

PATHWISE COORDINATE OPTIMIZATION FOR SPARSE LEARNING: ALGORITHM AND THEORY¹

BY TUO ZHAO, HAN LIU AND TONG ZHANG

Georgia Institute of Technology, Princeton University and Tencent AI Lab

The pathwise coordinate optimization is one of the most important computational frameworks for high dimensional convex and nonconvex sparse learning problems. It differs from the classical coordinate optimization algorithms in three salient features: *warm start initialization*, *active set updating* and *strong rule for coordinate preselection*. Such a complex algorithmic structure grants superior empirical performance, but also poses significant challenge to theoretical analysis. To tackle this long lasting problem, we develop a new theory showing that these three features play pivotal roles in guaranteeing the outstanding statistical and computational performance of the pathwise coordinate optimization framework. Particularly, we analyze the existing pathwise coordinate optimization algorithms and provide new theoretical insights into them. The obtained insights further motivate the development of several modifications to improve the pathwise coordinate optimization framework, which guarantees linear convergence to a unique sparse local optimum with optimal statistical properties in parameter estimation and support recovery. This is the first result on the computational and statistical guarantees of the pathwise coordinate optimization framework in high dimensions. Thorough numerical experiments are provided to support our theory.

1. Introduction. Modern data acquisition routinely produces massive amount of high dimensional data, where the number of variables d greatly exceeds the sample size n , such as high-throughput genomic data [Neale et al. (2012)] and image data from functional magnetic resonance imaging [Eloyan et al. (2012), Liu, Wang and Zhao (2015)]. To handle high dimensionality, we often assume that only a small subset of variables are relevant in modeling [Tibshirani (1996)]. Such a parsimonious assumption motivates various sparse learning approaches. Taking sparse linear regression as an example, we consider a linear model $y = X\theta^* + \varepsilon$, where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times d}$ is the design matrix, $\theta^* = (\theta_1, \dots, \theta_d)^\top \in \mathbb{R}^d$ is the unknown sparse regression coefficient vector, and $\varepsilon \sim N(0, \sigma^2 I_n)$ is the random noise. Here, $I_n \in \mathbb{R}^{n \times n}$ is an identity matrix. Let

Received August 2016; revised January 2017.

¹Supported by NSF IIS 1250985, NSF IIS 1407939, NSF DMS1454377 CAREER, NSF IIS 1546482 BIGDATA, NIH R01MH102339, NSF IIS 1408910, NSF IIS 1332109, NIH R01GM083084.

MSC2010 subject classifications. Primary 62F30, 90C26; secondary 62J12, 90C52.

Key words and phrases. Nonconvex sparse learning, pathwise coordinate optimization, global linear convergence, optimal statistical rates of convergence, oracle property, active set, strong rule.

$\|\cdot\|_2$ denote the ℓ_2 norm, and $\mathcal{R}_\lambda(\theta)$ denote a sparsity-inducing regularizer with a regularization parameter $\lambda > 0$. We can obtain a sparse estimator of θ^* by solving the following regularized least square optimization problem:

$$(1.1) \quad \min_{\theta \in \mathbb{R}^d} \mathcal{F}_\lambda(\theta) \quad \text{where } \mathcal{F}_\lambda(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2 + \mathcal{R}_\lambda(\theta).$$

Popular choices of $\mathcal{R}_\lambda(\theta)$ are usually coordinate decomposable, $\mathcal{R}_\lambda(\theta) = \sum_{j=1}^d r_\lambda(\theta_j)$, including the ℓ_1 [Lasso, Tibshirani (1996)], SCAD [Smooth Clipped Absolute Deviation, Fan and Li (2001)] and MCP [Minimax Concavity Penalty, Zhang (2010)] regularizers. For example, the ℓ_1 regularizer takes $\mathcal{R}_\lambda(\theta) = \lambda \|\theta\|_1 = \lambda \sum_j |\theta_j|$ with $r_\lambda(|\theta_j|) = \lambda |\theta_j|$ for $j = 1, \dots, d$.

The ℓ_1 regularizer is convex and computationally tractable, but often induces large estimation bias, and requires a restrictive irrepresentable condition to attain variable selection consistency [Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Zou (2006)]. To address this issue, nonconvex regularizers such as SCAD and MCP have been proposed to obtain nearly unbiased estimators. Throughout the rest of the paper, we only consider MCP as an example due to space limit, but the extension to SCAD is straightforward. Particularly, let \mathcal{E} be an event, we define $\mathbb{1}_{\{\mathcal{E}\}}$ as an indicator function with $\mathbb{1}_{\{\mathcal{E}\}} = 1$ if \mathcal{E} holds and $\mathbb{1}_{\{\mathcal{E}\}} = 0$ otherwise. Given $\gamma > 1$, MCP has

$$(1.2) \quad r_\lambda(|\theta_j|) = \lambda \left(|\theta_j| - \frac{\theta_j^2}{2\lambda\gamma} \right) \cdot \mathbb{1}_{\{|\theta_j| < \lambda\gamma\}} + \frac{\lambda^2\gamma}{2} \cdot \mathbb{1}_{\{|\theta_j| \geq \lambda\gamma\}}.$$

We call γ the concavity parameter of MCP, since it essentially characterizes the concavity of the MCP regularizer: A larger γ implies that the regularizer is less concave. We observe that the MCP regularizer can be written as

$$(1.3) \quad \mathcal{R}_\lambda(\theta) = \lambda \|\theta\|_1 + \mathcal{H}_\lambda(\theta),$$

where $\mathcal{H}_\lambda(\theta) = \sum_{j=1}^d h_\lambda(|\theta_j|)$ is a smooth, concave and also coordinate decomposable function with

$$(1.4) \quad h_\lambda(|\theta_j|) = -\frac{\theta_j^2}{2\gamma} \cdot \mathbb{1}_{\{|\theta_j| < \lambda\gamma\}} + \frac{\lambda^2\gamma - 2\lambda|\theta_j|}{2} \cdot \mathbb{1}_{\{|\theta_j| \geq \lambda\gamma\}}.$$

We present several examples of the MCP regularizer in Figure 1. Fan and Li (2001), Zhang (2010) show that the nonconvex regularizer effectively reduces the estimation bias, and achieve better performance than the ℓ_1 regularizer in both parameter estimation and support recovery. Particularly, given a suitable chosen $\gamma < \infty$, they show that there exists a local optimum to (1.1), which attains the oracle properties under much weaker conditions. However, they cannot provide specific algorithms that guarantee such a local optimum in polynomial time due to the nonconvexity.

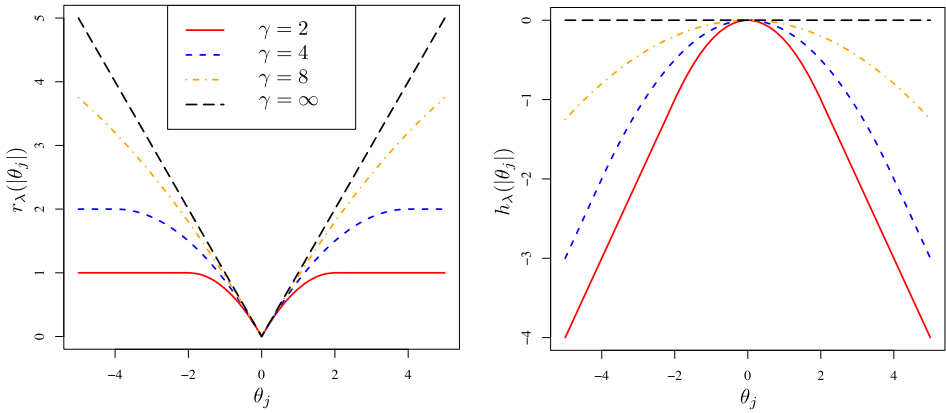


FIG. 1. Several examples of the MCP regularizer with $\lambda = 1$ and $\gamma = 2, 4, 8$ and ∞ (Lasso). The MCP regularizer reduces the estimation bias and achieve better performance than the ℓ_1 regularizer in both parameter estimation and support recovery, but imposes great computational challenge.

Typical algorithms for solving (1.1) developed in existing optimization literature include proximal gradient algorithms [Nesterov (2013)] and coordinate optimization algorithms [Luo and Tseng (1992), Shalev-Shwartz and Tewari (2011), Richtárik and Takáč (2014), Lu and Xiao (2015)]. The proximal gradient algorithms need to access all entries of the design matrix X in each iteration for computing a full gradient and a sophisticated line search step. Thus, they are often not scalable and efficient in practice when d is large. To address this issue, many researchers resort to the coordinate optimization algorithms for better computational efficiency and scalability.

The classical coordinate optimization algorithm is straightforward and much simpler than the proximal gradient algorithms in each iteration: Given $\theta^{(t)}$ at the t th iteration, we select a coordinate j , and then take an exact coordinate minimization step

$$(1.5) \quad \theta_j^{(t+1)} = \underset{\theta_j}{\operatorname{argmin}} \mathcal{F}_\lambda(\theta_j, \theta_{\setminus j}^{(t)}),$$

where $\theta_{\setminus j}$ is a subvector of θ with the j th entry removed. For the ℓ_1 , SCAD, and MCP regularizers, (1.5) admits a closed-form solution. For notational simplicity, we denote $\theta_j^{(t+1)} = \mathcal{T}_{\lambda,j}(\theta^{(t)})$. Then (1.5) can be rewritten as

$$(1.6) \quad \theta_j^{(t+1)} = \mathcal{T}_{\lambda,j}(\theta^{(t)}) = \underset{\theta_j}{\operatorname{argmin}} \frac{1}{2n} \|z^{(t)} - X_{*j}\theta_j\|_2^2 + r_\lambda(\theta_j),$$

where X_{*j} denotes the j th column of X and $z^{(t)} = y - X\theta^{(t)} + X_{*j}\theta_j^{(t)}$ is the partial residual. Without loss of generality, we assume that X satisfies the column

normalization condition $\|X_{*j}\|_2 = \sqrt{n}$ for all $j = 1, \dots, d$. Let $\tilde{\theta}_j^{(t)} = \frac{1}{n} X_{*j}^\top z^{(t)}$. Then for MCP, we obtain $\theta_j^{(t+1)}$ by

$$(1.7) \quad \theta_j^{(t+1)} = \tilde{\theta}_j^{(t)} \cdot \mathbb{1}_{\{|\tilde{\theta}_j^{(t)}| \geq \gamma\lambda\}} + \frac{\mathcal{S}_\lambda(\tilde{\theta}_j^{(t)})}{1 - 1/\gamma} \cdot \mathbb{1}_{\{|\tilde{\theta}_j^{(t)}| < \gamma\lambda\}},$$

where $\mathcal{S}_\lambda(a) = \text{sign}(a) \cdot \max\{|a| - \lambda, 0\}$. As shown in the supplementary materials [Zhao, Liu and Zhang (2018)], (1.7) can be efficiently calculated by a simple partial residual update trick, which only requires the access to one single column of the design matrix X_{*j} (Recall the proximal gradient algorithms need to access the entire design matrix). Once we obtain $\theta_j^{(t+1)}$, we take $\theta_{\setminus j}^{(t+1)} = \theta_{\setminus j}^{(t)}$. Such a coordinate optimization algorithm, though simple, is not necessarily efficient in theory and practice. Existing optimization theory only shows its sublinear convergence to local optima in high dimensions if we select coordinates from 1 to d in a cyclic order throughout all iterations [Razaviyayn, Hong and Luo (2013), Li et al. (2016)]. Moreover, no theoretical guarantee has been established on statistical properties of the obtained estimators for nonconvex regularizers in parameter estimation and support recovery. Thus, the coordinate optimization algorithms were almost neglected until recent rediscovery by Friedman et al. (2007), Mazumder, Friedman and Hastie (2011), Tibshirani et al. (2012).²

REMARK 1.1 (Connection between MCP and Lasso). Let $\frac{c}{\infty} = 0$ for any constant c . As can be seen from (1.2), for $\gamma = \infty$, MCP is reduced to the ℓ_1 regularizer, that is, $r_\lambda(|\theta_j|) = \lambda|\theta_j|$ with $h_\lambda(|\theta_j|) = 0$. Accordingly, (1.7) is reduced to $\theta_j^{(t+1)} = \mathcal{S}_\lambda(\tilde{\theta}_j^{(t)})$, which is identical to the updating formula of the coordinate optimization algorithm proposed in Fu (1998) for Lasso. Thus, throughout the rest of the paper, we just simply consider the ℓ_1 regularizer as a special case of MCP, unless we clearly specify the difference between $\gamma < \infty$ and $\gamma = \infty$ for MCP.

As illustrated in Figure 2, Hastie and Tibshirani (1990), Mazumder, Friedman and Hastie (2011), Tibshirani et al. (2012) propose a pathwise coordinate optimization framework with three nested loops, which integrates the warm start initialization, active set updating strategy, and strong rule for coordinate preselection into the classical coordinate optimization.

Particularly, in the *outer loop*, the warm start initialization optimizes (1.1) with a sequence of decreasing regularization parameters in a multistage manner, and yields solutions from sparse to dense. Within each stage of the warm start initialization (an iteration of the outer loop), the algorithm uses the solution from the previous stage for initialization, and then adopts the active set updating strategy

²A brief history on applying coordinate optimization to sparse learning problems is presented in Hastie (2009).

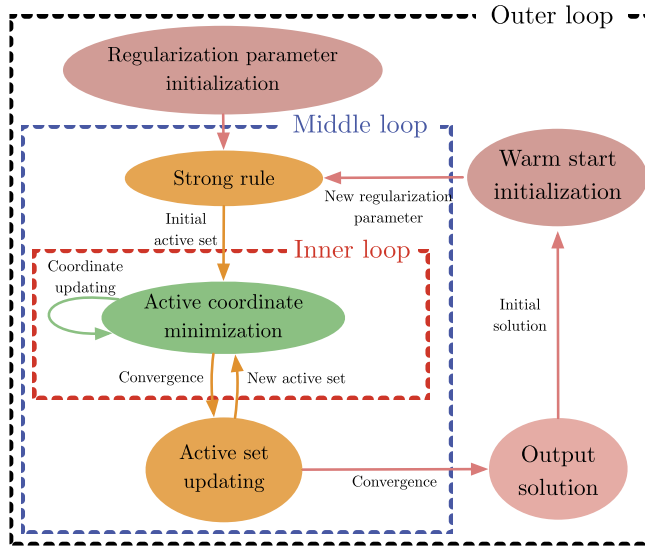


FIG. 2. The pathwise coordinate optimization framework contains 3 nested loops: (I) Warm start initialization; (II) Active set updating and strong rule for coordinate preselection; (III) Active coordinate minimization. Many empirical results have corroborated its outstanding performance. Detailed descriptions of the loops are presented in Section 2.

to exploit the solution sparsity to speed up computation. The active set updating strategy contains two consequent nested loops: In the *middle loop*, the algorithm first divides all coordinates into active ones (active set) and inactive ones (inactive set) based on some heuristic coordinate gradient thresholding rule [strong rule, Tibshirani et al. (2012)]. Then within each iteration of the middle loop, an *inner loop* is called to conduct coordinate optimization. In general, the algorithm runs an inner loop on the current active coordinates until convergence, with all inactive coordinates remain zero. The algorithm then exploits some heuristic rule to identify a new active set, which further decreases the objective value and repeats the inner loops. The iteration within each stage terminates when the active set in the middle loop no longer changes. In practice, the warm start initialization, active set updating strategy, and strong rule for coordinate preselection encourage the algorithm to iterate over a small active set involving only a small number of coordinates and, therefore, significantly boost the computational efficiency and scalability. Software packages such as GLMNET, SparseNet and HUGE have been developed and widely applied to many research areas [Hastie and Tibshirani (1990), Zhao et al. (2012)].

