Electronic Journal of Statistics
Vol. 11 (2017) 752–799
ISSN: 1935-7524
DOI: 10.1214/17-EJS1233

Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators

Yuchen Zhang

Computer Science Department, Stanford University
Stanford, California 94305
e-mail: zhangyuc@cs.stanford.edu

Martin J. Wainwright and Michael I. Jordan

Department of Statistics, UC Berkeley
Berkeley, California 94720
e-mail: wainwrig@stat.berkeley.edu; jordan@stat.berkeley.edu

Abstract: For the problem of high-dimensional sparse linear regression, it is known that an ℓ_0-based estimator can achieve a $1/n$ "fast" rate for prediction error without any conditions on the design matrix, whereas in the absence of restrictive conditions on the design matrix, popular polynomial-time methods only guarantee the $1/\sqrt{n}$ "slow" rate. In this paper, we show that the slow rate is intrinsic to a broad class of M-estimators. In particular, for estimators based on minimizing a least-squares cost function together with a (possibly nonconvex) coordinate-wise separable regularizer, there is always a "bad" local optimum such that the associated prediction error is lower bounded by a constant multiple of $1/\sqrt{n}$. For convex regularizers, this lower bound applies to all global optima. The theory is applicable to many popular estimators, including convex ℓ_1-based methods as well as M-estimators based on nonconvex regularizers, including the SCAD penalty or the MCP regularizer. In addition, we show that bad local optima are very common, in that a broad class of local minimization algorithms with random initialization typically converge to a bad solution.

MSC 2010 subject classifications: Primary 62F12; secondary 62J05.
Keywords and phrases: Sparse linear regression, high-dimensional statistics, computationally-constrained minimax theory, nonconvex optimization.

Received November 2015.

1. Introduction

The classical notion of minimax risk, which plays a central role in decision theory, is agnostic to the computational cost of estimators. In many modern inference problems, computational cost is an important consideration, driven by the growing size of modern data sets and the need to impose constraints on the amount of time that an analysis can take. Thus it has become increasingly important to study computationally-constrained analogues of the minimax estimator, in which the choice of estimator is restricted to a subset of computationally efficient estimators, and to study tradeoffs between computation and

risk [22, 39]. A fundamental question is when such computationally-constrained forms of minimax risk estimation either coincide or differ in a fundamental way from their classical counterparts.

The goal of this paper is to explore such relationships—between classical and computationally practical minimax risks—in the context of prediction error for high-dimensional sparse regression. Our main contribution is to establish a fundamental gap between the classical minimax prediction risk and the best possible risk achievable by a broad class of M-estimators based on coordinate-separable regularizers, one which includes various nonconvex regularizers that are used in practice.

In more detail, the classical linear regression model is based on a response vector $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times d}$ that are linked via the relationship

$$y = X\theta^* + w, \tag{1.1}$$

where the vector $w \in \mathbb{R}^n$ is random. Our goal is to estimate the unknown regression vector $\theta^* \in \mathbb{R}^d$. Throughout this paper, we focus on the standard Gaussian model, in which the entries of the noise vector w are i.i.d. $N(0, \sigma^2)$ variates, and the case of deterministic design, in which the matrix X is viewed as non-random. In the sparse variant of this model, the regression vector is assumed to have a small number of non-zero coefficients. In particular, for some positive integer $k < d$, the vector θ^* is said to be k-sparse if it has at most k non-zero coefficients. Thus, the model is parameterized by the triple (n, d, k) of sample size n, ambient dimension d, and sparsity k. We use $\mathbb{B}_0(k)$ to denote the ℓ_0-"ball" of all d-dimensional vectors with at most k non-zero entries.

An estimator $\widehat{\theta}$ is a measurable function of the pair (y, X), taking values in $\mathbb{R}^d$, and its quality can be assessed in different ways. In this paper, we focus on its *fixed design prediction error*, given by $\mathbb{E}\big[\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2\big]$, a quantity that measures how well $\widehat{\theta}$ can be used to predict the vector $X\theta^*$ of noiseless responses. The worst-case prediction error of an estimator $\widehat{\theta}$ over the set $\mathbb{B}_0(k)$ is given by

$$\mathcal{M}_{n,k,d}(\widehat{\theta}; X) := \sup_{\theta^* \in \mathbb{B}_0(k)} \frac{1}{n}\mathbb{E}[\|X(\widehat{\theta} - \theta^*)\|_2^2]. \tag{1.2}$$

Given that θ^* is k-sparse, the most direct approach would be to seek a k-sparse minimizer to the least-squares cost $\|y - X\theta\|_2^2$, thereby obtaining the ℓ_0-based estimator

$$\widehat{\theta}_{\ell_0} \in \arg\min_{\theta \in \mathbb{B}_0(k)} \|y - X\theta\|_2^2. \tag{1.3}$$

The ℓ_0-based estimator $\widehat{\theta}_{\ell_0}$ is known [7, 33] to satisfy a bound of the form

$$\mathcal{M}_{n,k,d}(\widehat{\theta}_{\ell_0}; X) \precsim \frac{\sigma^2\, k \log d}{n}, \tag{1.4}$$

where $\precsim$ denotes an inequality up to constant factors—that is, independent of the triple (n, d, k) as well as the standard deviation σ. However, it is not tractable to compute this estimator, since there are $\binom{d}{k}$ subsets of size k to consider.

The computational intractability of the ℓ_0-based estimator has motivated the use of various heuristic algorithms and approximations, including the basis pursuit method [10], the Dantzig selector [8], as well as the extended family of Lasso estimators [37, 10, 43, 2]. Essentially, these methods are based on replacing the ℓ_0-constraint with its ℓ_1-equivalent, in either a constrained or penalized form. There is now a very large body of work on the performance of such methods, covering different criteria including support recovery, ℓ_2-norm error and prediction error (see, e.g., the book [6] and references therein).

For the case of fixed design prediction error that is the primary focus here, such ℓ_1-based estimators are known to achieve the bound (1.4) only if the design matrix X satisfies certain conditions, such as the restricted eigenvalue (RE) condition or compatibility condition [4, 38] or the stronger restricted isometry property [8]; see the paper [38] for an overview of these various conditions, and their inter-relationships. Without such conditions, the best known guarantees for ℓ_1-based estimators are of the form

$$\mathcal{M}_{n,k,d}(\widehat{\theta}_{\ell_1}; X) \precsim \sigma R \sqrt{\frac{\log d}{n}}, \tag{1.5}$$

a bound that is valid without any RE conditions on the design matrix X whenever the k-sparse regression vector θ^* has ℓ_1-norm bounded by R (see, e.g., the papers [7, 30, 33].)

The substantial gap between the "fast" rate (1.4) and the "slow" rate (1.5) leaves open a fundamental question: is there a computationally efficient estimator attaining the bound (1.4) for general design matrices? In the following subsections, we provide an overview of the currently known results on this gap, and we then provide a high-level statement of the main result of this paper.

1.1. Lower bounds for Lasso

Given the gap between the fast rate (1.4) and Lasso's slower rate (1.5), one possibility might be that existing analyses of prediction error are overly conservative, and ℓ_1-based methods can actually achieve the bound (1.4), without additional constraints on X. Some past work has cast doubt upon this possibility. Foygel and Srebro [17] constructed a 2-sparse regression vector and a random design matrix for which the Lasso prediction error, with any choice of regularization parameter λ_n, is lower bounded by $1/\sqrt{n}$. In particular, their proposed regression vector is $\theta^* = (0, \ldots, 0, \frac{1}{2}, \frac{1}{2})$. In their design matrix, the columns are randomly generated with distinct covariances and the rightmost column is strongly correlated with the other two columns on its left. With this particular regression vector and design matrix, they show that Lasso's prediction error is lower bounded by $1/\sqrt{n}$ for *any* choice of Lasso regularization parameter λ. This construction is explicit for Lasso, and does not apply to more general M-estimators. Moreover, for this particular counterexample, there is a one-to-one correspondence between the regression vector and the design matrix, so that one can identify the non-zero coordinates of θ^* by examining the design matrix.

Consequently, for this construction, a simple reweighted form of the Lasso can be used to achieve the fast rate. In particular, the reweighted Lasso estimator

$$\widehat{\theta}_{\mathrm{wl}} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \|y - X\theta\|_2^2 + \lambda \sum_{j=1}^{d} \alpha_j |\theta_j| \right\}, \tag{1.6}$$

with λ chosen in the usual manner ($\lambda \asymp \sigma\sqrt{\frac{\log d}{n}}$), weights $\alpha_{d-1} = \alpha_d = 1$, and the remaining weights $\{\alpha_1, \ldots, \alpha_{d-2}\}$ chosen to be sufficiently large, has this property. Dalalyan et al. [12] construct a stronger counter-example, for which the prediction error of Lasso is again lower bounded by $1/\sqrt{n}$. For this counterexample, there is no obvious correspondence between the regression vector and the design matrix. Nevertheless, as we show in Appendix A, the reweighted Lasso estimator (1.6) with a proper choice of the regularization coefficients still achieves the fast rate on this example. Another related piece of work is by Candès and Plan [9]. They construct a design matrix for which the Lasso estimator, when applied with the usual choice of regularization parameter $\lambda \asymp \sigma(\frac{\log d}{n})^{1/2}$, has sub-optimal prediction error. Their matrix construction is spiritually similar to ours, but the theoretical analysis is limited to the Lasso for a particular choice of regularization parameter. Consequently, it does not rule out the possibility that the Lasso with other choices of regularization parameters, or alternatively a different polynomial-time estimators might be able to achieve the fast rate. In contrast, our hardness result applies to general M-estimators based on coordinatewise separable regularizers, and it allows for arbitrary regularization parameters.

1.2. Complexity-theoretic lower bound for polynomial-time sparse estimators

In our own recent work [42], we provided a complexity-theoretic lower bound that applies to a very broad class of polynomial-time estimators. The analysis is performed under a standard complexity-theoretic condition—namely, that the class **NP** is not a subset of the class **P/poly**—and shows that there is no polynomial-time algorithm that returns a k-sparse vector that achieves the fast rate. The lower bound is established as a function of the restricted eigenvalue of the design matrix. Given sufficiently large (n, k, d) and any $\gamma > 0$, a design matrix X with restricted eigenvalue γ can be constructed, such that every polynomial-time k-sparse estimator $\widehat{\theta}_{\mathrm{poly}}$ has its minimax prediction risk lower bounded as

$$\mathcal{M}_{n,k,d}(\widehat{\theta}_{\mathrm{poly}}; X) \gtrsim \frac{\sigma^2\, k^{1-\delta} \log d}{\gamma n}, \tag{1.7}$$

where $\delta > 0$ is an arbitrarily small positive scalar. Note that the fraction $k^{-\delta}/\gamma$, which characterizes the gap between the fast rate and the rate (1.7), could be

arbitrarily large. The lower bound has the following consequence: any estimator that achieves the fast rate must either not be polynomial-time, or must return a regression vector that is not k-sparse.

The condition that the estimator is k-sparse is essential in the proof of lower bound (1.7). In particular, the proof relies on a reduction between estimators with small prediction error in the sparse linear regression model and methods that can solve the 3-set covering problem [28], a classical problem that is known to be NP-hard. The 3-set covering problem takes as input a list of 3-sets, which are subsets of a set $\mathcal{S}$ whose cardinality is $3k$. The goal is to choose k of these subsets in order to cover the set $\mathcal{S}$. The lower bound (1.7) is established by showing that if there is a k-sparse estimator achieving better prediction error, then it provides a solution to the 3-set covering problem, as every non-zero coordinate of the estimate corresponds to a chosen subset. This hardness result does not eliminate the possibility of finding a polynomial-time estimator that returns dense vectors satisfying the fast rate. In particular, it is possible that a dense estimator cannot be used to recover a a good solution to the 3-set covering problem, implying that it is not possible to use the hardness of 3-set covering to assert the hardness of achieving small prediction error in sparse regression.

At the same time, there is some evidence that better prediction error can be achieved by dense estimators. For instance, suppose that we consider a sequence of high-dimensional sparse linear regression problems, such that the restricted eigenvalue $\gamma = \gamma_n$ of the design matrix $X \in \mathbb{R}^{n \times d}$ decays to zero at the rate $\gamma_n = 1/n^2$. For such a sequence of problems, as n diverges to infinity, the lower bound (1.7), which applies to k-sparse estimators, goes to infinity, whereas the Lasso upper bound (1.5) converges to zero. Although this behavior is somewhat mysterious, it is not a contradiction. Indeed, what makes Lasso's performance better than the lower bound (1.7) is that it allows for non-sparse estimates. In this example, truncating the Lasso's estimate to be k-sparse will substantially hurt the prediction error. In this way, we see that proving lower bounds for non-sparse estimators—the problem to be addressed in this paper—is a substantially more challenging task than proving lower bounds for estimators that must return sparse outputs.

1.3. Main results of this paper

Let us now turn to a high-level statement of the main results of this paper. Our contribution is to provide additional evidence against the polynomial achievability of the fast rate (1.4), in particular by showing that the slow rate (1.5) is a lower bound for a broad class of M-estimators, namely those based on minimizing a least-squares cost function together with a coordinate-wise decomposable regularizer. In particular, we consider estimators that are based on an objective function of the form $L(\theta; \lambda) = \frac{1}{n}\|y - X\theta\|_2^2 + \lambda\,\rho(\theta)$, for a weighted regularizer $\rho : \mathbb{R}^d \to \mathbb{R}$ that is coordinate-separable. See Section 2.1 for a precise definition of this class of estimators. Our first main result (Theorem 1) establishes that

in the regime $n < d$, there is always a matrix $X \in \mathbb{R}^{n \times d}$ such that for any coordinate-wise separable function ρ and for any choice of weight $\lambda \geq 0$, the objective L always has at least one local optimum $\widehat{\theta}_\lambda$ such that

$$\sup_{\theta^* \in \mathbb{B}_0(k)} \mathbb{E}\Big[\frac{1}{n}\|X(\widehat{\theta}_\lambda - \theta^*)\|_2^2\Big] \gtrsim \sigma \sqrt{\frac{\log n}{n}}. \tag{1.8}$$

Moreover, if the regularizer ρ is convex, then this lower bound applies to all global optima of the convex criterion L. This lower bound is applicable to many popular estimators, including the ridge regression estimator [21], the basis pursuit method [10], the Lasso estimator [37], the weighted Lasso estimator [43], the square-root Lasso estimator [2], and least squares based on nonconvex regularizers such as the SCAD penalty [16] or the MCP penalty [41].

Next, we extend the hardness result from a particular "bad" matrices to arbitrary design matrices, by presenting sufficient conditions under which the lower bound (1.8) must hold. Our second result (Theorem 2) provides a general set of conditions on the design matrix that are sufficient to ensure that any regularized M-estimator can at best achieve the slow rate. As a corollary, we prove that there are covariance matrices $\Sigma \in \mathbb{R}^{d \times d}$ for which, when a random design matrix X is generated by sampling its rows in an i.i.d. fashion from the multivariate Gaussian $\mathcal{N}(0, \Sigma)$ distribution, the slow rate is still a fundamental barrier. This negative result for random design matrices is complementary to the line of work that shows the optimality of ℓ_1-based methods on random Gaussian designs that are generated with an incoherent covariance [26, 32].

In the nonconvex setting, it is impossible (in general) to guarantee anything beyond local optimality for any solution found by a polynomial-time algorithm [19]. Nevertheless, to play the devil's advocate, one might argue that the assumption that an adversary is allowed to pick a bad local optimum could be overly pessimistic for statistical problems. In order to address this concern, we prove a third result (Theorem 3) that demonstrates that bad local solutions are difficult to avoid. Focusing on a class of local descent methods, we show that given a random isotropic initialization centered at the origin, the resulting stationary points have poor mean-squared error—that is, they can only achieve the slow rate. In this way, this paper shows that the gap between the fast and slow rates in high-dimensional sparse regression cannot be closed via standard application of a very broad class of methods. In conjunction with our earlier complexity-theoretic paper [42], it adds further weight to the conjecture that there is a fundamental gap between the performance of polynomial-time and exponential-time methods for sparse prediction.

The remainder of this paper is organized as follows. We begin in Section 2 with further background, including a precise definition of the family of M-estimators considered in this paper, some illustrative examples, and discussion of the prediction error bound achieved by the Lasso. Section 3 is devoted to the statements of our main results, along with discussion of their consequences. In Section 4, we provide the proofs of our main results, with some technical lemmas deferred to the appendices. We conclude with a discussion in Section 5.

2. Background and problem set-up

As previously described, an instance of the sparse linear regression problem is based on observing a pair $(X, y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ of instances that are linked via the linear model (1.1), where the unknown regressor θ^* is assumed to be k-sparse, and so belongs to the ℓ_0-ball $\mathbb{B}_0(k)$. Our goal is to find a good predictor, meaning a vector $\widehat{\theta}$ such that the mean-squared prediction error $\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2$ is small.

2.1. Least squares with coordinate-separable regularizers

The analysis of this paper applies to estimators that are based on minimizing a cost function of the form

$$L(\theta; \lambda) = \frac{1}{n}\|y - X\theta\|_2^2 + \lambda\, \rho(\theta), \tag{2.1}$$

where $\rho : \mathbb{R}^d \to \mathbb{R}$ is a *regularizer*, and $\lambda \geq 0$ is a regularization weight. We consider the following family $\mathcal{F}$ of coordinate-separable regularizers:

(i) The function $\rho : \mathbb{R}^d \to \mathbb{R}$ is coordinate-wise decomposable, meaning that $\rho(\theta) = \sum_{j=1}^d \rho_j(\theta_j)$ for some univariate functions $\rho_j : \mathbb{R} \to \mathbb{R}$.
(ii) Each univariate function satisfies $\rho_j(0) = 0$ and is symmetric around zero (i.e., $\rho_j(t) = \rho_j(-t)$ for all $t \in \mathbb{R}$).
(iii) On the nonnegative real line, $[0, +\infty)$, each function ρ_j is nondecreasing.

Let us consider some examples to illustrate this definition.

