Journal of Machine Learning Research 19 (2018) 1-37 Submitted 4/17; Revised 6/18; Published 8/18

A Constructive Approach to L_0 Penalized Regression

Jian Huang J.HUANG@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon
Hong Kong, China

Yuling Jiao[*] YULINGJIAOMATH@WHU.EDU.CN
School of Statistics and Mathematics
Zhongnan University of Economics and Law
Wuhan, 430063, China

Yanyan Liu LIUYY@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan, 430072, China

Xiliang Lu[†] XLLV.MATH@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan, 430072, China

Editor: Tong Zhang

Abstract

We propose a constructive approach to estimating sparse, high-dimensional linear regression models. The approach is a computational algorithm motivated from the KKT conditions for the ℓ_0-penalized least squares solutions. It generates a sequence of solutions iteratively, based on support detection using primal and dual information and root finding. We refer to the algorithm as SDAR for brevity. Under a sparse Riesz condition on the design matrix and certain other conditions, we show that with high probability, the ℓ_2 estimation error of the solution sequence decays exponentially to the minimax error bound in $O(\log(R\sqrt{J}))$ iterations, where J is the number of important predictors and R is the relative magnitude of the nonzero target coefficients; and under a mutual coherence condition and certain other conditions, the ℓ_∞ estimation error decays to the optimal error bound in $O(\log(R))$ iterations. Moreover the SDAR solution recovers the oracle least squares estimator within a finite number of iterations with high probability if the sparsity level is known. Computational complexity analysis shows that the cost of SDAR is $O(np)$ per iteration. We also consider an adaptive version of SDAR for use in practical applications where the true sparsity level is unknown. Simulation studies demonstrate that SDAR outperforms Lasso, MCP and two greedy methods in accuracy and efficiency.

Keywords: Geometrical convergence, KKT conditions, nonasymptotic error bounds, oracle property, root finding, support detection

[*]. Also in the Institute of Big Data of Zhongnan University of Economics and Law

[†]. Also in the Hubei Key Laboratory of Computational Science

©2018 Jian Huang, Yuling Jiao, Yanyan Liu, and Xiliang Lu.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v19/17-194.html`.

1. Introduction

Consider the linear regression model

$$y = X\beta^* + \eta \tag{1}$$

where $y \in \mathbb{R}^n$ is a response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix with $\sqrt{n}$-normalized columns, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$ is the vector of the underlying regression coefficients and $\eta \in \mathbb{R}^n$ is a vector of random noises. We focus on the case where $p \gg n$ and the model is sparse in the sense that only a relatively small number of predictors are important. Without any constraints on β^* there exist infinitely many least squares solutions for (1) since it is a highly undetermined linear system when $p \gg n$. These solutions usually over-fit the data. Under the assumption that β^* is sparse in the sense that the number of important nonzero elements of β^* is small relative to n, we can estimate β^* by the solution of the ℓ_0 minimization problem

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|X\beta - y\|_2^2, \quad \text{subject to} \quad \|\beta\|_0 \leq s, \tag{2}$$

where $s > 0$ controls the sparsity level. However, (2) is generally NP hard (Natarajan, 1995; Chen et al., 2014), hence it is not tractable to design a stable and fast algorithm to solve it, especially in high-dimensional settings.

In this paper we propose a constructive approach to approximating the ℓ_0-penalized solution to (1). The approach is a computational algorithm motivated from the necessary KKT conditions for the Lagrangian form of (2). It finds an approximate sequence of solutions to the KKT equations iteratively based on support detection and root finding until convergence is achieved. For brevity, we refer to the proposed approach as SDAR.

1.1 Literature review

Several approaches have been proposed to approximate (2). Among them the Lasso (Tibshirani, 1996; Chen et al., 1998), which uses the ℓ_1 norm of β in the constraint instead of the ℓ_0 norm in (2), is a popular method. Under the irrepresentable condition on the design matrix X and a sparsity assumption on β^*, Lasso is model selection (and sign) consistent (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009). Lasso is a convex minimization problem. Several fast algorithms have been proposed, including LARS (Homotopy) (Osborne et al., 2000; Efron et al., 2004; Donoho and Tsaig, 2008), coordinate descent (Fu, 1998; Friedman et al., 2007; Wu and Lange, 2008), and proximal gradient descent (Agarwal et al., 2012; Xiao and Zhang, 2013; Nesterov, 2013).

However, Lasso tends to overshrink large coefficients, which leads to biased estimates (Fan and Li, 2001; Fan and Peng, 2004). The adaptive Lasso proposed by Zou (2006) and analyzed by Huang et al. (2008b) in high-dimensions can achieve the oracle property under certain conditions, but its requirements on the minimum value of the nonzero coefficients are not optimal. Nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax concave penalty (MCP) (Zhang, 2010a) and the capped ℓ_1 penalty (Zhang, 2010b) were proposed to remedy these problems (but these methods still require a minimum signal strength in order to achieve support recovery).

Although the global minimizers (also certain local minimizers) of these nonconvex regularized models can eliminate the estimation bias and enjoy the oracle property (Zhang and Zhang, 2012), computing the global or local minimizers with the desired statistical properties is challenging, since the optimization problem is nonconvex, nonsmooth and large scale in general.

There are several numerical algorithms for nonconvex regularized problems. The first kind of such methods can be considered a special case (or variant) of minimization maximization algorithm (Lange et al., 2000; Hunter and Li, 2005) or of multi-stage convex relaxation (Zhang, 2010b). Examples include local quadratic approximation (LQA) (Fan and Li, 2001), local linear approximation (LLA) (Zou and Li, 2008), decomposing the penalty into a difference of two convex terms (CCCP) (Kim et al., 2008; Gasso et al., 2009). The second type of methods is the coordinate descent algorithms, including coordinate descent of the Gauss-Seidel version (Breheny and Huang, 2011; Mazumder et al., 2011) and coordinate descent of the Jacobian version, i.e., the iterative thresholding method (Blumensath and Davies, 2008; She, 2009). These algorithms generate a sequence of solutions at which the objective functions are nonincreasing, but the convergence of the sequence itself is generally unknown. Moreover, if the sequence generated from multi-stage convex relaxation (starts from a Lasso solution) converges, it converges to some stationary point which may enjoy certain oracle statistical properties with the cost of a Lasso solver per iteration (Zhang, 2010b; Fan et al., 2014). Huang et al. (2018) proposed a globally convergent primal dual active set algorithm for a class of nonconvex regularized problems. Recently, there has been much effort to show that CCCP, LLA and the path following proximal-gradient method can track the local minimizers with the desired statistical properties (Wang et al., 2013; Fan et al., 2014; Wang et al., 2014; Loh and Wainwright, 2015).

Another line of research concerns the greedy methods such as the orthogonal matching pursuit (OMP) (Mallat and Zhang, 1993) for solving (2) approximately. The main idea is to iteratively select one variable with the strongest correlation with the current residual at a time. Roughly speaking, the performance of OMP can be guaranteed if the small submatrices of X are well conditioned like orthogonal matrices (Tropp, 2004; Donoho et al., 2006; Cai and Wang, 2011; Zhang, 2011b). Fan and Lv (2008) proposed a marginal correlation learning method called sure independence screening (SIS), see also Huang et al. (2008a) for an equivalent formulation that uses penalized univariate regression for screening. Fan and Lv (2008) recommended an iterative SIS to improve the finite-sample performance. As they discussed the iterative SIS also uses the core idea of OMP but it can select more features at each iteration. There are several more recently developed greedy methods aimed at selecting several variables a time or removing variables adaptively, such as iterative hard thresholding (IHT) (Blumensath and Davies, 2009; Jain et al., 2014) or hard thresholding gradient descent (GraDes) (Garg and Khandekar, 2009), adaptive forward-backward selection (FoBa) (Zhang, 2011a).

Liu and Wu (2007) proposed a Mixed Integer Optimization (MIO) approach for solving penalized classification and regression problems with a penalty that is a combination of ℓ_0 and ℓ_1 penalties. However, they only considered low-dimensional problems with p in the 10s and n in the 100s. Bertsimas et al. (2016) also considered an MIO approach for solving the best subset selection problem in linear regression with a possible side constraint. Their approach can solve problems with moderate sample sizes and moderate dimensions in minutes, for example, for $(n, p) \approx (100, 1000)$ or $(n, p) \approx (1000, 100)$. For the $p > n$ examples, the authors carried out all the computations on Columbia University's high performance computing facility using a commercial MIO solver Gurobi (Gurobi Optimization, 2015). In comparison, our proposed approach can deal with high-dimensional models. For the examples we consider in our simulation studies with $(n, p) = (5000, 50000)$, it can find the solution in seconds on a personal laptop computer.

1.2 Contributions

SDAR is a new approach for fitting sparse, high-dimensional regression models. Compared with the penalized methods, SDAR generates a sequence of solutions $\{\beta^k, k \geq 1\}$ to the KKT system of the ℓ_0 penalized criterion, which can be viewed as a primal-dual active set method for solving the ℓ_0 regularized least squares problem with a changing regularization parameter λ in each iteration (this will be explained in detail in Section 2).

We show that SDAR achieves sharp estimation error bounds within a finite number of iterations. Specifically, we show that: (a) under a sparse Riesz condition on X and a sparsity assumption on β^*, $\|\beta^k - \beta^*\|_2$ achieves the minimax error bound up to a constant factor with high probability in $O(\sqrt{J}\log(R))$ iterations, where J is the number of important predictors and R is the relative magnitude of the nonzero target coefficients (the exact definitions of J and R are given in Section 3); (b) under a mutual coherence condition on X and a sparsity assumption on β^*, the $\|\beta^k - \beta^*\|_\infty$ achieves the optimal error bound $O(\sigma\sqrt{\log(p)/n})$ in $O(\log(R))$ iterations; (c) under the conditions in (a) and (b), with high probability, β^k coincides with the oracle least squares estimator in $O(\sqrt{J}\log(R))$ and $O(\log(R))$ iterations, respectively, if J is available and the minimum magnitude of the nonzero elements of β^* is of the order $O(\sigma\sqrt{2\log(p)/n})$, which is the optimal magnitude of detectable signal.

An interesting aspect of the result in (b) is that the number of iterations for SDAR to achieve the optimal error bound is $O(\log(R))$, which does not depend on the underlying sparsity level. This is an appealing feature for the problems with a large triple (n, p, J). We also analyze the computational cost of SDAR and show that it is $O(np)$ per iteration, comparable to the existing penalized methods and the greedy methods.

In summary, the main contributions of this paper are as follows.

- We propose a new approach to fitting sparse, high-dimensional regression models. The approach seeks to directly approximate the solutions to the KKT equations for the ℓ_0 penalized problem.

- We show that the sequence of solutions $\{\beta^k, k \geq 1\}$ generated by the SDAR achieves sharp error bounds within a finite number of iterations.

- We also consider an adaptive version of SDAR, or simply ASDAR, by tuning the size of the fitted model based on a data driven procedure such as the BIC. Our simulation studies demonstrate that SDAR/ASDAR outperforms the Lasso, MCP and several greedy methods in terms of accuracy and efficiency in the generating models we considered.

1.3 Notation

For a column vector $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$, denote its q-norm by $\|\beta\|_q = (\sum_{i=1}^p |\beta_i|^q)^{1/q}, q \in [1, \infty]$, and its number of nonzero elements by $\|\beta\|_0$. Let $\mathbf{0}$ denote a column vector in $\mathbb{R}^p$ or a matrix whose elements are all 0. Let $S = \{1, 2, ..., p\}$. For any A and $B \subseteq S$ with length $|A|$ and $|B|$, let $\beta_A = (\beta_i, i \in A) \in \mathbb{R}^{|A|}$, $X_A = (X_i, i \in A) \in \mathbb{R}^{n \times |A|}$, and let $X_{AB} \in \mathbb{R}^{|A| \times |B|}$ be a submatrix of X whose rows and columns are listed in A and B, respectively. Let $\beta|_A \in \mathbb{R}^p$ be a vector with its i-th element $(\beta|_A)_i = \beta_i \mathbf{1}(i \in A)$, where $\mathbf{1}(\cdot)$ is the indicator function. Denote the support of β by $\mathrm{supp}(\beta)$. Denote $A^* = \mathrm{supp}(\beta^*)$ and $K = \|\beta^*\|_0$. Let $\|\beta\|_{k,\infty}$ and $|\beta|_{\min}$ be the kth largest elements (in absolute value) and the minimum absolute value of β, respectively. Denote the operator norm of X induced by the vector 2-norm by $\|X\|$. Let $\mathbb{I}$ be an identity matrix.

1.4 Organization

In Section 2 we develop the SDAR algorithm based on the necessary conditions for the ℓ_0 penalized solutions. In Section 3 we establish the nonasymptotic error bounds of the SDAR solutions. In Section 4 we describe the adaptive SDAR, or ASDAR. In Section 5 we analyze the computational complexity of SDAR and ASDAR. In Section 6 we compare SDAR with several greedy methods and a screening method. In Section 7 we conduct simulation studies to evaluate the performance of SDAR/ASDAR and compare it with Lasso, MCP, FoBa and DesGras. We conclude in Section 8 with some final remarks. The proofs are given in the Appendix.