Despite of the popularity of the pathwise coordinate optimization framework, we are still in lack of adequate theory to justify its superior computational performance due to its complex algorithmic structure. The warm start initialization, active set updating strategy, and strong rule for coordinate preselection are only

considered as engineering heuristics in existing literature. On the other hand, many experimental results have shown that the pathwise coordinate optimization framework is effective at finding local optima with good empirical performance, yet no theoretical guarantee has been established. Thus, a gap exists between theory and practice.

To bridge this gap, we propose a new algorithm, named PICASSO (Path-wise CalibrAted Sparse Shooting algOrithm), which improves the existing pathwise coordinate optimization framework. Particularly, we propose a new greedy selection rule for active set updating and a new convex relaxation based warm start initialization (for sparse learning problems using general loss functions beyond the least square loss). These modifications though simple, have a profound impact: The solution sparsity and restricted strong convexity can be ensured throughout all iterations, which allows us to establish statistical and computational guarantees of PICASSO in high dimensions [Zhang and Huang (2008), Bickel, Ritov and Tsybakov (2009), Raskutti, Wainwright and Yu (2010)]. Eventually, we prove that PICASSO attains a linear convergence to a unique sparse local optimum with optimal statistical properties in parameter estimation and support recovery (See more details in Section 3). To the best of our knowledge, this is the first result on the computational and statistical guarantees for the pathwise coordinate optimization framework in high dimensions.

Several proximal gradient algorithms are closely related to PICASSO. By exploiting similar sparsity structures of the optimization problem, Wang, Liu and Zhang (2014), Zhao and Liu (2016), Loh and Wainwright (2015) show that these proximal gradient algorithms also attain linear convergence to (approximate) local optima with guaranteed statistical properties. We will compare these algorithms with PICASSO in Section 6.

The rest of this paper is organized as follows: In Section 2, we present the PICASSO algorithm. In Section 3, we present a new theory for analyzing the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for sparse linear regression. In Section 4, we extend PICASSO to other sparse learning problems with general loss functions, and provide theoretical guarantees; In Section 5, we present thorough numerical experiments to support our theory. In Section 6, we discuss related work. In Section 7, we present the proofs of the theorems. Due to space limit, the proofs of all lemmas are deferred to the supplementary materials [Zhao, Liu and Zhang (2018)].

NOTATION. Given a vector $v = (v_1, \dots, v_d)^\top \in \mathbb{R}^d$, we define vector norms: $\|v\|_1 = \sum_j |v_j|$, $\|v\|_2^2 = \sum_j v_j^2$, and $\|v\|_\infty = \max_j |v_j|$. We denote the number of nonzero entries in v as $\|v\|_0 = \sum_j \mathbb{1}_{\{v_j \neq 0\}}$. We define the soft-thresholding function and operator as $\mathcal{S}_\lambda(v_j) = \text{sign}(v_j) \cdot \max\{|v_j| - \lambda, 0\}$ and $\mathcal{S}_\lambda(v) = (\mathcal{S}_\lambda(v_1), \dots, \mathcal{S}_\lambda(v_d))^\top$. We denote $v_{\setminus j} = (v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_d)^\top \in \mathbb{R}^{d-1}$ as the subvector of v with the j th entry removed. Let $\mathcal{A} \subseteq \{1, \dots, d\}$ be an index set. We use $\overline{\mathcal{A}}$ to denote the complementary set to \mathcal{A} , that is, $\overline{\mathcal{A}} = \{j \mid j \in$

$\{1, \dots, d\}$, $j \notin \mathcal{A}$). We use $v_{\mathcal{A}}$ to denote a subvector of v by extracting all entries of v with indices in \mathcal{A} . Given a matrix $A \in \mathbb{R}^{d \times d}$, we use $A_{*j} = (A_{1j}, \dots, A_{dj})^\top$ to denote the j th column of A , and $A_{k*} = (A_{k1}, \dots, A_{kd})^\top$ to denote the k th row of A . Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ be the largest and smallest eigenvalues of A . We define the matrix norms $\|A\|_{\text{F}}^2 = \sum_j \|A_{*j}\|_2^2$ and $\|A\|_2$ as the largest singular value of A . We denote $A_{\setminus i \setminus j}$ as the submatrix of A with the i th row and the j th column removed. We denote $A_{i \setminus j}$ as the i th row of A with its j th entry removed. Let $\mathcal{A} \subseteq \{1, \dots, d\}$ be an index set. We use $A_{\mathcal{A}\mathcal{A}}$ to denote a submatrix of A by extracting all entries of A with both row and column indices in \mathcal{A} .

2. Pathwise calibrated sparse shooting algorithm. We introduce the PICASSO algorithm for sparse linear regression. PICASSO is a pathwise coordinate optimization algorithm and contains three nested loops (as illustrated in Figure 2). For simplicity, we first introduce its inner loop, then its middle loop and at last its outer loop.

2.1. *Inner loop: Iterates over coordinates within an active set.* We start with the inner loop of PICASSO, which is the active coordinate minimization (ActCooMin) algorithm. The iteration index for the inner loop is (t) , where $t = 0, 1, 2, \dots$. As illustrated in Algorithm 1, the ActCooMin algorithm solves (1.1) by iteratively conducting exact coordinate minimization, but it is only allowed to iterate over a subset of all coordinates, which is called “the active set.” Accordingly, the complementary set to the active set is called “the inactive set,” because the values of these coordinates do not change throughout all iterations of

Algorithm 1: *The active coordinate minimization algorithm (ActCooMin) is the inner loop of PICASSO. It iterates over only a small subset of all coordinates in a cyclic order. Thus, its computation is scalable and efficient. Without loss of generality, we assume $|\mathcal{A}| = s$ and $\mathcal{A} = \{j_1, \dots, j_s\} \subseteq \{1, \dots, d\}$, where $j_1 \leq j_2 \leq \dots \leq j_s$.*

Algorithm: $\hat{\theta} \leftarrow \text{ActCooMin}(\lambda, \theta^{(0)}, \mathcal{A}, \tau)$

Initialize: $t \leftarrow 0$

Repeat

$w^{(t+1,0)} \leftarrow \theta^{(t)}$
For $k \leftarrow 1, \dots, s$

 $w_{j_k}^{(t+1,k)} \leftarrow \mathcal{T}_{\lambda, j_k}(w^{(t+1,k-1)}), w_{\setminus j_k}^{(t+1,k)} \leftarrow w_{\setminus j_k}^{(t+1,k-1)}$

 $\theta^{(t+1)} \leftarrow w^{(t+1,s)}$
 $t \leftarrow t + 1$

Until $\|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau\lambda$

Return: $\hat{\theta} \leftarrow \theta^{(t)}$

the inner loop. Since the active set usually contains a very small number of coordinates, the active set coordinate minimization algorithm is very scalable and efficient.

For notational simplicity, we denote the active and inactive sets by \mathcal{A} and $\overline{\mathcal{A}}$, respectively. Here, we select \mathcal{A} and $\overline{\mathcal{A}}$ based on the sparsity pattern of the initial solution of the inner loop $\theta^{(0)}$,

$$\mathcal{A} = \{j \mid \theta_j^{(0)} \neq 0\} \quad \text{and} \quad \overline{\mathcal{A}} = \{j \mid \theta_j^{(0)} = 0\}.$$

The ActCooMin algorithm then minimizes (1.1) with all coordinates of $\overline{\mathcal{A}}$ staying at zero values,

$$(2.1) \quad \min_{\theta \in \mathbb{R}^d} \mathcal{F}_\lambda(\theta) \quad \text{subject to } \theta_{\overline{\mathcal{A}}} = 0.$$

The ActCooMin algorithm iterates over all active coordinates in a cyclic order at each iteration. Without loss of generality, we assume

$$|\mathcal{A}| = s \quad \text{and} \quad \mathcal{A} = \{j_1, \dots, j_s\} \subseteq \{1, \dots, d\},$$

where $j_1 \leq j_2 \leq \dots \leq j_s$. Given a solution $\theta^{(t)}$ at the t th iteration, we construct a sequence of auxiliary solutions $\{w^{(t+1,k)}\}_{k=0}^s$ to obtain $\theta^{(t+1)}$. Particularly, for $k=0$, we take $w^{(t+1,0)} = \theta^{(t)}$; For $k=1, \dots, s$, we take

$$w_{j_k}^{(t+1,k)} = \mathcal{T}_{\lambda, j_k}(w^{(t+1,k-1)}) \quad \text{and} \quad w_{\setminus j_k}^{(t+1,k)} = w_{\setminus j_k}^{(t+1,k-1)},$$

where $\mathcal{T}_{\lambda, j_k}(\cdot)$ is defined in (1.6). We then set $\theta^{(t+1)} = w^{(t+1,s)}$ for the next iteration. Given τ as a small convergence parameter (e.g., 10^{-5}), we terminate the ActCooMin algorithm when

$$(2.2) \quad \|\theta^{(t+1)} - \theta^{(t)}\|_2 \leq \tau \lambda.$$

We then take the output solution as $\hat{\theta} = \theta^{(t+1)}$.

The ActCooMin algorithm only converges to a local optimum of (2.1), which is not necessarily a local optimum of (1.1). Thus, PICASSO needs to combine this inner loop with some active set updating scheme, which allows the active set to change. This leads to the middle loop of PICASSO.

2.2. Middle loop: Iteratively updates active sets. We then introduce the middle loop of PICASSO, which is the iterative active set updating (IteActUpd) algorithm. The iteration index of the middle loop is $[m]$, where $m = 0, 1, 2, \dots$. As illustrated in Algorithm 2, the IteActUpd algorithm simultaneously decreases the objective value and iteratively changes the active set to ensure convergence to a local optimum to (1.1). For notational simplicity, we denote the least square loss function and its gradient as $\mathcal{L}(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2$ and $\nabla \mathcal{L}(\theta) = \frac{1}{n} X^\top (X\theta - y)$.

Algorithm 2: *The iterative active set updating (IteActUpd) algorithm is the middle loop of PICASSO. It simultaneously decreases the objective value and iteratively changes the active set. To encourage the sparsity of the active set, the greedy selection rule moves only one inactive coordinate to the active set in each iteration.*

Algorithm: $\hat{\theta} \leftarrow \text{IteActUpd}(\lambda, \theta^{[0]}, \delta, \tau, \varphi)$

Initialize: $m \leftarrow 0, \mathcal{A}_0 \leftarrow \{j \mid \theta_j^{[0]} = 0, |\nabla_j \mathcal{L}(\theta^{[0]})| \geq (1 - \varphi)\lambda\} \cup \{j \mid \theta_j^{[0]} \neq 0\}$

Repeat

$\theta^{[m+0.5]} \leftarrow \text{ActCooMin}(\lambda, \theta^{[m]}, \mathcal{A}_m, \tau)$
 $\mathcal{A}_{m+0.5} \leftarrow \{j \mid \theta_j^{[m+0.5]} \neq 0\}, \bar{\mathcal{A}}_{m+0.5} \leftarrow \{j \mid \theta_j^{[m+0.5]} = 0\}$
 $k_m \leftarrow \text{argmax}_{k \in \bar{\mathcal{A}}_{m+0.5}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})|$
 $\theta_{k_m}^{[m+1]} \leftarrow \mathcal{T}_{\lambda, k_m}(\theta^{[m+0.5]}), \theta_{\setminus k_m}^{[m+1]} \leftarrow \theta_{\setminus k_m}^{[m+0.5]}$
 $\mathcal{A}_{m+1} \leftarrow \mathcal{A}_{m+0.5} \cup \{k_m\}, \bar{\mathcal{A}}_{m+1} \leftarrow \bar{\mathcal{A}}_{m+0.5} \setminus \{k_m\}$
 $m \leftarrow m + 1$

Until $|\nabla_{k_m} \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda$

Return: $\hat{\theta} \leftarrow \theta^{[m]}$

(I) **Active Set Initialization by Strong Rule:** We first introduce how PICASSO initializes the active set for each middle loop. Suppose an initial solution $\theta^{[0]}$ is supplied to the middle loop of PICASSO. [Friedman et al. \(2007\)](#) suggest a straightforward “simple rule” to initialize the active set based on the sparsity pattern of $\theta^{[0]}$,

$$(2.3) \quad \mathcal{A}_0 = \{j \mid \theta_j^{[0]} \neq 0\} \quad \text{and} \quad \bar{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0\}.$$

[Tibshirani et al. \(2012\)](#) further show that (2.3) is sometimes too conservative, and suggest a more aggressive active set initialization procedure using a “strong rule,” which often leads to better computational performance in practice. Specifically, given an active set initialization parameter $\varphi \in (0, 1)$, the strong rule³ for PICASSO initializes \mathcal{A}_0 and $\bar{\mathcal{A}}_0$ as

$$(2.4) \quad \mathcal{A}_0 = \{j \mid \theta_j^{[0]} = 0, |\nabla_j \mathcal{L}(\theta^{[0]})| \geq (1 - \varphi)\lambda\} \cup \{j \mid \theta_j^{[0]} \neq 0\},$$

$$(2.5) \quad \bar{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0, |\nabla_j \mathcal{L}(\theta^{[0]})| < (1 - \varphi)\lambda\},$$

where $\nabla_j \mathcal{L}(\theta^{[0]})$ denotes the j th entry of $\nabla \mathcal{L}(\theta^{[0]})$. As can be seen from (2.4), the strong rule yields an active set, which is no smaller than the simple rule. Note that we need the initialization parameter φ to be a reasonably small value (e.g., 0.1).

³Our proposed strong rule for PICASSO is slightly different from the sequential strong rule proposed in [Tibshirani et al. \(2012\)](#). See more details in Remark 2.2.

Otherwise, the strong rule may select too many active coordinates and compromise the solution sparsity.

(II) **Active Set Updating Strategy:** We then introduce how PICASSO updates the active set at each iteration of the middle loop. Suppose at the m th iteration ($m \geq 1$), we are supplied with a solution $\theta^{[m]}$ with a pair of active and inactive sets defined as

$$\mathcal{A}_m = \{j \mid \theta_j^{[m]} \neq 0\} \quad \text{and} \quad \bar{\mathcal{A}}_m = \{j \mid \theta_j^{[m]} = 0\}.$$

Each iteration of the `IteActUpd` algorithm contains two stages. The first stage conducts the active coordinate minimization algorithm over the active set \mathcal{A}_m until convergence, and returns a solution $\theta^{[m+0.5]}$. Note that the active coordinate minimization algorithm may yield zero values for some active coordinates. Accordingly, we remove those coordinates from the active set, and obtain a new pair of active and inactive sets as

$$\mathcal{A}_{m+0.5} = \{j \mid \theta_j^{[m+0.5]} \neq 0\} \quad \text{and} \quad \bar{\mathcal{A}}_{m+0.5} = \{j \mid \theta_j^{[m+0.5]} = 0\}.$$

The second stage checks which inactive coordinates of $\bar{\mathcal{A}}_{m+0.5}$ should be added into the active set. Existing pathwise coordinate optimization algorithms usually add inactive coordinates into the active set based on a *cyclic selection rule* [Friedman et al. (2007), Mazumder, Friedman and Hastie (2011)]. Particularly, they conduct the exact coordinate minimization over all coordinates of $\bar{\mathcal{A}}_{m+0.5}$ in a cyclic order. Accordingly, an inactive coordinate is added into the active set if the corresponding exact coordinate minimization step yields a nonzero value. Such a cyclic selection rule, however, has no control over the solution sparsity. It may add too many inactive coordinates into the active set, and compromise the solution sparsity.