Bridge regression: The family of bridge regression estimates [18] take the form

$$\widehat{\theta}_{\text{bidge}} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{n}\|y - X\theta\|_2^2 + \lambda \sum_{i=1}^d |\theta|^\gamma \Big\}.$$

Note that this is a special case of the objective function (2.1) with $\rho_j(\cdot) = |\cdot|^\gamma$ for each coordinate. When $\gamma \in \{1, 2\}$, it corresponds to the Lasso estimator and the ridge regression estimator respectively. The analysis of this paper provides lower bounds for both estimators, uniformly over the choice of λ.

Weighted Lasso: The weighted Lasso estimator [43] uses a weighted ℓ_1-norm to regularize the empirical risk, and leads to the estimator

$$\widehat{\theta}_{\text{wl}} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{n}\|y - X\theta\|_2^2 + \lambda \sum_{i=1}^d \alpha_i |\theta_i| \Big\}.$$

Here $\alpha_1, \ldots, \alpha_d$ are weights that can be adaptively chosen with respect to the design matrix X. The weighted Lasso can perform better than the ordinary

Lasso, corresponding to the special case in which all α_j are all equal. For instance, on the counter-example proposed by Foygel and Srebro [17], for which the ordinary Lasso estimator achieves only the slow $1/\sqrt{n}$ rate, the weighted Lasso estimator achieves the $1/n$ convergence rate. Nonetheless, the analysis of this paper shows that there are design matrices for which the weighted Lasso, even when the weights are chosen adaptively with respect to the design, has prediction error at least a constant multiple of $1/\sqrt{n}$.

Square-root Lasso: The square-root Lasso estimator [2] is defined by minimizing the criterion

$$\widehat{\theta}_{\text{sqrt}} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{\sqrt{n}} \|y - X\theta\|_2 + \lambda \|\theta\|_1 \right\}.$$

This criterion is slightly different from our general objective function (2.1), since it involves the square root of the least-squares error. Relative to the Lasso, its primary advantage is that the optimal setting of the regularization parameter does not require the knowledge of the standard deviation of the noise. For the purposes of the current analysis, it suffices to note that by Lagrangian duality, every square-root Lasso estimate $\widehat{\theta}_{\text{sqrt}}$ is a minimizer of the least-squares criterion $\|y - X\theta\|_2$, subject to $\|\theta\|_1 \leq R$, for some radius $R \geq 0$ depending on λ. Consequently, as the weight λ is varied over the interval $[0, \infty)$, the square root Lasso yields the same solution path as the Lasso. Since our lower bounds apply to the Lasso for any choice of $\lambda \geq 0$, they also apply to all square-root Lasso solutions.

For Lasso and Square-root Lasso, it is a common practice to select the parameter λ by an iterative algorithm [36], or sample it from a data-dependent distribution [2]. In either case, the value of λ is not pre-determined. Our lower bounds capture these estimators by holding uniformly over all choices of λ.

SCAD penalty or MCP regularizer: Due to the intrinsic bias induced by ℓ_1-regularization, various forms of nonconvex regularization are widely used. Two of the most popular are the SCAD penalty, due to Fan and Li [16], and the MCP penalty, due to Zhang et al. [41]. The family of SCAD penalties takes the form

$$\phi_\lambda(t) := \frac{1}{\lambda} \begin{cases} \lambda|t| & \text{for } |t| \leq \lambda, \\ -(t^2 - 2a\lambda|t| + \lambda^2)/(2a-2) & \text{for } \lambda < |t| \leq a\lambda, \\ (a+1)\lambda^2/2 & \text{for } |t| \geq a\lambda, \end{cases}$$

where $a > 2$ is a fixed parameter. When used with the least-squares objective, it is a special case of our general set-up with $\rho_j(\theta_j) = \phi_\lambda(\theta_j)$ for each coordinate $j = 1, \ldots, d$. Similarly, the MCP penalty takes the form

$$\rho_\lambda(t) := \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

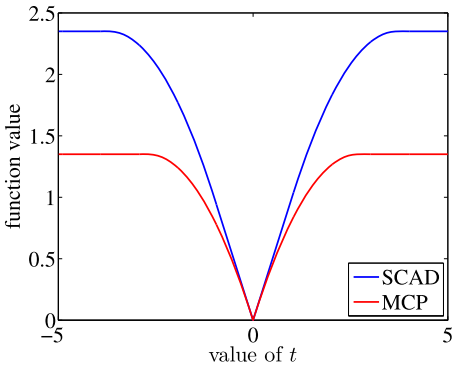

FIG 1. *Plots with regularization weight* $\lambda = 1$*, and parameters* $a = 3.7$ *for SCAD, and* $b = 2.7$ *for MCP.*

where $b > 0$ is a fixed parameter. It can be verified that both the SCAD penalty and the MCP regularizer belong to the function class $\mathcal{F}$ previously defined. See Figure 1 for a graphical illustration of the SCAD penalty and the MCP regularizer.

2.2. Prediction error for the Lasso

We now turn to a precise statement of the best known upper bounds for the Lasso prediction error. We assume that the design matrix satisfies the column normalization condition. More precisely, letting $X_j \in \mathbb{R}^n$ denote the j^{th} column of the design matrix X, we say that it is 1-*column normalized* if

$$\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 \qquad \text{for } j = 1, 2, \ldots, d. \tag{2.2}$$

Our choice of the constant 1 is to simplify notation; the more general notion allows for an arbitrary constant C in this bound.

In addition to the column normalization condition, if the design matrix further satisfies a restricted eigenvalue (RE) condition [4, 38], then the Lasso is known to achieve the fast rate (1.4) for prediction error. More precisely, restricted eigenvalues are defined in terms of subsets S of the index set $\{1, 2, \ldots, d\}$, and a cone associated with any such subset. In particular, letting S^c denote the complement of S, we define the cone

$$\mathbb{C}(S) := \big\{\theta \in \mathbb{R}^d \mid \|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1\big\}.$$

Here $\|\theta_{S^c}\|_1 := \sum_{j \in S^c} |\theta_j|$ corresponds to the ℓ_1-norm of the coefficients indexed by S^c, with $\|\theta_S\|_1$ defined similarly. Note that any vector θ^* supported on S belongs to the cone $\mathbb{C}(S)$; in addition, it includes vectors whose ℓ_1-norm on the "bad" set S^c is small relative to their ℓ_1-norm on S. Given triplet (n, d, k),

the matrix $X \in \mathbb{R}^{n \times d}$ is said to satisfy a γ-RE condition (also known as a compatibility condition) if

$$\frac{1}{n}\|X\theta\|_2^2 \geq \gamma \|\theta\|_2^2 \qquad \text{for all } \theta \in \bigcup_{|S|=k} \mathbb{C}(S). \tag{2.3}$$

The following result [4, 29, 6] provides a bound on the prediction error for the Lasso estimator:

Proposition 1 (Prediction error for Lasso with RE condition)**.** *Consider the standard linear model for a design matrix X satisfying the column normalization condition (2.2) and the γ-RE condition. Then, for any vector $\theta^* \in \mathbb{B}_0(k)$ and $\delta \in (0, 1]$, the Lasso estimator $\widehat{\theta}_{\lambda_n}$ with $\lambda_n = 4\sigma\sqrt{\frac{2\log d}{n}} + \delta$ satisfies*

$$\frac{1}{n}\|X\widehat{\theta}_{\lambda_n} - X\theta^*\|_2^2 \leq \frac{c\,k}{\gamma^2}\left\{\frac{\sigma^2 \log d}{n} + \delta^2\right\} \qquad \text{for any } \theta^* \in \mathbb{B}_0(k), \tag{2.4}$$

with probability at least $1 - c_1 d e^{-c_2 n \delta^2}$.

The Lasso rate (2.4) will match the optimal rate (1.4) if the RE constant γ is bounded away from zero. If γ is close to zero, then the Lasso rate could be arbitrarily worse than the optimal rate. It is known that the RE condition is necessary for recovering the true vector θ^* [see, e.g., 33], but minimizing the prediction error should be easier than recovering the true vector. In particular, strong correlations between the columns of X, which lead to violations of the RE conditions, should have no effect on the intrinsic difficulty of the prediction problem. Recall that the ℓ_0-based estimator $\widehat{\theta}_{\ell_0}$ satisfies the prediction error upper bound (1.4) without any constraint on the design matrix. Moreover, Raskutti et al. [33] show that many problems with strongly correlated columns are actually easy from the prediction point of view.

In the absence of RE conditions, ℓ_1-based methods are known to achieve the slow $1/\sqrt{n}$ rate, with the only constraint on the design matrix being a uniform column bound [4]:

Proposition 2 (Prediction error for Lasso without RE condition)**.** *Consider the standard linear model for a design matrix X satisfying the column normalization condition (2.2). Then for any vector $\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_1(R)$ and $\delta \in (0, 1]$, the Lasso estimator $\widehat{\theta}_{\lambda_n}$ with $\lambda_n = 4\sigma\sqrt{\frac{2\log d}{n}} + \delta$ satisfies the bound*

$$\frac{1}{n}\|X(\widehat{\theta}_{\lambda_n} - \theta^*)\|_2^2 \leq c\, R\Big(\sigma\sqrt{\frac{2\log d}{n}} + \delta\Big), \tag{2.5}$$

with probability at least $1 - c_1 d\, e^{-c_2 n \delta^2}$.

Combining the bounds of Proposition 1 and Proposition 2, we have

$$\mathcal{M}_{n,k,d}(\widehat{\theta}_{\ell_1}; X) \leq c' \min\left\{\frac{\sigma^2 k \log d}{\gamma^2 n}, \sigma R\sqrt{\frac{\log d}{n}}\right\}. \tag{2.6}$$

If the RE constant γ is sufficiently close to zero, then the second term on the right-hand side will dominate the first term. In that case, the $1/\sqrt{n}$ achievable rate is substantially slower than the $1/n$ optimal rate for reasonable ranges of (k, R). One might wonder whether the analysis leading to the bound (2.6) could be sharpened so as to obtain the fast rate. Among other consequences, our first main result (Theorem 1 below) shows that no substantial sharpening is possible.

3. Main results

We now turn to statements of our main results, and discussion of their consequences.

3.1. Lower bound for a particular family of design matrices

In this section, we show that there exists a particular family of "bad" design matrices, in which the $1/\sqrt{n}$ rate is unavoidable for any regularized M-estimator. Our analysis applies to the set of local minimas of the objective function L defined in equation (2.1). More precisely, a vector $\widetilde{\theta}$ is a local minimum of the function $\theta \mapsto L(\theta; \lambda)$ if there is an open ball $\mathbb{B}$ centered at $\widetilde{\theta}$ such that $\widetilde{\theta} \in \arg\min_{\theta \in \mathbb{B}} L(\theta; \lambda)$. We then define the set

$$\widehat{\Theta}_\lambda := \Big\{ \theta \in \mathbb{R}^d \mid \theta \text{ is a local minimum of the function } \theta \mapsto L(\theta; \lambda) \Big\}, \quad (3.1)$$

an object that depends on the triplet (X, y, ρ) as well as the choice of regularization weight λ. The set $\widehat{\Theta}_\lambda$ can contain multiple elements, and it may contain both global and local minima.

At best, a typical descent method applied to the objective L can be guaranteed to converge to some element of $\widehat{\Theta}_\lambda$. The following theorem provides a lower bound, applicable to any method that always returns some local minimum of the objective function (2.1).

Theorem 1. *For any pair (n, d) such that $d \geq n \geq \max\{4, (\frac{16\sigma}{R})^2 \log(n), (\frac{R}{\sigma})^4\}$ and any sparsity level $k \geq 2$, there is a design matrix $X \in \mathbb{R}^{n \times d}$ satisfying the column normalization condition (2.2) such that for any coordinate-separable penalty, we have*

$$\sup_{\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_1(R)} \mathbb{E}\left[\inf_{\lambda \geq 0} \sup_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n} \|X(\theta - \theta^*)\|_2^2 \right] \geq c\sigma R \sqrt{\frac{\log n}{n}}. \quad (3.2a)$$

Moreover, for any convex coordinate-separable penalty, we have

$$\sup_{\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_1(R)} \mathbb{E}\left[\inf_{\lambda \geq 0} \inf_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n} \|X(\theta - \theta^*)\|_2^2 \right] \geq c\sigma R \sqrt{\frac{\log n}{n}}. \quad (3.2b)$$

In both of these statements, the constant c is universal, independent of (n, d, k, σ, R) as well as the design matrix. See Section 4.1 for the proof.

In order to interpret the lower bound (3.2a), consider any estimator $\widehat{\theta}$ that takes values in the set $\widehat{\Theta}_\lambda$, corresponding to local minima of L. The result is of a game-theoretic flavor: the statistician is allowed to adaptively choose λ based on the observations (y, X), whereas nature is allowed to act adversarially in choosing a local minimum for every execution of $\widehat{\theta}_\lambda$. Under this setting, Theorem 1 implies that

$$\sup_{\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_1(R)} \frac{1}{n} \mathbb{E}\left[\|X\widehat{\theta}_\lambda - X\theta^*\|_2^2 \right] \geq c\, \sigma R \sqrt{\frac{\log n}{n}}. \tag{3.3}$$

For any convex regularizer (such as the ℓ_1-penalty underlying the Lasso estimate), equation (3.2b) provides a stronger lower bound, one that holds uniformly over all choices of $\lambda \geq 0$ and all (global) minima. For the Lasso estimator, the lower bound of Theorem 1 matches the upper bound (2.5) up to the logarithmic term $(\frac{\log d}{\log n})^{1/2}$. Thus, our lower bound is tight as long as the dimension d is bounded by a constant-degree polynomial function of n. Closing the gap for problems of super-polynomial dimensions remains an open problem.

3.2. Lower bound for design matrices under RE conditions

One potential concern with Theorem 1 is that lower bound might apply only to extremely ill-conditioned design matrices, even with sparsity constraints; such matrices might not be as likely to arise in practice. As noted earlier, control of restricted eigenvalues provides guarantees on the "sparse condition number" of the design matrix, and such control plays an important role in the theory of sparse estimation. Accordingly, it is natural to wonder whether it is also possible to prove a non-trivial lower bound when the restricted eigenvalues are bounded above zero. Recall that under the RE condition with a positive constant γ, the Lasso will achieve the rate (2.6), as defined by the minimum of a scaled fast rate $1/(\gamma^2 n)$ and the familiar slow rate $1/\sqrt{n}$. The following result shows that the Lasso rate cannot be improved to match the fast rate.

Corollary 1. *For any sparsity level $k \geq 2$, any constant $\gamma \in (0, 1]$ and any pair (n, d) such that $d = n \geq \max\{2k^2, k(\frac{16\sigma}{R})^2 \log(n), k(\frac{R}{\sigma})^4\}$, there is a design matrix $X \in \mathbb{R}^{n \times d}$ satisfying the column normalization condition (2.2) and the γ-RE condition, such that for any coordinate-separable penalty, we have*

$$\sup_{\theta^* \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(R)} \mathbb{E}\left[\inf_{\lambda \geq 0} \sup_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n} \|X(\theta - \theta^*)\|_2^2 \right] \geq c \min\left\{ \frac{\sigma^2 k \log n}{\gamma n}, \sigma R \sqrt{\frac{\log n}{n}} \right\}. \tag{3.4a}$$

Moreover, for any convex coordinate-separable penalty, we have

$$\sup_{\theta^* \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(R)} \mathbb{E}\left[\inf_{\lambda \geq 0} \inf_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n} \|X(\theta - \theta^*)\|_2^2 \right] \geq c \min\left\{ \frac{\sigma^2 k \log n}{\gamma n}, \sigma R \sqrt{\frac{\log n}{n}} \right\}. \tag{3.4b}$$

Since none of the terms on the right-hand side of inequalities (3.4a) and (3.4b) match the optimal rate (1.4), the corollary implies that the optimal rate is not achievable even if the restricted eigenvalues are bounded above zero. Comparing this lower bound to the Lasso upper bound (2.6), we observe that the upper bound has a term that is proportional to $1/\gamma^2$, but the corresponding term in the lower bound is proportional to $1/\gamma$. Proving a sharper lower bound that closes this gap remains an open problem.

We remark that Corollary 1 follows by a refinement of the proof of Theorem 1. In particular, we first show that the design matrix underlying Theorem 1—call it X_{bad}—satisfies the γ_n-RE condition, where the quantity γ_n converges to zero as a function of sample size n. In order to prove Corollary 1, we construct a new block-diagonal design matrix such that each block corresponds to a version of X_{bad}. The size of these blocks are then chosen so that, given a predefined quantity $\gamma > 0$, the new matrix satisfies the γ-RE condition. We then lower bound the prediction error of this new matrix, using Theorem 1 to lower bound the prediction error of each of the blocks. We refer the reader to Appendix C for the full proof.

3.3. Lower bound for general design matrices

Our proof of Theorem 1 is based on a particular construction of "bad" design matrices, but at a deeper level, there is actually a fairly broad class of matrices that also lead to the slow rate. In particular, in this section, we describe a set of general conditions on design matrices that are sufficient to ensure that the slow rate is unavoidable by any regularized M-estimator. We then show that the family of "bad" design matrices from the previous sections can be understood as a special case. We further show that our theory also encompasses certain ensembles of random matrices, for instance, those with rows sampled in an i.i.d. manner from a multivariate normal distribution with a suitable covariance matrix.

Before presenting the main result of this subsection, let us introduce some shorthand notation that is useful. For an arbitrary subset of integers $J \subseteq \{1, \ldots, d\}$ and arbitrary matrix A (or vector v), we denote by A_J the sub-matrix with column indices in J (or by v_J the sub-vector with coordinate indices in J). We use A_{-J} (or v_{-J}) as shorthand notation for the sub-matrix (or the sub-vector) whose column indices (or coordinate indices) are not in J. Similarly, we denote by $\rho_J(\theta) := \sum_{j \in J} \rho_j(\theta_j)$ and $\rho_{-J}(\theta) := \sum_{j \notin J} \rho_j(\theta_j)$ the penalty function for coordinates in and outside of the set J. We use the standard order notation $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$, where we suppress all constants that do not depend on the triple (n, d, k).