2. Derivation of SDAR

Consider the Lagrangian form of the ℓ_0 regularized minimization problem (2),

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|X\beta - y\|_2^2 + \lambda \|\beta\|_0. \tag{3}$$

Lemma 1 *Let $\beta^\diamond$ be a coordinate-wise minimizer of (3). Then $\beta^\diamond$ satisfies:*

$$\begin{cases} d^\diamond = X'(y - X\beta^\diamond)/n, \\ \beta^\diamond = H_\lambda(\beta^\diamond + d^\diamond), \end{cases} \tag{4}$$

where $H_\lambda(\cdot)$ is the hard thresholding operator defined by

$$(H_\lambda(\beta))_i = \begin{cases} 0, & \text{if } |\beta_i| < \sqrt{2\lambda}, \\ \beta_i, & \text{if } |\beta_i| \geq \sqrt{2\lambda}. \end{cases} \tag{5}$$

Conversely, if $\beta^\diamond$ and $d^\diamond$ satisfy (4)*, then $\beta^\diamond$ is a local minimizer of* (3)*.*

Remark 2 *Lemma 1 gives the KKT condition of the ℓ_0 regularized minimization problem (3), which is also derived in Jiao et al. (2015). Similar results for SCAD, MCP and capped-ℓ_1 regularized least squares models can be derived by replacing the hard thresholding operator in* (4) *with their corresponding thresholding operators, see Huang et al. (2018) for details.*

Let $A^\diamond = \text{supp}(\beta^\diamond)$ and $I^\diamond = (A^\diamond)^c$. Suppose that the rank of $X_{A^\diamond}$ is $|A^\diamond|$. From the definition of $H_\lambda(\cdot)$ and (4) it follows that

$$A^\diamond = \Big\{i \in S \big| \, |\beta_i^\diamond + d_i^\diamond| \geq \sqrt{2\lambda}\Big\}, \quad I^\diamond = \Big\{i \in S \big| \, |\beta_i^\diamond + d_i^\diamond| < \sqrt{2\lambda}\Big\},$$

and

$$\begin{cases} \beta_{I^\diamond}^\diamond = \mathbf{0}, \\ d_{A^\diamond}^\diamond = \mathbf{0}, \\ \beta_{A^\diamond}^\diamond = (X'_{A^\diamond} X_{A^\diamond})^{-1} X'_{A^\diamond} y, \\ d_{I^\diamond}^\diamond = X'_{I^\diamond}(y - X_{A^\diamond}\beta_{A^\diamond}^\diamond)/n. \end{cases}$$

We solve this system of equations iteratively. Let $\{\beta^k, d^k\}$ be the solution at the kth iteration. We approximate $\{A^\diamond, I^\diamond\}$ by

$$A^k = \Big\{i \in S \big| |\beta_i^k + d_i^k| \geq \sqrt{2\lambda}\Big\}, \quad I^k = (A^k)^c. \tag{6}$$

Then we can obtain an updated approximation pair $\{\beta^{k+1}, d^{k+1}\}$ by

$$\begin{cases} \beta_{I^k}^{k+1} = \mathbf{0}, \\ d_{A^k}^{k+1} = \mathbf{0}, \\ \beta_{A^k}^{k+1} = (X'_{A^k} X_{A^k})^{-1} X'_{A^k} y, \\ d_{I^k}^{k+1} = X'_{I^k}(y - X_{A^k}\beta_{A^k}^{k+1})/n. \end{cases} \tag{7}$$

Now suppose we want the support of the solutions to have the size T, where $T \geq 1$ is a given integer. We can choose

$$\sqrt{2\lambda^k} \triangleq \|\beta^k + d^k\|_{T,\infty} \tag{8}$$

in (6). With this choice of λ, we have $|A^k| = T, k \geq 1$. Then with an initial β^0 and using (6) and (7) with the λ^k in (8), we obtain a sequence of solutions $\{\beta^k, k \geq 1\}$.

There are two key aspects of SDAR. In (6) we detect the support of the solution based on the sum of the primal (β^k) and dual (d^k) approximations and, in (7) we calculate the least squares solution on the detected support. Therefore, SDAR can be considered an iterative method for solving the KKT equations (4) with an important modification: a different λ value given in (8) in each step of the iteration is used. Thus we can also view SDAR as a method that combines adaptive thresholding using primal and dual information and least-squares fitting. We summarize SDAR in Algorithm 1.

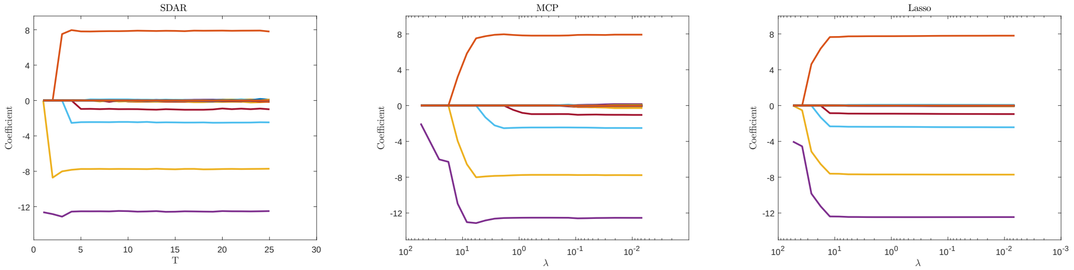

Figure 1: The solution paths of SDAR, MCP and Lasso. We see that large components were selected in by SDAR gradually when T increases. This is similar to Lasso and MCP as λ decreases.

As an example, Figure 1 shows the solution path of SDAR with $T = 1, 2, \ldots, 5K$ along with the MCP and the Lasso paths on $5K$ different λ values for a data set generated from a model with $(n = 50, p = 100, K = 5, \sigma = 0.3, \rho = 0.5, R = 10)$, which will be described in Section 7. The Lasso path is computed using LARS (Efron et al., 2004). Note that the SDAR path is a function of the fitted model size $T = 1, \ldots, L$, where L is the size of the largest fitted model. In comparison, the paths of MCP and Lasso are functions of the penalty parameter λ in a prespecified interval. In this example, when $T \leq K$, SDAR selects the first T largest components of β^* correctly. When $T > K$, there will be spurious elements included in the estimated support, the exact number of such elements is $T - K$. In Figure 1, the estimated coefficients of the spurious elements are close to zero.

Algorithm 1 Support detection and root finding (SDAR)

```
Require: $\beta^0$, $d^0 = X'(y - X\beta^0)/n$, $T$; set $k = 0$.
for $k = 0, 1, 2, \cdots$ do
    $A^k = \{i \in S \big| |\beta_i^k + d_i^k| \geq \|\beta^k + d^k\|_{T,\infty}\}, I^k = (A^k)^c$
    $\beta_{I^k}^{k+1} = \mathbf{0}$
    $d_{A^k}^{k+1} = \mathbf{0}$
    $\beta_{A^k}^{k+1} = (X'_{A^k} X_{A^k})^{-1} X'_{A^k} y$
    $d_{I^k}^{k+1} = X'_{I^k}(y - X_{A^k}\beta_{A^k}^{k+1})/n$
    if $A^{k+1} = A^k$, then
        Stop and denote the last iteration by $\beta_{\hat{A}}, \beta_{\hat{I}}, d_{\hat{A}}, d_{\hat{I}}$
    else
        $k = k + 1$
    end if
end for
Ensure: $\hat{\beta} = (\beta'_{\hat{A}}, \beta'_{\hat{I}})'$ as the estimate of $\beta^*$.
```

Remark 3 *If $A^{k+1} = A^k$ for some k we stop SDAR since the sequences generated by SDAR will not change. Under certain conditions, we will show that $A^{k+1} = A^k = supp(\beta^*)$ if k is large enough, i.e., the stop condition in SDAR will be active and the output is the oracle estimator when it stops.*

3. Nonasymptotic error bounds

In this section we present the nonasymptotic ℓ_2 and ℓ_∞ error bounds for the solution sequence generated by SDAR as given in Algorithm 1.

We say that X satisfies the sparse Rieze condition (SRC) (Zhang and Huang, 2008; Zhang, 2010a) with order s and spectrum bounds $\{c_-(s), c_+(s)\}$ if

$$0 < c_-(s) \leq \frac{\|X_A u\|_2^2}{n\|u\|_2^2} \leq c_+(s) < \infty, \forall\, 0 \neq u \in \mathbb{R}^{|A|} \text{ with } A \subset S \text{ and } |A| \leq s.$$

We denote this condition by $X \sim \text{SRC}\{s, c_-(s), c_+(s)\}$. The SRC gives the range of the spectrum of the diagonal sub-matrices of the Gram matrix $G = X'X/n$. The spectrum of the off diagonal sub-matrices of G can be bounded by the sparse orthogonality constant $\theta_{a,b}$ defined as the smallest number such that

$$\theta_{a,b} \geq \frac{\|X_A' X_B u\|_2}{n\|u\|_2}, \forall\, 0 \neq u \in \mathbb{R}^{|B|} \text{ with } A, B \subset S, |A| \leq a, |B| \leq b, \text{ and } A \cap B = \emptyset.$$

Another useful quantity is the mutual coherence μ defined as $\mu = \max_{i \neq j} |G_{i,j}|$, which characterizes the minimum angle between different columns of $X/\sqrt{n}$. Some useful properties of these quantities are summarized in Lemma 20 in the Appendix.

In addition to the regularity conditions on the design matrix, another key condition is the sparsity of the regression parameter β^*. The usual sparsity condition is to assume that the regression parameter β_i^* is either nonzero or zero and that the number of nonzero coefficients is relatively small. This strict sparsity condition is not realistic in many problems. Here we allow that β^* may not be strictly sparse but most of its elements are small. Let $A_J^* = \{i \in S : |\beta_i^*| \geq \|\beta^*\|_{J,\infty}\}$ be the set of the indices of the first J largest components of β^*. Typically, we have $J \ll n$. Let

$$R = \frac{\bar{M}}{\bar{m}}, \tag{9}$$

where $\bar{m} = \min\{|\beta_i^*|, i \in A_J^*\}$ and $\bar{M} = \max\{|\beta_i^*|, i \in A_J^*\}$. Since $\beta^* = \beta^*|_{A_J^*} + \beta^*|_{(A_J^*)^c}$, we can transform the non-exactly sparse model (1) to the following exactly sparse model by including the small components of β^* in the noise,

$$y = X\bar{\beta}^* + \bar{\eta}, \tag{10}$$

where

$$\bar{\beta}^* = \beta^*|_{A_J^*} \text{ and } \bar{\eta} = X\beta^*|_{(A_J^*)^c} + \eta. \tag{11}$$

Let $R_J = \|\beta^*|_{(A_J^*)^c}\|_2 + \|\beta^*|_{(A_J^*)^c}\|_1/\sqrt{J}$, which is a measure of the magnitude of the small components of β^* outside A_J^*. Of course, $R_J = 0$ if β^* is exactly K-sparse with $K \leq J$. Without loss of generality, we let $J = K$, $m = \bar{m}$ and $M = \bar{M}$ for simplicity if β^* is exactly K-sparse.

Let $\beta^{J,o}$ be the oracle estimator defined as $\beta^{J,o} = \arg\min_\beta\{\frac{1}{2n}\|y - X\beta\|_2^2, \beta_j = 0, j \notin A_J^*\}$, that is, $\beta_{A_J^*}^{J,o} = X_{A_J^*}^\dagger y$ and $\beta_{(A_J^*)^c}^{J,o} = \mathbf{0}$, where $X_{A_J^*}^\dagger$ is the generalized inverse of $X_{A_J^*}$ and equals to $(X_{A_J^*}' X_{A_J^*})^{-1} X_{A_J^*}'$ if $X_{A_J^*}$ is of full column rank. So $\beta^{J,o}$ is obtained by keeping the predictors corresponding to the J largest components of β^* in the model and dropping the other predictors. Obviously, $\beta^{J,o} = \beta^o$ if β^* is exactly K-sparse, where $\beta_{A^*}^o = X_{A^*}^\dagger y, \beta_{(A^*)^c}^o = \mathbf{0}$.

3.1 ℓ_2 error bounds

Let $1 \le T \le p$ be a given integer used in Algorithm 1. We require the following basic assumptions on the design matrix X and the error vector η.

(A1) The input integer T used in Algorithm 1 satisfies $T \ge J$.

(A2) For the input integer T used in Algorithm 1, $X \sim \mathrm{SRC}\{2T, c_-(2T), c_+(2T)\}$.

(A3) The random errors $\eta_1, \ldots, \eta_n$ are independent and identically distributed with mean zero and sub-Gaussian tails, that is, there exists a $\sigma \ge 0$ such that $E[\exp(t\eta_i)] \le \exp(\sigma^2 t^2/2)$ for $t \in \mathbb{R}^1$, $i = 1, \ldots, n$.

Let

$$\gamma = \frac{2\theta_{T,T} + (1+\sqrt{2})\theta_{T,T}^2}{c_-(T)^2} + \frac{(1+\sqrt{2})\theta_{T,T}}{c_-(T)}.$$

Define

$$h_2(T) = \max_{A \subseteq S : |A| \le T} \|X_A' \bar{\eta}\|_2 / n, \tag{12}$$

where $\bar{\eta}$ is defined in (11).

Theorem 4 *Let T be the input integer used in Algorithm 1, where $1 \le T \le p$. Suppose $\gamma < 1$.*

(i) Assume (A1) and (A2) hold. We have

$$\|\bar{\beta}^*|_{A_J^* \setminus A^{k+1}}\|_2 \le \gamma^{k+1}\|\bar{\beta}^*\|_2 + \frac{\gamma}{(1-\gamma)\theta_{T,T}} h_2(T), \tag{13}$$

$$\|\beta^{k+1} - \bar{\beta}^*\|_2 \le b_1 \gamma^k \|\bar{\beta}^*\|_2 + b_2 h_2(T), \tag{14}$$

where

$$b_1 = 1 + \frac{\theta_{T,T}}{c_-(T)} \quad \text{and} \quad b_2 = \frac{\gamma}{(1-\gamma)\theta_{T,T}} b_1 + \frac{1}{c_-(T)}. \tag{15}$$

(ii) Assume (A1)-(A3) hold. Then for any $\alpha \in (0, 1/2)$, with probability at least $1 - 2\alpha$,

$$\|\bar{\beta}^*|_{A_J^* \setminus A^{k+1}}\|_2 \le \gamma^{k+1}\|\bar{\beta}^*\|_2 + \frac{\gamma}{(1-\gamma)\theta_{T,T}} \varepsilon_1, \tag{16}$$

$$\|\beta^{k+1} - \bar{\beta}^*\|_2 \le b_1 \gamma^k \|\bar{\beta}^*\|_2 + b_2 \varepsilon_1, \tag{17}$$

where

$$\varepsilon_1 = c_+(J) R_J + \sigma\sqrt{T}\sqrt{2\log(p/\alpha)/n}. \tag{18}$$

Remark 5 *Part (i) of Theorem 4 establishes the ℓ_2 bounds for the approximation errors of the solution sequence generated by the SDAR algorithm at the $(k+1)$th iteration for a general noise vector η. In particular, (13) gives the ℓ_2 bound of the elements in A_J^* not included in the active set in the $(k+1)$th iteration, and (14) provides an upper bound for the ℓ_2 estimation error of β^{k+1}. These error bounds decay geometrically to the model error measured by $h_2(T)$ up to a constant factor. Part (ii) specializes these results to the case where the noise terms are sub-Gaussian.*