To address this issue, we propose a new greedy selection rule for updating the active set. Particularly, let $\nabla_j \mathcal{L}(\theta^{[m+0.5]})$ denote the j th entry of $\nabla \mathcal{L}(\theta^{[m+0.5]})$. We select a coordinate by

$$k_m = \underset{k \in \bar{\mathcal{A}}_{m+0.5}}{\operatorname{argmax}} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})|.$$

We then terminate the `IteActUpd` algorithm if

$$(2.6) \quad |\nabla_{k_m} \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda,$$

where δ is a small convergence parameter (e.g., 10^{-5}). Otherwise, we take

$$\theta_{k_m}^{[m+1]} = \mathcal{T}_{\lambda, k_m}(\theta^{[m+0.5]}) \quad \text{and} \quad \theta_{\setminus k_m}^{[m+1]} = \theta_{\setminus k_m}^{[m+0.5]},$$

and set the new active and inactive sets as

$$\mathcal{A}_{m+1} = \mathcal{A}_{m+0.5} \cup \{k_m\} \quad \text{and} \quad \bar{\mathcal{A}}_{m+1} = \bar{\mathcal{A}}_{m+0.5} \setminus \{k_m\}.$$

REMARK 2.1. Beside the greedy selection rule, we also propose a randomized selection rule and a truncated cyclic selection rule for updating the active set. Due to space limit, we defer the details to the supplementary materials [Zhao, Liu and Zhang (2018)].

The IteActUpd algorithm, though equipped with the proposed greedy selection rule and strong rule for coordinate preselection, ensures the solution sparsity throughout iterations only for a sufficiently large regularization parameter.⁴ Otherwise, given an insufficiently large regularization parameter, the IteActUpd algorithm may still overselect the active coordinates. To address this issue, we combine the IteActUpd algorithm with a sequence of decreasing regularization parameters, which leads to the outer loop of PICASSO.

2.3. *Outer loop: Iterates over regularization parameters.* The outer loop of PICASSO is the warm start initialization (WarmStartInt). The iteration index of the outer loop is $\{K\}$, where $K = 1, \dots, N$. As illustrated in Algorithm 3, the warm start initialization solves (1.1) indexed by a geometrically decreasing sequence of regularization parameters $\{\lambda_K = \lambda_0 \eta^K\}_{K=0}^N$ with a common ratio $\eta \in (0, 1)$, and outputs a sequence of $N + 1$ solutions $\{\hat{\theta}^{\{K\}}\}_{K=0}^N$, which is also called the solution path.

Algorithm 3: *The warm start initialization is the **outer loop** of PICASSO. It solves (1.1) with respect to a decreasing sequence of regularization parameters $\{\lambda_K\}_{K=0}^N$. The leading regularization parameter λ_0 is chosen as $\lambda_0 = \|\nabla \mathcal{L}(0)\|_\infty$, which yields an all zero output solution $\hat{\theta}^{\{0\}} = 0$. For $K = 1, \dots, N$, we solve (1.1) for λ_K using $\hat{\theta}^{\{K-1\}}$ as an initial solution. $\{\tau_K\}_{K=1}^N$ and $\{\delta_K\}_{K=1}^N$ are two sequences of small convergence parameters, where τ_K and δ_K correspond to the K th outer loop iteration with the regularization parameter λ_K .*

Algorithm: $\{\hat{\theta}^{\{K\}}\}_{K=0}^N \leftarrow \text{WarmStartInt}(\{\lambda_K\}_{K=0}^N)$

Parameter: $\eta, \varphi, \{\tau_K\}_{K=1}^N, \{\delta_K\}_{K=1}^N$

Initialize: $\lambda_0 \leftarrow \|\nabla \mathcal{L}(0)\|_\infty, \hat{\theta}^{\{0\}} \leftarrow 0$

For $K \leftarrow 1, 2, \dots, N$

$\lambda_K \leftarrow \eta \lambda_{K-1}$
 $\hat{\theta}^{\{K\}} \leftarrow \text{IteActUpd}(\lambda_K, \hat{\theta}^{\{K-1\}}, \delta_K, \tau_K, \varphi)$

Return: $\{\hat{\theta}^{\{K\}}\}_{K=0}^N$

⁴As will be shown in Section 3, the choice of λ is determined by the initial solution of the middle loop.

For sparse linear regression,⁵ the warm start initialization chooses the leading regularization parameter λ_0 as $\lambda_0 = \|\nabla\mathcal{L}(0)\|_\infty = \|\frac{1}{n}X^\top y\|_\infty$. Recall $\mathcal{H}_\lambda(\theta)$ is defined in (1.3). By verifying the KKT condition, we have

$$\min_{\xi \in \partial\|0\|_1} \|\nabla\mathcal{L}(0) + \nabla\mathcal{H}_{\lambda_0}(0) + \lambda_0\xi\|_\infty = \min_{\xi \in \partial\|0\|_1} \|\nabla\mathcal{L}(0) + \lambda_0\xi\|_\infty = 0,$$

where the first equality comes from $\nabla\mathcal{H}_{\lambda_0}(0) = 0$ for the MCP regularizer (See more details in the supplementary materials [Zhao, Liu and Zhang (2018)]). This indicates that 0 is a local optimum of (1.1). Accordingly, we set $\hat{\theta}^{(0)} = 0$. Then for $K = 1, 2, \dots, N$, we solve (1.1) for λ_K using $\hat{\theta}^{\{K-1\}}$ as initialization.

The warm start initialization starts with large regularization parameters to suppress the overselection of the irrelevant coordinates $\{j \mid \theta_j^* = 0\}$ (in conjunction with the `IteActUpd` algorithm). Thus, the solution sparsity ensures the restricted convexity throughout all iterations, making the algorithm behave as if minimizing a strongly convex function. Though large regularization parameters may also yield zero values for many relevant coordinates $\{j \mid \theta_j^* \neq 0\}$ and result in larger estimation errors, this can be compensated by the decreasing regularization sequence. Eventually, PICASSO gradually recovers the relevant coordinates, reduces the estimation error of each output solution and attains a sparse output solution with optimal statistical properties in parameter estimation and support recovery.

REMARK 2.2 (Connection to the sequential strong rule). Tibshirani et al. (2012) propose a sequential strong rule for coordinate preselection, which initializes the active set for λ_K as

$$(2.7) \quad \mathcal{A}_0 = \{j \mid \theta_j^{[0]} = 0, |\nabla_j\mathcal{L}(\theta^{[0]})| \geq 2\lambda_K - \lambda_{K-1}\} \cup \{j \mid \theta_j^{[0]} \neq 0\},$$

$$(2.8) \quad \bar{\mathcal{A}}_0 = \{j \mid \theta_j^{[0]} = 0, |\nabla_j\mathcal{L}(\theta^{[0]})| < 2\lambda_K - \lambda_{K-1}\}.$$

Recall $\lambda_K = \eta\lambda_{K-1}$. Then we have $2\lambda_K - \lambda_{K-1} = (1 - (1 - \eta)/\eta)\lambda_K$. This indicates that the sequential strong rule is a special case of our strong rule for PICASSO with $\varphi = (1 - \eta)/\eta$.

3. Computational and statistical theory. We develop a new theory to analyze the pathwise coordinate optimization framework, and establish the computational and statistical properties of PICASSO for sparse linear regression. Recall our linear model assumption is $y = X\theta^* + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I_n)$.⁶ Moreover, in (1.3), we rewrite the nonconvex regularizer as $\mathcal{R}_\lambda(\theta) = \lambda\|\theta\|_1 + \mathcal{H}_\lambda(\theta)$, where

⁵When dealing with general loss functions, we need a new convex relaxation based warm start initialization approach, which will be introduced in Section 4.2.

⁶For simplicity, we only consider the Gaussian noise setting, but it is straightforward to extend our analysis to the sub-Gaussian noise setting.

$\mathcal{H}_\lambda(\theta) = \sum_{j=1}^d h_\lambda(|\theta_j|)$ is a smooth, concave and coordinate decomposable function. For notational simplicity, we define $\tilde{\mathcal{L}}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{H}_\lambda(\theta)$. Accordingly, we rewrite $\mathcal{F}_\lambda(\theta)$ as

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) = \tilde{\mathcal{L}}_\lambda(\theta) + \lambda \|\theta\|_1.$$

3.1. Computational theory. We first introduce three assumptions. The first assumption requires λ_N to be sufficiently large.

ASSUMPTION 3.1. We require that the regularization sequence satisfies

$$(3.1) \quad \lambda_N = 8\sigma \sqrt{\frac{\log d}{n}} \geq 4 \|\nabla \mathcal{L}(\theta^*)\|_\infty = \frac{4}{n} \|X^\top \varepsilon\|_\infty.$$

Moreover, we require $\eta \in [0.96, 1)$.

Assumption 3.1 ensures that all regularization parameters are sufficiently large to eliminate the irrelevant coordinates for PICASSO.

REMARK 3.2. Note that Assumption 3.1 is a deterministic bound for our chosen λ_N . As will be shown in Lemma 3.13, since $\|X^\top \varepsilon\|_\infty$ is random, we need to verify that (3.1) holds with high probability when applying PICASSO to sparse linear regression.

Before we present the second assumption, we define the largest and smallest s sparse eigenvalues of the Hessian matrix $\nabla^2 \mathcal{L}(\theta) = \frac{1}{n} X^\top X$ as follows.

DEFINITION 3.3. Given an integer $s \geq 1$, we define the largest and smallest s sparse eigenvalues as

$$\rho_+(s) = \sup_{\|v\|_0 \leq s} \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2} \quad \text{and} \quad \rho_-(s) = \inf_{\|v\|_0 \leq s} \frac{v^\top \nabla^2 \mathcal{L}(\theta) v}{\|v\|_2^2}.$$

The next lemma connects the largest and smallest s sparse eigenvalues to the restricted strong convexity and smoothness.

LEMMA 3.4. *Suppose there exists an integer s such that $0 < \rho_-(s) \leq \rho_+(s) < \infty$. For any $\theta, \theta' \in \mathbb{R}^d$ satisfying $\|\theta - \theta'\|_0 \leq s$, $\mathcal{L}(\theta)$ is restricted strongly convex and smooth,*

$$(3.2) \quad \frac{\rho_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{L}(\theta') - \mathcal{L}(\theta) - (\theta' - \theta)^\top \nabla \mathcal{L}(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2.$$

Moreover, given $\alpha = 1/\gamma \leq \rho_-(s)$ and $\tilde{\rho}_-(s) = \rho_-(s) - \alpha > 0$, where γ is the concavity parameter of MCP defined in (1.2), for any $\theta, \theta' \in \mathbb{R}^d$ satisfying $\|\theta - \theta'\|_0 \leq s$, $\tilde{\mathcal{L}}_\lambda(\theta)$ is restricted strongly convex and smooth,

$$\frac{\tilde{\rho}_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \tilde{\mathcal{L}}_\lambda(\theta') - \tilde{\mathcal{L}}_\lambda(\theta) - (\theta' - \theta)^\top \nabla \tilde{\mathcal{L}}_\lambda(\theta) \leq \frac{\rho_+(s)}{2} \|\theta' - \theta\|_2^2.$$

Meanwhile, for any $\xi \in \partial \|\theta\|_1$, $\mathcal{F}_\lambda(\theta)$ is restricted strongly convex,

$$\frac{\tilde{\rho}_-(s)}{2} \|\theta' - \theta\|_2^2 \leq \mathcal{F}_\lambda(\theta') - \mathcal{F}_\lambda(\theta) - (\theta' - \theta)^\top (\nabla \tilde{\mathcal{L}}_\lambda(\theta) + \lambda \xi).$$

Lemma 3.4 indicates the importance of the solution sparsity: When θ is sufficiently sparse, the restricted strong convexity of $\mathcal{L}(\theta)$ dominates the concavity of $\mathcal{H}_\lambda(\theta)$. Thus, if an algorithm ensures the solution sparsity throughout all iterations, it will behave like minimizing a strongly convex optimization problem. Accordingly, a linear convergence can be established. Note that Lemma 3.4 is also applicable to Lasso, since Lasso satisfies $\alpha = 1/\gamma = 1/\infty = 0$. Now we introduce the second assumption.

ASSUMPTION 3.5. Given $\|\theta^*\|_0 \leq s^*$, there exists an integer \tilde{s} such that

$$\tilde{s} \geq (484\kappa^2 + 100\kappa)s^*, \quad \rho_+(s^* + 2\tilde{s}) < \infty, \quad \text{and} \quad \tilde{\rho}_-(s^* + 2\tilde{s}) > 0,$$

where κ is defined as $\kappa = \rho_+(s^* + 2\tilde{s})/\tilde{\rho}_-(s^* + 2\tilde{s})$.

Assumption 3.5 guarantees that the optimization problem satisfies the restricted strong convexity as long as the number of active irrelevant coordinates never exceeds \tilde{s} throughout all iterations.

REMARK 3.6. Assumptions 3.1 and 3.5 are closely related to high dimensional statistical theories for sparse linear regression in existing literature. See more details in Zhang and Huang (2008), Bickel, Ritov and Tsybakov (2009), Zhang (2010), Negahban et al. (2012).

Now we introduce the last assumption on the computational parameters.

ASSUMPTION 3.7. Recall the convergence parameters δ_K 's and τ_K 's are defined in Algorithm 3, and the active set initialization parameter φ is defined in (2.4). We require for all $K = 1, \dots, N$,

$$\delta_K \leq \frac{1}{8}, \quad \tau_K \leq \frac{\delta_K}{\rho_+(s^* + 2\tilde{s})} \sqrt{\frac{\tilde{\rho}_-(1)}{\rho_+(1)(s^* + 2\tilde{s})}} \quad \text{and} \quad \varphi \leq \frac{1}{8}.$$

Assumption 3.7 guarantees that all middle and inner loops of PICASSO attain adequate precisions such that their output solutions satisfy the desired computational and statistical properties.

REMARK 3.8. All constants in our technical assumptions and proof are for providing insights of PICASSO. We do not make efforts on optimizing any of these constants. Taking Assumption 3.1 as an example, we choose $\eta \in [0.96, 1)$ just for easing our analysis. We can also choose η to be any other constant, for example, 0.95, as long as it is sufficiently close to 1. Such a change in η only affects the required sample complexity, iteration complexity and statistical rates of convergence up to small constant factors.

Now, we start with the convergence analysis for the inner loop of PICASSO. The following theorem presents the convergence rate in terms of the objective value. For notational simplicity, we omit the outer loop index K , and denote λ_K and τ_K by λ and τ , respectively.

THEOREM 3.9 (Inner loop). *Suppose Assumption 3.5 holds. If the initial active set satisfies $|\mathcal{A}| = s \leq s^* + 2\tilde{s}$, then (2.1) is essentially strongly convex. For $t = 1, 2, \dots$, we have*

$$\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\bar{\theta}) \leq \left(\frac{s\rho_+^2(s)}{s\rho_+^2(s) + \tilde{\rho}_-(s)\tilde{\rho}_-(1)} \right)^t [\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})],$$

where $\bar{\theta}$ is a unique global optimum to (2.1). Moreover, we need at most

$$\left(1 + \frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)\tilde{\rho}_-(1)} \right) \cdot \log \left(\frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})]}{\tilde{\rho}_-(1)\tau^2\lambda^2} \right)$$

iterations to terminate the ActCooMin algorithm, where τ is defined in (2.2).

Theorem 3.9 guarantees that given a sufficiently sparse active set, Algorithm 1 essentially minimizes a strongly convex optimization problem, though (1.1) is globally nonconvex. Thus, it attains a linear convergence to a unique global optimum.