Given this notation, we are ready to describe our sufficient conditions on the design matrix:

Assumption A (Sufficient conditions for "badness")**.**

- *First, the matrix $X \in \mathbb{R}^{n \times d}$ satisfies the column normalization condition (2.2), and has rank at least $r = \Omega(n^{1/2})$.*

- *Second, there are integers $s = \mathcal{O}(1)$ and $m = \Omega(r)$ as well as disjoint sets $J_1, \ldots, J_m \subseteq [d]$, each of cardinality s, such that:*
  - *(a) For each subset J in the collection $\{J_1, \ldots, J_m\}$, there is a vector $u \in \mathbb{R}^s$ such that $\|u\|_1 = 1$ and $\frac{1}{n}\|X_J^T X_J u\|_2 = \mathcal{O}(n^{-1/2})$.*
  - *(b) For each $J \in \{J_1, \ldots, J_m\}$, any unit vector v in the column space of X_J and any unit vector v' in the column space of X_{-J}, we have $|\langle v, v' \rangle| \leq \mathcal{O}(r^{-1/2})$.*
  - *(c) For each index $j \in \cup_{i=1}^m J_i$, we have $\frac{1}{\sqrt{n}}\|X_j\|_2 = \Omega(1)$. Moreover, the singular values of matrix $\frac{1}{\sqrt{n}} X_{(\cup_{i=1}^m J_i)}$ are lower bounded as $\Omega(n^{-1/4})$ and upper bounded as $\mathcal{O}(1)$.*

It is worthwhile making a few comments on the meaning of these different requirements. First, condition (a) implies there is an s-dimensional vector u such that the squared norm $\frac{1}{n}\|X_J u\|_2^2$ is bounded by $\gamma\|u\|_2^2$ where $\gamma = \mathcal{O}(n^{-1/2})$. Comparing with inequality (2.3), we observe that the Restricted Eigenvalue (RE) condition is violated. Since the RE condition is sufficient for the fast rate, it is obvious that condition (a) is necessary for the slow rate. On the other hand, condition (b) implies that the column space of X_J is roughly orthogonal to the space spanned by other columns in the design matrix. The consequence of this assumption is that the original d-dimensional regression problem contains m weakly correlated sub-problems of dimension s, so that each of them can be separately studied. Condition (c) implies that the singular values of the joint design matrix of these sub-problems—namely, the singular values of $\frac{1}{\sqrt{n}} X_{(\cup_{i=1}^m J_i)}$—are neither too large nor too small.

We present a general theorem before giving concrete examples. In order to simplify the statement, we assume the noise variance $\sigma^2 = 1$, and that the true vector θ^* is bounded by $\|\theta^*\|_1 \leq 1$. The theorem holds if these two quantities are assigned by any other constants.

Theorem 2. *Consider any design matrix X that satisfies the conditions in Assumption A. For any sufficiently large sample size $n \geq c_1$, any sparsity level $k \geq s$, and for any coordinate-separable penalty, we have*

$$\sup_{\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_1(1)} \mathbb{E}\left[\inf_{\lambda \geq 0} \sup_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n}\|X(\theta - \theta^*)\|_2^2\right] \geq c_2 n^{-1/2}, \tag{3.5}$$

where $c_1, c_2 > 0$ denote constants independent of (n, d, k).

Relative to our earlier results, the proof of Theorem 2 is significantly more challenging, because we have no explicit control over the design matrix. In particular, if a matrix X satisfies Assumption A, then the assumption will remain valid for any small perturbation of the matrix. This property enables the lower bound to capture random design matrices, but adds challenges to the proof. See Section 4.2 for the proof of the theorem.

Example 1 As a concrete example, let us demonstrate that the fixed design matrix defined in the proof of Theorem 1 satisfies Assumption A. As detailed

in Section 4.1, the proof of this result is based on constructing a block-diagonal design matrix $X \in \mathbb{R}^{n \times d}$, given by

$$X = \Big[\text{blkdiag}\underbrace{\big\{\sqrt{n}A, \sqrt{n}A, \ldots, \sqrt{n}A\big\}}_{n/2 \text{ copies}} \; \mathbf{0}\Big] \in \mathbb{R}^{n \times d},$$

where the sub-matrix A takes the form

$$A = \begin{bmatrix} \cos(\alpha) & -\cos(\alpha) \\ \sin(\alpha) & \sin(\alpha) \end{bmatrix}, \quad \text{with} \quad \alpha := \arcsin(n^{-1/4}).$$

It is straightforward to verify that the matrix X satisfies the column normalization condition ((2.2)) and has rank n. It remains to verify the conditions (a), (b) and (c) of Assumption A. Choosing $s = 2$ and $J_i = \{2i-1, 2i\}$ for $i = 1, 2, \ldots, n/2$, then condition (a) holds with vector $u := (1/2, 1/2)$. Condition (b) is satisfied because the column spaces of X_{J_i} and X_{-J_i} are orthogonal spaces for every $i \in [n/2]$. Condition (c) is satisfied because the matrix $n^{-1/2} X_{(\cup_{i=1}^m J_i)}$ is a block-diagonal matrix consisting of sub-matrices A, and each submatrix's singular values are lower bounded by $\Omega(n^{-1/4})$ and upper bounded by 1.

Example 2 As a second illustration, consider a random design $X \in \mathbb{R}^{n \times d}$ with rows drawn in an i.i.d. manner from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$ where the covariance matrix is defined by:

$$\Sigma = \text{diag}\underbrace{(A, A, \ldots, A)}_{d/2 \text{ copies}} \quad \text{where} \quad A := \begin{bmatrix} 1/2 & 1/2 - n^{-1/2} \\ 1/2 - n^{-1/2} & 1/2 \end{bmatrix}. \tag{3.6}$$

For such correlated Gaussian designs, sufficient conditions for achieving the optimal rate (1.4) have been extensively studied (e.g., [26, 32]). However, it has remained unknown if the optimal rate can be achieved for random design matrices drawn from general covariance matrices Σ. Here we provide a negative answer to this question by showing that if the covariance matrix is unfavorably chosen, then any regularized M-estimator can (at best) achieve the slow rate.

Proposition 3. *For $d = \sqrt{n}$, consider a random design matrix $X \in \mathbb{R}^{n \times d}$ with each row sampled in an i.i.d. manner from the multivariate normal $\mathcal{N}(0, \Sigma)$ with the covariance Σ from equation (3.6). Then with probability at least $1 - e^{-\Omega(\sqrt{n})}$, the matrix X satisfies Assumption A with sparsity level $s = 2$.*

See Appendix D for the proof.

3.4. Lower bound for local descent methods

For any least-squares cost with a coordinate-wise separable regularizer, Theorem 1 establishes the existence of at least one "bad" local minimum such that

the associated prediction error is lower bounded by $1/\sqrt{n}$. One might argue that this result could be overly pessimistic, in that the adversary is given too much power in choosing local minima. Indeed, the mere existence of bad local minima need not be a practical concern unless it can be shown that a typical optimization algorithm will frequently converge to one of them.

Steepest descent is a standard first-order algorithm for minimizing a convex cost function [3, 5]. However, for nonconvex and non-differentiable loss functions, it is known that the steepest descent method does not necessarily yield convergence to a local minimum [14, 40]. Although there exist provably convergent first-order methods for general nonsmooth optimization (e.g., [27, 23]), the paths defined by their iterations are difficult to characterize, and it is also difficult to predict the point to which such an algorithm eventually converges.

In order to address a broad class of methods in a unified manner, we begin by observing that most first-order methods can be seen as iteratively and approximately solving a local minimization problem. For example, given a stepsize parameter $\eta > 0$, the method of steepest descent iteratively approximates the minimizer of the objective over a ball of radius η. Similarly, the convergence of algorithms for nonconvex optimization is based on the fact that they guarantee decrease of the function value in the local neighborhood of the current iterate [27, 23]. We thus study an iterative local descent algorithm taking the form:

$$\theta^{t+1} \in \arg \min_{\theta \in \mathbb{B}_2(\eta; \theta^t)} L(\theta; \lambda), \tag{3.7}$$

where $\eta > 0$ is a given radius, and $\mathbb{B}_2(\eta; \theta^t) := \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^t\|_2 \leq \eta\}$ is the ball of radius η around the current iterate. If there are multiple points achieving the optimum, the algorithm chooses the one that is closest to θ^t, resolving any remaining ties by randomization. The algorithm terminates when there is a minimizer belonging to the interior of the ball $\mathbb{B}_2(\eta; \theta^t)$—that is, exactly when θ^{t+1} is a local minimum of the loss function.

It should be noted that the algorithm (3.7) defines a powerful algorithm—one that might not be easy to implement in polynomial time—since it is guaranteed to return the global minimum of a nonconvex program over the ball $\mathbb{B}_2(\eta; \theta^t)$. In a certain sense, it is more powerful than any first-order optimization method, since it will always decrease the function value at least as much as a descent step with stepsize related to η. Since we are proving lower bounds, these observations only strengthen our result. We impose two additional conditions on the regularizers:

(iv) Each component function ρ_j is continuous at the origin.
(v) There is a constant H such that $|\rho_j'(x) - \rho_j'(\tilde{x})| \leq H|x - \tilde{x}|$ for any pair $x, \tilde{x} \in (0, \infty)$.

Assumptions (i)-(v) are more restrictive than assumptions (i)-(iii), but they are satisfied by many popular penalties. As illustrative examples, for the ℓ_1-norm, we have $H = 0$. For the SCAD penalty, we have $H = 1/(a-1)$, whereas for the MCP regularizer, we have $H = 1/b$. Finally, in order to prevent the

update (3.7) being so powerful that it reaches the global minimum in one single step, we impose an additional condition on the stepsize, namely that

$$\eta \leq \min\left\{B, \frac{B}{\lambda H}\right\}, \quad \text{where } B \asymp \tfrac{\sqrt{\log n}\sigma}{\sqrt{n}}. \tag{3.8}$$

It is reasonable to assume that the stepsize bounded by a time-invariant constant, as we can always partition a single-step update into a finite number of smaller steps, increasing the algorithm's time complexity by a multiplicative constant. On the other hand, the $\mathcal{O}(1/\sqrt{n})$ stepsize is achieved by popular first-order methods. Under these assumptions, we have the following theorem, which applies to any regularizer ρ that satisfies Assumptions (i)-(v).

Theorem 3. *For any pair (n, d) such that $d \geq n \geq \max\{4, (\frac{4\sigma}{R})^2 \log(n), (\frac{R}{\sigma})^4\}$, and any sparsity level $k \geq 2$, there is a design matrix $X \in \mathbb{R}^{n\times d}$ satisfying the column normalization condition (2.2) such that*

(a) The update (3.7) terminates after a finite number of steps T at a vector $\widehat{\theta} = \theta^{T+1}$ that is a local minimum of the loss function.

(b) Given a random initialization $\theta^0 \sim N(0, \gamma^2 I_{d\times d})$, the local minimum satisfies the lower bound

$$\sup_{\theta^* \in \mathbb{B}_0(k)\cap\mathbb{B}_1(R)} \mathbb{E}\left[\inf_{\lambda\geq 0} \frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2\right] \geq c\,\sigma R\sqrt{\frac{\log n}{n}}.$$

Part (a) shows that the local descent method always returns a local minimizer. For convex loss functions, any local minimum is a global minimum, thus the algorithm is able to exactly compute the Lasso estimator. As a consequence, the prediction error established by Proposition 2, namely $\sigma R\sqrt{\log(d)/n}$, provides an upper bound on the prediction error for the family of local descent algorithms. It matches the lower bound in Part (b) up to a constant factor whenever $\frac{\log d}{\log n} = \mathcal{O}(1)$.

Theorem 3 shows that local descent methods based on a random initialization do not lead to local optima that achieve the fast rate. This conclusion provides stronger negative evidence than Theorem 1, since it shows that bad local minima not only exist, but are difficult to avoid.

3.5. Simulations

In the proof of Theorem 1 and Theorem 3, we construct specific design matrices to make the problem hard to solve. In this section, we apply several popular algorithms to the solution of the sparse linear regression problem on these "hard" examples, and compare their performance with the ℓ_0-based estimator (1.3). More specifically, focusing on the special case $n = d$, we perform simulations for the design matrix $X \in \mathbb{R}^{n\times n}$ used in the proof of Theorem 3. It is given by

$$X = \Big[\text{blkdiag}\underbrace{\{\sqrt{n}A, \sqrt{n}A, \ldots, \sqrt{n}A\}}_{n/2 \text{ copies}}\Big],$$

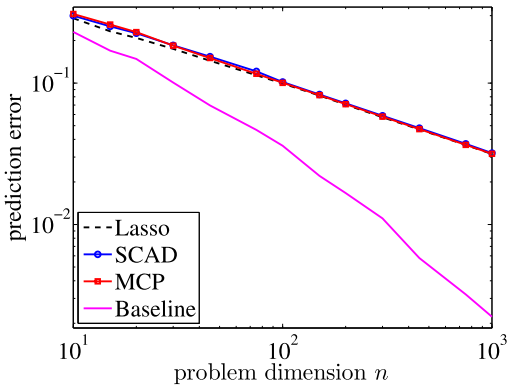

FIG 2. *Problem scale n versus the prediction error $\mathbb{E}[\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2]$. The expectation is computed by averaging 100 independent runs of the algorithm. Both the sample size n and the prediction error are plotted on a logarithmic scale.*

where the sub-matrix A takes the form

$$A = \begin{bmatrix} \cos(\alpha) & -\cos(\alpha) \\ \sin(\alpha) & \sin(\alpha) \end{bmatrix}, \quad \text{where} \quad \alpha = \arcsin(n^{-1/4}).$$

Given the 2-sparse regression vector $\theta^* = \big(0.5, 0.5, 0, \ldots, 0\big)$, we form the response vector $y = X\theta^* + w$, where $w \sim N(0, I_{n\times n})$.

We compare the ℓ_0-based estimator, referred to as the *baseline estimator*, with three other methods: the Lasso estimator [37], the estimator based on the SCAD penalty [16] and the estimator based on the MCP penalty [41]. In implementing the ℓ_0-based estimator, we provide it with the knowledge that $k = 2$, since the true vector θ^* is 2-sparse. For Lasso, we adopt the `MATLAB` implementation [1], which generates a Lasso solution path evaluated at 100 different regularization parameters, and we choose the estimate that yields the smallest prediction error. For the SCAD penalty, we choose $a = 3.7$ as suggested by Fan and Li [16]. For the MCP penalty, we choose $b = 2.7$, so that the maximum concavity of the MCP penalty matches that of the SCAD penalty. For the SCAD penalty and the MCP penalty (and recalling that $d = n$), we studied choices of the regularization weight of the form $\lambda = C\sqrt{\frac{\log n}{n}}$ for a pre-factor C to be determined. As shown in past work on nonconvex regularizers [25], such choices of λ lead to low ℓ_2-error. By manually tuning the parameter C to optimize the prediction error, we found that $C = 0.1$ is a reasonable choice. We used routines from the GIST package [20] to optimize these nonconvex objectives.

By varying the sample size over the range 10 to 1000, we obtained the results plotted in Figure 2, in which the prediction error $\mathbb{E}[\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2]$ and sample size n are both plotted on a logarithmic scale. The performance of the Lasso, SCAD-based estimate, and MCP-based estimate are all similar. For all of the three methods, the prediction error scales as $1/\sqrt{n}$, as confirmed by the slopes of the corresponding lines in Figure 2, which are very close to 0.5. In fact,

by examining the estimator's output, we find that in many cases, all three estimators output $\widehat{\theta} = 0$, leading to the prediction error $\frac{1}{n}\|X(0-\theta^*)\|_2^2 = \frac{1}{\sqrt{n}}$. Since the regularization parameters have been chosen to optimize the prediction error, this scaling is the best rate that the three estimators are able to achieve, and it matches the theoretical prediction of Theorem 1 and Theorem 3.

In contrast, the ℓ_0-based estimator achieves a substantially better error rate. The slope of the corresponding line in Figure 2 is very close to 1. It means that the prediction error of the ℓ_0-based estimator scales as $1/n$, thereby matching the theoretically-predicted scaling (1.4).

4. Proofs

We now turn to the proofs of our theorems. In each case, we defer the proofs of more technical results to the appendices.

4.1. Proof of Theorem 1

For a given triplet (n, σ, R), we define an angle $\alpha := \arcsin\left(\frac{\sqrt{s_n}\sigma}{n^{1/4}\sqrt{32R}}\right)$, where $s_n \in [1, \sqrt{\log n}]$ is a scaling factor (depending on n) to be specified later. Then we define a two-by-two matrix

$$A = \begin{bmatrix} \cos(\alpha) & -\cos(\alpha) \\ \sin(\alpha) & \sin(\alpha) \end{bmatrix}. \tag{4.1a}$$

Using the matrix $A \in \mathbb{R}^{2\times 2}$ as a building block, we construct a design matrix $X \in \mathbb{R}^{n\times d}$. Without loss of generality, we may assume that n is divisible by two. (If n is not divisible by two, constructing a $(n-1)$-by-d design matrix concatenated by a row of zeros only changes the result by a constant.) We then define the matrix

$$X = \Big[\text{blkdiag}\underbrace{\big\{\sqrt{n}A, \sqrt{n}A, \ldots, \sqrt{n}A\big\}}_{n/2 \text{ copies}} \quad \mathbf{0}\Big] \in \mathbb{R}^{n\times d}, \tag{4.1b}$$

where the all-zeroes matrix on the right side has dimensions $n \times (d-n)$. It is easy to verify that the matrix X defined in this way satisfies the column normalization condition (2.2).

Next, we prove the lower bound (3.2a). For any integers $i, j \in [d]$ with $i < j$, let θ_i denote the i^{th} coordinate of θ, and let $\theta_{i:j}$ denote the subvector with entries $\{\theta_i, \ldots, \theta_j\}$. Since the matrix A appears in diagonal blocks of X, we have

$$\inf_{\lambda\geq 0}\sup_{\theta\in\widehat{\Theta}_\lambda}\frac{1}{n}\|X(\theta-\theta^*)\|_2^2 = \inf_{\lambda\geq 0}\sup_{\theta\in\widehat{\Theta}_\lambda}\sum_{i=1}^{n/2}\|A\big(\theta_{(2i-1):2i} - \theta^*_{(2i-1):2i}\big)\|_2^2, \tag{4.2}$$

and it suffices to lower bound the right-hand side of the above equation.