Remark 6 *Assumption (A1) is necessary for SDAR to select at least J nonzero features. The SRC in (A2) has been used in the analysis of the Lasso and MCP (Zhang and Huang, 2008; Zhang, 2010a). Sufficient conditions are provided for a design matrix to satisfy the SRC in Propositions 4.1 and 4.2 in Zhang and Huang (2008). For example, the SRC would follow from a mutual coherence condition. Let $c(T) = (1 - c_-(2T)) \vee (c_+(2T) - 1)$, which is closely related to the the RIP (restricted isometry property) constant δ_{2T} for X (Candes and Tao, 2005). By* (43) *in the Appendix, it can be verified that a sufficient condition for $\gamma < 1$ is $c(T) \leq 0.1599$, i.e., $c_+(2T) \leq 1.1599$, $c_-(2T) \geq 0.8401$. The sub-Gaussian condition (A3) is often assumed in the literature on sparse estimation and slightly weaker than the standard normality assumption. It is used to calculate the tail probabilities of certain maximal functions of the noise vector η.*

Remark 7 *Several greedy algorithms have also been studied under the assumptions related to the sparse Riesz condition. For example, Zhang (2011b) studied OMP under the condition $c_+(T)/c_-(31T) \leq 2$. Zhang (2011a) analyzed the forward-backward greedy algorithm (FoBa) under the condition $8(T+1) \leq (s-2)Tc_-^2(sT)$, where $s > 0$ is a properly chosen parameter. GraDes has been analyzed under the RIP condition $\delta_{2T} \leq 1/3$ (Garg and Khandekar, 2009). These conditions and (A2) are related but do not imply each other. The order of ℓ_2-norm estimation error of SDAR is at least as good as that of the above mentioned greedy methods since it achieves the minimax error bound, see, Remark 10 below. A high level comparison between SDAR and the greedy algorithms will be given in Section 6.*

Corollary 8 *(i) Suppose (A1) and (A2) hold. Then*

$$\|\beta^k - \bar{\beta}^*\|_2 \leq ch_2(T) \quad \textit{if} \quad k \geq \log_{\frac{1}{\gamma}} \frac{\sqrt{J}\bar{M}}{h_2(T)}, \tag{19}$$

where $c = b_1 + b_2$ with b_1 and b_2 defined in (15).

Furthermore, assume $\bar{m} \geq \frac{\gamma h_2(T)}{(1-\gamma)\theta_{T,T}\xi}$ for some $0 < \xi < 1$, then we have

$$A^k \supseteq A_J^* \quad \textit{if} \quad k \geq \log_{\frac{1}{\gamma}} \frac{\sqrt{J}R}{1-\xi}. \tag{20}$$

(ii) Suppose (A1)-(A3) hold. Then, for any $\alpha \in (0, 1/2)$, with probability at least $1 - 2\alpha$, we have

$$\|\beta^k - \bar{\beta}^*\|_2 \leq c\varepsilon_1 \quad \textit{if} \quad k \geq \log_{\frac{1}{\gamma}} \frac{\sqrt{J}\bar{M}}{\varepsilon_1}, \tag{21}$$

where ε_1 is defined in (18). Furthermore, assume $\bar{m} \geq \frac{\varepsilon_1\gamma}{(1-\gamma)\theta_{T,T}\xi}$ for some $0 < \xi < 1$, then with probability at least $1 - 2\alpha$, we have

$$A^k \supseteq A_J^* \quad \textit{if} \quad k \geq \log_{\frac{1}{\gamma}} \frac{\sqrt{J}R}{1-\xi}. \tag{22}$$

(iii) Suppose β^* is exactly K-sparse. Let $T = K$ in SDAR. Suppose (A1)-(A3) hold and $m \geq \frac{\gamma}{(1-\gamma)\theta_{T,T}\xi}\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}$ for some $0 < \xi < 1$, we have with probability

at least $1-2\alpha$*,* $A^k = A^{k+1} = A^*$ *if* $k \geq \log_{\frac{1}{\gamma}}(\sqrt{K}R/(1-\xi))$*, i.e., with at most* $O(\log\sqrt{K}R)$ *iterations, SDAR stops and the output is the oracle least squares estimator* β^o*.*

Remark 9 *Parts (i) and (ii) in Corollary 8 show that the SDAR solution sequence achieves the minimax* ℓ_2 *error bound up to a constant factor and its support covers* A^*_J *within a finite number of iterations. In particular, the number of iterations required is* $O(\log(\sqrt{J}R))$*, depending on the sparsity level* J *and the relative magnitude* R *of the coefficients of the important predictors. In the case of exact sparsity with* K *nonzero coefficients in the model, part (iii) provides conditions under which the SDAR solution is the same as the oracle least squares estimator in* $O(\log(\sqrt{K}R))$ *iterations with high probability.*

Remark 10 *Suppose* β^* *is exactly* K*-sparse. In the event* $\|\eta\|_2 \leq \varepsilon$*, part (i) of Corollary 8 implies* $\|\beta^k - \beta^*\|_2 = O(\varepsilon/\sqrt{n})$ *if* k *is sufficiently large. Under certain conditions on the RIP constant of* X*, Candes et al. (2006) showed that* $\|\hat{\beta} - \beta^*\|_2 = O(\varepsilon/\sqrt{n})$*, where* $\hat{\beta}$ *solves*

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \textit{ subject to } \|X\beta - y\|_2 \leq \varepsilon. \tag{23}$$

So the result here is similar to that of Candes et al. (2006) (they assumed the columns of X *are unit-length normalized, here the result is stated for the case where the columns of* X *are* $\sqrt{n}$*-length normalized). However, it is a nontrivial task to solve* (23) *in high-dimensional settings. In comparison, SDAR only involves simple computational steps.*

Remark 11 *If* β^* *is exactly* K*-sparse and* $T = K$*, part (ii) of Corollary 8 implies that SDAR achieves the minimax error bound (Raskutti et al., 2011), that is,*

$$\|\beta^k - \beta^*\|_2 \leq c\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}$$

with high probability if $k \geq \log_{\frac{1}{\gamma}} \frac{\sqrt{K}M}{\sigma\sqrt{T}\sqrt{2\log(p/\alpha)/n}}$*.*

3.2 ℓ_∞ error bounds

We now consider the ℓ_∞ error bounds of SDAR. We replace condition (A2) by

(A2*) The mutual coherence μ of X satisfies $T\mu \leq 1/4$.

Let

$$\gamma_\mu = \frac{(1+2T\mu)T\mu}{1-(T-1)\mu} + 2T\mu \text{ and } c_\mu = \frac{16}{3(1-\gamma_\mu)} + \frac{5}{3}.$$

Define

$$h_\infty(T) = \max_{A\subseteq S:|A|\leq T} \|X'_A\bar{\eta}\|_\infty/n, \tag{24}$$

where $\bar{\eta}$ is defined in (11).

Theorem 12 *Let* T *be the input integer used in Algorithm 1, where* $1 \leq T \leq p$*.*

(i) *Assume (A1) and (A2*) hold. We have*

$$\|\bar{\beta}^*|_{A_J^*\backslash A^{k+1}}\|_\infty < \gamma_\mu^{k+1}\|\bar{\beta}^*\|_\infty + \frac{4}{1-\gamma_\mu}h_\infty(T), \tag{25}$$

$$\|\beta^{k+1} - \bar{\beta}^*\|_\infty < \frac{4}{3}\gamma_\mu^k\|\bar{\beta}^*\|_\infty + \frac{4}{3}(\frac{4}{1-\gamma_\mu}+1)h_\infty(T), \tag{26}$$

(ii) *Assume (A1), (A2*) and (A3) hold. For any $\alpha \in (0, 1/2)$, with probability at least $1 - 2\alpha$,*

$$\|\bar{\beta}^*|_{A_J^*\backslash A^{k+1}}\|_\infty < \gamma_\mu^{k+1}\|\bar{\beta}^*\|_\infty + \frac{4}{1-\gamma_\mu}\varepsilon_2, \tag{27}$$

$$\|\beta^{k+1} - \bar{\beta}^*\|_\infty < \frac{4}{3}\gamma_\mu^k\|\bar{\beta}^*\|_\infty + \frac{4}{3}(\frac{4}{1-\gamma_\mu}+1)\varepsilon_2, \tag{28}$$

where

$$\varepsilon_2 = (1 + (T-1)\mu)R_J + \sigma\sqrt{2\log(p/\alpha)/n}. \tag{29}$$

Remark 13 *Part (i) of Theorem 12 establishes the ℓ_∞ bounds for the approximation errors of the solution sequence at the $(k+1)$th iteration for a general noise vector η. In particular, (25) gives the ℓ_∞ bound of the elements in A_J^* not selected at the $(k+1)$th iteration, and (26) provides an upper bound for the ℓ_∞ estimation error of β^{k+1}. These errors bounds decay geometrically to the model error measured by $h_\infty(T)$ up to a constant factor. Part (ii) specializes these to the case where the noise terms are sub-Gaussian.*

Corollary 14 (i) *Suppose (A1) and (A2*) hold. Then*

$$\|\beta^k - \bar{\beta}^*\|_\infty \leq c_\mu h_\infty(T) \quad \text{if} \quad k \geq \log_{\frac{1}{\gamma_\mu}} \frac{4\bar{M}}{h_\infty(T)}. \tag{30}$$

Furthermore, assume $\bar{m} \geq \frac{4h_\infty(T)}{(1-\gamma_\mu)\xi}$ with $\xi < 1$, then we have

$$A^k \supseteq A_J^* \quad \text{if} \quad k \geq \log_{\frac{1}{\gamma_\mu}} \frac{R}{1-\xi}. \tag{31}$$

(ii) *Suppose (A1), (A2*) and (A3) hold. Then for any $\alpha \in (0, 1/2)$, with probability at least $1 - 2\alpha$,*

$$\|\beta^k - \bar{\beta}^*\|_\infty \leq c_\mu \varepsilon_2 \quad \text{if} \quad k \geq \log_{\frac{1}{\gamma_\mu}} \frac{4\bar{M}}{\varepsilon_2}, \tag{32}$$

where ε_2 is given in (29).

Furthermore, assume $\bar{m} \geq \frac{4\varepsilon_2}{\xi(1-\gamma_\mu)}$ for some $0 < \xi < 1$, then

$$A^k \supseteq A_J^* \quad \text{if} \quad k \geq \log_{\frac{1}{\gamma_\mu}} \frac{R}{1-\xi}. \tag{33}$$

(iii) *Suppose β^* is exactly K-sparse. Let $T = K$ in SDAR. Suppose (A1), (A2*) and (A3) hold and $m \geq \frac{4}{\xi(1-\gamma_\mu)}\sigma\sqrt{2\log(p/\alpha)/n}$ for some $0 < \xi < 1$. We have with probability at least $1-2\alpha$, $A^k = A^{k+1} = A^*$ if $k \geq \log_{\frac{1}{\gamma_\mu}} \frac{R}{1-\xi}$, i.e., with at most $O(\log R)$ iterations, SDAR stops and the output is the oracle least squares estimator β^o.*

Remark 15 *Theorem 4 and Corollary 8 can be derived from Theorem 12 and Corollary 14, respectively, by using the relationship between the ℓ_∞ norm and the ℓ_2 norm. Here we present them separately because (A2) is weaker than (A2*). The stronger assumption (A2*) brings us some new insights into the SDAR, i.e., the sharp ℓ_∞ error bound, based on which we can show that the worst case iteration complexity of SDAR does not depend on the underlying sparsity level, as stated in parts (ii) and (iii) of Corollary 14.*

Remark 16 *The mutual coherence condition $s\mu \leq 1$ with $s \geq 2K-1$ is used in the study of OMP and Lasso under the assumption that β^* is exactly K-sparse. In the noiseless case with $\eta = 0$, Tropp (2004); Donoho and Tsaig (2008) showed that under the condition $(2K-1)\mu < 1$, OMP can recover β^* exactly in K steps. In the noisy case with $\|\eta\|_2 \leq \varepsilon$, Donoho et al. (2006) proved that OMP can recover the true support if $(2K-1)\mu \leq 1-(2\varepsilon/m)$. Cai and Wang (2011) gave a sharp analysis of OMP under the condition $(2K-1)\mu < 1$. The mutual coherence condition $T\mu \leq 1/4$ in (A2*) is a little stronger than those used in the analysis of the OMP. However, under (A2*) we obtain a sharp ℓ_∞ error bound, which is not available for OMP in the literature. Furthermore, Corollary 14 implies that the number of iterations of SDAR does not depend on the sparsity level, which is a surprising result and does not appear in the literature on greedy methods, see Remark 18 below. Lounici (2008); Zhang (2009) derived an ℓ_∞ estimation error bound for the Lasso under the conditions $K\mu < 1/7$ and $K\mu \leq 1/4$, respectively. However, they needed a nontrivial Lasso solver for computing an approximate solution while SDAR only involves simple computational steps.*

Remark 17 *Suppose β^* is exactly K-sparse. Part (ii) of Corollary 14 implies that the sharp error bound*

$$\|\beta^k - \beta^*\|_\infty \leq c_\mu \sigma\sqrt{2\log(p/\alpha)/n} \tag{34}$$

is achieved with high probability if $k \geq \log_{\frac{1}{\gamma_\mu}} \frac{M}{\sigma\sqrt{2\log(p/\alpha)/n}}$.

Remark 18 *Suppose β^* is exactly K-sparse. Part (iii) of Corollary 14 implies that with high probability, the oracle estimator can be recovered in no more than $O(\log R)$ steps if we set $T = K$ in SDAR and the minimum magnitude of the nonzero elements of β^* is $O(\sigma\sqrt{2\log(p)/n})$, which is the optimal magnitude of detectable signals.*

Remark 19 *The number of iterations in Corollary 14 depends on the relative magnitude R, but not the sparsity level K, see Figure 2 for the numerical results supporting this. This improves the result in part (iii) of Corollary 8. This is a surprising result since as far as we know the number of iterations for the greedy methods to recover A^* depends on K, see for example, Garg and Khandekar (2009).*

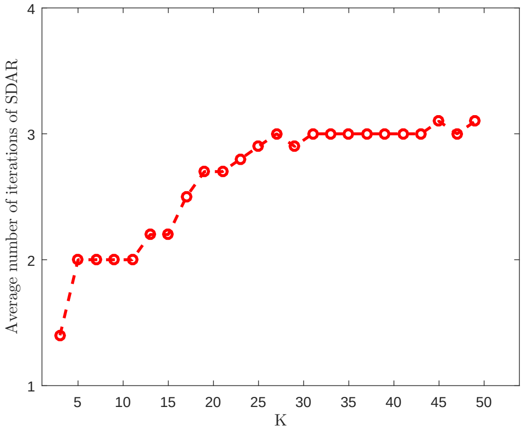

Figure 2: The average number of iterations of SDAR as K increases.