Then we proceed with the convergence analysis for the middle loop of PICASSO. The following theorem presents the convergence rate in terms of the objective value. For notational simplicity, we omit the outer loop index K , and denote λ_K and δ_K by λ and δ , respectively. Moreover, we define

$$(3.3) \quad \Delta_\lambda = \frac{4\lambda^2 s^*}{\tilde{\rho}_-(s^* + \tilde{s})}, \quad \mathcal{S} = \{j \mid \theta_j^* \neq 0\} \quad \text{and} \quad \bar{\mathcal{S}} = \{j \mid \theta_j^* = 0\}.$$

THEOREM 3.10 (Middle loop). *Suppose Assumptions 3.1, 3.5 and 3.7 hold. For any $\lambda \geq \lambda_N$, if the initial solution $\theta^{[0]}$ satisfies $\|\theta_{\bar{\mathcal{S}}}^{[0]}\|_0 \leq \tilde{s}$ and $\mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \Delta_\lambda$, then regardless the active set initialized by either the strong rule or simple rule, we have $|\mathcal{A}_0 \cap \bar{\mathcal{S}}| \leq \tilde{s}$. Meanwhile, for $m = 0, 1, 2, \dots$, we also have $\|\theta_{\bar{\mathcal{S}}}^{[m]}\|_0 \leq \tilde{s} + 1$, $\|\theta_{\bar{\mathcal{S}}}^{[m+0.5]}\|_0 \leq \tilde{s}$, and*

$$\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \leq \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})\rho_+(1)} \right)^m [\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)],$$

where $\bar{\theta}^\lambda$ is a unique sparse local optimum of (1.1) satisfying

$$(3.4) \quad \mathcal{K}_\lambda(\bar{\theta}^\lambda) = \min_{\xi \in \partial \|\bar{\theta}^\lambda\|_1} \|\nabla \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda) + \lambda \xi\|_\infty = 0 \quad \text{and} \quad \|\bar{\theta}^\lambda\|_0 \leq \tilde{s}.$$

Moreover, recall δ is defined in (2.6), we need at most

$$\frac{(s^* + 2\tilde{s})\rho_+(1)}{\tilde{\rho}_-(s^* + 2\tilde{s})} \cdot \log\left(\frac{3\rho_+(1)[\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)]}{\delta^2\lambda^2}\right)$$

active set updating iterations to terminate the *IteActUpd* algorithm. Meanwhile, we have the output solution $\hat{\theta}^\lambda$ satisfying $\mathcal{K}_\lambda(\hat{\theta}^\lambda) \leq \delta\lambda$.

Theorem 3.10 guarantees that when supplied a proper initial solution, the middle loop of PICASSO attains a linear convergence to a unique sparse local optimum. Moreover, Theorem 3.10 has three important implications:

(I) The greedy rule is conservative and only select one coordinate each time. This mechanism prevents the overselection of the irrelevant coordinates and encourages the solution sparsity. In contrast, the cyclic selection rule in Mazumder, Friedman and Hastie (2011) may overselect the irrelevant coordinates and compromise the restricted convexity. An illustration is provided in Figure 3.

(II) Besides decreasing the objective value, the active coordinate minimization algorithm can remove some irrelevant coordinates from the active set. Thus, in conjunction with the greedy selection rule, the solution sparsity is ensured throughout all iterations. An illustration is provided in Figure 4. To the best of our knowledge,

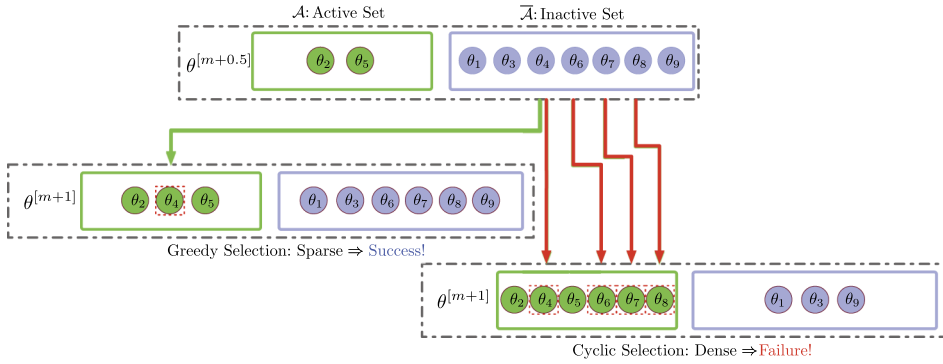


FIG. 3. An illustration of the failure of the cyclic selection rule. The green and blue circles denote the active and inactive coordinates, respectively. Suppose we have 9 coordinates and the maximum number of active coordinates we can tolerate is 4. The greedy selection rule is conservative, and only add one coordinate to the active set each time. Thus, it eventually increases the number of active coordinates from 2 to 3, and prevents from overselecting coordinates. In contrast, the cyclic selection rule used in Friedman et al. (2007), Mazumder, Friedman and Hastie (2011) leads to overselecting coordinates, which eventually increases the number of active coordinates to 6. Thus, it fails to preserve the restricted strong convexity.

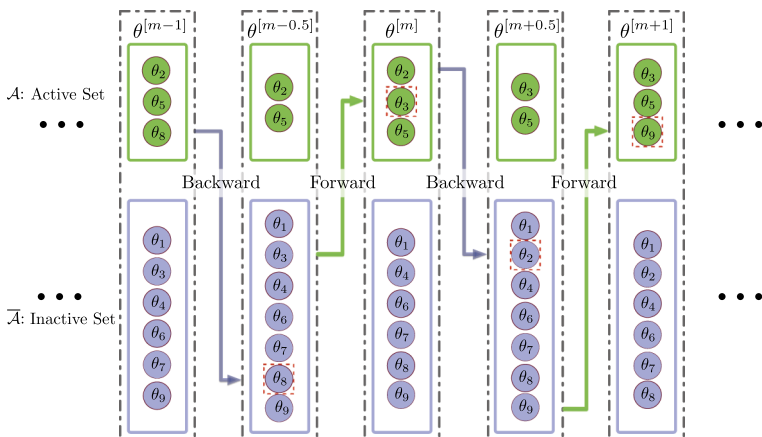


FIG. 4. An illustration of the active set updating algorithm. The green and blue circles denote the active and inactive coordinates respectively. Suppose we have 9 coordinates, and the maximum number of active coordinates we can tolerate is 4. The active set updating iteration first removes some active coordinates from the active set, then add some inactive coordinates into the active set. Thus, the number of active coordinates is ensured to never exceed 4 throughout all iterations. To the best of our knowledge, such a “forward-backward” phenomenon has not been discovered and rigorously characterized in existing literature.

such a “forward-backward” phenomenon has not been discovered and rigorously characterized in existing literature.

(III) The strong rule for coordinate preselection in PICASSO put some coordinates with zero values to the active set, only when their corresponding coordinate gradients have sufficiently large magnitudes. Thus, it can also prevent the overselection of the irrelevant coordinates and ensure the solution sparsity.

Next, we proceed with the convergence analysis for the outer loop of PICASSO. As has been shown in Theorem 3.10, each middle loop of PICASSO requires a proper initialization. Since θ^* and \mathcal{S} are unknown in practice, it is difficult to manually pick such an initial solution. The next theorem shows that the warm start initialization guides PICASSO to attain such a proper initialization for every middle loop without any prior knowledge.

LEMMA 3.11 (Outer loop). Recall Δ_{λ_K} and $\mathcal{K}_{\lambda_K}(\theta)$ are defined in (3.3) and (3.4), respectively. Suppose Assumptions 3.1, 3.5 and 3.7 hold. If θ satisfies $\|\theta_{\bar{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\mathcal{K}_{\lambda_{K-1}}(\theta) \leq \delta_{K-1}\lambda_{K-1}$, then we have

$$\begin{aligned} \|\widehat{\Delta}\|_1 &\leq 11\|\widehat{\Delta}_{\mathcal{S}}\|_1 \leq 11\sqrt{s^*}\|\widehat{\Delta}\|_2, \\ \mathcal{K}_{\lambda_K}(\theta) &\leq \frac{\lambda_K}{4}, \\ \mathcal{F}_{\lambda_K}(\theta) &\leq \mathcal{F}_{\lambda_K}(\theta^*) + \Delta_{\lambda_K}. \end{aligned}$$

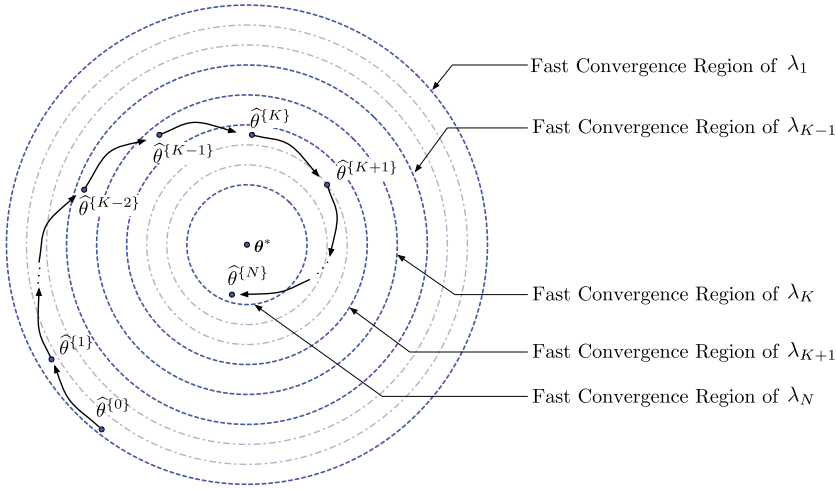


FIG. 5. An illustration of the warm start initialization (the outer loop). From an intuitive geometric perspective, the warm start initialization yields a sequence of nested fast convergence regions. We start with large regularization parameters. This suppresses the overselection of the irrelevant coordinates $\{j \mid \theta_j^* = 0\}$ and yields highly sparse solutions. With the decrease of the regularization parameter, PICASSO gradually recovers the relevant coordinates, and eventually obtains a sparse estimator $\hat{\theta}^{(N)}$ with optimal statistical properties in both parameter estimation and support recovery.

The warm start initialization starts with an all zero local optimum and a sufficiently large λ_0 , which naturally satisfy all requirements:

$$\|0_{\bar{S}}\|_0 \leq \tilde{s} \quad \text{and} \quad \mathcal{K}_{\lambda_0}(0) = 0.$$

Thus, $\theta^{[0]} = 0$ is a proper initial solution for λ_1 . Then combining Theorems 3.10 and 3.11, we show by induction that the output solution of each middle loop is always a proper initial solution for the next middle loop. The warm start initialization is illustrated in Figure 5.

Combining Theorems 3.9 and 3.10 with Lemma 3.11, we establish the global convergence in terms of the objective value for PICASSO.

THEOREM 3.12 (Main theorem). *Suppose Assumptions 3.1, 3.5 and 3.7 hold. Recall $\alpha = 1/\gamma$ and γ is the concavity parameter defined in (1.2), δ_K 's and τ_K 's are the convergence parameters for the middle and inner loops within the K th iteration of the outer loop, and κ and \tilde{s} are defined in Assumption 3.5. For $K = 1, \dots, N$, we have:*

(I) *At the K th iteration of the outer loop, the number of exact coordinate minimization iterations within each inner loop is at most*

$$\left(s^* + 2\tilde{s} + \frac{(s^* + 2\tilde{s})^2 \rho_+^2 (s^* + 2\tilde{s})}{\tilde{\rho}_- (s^* + 2\tilde{s}) \tilde{\rho}_- (1)} \right) \cdot \log \left(\frac{50s^*}{\tilde{\rho}_- (1) \tau_K^2 \tilde{\rho}_- (s^* + \tilde{s})} \right).$$

(II) *At the K th iteration of the outer loop, the number of active set updating iterations is at most*

$$\frac{(s^* + 2\tilde{s})\rho_+(1)}{\tilde{\rho}_-(s^* + 2\tilde{s})} \cdot \log\left(\frac{75s^*\rho_+(1)}{\delta_K^2 \tilde{\rho}_-(s^* + \tilde{s})}\right).$$

(III) *At the K th iteration of the outer loop, we have*

$$\mathcal{F}_{\lambda_N}(\hat{\theta}^{(K)}) - \mathcal{F}_{\lambda_N}(\bar{\theta}^{\lambda_N}) \leq [\mathbb{1}_{\{K < N\}} + \mathbb{1}_{\{K = N\}}] \cdot \delta_N \frac{50\lambda_K^2 s^*}{\tilde{\rho}_-(s^* + \tilde{s})}.$$

Theorem 3.12 guarantees that PICASSO attains a linear convergence to a unique sparse local optimum, which is a significant improvement over sublinear convergence of the randomized coordinate minimization algorithms established in existing literature. To the best of our knowledge, this is the first result establishing the convergence properties of the pathwise coordinate optimization framework in high dimensions.

3.2. *Statistical theory.* Finally, we analyze the statistical properties of the estimator obtained by PICASSO for sparse linear regression. We assume $\|\theta^*\|_0 \leq s^*$, and for any $v \neq 0$, the design matrix X satisfies

$$(3.5) \quad \psi_\ell \|v\|_2^2 - \gamma_\ell \cdot \frac{\log d}{n} \|v\|_1^2 \leq \frac{\|Xv\|_2^2}{n} \leq \psi_u \|v\|_2^2 + \gamma_u \cdot \frac{\log d}{n} \|v\|_1^2,$$

where $\gamma_\ell, \gamma_u, \psi_\ell$ and ψ_u are positive constants, and do not scale with (s^*, n, d) . Existing literature has shown that (3.5) is satisfied by many common examples of sub-Gaussian random design with high probability [Raskutti, Wainwright and Yu (2010), Negahban et al. (2012)].

We then verify Assumptions 3.1 and 3.5 by the following lemma.

LEMMA 3.13. *Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$ and (3.5) holds. Given $\lambda_N = 8\sigma \times \sqrt{\log d/n}$, we have*

$$\mathbb{P}\left(\lambda_N \geq 4\|\nabla\mathcal{L}(\theta^*)\|_\infty = \frac{4}{n}\|X^\top \varepsilon\|_\infty\right) \geq 1 - 2d^{-2}.$$

Moreover, given $\|\frac{1}{n}X^\top X\|_1 = \mathcal{O}(d)$, $\|\theta^*\|_\infty = \mathcal{O}(d)$, $\gamma \geq 4/\psi_\ell$, and large enough n , there exists a generic constant C_1 such that we have $N = \mathcal{O}_P(\log d)$,

$$\tilde{s} = C_1 s^* \geq [484\kappa^2 + 100\kappa] \cdot s^*,$$

$$\tilde{\rho}_-(s^* + 2\tilde{s}) \geq \frac{\psi_\ell}{4} \quad \text{and}$$

$$\rho_+(s^* + 2\tilde{s}) \leq \frac{5\psi_u}{4}.$$

Lemma 3.13 guarantees that the regularization sequence satisfies Assumption 3.1 with high probability, and Assumption 3.5 holds when the design matrix satisfies (3.5). Thus, by Theorem 3.12, we know that with high probability, PICASSO attains a linear convergence to a unique sparse local optimum for sparse linear regression. Moreover, Lemma 3.13 also implies that the number of regularization parameters only needs to be the order of $\log d$. Thus, solving the optimization problem with a sequence of regularization parameters does not require much additional efforts.

We then characterize the statistical rate of convergence in parameter estimation for the estimator obtained by PICASSO.

THEOREM 3.14 (Parameter estimation). *Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$ and (3.5) holds. Given $\gamma \geq 4/\psi_\ell$ and $\lambda_N = 8\sigma \sqrt{\log d/n}$, for small enough δ_N and large enough n such that $n \geq C_2 s^* \log d$ for a generic constant C_2 , we have*

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma \sqrt{\frac{s_1^*}{n}} + \sigma \sqrt{\frac{s_2^* \log d}{n}}\right),$$

where $s_1^* = |\{j \mid |\theta_j^*| \geq \gamma \lambda_N\}|$ and $s_2^* = |\{j \mid 0 < |\theta_j^*| < \gamma \lambda_N\}|$.

By dividing all nonzero θ_j^* 's into strong signals and weak signals by their magnitudes, Theorem 3.14 shows that the MCP regularizer reduces the estimation error for strong signal with magnitudes larger than $\gamma \lambda_N$ and, therefore, attains a faster statistical rate of convergence than Lasso.