For the sake of simplicity, we introduce the shorthand $B := \frac{4s_n\sigma}{\sqrt{n}}$, and define the scalars

$$\gamma_i = \min\{\rho_{2i-1}(B), \rho_{2i}(B)\} \qquad \text{for each } i = 1, \ldots, n/2.$$

Furthermore, we define

$$a_i := \begin{cases} (\cos\alpha, \sin\alpha) & \text{if } \gamma_i = \rho_{2i-1}(B) \\ (-\cos\alpha, \sin\alpha) & \text{if } \gamma_i = \rho_{2i}(B) \end{cases} \qquad \text{and} \qquad w_i' := \frac{\langle a_i, w_{(2i-1):2i}\rangle}{\sqrt{n}}. \tag{4.3a}$$

Without loss of generality, we may assume that $\gamma_1 = \max_{i \in [n/2]} \{\gamma_i\}$ and $\gamma_i = \rho_{2i-1}(B)$ for all $i \in [n/2]$. If this condition does not hold, we can simply re-index the columns of X to make these properties hold. Note that when we swap the columns $2i-1$ and $2i$, the value of a_i doesn't change; it is always associated with the column whose regularization term is equal to γ_i.

Finally, we define the regression vector $\theta^* = \begin{bmatrix} \frac{R}{2} & \frac{R}{2} & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^d$. Given these definitions, the following lemma lower bounds each term on the right-hand side of equation (4.2).

Lemma 1. *For any $\lambda \geq 0$, there is a local minimum $\widehat{\theta}_\lambda$ of the objective function $L(\theta; \lambda)$ such that $\frac{1}{n}\|X(\widehat{\theta}_\lambda - \theta^*)\|_2^2 \geq T_1 + T_2$, where*

$$T_1 := \mathbb{I}\Big[\lambda\gamma_1 > 4B(\sin^2(\alpha)R + \frac{\|w_{1:2}\|_2}{\sqrt{n}})\Big] \sin^2(\alpha)(R - 2B)_+^2 \quad \textit{and} \tag{4.4a}$$

$$T_2 := \sum_{i=2}^{n/2} \mathbb{I}\Big[B/2 \leq w_i' \leq B\Big]\Big(\frac{B^2}{4} - \lambda\gamma_1\Big)_+. \tag{4.4b}$$

Moreover, if the regularizer ρ is convex, then every minimizer $\widehat{\theta}_\lambda$ satisfies this lower bound.

See Appendix B for the proof of this claim.

Using Lemma 1, we can now complete the proof of the theorem. It is convenient to condition on the event $\mathcal{E} := \{\|w_{1:2}\|_2 \leq \frac{\sigma}{32}\}$. Since $\|w_{1:2}\|_2^2/\sigma^2$ follows a chi-square distribution with two degrees of freedom, we have $\mathbb{P}[\mathcal{E}] > 0$. Conditioned on this event, we now consider two separate cases:

Case 1: First, suppose that $\lambda\gamma_1 > (s_n\sigma)^2/n$. In this case, we have

$$4B\Big\{\sin^2(\alpha)R + \frac{\|w_{1:2}\|_2}{\sqrt{n}}\Big\} \leq \frac{16 s_n\sigma}{\sqrt{n}}\left(\frac{s_n\sigma}{32\sqrt{n}} + \frac{\sigma}{32\sqrt{n}}\right) = \frac{(s_n\sigma)^2}{n} < \lambda\gamma_1,$$

and consequently

$$T_1 + T_2 \geq T_1 = \sin^2(\alpha)(R - 2B)_+^2 = \frac{s_n\sigma}{32\sqrt{n}R}\left(R - \frac{8s_n\sigma}{\sqrt{n}}\right)_+^2 \geq \frac{s_n\sigma R}{128\sqrt{n}}. \tag{4.5a}$$

The last inequality holds because we assumed $n \geq (16\sigma/R)^2\log(n)$, and as a consequence, the radius R is lower bounded by $R/2 \geq 8\sigma\sqrt{\log(n)/n} \geq 8s_n\sigma/\sqrt{n}$.

Case 2: Otherwise, we may assume that $\lambda\gamma_1 \leq (s_n\sigma)^2/n$. In this case, we have

$$T_1 + T_2 \geq T_2 = \sum_{i=2}^{n/2} \mathbb{I}\left(B/2 \leq w_i' \leq B\right) \frac{3(s_n\sigma)^2}{n}. \tag{4.5b}$$

Combining the two lower bounds (4.5a) and (4.5b), we find

$$\begin{aligned} &\mathbb{E}\left[\inf_{\lambda\geq 0} \sup_{\theta\in\widehat{\Theta}_\lambda} \frac{1}{n}\|X\theta - X\theta^*\|_2^2 \Big| \mathcal{E}\right] \\ &\qquad \geq \underbrace{\mathbb{E}\left[\min\left\{\frac{s_n\sigma R}{128\sqrt{n}}, \sum_{i=2}^{n/2} \mathbb{I}\left[B/2 \leq w_i' \leq B/2\right] \frac{3(s_n\sigma)^2}{n}\right\}\right]}_{T_3}, \end{aligned} \tag{4.6}$$

where we have used the fact that $\{w_i'\}_{i=2}^{n/2}$ are independent of the event $\mathcal{E}$. To lower bound the right-hand side, we partition the integer set $\{1, 2, \ldots, n/2\}$ into $m := \lfloor n^{3/4} \rfloor$ disjoint subsets, such that each subset contains at least $\lfloor n/(2m) \rfloor$ integers. Let these subsets be called $S_1, \ldots, S_m$. Using the inequality $\min\left\{a, \sum_{i=2}^{n/2} b_i\right\} \geq \sum_{j=1}^{m} \min\{\frac{|S_j|}{n/2}a, \sum_{i\in S_j} b_i\}$, valid for scalars a and $\{b_i\}_{i=2}^{n/2}$, we see that

$$T_3 \geq \sum_{j=1}^{m} \mathbb{P}\left[\sum_{i\in S_j} \mathbb{I}\Big[\frac{2s_n\sigma}{\sqrt{n}} \leq w_i' \leq \frac{4s_n\sigma}{\sqrt{n}}\Big] \geq 1\right] \min\left\{\frac{\lfloor n/(2m)\rfloor}{n/2} \cdot \frac{s_n\sigma R}{128\sqrt{n}}, \frac{3(s_n\sigma)^2}{n}\right\},$$

where we have used the definition $B := \frac{4s_n\sigma}{\sqrt{n}}$.

Since $w_i' \sim N(0, \sigma^2/n)$, we have

$$\mathbb{P}[2s_n\sigma/\sqrt{n} \leq w_i' \leq 4s_n\sigma/\sqrt{n}] = \Phi(-2s_n) - \Phi(-4s_n),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. For any $t > 0$, the function Φ is sandwiched [15] as

$$\frac{(t^{-1} - t^{-3})e^{-t^2}}{\sqrt{2\pi}} \leq \Phi(-t) \leq \frac{t^{-1}e^{-t^2}}{\sqrt{2\pi}}.$$

By choosing $s_n = \max\{1, c_1\sqrt{\log n}\}$ with a sufficiently small universal constant c_1, we guarantee that $\Phi(-2s_n) - \Phi(-4s_n) \geq c_2 n^{-1/4}$ for a universal constant $c_2 > 0$. Since $|S_j| = \Omega(n^{1/4})$, there is a universal constant $c_3 > 0$ such that $\mathbb{P}[\sum_{i\in S_j} \mathbb{I}[\frac{2s_n\sigma}{\sqrt{n}} \leq w_i' \leq \frac{4s_n\sigma}{\sqrt{n}}] \geq 1] \geq c_3$. Putting together the pieces, we have shown that

$$\mathbb{E}\left[\inf_{\lambda\geq 0} \sup_{\theta\in\widehat{\Theta}_\lambda} \frac{1}{n}\|X\theta - X\theta^*\|_2^2\right] \geq \mathbb{P}[\mathcal{E}]\, T_3 \geq c \min\left\{\sigma R\sqrt{\frac{\log(n)}{n}}, \frac{\log(n)\sigma^2}{n^{1/4}}\right\}.$$

The assumption $n \geq (R/\sigma)^4$ implies that the first-term on the right-hand side is smaller than the second term. Hence we obtain the desired lower bound.

4.2. Proof of Theorem 2

The proof involves four positive constants, denoted by C_1, C_2, C_3 and C_4, whose precise values are specified in later parts of the proof. The constant $C_1 \in (0, 1]$ is used to define the quantity $B := C_1 \cdot n^{-1/2}$, which specifies a cut-off point used in the argument. More precisely, for each index set $J_i \in \{J_1, \ldots, J_m\}$, we let

$$\gamma_i := \min_{j \in J_i} \{\rho_j(B)\}$$

be the minimum value of $\rho_j(B)$ within the index set. In other words, the penalty at point B for coordinates J_i is lower bounded by γ_i. As a shorthand notation, let i^* be the index such that $\gamma_{i^*} = \max_{i \in [m]} \{\gamma_i\}$; that is, the index where the lower bounds reach the maximal value. Throughout this proof, we write the index set J_{i^*} and the quantity γ_{i^*} simply by J and γ. As additional shorthand notations, let $\Pi(\cdot)$ represent the orthogonal operator that projects onto the column space of matrix X, and let $\Pi_J(\cdot)$ represent the orthogonal operator that projects onto the column space of matrix X_J.

For every index $i \neq i^*$, by definition we can find an index j from the index set J_i such that $\rho_j(B) = \gamma_i$. The set of these $m - 1$ indices is denoted by I. The largest singular value of matrix X_I is upper bounded by that of the matrix $X_{(\cup_{i=1}^m J_i)}$, and by the condition (c) of Assumption A, bounded by $\mathcal{O}(n^{1/2})$. Thus, the largest singular value of X_I is at most $Ln^{1/2}$ for some constant L.

The constant C_2 will be chosen in the interval $(0, 1]$ to define the true vector θ^*. The condition (a) of Assumption A shows that there is a vector $u \in \mathbb{R}^s$ satisfying $\|u\|_1 = 1$ and $\|X_J^T X_J u\|_2 = \mathcal{O}(n^{1/2})$. Based on this vector u, we construct the ground truth vector θ^* by setting $\theta^*_J := C_2 u$ and $\theta^*_{-J} := 0$. This choice ensures that response vector y satisfies the linear equation $y = C_2 X_J u + w$.

Finally, the constants C_3 and C_4 are used to define two events that play an important role in the proof. Recalling that r denotes the rank of matrix X, we define the events

$$\mathcal{E}_1 := \Big\{ \|\Pi(w)\|_2^2 \leq 2r \quad \text{and} \quad \|\Pi_J(w)\|_2^2 \leq s, \Big\}, \qquad \text{and}$$

$$\mathcal{E}_2 := \Big\{ \text{there exists } I' \subseteq I \text{ such that } |X_i^T w| \geq C_3 n^{1/2} \text{ for any } i \in I' \text{ and } |I'| \geq C_4 m \Big\}.$$

At high level, the remainder of the proof consists of the following steps:

- First, we condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, and split the analysis into two cases, depending on whether $\lambda\gamma > 1/n$ or $\lambda\gamma \leq 1/n$;
- Second, we prove that the event $\mathcal{E}_1 \cap \mathcal{E}_2$ holds with positive probability.

Recall that the definition of γ relies on the constant C_1. In fact, we can choose a concrete value for the constant C_1 such that whenever $\lambda\gamma > 1/n$ (referred to as Case I), the prediction error is lower bounded by $n^{-1/2}$. This claim is formalized in the following lemma:

Lemma 2. *Assume that the event $\mathcal{E}_1$ holds. There is a setting of the constant $C_1 \in (0, 1]$ such that if $\lambda\gamma > 1/n$ holds, then there is a local minimum $\widehat{\theta}$ of the objective function that satisfies $\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 = \Omega(n^{-1/2})$.*

See Appendix E for the proof of this claim.

Once the constant C_1 is chosen so that the guarantee of Lemma 2 holds, our next step is to choose the other constants C_2, C_3 and C_4 so that the prediction error is also lower bounded by $n^{-1/2}$ in the alternative case $\lambda\gamma \leq 1/n$ (referred to as Case II). From here onwards, we fix the choice

$$C_3 := \frac{1}{C_1} + \frac{LC_1}{2}.$$

As for the constants (C_2, C_4), our proof exploits the fact that no matter how we choose the value of C_4, there is always an setting of C_2 (possibly depending on C_4) such that the lower bound on the prediction error holds. This fact is formalized in the following lemma.

Lemma 3. *Assume that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Then, given an arbitrary value $C_4 > 0$, there is a setting of the constant $C_2 \in (0, 1]$ such that if $\lambda\gamma \leq 1/n$, then the global minimizer $\widehat{\theta}$ of the objective function satisfies $\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2 = \Omega(n^{-1/2})$.*

We give the proof of this claim in Appendix E. In conjunction, Lemmas 2 and 3 guarantee that, with appropriate choices of the constants, the prediction error is lower bounded as $\Omega(n^{-1/2})$ in both Case I and Case II.

Our final step is to ensure that the event $\mathcal{E}_1 \cap \mathcal{E}_2$ holds with a positive constant probability. Recalling that the constant C_4 has not been chosen, we use this remaining degree of freedom to control this probability, as formalized in the following:

Lemma 4. *Given an arbitrary value $C_3 > 0$, there is a constant $p > 0$ and an assignment for $C_4 > 0$ such that $P(\mathcal{E}_1 \cap \mathcal{E}_2) \geq p$.*

Note that our choice of constant C_4, made to ensure that the conclusion of Lemma 4 holds, depends on that of C_3, because both C_3 and C_4 appear in the definition of event $\mathcal{E}_2$. Finally, based on the value of C_4, we choose C_2 to be the constant that makes Lemma 3 hold. Putting all pieces together, we conclude that

$$\mathbb{E}\left[\inf_{\lambda \geq 0} \sup_{\theta \in \widehat{\Theta}_\lambda} \frac{1}{n}\|X(\theta - \theta^*)\|_2^2\right] \geq \Omega(n^{-1/2}) \cdot \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) = \Omega(n^{-1/2}),$$

which completes the proof of the theorem.

4.3. Proof of Theorem 3

The proof of Theorem 3 is conceptually similar to the proof of Theorem 1, but differs in some key details. We begin with the definitions

$$\alpha := \arcsin\left(\frac{\sqrt{s_n\sigma}}{n^{1/4}\sqrt{R}}\right) \quad \text{and} \quad B := \frac{s_n\sigma}{\sqrt{n}},$$

where $s_n \in [1, \sqrt{\log n}]$ is a scaling factor to be specified later. Given our assumption $n \geq \log(n)(4\sigma/R)^2$, note that we are guaranteed that the inequalities $\sin(\alpha) \leq 1/2$ and $2B \leq 2\sigma\sqrt{\log(n)/n} \leq R/2$ hold. We then define the matrix $A \in \mathbb{R}^{2\times 2}$ and the matrix $X \in \mathbb{R}^{n\times d}$ by equations (4.1a) and (4.1b).

4.3.1. Proof of part (a)

Let $\{\theta^t\}_{t=0}^{\infty}$ be the sequence of iterates generated by equation (3.7). We proceed via proof by contradiction, assuming that the sequence does not terminate finitely, and then deriving a contradiction. We begin with a lemma.

Lemma 5. *If the sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ is not finitely convergent, then it is unbounded.*

We defer the proof of this claim to the end of this section. Based on Lemma 5, it suffices to show that, in fact, the sequence $\{\theta^t\}_{t=0}^{\infty}$ is bounded. Partitioning the full vector as $\theta := \big(\theta_{1:n}, \theta_{n+1:d}\big)$, we control the two sequences $\{\theta^t_{1:n}\}_{t=0}^{\infty}$ and $\{\theta^t_{n+1:d}\}_{t=0}^{\infty}$.

Beginning with the former sequence, notice that the objective function can be written in the form

$$L(\theta;\lambda) = \frac{1}{n}\|y - X_{1:n}\theta_{1:n}\|_2^2 + \sum_{i=1}^{d}\lambda\rho_i(\theta_i),$$

where $X_{1:n}$ represents the first n columns of matrix X. The conditions ((4.1a)) and ((4.1b)) guarantee that the Gram matrix $X_{1:n}^T X_{1:n}$ is positive definite, which implies that the quadratic function $\theta_{1:n} \mapsto \|y - X_{1:n}\theta_{1:n}\|_2^2$ is strongly convex. Thus, if the sequence $\{\theta^t_{1:n}\}_{t=0}^{\infty}$ were unbounded, then the associated cost sequence $\{L(\theta^t;\lambda)\}_{t=0}^{\infty}$ would also be unbounded. But this is not possible since $L(\theta^t;\lambda) \leq L(\theta^0;\lambda)$ for all iterations $t = 1, 2, \ldots$. Consequently, we are guaranteed that the sequence $\{\theta^t_{1:n}\}_{t=0}^{\infty}$ must be bounded.

It remains to control the sequence $\{\theta^t_{n+1:d}\}_{t=0}^{\infty}$. We claim that for any $i \in \{n+1, \ldots, d\}$, the sequence $\{|\theta_i^t|\}_{t=0}^{\infty}$ is non-increasing, which implies the boundedness condition. Proceeding via proof by contradiction, suppose that $|\theta_i^t| < |\theta_i^{t+1}|$ for some index $i \in \{n+1, \ldots, d\}$ and iteration number $t \geq 0$. Under this condition, define the vector

$$\widetilde{\theta}_j^{t+1} := \begin{cases} \theta_j^{t+1} & \text{if } j \neq i \\ \theta_j^t & \text{if } j = i. \end{cases}$$

Since ρ_j is a monotonically non-decreasing function of $|x|$, we are guaranteed that $L(\widetilde{\theta}^{t+1};\lambda) \leq L(\theta^{t+1};\lambda)$, which implies that $\widetilde{\theta}^{t+1}$ is also a constrained minimum point over the ball $\mathbb{B}_2(\eta;\theta^t)$. In addition, we have

$$\|\widetilde{\theta}^{t+1} - \theta^t\|_2 = \|\theta^{t+1} - \theta^t\|_2 - |\theta_i^t - \theta_i^{t+1}| < \eta,$$

so that $\widetilde{\theta}_j^{t+1}$ is strictly closer to θ^t. This contradicts the specification of the algorithm, in that it chooses the minimum closest to θ^t.