Figure 2 shows the average number of iterations of SDAR with $T = K$ based on 100 independent replications on data sets generated from a model with $(n = 500, p = 1000, K = 3 : 2 : 50, \sigma = 0.01, \rho = 0.1, R = 1)$, which will be described in Section 7.4. We can see that as the sparsity level increases from 3 to 50 the average number of iterations of SDAR remains stable, ranging from 1 to 3, which supports the assertion in Corollary 14. More numerical comparison on number of iterations with greedy methods are shown in Section 7.2.

3.3 A brief high-level description of the proofs

The detailed proofs of Theorems 4 and 12 and their corollaries are given in the Appendix. Here we describe the main ideas behind the proofs and point out the places where the SRC and the mutual coherence condition are used.

SDAR iteratively detects the support of the solution and then solves a least squares problem on the support. Therefore, to study the convergence properties of the sequence generated by SDAR, the key is to show that the sequence of active sets A^k can approximate A_J^* more and more accurately as k increases. Let

$$D(A^k) = \|\bar{\beta}^*|_{A_J^* \backslash A^k}\|, \tag{35}$$

where $\|\cdot\|$ can be either the ℓ_2 norm or the ℓ_∞ norm. This is a measure of the difference between A_J^* and A^k at the kth iteration in terms of the norm of the coefficients in A_J^* but not in A^k. A crucial step is to show that $D(A^k)$ decays geometrically to a value bounded by $h(T)$ up to a constant factor, where $h(T)$ is $h_2(T)$ defined in (12) or $h_\infty(T)$ in (24). Here $h(T)$ is a measure of the intrinsic error due to the noise η and the approximate error in (10). Specifically, much effort is spent on establishing the inequality (Lemma 27 in the Appendix)

$$D(A^{k+1}) \leq \gamma^* D(A^k) + c^* h(T), k = 0, 1, 2, \ldots, \tag{36}$$

where $\gamma^* = \gamma$ for the ℓ_2 results in Theorem 4 and $\gamma^* = \gamma_\mu$ for the ℓ_∞ results in Theorem 12, and $c^* > 0$ is a constant depending on the design matrix. The SRC (A2) and the mutual coherence condition (A2*) play a critical role in establishing (36). Clearly, for this inequality to be useful, we need $0 < \gamma^* < 1$.

Another useful inequality is

$$\|\beta^{k+1} - \bar{\beta}^*\| \leq c_1 D(A^k) + c_2 h(T), \tag{37}$$

where c_1 and c_2 are positive constants depending on the design matrix, see Lemma 23 in the Appendix. The SRC and the mutual coherence condition are needed to establish this inequality for the ℓ_2 norm and the ℓ_∞ norm, respectively. Then combining (36) and (37), we can show part (*i*) of Theorem 4 and part (*i*) of Theorem 12.

The inequalities (36) and (37) hold for any noise vector η. Under the sub-Gaussian assumption for η, $h(T)$ can be controlled by the sum of unrecoverable approximation error R_J and the universal noise level $O(\sigma\sqrt{2\log(p)/n})$ with high probability. This leads to the results in the remaining parts of Theorems 4 and 12, as well as Corollaries 8 and 14.

4. Adaptive SDAR

In practice, because the sparsity level of the model is usually unknown, we can use a data driven procedure to determine an upper bound, T, for the number of important variables, J, used in SDAR (Algorithm 1). The idea is to take T as a tuning parameter, so T plays a role similar to the penalty parameter λ in a penalized method. We can run SDAR from $T = 1$ to a large $T = L$. For example, we can take $L = O(n/\log(n))$ as suggested by Fan and Lv (2008), which is an upper bound of the largest possible model that can be consistently estimated with sample size n. By doing so we obtain a solution path $\{\hat{\beta}(T) : T = 0, 1, \ldots, L\}$, where $\hat{\beta}(0) = 0$, that is, $T = 0$ corresponds to the null model. Then we use a data driven criterion, such as HBIC (Wang et al., 2013), to select a $T = \hat{T}$ and use $\hat{\beta}(\hat{T})$ as the final estimate. The overall computational complexity of this process is $O(Lnp\log(R))$, see Section 5.

We can also compute the path by increasing T along a subsequence of the integers in $[1, L]$, for example, by taking a geometrically increasing subsequence. This will reduce the computational cost, but here we consider the worst-case scenario.

We note that tuning T is no more difficult than tuning a continuous penalty parameter λ in a penalized method. Indeed, we can simply increase T one by one from $T = 0$ to $T = L$ (or along a subsequence). In comparison, in tuning the value of λ based on a pathwise solution over an interval $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max}$ corresponds to the null model and $\lambda_{\min} > 0$ is a small value, we need to determine the grid of λ values on $[\lambda_{\min}, \lambda_{\max}]$ as well as $\lambda_{\min}$. Here $\lambda_{\min}$ corresponds to the largest model on the solution path. In the numerical implementation of the coordinate descent algorithms for the Lasso (Friedman et al., 2007), MCP and SCAD (Breheny and Huang, 2011), $\lambda_{\min} = \alpha\lambda_{\max}$ for a small α, for example, $\alpha = 0.0001$. Determining the value of L is somewhat similar to determining $\lambda_{\min}$. However, L has the meaning of the model size, but the meaning of $\lambda_{\min}$ is less explicit.

We also have the option to stop the iteration early according to other criterions. For example, we can run SDAR by gradually increasing T until the change in the consecutive solutions is smaller than a given value. Candes et al. (2006) proposed to recover β^* based on (23) by finding the most sparse solution whose residual sum of squares is smaller than a prespecified noise level ε. Inspired by this, we can also run SDAR by increasing T gradually until the residual sum of squares is smaller than a prespecified value ε.

We summarize these ideas in Algorithm 2 below.

Algorithm 2 Adaptive SDAR (ASDAR)

```
Require: Initial guess $\beta^0, d^0$, an integer $\tau$, an integer $L$, and an early stopping criterion
    (optional). Set $k = 1$.
for $k = 1, 2, \cdots$ do
    Run Algorithm 1 with $T = \tau k$ and with initial value $(\beta^{k-1}, d^{k-1})$. Denote the output
    by $(\beta^k, d^k)$.
    if the early stopping criterion is satisfied or $T > L$ then
        stop
    else
        $k = k + 1$.
    end if
end for
Ensure: $\hat{\beta}(\hat{T})$ as estimations of $\beta^*$.
```

5. Computational complexity

We look at the number of floating point operations line by line in Algorithm 1. Clearly it takes $O(p)$ flops to finish step 2-4. In step 5, we use conjugate gradient (CG) method (Golub and Van Loan, 2012) to solve the linear equation iteratively. During the CG iterations the main operation include two matrix-vector multiplications, which cost $2n|A_{k+1}|$ flops (the term $X'y$ on the right-hand side can be precomputed and stored). Therefore the number of CG iterations is smaller than $p/(2|A_{k+1}|)$, this ensures that the number of flops in step 5 is $O(np)$. In step 6, calculating the matrix-vector product costs np flops. In step 7, checking the stopping condition needs $O(p)$ flops. So the the overall cost per iteration of Algorithm 1 is $O(np)$. By Corollary 14 it needs no more than $O(\log(R))$ iterations to get a good solution for Algorithm 1 under the certain conditions. Therefore the overall cost of Algorithm 1 is $O(np\log(R))$ for exactly sparse and approximately sparse case under proper conditions.

Now we consider the cost of ASDAR (Algorithm 2). Assume ASDAR is stopped when $k = L$. Then the above discussion shows the the overall cost of Algorithm 2 is bounded by $O(Lnp\log(R))$ which is very efficient for large scale high dimension problem since the cost increases linearly in the ambient dimension p.

6. Comparison with greedy and screening methods

We give a high level comparison between SDAR and several greedy and screening methods, including OMP (Mallat and Zhang, 1993; Tropp, 2004; Donoho et al., 2006; Cai and Wang, 2011; Zhang, 2011b), FoBa (Zhang, 2011a), IHT (Blumensath and Davies, 2009; Jain et al., 2014) or GraDes (Garg and Khandekar, 2009), and SIS (Fan and Lv, 2008). These greedy methods iteratively select/remove one or more variables and project the response vector onto the linear subspace spanned by the variables that have already been selected. From this point of view, they and SDAR share a similar characteristic. However, OMP and FoBa, select one variable per iteration based on the current correlation, i.e., the dual variable d^k in our notation, while SDAR selects T variables at a time based on the sum of primal (β^k) and dual (d^k) information. The following interpretation in a low-dimensional setting with a small noise term may clarify the differences between these two approaches. If $X'X/n \approx \mathbb{I}$ and $\eta \approx 0$, we have

$$d^k = X'(y - X\beta^k)/n = X'(X\beta^* + \eta - X\beta^k)/n \approx \beta^* - \beta^k + X'\eta/n \approx \beta^* - \beta^k,$$

and

$$\beta^k + d^k \approx \beta^*.$$

Hence, SDAR can approximate the underlying support A^* more accurately than OMP and Foba. This is supported by the simulation results given in Section 7.

IHT (Blumensath and Davies, 2009; Jain et al., 2014) or GraDes (Garg and Khandekar, 2009), can be formulated as

$$\beta^{k+1} = H_K(\beta^k + s_k d^k), \tag{38}$$

where $H_K(\cdot)$ is the hard thresholding operator by keeping the first K largest elements and setting others to 0. The step size s_k is chosen as $s_k = 1$ and $s_k = 1/(1 + \delta_{2K})$ (where δ_{2K} is the RIP constant) for IHT and GraDes, respectively. IHT and GraDes use both primal and dual information to detect the support of the solution, which is similar to SDAR. But when the approximate active set is given, SDAR uses least squares fitting, which is more accurate than just keeping the largest elements by hard thresholding. This is supported by the simulation results given in Section 7. Jain et al. (2014) proposed an iterative hard thresholding algorithm for general high-dimensional sparse regressions. In the linear regression setting, the algorithm proposed in Jain et al. (2014) is the same as GraDes. Jain et al. (2014) also considered a two-stage IHT, which involves a refit step on the detected support. Yuan et al. (2018) extended gradient hard thresholding for least squares loss to a general class of convex losses and analyzed the estimation and sparsity recovery performance of their proposed method. Under restricted strongly convexity (RSS) and restricted strongly smoothness conditions (RSC), Jain et al. (2014) derived an error estimate between the approximate solutions and the oracle solution in ℓ_2 norm, which has the same order as our result in Section 3.1. There are some differences between SDAR and the two-stage IHT proposed in Jain et al. (2014). First, SDAR solves an $n \times K$ least squares problem at each iteration while the two-stage IHT involves two least-squares problems with larger sizes. The regularity conditions on X for SDAR concerns $2K \times 2K$ submatrices of X, while the regularity conditions for the two-stage IHT involves larger submatries of X. Second, our results are applicable to approximately sparse models. Jain et al. (2014) only considered exact sparse case. Third, we showed in (iii) of Corollary 3.1 that the iteration complexity of SDAR is $\mathcal{O}(\log K)$. In comparison, the iteration complexity of the two-stage IHT is $\mathcal{O}(K)$. We also established an ℓ_∞ norm estimation result and showed that the number of iterations of SDAR is independent of the sparsity level, see (iii) of Corollary 3.2. Last, we showed that the stopping criterion for SDAR can be archived in finitely many steps (Corollary 3.1 (iii) and Corollary 3.2. (iii)). However, Jain et al. (2014) did not discuss this issue.

Fan and Lv (2008) proposed SIS for dimension reduction in ultrahigh dimensional liner regression problems. This method selects variables with the T largest absolute values of $X'y$. To improve the performance of SIS, Fan and Lv (2008) also considered an iterative SIS, which iteratively selects more than one feature at a time until a desired number of variables are selected. They reported that the iterative SIS outperforms SIS numerically. However, the iterative SIS lacks a theoretical analysis. Interestingly, the first step in SDAR initialized with $\mathbf{0}$ is exactly the same as the SIS. But again the process of SDAR is different from the iterative SIS in that the active set of SDAR is determined based on the sum of primal and dual approximations while the iterative SIS uses dual only.

7. Simulation Studies

7.1 Implementation

We implemented SDAR/ASDAR, FoBa, GraDes and MCP in MatLab. For FoBa, our MatLab implementation follows the R package developed by Zhang (2011a). We optimize it by keeping track of rank-one updates after each greedy step. Our implementation of MCP uses the iterative threshholding algorithm (She, 2009) with warm starts. Publicly available Matlab packages for LARS (included in the SparseLab package) are used. Since LARS and FoBa add one variable at a time, we stop them when K variables are selected in addition to their default stopping conditions. Of course, doing so will reduce the computation time for these algorithms as well as improve accuracy by preventing overfitting.

In GraDes, the optimal gradient step length s_k depends on the RIP constant of X, which is NP hard to compute (Tillmann and Pfetsch, 2014). Here, we set $s_k = 1/3$ following Garg and Khandekar (2009). We stop GraDes when the residual norm is smaller than $\varepsilon = \sqrt{n}\sigma$, or the maximum number of iterations is greater than $n/2$. We compute the MCP solution path and select an optimal solution using the HBIC (Wang et al., 2013). We stop the iteration when the residual norm is smaller than $\varepsilon = \|\eta\|_2$, or the estimated support size is greater than $L = n/\log(n)$. In ASDAR (Algorithm 2), we set $\tau = 50$ and we stop the iteration if the residual $\|y - X\beta^k\|$ is smaller than $\varepsilon = \sqrt{n}\sigma$ or $k \geq L = n/\log(n)$.

7.2 Accuracy and efficiency

We compare the accuracy and efficiency of SDAR/ASDAR with Lasso (LARS), MCP, GraDes and FoBa.

We consider a moderately large scale setting with $n = 5000$ and $p = 50000$. The number of nonzero coefficients is set to be $K = 400$. So the sample size n is about $O(K\log(p-K))$. The dimension of the model is nearly at the limit where β^* can be reasonably well estimated by the Lasso (Wainwright, 2009).