REMARK 3.15 (Parameter estimation for Lasso). Theorem 3.14 is also applicable to Lasso with $\gamma = \infty$. As a result, all nonzero θ_j^* 's are considered as weak signals $|\theta_j^*| < \infty$ for all $j = 1, \dots, d$, that is, $s_1^* = 0$ and $s_2^* = s^*$. Theorem 3.14 only guarantees a slower statistical rate of convergence for Lasso,

$$\|\widehat{\theta}^{\{N\}} - \theta^*\|_2 = \mathcal{O}_P\left(\sigma \sqrt{\frac{s_2^* \log d}{n}}\right) = \mathcal{O}_P\left(\sigma \sqrt{\frac{s^* \log d}{n}}\right) \quad \text{for } \gamma = \infty.$$

We then proceed to show that the statistical rate of convergence in Theorem 3.14 is minimax optimal in parameter estimation for a suitably chosen $\gamma < \infty$. Particularly, we consider a class of sparse vectors:

$$(3.6) \quad \Theta(s_1^*, s_2^*, d) = \left\{ \theta^* \mid \theta^* \in \mathbb{R}^d, \sum_{j=1}^d \mathbb{1}_{\{|\theta_j^*| \geq \theta_{\min}\}} \leq s_1^*, \sum_{j=1}^d \mathbb{1}_{\{0 < |\theta_j^*| < \theta_{\min}\}} \leq s_2^* \right\},$$

where $\theta_{\min} = \frac{8\gamma\sigma}{\sqrt{C_2(s_1^* + s_2^*)}}$ is the threshold between strong and weak signals for some generic constant C_2 and $\gamma < \infty$. Given $s^* = s_1^* + s_2^*$ and $n \geq C_2 s^* \log d$, we have

$$\theta_{\min} = \frac{8\gamma\sigma}{\sqrt{C_2(s_1^* + s_2^*)}} \geq 8\gamma\sigma \sqrt{\frac{\log d}{n}} = \gamma\lambda_N,$$

which matches the threshold for dividing signals in Theorem 3.14. The next theorem establishes a lower bound for parameter estimation.

THEOREM 3.16 (Lower bound). *Let $\hat{\theta}$ denote any estimator of θ^* based on $y \sim N(X\theta^*, \sigma^2 I_n)$, where $\theta^* \in \Theta(s_1^*, s_2^*, d)$. Then there exists a generic constant C_4 such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\hat{\theta} - \theta^*\|_2 \geq C_4 \left(\sigma \sqrt{\frac{s_1^*}{n}} + \sigma \sqrt{\frac{s_2^* \log d}{n}} \right).$$

Theorem 3.16 guarantees that the estimator obtained by PICASSO attains the minimax optimal rates of convergence over $\Theta(s_1^*, s_2^*, d)$. The convex ℓ_1 regularizer, however, only attains a suboptimal statistical rate of convergence due to the universal estimation bias regardless the signal strength. See more details in Zhang and Huang (2008), Bickel, Ritov and Tsybakov (2009).

To analyze the support recovery performance for the estimator obtained by PICASSO, we define the oracle least square estimator $\hat{\theta}^o$ as

$$(3.7) \quad \hat{\theta}_S^o = \operatorname{argmin}_{\theta_S} \frac{1}{2n} \|y - X_{*S}\theta_S\|_2^2 \quad \text{and} \quad \hat{\theta}_{\bar{S}}^o = 0,$$

where S and \bar{S} are defined in (3.3). Recall $\bar{\theta}^{\lambda_N}$ is the unique sparse local minimizer to (1.1) with λ_N . The following theorem shows that $\bar{\theta}^{\lambda_N}$ is identical to the oracle least square estimator $\hat{\theta}^o$ with high probability.

THEOREM 3.17 (Support recovery). *Suppose (3.5) holds,*

$$(3.8) \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad \text{and} \quad \min_{j \in S} |\theta_j^*| \geq C_5 \gamma \sigma \sqrt{\frac{\log d}{n}}$$

for a generic constant C_5 . Given $4/\psi_\ell \leq \gamma < \infty$ and $\lambda_N = 8\sigma \sqrt{\log d/n}$, for large enough n , there exists a generic constant C_3 such that $\mathbb{P}(\bar{\theta}^{\lambda_N} = \hat{\theta}^o) \geq 1 - 4d^{-2}$. Meanwhile, with probability at least $1 - 4d^{-2}$, we also have

$$\|\hat{\theta}^{\{N\}} - \theta^*\|_2 \leq C_3 \sigma \sqrt{\frac{s^*}{n}} \quad \text{and} \quad \operatorname{supp}(\hat{\theta}^{\{N\}}) = \operatorname{supp}(\theta^*).$$

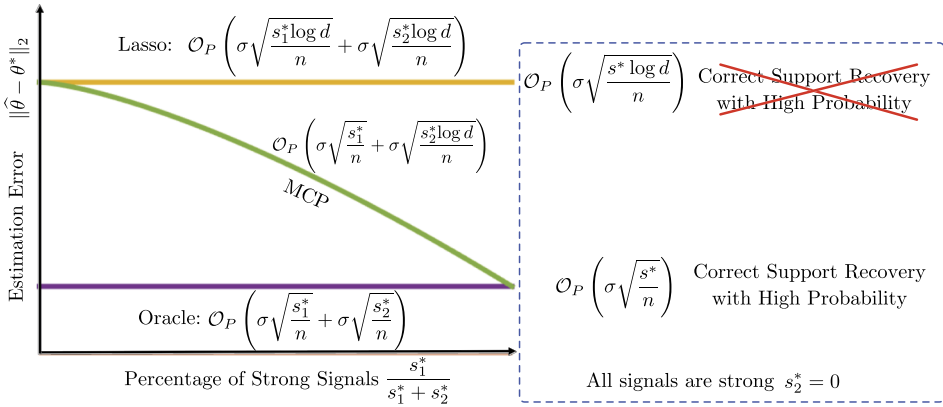


FIG. 6. An illustration of the statistical rates of convergence in parameter estimation and support recovery for the Lasso, MCP and oracle estimators. Recall s_1^* and s_2^* are defined in (3.6), and $s^* = s_1^* + s_2^*$. When all the signals are weak ($s_1^* = 0$, $s^* = s_2^*$), both the Lasso and MCP estimators attain the same estimation error bound $\mathcal{O}_P(\sigma\sqrt{s^*\log d/n})$. When some signals are strong, the MCP-regularized estimator attains a better estimation error bound $\mathcal{O}_P(\sigma\sqrt{s_1^*/n} + \sigma\sqrt{s_2^*\log d/n})$ than Lasso, because it reduces the estimation bias for the strong signals. Eventually, when all the signals are strong ($s_2^* = 0$, $s^* = s_1^*$), the MCP estimator attains the same estimation error bound as the oracle estimator $\mathcal{O}_P(\sigma\sqrt{s^*/n})$, and also correctly identify the support true support with high probability.

Theorem 3.17 guarantees that PICASSO converges to $\hat{\theta}^0$ with high probability, which is often referred to the oracle property in existing literature [Fan and Li (2001)]. Besides, we also guarantee that the estimator $\hat{\theta}^{(N)}$ obtained by PICASSO is nearly unbiased and correctly identifies the true support with high probability. Although the ℓ_1 regularizer can be viewed as a special case of MCP, such an oracle property does not hold Lasso. This is because we require $\gamma < \infty$ such that the estimation bias can be eliminated for strong signals. Thus, Lasso cannot guarantee the correct support recovery [unless the design matrix satisfies a restrictive irrepresentable condition; see more details in Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Zou (2006)]. We present an illustration of Theorems 3.14 and 3.17 in Figure 6.

4. Extension to general loss functions. PICASSO can be further extended to other regularized M-estimation problems. Taking sparse logistic regression as an example,⁷ we denote the binary response vector by $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, and the design matrix by $X \in \mathbb{R}^{n \times d}$. We consider a logistic model with $\mathbb{P}(y_i = 1) =$

⁷Due to space limit, we only present sparse logistic regression as an example. Please see more details on sparse robust regression using the huber loss function in the supplementary materials [Zhao, Liu and Zhang (2018)].

$\pi_i(\theta^*)$ and $\mathbb{P}(y_i = -1) = 1 - \pi_i(\theta^*)$, where

$$(4.1) \quad \pi_i(\theta) = \frac{1}{1 + \exp(-X_{i*}^\top \theta)} \quad \text{for } i = 1, \dots, n.$$

When θ^* is sparse, we consider the optimization problem

$$(4.2) \quad \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{where } \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n [\log(1 + \exp(-y_i X_{i*}^\top \theta))].$$

For notational simplicity, we denote the logistic loss function in (4.2) as $\mathcal{L}(\theta)$, and define $\tilde{\mathcal{L}}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{H}_\lambda(\theta)$. Then similar to sparse linear regression, we write $\mathcal{F}_\lambda(\theta)$ as

$$\mathcal{F}_\lambda(\theta) = \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) = \tilde{\mathcal{L}}_\lambda(\theta) + \lambda \|\theta\|_1.$$

The logistic loss function is twice differentiable with

$$\nabla \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n [1 - \pi_i(\theta)] y_i X_{i*} \quad \text{and} \quad \nabla^2 \mathcal{L}(\theta) = \frac{1}{n} X^\top P X,$$

where $P = \text{diag}([1 - \pi_1(\theta)]\pi_1(\theta), \dots, [1 - \pi_n(\theta)]\pi_n(\theta)) \in \mathbb{R}^{n \times n}$. Similar to sparse linear regression, we also assume that the design matrix X satisfies the column normalization condition $\|X_{*j}\|_2 = \sqrt{n}$ for all $j = 1, \dots, d$.

4.1. Proximal coordinate gradient descent. For sparse logistic regression, directly taking the minimum with respect to a selected coordinate does not admit a closed-form solution and, therefore, may involve some sophisticated algorithm such as the root-finding method.

To address this issue, [Razaviyayn, Hong and Luo \(2013\)](#) suggest a more convenient approach, which takes a proximal coordinate gradient descent iteration. For example, we select a coordinate j at the t th iteration and consider a quadratic approximation of $\mathcal{F}_\lambda(\theta_j; \theta_j^{(t)})$,

$$\mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) + \lambda |\theta_j| + \lambda \|\theta_j^{(t)}\|_1,$$

where $L > 0$ is a step size parameter, and $\mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)})$ is defined as

$$\mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \tilde{\mathcal{L}}_\lambda(\theta^{(t)}) + (\theta_j - \theta_j^{(t)}) \nabla_j \tilde{\mathcal{L}}_\lambda(\theta^{(t)}) + \frac{L}{2} (\theta_j - \theta_j^{(t)})^2.$$

Here, we choose the step size parameter L such that $\mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) \geq \mathcal{F}_\lambda(\theta_j, \theta_j^{(t)})$ for all $j = 1, \dots, d$. We then take

$$(4.3) \quad \theta_j^{(t+1)} = \underset{\theta_j}{\operatorname{argmin}} \mathcal{Q}_{\lambda,j,L}(\theta_j; \theta^{(t)}) = \underset{\theta_j}{\operatorname{argmin}} \mathcal{V}_{\lambda,j,L}(\theta_j; \theta^{(t)}) + \lambda |\theta_j|.$$

Different from the exact coordinate minimization, (4.3) always has a closed-form solution obtained by soft thresholding. Particularly, we define $\tilde{\theta}_j^{(t)} = \theta_j^{(t)} - \nabla_j \tilde{\mathcal{L}}_\lambda(\theta^{(t)})/L$. Then we have

$$\theta_j^{(t+1)} = \operatorname{argmin}_{\theta_j} \frac{1}{2}(\theta_j - \tilde{\theta}_j^{(t)})^2 + \frac{\lambda}{L}|\theta_j| = \mathcal{S}_{\lambda/L}(\tilde{\theta}_j^{(t)}) \quad \text{and} \quad \theta_{\setminus j}^{(t+1)} = \theta_{\setminus j}^{(t)}.$$

For notational convenience, we write $\theta_j^{(t+1)} = \mathcal{T}_{\lambda, j, L}(\theta^{(t)})$. When applying PICASSO to solve sparse logistic regression, we only need to replace $\mathcal{T}_{\lambda, j}(\cdot)$ with $\mathcal{T}_{\lambda, j, L}(\cdot)$ in Algorithms 1–3.

REMARK 4.1. For sparse logistic regression, we have $\nabla_{jj}^2 \mathcal{L}(\theta) = \frac{1}{n} X_{*j}^\top P X_{*j}$. Since P is a diagonal matrix, and $\pi_i(\theta) \in (0, 1)$ for any $\theta \in \mathbb{R}^d$, we have $\|P\|_2 = \max_i P_{ii} \in (0, 1/4]$ for all $i = 1, \dots, n$. Then we have $X_{*j}^\top P X_{*j} \leq \|P\|_2 \|X_{*j}\|_2^2 \leq n/4$, where the last inequality comes from the column normalization condition of X . Thus, we choose $L = \sup_\theta \max_j \nabla_{jj}^2 \mathcal{L}(\theta) = 1/4$.

We then analyze the computational and statistical properties of the estimator obtained by PICASSO for sparse logistic regression.

4.2. *Convex relaxation based warm start initialization.* We assume that $\|\theta^*\|_0 \leq s^*$, and for any $v \neq 0$ and any θ such that $\|\theta - \theta^*\|_2 \leq R$, we have

$$(4.4) \quad \psi_\ell \|v\|_2^2 - \gamma_\ell \sqrt{\frac{\log d}{n}} \|v\|_1^2 \leq v^\top \nabla^2 \mathcal{L}(\theta) v \leq \psi_u \|v\|_2^2 + \gamma_u \sqrt{\frac{\log d}{n}} \|v\|_1^2,$$

where $\gamma_\ell, \gamma_u, \psi_\ell, \psi_u$ and R are positive constants, and do not scale with (s^*, n, d) . Existing literature has shown that many common examples of sub-Gaussian random design satisfy (4.4) with high probability [Raskutti, Wainwright and Yu (2010), Negahban et al. (2012), Loh and Wainwright (2015)].

Similar to sparse linear regression, we need to verify Assumptions 3.1 and 3.5 for sparse logistic regression by the following lemma.

LEMMA 4.2. *Suppose (4.4) holds. Given $\lambda_N = 16\sqrt{\log d/n}$, we have*

$$\mathbb{P}\left(\lambda_N \geq 4\|\nabla \mathcal{L}(\theta^*)\|_\infty = \frac{4}{n}\|X^\top w\|_\infty\right) \geq 1 - d^{-7},$$

where $w = ([1 - \pi_1(\theta^*)]y_1, \dots, [1 - \pi_n(\theta^*)]y_n)^\top$ with $\pi_i(\theta)$'s defined in (4.1). Moreover, given $\gamma \geq 4/\psi_\ell$ and $\|\theta - \theta^*\|_2 \leq R$, there exists some generic constant C_1 such that for large enough n , we have

$$\tilde{s} = C_1 s^* \geq [484\kappa^2 + 100\kappa]s^*, \quad \tilde{\rho}_-(s^* + 2\tilde{s}) \geq \frac{\psi_\ell}{2}, \quad \rho_+(s^* + 2\tilde{s}) \leq \frac{5\psi_u}{4}.$$

The proof of Lemma 4.2 directly follows the proof of Lemma 3.13 and Loh and Wainwright (2015) and, therefore, is omitted. Lemma 4.2 guarantees that the regularization sequence satisfies Assumption 3.1 with high probability, and Assumption 3.5 holds when the design matrix satisfies (4.4).

Different from sparse linear regression, however, the restricted convexity and smoothness only hold over an ℓ_2 ball centered at θ^* for sparse logistic regression. Thus, directly choosing $\widehat{\theta}^{(0)} = 0$ may violate the restricted strong convexity. A simple counterexample is $\|\theta^*\|_2 > R$, which results in $\|0 - \theta^*\|_2 > R$. To address this issue, we propose a new convex relaxation based warm start initialization to obtain an initial solution for λ_0 . Particularly, we solve the following convex relaxation of (1.1):

$$(4.5) \quad \min_{\theta \in \mathbb{R}^d} \widetilde{\mathcal{F}}_{\lambda_0}(\theta) \quad \text{where } \widetilde{\mathcal{F}}_{\lambda_0}(\theta) = \mathcal{L}(\theta) + \lambda_0 \|\theta\|_1$$

up to an adequate precision. For example, we choose θ^{relax} satisfying the approximate KKT condition of (4.5) as follows:

$$(4.6) \quad \min_{\xi \in \partial \|\theta^{\text{relax}}\|_1} \|\nabla \mathcal{L}(\theta^{\text{relax}}) + \lambda_0 \xi\|_\infty \leq \delta_0 \lambda_0,$$

where $\delta_0 \in (0, 1)$ is the initial precision parameter for λ_0 . Since δ_0 in (4.6) can be chosen as a sufficiently large value (e.g., $\delta_0 = 1/8$), computing θ^{relax} becomes very efficient even for algorithms with only sublinear rates of convergence to global optima, for example, classical coordinate minimization and proximal gradient algorithms. For notational convenience, we call the above initialization procedure the convex relaxation based warm initialization.