Proof of Lemma 5: The final remaining step is to prove Lemma 5. We first claim that $\|\theta^s - \theta^t\|_2 \geq \eta$ for all pairs $s < t$. If not, we could find some pair $s < t$ such that $\|\theta^s - \theta^t\|_2 < \eta$. But since $t > s$, we are guaranteed that $L(\theta^t; \lambda) \leq L(\theta^{s+1}; \lambda)$. Since θ^{s+1} is a global minimum over the ball $\mathbb{B}_2(\eta; \theta^s)$ and $\|\theta^s - \theta^t\|_2 < \eta$, the point θ^t is also a global minimum, and this contradicts the definition of the algorithm (since it always chooses the constrained global minimum closest to the current iterate).

Using this property, we now show that the sequence $\{\theta^t\}_{t=0}^{\infty}$ is unbounded. For each iteration $t = 0, 1, 2 \ldots$, we use $\mathbb{B}^t = \mathbb{B}_2(\eta/3; \theta^t)$ to denote the Euclidean ball of radius $\eta/3$ centered at θ^t. Since $\|\theta^s - \theta^t\|_2 \geq \eta$ for all $s \neq t$, the balls $\{\mathbb{B}^t\}_{t=0}^{\infty}$ are all disjoint, and hence there is a numerical constant $C > 0$ such that for each $T \geq 1$, we have

$$\text{vol}\Big(\cup_{t=0}^{T} \mathbb{B}^t \Big) = \sum_{t=0}^{T} \text{vol}(\mathbb{B}^t) = C \sum_{t=0}^{T} \eta^d.$$

Since this volume diverges as $T \to \infty$, we conclude that the set $\mathbb{B} := \cup_{t=0}^{\infty} \mathbb{B}^t$ must be unbounded. By construction, any point in $\mathbb{B}$ is within $\eta/3$ of some element of the sequence $\{\theta^t\}_{t=0}^{\infty}$, so this sequence must be unbounded, as claimed.

4.3.2. Proof of part (b)

We now prove a lower bound on the prediction error corresponding the local minimum to which the algorithm converges, as claimed in part (b) of the theorem statement. In order to do so, we begin by introducing the shorthand notation

$$\gamma_i = \min \Big\{ \sup_{u \in (0,B]} \rho'_{2i-1}(u), \quad \sup_{u \in (0,B]} \rho'_{2i}(u) \Big\} \qquad \text{for each } i = 1, \ldots, n/2.$$

Then we define the quantities a_i and w'_i by equations (4.3a). Similar to the proof of Theorem 1, we assume (without loss of generality, re-indexing as needed) that $\gamma_i = \sup_{u \in (0,B]} \rho'_{2i-1}(u)$ and that $\gamma_1 = \max_{i \in [n/2]} \{\gamma_i\}$.

Consider the regression vector $\theta^* := \begin{bmatrix} \frac{R}{2} & \frac{R}{2} & 0 & \cdots & 0 \end{bmatrix}$. Since the matrix A appears in diagonal blocks of X, the algorithm's output $\widehat{\theta}$ has error

$$\inf_{\lambda \geq 0} \frac{1}{n} \|X(\widehat{\theta} - \theta^*)\|_2^2 = \inf_{\lambda \geq 0} \sum_{i=1}^{n/2} \|A\big(\widehat{\theta}_{(2i-1):2i} - \theta^*_{(2i-1):2i}\big)\|_2^2. \tag{4.7}$$

Given the random initialization θ^0, we define the events

$$\mathcal{E}_0 := \Big\{ \max\{\theta_1^0, \theta_2^0\} \leq 0 \text{ and } \frac{\|w_{1:2}\|_2}{\sqrt{n}} \leq B \Big\} \quad \text{and} \quad \mathcal{E}_1 := \Big\{ \lambda \gamma_1 > 2 \sin^2(\alpha) R + 5B \Big\},$$

as well as the (random) subsets

$$\mathbb{S}_1 := \Big\{ i \in \{2, \ldots, n/2\} \mid \lambda\gamma_1 \leq \frac{w_{2i-1}}{\sqrt{n}} - B \text{ and } \frac{|w_{2i}|}{\sqrt{n}} \leq B \Big\}, \quad \text{and} \tag{4.8a}$$

$$\mathbb{S}_2 := \Big\{ i \in \{2, \ldots, n/2\} \mid 2\sin^2(\alpha)R + 5B \leq \frac{w_{2i-1}}{\sqrt{n}} - B \text{ and } \frac{|w_{2i}|}{\sqrt{n}} \leq B \Big\}. \tag{4.8b}$$

Given these definitions, the following lemma provides lower bounds on the decomposition (4.7) for the vector $\widehat{\theta}$ after convergence.

Lemma 6. *(a) If $\mathcal{E}_0 \cap \mathcal{E}_1$ holds, then $\|A\left(\widehat{\theta}_{1:2} - \theta^*_{1:2}\right)\|_2^2 \geq \frac{s_n \sigma R}{4\sqrt{n}}$.*

(b) For any index $i \in \mathbb{S}_1$, we have $\|A\left(\widehat{\theta}_{2i-1:2i} - \theta^*_{2i-1:2i}\right)\|_2^2 \geq \frac{s_n^2 \sigma^2}{20n}$.

See Appendix F for the proof of this claim.

Conditioned on event $\mathcal{E}_0$, for any index $i \in \mathbb{S}_2$, either the event $\mathcal{E}_0 \cap \mathcal{E}_1$ holds, or we have

$$\lambda\gamma_1 < 2\sin^2(\alpha)R + 5B \leq \frac{w_{2i-1}}{\sqrt{n}} - B \quad \text{and} \quad \frac{|w_{2i}|}{\sqrt{n}} \leq B$$

which means that $i \in \mathbb{S}_1$ holds. Applying Lemma 6 yields the lower bound

$$\inf_{\lambda \geq 0} \sum_{i=1}^{n/2} \|A\widehat{\theta}_{(2i-1):2i} - A\theta^*_{(2i-1):2i}\|_2^2 \geq \mathbb{I}[\mathcal{E}_0] \ \min\Big\{ \frac{s_n \sigma R}{4\sqrt{n}}, \ \frac{s_n^2 \sigma^2}{20n} \sum_{i=2}^{[n/2]} \mathbb{I}[i \in \mathbb{S}_2] \Big\}.$$

The random variables $\mathbb{I}[\mathcal{E}_0]$ and $\{\mathbb{I}[i \in \mathbb{S}_2]\}_{i=2}^{[n/2]}$ are mutually independent. Plugging in the definition of quantity B, it is easy to see that $\mathbb{P}[\mathcal{E}_0]$ is lower bounded by a universal constant $c_0 > 0$, so that

$$\begin{aligned} &\mathbb{E}\Big[\inf_{\lambda \geq 0} \sum_{i=1}^{n/2} \|A\widehat{\theta}_{(2i-1):2i} - A\theta^*_{(2i-1):2i}\|_2^2 \Big] \\ &\qquad \geq c_0 \, \mathbb{E}\Big[\min\Big\{ \frac{s_n \sigma R}{4\sqrt{n}}, \ \frac{s_n^2 \sigma^2}{20n} \sum_{i=2}^{[n/2]} \mathbb{I}[i \in \mathbb{S}_2] \Big\} \Big] \end{aligned} \tag{4.9}$$

The right-hand side of inequality (4.9) takes the same form as that of the right-hand side of inequality (4.6). Thus, we can follow the steps in the proof of Theorem 1 to lower bound this term. Each event $i \in \mathbb{S}_2$ is the intersection of two independent events $w_{2i-1} \geq 8 s_n \sigma$ and $|w_{2i}| \leq s_n \sigma$. Recall that both w_{2i-1} and w_{2i} have the distribution $N(0, \sigma^2)$. In the proof of Theorem 1, we have shown that by choosing $s_n := \max\{1, c\sqrt{\log(n)}\}$ for a sufficiently small universal constant $c_1 > 0$, the first event happens with probability at least $c_2/n^{1/4}$, where $c_2 > 0$ is a universal constant. The second event happens with probability at least $\mathbb{P}[|w_{2i}| \leq \sigma] > 0.68$. Therefore, the event $i \in \mathbb{S}_2$ happens with probability at least $c_3 n^{-1/4}$ for some universal constant $c_3 > 0$.

Having lower bounded the probability of events $i \in \mathbb{S}_2$, we lower bound the right-hand side of inequality (4.9) using the same argument that we used for lower bounding the right-hand side of inequality (4.6), which implies

$$\mathbb{E}\Big[\inf_{\lambda \geq 0} \sum_{i=1}^{n/2} \|A\widehat{\theta}_{(2i-1):2i} - A\theta^*_{(2i-1):2i}\|_2^2\Big] \geq c\,\sigma R\sqrt{\frac{\log(n)}{n}},$$

for a universal constant $c > 0$. This completes the proof.

5. Discussion

In this paper, we have demonstrated a fundamental gap in sparse linear regression: the best prediction risk achieved by a class of M-estimators based on coordinate-wise separable regularizers is strictly larger than the classical minimax prediction risk, achieved for instance by minimization over the ℓ_0-ball. This gap applies to a range of methods used in practice, including the Lasso in its ordinary and weighted forms, as well as estimators based on nonconvex penalties such as the MCP and SCAD penalties.

Several open questions remain, and we discuss a few of them here. When the penalty function ρ is convex, the M-estimator minimizing function (2.1) can be understood as a particular convex relaxation of the ℓ_0-based estimator (1.3). It would be interesting to consider other forms of convex relaxations for the ℓ_0-based problem. For instance, Pilanci et al. [31] show how a broad class of ℓ_0-regularized problems can be reformulated exactly as optimization problems involving convex functions in Boolean variables. This exact reformulation allows for the direct application of many standard hierarchies for Boolean polynomial programming, including the Lasserre hierarchy [24] as well as the Sherali-Adams hierarchy [35]. Other relaxations are possible, including those that are based on introducing auxiliary variables for the pairwise interactions (e.g., $\gamma_{ij} = \theta_i\theta_j$), and so incorporating these constraints as polynomials in the constraint set. We conjecture that for any fixed natural number t, if the t-th level Lasserre (or Sherali-Adams) relaxation is applied to such a reformulation, it still does not yield an estimator that achieves the fast rate (1.4). Since a t^{th}-level relaxation involves $\mathcal{O}(d^t)$ variables, this would imply that these hierarchies do not contain polynomial-time algorithms that achieve the classical minimax risk. Proving or disproving this conjecture remains an open problem.

Finally, when the penalty function ρ is concave, concurrent work by Ge et al. [19] shows that finding the global minimum of the loss function (2.1) is strongly NP-hard. This result implies that no polynomial-time algorithm computes the global minimum unless $\mathbf{NP} = \mathbf{P}$. The result given here is complementary in nature: it shows that bad local minima exist, and that local descent methods converge to these bad local minima. It would be interesting to extend this algorithmic lower bound to a broader class of first-order methods. For instance, we suspect that any algorithm that relies on an oracle giving first-order information will inevitably converge to a bad local minimum for a broad class of random initializations.

Appendix A: Fast rate for the bad example of Dalalyan et al. [12]

In this appendix, we describe the bad example of Dalayan et al. [12], and show that a reweighted form of the Lasso achieves the fast rate. For a given sample size $n \geq 4$, they consider a linear regression model $y = X\theta^* + w$, where $X \in \mathbb{R}^{n \times 2m}$ with $m = n - 1$, and the noise vector $w \in \{-1, 1\}^n$ has i.i.d. Rademacher entries (equiprobably chosen in $\{-1, 1\}$). In the construction, the true vector θ^* is 2-sparse, and the design matrix $X \in \mathbb{R}^{n \times 2m}$ is given by

$$X = \sqrt{n} \begin{bmatrix} 1_m^T & \mathbf{1}_m^T \\ I_{m \times m} & -I_{m \times m} \end{bmatrix},$$

where $1_m \in \mathbb{R}^m$ is a vector of all ones. Notice that this construction has $n = m + 1$.

In this appendix, we analyze the performance of the following estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^{2m}} \frac{1}{n} \|X\theta - y\|_2^2 + \lambda \sum_{i=2}^{m} (|\theta_i| + |\theta_{m+i}|). \tag{A.1}$$

It is a reweighted form of the Lasso based on ℓ_1-norm regularization, but one that imposes *no* constraint on the first and the $(m+1)$-th coordinate. We claim that with an appropriate choice of λ, this estimator achieves the fast rate for any 2-sparse vector θ^*.

Letting $\widehat{\theta}$ be a minimizer of function (A.1), we first observe that no matter what value it attains, the minimizer always chooses $\widehat{\theta}_1$ and $\widehat{\theta}_{m+1}$ so that $(X\widehat{\theta})_{1:2} = y_{1:2}$. This property occurs because:

- There is no penalty term associated with $\widehat{\theta}_1$ and $\widehat{\theta}_{m+1}$.
- By the definition of X, changes in the coordinates $\widehat{\theta}_1$ and $\widehat{\theta}_{m+1}$ only affect the first two coordinates of $X\widehat{\theta}$ by the additive term

$$\sqrt{n} \begin{bmatrix} 1 & 1 \\ e_1 & -e_1 \end{bmatrix} \cdot \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_{m+1} \end{bmatrix}$$

  Since the above 2-by-2 matrix is non-singular, there is always an assignment to $(\widehat{\theta}_1, \widehat{\theta}_{m+1})$ so that $(X\widehat{\theta})_{1:2} - y_{1:2} = 0$.

Thus, only the last $n - 2$ coordinates of $X\widehat{\theta} - y$ might be non-zero, so that we may rewrite the objective function (A.1) as

$$\begin{aligned} &\frac{1}{n} \|X\widehat{\theta} - y\|_2^2 + \lambda \sum_{i=2}^{m} (|\widehat{\theta}_i| + |\widehat{\theta}_{m+i}|) \\ &\quad = \frac{1}{n} \sum_{i=2}^{m} \left((\sqrt{n}\widehat{\theta}_i - \sqrt{n}\widehat{\theta}_{m+i} - y_i)^2 + \lambda(|\widehat{\theta}_i| + |\widehat{\theta}_{m+i}|) \right). \end{aligned} \tag{A.2}$$

The function (A.2) is not strictly convex so that there are multiple equivalent solutions. Essentially, we need to break symmetry by choosing to vary one of $\widehat{\theta}_i$ or $\widehat{\theta}_{m+i}$, for each $i \in \{2, \ldots, m\}$. Without loss of generality, we assume that $\widehat{\theta}_{m+2} = \widehat{\theta}_{m+3} = \cdots = \widehat{\theta}_{2m} = 0$, so that the equation is simplified as

$$\frac{1}{n}\|X\widehat{\theta} - y\|_2^2 + \lambda \sum_{i=2}^{m}(|\widehat{\theta}_i| + |\widehat{\theta}_{m+i}|) = \frac{1}{n}\sum_{i=2}^{m}\left((\sqrt{n}\widehat{\theta}_i - y_i)^2 + \lambda|\widehat{\theta}_i|)\right). \quad \text{(A.3)}$$

Moreover, with this choice, we can write the prediction error as

$$R(\widehat{\theta}) := \frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2 = \frac{2}{n} + \frac{1}{n}\sum_{i=2}^{m}(\sqrt{n}\widehat{\theta}_i - \sqrt{n}(\theta_i^* - \theta_{m+i}^*))^2. \quad \text{(A.4)}$$

The first term on the right-hand side is obtained from the fact $\|(X\widehat{\theta} - X\theta^*)_{1:2}\|_2^2 = \|w_{1:2}\|_2^2 = 2$, recalling that the set-up assumes that the noise elements takes values in $\{-1, 1\}$.

The right-hand side of equation (A.3) is a Lasso objective function with design matrix $\sqrt{n}I_{(m-1)\times(m-1)}$. The second term on the right-hand side of equation (A.4) is the associated prediction error. By choosing a proper λ and using the fact that θ^* is 2-sparse, it is well-known that the prediction error scales as $\mathcal{O}(\frac{\log n}{n})$, which corresponds to the fast rate. (Here we have recalled that the dimension of the Lasso problem is $m - 1 = n - 2$.)

Appendix B: Proof of Lemma 1

Given our definition of X in terms of the matrix $A \in \mathbb{R}^{2\times 2}$, it suffices to prove the two lower bounds

$$\|A(\widehat{\theta}_\lambda)_{1:2} - A\theta^*_{1:2}\|_2^2 \geq \mathbb{I}\Big[\lambda\gamma_1 > 4B(\sin^2(\alpha)R + \frac{\|w_{1:2}\|_2}{\sqrt{n}})\Big]\sin^2(\alpha)(R - 2B)_+^2 \quad \text{and,} \quad \text{(B.1a)}$$

$$\|A(\widehat{\theta}_\lambda)_{2i-1:2i} - A\theta^*_{2i-1:2i}\|_2^2 \geq \mathbb{I}\Big[0 \leq w_i' \leq B\Big](\frac{B^2}{4} - \lambda\gamma_1)_+ \quad \text{for } i = 2, 3, \ldots, n/2. \quad \text{(B.1b)}$$

In the proofs to follow, it is convenient to omit reference to the index i. In particular, viewing the index i as fixed a priori, we let $\widehat{u}$ and u^* be shorthand representations of the sub-vectors $(\widehat{\theta}_\lambda)_{2i-1,2i}$, and $\theta^*_{2i-1,2i}$, respectively. We introduce the normalized noise $\varepsilon := w_{2i-1,2i}/\sqrt{n}$. By our construction of the design matrix X in terms of A, the vector $\widehat{\theta}_\lambda$ is a local minimizer of the objective function if and only if $\widehat{u}$ is a local minimum of the following loss:

$$\ell(u; \lambda) := \|Au - Au^* - \varepsilon\|_2^2 + \lambda\rho_{2i-1}(u_1) + \lambda\rho_{2i}(u_2),$$

where this statement should hold for each $i \in [n/2]$. Hence, it suffices to find a local minimum of $\ell(u; \lambda)$ such that the bounds (B.1a) and (B.1b) hold.

B.1. Proof of inequality (B.1a)

If $\lambda\gamma_1 \leq 4B(\sin^2(\alpha)R + \|\varepsilon\|_2)$, then the lower bound (B.1a) is trivial, and in particular, it holds for $\widehat{u} := \arg\min_u \ell(u;\lambda)$.