To generate the design matrix X, we first generate an $n \times p$ random Gaussian matrix $\bar{X}$ whose entries are i.i.d. $\mathcal{N}(0,1)$ and then normalize its columns to the $\sqrt{n}$ length. Then X is generated with $X_1 = \bar{X}_1$, $X_j = \bar{X}_j + \rho(\bar{X}_{j+1} + \bar{X}_{j-1}), j = 2, \ldots, p-1$ and $X_p = \bar{X}_p$. The underlying regression coefficient β^* is generated with the nonzero coefficients uniformly distributed in $[m, M]$, where $m = \sigma\sqrt{2\log(p)/n}$ and $M = 100m$. Then the observation vector $y = X\beta^* + \eta$ with $\eta_1, \ldots, \eta_n$ generated independently from $\mathcal{N}(0, \sigma^2)$. We set $R = 100, \sigma = 1$ and $\rho = 0.2, 0.4$ and 0.6.

Table 1 shows the results based on 100 independent replications. The first column gives the correlation value ρ and the second column shows the methods in the comparison. The third and the fourth columns give the averaged relative error, defined as ReErr $= \sum \|\hat{\beta} - \beta^*\|/\|\beta^*\|$, and the averaged CPU time (in seconds), The standard deviations of the CPU times and the relative errors are shown in the parentheses. In each column of Table 1, the numbers in boldface indicate the best performers.

Table 1: Numerical results (relative errors, CPU times) on data sets with $n = 5000, p = 50000, K = 400, R = 100, \sigma = 1, \rho = 0.2 : 0.2 : 0.6$.

ρ	Method	ReErr	time(s)
0.2	LARS	1.1e-1 (2.5e-2)	4.8e+1 (9.8e-1)
	MCP	**7.5e-4** (**3.6e-5**)	9.3e+2 (2.4e+3)
	GraDes	1.1e-3 (7.0e-5)	2.3e+1 (9.0e-1)
	FoBa	**7.5e-4** (7.0e-5)	4.9e+1 (3.9e-1)
	ASDAR	**7.5e-4** (4.0e-5)	8.4e+0 (4.5e-1)
	SDAR	**7.5e-4** (4.0e-5)	**1.4e+0** (**5.1e-2**)
0.4	LARS	1.8e-1 (1.2e-2)	4.8e+1 (1.8e-1)
	MCP	6.2e-4 (3.6e-5)	2.2e+2 (1.6e+1)
	GraDes	8.8e-4 (5.7e-5)	8.7e+2 (2.6e+3)
	FoBa	1.0e-2 (1.4e-2)	5.0e+1 (4.2e-1)
	ASDAR	**6.0e-4** (**2.6e-5**)	8.8e+0 (**3.2e-1**)
	SDAR	**6.0e-4** (**2.6e-5**)	**2.3e+0** (1.7e+0)
0.6	LARS	3.0e-1 (2.5e-2)	4.8e+1 (3.5e-1)
	MCP	4.5e-4 (**2.5e-5**)	4.6e+2 (5.1e+2)
	GraDes	7.8e-4 (1.1e-4)	1.5e+2 (2.3e+2)
	FoBa	8.3e-3 (1.3e-2)	5.1e+1 (1.1e+0)
	ASDAR	**4.3e-4** (3.0e-5)	1.1e+1 (5.1e-1)
	SDAR	**4.3e-4** (3.0e-5)	**2.1e+0** (**8.6e-2**)

We see that when the correlation ρ is low, i.e., $\rho = 0.2$, MCP, FoBa, SDAR and ASDAR are on the top of the list in average error (ReErr). In terms of speed, SDAR/ASDAR is about 3 to 100 times faster than the other methods. As the correlation ρ increases to $\rho = 0.4$ and $\rho = 0.6$, FoBa becomes less accurate than SDAR/ASDAR. MCP is similar to SDAR/ASDAR in terms of accuracy, but it is 20 to 100 times slower than SDAR/ASDAR. The standard deviations of the CPU times and the relative errors of MCP and SDAR/ASDAR are similar and smaller than those of the other methods in all the three settings.

7.3 Influence of the model parameters

We now consider the effects of each of the model parameters on the performance of ASDAR, LARS, MCP, GraDes and FoBa more closely.

In this set of simulations, the rows of the design matrix X are drawn independently from $\mathcal{N}(0,\Sigma)$ with $\Sigma_{jk} = \rho^{|j-k|}, 1 \leq j,k \leq p$. The elements of the error vector η are generated independently with $\eta_i \sim \mathcal{N}(0,\sigma^2)$, $i = 1,\ldots,n$. Let $R = M/m$, where, $M = \max\{|\beta^*_{A^*}|\}, m = \min\{|\beta^*_{A^*}|\} = 1$. The underlying regression coefficient vector $\beta^* \in \mathbb{R}^p$ is generated in such a way that A^* is a randomly chosen subset of $\{1,2,...,p\}$ with $|A^*| = K < n$ and $R \in [1, 10^3]$. Then the observation vector $y = X\beta^* + \eta$. We use $\{n,p,K,\sigma,\rho,R\}$ to indicate the parameters used in the data generating model described above. We run ASDAR with $\tau = 5, L = n/\log(n)$ (if not specified). We use the HBIC (Wang et al., 2013) to select the tuning parameter T. The simulation results given in Figure 3 are based on 100 independent replications.

7.3.1 Influence of the sparsity level K

The top left panel of Figure 3 shows the results of the influence of sparsity level K on the probability of exact recovery of A^* of ASDAR, LARS, MCP, GraDes and FoBa. Data are generated from the model with $(n = 500, p = 1000, K = 10:50:360, \sigma = 0.5, \rho = 0.1, R = 10^3)$. Here $K = 10:50:360$ means the sample size starts from 10 to 360 with an increment of 50. We use $L = 0.8n$ for both ASDAR and MCP to eliminate the effect of stopping rule since the maximum $K = 360$. When the sparsity level $K = 10$, all the solvers performed well in recovering the true support. As K increases, LARS was the first one that failed to recover the support and vanished when $K = 60$ (this phenomenon had also been observed in Garg and Khandekar (2009), MCP began to fail when $K > 110$, GraDes and FoBa began to fail when $K > 160$. In comparison, ASDAR was still able to do well even when $K = 260$.

7.3.2 Influence of the sample size n

The top right panel of Figure 3 shows the influence of the sample size n on the probability of correctly estimating A^*. Data are generated from the model with $(n = 30:20:200, p = 500, K = 10, \sigma = 0.1, \rho = 0.1, R = 10)$. We see that the performance of all the five methods becomes better as n increases. However, ASDAR performs better than the others when $n = 30$ and 50. These simulation results indicate that ASDAR is more capable of handling high-dimensional data when p/n is large in the generating models considered here

7.3.3 Influence of the ambient dimension p

The bottom left panel of Figure 3 shows the influence of ambient dimension p on the performance of ASDAR, LARS, MCP, GraDes and FoBa. Data are generated from the model with $(n = 100, p = 200:200:1000, K = 20, \sigma = 1, \rho = 0.3, R = 10)$. We see that the probabilities of exactly recovering the support of the underlying coefficients of ASDAR and MCP are higher than those of the other solvers as p increasing, which indicate that ASDAR and MCP are more robust to the ambient dimension.

7.3.4 Influence of correlation ρ

The bottom right panel of Figure 3 shows the influence of correlation ρ on the performance of ASDAR, LARS, MCP, GraDes and FoBa. Data are generated from the model with $(n = 150, p = 500, K = 25, \sigma = 0.1, \rho = 0.05:0.1:0.95, R = 10^2)$. The performance of all the solvers becomes worse when the correlation ρ increases. However, ASDAR generally performed better than the other methods as ρ increases.

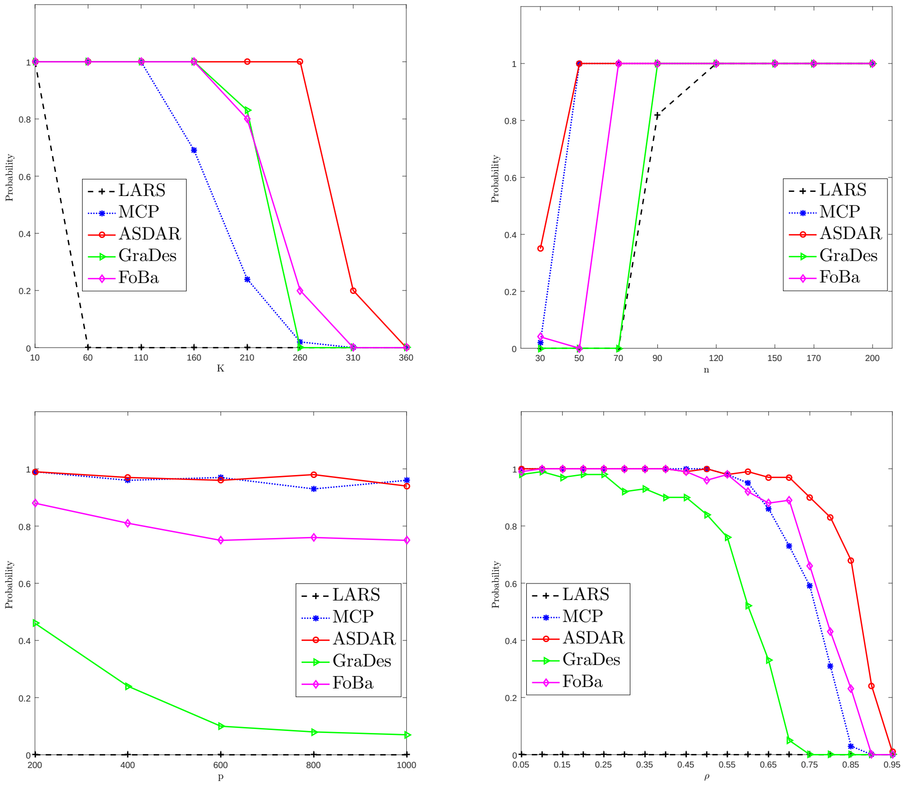

Figure 3: Numerical results of the influence of sparsity level K (top left panel), sample size n (top right panel), ambient dimension p (bottom left panel) and correlation ρ (bottom right panel) on the probability of exact recovery of the true support of all the solvers considered here.

In summary, our simulation studies demonstrate that SDAR/ASDAR is generally more accurate, more efficient and more stable than Lasso, MCP, FoBa and GraDes.

7.4 Number of iterations

In this subsection we compare SDAR with GraDes (IHT) in terms of the number of iterations. We run 100 independent replications on data sets generated from the models with $(n = 500, p = 1000, K = 5 : 5 : 55, \sigma = 0.05, \rho = 0, R = 1)$ and $(n = 2000, p = 5000, K = 10 : 20 : 250, \sigma = 0.05, \rho = 0, R = 1)$ described in Section 7.3. The average number of iteration (left column) and average absolute error in ℓ_∞ norm (right column) are displayed in Figure 4. We can see that the number of iterations of GraDes increases almost sublinearly as the sparsity level K increases while that of SDAR almost varies little. And in terms of the average error, SDAR is serval times more accurate than GraDes. This provides empirical support for our theoretical results in Corollary 30.

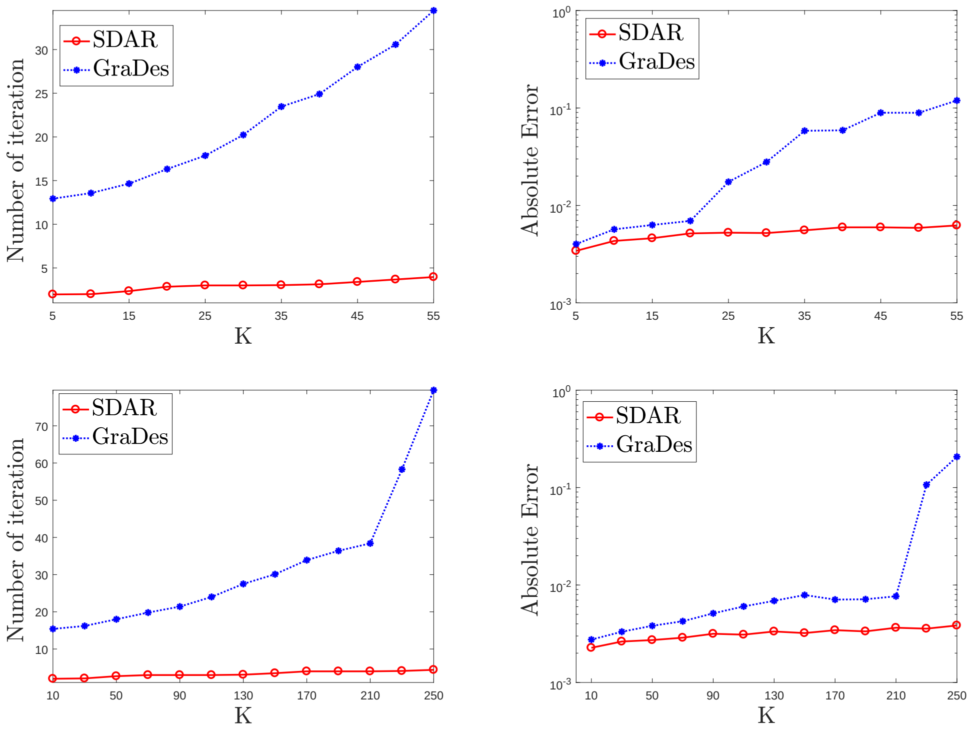

Figure 4: Comparisons the dependence of number of iterations (left panels) and accuracy (right panels) on sparsity level K with data set $(n = 500, p = 1000, K = 5 : 5 : 55, \sigma = 0.05, \rho = 0.3, R = 1)$ and $(n = 2000, p = 5000, K = 10 : 20 : 250, \sigma = 0.05, \rho = 0.3, R = 1)$.