LEMMA 4.3. *Suppose Assumption 3.5 holds only for $\|\theta - \theta^*\|_2 \leq R$. Given $\rho_-(s^* + \widetilde{s})R \geq 9\lambda_0\sqrt{s^*} \geq 18\lambda_N\sqrt{s^*}$ and $\delta_0 = 1/8$, we have*

$$\|\theta_S^{\text{relax}}\|_0 \leq \widetilde{s}, \quad \|\theta^{\text{relax}} - \theta^*\|_2 \leq R, \quad \text{and} \quad \mathcal{F}_{\lambda_0}(\theta^{\text{relax}}) \leq \mathcal{F}_{\lambda_0}(\theta^*) + \Delta_{\lambda_0}.$$

Lemma 4.3 guarantees that θ^{relax} is a proper initial solution for λ_0 . Thus, all convergence analysis in Theorem 3.12 directly follows, and PICASSO attains a linear convergence to a unique sparse local optimum with high probability. The statistical properties can also be established accordingly. An illustration of the convex relaxation based warm start initialization is provided in Figure 7.

5. Numerical experiments. We evaluate the computational and statistical performance of PICASSO through numerical simulations. We compare PICASSO with five competitors: (1) SparseNet [Mazumder, Friedman and Hastie (2011)]; (2) Path-following Iterative Shrinkage Thresholding Algorithm [PISTA, Wang, Liu and Zhang (2014)]; (3) Accelerated PISTA [A-PISTA, Zhao and Liu (2016)]; (4) Multistage Convex Relaxation Method [Mcvx, Zhang (2013)]; (5) Local Linear Approximation [LLA, Zou and Li (2008)]. Note that each subproblem of Mcvx and LLA is solved by proximal gradient algorithms with backtracking line search.

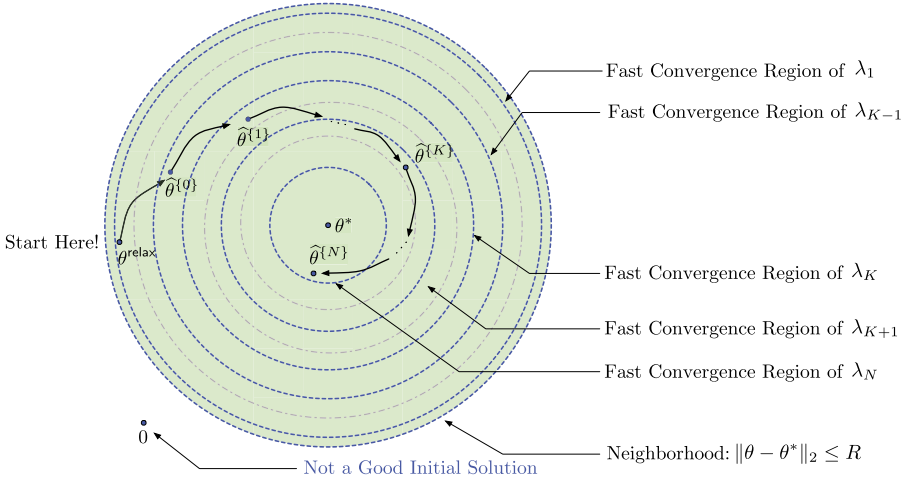


FIG. 7. An illustration of the convex relaxation based warm start initialization. When the restricted convexity and smoothness only hold over a neighborhood around θ^* (green region). Directly choosing 0 as the initial solution may violate the restricted strong convexity. Thus, we adopt a convex relaxation approach to obtain an initial solution, which is ensured to be sparse and belong to the desired neighborhood.

All experiments are conducted on a PC with Intel Core i5 3.3 GHz and 16 GB memory. All programs are coded in double precision C, called from an R wrapper. We optimize the computation by exploiting the vector and matrix sparsity, which gains a significant speedup in vector and matrix manipulations (e.g., computing the gradient and evaluating the objective value). We apply PICASSO to sparse linear regression with the MCP regularizer.

We generate each row of the design matrix X_{i*} independently from a d -dimensional Gaussian distribution with mean 0 and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, where $\Sigma_{kj} = 0.75$ and $\Sigma_{kk} = 1$ for all $j, k = 1, \dots, d$ and $k \neq j$. We then normalize each column of the design matrix X_{*j} such that $\|X_{*j}\|_2 = \sqrt{n}$. The response vector is generated from the linear model $y = X\theta^* + \varepsilon$, where $\theta^* \in \mathbb{R}^d$ is the regression coefficient vector, and ε is generated from a n -dimensional Gaussian distribution with mean 0 and covariance matrix $\sigma^2 I_n$. We set $n = 300$, $d = 18,000$, $s^* = 18$ and $\sigma^2 = 4$. θ^* has 18 nonzero entries, which are $\theta_{1000}^* = \theta_{7000}^* = \theta_{13,000}^* = 3$, $\theta_{2000}^* = \theta_{8000}^* = \theta_{14,000}^* = 2$, $\theta_{3000}^* = \theta_{9000}^* = \theta_{15,000}^* = 1.5$, $\theta_{4000}^* = \theta_{10,000}^* = \theta_{16,000}^* = -3$, $\theta_{5000}^* = \theta_{11,000}^* = \theta_{17,000}^* = -2$ and $\theta_{6000}^* = \theta_{12,000}^* = \theta_{18,000}^* = -1.5$ for $k = 0, \dots, 2$. We then set $\gamma = 1.25$, $N = 70$, $\lambda_N = 0.25\sigma\sqrt{\log d/n}$, $\varphi = 0.05$, $\delta_K = 10^{-3}$ and $\tau_K = 10^{-6}$ for all $1 \leq K \leq N$.

We present the numerical results averaged over 1000 simulations. Specifically, we create a validation set using the same design matrix as the training set for regularization parameter selection. We then tune the regularization parameter over the selected regularization sequence. We denote the response vector of the valida-

TABLE 1

Quantitative comparison on the simulated data set ($n = 300$, $d = 18,000$, $s^ = 18$, $\sigma^2 = 4$). In terms of timing performance, PICASSO slightly outperforms SparseNet, outperforms A-PISTA and greatly outperforms PISTA, LLA and Mcvx. In terms of support recovery and parameter estimation, PICASSO slightly outperforms A-PISTA, PISTA and Mcvx, and greatly outperforms SparseNet and LLA*

Method	$\ \hat{\theta} - \theta^*\ _2$	$\ \hat{\theta}_{\mathcal{S}}\ _0$	$\ \hat{\theta}_{\mathcal{S}^c}\ _0$	Correct	Timing
PICASSO	1.258 (0.515)	17.79 (0.54)	0.48 (0.52)	616/1000	1.062 (0.084)
SparseNet	1.602 (0.791)	17.64 (0.85)	2.07 (1.41)	248/1000	1.109 (0.088)
PISTA	1.267 (0.528)	17.76 (0.54)	0.55 (0.51)	614/1000	52.358 (5.920)
A-PISTA	1.276 (0.530)	17.76 (0.54)	0.57 (0.57)	613/1000	6.358 (0.865)
Mcvx	1.293 (0.529)	17.76 (0.52)	0.58 (0.52)	615/1000	67.247 (7.128)
LLA	1.517 (0.949)	17.50 (0.61)	1.28 (0.85)	365/1000	31.247 (3.870)

tion set as $\tilde{y} \in \mathbb{R}^n$. Let $\hat{\theta}^\lambda$ denote the obtained estimator using the regularization parameter λ . We then choose the optimal regularization parameter $\hat{\lambda}$ by

$$\hat{\lambda} = \underset{\lambda \in \{\lambda_1, \dots, \lambda_N\}}{\operatorname{argmin}} \|\tilde{y} - X\hat{\theta}^\lambda\|_2^2.$$

We repeat the simulation for 1000 times and summarize the averaged results in Table 1. In terms of timing performance, PICASSO slightly outperforms SparseNet, outperforms A-PISTA, and greatly outperforms PISTA, LLA and Mcvx. In terms of support recovery and parameter estimation, PICASSO slightly outperforms A-PISTA, PISTA and Mcvx, and greatly outperforms SparseNet and LLA.

To further demonstrate the superiority of PICASSO, we present a typical failure example of SparseNet using the heuristic cyclic selection rule. This example is chosen from our 1000 simulations, and illustrated in Figure 8. We see that the heuristic cyclic selection rule in SparseNet always needs to iterate over many irrelevant variables before getting to the relevant variable when identifying a new active set. Since these irrelevant variables are highly correlated with the relevant variables in our experiment, the heuristic cyclic selection rule tends to overselect the irrelevant variables and miss some relevant variables. In contrast, PICASSO, PISTA and A-PISTA have mechanisms to prevent from overselecting the irrelevant coordinates when identifying active sets. This eventually makes them outperform SparseNet in both parameter estimation and support recovery. Moreover, we also see that PISTA is much slower than other algorithms, because PISTA needs to calculate a full gradient and conduct a sophisticated line search in every iteration, which are computationally expensive. Though A-PISTA adopts the coordinate minimization to further accelerate PISTA, it still suffers from the computationally expensive line search when identifying active sets. This eventually leads to less competitive timing performance than PICASSO.

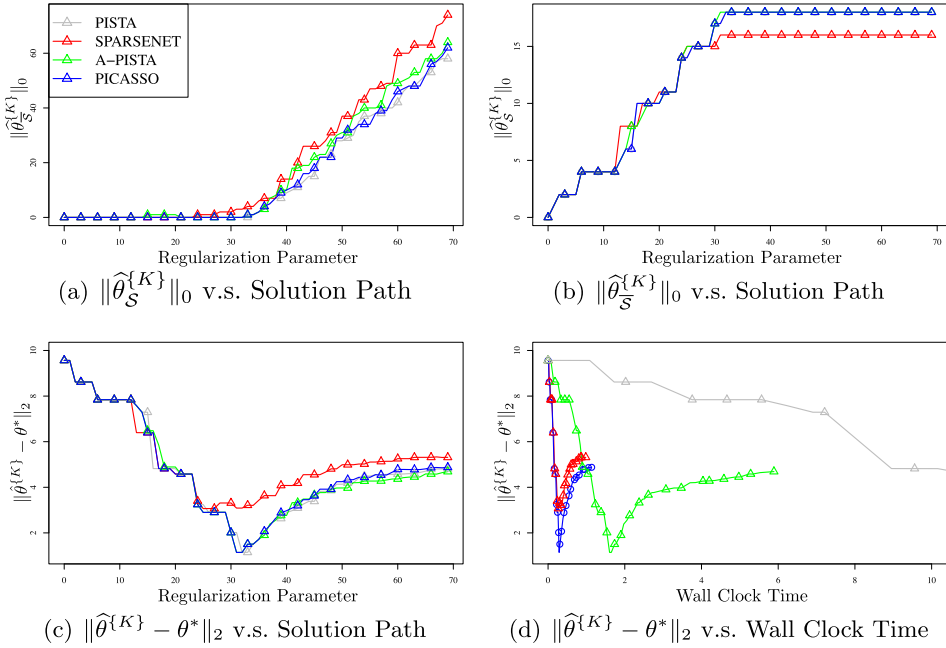


FIG. 8. A typical failure example of SparseNet using the heuristic cyclic selection rule, which is chosen from our 1000 simulations. We see that cyclic selection rule tends to overselect the irrelevant coordinates and miss some relevant coordinates when updating the active set. Thus, SparseNet eventually yields denser solutions with worse performance in parameter estimation and support recovery than PICASSO, PISTA and A-PISTA.

6. Discussions and future work. Here, we discuss several existing methods related to PICASSO, including the multistage convex relaxation method (Mcvx), local linear approximation method (LLA), path-following iterative shrinkage thresholding algorithm (PISTA), accelerated path-following iterative shrinkage thresholding algorithm (A-PISTA) and proximal gradient algorithm.

The multistage convex relaxation method is proposed in Zhang (2013). It solves a sequence of convex relaxation problems of (1.1). Zhang (2013) show that the obtained estimator enjoys similar statistical guarantees to those of PICASSO for sparse linear regression. However, there is only sublinear guarantee on its convergence rate to a local optimum. Moreover, since each relaxed problem is still lack of strong convexity, the multistage convex relaxation method needs to be combined with some efficient computational algorithms such as PICASSO.

The local linear approximation method is proposed in Zou and Li (2008), Wang, Kim and Li (2013), Fan, Xue and Zou (2014). It is essentially a special case of the multistage convex relaxation with only two iterations. Similar to the multistage convex relaxation method, it also needs an efficient computational algorithm to solve each relaxed problem. Moreover, in order to obtain the variable selection

consistency, the local linear approximation method requires a stronger minimum signal strength. Taking sparse linear regression as an example, Wang, Kim and Li (2013), Fan, Xue and Zou (2014) requires a minimum signal strength of order of $\sigma\sqrt{s^*\log d/n}$, while PICASSO only requires a minimum signal strength of order of $\sigma\sqrt{\log d/n}$.

The path-following iterative shrinkage thresholding algorithm (PISTA) is proposed in Wang, Liu and Zhang (2014). PISTA is essentially a proximal gradient algorithm combined with the warm start initialization. PISTA needs to calculate the entire (d -dimensional) gradient vector and requires a sophisticated backtracking line search procedure in every iteration. Thus, PICASSO is computationally much more efficient and scalable than PISTA in practice, although PISTA and PICASSO enjoy similar theoretical guarantees. Besides, the implementation of PISTA requires subtle control over the step size, and often yield slow empirical convergence. An accelerated PISTA algorithm (A-PISTA) is proposed in Zhao and Liu (2016), which uses coordinate minimization algorithms to accelerated PISTA. It shows an improved computational performance over PISTA in our numerical simulations, but not as competitive as PICASSO.

Moreover, when extending PISTA to general loss functions, Wang, Liu and Zhang (2014) propose a contained formulation. Particularly, they solve (1.1) with an additional constraint,

$$(6.1) \quad \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to } \|\theta\|_2 \leq R/2.$$

The additional constraint guarantees that the solution always stays in the restricted strongly convex region (a small neighborhood around θ^*), only under the assumption $\|\theta^*\|_2 \leq R/2$, where R is a constant and cannot scale with (n, s^*, d) . This assumption is very restrictive, and also introduces an additional tuning parameter. In contrast, our proposed convex relaxation based warm start initialization avoids this assumption, and allows $\|\theta^*\|_2$ to be arbitrarily large. Furthermore, we want to emphasize that PISTA exploits an explicit soft-thresholding procedure to directly control the solution sparsity in each iteration, while PICASSO adopts an algorithmic strategy to control the sparsity of the active set.

Other researchers focus on solving (1.1) with an additional constraint,

$$(6.2) \quad \min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \mathcal{R}_\lambda(\theta) \quad \text{subject to } \|\theta\|_1 \leq M,$$

where $M > 0$ is an extra tuning parameter. Loh and Wainwright (2015) show that the proximal gradient algorithm attains a linear convergence to a ball centered at θ^* to (6.2) with a radius approximately equal to the statistical error. However, the analysis of Loh and Wainwright (2015) does not justify the advantage of nonconvex regularization: They only provide a slower statistical rate of convergence than PICASSO in parameter estimation for their obtained estimator, and no support recovery guarantee is established. Besides, their analysis for general loss functions also requires the restrictive assumption: $\|\theta^*\|_2 \leq R/2$, where R is a constant and

does not scale with (n, s^*, d) . Nevertheless, PICASSO does not require this assumption.