Otherwise, we may assume that $\lambda\gamma_1 > 4B(\sin^2(\alpha)R + \|\varepsilon\|_2)$. In this case ($i = 1$), we have $u^* = (R/2, R/2)$. Defining the vectors $v^* := Au^* = (0, \sin(\alpha)R)$ and $\widetilde{u} := (0,0)$, we have

$$\ell(\widetilde{u};\lambda) = \|A\widetilde{u} - v^* - \varepsilon\|_2^2 + \lambda\rho_1(\widetilde{u}_1) + \lambda\rho_2(\widetilde{u}_2) = \|v^* + \varepsilon\|_2^2. \tag{B.2}$$

We claim that

$$\inf_{u\in\partial U} \ell(u;\lambda) > \ell(\widetilde{u};\lambda), \tag{B.3}$$

where $U := \{u \in \mathbb{R}^2 \mid |u_1| \leq B \text{ and } |u_2| \leq B\}$, and ∂U denotes its boundary. If ρ is a convex function, then the lower bound (B.3) implies that any minimizers of the function $\ell(\cdot;\lambda)$ lie in the interior of U. Otherwise, it implies that at least one local minimum—say $\widehat{u}$—lies in the interior of U. Since $\widehat{u}_1 \leq B$ and $\widehat{u}_2 \leq B$, we have the lower bound

$$\begin{aligned}\|A(\widehat{\theta}_\lambda)_{1:2} - A\theta^*_{1:2}\|_2^2 = \|A\widehat{u} - v^*\|_2 &= \cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 + \sin^2(\alpha)(R - \widehat{u}_1 - \widehat{u}_2)^2 \\ &\geq \sin^2(\alpha)(R - \widehat{u}_1 - \widehat{u}_2)^2 \geq \sin^2(\alpha)(R-2B)_+^2,\end{aligned}$$

which completes the proof.

It remains to prove the lower bound (B.3). For any $u \in \partial U$, we have

$$\begin{aligned}\ell(u;\lambda) = \|Au - v^* - \varepsilon\|_2^2 + \lambda\rho_1(\widetilde{u}_1) + \lambda\rho_2(\widetilde{u}_2) &\overset{(i)}{\geq} \|Au + v^* + \varepsilon\|_2^2 + \lambda\gamma_1 \\ &\overset{(ii)}{\geq} \|v^* + \varepsilon\|_2^2 + 2(v^* + \varepsilon)^T Au + \lambda\gamma_1.\end{aligned} \tag{B.4}$$

Inequality (i) holds since either $\widetilde{u}_1$ or $\widetilde{u}_2$ is equal to B, and $\min\{\rho_1(B), \rho_2(B)\} \geq \gamma_1$ by definition, whereas inequality (ii) holds since $\|a+b\|_2^2 \geq \|b\|_2^2 + 2b^T a$. We notice that

$$\begin{aligned}&\inf_{u\in\partial U} 2(v^* + \varepsilon)^T Au \geq \inf_{u\in\partial U} \{2\langle v^*, Au\rangle - 2\|\varepsilon\|_2 \|Au\|_2\} \\ &= \inf_{u\in\partial U} \left\{2\sin^2(\alpha)R(u_1+u_2) - 2\|\varepsilon\|_2\sqrt{\cos^2(\alpha)(u_1-u_2)^2 + \sin^2(\alpha)(u_1+u_2)^2}\right\} \\ &\geq -4B(\sin^2(\alpha)R + \|\varepsilon\|_2).\end{aligned}$$

Combining this lower bound with inequality (B.4) and the bound

$$\lambda\gamma_1 > 4B(\sin^2(\alpha)R + \|\varepsilon\|_2)$$

yields the claim (B.3).

B.2. Proof of inequality (B.1b)

For $i = 2, 3, \ldots, n/2$, consider $u^* = (0, 0)$ and recall the vector $a_i = (\cos\alpha, \sin\alpha)$, as well as our assumption that

$$\gamma_i := \rho_{2i-1}(B)/B \leq \frac{\rho_{2i}(B)}{B}.$$

Define the vector $\widetilde{u} := (a_i^T\varepsilon, 0)$ and let $\widehat{u} = \arg\min_{u\in\mathbb{R}^2} \ell(u;\lambda)$ be an arbitrary global minimizer. We then have

$$\|A\widehat{u} - \varepsilon\|_2^2 \leq \|A\widehat{u} - \varepsilon\|_2^2 + \lambda\rho_1(\widehat{u}_1) + \lambda\rho_2(\widehat{u}_2) \leq \|A\widetilde{u} - \varepsilon\|_2^2 + \lambda\rho_1(\widetilde{u}_1) + \lambda\rho_2(\widetilde{u}_2),$$

since the regularizer is non-negative, and $\widehat{u}$ is a global minimum. Using the definition of $\widetilde{u}$, we find that

$$\|A\widehat{u} - \varepsilon\|_2^2 \leq \|a_i a_i^T\varepsilon - \varepsilon\|_2^2 + \lambda\rho_1(a_i^T\varepsilon) = \|\varepsilon\|_2^2 - (a_i^T\varepsilon)^2 + \lambda\rho_1(a_i^T\varepsilon),$$

where the final equality holds since $a_i a_i^T$ defines an orthogonal projection. By the triangle inequality, we find that $\|A\widehat{u} - \varepsilon\|_2^2 \geq \|\varepsilon\|_2^2 - \|A\widehat{u}\|_2^2$, and combining with the previous inequality yields

$$\|A\widehat{u}\|_2^2 \geq (a_i^T\varepsilon)^2 - \lambda\rho_1(a_i^T\varepsilon). \tag{B.5}$$

Now if $B/2 \leq a^T\varepsilon \leq B$, then we have $\rho_1(a_i^T\varepsilon) \leq \rho_1(B) = \gamma_i \leq \gamma_1$. Substituting this relation into inequality (B.5), we have

$$\|A\widehat{u} - \varepsilon\|_2^2 \geq \mathbb{I}(B/2 \leq a_i^T\varepsilon \leq B)\left(B^2/4 - \lambda\gamma_1\right)_+,$$

which completes the proof.

Appendix C: Proof of Corollary 1

Here we provide a detailed proof of inequality (3.4a). We note that inequality (3.4b) follows by an essentially identical series of steps, so that we omit the details.

Let m be an even integer and let $X_m \in \mathbb{R}^{m\times m}$ denote the design matrix constructed in the proof of Theorem 1. In order to avoid confusion, we rename the parameters (n, d, R) in the construction (4.1b) by (n', d', R'), and set them equal to

$$(n', d', R') := \left(m, m, \min\left\{\frac{R\sqrt{n}}{k\sqrt{m}}, \frac{s_m\sigma}{16\gamma\sqrt{m}}\right\}\right), \tag{C.1}$$

where the quantities (k, n, R, σ) are defined in the statement of Corollary 1, and the scaling factor $s_m := c_1\sqrt{\log m}$ was defined in the proof of Theorem 1. Note that X_m is a square matrix, and according to equation (4.1b), all of its

eigenvalues are lower bounded by $(\frac{m^{1/2} s_m \sigma}{16 R'})^{1/2}$. By equation (C.1), this quantity is lower bounded by $\sqrt{m\gamma}$.

Using the matrix X_m as a building block, we now construct a larger design matrix $X \in \mathbb{R}^{n \times n}$ that we then use to prove the corollary. Let m be the greatest integer divisible by two such that $km \leq n$. We may construct the $n \times n$ dimensional matrix

$$X := \text{blkdiag}\Big\{ \underbrace{\sqrt{n/m} X_m, \ldots, \sqrt{n/m} X_m}_{k \text{ copies}}, \sqrt{n} I_{n-km} \Big\} \in \mathbb{R}^{n \times n}, \tag{C.2}$$

where I_{n-km} is the $(n-km)$-dimensional identity matrix. It is easy to verify the matrix X satisfies the column normalization condition. Since all eigenvalues of X_m are lower bounded by $\sqrt{m\gamma}$, we are guaranteed that all eigenvalues of X are lower bounded by $\sqrt{n\gamma}$. Thus, the matrix X satisfies the γ-RE condition.

It remains to prove a lower bound on the prediction error, and in order to do so, it is helpful to introduce some shorthand notation. Given an arbitrary vector $u \in \mathbb{R}^n$, for each integer $i \in \{1, \ldots, k\}$, we let $u_{(i)} \in \mathbb{R}^m$ denote the sub-vector consisting of the $((i-1)m+1)$-th to the (im)-th elements of vector u, and we let $u_{(k+1)}$ denote the sub-vector consisting of the last $n - km$ elements. We also introduce similar notation for the function $\rho(x) = \rho_1(x_1) + \cdots + \rho_n(x_n)$; specifically, for each $i \in \{1, \ldots, k\}$, we define the function $\rho_{(i)} : \mathbb{R}^m \to \mathbb{R}$ via $\rho_{(i)}(\theta) := \sum_{j=1}^m \rho_{(i-1)m+j}(\theta_j)$.

Using this notation, we may rewrite the cost function as:

$$L(\theta; \lambda) = \frac{1}{n} \sum_{i=1}^k \Big(\|\sqrt{n/m} X_m \theta_{(i)} - y_{(i)}\|_2^2 + n\lambda \rho_{(i)}(\theta_{(i)}) \Big) + h(\theta_{(k+1)}),$$

where h is a function that only depends on $\theta_{(k+1)}$. If we define $\theta'_{(i)} := \sqrt{n/m}\theta_{(i)}$ as well as $\rho'_{(i)}(\theta) := \frac{n}{m}\rho_{(i)}(\sqrt{m/n}\theta)$, then substituting them into the above expression, the cost function can be rewritten as

$$G(\theta'; \lambda) := \frac{m}{n} \sum_{i=1}^k \Big(\frac{1}{m} \|X_m \theta'_{(i)} - y_{(i)}\|_2^2 + \lambda \rho'_{(i)}(\theta'_{(i)}) \Big) + h(\sqrt{m/n}\, \theta'_{(k+1)}).$$

Note that if the vector $\widehat{\theta}$ is a local minimum of the function $\theta \mapsto L(\theta; \lambda)$, then the rescaled vector $\widehat{\theta}' := \sqrt{n/m}\, \widehat{\theta}$ is a local minimum of the function $\theta' \mapsto G(\theta'; \lambda)$. Consequently, the sub-vector $\widehat{\theta}'_{(i)}$ must be a local minimum of the function

$$\frac{1}{m} \|X_m \theta'_{(i)} - y_{(i)}\|_2^2 + \rho'_{(i)}(\theta'_{(i)}). \tag{C.3}$$

Thus, the sub-vector $\widehat{\theta}'_{(i)}$ is the solution of a regularized sparse linear regression problem with design matrix X_m.

Defining the rescaled true regression vector $(\theta^*)' := \sqrt{n/m}\, \theta^*$, we can then write the prediction error as

$$\frac{1}{n} \|X(\widehat{\theta} - \theta^*)\|_2^2 = \frac{1}{n} \sum_{i=1}^k \Big(\|X_m(\widehat{\theta}'_{(i)} - (\theta^*)'_{(i)})\|_2^2 \Big) + \|\widehat{\theta}_{(k+1)} - \theta^*_{(k+1)}\|_2^2$$

$$\geq \frac{m}{n}\sum_{i=1}^{k}\Big(\frac{1}{m}\|X_m(\widehat{\theta}'_{(i)} - (\theta^*)'_{(i)})\|_2^2\Big). \tag{C.4}$$

Consequently, the overall prediction error is lower bounded by a scaled sum of the prediction errors associated with the design matrix X_m. Moreover, each term $\frac{1}{m}\|X_m(\widehat{\theta}'_{(i)} - (\theta^*)'_{(i)})\|_2^2$ can be bounded by Theorem 1.

More precisely, let $\mathcal{Q}(X, 2k, R)$ denote the left-hand side of inequality (3.4a). The above analysis shows that the sparse linear regression problem on the design matrix X and the constraint $\theta^* \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(R)$ can be decomposed into smaller-scale problems on the design matrix X_m and constraints on the scaled vector $(\theta^*)'$. By the rescaled definition of $(\theta^*)'$, the constraint $\theta^* \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(R)$ holds if and only if $(\theta^*)' \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(\sqrt{n/m}R)$. Recalling the definition of the radius R' from equation (C.1), we can ensure that $(\theta^*)' \in \mathbb{B}_0(2k) \cap \mathbb{B}_1(\sqrt{n/m}R)$ by requiring that $(\theta^*)'_{(i)} \in \mathbb{B}_0(2) \cap \mathbb{B}_1(R')$ for each index $i \in \{1, \ldots, k\}$. Combining expressions (C.3) and (C.4), the quantity $\mathcal{Q}(X, 2k, R)$ can be lower bounded by the sum

$$\mathcal{Q}(X, 2k, R) \geq \frac{m}{n}\sum_{i=1}^{k}\mathcal{Q}(X_m, 2, R'). \tag{C.5a}$$

By the assumption of the corollary, we have

$$m = \lfloor n/k \rfloor \geq \max\{4, \frac{8\sigma}{R})^2 \log m, (\frac{R}{\sigma})^4\},$$

satisfying the sample size assumption of Theorem 1. Thus, Theorem 1 implies

$$\mathcal{Q}(X_m, 2, R') \geq c\,\sigma R'\sqrt{\frac{\log m}{m}} = c' \min\left\{\frac{\sigma^2 \log n}{16\gamma m}, \frac{\sigma R\sqrt{n \log n}}{km}\right\}, \tag{C.5b}$$

where the second equality follows from our choce of R' from equation (C.1), and uses the fact that $\log m = \log\lfloor n/k \rfloor \geq \log\lfloor (2n)^{1/2} \rfloor = \Omega(\log n)$. Combining the lower bounds (C.5a) and (C.5b) completes the proof.

Appendix D: Proof of Proposition 3

Throughout this proof, if an event holds with probability at least $1 - e^{-c\sqrt{n}}$ for some universal constant $c > 0$, then we say that this event holds with high probability. It is clear that if there are poly(n) events which hold with high probability, then these events simultaneously hold with high probability.

Note that each column X_j ($j \in [d]$) follows the n-dimensional normal distribution $\mathcal{N}(0, \frac{1}{2}I_{n\times n})$. The squared norm of $\sqrt{2}X_j$ follows a chi-square distribution with n degrees of freedom. The concentration of chi-square random variables [see, e.g. 13, Lemma 2.2] implies that $2\|X_j\|_2^2$ is bounded by $2n$ with high probability, so that X_j satisfies the column normalization condition with high probability. In addition, the rank of matrix X is equal to $n^{1/2}$ almost surely.

By the definition of the matrix X, the vectors X_{2i-1} and X_{2i} are correlated, but independent from all other columns. Since all columns are joint Gaussian vectors, we assume that there is an n-by-d matrix W, where each entry has the standard normal distribution, such that

$$X_{2i-1:2i} = W_{2i-1:2i}B \quad \text{where} \quad B := \frac{1}{\sqrt{2}} \begin{bmatrix} (1-n^{-1/2})^{1/2} & (1-n^{-1/2})^{1/2} \\ n^{-1/4} & -n^{-1/4} \end{bmatrix}. \tag{D.1}$$

It is easy to verify that the vector $X_{2i-1,2i}$ written in this way has mean zero and covariance matrix A. As a consequence, the matrix X can be written in the form

$$X = W \cdot \operatorname{diag}(\underbrace{B, B, \ldots, B}_{d/2 \text{ copies}}).$$

In order to verify that the conditions (a), (b) and (c) hold, we choose $s := 2$ and let $J_i := \{2i-1, 2i\}$ for $i = 1, \ldots, d/2$. For each $i \in [d/2]$, let $u := (1/2, -1/2)$ be a vector of unit ℓ_1-norm, then equation (D.1) implies that

$$\begin{aligned} X_{J_i}^T X_{J_i} u &= B^T W_{2i-1:2i}^T \Big(\frac{1}{\sqrt{2}} n^{-1/4} W_{2i} \Big) \\ &= \frac{n^{-1/4}}{2} \begin{bmatrix} (1-n^{-1/2})^{1/2} W_{2i-1}^T W_{2i} + n^{-1/4} \|W_{2i}\|_2^2 \\ (1-n^{-1/2})^{1/2} W_{2i-1}^T W_{2i} - n^{-1/4} \|W_{2i}\|_2^2 \end{bmatrix}. \end{aligned}$$

In order to bound the inner product $W_{2i-1}^T W_{2i}$, we note that it can be written as

$$\|W_{2i-1}\|_2 \cdot \Big\langle \frac{W_{2i-1}}{\|W_{2i-1}\|_2}, W_{2i} \Big\rangle.$$

The first term is bounded by $\mathcal{O}(n^{1/2})$ with high probability due to the concentration of chi-square random variables [13]. The second term satisfies a standard normal distribution, bounded by $\mathcal{O}(n^{1/4})$ with high probability. Thus the inner product is bounded by $\mathcal{O}(n^{3/4})$ with high probability. In addition, the squared norm $\|W_{2i}\|_2^2$ is bounded by $\mathcal{O}(n)$ with high probability. Combining these facts, we find that $\|X_{J_i}^T X_{J_i} u\|_2$ is bounded by $\mathcal{O}(n^{1/2})$ with high probability, thus condition (a) holds.

For each $i \in [d/2]$, Let S and S' be the column space of X_{J_i} and X_{-J_i}, respectively. For any unit vector $v \in S$ and $v' \in S'$, let $\Pi(\cdot)$ be the projection operator onto the space S', then we have $\langle v, v' \rangle = \langle \Pi(v), v' \rangle \leq \|\Pi(v)\|_2$. To upper bound the right-hand side, we let $\{b_1, \ldots, b_s\}$ be an orthogonal basis of the space S. Then the vector v can be represented by $v := \sum_{i=1}^s \alpha_i b_i$ with the constraint $\sum_{i=1}^s \alpha_i^2 = 1$. This representation implies that

$$\|\Pi(v)\|_2 = \Big\| \sum_{i=1}^s \alpha_i \Pi(b_i) \Big\|_2 \leq \sum_{i=1}^s \alpha_i \|\Pi(b_i)\|_2.$$

Notice that each b_i is an n-dimensional random unit vector that is independent of the space S'. Dasgupta and Gupta [13, Lemma 2.2] prove that $\|\Pi(b_i)\|_2$ is bounded by $2(d/n)^{1/2}$ with probability at least $1 - e^{c \cdot d}$ for some universal constant $c > 0$. Consequently, the inner product $\langle v, v' \rangle$ is bounded by $2(sd/n)^{1/2}$ with high probability. Plugging in $d = n^{1/2}$ verifies condition (b).