8. Concluding remarks

SDAR is a constructive approach for fitting sparse, high-dimensional linear regression models. Under appropriate conditions, we established the nonasymptotic minimax ℓ_2 error bound and optimal ℓ_∞ error bound of the solution sequence generated by SDAR. We also calculated the number of iterations required to achieve these bounds. In particular, an interesting and surprising aspect of our results is that, under a mutual coherence condition on the design matrix, the number of iterations required for the SDAR to achieve the optimal ℓ_∞ bound does not depend on the underlying sparsity level. In addition, SDAR has the same computational complexity per iteration as LARS, coordinate descent and greedy methods. Our simulation studies demonstrate that SDAR/ASDAR is accurate, fast, stable and easy to implement, and it is competitive with or outperforms the Lasso, MCP and two greedy methods in efficiency and accuracy in the generating models we considered. These theoretical and numerical results suggest that SDAR/ASDAR is a useful addition to the literature on sparse modeling.

We have only considered the linear regression model. It would be interesting to generalize SDAR to models with more general loss functions or with other types of sparsity structures.

It would also be interesting to develop parallel or distributed versions of SDAR that can run on multiple cores for data sets with big n and large p or for data that are distributively stored.

We have implemented SDAR in a Matlab package *sdar*, which is available at `http://homepage.stat.uiowa.edu/~jian/`.

Acknowledgments

We are grateful to the action editor and the anonymous reviewers for their detailed and constructive comments which led to considerable improvements in the paper. We also thank Patrick Breheny for his critical reading of the paper and providing helpful comments. This research is supported in part by the National Science Foundation of China (NSFC) grants 11501579, 11571263, 11471253, 91630313, 11871385, 11871474 and 11801531.

Appendix A

Proof of Lemma 1.

Proof Let $L_\lambda(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_0$. Suppose $\beta^\diamond$ is a coordinate-wise minimizer of L_λ. Then

$$\begin{aligned} & \beta_i^\diamond \in \operatorname*{argmin}_{t\in\mathbb{R}} L_\lambda(\beta_1^\diamond, ..., \beta_{i-1}^\diamond, t, \beta_{i+1}^\diamond, ..., \beta_p^\diamond) \\ \Rightarrow \quad & \beta_i^\diamond \in \operatorname*{argmin}_{t\in\mathbb{R}} \tfrac{1}{2n}\|X\beta^\diamond - y + (t-\beta_i^\diamond)X_i\|_2^2 + \lambda|t|_0 \\ \Rightarrow \quad & \beta_i^\diamond \in \operatorname*{argmin}_{t\in\mathbb{R}} \tfrac{1}{2}(t-\beta_i^\diamond)^2 + \tfrac{1}{n}(t-\beta_i^\diamond)X_i'(X\beta^\diamond - y) + \lambda|t|_0 \\ \Rightarrow \quad & \beta_i^\diamond \in \operatorname*{argmin}_{t\in\mathbb{R}} \tfrac{1}{2}(t - (\beta_i^\diamond + X_i'(y - X\beta^\diamond)/n))^2 + \lambda|t|_0. \end{aligned}$$

Let $d^\diamond = X'(y - X\beta^\diamond)/n$. By the definition of the hard thresholding operator $H_\lambda(\cdot)$ in (5), we have

$$\beta_i^\diamond = H_\lambda(\beta_i^\diamond + d_i^\diamond) \qquad \text{for } i = 1, ..., p,$$

which shows (4) holds.

Conversely, suppose (4) holds. Let

$$A^\diamond = \Big\{ i \in S \Big| \, |\beta_i^\diamond + d_i^\diamond| \geq \sqrt{2\lambda} \Big\}.$$

By (4) and the definition of $H_\lambda(\cdot)$ in (5), we deduce that for $i \in A^\diamond$, $|\beta_i^\diamond| \geq \sqrt{2\lambda}$. Furthermore, $\mathbf{0} = d_{A^\diamond}^\diamond = X_{A^\diamond}'(y - X_{A^\diamond}\beta_{A^\diamond}^\diamond)/n$, which is equivalent to

$$\beta_{A^\diamond}^\diamond \in \operatorname{argmin} \tfrac{1}{2n}\|X_{A^\diamond}\beta_{A^\diamond} - y\|_2^2. \tag{39}$$

Next we show $L_\lambda(\beta^\diamond + h) \geq L_\lambda(\beta^\diamond)$ if h is small enough with $\|h\|_\infty < \sqrt{2\lambda}$. We consider two cases. If $h_{(A^\diamond)^c} \neq 0$, then

$$L_\lambda(\beta^\diamond + h) - L_\lambda(\beta^\diamond) \geq \tfrac{1}{2n}\|X\beta^\diamond - y + Xh\|_2^2 - \tfrac{1}{2n}\|X\beta^\diamond - y\|_2^2 + \lambda \geq \lambda - |\langle h, d^\diamond\rangle|,$$

which is positive for sufficiently small h. If $h_{(A^\diamond)^c} = 0$, by the minimizing property of $\beta^\diamond_{A^\diamond}$ in (39) we deduce that $L_\lambda(\beta^\diamond + h) \geq L_\lambda(\beta^\diamond)$. This completes the proof of Lemma 1. ■

Lemma 20 *Let A and B be disjoint subsets of S, with $|A| = a$ and $|B| = b$. Assume $X \sim SRC(a+b, c_-(a+b), c_+(a+b))$. Let $\theta_{a,b}$ be the sparse orthogonality constant and let μ be the mutual coherence of X. Then we have*

$$nc_-(a) \leq \left\|X_A^T X_A\right\| \leq nc_+(a), \tag{40}$$

$$\frac{1}{nc_+(a)} \leq \left\|(X_A^T X_A)^{-1}\right\| \leq \frac{1}{nc_-(a)}, \tag{41}$$

$$\left\|X_A'\right\| \leq \sqrt{nc_+(a)} \tag{42}$$

$$\theta_{a,b} \leq (c_+(a+b) - 1) \vee (1 - c_-(a+b)) \tag{43}$$

$$\|X_B' X_A u\|_\infty \leq na\mu\|u\|_\infty, \quad \forall u \in \mathbb{R}^{|A|}, \tag{44}$$

$$\|X_A\| = \left\|X_A'\right\| \leq \sqrt{n(1 + (a-1)\mu)}. \tag{45}$$

Furthermore, if $\mu < 1/(a-1)$, then

$$\|(X_A' X_A)^{-1} u\|_\infty \leq \frac{\|u\|_\infty}{n(1-(a-1)\mu)}, \quad \forall u \in \mathbb{R}^{|A|}. \tag{46}$$

Moreover, $c_+(s)$ is an increasing function of s, $c_-(s)$ a decreasing function of s and $\theta_{a,b}$ an increasing function of a and b.

Proof The assumption $X \sim \mathrm{SRC}(a, c_-(a), c_+(a))$ implies the spectrum of $X_A' X_A/n$ is contained in $[c_-(a), c_+(a)]$. So (40) - (42) hold. Let $\mathbb{I}$ be an $(a+b) \times (a+b)$ identity matrix. (43) follows from the fact that $X_A' X_B/n$ is a submatrix of $X_{A\cup B}' X_{A\cup B}/n - \mathbb{I}$ whose spectrum norm is less than $(1 - c_-(a+b)) \vee (c_+(a+b) - 1)$. Let $G = X'X/n$. Then, $|\sum_{j=1}^a G_{i,j} u_j| \leq \mu a \|u\|_\infty$, for all $i \in B$, which implies (44). By Gerschgorin's disk theorem,

$$|\, \|G_{A,A}\| - G_{i,i}| \leq \sum_{i \neq j = 1}^{a} |G_{i,j}| \leq (a-1)\mu \quad \forall i \in A,$$

thus (45) holds. For (46), it suffices to show $\|G_{A,A} u\|_\infty \geq (1-(a-1)\mu)\|u\|_\infty$ if $\mu < 1/(a-1)$. In fact, let $i \in A$ such that $\|u\|_\infty = |u_i|$, then

$$\|G_{A,A} u\|_\infty \geq |\sum_{j=1}^{a} G_{i,j} u_j| \geq |u_i| - \sum_{i \neq j = 1}^{a} |G_{i,j}|\, |u_j| \geq \|u\|_\infty - \mu(a-1)\|u\|_\infty.$$

The last assertion follows from their definitions. This completes the proof of Lemma 20. ■

Lemma 21 *Suppose (A3) holds. We have for any $\alpha \in (0, 1/2)$,*

$$\mathbf{P}\Big(\|X'\eta\|_\infty \leq \sigma\sqrt{2\log(p/\alpha)n}\Big) \geq 1 - 2\alpha, \tag{47}$$

$$\mathbf{P}\Big(\max_{|A| \leq T} \|X_A'\eta\|_2 \leq \sigma\sqrt{T}\sqrt{2\log(p/\alpha)n}\Big) \geq 1 - 2\alpha. \tag{48}$$

Proof This lemma follows from the sub-Gaussian assumption (A3) and standard probability calculation, see Zhang and Huang (2008); Wainwright (2009) for details. ■

We now define some notation that will be useful in proving Theorems 4 and 12. For any given integers T and J with $T \geq J$ and $F \subseteq S$ with $|F| = T - J$, let $A^\circ = A_J^* \cup F$ and $I^\circ = (A^\circ)^c$. Let $\{A^k\}_k$ be the sequence of active sets generated by SDAR (Algorithm 1).

Define

$$D_2(A^k) = \|\bar{\beta}^*|_{A_J^* \backslash A^k}\|_2 \quad \text{and} \quad D_\infty(A^k) = \|\bar{\beta}^*|_{A^* \backslash A^k}\|_\infty.$$

These quantities measure the differences between A_k and A_J^* in terms of the ℓ_2 and ℓ_∞ norms of the coefficients in A_J^* but not in A^k. A crucial step in our proofs is to control the sizes of these measures.

Let

$$A_1^k = A^k \cap A^\circ, A_2^k = A^\circ \backslash A_1^k, I_3^k = A^k \cap I^\circ, I_4^k = I^\circ \backslash I_3^k.$$

Denote the cardinality of I_3^k by $l_k = |I_3^k|$. Let

$$A_{11}^k = A_1^k \backslash (A^{k+1} \cap A_1^k), A_{22}^k = A_2^k \backslash (A^{k+1} \cap A_2^k), I_{33}^k = A^{k+1} \cap I_3^k, I_{44}^k = A^{k+1} \cap I_4^k,$$

and

$$\triangle^k = \beta^{k+1} - \bar{\beta}^*|_{A^k}.$$

These notation can be easily understood in the case $T = J$. For example, $D_2(A^k)$ and $D_\infty(A^k)$ are measures of the difference between the active set A^k and the target support A_J^*. A_1^k and I_3^k contain the correct indices and incorrect indices in A^k, respectively. A_{11}^k and A_{22}^k include the indices in A° that will be lost from the kth iteration to the $(k+1)$th iteration. I_{33}^k and I_{44}^k contain the indices included in I° that will be gained from the kth iteration to the $(k+1)$th iteration. By Algorithm 1, we have $|A^k| = |A^{k+1}| = T$, $A^k = A_1^k \cup I_3^k$, $|A_2^k| = |A^\circ| - |A_1^k| = |A^\circ| - |I_3^k| = T - (T - l_k) = l_k \leq T$, and

$$|A_{11}^k| + |A_{22}^k| = |I_{33}^k| + |I_{44}^k|, \tag{49}$$

$$D_2(A^k) = \|\bar{\beta}^*|_{A^\circ \backslash A^k}\|_2 = \|\bar{\beta}^*|_{A_2^k}\|_2, \tag{50}$$

$$D_\infty(A^k) = \|\bar{\beta}^*|_{A^\circ \backslash A^k}\|_\infty = \|\bar{\beta}^*|_{A_2^k}\|_\infty. \tag{51}$$

In Subsection 3.3, we described the overall approach for proving Theorems 4 and 12. Before proceeding to the proofs, we break down the argument into the following steps.

1. In Lemma 22 we show that the effects of the noise and the approximation model (10) measured by $h_2(T)$ and $h_\infty(T)$ can be controlled by the sum of unrecoverable approximation error R_J and the universal noise level $O(\sigma\sqrt{2\log(p)/n})$ with high probability, provided that η is sub-Gaussian.

2. In Lemma 23 we show that the ℓ_2 norms (ℓ_∞ norms) of $\triangle^k$ and $\beta^k - \bar{\beta}^*$ are controlled in terms of $D_2(A^k)$ and $h_2(T)$ ($D_\infty(A^k)$ and $h_\infty(T)$).

3. In Lemma 24 we show that $D_2(A^{k+1})$ ($D_\infty(A^{k+1})$) can be bounded by the norm of $\bar{\beta}^*$ on the lost indices, which in turn can be controlled in terms of $D_2(A^k)$ and $h_2(T)$ ($D_\infty(A^k)$ and $h_\infty(T)$) and the norms of $\triangle^k$, β^{k+1} and d^{k+1} on the lost indices.

4. In Lemma 25 we make use of the orthogonality between β^k and d^k to show that the norms of β^{k+1} and d^{k+1} on the lost indices can be bounded by the norm on the gained indices. Lemma 26 gives the upper bound of the norms of β^{k+1} and d^{k+1} on the gained indices by the sum of $D_2(A^k)$, $h_2(T)$ ($D_\infty(A^k)$, $h_\infty(T)$), and the norm of $\triangle^k$.

5. We combine Lemmas 22-26 and get the desired relations between $D_2(A^{k+1})$ and $D_2(A^k)$ ($D_\infty(A^{k+1})$ and $D_\infty(A^k)$) in Lemma 27.

Then we prove Theorems 4 and 12 based on Lemma 27, (56) and (58).

Lemma 22 *Let $A \subset S$ with $|A| \leq T$. Suppose (A1) and (A3) holds. Then for $\alpha \in (0, 1/2)$ with probability at least $1 - 2\alpha$, we have*

(i) If $X \sim SRC(T, c_-(T), c_+(T))$, then

$$h_2(T) \leq \varepsilon_1, \tag{52}$$

where ε_1 is defined in (18).

(ii) We have

$$h_\infty(T) \leq \varepsilon_2, \tag{53}$$

where ε_2 is defined in (29).