For future work, we are interested in several possible extensions: (I) Extension to more complicated regularizers such as grouping regularizers for variable clustering; (II) Extension to more complicated (possibly nonconvex) loss functions such as sparse phase retrieval and sparse coding problems; (III) Extension to asynchronous parallel optimization setting with shared memory or communication-efficient distributed optimization setting; (IV) Extension to second-order algorithms such as the regularized iterative reweighted least square optimization algorithm for sparse generalized linear model estimation (proximal Newton). These extensions will lead to more efficient and scalable coordinate optimization algorithms for more sophisticated nonconvex optimization problems.

7. Proof of main results. We present the proof sketch of our computational and statistical theories. Some lemmas are deferred to the supplementary materials [Zhao, Liu and Zhang (2018)]. To unify the convergence analysis of PICASSO using the exact coordinate minimization (1.6) and proximal coordinate gradient descent (4.3), we define two auxiliary parameters $\nu_+(1)$ and $\nu_-(1)$. Specifically, we choose $\nu_+(1) = \nu_-(1) = L$ for the proximal coordinate gradient descent, and $\nu_+(1) = \rho_+(1)$ and $\nu_-(1) = \tilde{\rho}_-(1)$ for the exact coordinate minimization.

7.1. Proof of Theorem 3.9.

PROOF. Since $\|\theta^{(0)}\|_0 = s \leq s^* + 2\tilde{s}$, by Assumption 3.5 and Lemma 3.4, we know that (2.1) is a strongly convex optimization problem. Thus, its minimizer $\bar{\theta}$ is unique. We then introduce the following lemmas.

LEMMA 7.1. Suppose Assumption 3.5 holds, and $|\mathcal{A}| = s \leq s^* + 2\tilde{s}$. For $t = 0, 1, 2, \dots$, we have $\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\theta^{(t+1)}) \geq \frac{\nu_-(1)}{2} \|\theta^{(t)} - \theta^{(t+1)}\|_2^2$.

LEMMA 7.2. Suppose Assumption 3.5 holds, and $|\mathcal{A}| = s \leq s^* + 2\tilde{s}$. For $t = 0, 1, 2, \dots$, we have $\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\bar{\theta}) \leq \frac{s\rho_+^2(s)}{2\tilde{\rho}_-(s)} \|\theta^{(t+1)} - \theta^{(t)}\|_2^2$.

Lemmas 7.1 and 7.2 characterize the successive descent and the gap toward the optimal objective value after each iteration, respectively.

(Linear Convergence) Combining Lemmas 7.1 and 7.2, we obtain

$$\begin{aligned}
 \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\bar{\theta}) &\leq \frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)\nu_-(1)} [\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\bar{\theta})] \\
 &\quad - \frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)\nu_-(1)} [\mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\bar{\theta})].
 \end{aligned}
 \tag{7.1}$$

By simple manipulation, (7.1) implies

$$(7.2) \quad \begin{aligned} \mathcal{F}_\lambda(\theta^{(t+1)}) - \mathcal{F}_\lambda(\bar{\theta}) &\stackrel{(i)}{\leq} \left(\frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)v_-(1) + s\rho_+^2(s)} \right) [\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\bar{\theta})] \\ &\stackrel{(ii)}{\leq} \left(\frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)v_-(1) + s\rho_+^2(s)} \right)^{t+1} [\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})], \end{aligned}$$

where (ii) comes from recursively using (i).

(Number of Iterations) Combining (7.2) with Lemma 7.1, we obtain

$$\begin{aligned} \|\theta^{(t)} - \theta^{(t+1)}\|_2^2 &\stackrel{(i)}{\leq} \frac{2[\mathcal{F}_\lambda(\theta^{(t)}) - \mathcal{F}_\lambda(\bar{\theta})]}{v_-(1)} \\ &\leq \left(\frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)v_-(1) + s\rho_+^2(s)} \right)^t \frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})]}{v_-(1)}, \end{aligned}$$

where (i) comes from $\mathcal{F}_\lambda(\theta^{(t)}) \geq \mathcal{F}_\lambda(\bar{\theta})$. Thus, we need at most

$$t = \log^{-1} \left(\frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)v_-(1) + s\rho_+^2(s)} \right) \log \left(\frac{v_-(1)\tau^2\lambda^2}{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})]} \right)$$

iterations such that

$$\|\theta^{(t+1)} - \theta^{(t)}\|_2^2 \leq \left(\frac{s\rho_+^2(s)}{\tilde{\rho}_-(s)v_-(1) + s\rho_+^2(s)} \right)^t \frac{2[\mathcal{F}_\lambda(\theta^{(0)}) - \mathcal{F}_\lambda(\bar{\theta})]}{v_-(1)} \leq \tau^2\lambda^2. \quad \square$$

7.2. Proof of Theorem 3.10.

PROOF. Before we start the proof, we introduce the following lemmas.

LEMMA 7.3. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. There exists a unique sparse local optimum $\bar{\theta}^\lambda$ satisfying $\|\bar{\theta}_{\bar{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\mathcal{K}_\lambda(\bar{\theta}^\lambda) = 0$.*

LEMMA 7.4. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. If the initial solution $\theta^{(0)}$ in Algorithm 1 satisfies $\|\theta_{\bar{\mathcal{S}}}^{(0)}\|_0 \leq 2\tilde{s}$ and $\mathcal{F}_\lambda(\theta^{(0)}) \leq \mathcal{F}_\lambda(\theta^*) + \Delta_\lambda$, the output solution $\hat{\theta}$ satisfies*

$$(7.3) \quad \min_{\xi_{\mathcal{A}} \in \partial \|\hat{\theta}_{\mathcal{A}}\|_1} \|\nabla_{\mathcal{A}} \tilde{\mathcal{L}}_\lambda(\hat{\theta}) + \lambda \xi_{\mathcal{A}}\|_\infty \leq \delta\lambda \quad \text{and} \quad \|\hat{\theta}_{\bar{\mathcal{S}}}\|_0 \leq \tilde{s}.$$

LEMMA 7.5. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. If the initial solution $\theta^{[0]}$ satisfies $\|\theta_{\bar{\mathcal{S}}}^{[0]}\|_0 \leq \tilde{s}$ and $\mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \Delta_\lambda$. Then regardless the simple rule or strong rule, we have $|\mathcal{A}_0 \cap \bar{\mathcal{S}}| \leq \tilde{s}$.*

The proofs of Lemmas 7.3, 7.4 and 7.5 are provided in the supplementary materials [Zhao, Liu and Zhang (2018)]. Lemma 7.3 verifies the existence of the unique sparse local optimum. Lemma 7.4 implies that the inner loop of PICASSO removes some irrelevant coordinates, and encourages the output solution sparsity. Lemma 7.5 implies that the initial active set is sufficiently sparse for both simple and strong rules.

(Solution Sparsity) Since the objective always decreases, we have

$$(7.4) \quad \mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\theta^{[m+0.5]}) \leq \mathcal{F}_\lambda(\theta^{[0]}) \leq \mathcal{F}_\lambda(\theta^*) + \Delta_\lambda$$

for all $m = 0, 1, 2, \dots$. Since $\theta^{[0]}$ satisfies $\|\theta_{\bar{S}}^{[0]}\|_0 \leq \tilde{s}$, by Lemma 7.5, we have $|\mathcal{A}_0 \cap \bar{S}| \leq \tilde{s}$. Then by Lemma 7.4, we have $\|\theta_{\bar{S}}^{[0.5]}\|_0 \leq \tilde{s}$. Moreover, the greedy selection rule moves only one inactive coordinate to the active set and, therefore, guarantees $\|\theta_{\bar{S}}^{[1]}\|_0 \leq \tilde{s} + 1$. By induction, we prove $\|\theta_{\bar{S}}^{[m]}\|_0 \leq \tilde{s} + 1$ and $\|\theta_{\bar{S}}^{[m+0.5]}\|_0 \leq \tilde{s}$ for all $m = 0, 1, 2, \dots$.

(Linear Convergence) We first prove the linear convergence for the proximal coordinate gradient descent. We need to construct an auxiliary solution:

$$\begin{aligned} w^{[m+1]} &= \operatorname{argmin}_{w \in \mathbb{R}^d} \mathcal{J}_{\lambda, L}(w; \theta^{[m+0.5]}) \\ &= \operatorname{argmin}_{w \in \mathbb{R}^d} \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + (w - \theta^{[m+0.5]})^\top \nabla \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) \\ &\quad + \frac{L}{2} \|w - \theta^{[m+0.5]}\|_2^2 + \lambda \|w\|_1. \end{aligned}$$

We can verify $w_k^{[m+1]} = \operatorname{argmin}_{\theta_k} \mathcal{Q}_{\lambda, k, L}(\theta_k; \theta^{[m+0.5]})$ for $j = 1, \dots, d$. For notational simplicity, we define $w^{[m+1]} = \mathcal{T}_{\lambda, L}(\theta^{[m+0.5]})$. Before we proceed, we introduce the following lemmas.

LEMMA 7.6. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. For the proximal coordinate gradient descent and $m = 0, 1, 2, \dots$, we have*

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{1}{s^* + 2\tilde{s}} [\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda, L}(w^{[m+1]}; \theta^{[m+0.5]})].$$

LEMMA 7.7. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. For the proximal coordinate gradient descent and $m = 0, 1, 2, \dots$, we have*

$$\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \leq \frac{L}{\tilde{\rho}_-(s^* + 2\tilde{s})} [\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{J}_{\lambda, L}(w^{[m+1]}; \theta^{[m+0.5]})].$$

The proofs of Lemmas 7.6 and 7.7 are presented in the supplementary materials [Zhao, Liu and Zhang (2018)]. Lemmas 7.6 and 7.7 characterize the successive

descent in each iteration and the gap toward the optimal objective value after each iteration respectively. Combining Lemmas 7.6 and 7.7, we obtain

$$(7.5) \quad \begin{aligned} & \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \\ & \leq \frac{(s^* + 2\tilde{s})L}{\tilde{\rho}_-(s^* + 2\tilde{s})} ([\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)] - [\mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)]). \end{aligned}$$

By simple manipulation, (7.5) implies

$$(7.6) \quad \begin{aligned} \mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) & \leq \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^2 + 2\tilde{s})L}\right) [\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)] \\ & \leq^{(i)} \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})L}\right) [\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)] \\ & \leq^{(ii)} \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})L}\right)^{m+1} [\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)], \end{aligned}$$

where (i) comes from (7.4), and (ii) comes from recursively applying (i).

For the exact coordinate minimization, at the m th iteration, we only need to conduct a proximal coordinate gradient descent iteration with $L = \rho_+(1)$, and obtain an auxiliary solution $\tilde{\theta}^{[m+1]}$. Since $\mathcal{F}_\lambda(\theta^{[m+1]}) \leq \mathcal{F}_\lambda(\tilde{\theta}^{[m+1]})$, by (7.6), we further have

$$(7.7) \quad \mathcal{F}_\lambda(\theta^{[m+1]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \leq \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})\rho_+(1)}\right) [\mathcal{F}_\lambda(\theta^{[m]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)].$$

(Number of Iterations) Before we proceed, we introduce the following lemma.

LEMMA 7.8. *Suppose Assumption 3.5 holds. For any θ , we conduct an exact coordinate minimization or proximal coordinate gradient descent iteration over a coordinate k , and obtain w . Then we have $\mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{\nu_-(1)}{2}(w_k - \theta_k)^2$. Moreover, if $\theta_k = 0$ and $|\nabla_k \mathcal{L}(\theta)| \geq (1 + \delta)\lambda$, we have*

$$|w_k| \geq \frac{\delta\lambda}{L} \quad \text{and} \quad \mathcal{F}_\lambda(\theta) - \mathcal{F}_\lambda(w) \geq \frac{\delta^2\lambda^2}{2\nu_+(1)}.$$

Lemma 7.8 characterizes the sufficient descent when adding the selected inactive coordinate k into the active set. Assume that the selected coordinate k_m satisfies $|\nabla_{k_m} \mathcal{L}(\theta^{[m+0.5]})| \geq (1 + \delta)\lambda$. Then by Lemma 7.8 we have

$$(7.8) \quad \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \geq \mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\theta^{[m+1]}) \geq \frac{\delta^2\lambda^2}{2\nu_+(1)}.$$

Moreover, by (7.6) and (7.7), we need at most

$$m = \log^{-1} \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})\nu_+(1)}\right) \log \left(\frac{\delta^2\lambda^2}{3\nu_+(1)[\mathcal{F}_\lambda(\theta^{[0]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda)]}\right)$$

iterations such that $\mathcal{F}_\lambda(\theta^{[m+0.5]}) - \mathcal{F}_\lambda(\bar{\theta}^\lambda) \leq \frac{\delta^2 \lambda^2}{3\nu_+(1)}$, which is contradicted by (7.8). Thus, we must have $\max_{k \in \bar{\mathcal{A}}_m} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda$, and the algorithm is terminated.

(Approximately Optimal Output Solution) By Lemma 7.4, we know that when every inner loop terminates, the approximate KKT condition must hold over the active set. Since $\nabla_{\bar{\mathcal{A}}_m} \mathcal{H}_\lambda(\theta^{[m+0.5]}) = 0$, the stopping criterion

$$\max_{k \in \bar{\mathcal{A}}_m} |\nabla_k \mathcal{L}(\theta^{[m+0.5]})| \leq (1 + \delta)\lambda$$

implies that the approximate KKT condition holds over the inactive set

$$\min_{\xi_{\bar{\mathcal{A}}_m} \in \partial \|\theta_{\bar{\mathcal{A}}_m}^{[m+0.5]}\|_1} \|\nabla_{\bar{\mathcal{A}}_m} \tilde{\mathcal{L}}_\lambda(\theta^{[m+0.5]}) + \lambda \xi_{\bar{\mathcal{A}}_m}\|_\infty \leq \delta\lambda.$$

The above two approximate KKT conditions implies that $\theta^{[m+0.5]}$ satisfies the approximate KKT condition $\mathcal{K}_\lambda(\theta^{[m+0.5]}) \leq \delta\lambda$. \square

7.3. Proof of Theorem 3.12.

PROOF. [Result (I)] Before we proceed, we introduce the following lemma.