In order to verify condition (c), we use the concentration of chi-square random variables [13] to establish that $\frac{1}{\sqrt{n}}\|X_j\|_2 = \Omega(1)$ with high probability. Since $d = n^{1/2}$, Rudelson and Vershynin [34] show that the singular values of matrix W lie in the interval $[\frac{1}{2}(n^{1/2} - n^{1/4}), 2(n^{1/2} + n^{1/4})]$ with probability at least $1 - e^{-cn}$ for a universal constant $c > 0$. On the other hand, the singular values of matrix B lie between $\Omega(n^{-1/4})$ and $\mathcal{O}(1)$. Combining these facts, inequality (D.1) implies that the singular values of matrix $\frac{1}{\sqrt{n}}X$ are lower bounded by $\Omega(n^{-1/4})$ and upper bounded by $\mathcal{O}(1)$ with high probability. These properties establish condition (c).

Appendix E: Auxiliary results for Theorem 2

In this appendix, we prove lemmas that were used in the proof of Theorem 2.

E.1. Proof of Lemma 2

Consider the subset $U := \{\theta \in \mathbb{R}^d \mid \|\theta_J\|_\infty \leq B\}$, and let ∂U denote its boundary. Recall our previous definition $B := C_1 \cdot n^{-1/2}$, where C_1 is a constant to be specified.

The remainder of our argument consists of showing that the objective function has a local minimum in the interior of the subset U, such that the prediction error on this local minimum is lower bounded by $\Omega(n^{-1/2})$. To this end, we define a vector:

$$\overline{\theta} \in \arg\min_{\theta \in \partial U} \frac{1}{n}\|y - X\theta\|_2^2 + \lambda\rho(\theta).$$

The minimizer $\overline{\theta}$ lies on the boundary of subset U so that its satisfies $\|\overline{\theta}_J\|_\infty = B$. We define another vector $\widetilde{\theta}$ belonging to the interior of subset U:

$$\text{Let} \quad \widetilde{\theta}_J := 0 \quad \text{and} \quad \widetilde{\theta}_{-J} := \overline{\theta}_{-J}.$$

The vectors $\overline{\theta}$ and $\widetilde{\theta}$ differ only at the coordinates indexed by set J. We make the following claim:

Claim 1. *If we choose $B = C_1 \cdot n^{-1/2}$ for a sufficiently small constant $C_1 \in (0, 1]$, then we have*

$$\frac{1}{n}\|y - X\widetilde{\theta}\|_2^2 \leq \frac{1}{n}\|y - X\overline{\theta}\|_2^2 + \frac{1}{n}.$$

We defer the proof of Claim 1 to the end of this section and focus on its consequences. If Claim 1 holds, then we have

$$\begin{aligned}\Big(\frac{1}{n}\|y - X\widetilde{\theta}\|_2^2 + \lambda\rho(\widetilde{\theta})\Big) - \Big(\frac{1}{n}\|y - X\overline{\theta}\|_2^2 + \lambda\rho(\overline{\theta})\Big) \\ = \frac{1}{n}\|y - X\widetilde{\theta}\|_2^2 - \frac{1}{n}\|y - X\overline{\theta}\|_2^2 - \lambda\rho_J(\overline{\theta}) \\ \leq \frac{1}{n} - \lambda\rho_J(\overline{\theta}) \leq \frac{1}{n} - \lambda\gamma,\end{aligned}$$

The last inequality follows since there is an index $j \in J$ such that $|\overline{\theta}_j| = B$, and as a consequence $\rho_J(\overline{\theta}) \geq \rho_j(B) \geq \min_{j\in J}\{\rho_j(B)\} = \gamma$.

Thus, as long as $\lambda\gamma > 1/n$, the objective function value of $\widetilde{\theta}$ is strictly smaller than any objective value on the boundary of set U. It implies that there is a local minimum inside the region U. Let $\widehat{\theta} \in U$ be such a local minimum. Let $\Delta := \widehat{\theta} - \theta^*$ be a shorthand notation. The prediction error on this local minimum is equal to

$$\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 = \frac{1}{n}\Big(\|X_J\Delta_J\|_2^2 + \|X_{-J}\Delta_{-J}\|_2^2 + 2\langle X_J\Delta_J, X_{-J}\Delta_{-J}\rangle\Big).$$

Note that the vectors $X_J\Delta_J$ and $X_{-J}\Delta_{-J}$ are in the column space of X_J and X_{-J}, respectively. By the condition (b) of Assumption A, there is a constant c such that

$$|2\langle X_J\Delta_J, X_{-J}\Delta_{-J}\rangle| \leq cn^{-1/4} \cdot \|X_J\Delta_J\|_2 \cdot \|X_{-J}\Delta_{-J}\|_2.$$

If $n \geq c^4$, then we have $cn^{-1/4} \leq 1$, and as a consequence,

$$\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 \geq \frac{1}{2n}\Big(\|X_J\Delta_J\|_2^2 + \|X_{-J}\Delta_{-J}\|_2^2\Big) \geq \Omega(n^{-1/2}) \cdot \|\Delta_J\|_2^2. \quad \text{(E.1)}$$

The last inequality holds since the condition (c) of Assumption A guarantees that the smallest singular value of matrix X_J is lower bounded by $\Omega(n^{-1/4})$.

Plugging in the definition of the vector Δ, we find that

$$\|\Delta_J\|_2 \geq \frac{1}{\sqrt{s}}\,\|\Delta_J\|_1 = \frac{1}{\sqrt{s}}\,\|\widehat{\theta}_J - \theta^*_J\|_1 \geq \frac{\|\theta^*_J\|_1 - \|\widehat{\theta}_J\|_1}{\sqrt{s}}.$$

Recall that $\|\theta^*_J\|_1 = C_2$ is a positive constant, and $\|\widehat{\theta}_J\|_1 \leq sB = \mathcal{O}(n^{-1/2})$. Thus, for any sample size n greater than a sufficiently large constant, the difference $\|\theta^*_J\|_1 - \|\widehat{\theta}_J\|_1$ is lower bounded by $C_2/2$, so that we have $\|\Delta_J\|_2 = \Omega(1)$. Combining this lower bound with inequality (E.1) completes the proof.

Proof of Claim 1: In order to prove the claim, we first prove that the norm of $X_{-J}\overline{\theta}_{-J}$ is bounded. More concretely, we prove that

$$\|X_{-J}\overline{\theta}_{-J}\|_2 = \mathcal{O}(n^{1/4} + r^{1/2}) = \mathcal{O}(r^{1/2}). \quad \text{(E.2)}$$

Recall that $\overline{\theta}$ minimizes the objective function on the boundary. If we replace the coordinates of $\overline{\theta}_{-J}$ by zeros, then the new vector stays in ∂U, so that the objective function won't decrease. As a consequence, we have

$$\frac{1}{n}\|y - X\overline{\theta}\|_2^2 + \lambda\rho_J(\overline{\theta}) + \lambda\rho_{-J}(\overline{\theta}) \leq \frac{1}{n}\|y - X_J\overline{\theta}_J - X_{-J}0\|_2^2 + \lambda\rho_J(\overline{\theta}) + \lambda\rho_{-J}(0).$$

Recalling that $\lambda\rho_{-J}(\overline{\theta}) \geq 0 = \lambda\rho_{-J}(0)$, this inequality ensures that

$$\|y - X\overline{\theta}\|_2 \leq \|y - X_J\overline{\theta}_J\|_2.$$

Note that $\Pi(\cdot)$ is the orthogonal projection operator onto the column space of matrix X. Since the vectors $y - X\overline{\theta}$ and $y - X_J\overline{\theta}_J$ differ only in the column space of matrix X, we have

$$\begin{aligned} & \|\Pi(y) - X_J\overline{\theta}_J - X_{-J}\overline{\theta}_{-J}\|_2 \leq \|\Pi(y) - X_J\overline{\theta}_J\|_2 \\ \Rightarrow\ & \|X_{-J}\overline{\theta}_{-J}\|_2 - \|\Pi(y) - X_J\overline{\theta}_J\|_2 \leq \|\Pi(y) - X_J\overline{\theta}_J\|_2 \\ \Rightarrow\ & \|X_{-J}\overline{\theta}_{-J}\|_2 \leq 2\|\Pi(y) - X_J\overline{\theta}_J\|_2. \end{aligned} \tag{E.3}$$

Plugging the equation $y = X\theta^* + w$ into inequality (E.3), and using the fact that $\|X_J\overline{\theta}_J\|_2 \leq \sum_{j\in J}\|X_j\overline{\theta}_j\|_2 \leq sBn^{1/2} = C_1 s$, we obtain

$$\|X_{-J}\overline{\theta}_{-J}\|_2 \leq 2\Big(\|\Pi(w)\|_2 + \|X\theta^*\|_2 + C_1 s\Big). \tag{E.4}$$

The event $\mathcal{E}_1$ implies $\|\Pi(w)\|_2 = (2r)^{1/2}$, which bounds the first term on the right-hand side. For the second term, we notice that it is bounded by $\mathcal{O}(n^{1/4})$, since

$$\|X\theta^*\|_2^2 = C_2^2 \cdot \|u^T X_J^T X_J u\|_2 \leq \|u\|_2 \cdot \|X_J^T X_J u\|_2 = \mathcal{O}(n^{1/2}). \tag{E.5}$$

In the above deductions, the first inequality uses the fact that $C_2 \leq 1$, and the final bound uses condition (a) from Assumption A. Combining inequalities (E.4) and (E.5) with the conditions $s = \mathcal{O}(1)$, $r = \Omega(n^{1/2})$ and $C_1 \leq 1$, we obtain the upper bound (E.2).

Given inequality (E.2), we are ready to prove Claim 1. We expand the difference of the prediction errors as follows:

$$\begin{aligned} \frac{1}{n}\|y - X\widetilde{\theta}\|_2^2 - \frac{1}{n}\|y - X\overline{\theta}\|_2^2 &= \frac{2}{n}\langle X_J\overline{\theta}_J, X\theta^* + w - X_{-J}\overline{\theta}_{-J}\rangle - \frac{1}{n}\|X\overline{\theta}_J\|_2^2 \\ &\leq \frac{2}{n}\langle X_J\overline{\theta}_J, X\theta^* + w - X_{-J}\overline{\theta}_{-J}\rangle. \end{aligned} \tag{E.6}$$

It suffices to show that the inner product on the right-hand side is bounded by $1/2$. Indeed, the inner product is the sum of three inner products: $\langle X_J\overline{\theta}_J, X\theta^*\rangle$, $\langle X_J\overline{\theta}_J, w\rangle$ and $\langle X_J\overline{\theta}_J, X_{-J}\theta^*_{-J}\rangle$. Their absolute values are bounded by:

$$|\langle X_J\overline{\theta}_J, X\theta^*\rangle| \leq |\|\overline{\theta}_J\|_2 \cdot \|X_J^T X_J u\|_2$$

$$\langle |X_J \overline{\theta}_J, w\rangle| \leq \|X_J \overline{\theta}_J\|_2 \cdot |\|\Pi_J(w)\|_2,$$
$$\langle X_J \overline{\theta}_J, X_{-J}\theta^*_{-J}|\rangle| \leq \mathcal{O}(r^{-1/2}) \cdot \|X_J \overline{\theta}_J\|_2 \cdot |\|X_{-J}\overline{\theta}_{-J}\|_2.$$

The operator $\Pi_J(\cdot)$ represents the projection onto the column space of matrix X_J. The third inequality holds since $X_J\overline{\theta}_J$ and $X_{-J}\theta^*_{-J}$ belong to the column space of matrix X_J and X_{-J} respectively, then the condition (b) of Assumption A implies the upper bound.

Recall that $\|\theta_J\|_2 \leq sB = C_1 sn^{-1/2}$ and $\|X_J\overline{\theta}_J\|_2 \leq sBn^{1/2} = C_1 s$. The condition (a) of Assumption A implies $\|X_J^T X_J u\|_2 = \mathcal{O}(n^{1/2})$, so that the first inner product term is upper bounded by:

$$|\langle X_J \overline{\theta}_J, X\theta^*\rangle| \leq C_1 \cdot \mathcal{O}(1).$$

The event $\mathcal{E}_1$ implies $\|\Pi_J(w)\|_2 \leq s^{1/2} = \mathcal{O}(1)$, which further implies that $|\langle X_J\overline{\theta}_J, w\rangle| \leq C_1 \cdot \mathcal{O}(1)$. Finally, inequality (E.2) upper bounds the term $\|X_{-J}\overline{\theta}_{-J}\|_2$ by $\mathcal{O}(r^{1/2})$, and hence

$$|\langle X_J \overline{\theta}_J, X_{-J}\theta^*_{-J}\rangle| \leq C_1 \cdot \mathcal{O}(1).$$

Consequently, as long as we choose a sufficiently small constant C_1, the inner product on the right-hand side of inequality (E.6) is at most $1/2$, which establishes the upper bound of Claim 1.

E.2. Proof of Lemma 3

In order to prove the lemma, we establish the existence of a specific regression vector $\overline{\theta}$ such that the norm of $y - X\overline{\theta}$ is substantially smaller than that of y. This property is formalized by the following claim:

Claim 2. *Under the event $\mathcal{E}_1 \cap \mathcal{E}_2$, there is a subset $I' \subseteq I$ of cardinality $m' \geq C_4 m$, as well as a vector $\overline{\theta} \in \mathbb{R}^d$ with $\|\overline{\theta}_{I'}\|_\infty \leq B$ and $\overline{\theta}_{-I'} = 0$ such that*

$$\|y - X\overline{\theta}\|_2^2 \leq \|y\|_2^2 - 2m'.$$

We defer the proof of Claim 2 to the end of this section. Assuming Claim 2, we use the properties of the vector $\overline{\theta}$ to show that the M-estimator will overfit the empirical loss whenever $\lambda\gamma \leq 1/n$.

Let $\widehat{\theta}$ be a global minimizer of the objective function. Using the non-negativity of the penalty term and the properties of the vectors $\widehat{\theta}$ and $\overline{\theta}$, we find that

$$\begin{aligned}\frac{1}{n}\|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{n}\|y - X\widehat{\theta}\|_2^2 + \lambda\rho(\widehat{\theta}) &\leq \frac{1}{n}\|y - X\overline{\theta}\|_2^2 + \lambda\rho(\overline{\theta}) \\ &\leq \frac{1}{n}\|y\|_2^2 - \frac{2m'}{n} + \lambda\rho(\overline{\theta}).\end{aligned}$$

Since $\|\overline{\theta}_{I'}\|_\infty \leq B$ and $\lambda\gamma \leq 1/n$, we have

$$\lambda\rho(\overline{\theta}) \leq \lambda \sum_{i \in I'} \rho_i(B) \leq m'\lambda\gamma \leq \frac{m'}{n}.$$

Combining the above two inequalities, we find that

$$\frac{1}{n}\|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{n}\|y\|_2^2 - \frac{m'}{n}. \tag{E.7}$$

Let $\Pi(\cdot)$ be the operator that projects onto the column space of matrix X. Since the vectors $y - X\widehat{\theta}$ and y only differ in the column space of X, inequality (E.7) implies $\|\Pi(y) - X\widehat{\theta}\|_2^2 \leq \|\Pi(y)\|_2^2 - m'$. Combining this bound with the triangle inequality, we have

$$\|X\widehat{\theta}\|_2 \geq \|\Pi(y)\|_2 - \|\Pi(y) - X\widehat{\theta}\|_2 \geq \frac{m'}{\|\Pi(y)\|_2 + \|\Pi(y) - X\widehat{\theta}\|_2} \geq \frac{m'}{2\|\Pi(y)\|_2}. \tag{E.8}$$

Since $y = X\theta^* + w$, the norm $\|\Pi(y)\|_2$ is bounded by $\|\Pi(w)\|_2 + \|X\theta^*\|_2$. For the first term of this upper bound, the event $\mathcal{E}_1$ implies $\|\Pi(w)\|_2 \leq 2r^{1/2}$. For the second term, we have shown that $\|X\theta^*\|_2 = \mathcal{O}(n^{1/4})$ in inequality (E.5). Thus, the norm $\|\Pi(y)\|_2$ is bounded by $\mathcal{O}(r^{1/2} + n^{1/4})$. From the scaling $r = \Omega(n^{1/2})$ given byAssumption A, it is bounded as $\|\Pi(y)\|_2 = \mathcal{O}(r^{1/2})$. As a consequence, we have

$$\|X\widehat{\theta}\|_2 \geq \Omega(r^{-1/2}) \cdot m' = \Omega(r^{1/2}) = \Omega(n^{1/4}).$$

In order to lower bound the prediction error, we observe that by triangle inequality, we have the lower bound

$$\|X(\widehat{\theta} - \theta^*)\|_2 \geq \|X\widehat{\theta}\|_2 - \|X\theta^*\|_2.$$

Since we have lower bounded $\|X\widehat{\theta}\|_2$, it suffices to upper bound $\|X\theta^*\|_2$. Using condition (a) of Assumption A, the square of this norm is at most

$$\|X\theta^*\|_2^2 = C_2^2 \cdot \|u^T X_J^T X_J u\|_2 \leq C_2^2 \cdot \|u\|_2 \cdot \|X_J^T X_J u\|_2 = C_2^2 \cdot \mathcal{O}(n^{1/2}).$$

Consequently, we have $\|X\theta^*\|_2 = C_2 \cdot \mathcal{O}(n^{1/2})$, where the constant hidden by the big-O notation is independent of C_2. The lower bound on $\|X\widehat{\theta}\|_2$, according to the above proof, is independent of C_2 as long as $C_2 \leq 1$. Thus, if we choose a sufficiently small C_2, then the term $\|X\theta^*\|_2$ will be smaller than half of $\|X\widehat{\theta}\|_2$, so that the prediction error $\|X(\widehat{\theta} - \theta^*)\|_2$ will be lower bounded as $\Omega(n^{1/4})$, which establishes the claim.