Proof We first show

$$\|X\beta^*|_{(A_J^*)^c}\|_2 \leq \sqrt{nc_+(J)}R_J, \tag{54}$$

under the assumption of $X \sim SRC(c_-(T), c_+(T), T)$ and (A1). In fact, let β be an arbitrary vector in $\mathbb{R}^p$ and A_1 be the first J largest positions of β, A_2 be the next and so forth. Then

$$\begin{aligned}
\|X\beta\|_2 &\leq \|X\beta_{A_1}\|_2 + \sum_{i\geq 2}\|X\beta_{A_i}\|_2 \\
&\leq \sqrt{nc_+(J)}\|\beta_{A_1}\|_2 + \sqrt{nc_+(J)}\sum_{i\geq 2}\|\beta_{A_i}\|_2 \\
&\leq \sqrt{nc_+(J)}\|\beta\|_2 + \sqrt{nc_+(J)}\sum_{i\geq 1}\sqrt{\frac{1}{J}}\|\beta_{A_{i-1}}\|_1 \\
&\leq \sqrt{nc_+(J)}(\|\beta\|_2 + \sqrt{\frac{1}{J}}\|\beta\|_1),
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from (42), and the third and fourth ones follows from simple algebra. This implies (54) holds by observing the definition of R_J. By the triangle inequality, (42), (54) and (48), we have with probability at least $1 - 2\alpha$,

$$\begin{aligned}
\|X_A'\bar{\eta}\|_2/n &\leq \|X_A'X\beta^*|_{(A_J^*)^c}\|_2/n + \|X_A'\eta\|_2/n \\
&\leq c_+(J)R_J + \sigma\sqrt{T}\sqrt{2\log(p/\alpha)/n}.
\end{aligned}$$

Therefore, (52) follows by noticing the monotone increasing property of $c_+(\cdot)$, the definition of ε_1 in (18) and the arbitrariness of A.

By a similar argument for (54) and replacing $\sqrt{nc_+(J)}$ with $\sqrt{n(1+(J-1)\mu)}$ based on (45), we get

$$\|X\beta^*|_{(A_J^*)^c}\|_2 \leq \sqrt{n(1+(K-1)\mu)}R_J. \tag{55}$$

Therefore, by (45), (55) and (47), we have with probability at least $1-2\alpha$,

$$\begin{aligned}\|X_A'\bar{\eta}\|_\infty/n &\leq \|X_A'X\beta^*|_{(A_J^*)^c}\|_\infty/n + \|X_A'\eta\|_2/n \\ &\leq \|X_A'X\beta^*|_{(A_J^*)^c}\|_2/n + \|X_A'\eta\|_2/n \\ &\leq (1+(J-1)\mu)R_J + \sigma\sqrt{2\log(p/\alpha)/n}.\end{aligned}$$

This implies part (ii) of Lemma 22 by noticing the definition of ε_2 in (29) and the arbitrariness of A. This completes the proof of Lemma 22. ■

Lemma 23 *Let $A \subset S$ with $|A| \leq T$. Suppose (A1) holds.*

(i) If $X \sim SRC(T, c_-(T), c_+(T))$,

$$\|\beta^{k+1}-\bar{\beta}^*\|_2 \leq \Big(1+\frac{\theta_{T,T}}{c_-(T)}\Big)D_2(A^k) + \frac{h_2(T)}{c_-(T)}, \tag{56}$$

and

$$\|\triangle^k\|_2 \leq \frac{\theta_{T,T}}{c_-(T)}\|\bar{\beta}^*|_{A_2^k}\|_2 + \frac{h_2(T)}{c_-(T)}. \tag{57}$$

(ii) If $(T-1)\mu < 1$, then

$$\|\beta^{k+1}-\bar{\beta}^*\|_\infty \leq \frac{1+\mu}{1-(T-1)\mu}D_\infty(A^k) + \frac{h_\infty(T)}{1-(T-1)\mu}, \tag{58}$$

and

$$\|\triangle^k\|_\infty \leq \frac{T\mu}{1-(T-1)\mu}\|\bar{\beta}^*|_{A_2^k}\|_\infty + \frac{h_\infty(T)}{(1-(T-1)\mu)}. \tag{59}$$

Proof We have

$$\begin{aligned}\beta_{A^k}^{k+1} &= (X_{A^k}'X_{A^k})^{-1}X_{A^k}'y \\ &= (X_{A^k}'X_{A^k})^{-1}X_{A^k}'(X_{A_1^k}\bar{\beta}_{A_1^k}^* + X_{A_2^k}\bar{\beta}_{A_2^k}^* + \bar{\eta}), \qquad (60)\\ (\bar{\beta}^*|_{A^k})_{A^k} &= (X_{A^k}'X_{A^k})^{-1}X_{A^k}'X_{A^k}(\bar{\beta}^*|_{A^k})_{A^k} \\ &= (X_{A^k}'X_{A^k})^{-1}X_{A^k}'(X_{A_1^k}\bar{\beta}_{A_1^k}^*), \qquad (61)\end{aligned}$$

where the first equality uses the definition of β^{k+1} in Algorithm 1, the second equality follows from $y = X\bar{\beta}^* + \bar{\eta} = X_{A_1^k}\bar{\beta}^*_{A_1^k} + X_{A_2^k}\bar{\beta}^*_{A_2^k} + \bar{\eta}$, the third equality is simple algebra, and the last one uses the definition of A_1^k. Therefore,

$$\begin{aligned}
\|\triangle^k\|_2 &= \|\beta^{k+1}_{A^k} - (\bar{\beta}^*|_{A^k})_{A^k}\|_2 \\
&= \|(X'_{A^k}X_{A^k})^{-1}X'_{A^k}(X_{A_2^k}\bar{\beta}^*_{A_2^k} + \bar{\eta})\|_2 \\
&\leq \frac{1}{nc_-(T)}(\|X'_{A^k}X_{A_2^k}\bar{\beta}^*_{A_2^k}\|_2 + \|X'_{A^k}\bar{\eta}\|_2) \\
&\leq \frac{\theta_{T,T}}{c_-(T)}\|\bar{\beta}^*|_{A_2^k}\|_2 + \frac{h_2(T)}{c_-(T)},
\end{aligned}$$

where the first equality uses $\text{supp}(\beta^{k+1}) = A^k$, the second equality follows from (61) and (60), the first inequality follows from (41) and the triangle inequality, and the second inequality follows from (50), the definition of $\theta_{a,b}$ and the definition of $h_2(T)$. This proves (57). Then the triangle inequality $\|\beta^{k+1} - \bar{\beta}^*\|_2 \leq \|\beta^{k+1} - \bar{\beta}^*|_{A^k}\|_2 + \|\bar{\beta}^*|_{A^\circ\setminus A^k}\|_2$ and (57) imply (56).

Using an argument similar to the proof of (57) and by (46), (44) and (51), we can show (59). Thus (58) follows from the triangle inequality and (59). This completes the proof of Lemma 23. ■

Lemma 24

$$D_2(A^{k+1}) \leq \|\bar{\beta}^*_{A_{11}^k}\|_2 + \|\bar{\beta}^*_{A_{22}^k}\|_2, \tag{62}$$

$$D_\infty(A^{k+1}) \leq \|\bar{\beta}^*_{A_{11}^k}\|_\infty + \|\bar{\beta}^*_{A_{22}^k}\|_\infty. \tag{63}$$

$$\|\bar{\beta}^*_{A_{11}^k}\|_2 \leq \|\triangle^k_{A_{11}^k}\|_2 + \|\beta^{k+1}_{A_{11}^k}\|_2, \tag{64}$$

$$\|\bar{\beta}^*_{A_{11}^k}\|_\infty \leq \|\triangle^k_{A_{11}^k}\|_\infty + \|\beta^{k+1}_{A_{11}^k}\|_\infty. \tag{65}$$

Furthermore, assume (A1) holds. We have

$$\|\bar{\beta}^*_{A_{22}^k}\|_\infty \leq \|d^{k+1}_{A_{22}^k}\|_\infty + T\mu\|\triangle^k_{A^k}\|_\infty + T\mu D_\infty(A^k) + h_\infty(T), \tag{66}$$

$$\|\bar{\beta}^*_{A_{22}^k}\|_2 \leq \frac{\|d^{k+1}_{A_{22}^k}\|_2 + \theta_{T,T}\|\triangle^k_{A^k}\|_2 + \theta_{T,T}D_2(A^k) + h_2(T)}{c_-(T)} \quad \textit{if } X \sim \textit{SRC}(T, c_-(T), c_+(T)). \tag{67}$$

Proof By the definitions of $D_2(A^{k+1})$, A_{11}^k, A_{11}^k and A_{22}^k, we have

$$D_2(A^{k+1}) = \|\bar{\beta}^*|_{A^\circ\setminus A^{k+1}}\|_2 = \|\bar{\beta}^*|_{A_{11}^k\cup A_{22}^k}\|_2 \leq \|\bar{\beta}^*_{A_{11}^k}\|_2 + \|\bar{\beta}^*_{A_{22}^k}\|_2.$$

This proves (62). (63) can be proved similarly. To show (64), we note that $\triangle^k = \beta^{k+1} - \bar{\beta}^*|_{A^k}$. Thus

$$\|\beta^{k+1}_{A_{11}^k}\|_2 = \|\left(\bar{\beta}^*|_{A^k}\right)_{A_{11}^k} + \triangle^k_{A_{11}^k}\|_2 \geq \|\bar{\beta}^*_{A_{11}^k}\|_2 - \|\triangle^k_{A_{11}^k}\|_2.$$

This proves (64). (65) can be proved similarly.

Now consider (67). We have

$$\begin{aligned}
\|d^{k+1}_{A^k_{22}}\|_2 &= \|X'_{A^k_{22}}\Big(X_{A^k}\beta^{k+1}_{A^k} - y\Big)/n\|_2 \\
&= \|X'_{A^k_{22}}\Big(X_{A^k}\triangle^k_{A^k} + X_{A^k}\bar{\beta}^*_{A^k} - X_{A^\circ}\bar{\beta}^*_{A^\circ} - \bar{\eta}\Big)/n\|_2 \\
&= \|X'_{A^k_{22}}\Big(X_{A^k}\triangle^k_{A^k} - X_{A^k_{22}}\bar{\beta}^*_{A^k_{22}} - X_{A^k_2\backslash A^k_{22}}\bar{\beta}^*_{A^k_2\backslash A^k_{22}} - \bar{\eta}\Big)/n\|_2 \\
&\geq c_-(|A^k_{22}|)\|\bar{\beta}^*_{A^k_{22}}\|_2 - \theta_{|A^k_{22}|,T}\|\triangle^k_{A^k}\|_2 - \theta_{l_k,l_k-|A^k_{22}|}\|\bar{\beta}^*_{A^k_2\backslash A^k_{22}}\|_2 - \|X_{A^k_{22}}\bar{\eta}/n\|_2 \\
&\geq c_-(T)\|\bar{\beta}^*_{A^k_{22}}\|_2 - \theta_{T,T}\|\triangle^k_{A^k}\|_2 - \theta_{T,T}D_2(A^k) - h_2(T),
\end{aligned}$$

where the first equality uses the definition of d^{k+1}, the second equality uses the the definition of $\triangle^k$ and y, the third equality is simple algebra, the first inequality uses the triangle inequality, (40) and the definition of $\theta_{a,b}$, and the last inequality follows from the monotonicity property of $c_-(\cdot)$, $\theta_{a,b}$ and the definition of $h_2(T)$. This proves (67).

Finally, we show (66). Let $i_k \in A^k_{22}$ be an index satisfying $\left|\bar{\beta}^*_{i_k}\right| = \|\bar{\beta}^*_{A^k_{22}}\|_\infty$. Then

$$\begin{aligned}
\left|d^{k+1}_{i_k}\right| &= \|X'_{i_k}(X_{A^k}\triangle^k_{A^k} - X_{i_k}\bar{\beta}^*_{i_k} - X_{A^k_2\backslash\ i_k}\bar{\beta}^*_{A^k_2\backslash\ i_k} - \bar{\eta})/n\|_\infty \\
&\geq \left|\bar{\beta}^*_{i_k}\right| - T\mu\|\triangle^k_{A^k}\|_\infty - l_k\mu\|\bar{\beta}^*_{A^k_2\backslash\ i_k}\|_\infty - \|X'_{i_k}\bar{\eta}\|_\infty \\
&\geq \|\bar{\beta}^*_{A^k_{22}}\|_\infty - T\mu\|\triangle^k_{A^k}\|_\infty - T\mu D_\infty(A^k) - h_\infty(T),
\end{aligned}$$

where the first equality is derived from the first three equalities in the proof of (67) by replacing A^k_{22} with i_k, the first inequality follows from the triangle inequality and (44), and the last inequality follows from the definition of $h_\infty(T)$. Then (66) follows by rearranging the terms in the above inequality. This completes the proof of Lemma 24. ■

Lemma 25

$$\|\beta^k\|_\infty \vee \|d^k\|_\infty = \max\{|\beta^k_i| + |d^k_i| \big| i \in S\}, \forall k \geq 1. \tag{68}$$

$$\|\beta^{k+1}_{A^k_{11}}\|_\infty + \|d^{k+1}_{A^k_{22}}\|_\infty \leq \left|\beta^{k+1}_{I^k_{33}}\right|_{min} \wedge \left|d^{k+1}_{I^k_{44}}\right|_{min}. \tag{69}$$

$$\|\beta^{k+1}_{A^k_{11}}\|_2 + \|d^{k+1}_{A^k_{22}}\|_2 \leq \sqrt{2}\Big(\|\beta^{k+1}_{I^k_{33}}\|_2 + \|d^{k+1}_{I^k_{44}}\|_2\Big). \tag{70}$$

Proof By the definition of Algorithm 1 we have $\beta^k_i d^k_i = 0$, $\forall i \in S$, $\forall k \geq 1$, thus (68) holds. (69) follows from the definition of A^k_{11}, A^k_{22}, I^k_{33}, I^k_{44} and (68). Now

$$\begin{aligned}
\frac{1}{2}(\|\beta^{k+1}_{A^k_{11}}\|_2 + \|d^{k+1}_{A^k_{22}}\|_2)^2 &\leq \|\beta^{k+1}_{A^k_{11}}\|_2^2 + \|d^{k+1}_{A^k_{22}}\|_2^2 \\
&\leq (\|\beta^{k+1}_{I^k_{33}}\|_2 + \|d^{k+1}_{I^k_{44}}\|_2)^2,
\end{aligned}$$

where the first inequality follows from simple algebra, and the second inequality follows from (49) and (69). Thus (70) follows. This completes the proof of Lemma 25. ■