LEMMA 7.9. *Suppose Assumptions 3.1, 3.5 and 3.7 hold. For any $\lambda \geq \lambda_N$, if θ satisfies $\|\theta_{\bar{\mathcal{S}}}\|_0 \leq \tilde{s}$ and $\mathcal{K}_\lambda(\theta) \leq \delta\lambda$, where $\delta \leq 1/8$, then for any $\lambda' \in [\lambda_N, \lambda]$, we have*

$$\mathcal{F}_{\lambda'}(\theta) - \mathcal{F}_{\lambda'}(\bar{\theta}^{\lambda'}) \leq \frac{40(\mathcal{K}_\lambda(\theta) + 3(\lambda - \lambda'))(\lambda + \lambda')s^*}{\tilde{\rho}_-(s^* + \tilde{s})}.$$

The proof of Lemma 7.9 is provided in the supplementary materials [Zhao, Liu and Zhang (2018)]. If we take $\lambda = \lambda' = \lambda_K$ and $\theta = \hat{\theta}^{\{K-1\}}$, then Lemma 7.9 implies

$$(7.9) \quad \mathcal{F}_{\lambda_K}(\hat{\theta}^{\{K-1\}}) - \mathcal{F}_{\lambda_K}(\bar{\theta}^{\lambda_K}) \leq \frac{25s^* \lambda_K^2}{\tilde{\rho}_-(s^* + \tilde{s})}.$$

Since the objective value always decreases within each middle loop, for any inner loop with λ_K , we have $\mathcal{F}_{\lambda_K}(\theta^{(0)}) - \mathcal{F}_{\lambda_K}(\bar{\theta}) \leq \mathcal{F}_{\lambda_K}(\hat{\theta}^{\{K-1\}}) - \mathcal{F}_{\lambda_K}(\bar{\theta}^{\lambda_K})$. Thus, by Theorem 3.9 and (7.9), we know that the number of iterations within each inner loop is at most

$$\log^{-1} \left(\frac{\tilde{\rho}_-(s)\nu_-(1) + s\rho_+^2(s)}{s\rho_+^2(s)} \right) \log \left(\frac{\nu_-(1)\tau_K^2 \tilde{\rho}_-(s^* + \tilde{s})}{25s^*} \right).$$

[Results (II)] Combining Theorem 3.10 with (7.9), we know that the number of active set updating iterations within each middle loop is at most

$$\log^{-1} \left(1 - \frac{\tilde{\rho}_-(s^* + 2\tilde{s})}{(s^* + 2\tilde{s})\nu_+(1)} \right) \log \left(\frac{\delta_K^2 \tilde{\rho}_-(s^* + \tilde{s})}{75\nu_+(1)s^*} \right).$$

[Results (III)] For $K < N$, we take $\lambda' = \lambda_N$, $\lambda = \lambda_K$, and $\theta = \widehat{\theta}^{(K)}$. Then by Lemma 7.9, we have

$$\mathcal{F}_{\lambda_N}(\widehat{\theta}^{(K)}) - \mathcal{F}_{\lambda_N}(\bar{\theta}^{\lambda_N}) \leq \frac{25(\lambda_K + \lambda_N)(\mathcal{K}_{\lambda_K}(\widehat{\theta}^{(K)}) + 3(\lambda_K - \lambda_N))s^*}{\tilde{\rho}_-(s^* + \tilde{s})},$$

which completes the proof due to $\lambda_K > \lambda_N$ for $K = 0, \dots, N - 1$. \square

7.4. Proof of Theorem 3.16.

PROOF. For any θ^* , we consider a partition of \mathbb{R}^d as

$$\mathcal{S}_1 = \left\{ j \mid \theta_j^* \geq \frac{C_2\sigma}{\sqrt{s_1^* + s_2^*}} \right\}, \quad \text{and} \quad \mathcal{S}_0 = \left\{ j \mid \theta_j^* < \frac{C_2\sigma}{\sqrt{s_1^* + s_2^*}} \right\}.$$

We consider the first scenario, where $\mathcal{S}_0 = \emptyset$. Then we establish the lower bound for estimating $\theta_{\mathcal{S}_1}^*$ only. Let $\tilde{\theta}_{\mathcal{S}_1}$ denote any estimator of $\theta_{\mathcal{S}_1}^*$ based on $y \sim N(X_{*\mathcal{S}_1}\theta_{\mathcal{S}_1}^*, \sigma^2 I_n)$. This is essentially a low-dimensional linear regression problem since $s_1^* < n$. By the minimax lower bound for standard linear regression model in Duchi (2015), we have

$$\inf_{\tilde{\theta}_{\mathcal{S}_1}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_1} - \theta_{\mathcal{S}_1}^*\|_2 \geq C_6\sigma \sqrt{\frac{s_1^*}{n}}$$

for a generic constant C_6 . We then consider a second scenario, where $\mathcal{S}_1 = \emptyset$. Then we establish the lower bound for estimating $\theta_{\mathcal{S}_0}^*$ only. Let $\tilde{\theta}_{\mathcal{S}_0}$ denote any estimator of $\theta_{\mathcal{S}_0}^*$ based on $y \sim N(X_{*\mathcal{S}_0}\theta_{\mathcal{S}_0}^*, \sigma^2 I_n)$. This is essentially a high dimensional sparse linear regression problem. By the lower bound for sparse linear regression model established in Raskutti, Wainwright and Yu (2011), we have

$$\inf_{\tilde{\theta}_{\mathcal{S}_0}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_0} - \theta_{\mathcal{S}_0}^*\|_2 \geq 2C_7\sigma \sqrt{\frac{s_2^* \log(d - s_2^*)}{n}} \geq C_7\sigma \sqrt{\frac{s_2^* \log d}{n}},$$

where C_7 is a generic constant and the last inequality comes from the fact $s_2^* \ll d$. Combining two scenarios, we have

$$\begin{aligned} & \inf_{\hat{\theta}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\hat{\theta} - \theta^*\|_2 \\ & \geq \max \left\{ \inf_{\tilde{\theta}_{\mathcal{S}_1}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_1} - \theta_{\mathcal{S}_1}^*\|_2, \inf_{\tilde{\theta}_{\mathcal{S}_0}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_0} - \theta_{\mathcal{S}_0}^*\|_2 \right\} \\ & \geq \frac{1}{2} \inf_{\tilde{\theta}_{\mathcal{S}_1}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_1} - \theta_{\mathcal{S}_1}^*\|_2 + \frac{1}{2} \inf_{\tilde{\theta}_{\mathcal{S}_0}} \sup_{\theta \in \Theta(s_1^*, s_2^*, d)} \mathbb{E} \|\tilde{\theta}_{\mathcal{S}_0} - \theta_{\mathcal{S}_0}^*\|_2 \\ & \geq \frac{C_6}{2}\sigma \sqrt{\frac{s_1^*}{n}} + \frac{C_7}{2}\sigma \sqrt{\frac{s_2^* \log d}{n}} \geq C_4 \left(\sigma \sqrt{\frac{s_1^*}{n}} + \sigma \sqrt{\frac{s_2^* \log d}{n}} \right), \end{aligned}$$

where $C_4 = \min\{\frac{C_6}{2}, \frac{C_7}{2}\}$. \square

7.5. Proof of Theorem 3.17.

PROOF. For notational simplicity, we denote λ_N , $\widehat{\theta}^{\{N\}}$, and $\bar{\theta}^{\lambda_N}$ by λ , $\widehat{\theta}$, and $\bar{\theta}^\lambda$, respectively. Before we proceed, we introduce the following lemmas.

LEMMA 7.10. Suppose $\varepsilon \sim N(0, \sigma^2 I_n)$ and $\|X_{*j}\|_2 = \sqrt{n}$ for $j = 1, \dots, d$. Then we have

$$\mathbb{P}\left(\frac{1}{n}\|X^\top \varepsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}}\right) \leq 2d^{-2}.$$

LEMMA 7.11. Suppose Assumptions 3.1 and 3.5, and the following event

$$\mathcal{E}_1 = \left\{ \frac{1}{n}\|X^\top \varepsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}} \right\}$$

hold. We have

$$\frac{1}{n}X_{*\mathcal{S}}(y - X\widehat{\theta}^0) + \nabla_{\mathcal{S}}\mathcal{H}_\lambda(\widehat{\theta}^0) + \lambda\nabla\|\widehat{\theta}_{\mathcal{S}}^0\|_1 = 0.$$

LEMMA 7.12. Suppose Assumptions 3.1 and 3.5 and the following event

$$\mathcal{E}_2 = \left\{ \frac{1}{n}\|U^\top \varepsilon\|_\infty \geq 2\sigma\sqrt{\frac{\log d}{n}} \right\}$$

hold, where $U = X^\top(I_n - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1}X_{*\mathcal{S}}^\top)$. There exists some $\widehat{\xi}_{\mathcal{S}}^0 \in \partial\|\widehat{\theta}_{\mathcal{S}}^0\|_1$ such that

$$\frac{1}{n}X_{*\mathcal{S}}^\top(y - X\widehat{\theta}^0) + \nabla_{\mathcal{S}}\mathcal{H}_\lambda(\widehat{\theta}^0) + \lambda\widehat{\xi}_{\mathcal{S}}^0 = 0.$$

The proof of Lemma 7.10 is provided in Negahban et al. (2012), therefore, is omitted. The proofs of Lemmas 7.11 and 7.12 are presented in the supplementary materials [Zhao, Liu and Zhang (2018)]. Lemmas 7.11 and 7.12 imply that $\widehat{\theta}^0$ satisfies the KKT condition of (1.1) over \mathcal{S} and $\overline{\mathcal{S}}$, respectively. Note that the above results only depend on Conditions \mathcal{E}_1 and \mathcal{E}_2 . Meanwhile, we also have

$$\begin{aligned} \|U_{*j}\|_2 &= \|X_{*j}^\top(I_n - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1}X_{*\mathcal{S}}^\top)\|_2 \\ (7.10) \quad &\leq \|I_n - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1}X_{*\mathcal{S}}^\top\|_2 \|X_{*j}\|_2 \\ &\leq \|X_{*j}\|_2 = \sqrt{n}, \end{aligned}$$

where the last inequality comes from $\|I_n - X_{*\mathcal{S}}(X_{*\mathcal{S}}^\top X_{*\mathcal{S}})^{-1}X_{*\mathcal{S}}^\top\|_2 \leq 1$. Thus, (7.10) implies that Lemma 7.10 is also applicable to \mathcal{E}_2 . Moreover, since both $\widehat{\theta}^{\{N\}}$ and $\widehat{\theta}^0$ are sparse local optima, by Lemma C.1 of the supplementary materials, we further have $\mathbb{P}(\widehat{\theta}^0 = \bar{\theta}^\lambda) \geq 1 - 4d^{-2}$.

Moreover, since $\hat{\theta}$ converges to $\bar{\theta}^\lambda$, given a sufficiently small δ_N , we have

$$\|\nabla \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda) - \nabla \tilde{\mathcal{L}}_\lambda(\hat{\theta})\|_\infty \leq \|\tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda) - \tilde{\mathcal{L}}_\lambda(\hat{\theta})\|_2 \leq \rho_+(s^*) \|\bar{\theta}^\lambda - \hat{\theta}\|_2 \leq \omega \ll \frac{\lambda}{4}.$$

Since we have proved $\|\nabla_{\bar{S}} \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda)\|_\infty \leq \lambda/4$ in Lemma 7.12, we have

$$\|\tilde{\mathcal{L}}_\lambda(\hat{\theta})\|_\infty \leq \|\nabla_{\bar{S}} \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda)\|_\infty + \|\nabla \tilde{\mathcal{L}}_\lambda(\bar{\theta}^\lambda) - \nabla \tilde{\mathcal{L}}_\lambda(\hat{\theta})\|_\infty \leq \frac{\lambda}{4} + \omega.$$

Since $\hat{\theta}$ also satisfies the approximate KKT condition and $\delta \leq 1/8$, then we must have $\hat{\theta}_{\bar{S}} = 0$. Moreover, since we have also proved that there exists some constant C_8 such that $\min_{j \in S} |\hat{\theta}_j^\lambda| \geq C_8 \sigma \sqrt{\log d/n}$ in Lemma 7.11, then for $\omega/\rho_-(s^*) \ll C_8 \sigma \sqrt{\log d/n}$, we have

$$\min_{j \in S} |\hat{\theta}_j| = \min_{j \in S} |\bar{\theta}_j^\lambda| - \omega \geq C_8 \sigma \sqrt{\frac{\log d}{n}} > 0.$$

Combining with the fact $\hat{\theta}_{\bar{S}} = 0$, we have $\text{supp}(\hat{\theta}) = \text{supp}(\bar{\theta}^\lambda) = \text{supp}(\theta^*)$. Meanwhile, since all signals are strong enough, then by Theorem 3.14, we also have $\|\hat{\theta} - \theta^*\|_2 \leq C_3 \sigma \sqrt{\frac{s^*}{n}}$. \square

Acknowledgments. We sincerely thank Zhaoran Wang, Junwei Lu, Daniel Vainsencher, Mingyi Hong, Xingguo Li, Jason Ge and Qiang Sun for their helpful personal communications. We are grateful to the Editor, Associate Editor and anonymous referees for their insightful comments.

The R package PICASSO implementing the proposed algorithm is available on the Comprehensive R Archive Network <http://cran.r-project.org/web/packages/picasso/>.

SUPPLEMENTARY MATERIAL

Supplement to “Pathwise coordinate optimization for sparse learning: Algorithm and theory” (DOI: [10.1214/17-AOS1547SUPP](https://doi.org/10.1214/17-AOS1547SUPP); .pdf). The supplementary materials contain the supplementary proofs of the theoretical lemmas in the paper “Pathwise coordinate optimization for nonconvex sparse learning: Algorithm and theory.”

REFERENCES

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469
 DUCHI, J. (2015). Lecture Notes for Statistics and Information Theory. Available at http://stanford.edu/class/stats311/Lectures/full_notes.pdf.
 ELOYAN, A., MUSCHELLI, J., NEBEL, M. B., LIU, H., HAN, F., ZHAO, T., BARBER, A. D., JOEL, S., PEKAR, J. J., MOSTOFSKY, S. H. and CAFFO, B. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Front. Syst. Neurosci.* **6** 61.

- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581](#)
- FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. [MR3210988](#)
- FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. [MR2415737](#)
- FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. [MR1646710](#)
- HASTIE, T. (2009). Fast regularization paths via coordinate descent. In *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Denver 2008*. Plenary Talk.
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. [MR1082147](#)
- LI, X., ZHAO, T., ARORA, R., LIU, H. and HONG, M. (2016). An improved convergence analysis of cyclic block coordinate descent-type methods for strongly convex minimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.
- LIU, H., WANG, L. and ZHAO, T. (2015). Calibrated multivariate regression with application to neural semantic basis discovery. *J. Mach. Learn. Res.* **16** 1579–1606. [MR3417792](#)
- LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M -estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. [MR3335800](#)
- LU, Z. and XIAO, L. (2015). On the complexity analysis of randomized block-coordinate descent methods. *Math. Program.* **152** 615–642. [MR3369495](#)
- LUO, Z. Q. and TSENG, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.* **72** 7–35. [MR1141764](#)
- MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). *SparseNer*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. [MR2894769](#)
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#)
- NEALE, B. M., KOU, Y., LIU, L., MA'AYAN, A., SAMOCHA, K. E., SABO, A., LIN, C.-F., STEVENS, C., WANG, L.-S., MAKAROV, V. et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485** 242–245.
- NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. [MR3025133](#)
- NESTEROV, YU. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. [MR3071865](#)
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. [MR2719855](#)
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. [MR2882274](#)
- RAZAVIYAYN, M., HONG, M. and LUO, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM J. Optim.* **23** 1126–1153. [MR3063152](#)
- RIČHTÁRIK, P. and TAKÁČ, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.* **144** 1–38. [MR3179953](#)
- SHALEV-SHWARTZ, S. and TEWARI, A. (2011). Stochastic methods for ℓ_1 -regularized loss minimization. *J. Mach. Learn. Res.* **12** 1865–1892. [MR2819020](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)

- TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 245–266. [MR2899862](#)
- WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. [MR3127873](#)
- WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. [MR3269977](#)
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. [MR2604701](#)
- ZHANG, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293. [MR3160554](#)
- ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. [MR2435448](#)
- ZHAO, T. and LIU, H. (2016). Accelerated path-following iterative shrinkage thresholding algorithm with application to semiparametric graph estimation. *J. Comput. Graph. Statist.* **25** 1272–1296. [MR3572040](#)
- ZHAO, T., LIU, H. and ZHANG, T. (2018). Supplement to “Pathwise coordinate optimization for sparse learning: Algorithm and theory.” DOI:[10.1214/17-AOS1547SUPP](#).
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](#)
- ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. [MR2930633](#)
- ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. [MR2279469](#)
- ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. [MR2435443](#)

T. ZHAO
SCHOOL OF INDUSTRIAL AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
GROSECLOSE 336
755 FERST DRIVE NW
ATLANTA, GEORGIA 30332
USA
E-MAIL: tourzhao@gatech.edu

H. LIU
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
SHERRED HALL 224
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: hanliu@princeton.edu

T. ZHANG
TENCENT AI LAB
SHENNAN AVE., NANSHAN DISTRICT
SHEN ZHEN, GUANGDONG PROVINCE 518057
CHINA
E-MAIL: tongzhang@tongzhang-ml.org