control.* SIAM, Philadelphia, PA, 2008.

[25] K. Ito and K. Kunisch. A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse Problems*, 30(1):015001, 23 pp, 2014.

[26] Y. Jiao, B. Jin, and X. Lu. A primal dual active set algorithm for a class of nonconvex sparsity optimization. preprint, arXiv:1310.1147, 2013.

[27] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM J. Optim.*, 23(4):2448–2478, 2013.

[28] Q. Mo and Y. Shen. A remark on the restricted isometry property in orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 58(6):3654–3656, 2012.

[29] D. Needell and J. A. Tropp. CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2009.

[30] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.*, 9(3):317–334, 2009.

[31] M. Nikolova. Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer. *SIAM J. Imaging Sci.*, 6(2):904–937, 2013.

[32] M. Nikolova. Relationship between the optimal solutions of least squares regularized with ℓ_0-norm and constrained by k-sparsity. preprint, available at http://hal.archives-ouvertes.fr/hal-00944006/, 2014.

[33] N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):123–231, 2014.

[34] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The 27th Asilomar Conference on*, volume 1, pages 40–44. IEEE.

[35] M. C. Robini and I. E. Magnin. Optimization by stochastic continuation. *SIAM J. Imaging Sci.*, 3(4):1096–1121, 2010.

[36] M. C. Robini and P.-J. Reissman. From simulated annealing to stochastic continuation: a new trend in combinatorial optimization. *J. Global Optim.*, 56(1):185–215, 2013.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[38] A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inf. Theory*, 60(2):1248–1259, 2014.

[39] J. Tropp and S. Wright. Computational methods for sparse solution of linear inverse problems. *Proc. IEEE*, 98(6):948–958, 2010.

[40] J. A. Tropp. Greed is good: algorithmics results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(2):2231–2242, 2004.

[41] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.

[42] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[43] S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[44] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM J. Optim.*, 23(2):1062–1091, 2013.

[45] T. Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inform. Theory*, 57(9):6215–6221, 2011.