Lemma 26

$$\|\beta^{k+1}_{I^k_{33}}\|_2 \leq \|\triangle^k_{I^k_{33}}\|_2. \tag{71}$$

Furthermore, suppose (A1) holds. We have

$$\|d^{k+1}_{I^k_{44}}\|_\infty \leq T\mu\|\triangle^k_{A^k}\|_\infty + T\mu D_\infty(A^k) + h_\infty(T) \quad \textit{under the mutual coherence condition } (A^*), \tag{72}$$

$$\|d^{k+1}_{I^k_{44}}\|_2 \leq \theta_{T,T}\left\|\triangle^k_{A^k}\right\| + \theta_{T,T}D_2(A^k) + h_2(T) \quad \textit{if} \quad X \sim \textit{SRC}(T, c_-(T), c_+(T)). \tag{73}$$

Proof By the definition of $\triangle^k$, the triangle inequality and the fact that $\bar{\beta}^*$ vanishes on $A^k \cap I^k_{33}$, we have

$$\|\beta^{k+1}_{I^k_{33}}\|_2 = \|\triangle^k_{I^k_{33}} + \bar{\beta}^*_{I^k_{33}}\|_2 \leq \|\triangle^k_{I^k_{33}}\|_2 + \|\bar{\beta}^*_{A^k \cap I^k_{33}}\|_2 = \|\triangle^k_{I^k_{33}}\|_2.$$

So (71) follows. Now

$$\begin{aligned}
\|d^{k+1}_{I^k_{44}}\|_2 &= \|X'_{I^k_{44}}\left(X_{A^k}\triangle^k_{A^k} - X_{A^k_2}\bar{\beta}^*_{A^k_2} - \bar{\eta}\right)/n\|_2 \\
&\leq \theta_{|I^k_{44}|,T}\|\triangle^k_{A^k}\|_2 + \theta_{|I^k_{44}|,l_k}\|\bar{\beta}^*_{A^k_2}\|_2 + \|X'_{I^k_{44}}\bar{\eta}\|_2 \\
&\leq \theta_{T,T}\|\triangle^k_{A^k}\|_2 + \theta_{T,T}D_2(A^k) + h_2(T),
\end{aligned}$$

where the first equality is derived from the first three equalities in the proof of (67) by replacing A^k_{22} with I^k_{44}, the first inequality follows from the triangle inequality and the definition of $\theta_{a,b}$, and the last inequality follows from the monotonicity property of $\theta_{a,b}$ and $h_2(T)$. This implies (73). Finally, (72) can be proved similarly by using (44) and (53). This completes the proof of Lemma 26. ∎

Lemma 27 *Suppose (A1) holds.*

(i) If $X \sim \mathit{SRC}(T, c_-(T), c_+(T))$, then

$$D_2(A^{k+1}) \leq \gamma D_2(A^k) + \frac{\gamma}{\theta_{T,T}}h_2(T), \tag{74}$$

(ii) If $(T-1)\mu < 1$, then

$$D_\infty(A^{k+1}) \leq \gamma_\mu D_2(A^k) + \frac{3+2\mu}{1-(T-1)\mu}h_\infty(T). \tag{75}$$

Proof We have

$$\begin{aligned}
D_2(A^{k+1}) &\leq \|\bar{\beta}^*_{A^k_{11}}\|_2 + \|\bar{\beta}^*_{A^k_{22}}\|_2 \\
&\leq (\|\beta^{k+1}_{A^k_{11}}\|_2 + \|d^{k+1}_{A^k_{22}}\|_2 + \|\triangle^k_{A^k_{11}}\|_2 + \theta_{T,T}\|\triangle^k_{A^k}\|_2 + \theta_{T,T}D_2(A^k) + h_2(T))/c_-(T) \\
&\leq (\sqrt{2}(\|\beta^{k+1}_{I^k_{33}}\|_2 + \|d^{k+1}_{I^k_{44}}\|_2) + \|\triangle^k_{A^k_{11}}\|_2 + \theta_{T,T}\left\|\triangle^k_{A^k}\right\| + \theta_{T,T}D_2(A^k) + h_2(T))/c_-(T) \\
&\leq ((2 + (1+\sqrt{2})\theta_{T,T})\|\triangle^k\|_2 + (1+\sqrt{2})\theta_{T,T}D_2(A^k) + (1+\sqrt{2})h_2(T))/c_-(T) \\
&\leq (\frac{2\theta_{T,T} + (1+\sqrt{2})\theta^2_{T,T}}{c_-(T)^2} + \frac{(1+\sqrt{2})\theta_{T,T}}{c_-(T)})D_2(A^k) \\
&+ (\frac{2 + (1+\sqrt{2})\theta_{T,T}}{c_-(T)^2} + \frac{1+\sqrt{2}}{c_-(T)})h_2(T),
\end{aligned}$$

where the first inequality is (62), the second inequality follows from (64) and (67), the third inequality follows from (70), the fourth inequality uses the sum of (71) and (73), and the last inequality follows from (57). This implies (74) by noticing the definitions of γ.

Now

$$\begin{aligned}
D_\infty(A^{k+1}) &\leq \|\bar{\beta}^*_{A^k_{11}}\|_\infty + \|\bar{\beta}^*_{A^k_{22}}\|_2 \\
&\leq \|\beta^{k+1}_{A^k_{11}}\|_\infty + \|d^{k+1}_{A^k_{22}}\|_\infty + \|\triangle^k_{A^k_{11}}\|_\infty + T\mu\|\triangle^k_{A^k}\|_\infty + T\mu D_\infty(A^k) + h_\infty(T). \\
&\leq \|d^{k+1}_{I^k_{44}}\|_\infty + \|\triangle^k_{A^k_{11}}\|_\infty + T\mu\|\triangle^k_{A^k}\|_\infty + T\mu D_\infty(A^k) + h_\infty(T) \\
&\leq \|\triangle^k_{A^k_{11}}\|_\infty + 2T\mu\|\triangle^k_{A^k}\|_\infty + 2T\mu D_\infty(A^k) + 2h_\infty(T) \\
&\leq (\frac{(1+2T\mu)T\mu}{1-(T-1)\mu} + 2T\mu)D_\infty(A^k) + \frac{3+2\mu}{1-(T-1)\mu}h_\infty(T),
\end{aligned}$$

where the first inequality is (63), the second inequality follows from (65) and (66), the third inequality follows from (69), the fourth inequality follows from (72), and the last inequality follows from (59). Thus part (ii) of Lemma 27 follows by noticing the definition of γ_μ. This completes the proof of Lemma 27. ■

Proof of Theorem 4.

Proof Suppose $\gamma < 1$. By using (74) repeatedly,

$$\begin{aligned}
D_2(A^{k+1}) &\leq \gamma D_2(A^k) + \frac{\gamma}{\theta_{T,T}}h_2(T) \\
&\leq \gamma(\gamma D_2(A^{k-1}) + \frac{\gamma}{\theta_{T,T}}h_2(T)) + \gamma h_2(T) \\
&\leq \cdots \\
&\leq \gamma^{k+1}D_2(A^0) + \frac{\gamma}{\theta_{T,T}}(1 + \gamma + \cdots + \gamma^k)h_2(T) \\
&< \gamma^{k+1}\|\bar{\beta}^*\|_2 + \frac{\gamma}{(1-\gamma)\theta_{T,T}}h_2(T),
\end{aligned}$$

i.e., (13) holds. Now

$$\begin{aligned}\|\beta^{k+1}-\bar{\beta}^*\|_2 &\leq (1+\frac{\theta_{T,T}}{c_-(T)})D_2(A^k)+\frac{h_2(T)}{c_-(T)}\\ &\leq (1+\frac{\theta_{T,T}}{c_-(T)})\Big[\gamma^k\|\bar{\beta}^*\|_2+\frac{\gamma\theta_{T,T}}{1-\gamma}h_2(T)\Big]\\ &= (1+\frac{\theta_{T,T}}{c_-(T)})\gamma^k\|\bar{\beta}^*\|_2+\Big[\frac{\gamma\theta_{T,T}}{(1-\gamma)}(1+\frac{\theta_{T,T}}{c_-(T)})+\frac{1}{c_-(T)}\Big]h_2(T),\end{aligned}$$

where the first inequality follows from (56), the second inequality follows from (13), and the third line follows after some algebra. Thus (14) follows by noticing the definitions of b_1 and b_2. This completes the proof of part (i) of Theorem 4.

For part (ii), (16) follows from (13) and (52), (17) follows from (14) and (52). This completes the proof of Theorem 4. ■

Proof of Corollary 8.
Proof By (14),

$$\begin{aligned}\|\beta^{k+1}-\bar{\beta}^*\|_2 &\leq b_1\gamma_1^k\|\bar{\beta}^*\|_2+b_2h_2(T)\\ &\leq b_1h_2(T)+b_2h_2(T) \quad \text{if} \quad k\geq \log_{\frac{1}{\gamma}}\frac{\sqrt{J}\bar{M}}{h_2(T)}\end{aligned}$$

where the second inequality follows after some algebra. By (13),

$$\begin{aligned}\|\bar{\beta}^*|_{A_J^*\backslash A^k}\|_2 &\leq \gamma^k\|\bar{\beta}^*\|_2+\frac{\gamma\theta_{T,T}}{1-\gamma}h_2(T)\\ &\leq \gamma^k\sqrt{J}\bar{M}+\xi\bar{m}\\ &< \bar{m} \quad \text{if} \quad k\geq \log_{\frac{1}{\gamma}}\frac{\sqrt{J}R}{1-\xi},\end{aligned}$$

where the second inequality follows from the assumption $\bar{m}\geq\frac{\gamma h_2(T)}{(1-\gamma)\theta_{T,T}\xi}$ with $0<\xi<1$, and the last inequality follows after some simple algebra. This implies $A_J^*\subset A^k$ if $k\geq\log_{\frac{1}{\gamma}}\frac{\sqrt{J}R}{1-\xi}$. This proves part (i). The proof of part (ii) is similar to that of part (i) by using (52), we omit it here. For part (iii), suppose β^* is exactly K-sparse and $T=K$ in the SDAR algorithm (Algorithm 1). It follows from part (ii) that with probability at least $1-2\alpha$, $A^*=A^k$ if $k\geq\log_{\frac{1}{\gamma}}\frac{\sqrt{K}R}{1-\xi}$. Then part (iii) holds by showing that $A^{k+1}=A^*$. Indeed, by (74) and (52) we have

$$\begin{aligned}\|\bar{\beta}^*|_{A^*\backslash A^{k+1}}\|_2 &\leq \gamma\|\bar{\beta}^*|_{A^*\backslash A^k}\|_2+\frac{\gamma}{\theta_{T,T}}\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}\\ &= \frac{\gamma}{\theta_{T,T}}\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}.\end{aligned}$$

Then $A^{k+1}=A^*$ follows from the assumption that $m\geq\frac{\gamma}{(1-\gamma)\theta_{T,T}\xi}\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}>\frac{\gamma}{\theta_{T,T}}\sigma\sqrt{K}\sqrt{2\log(p/\alpha)/n}$. This completes the proof of Corollary 8. ■

Proof of Theorem 12.
Proof For μ satisfying $T\mu \leq 1/4$, some algebra shows $\gamma_\mu < 1$ and $\frac{1+\mu}{1-(T-1)\mu} < \frac{3+2\mu}{1-(T-1)\mu} < 4$. Now Theorem 12 can be proved similarly to Theorem 4 by using (75), (53) and (58). We omit it here. This completes the proof of Theorem 12. ■

Proof of Corollary 14.
Proof The proofs of part (i) and part (ii) are similar to those of Corollary 8, we omit them here. Suppose β^* is exactly K-sparse and $T = K$ in SDAR. It follows from part (ii) that with probability at least $1-2\alpha$, $A^* = A^k$ if $k \geq \log_{\frac{1}{\gamma_\mu}} \frac{R}{1-\xi}$. Then part (iii) holds by showing that $A^{k+1} = A^*$. By (75), (53) and $\frac{3+2\mu}{1-(T-1)\mu} < 4$ we have

$$\begin{aligned}\|\bar{\beta}^*|_{A^*\backslash A^{k+1}}\|_\infty &\leq \gamma_\mu \|\bar{\beta}^*|_{A^*\backslash A^k}\|_\infty + 4\sigma\sqrt{2\log(p/\alpha)/n}\\ &= 4\sigma\sqrt{2\log(p/\alpha)/n}.\end{aligned}$$

Then $A^{k+1} = A^*$ by the assumption that

$$m \geq \frac{4}{\xi(1-\gamma_\mu)}\sigma\sqrt{2\log(p/\alpha)/n} > 4\sigma\sqrt{2\log(p/\alpha)/n}.$$

This completes the proof of Corollary 14. ■

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40 (5):2452–2482, 2012.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14(5-6):629–654, 2008.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

T Tony Cai and Lie Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, 57(7):4680–4688, 2011.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33, 1998.

Xiaojun Chen, Dongdong Ge, Zizhuo Wang, and Yinyu Ye. Complexity of unconstrained l2-lp minimization. *Mathematical Programming*, 143(1-2):371–383, 2014.

David L Donoho and Yaakov Tsaig. Fast solution of l1 norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54(11):4789–4812, 2008.

David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819, 2014.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. pages 337–344, 2009.

Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, 2009.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Inc Gurobi Optimization. Gurobi optimizer reference manual. *URL http://www. gurobi. com*, 2015.

Jian Huang, Joel L Horowitz, and Shuangge Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008a.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618, 2008b.

Jian Huang, Yuling Jiao, Bangti Jin, Jin Liu, Xiliang Lu, and Can Yang. A unified primal dual active set algorithm for nonconvex sparse recovery. *arXiv:1310.1147v4*, 2018.

David R Hunter and Runze Li. Variable selection using mm algorithms. *Annals of statistics*, 33(4):1617, 2005.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Yuling Jiao, Bangti Jin, and Xiliang Lu. A primal dual active set with continuation algorithm for the l0-regularized optimization problem. *Applied and Computational Harmonic Analysis*, 39(3):400–426, 2015.

Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008.

Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

Yufeng Liu and Yichao Wu. Variable selection via a combination of the l0 and l1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(1):559–616, 2015.

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102, 2008.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495): 1125–1138, 2011.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34:1436–1462, 2006.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over lq balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Yiyuan She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of statistics*, 3:384–415, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505, 2013.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18:1–43, 2018.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010a.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pages 576–593, 2012.

Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5):2109–2144, 2009.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010b.

Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7):4689–4708, 2011a.

Tong Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011b.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7:2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.