Proof of Claim 2: Let θ be an arbitrary vector satisfying the following conditions: $\|\theta_{I'}\|_\infty \leq B$ and $\theta_{-I'} = 0$. For any index set $I' \subseteq I$, let m' indicate the cardinality of I'. The largest singular value of matrix $X_{I'}$ is bounded by that of the matrix X_I. It is bounded by $Ln^{1/2}$, as shown in the proof of Theorem 2.

Plugging in these facts, we find that

$$\|y - X\theta\|_2^2 - \|y\|_2^2 = \|X_{I'}\theta_{I'}\|_2^2 - 2 \cdot \langle X_{I'}\theta_{I'}, y\rangle$$

$$\leq L^2 B^2 m' n - 2 \cdot \langle X_{I'} \theta_{I'}, y \rangle$$
$$= L C_1^2 m' - 2 \cdot \langle X_{I'} \theta_{I'}, y \rangle. \tag{E.9}$$

The inner product $\langle X_{I'} \theta_{I'}, y \rangle$ on the right-hand side is equal to $\sum_{i \in I'} \theta_i^T X_i^T y$. Now suppose that we set

$$\theta_i := B \cdot \operatorname{sign}(X_i^T w) \quad \text{for each } i \in I'.$$

With these choices, the inner product $\langle X_{I'} \theta_{I'}, y \rangle$ is equal to $\sum_{i \in I'} |X_i^T y|$. The event $\mathcal{E}_2$ implies that there exists an index set I' with cardinality $m' \geq C_4 m$, such that $\sum_{i \in I'} |X_i^T y| \geq C_3 m' n^{1/2}$. This implies that

$$\langle X_{I'} \theta_{I'}, y \rangle \geq B C_3 m' n^{1/2}.$$

Substituting the definitions of B and C_3 yields

$$\langle X_{I'} \theta_{I'}, y \rangle \geq m' + \frac{L C_1^2 m'}{2}.$$

By combining this lower bound with inequality (E.9), we find that

$$\|y - X\theta\|_2^2 - \|y\|_2^2 \leq -2m',$$

which completes the proof.

E.3. Proof of Lemma 4

Since $w \in \mathbb{R}^n$ is a Gaussian vector and Π is a degree r projection matrix, the squared norm $\|\Pi(w)\|_2^2$ follows a chi-square distribution with r degrees of freedom. Thus, from standard results on χ^2-concentration [13], we have

$$\mathbb{P}\Big(\|\Pi(w)\|_2^2 \leq 2r \Big) \geq 1 - e^{-\Omega(r)}.$$

Let $\Pi_\perp = \Pi - \Pi_J$ be the operator that projects onto the subspace that is orthogonal to the column space of X_J. We decompose the vector $\Pi(w)$ into $w_\parallel := \Pi_J(w)$ and $w_\perp := \Pi_\perp(w)$. By the property of the multivariate normal distribution, the vector $w_\parallel$ is independent of the vector $w_\perp$. Since the squared norm $\|w_\parallel\|_2^2$ satisfies a chi-square distribution with s degrees of freedom, it satisfies $\|w_\parallel\|_2 \leq s$ with probability at least $1/2$. Thus event $\mathcal{E}_1$ holds with probability at least $1/2 - e^{-\Omega(r)}$.

Assuming the random event $\|w_\parallel\|_2 \leq s$, we lower bound the probability of event $\mathcal{E}_2$. For each index $i \in I$, the inner product $X_i^T w$ can be decomposed by $X_i^T w = X_i^T w_\parallel + X_i^T w_\perp$. Since $I \cap J = \emptyset$, The vectors $w_\parallel$ and X_i belong to the column space of X_J and X_{-J} respectively, so that the condition (b) of Assumption A implies

$$|X_i^T w_\parallel| \leq \mathcal{O}(n^{-1/4}) \cdot \|X_i\|_2 \cdot \|w_\parallel\|_2.$$

Plugging in the upper bounds $\|X_i\|_2 \le n^{1/2}$ and $\|w_\|\|_2 \le s = \mathcal{O}(1)$, there is a constant D such that $|X_i^T w_\| | \le D n^{-1/4}$.

On the other hand, the second term $X_i^T w_\perp$ is a zero-mean Gaussian with variance $\|\Pi_\perp(X_i)\|_2^2$. The variance is equal to $\|X_i\|_2^2 - \|\Pi_J(X_i)\|_2^2$. By condition (c) in Assumption A, we have $\|X_i\|_2^2 = \Omega(n)$. On the other hand, by condition (b) in Assumption A, we have

$$\|\Pi_J(X_i)\|_2 = \Big\langle \frac{\Pi_J(X_i)}{\|\Pi_J(X_i)\|_2}, X_i \Big\rangle = \mathcal{O}(n^{-1/4}) \cdot \|X_i\|_2 = \mathcal{O}(n^{1/4}).$$

Combining the pieces yields

$$\|\Pi_\perp(X_i)\|_2^2 = \|X_i\|_2^2 - \|\Pi_J(X_i)\|_2^2 = \Omega(n) - \mathcal{O}(n^{1/2}).$$

Thus, if the sample size n is greater than a sufficiently large constant, then the variance of the random variable $X_i^T w_\perp$ is lower bounded by $\Omega(n)$.

Let ϱ_i denote the random event that $|X_i^T w_\perp| \ge C_3 n^{1/2} + D \cdot n^{1/4}$. Since $X_i^T w_\perp$ is a normal random variable with $\Omega(n)$ variance, the probability of ϱ_i is lower bounded by a positive constant q. If event ϱ_i holds, then as a consequence we have

$$|X_i^T w| \ge |X_i^T w_\perp| - |X_i^T w_\| | \ge C_3 n^{1/2}. \tag{E.10}$$

Since $\mathbb{P}(\varrho_i)$ is lower bounded by q, the expectation $\mathbb{E}[\sum_{i\in I} \mathbb{I}(\varrho_i^c)]$ is upper bounded by $(1-q)(m-1)$. By Markov's inequality, for any constant $\alpha > 0$, we have

$$\mathbb{P}\Big[\sum_{i\in I} \mathbb{I}(\varrho_i^c) \le \frac{(1-q)(m-1)}{\alpha} \Big] \ge 1 - \alpha.$$

Setting $\alpha := 1 - q/2$ implies that with probability at least $q/2$, the following inequality holds:

$$\sum_{i\in I} \mathbb{I}(\varrho_i^c) \le \frac{1-q}{1-q/2}(m-1) \quad \Leftrightarrow \quad \sum_{i\in I} \mathbb{I}(\varrho_i) \ge \frac{q}{2-q}(m-1). \tag{E.11}$$

If inequality (E.11) holds, then combining with the lower bound (E.10) implies that there are at least $\frac{q}{2-q}(m-1)$ indices in I such that $|X_i^T w| \ge C_3 n^{1/2}$. This implies that event $\mathcal{E}_2$ holds with constant $C_4 = \frac{q}{4-2q}$ for any $m \ge 2$.

In summary, conditioning on the event $\|w_\|\|_2 \le s$, inequality (E.11) holds with probability at least $q/2$, which is a sufficient condition for the event $\mathcal{E}_2$. Note that whether inequality (E.11) holds is determined by the random variable $w_\perp$, independent of $w_\|$. Putting all pieces together, the probability that event $\mathcal{E}_1 \cap \mathcal{E}_2$ holds is lower bounded by:

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) &= \mathbb{P}(\|\Pi(w)\|_2^2 \le 2r,\ \|w_\|\|_2 \le s,\ \mathcal{E}_2) \\ &\ge \mathbb{P}(\|w_\|\|_2 \le s,\ \mathcal{E}_2) - \mathbb{P}(\|\Pi(w)\|_2^2 > 2r)\end{aligned}$$

$$= \mathbb{P}(\|w_{\|}\|_2 \leq s)\, \mathbb{P}(\mathcal{E}_2 \mid \|w_{\|}\|_2 \leq s) - \mathbb{P}(\|\Pi(w)\|_2^2 > 2r)$$
$$\geq \frac{1}{2} \times \frac{q}{2} - e^{-\Omega(r)}.$$

If r is greater than a sufficiently large constant, then this probability is lower bounded by a positive constant $p := q/8$.

Appendix F: Proof of Lemma 6

Similar to the proof of Lemma 1, it is convenient to omit reference to the index i. We let u^t and u^* be shorthand notation for the sub-vectors $\theta^t_{2i-1,2i}$, and $\theta^*_{2i-1,2i}$, respectively. We introduce the normalized noise $\varepsilon := w_{2i-1,2i}/\sqrt{n}$. By our construction of the design matrix X and the update formula (3.7), the vector u^t satisfies the recursion

$$u^{t+1} \in \operatorname*{argmin}_{\|u-u^t\|_2 \leq \beta} \ell(u; \lambda), \tag{F.1a}$$

where $\beta := \|u^{t+1} - u^t\|_2 \leq \eta$ and the loss function takes the form

$$\ell(u; \lambda) := \underbrace{\|Au - Au^* - \varepsilon\|_2^2}_{:=T} + \lambda\rho_{2i-1}(u_1) + \lambda\rho_{2i}(u_2). \tag{F.1b}$$

This statement holds for each $i \in [n/2]$. Hence, it suffices to study the update formula (F.1a).

F.1. Proof of part (a)

For the case of $i = 1$, we assume that the event $\mathcal{E}_0 \cap \mathcal{E}_1$ holds. Consequently, we have $\max\{u_1^0, u_2^0\} \leq 0$, $\|\varepsilon\|_2 \leq B$ and $\lambda\gamma_1 > 2\sin^2(\alpha)R + 5B$. The corresponding regression vector is $u^* = (R/2, R/2)$. Let us define

$$b_1 \in \arg\max_{u \in (0,B]} \rho_{2i-1}(u) \quad \text{and} \quad b_2 \in \arg\max_{u \in (0,B]} \rho_{2i}(u).$$

Our assumption implies $\rho'_{2i-1}(b_1) \geq \gamma_1$ and $\rho'_{2i}(b_2) \geq \gamma_1$. We claim that

$$u_k^t \leq b_k \leq B \quad \text{for } k = 1, 2 \text{ and for all iterations } t = 0, 1, 2, \ldots. \tag{F.2}$$

If the claim is true, then we have

$$\begin{aligned}\|A\theta^t_{1:2} - A\theta^*_{1:2}\|_2^2 &= \cos^2(\alpha)(u_1^t - u_2^t)^2 + \sin^2(\alpha)(R - u_1^t - u_2^t)^2 \\ &\geq \sin^2(\alpha)(R - u_1^t - u_2^t)^2 \geq \sin^2(\alpha)(R - 2B)_+^2 \\ &\geq \frac{s_n \sigma R}{4\sqrt{n}},\end{aligned}$$

where the final inequality follows from substituting the definition of α, and using the fact that $2B \leq R/2$. Thus, it suffices to prove the claim (F.2).

We prove the claim (F.2) by induction on the iteration number t. It is clearly true for $t = 0$. Assume that the claim is true for a specific integer $t \geq 0$, we establish it for integer $t+1$. Our strategy is as follows: suppose that the vector u^{t+1} minimizes the function $\ell(u;\lambda)$ inside the ball $\{u : \|u - u^t\|_2 \leq \beta\}$. Then, the scalar u_1^{t+1} satisfies

$$u_1^{t+1} = \underset{x:\|(x,u_2^{t+1})-u^t\|_2\leq\beta}{\operatorname{argmin}} f(x) \quad \text{where } f(x) := \ell((x, u_2^{t+1}); \lambda).$$

Let us now calculate the generalized derivative [11] of the function f at u_1^{t+1}. It turns out that

$$\text{either} \quad u_1^{t+1} \leq u_1^t \leq b_1, \quad \text{or} \quad \partial f(u_1^{t+1}) \cap (-\infty, 0] \neq \emptyset. \tag{F.3}$$

Otherwise, there is a sufficiently small scalar $\delta > 0$ such that

$$\|(u_1^{t+1} - \delta, u_2^{t+1}) - u^t\|_2 \leq \beta \quad \text{and} \quad f(u_1^{t+1} - \delta) < f(u_1^{t+1}),$$

contradicting the fact that u_1^{t+1} is the minimum point. In statement (F.3), if the first condition is true, then we have $u_1^{t+1} \leq b_1$. We claim that the second condition also implies $u_1^{t+1} \leq b_1$.

In order to prove the claim, we proceed via proof by contradiction. Suppose to the contrary that $u_1^{t+1} > b_1$ and $\partial f(u_1^{t+1}) \cap (-\infty, 0] \neq \emptyset$. Note that the function f is differentiable for all $x > 0$. In particular, for $u_1^{t+1} > b_1$, we have

$$\begin{aligned} f'(u_1^{t+1}) &= \frac{\partial T}{\partial u_1}\Big|_{u_1=u_1^{t+1}} + \lambda\rho'_{2i-1}(u_1^{t+1}) \\ &= -2(\sin^2(\alpha)R + a_i^T\varepsilon) + 2u_1^{t+1} - 2(1 - 2\sin^2(\alpha))u_2^{t+1} + \lambda\rho'_{2i-1}(u_1^{t+1}), \end{aligned} \tag{F.4}$$

where we have introduced the convenient shorthand $a_i = (\cos(\alpha), \sin(\alpha))$. Now make note of the inequalities

$$1 - 2\sin^2(\alpha) \leq 1a_i^T\varepsilon \leq \|\varepsilon\|_2, \qquad u_1^{t+1} > b_1 \quad \text{and} \quad u_2^{t+1} \leq u_2^t + \beta \leq b_2 + \beta.$$

Using these facts, equation (F.4) implies that

$$f'(u_1^{t+1}) \geq -2\sin^2(\alpha)R - 2\|\varepsilon\|_2 + 2(b_1 - b_2 - \beta) + \lambda\rho'_{2i-1}(u_1^{t+1}). \tag{F.5}$$

Recall that $b_1, b_2 \in [0, B]$, and also using the fact that

$$\rho'_1(u_1^{t+1}) \geq \rho'_1(b_1) - \beta H \geq \gamma_1 - \beta H,$$

we find that

$$\begin{aligned} f'(u_1^{t+1}) &\geq -2\sin^2(\alpha)R - 2\|\varepsilon\|_2 - 2(B + \beta) + \lambda(\gamma_1 - \beta H). \\ &\geq -2\sin^2(\alpha)R - 2\|\varepsilon\|_2 - 3B + \lambda\gamma_1. \end{aligned} \tag{F.6}$$

Here the second inequality follows since $\beta \leq \eta \leq \min\{B, \frac{B}{\lambda H}\}$. Since the inequalities $\|\varepsilon\|_2 \leq B$ and $\lambda\gamma_1 > 2\sin^2(\alpha)R + 5B$ holds, inequality (F.6) implies that $f'(u_1^{t+1}) > 0$. But this conclusion contradicts the assumption that

$$\partial f(u_1^{t+1}) \cap (-\infty, 0] \neq \emptyset.$$

Thus, in both cases, we have $u_1^{t+1} \leq b_1$.

The upper bound for u_2^{t+1} can be proved following the same argument. Thus, we have completed the induction.

F.2. Proof of part (b)

Recall the definition (4.8b) of $\mathbb{S}_2$. For the case of $i = 2, 3, \ldots, n/2$, we assume that the event $i \in \mathbb{S}_2$ holds. Consequently, we have $\varepsilon_1 \geq \lambda\gamma_1 + B$ and $|\varepsilon_2| \leq B$ as well as our assumption that

$$\sup_{u \in (0,B]} \rho'_{2i-1}(u) = \gamma_i \leq \gamma_1.$$

The corresponding regression vector is $u^* = (0, 0)$. Let $\widehat{u}$ be the stationary point to which the sequence $\{u^t\}_{t=0}^{\infty}$ converges. We claim that

$$\cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 + \sin^2(\alpha)(\widehat{u}_1 + \widehat{u}_2)^2 \geq \frac{B^2}{20}. \tag{F.7}$$

If the claim is true, then by the definition of the loss function, we have

$$\|A\widehat{\theta}_{2i-1:2i} - A\theta^*_{2i-1:2i}\|_2^2 = \cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 + \sin^2(\alpha)(\widehat{u}_1 + \widehat{u}_2)^2 \geq \frac{B^2}{20}.$$

This completes the proof of part (b). Thus, it suffices to prove the claim (F.7).

In order to establish the claim (F.7), we notice that $\widehat{u}$ is a local minimum of the loss function $\ell(\cdot; \lambda)$. Define the functions

$$f_1(x) := \ell((x, \widehat{u}_2); \lambda), \quad \text{and} \quad f_2(x) := \ell((\widehat{u}_1, x); \lambda),$$

corrresponding to the coordinate functions in which one argument of $\ell(\cdot; \lambda)$ to be either $\widehat{u}_2$ or $\widehat{u}_1$. Since $\widehat{u}$ is a local minimum of the function $\ell(\cdot; \lambda)$, the scalar $\widehat{u}_1$ must be a local minimum of f_1, and the scalar $\widehat{u}_2$ must be a local minimum of f_2. Consequently, the zero vector must belong to the generalized derivative [11] of f_1 and f_2, which we write as $0 \in \partial f_1(\widehat{u}_1)$ and $0 \in \partial f_2(\widehat{u}_2)$. We use this fact to prove the claim (F.7).

Calculating the generalized derivatives of f_1 and f_2, we have

$$\frac{1}{2}\lambda g_1 = \cos(\alpha)\varepsilon_1 + \sin(\alpha)\varepsilon_2 - \widehat{u}_1 + (1 - 2\sin^2(\alpha))\widehat{u}_2 \quad \text{for some} \quad g_1 \in \partial\rho_{2i-1}(\widehat{u}_1). \tag{F.8}$$

$$\frac{1}{2}\lambda g_2 = -\cos(\alpha)\varepsilon_1 + \sin(\alpha)\varepsilon_2 - \widehat{u}_2 + (1-2\sin^2(\alpha))\widehat{u}_1 \quad \text{for some} \quad g_2 \in \partial\rho_{2i}(\widehat{u}_2). \tag{F.9}$$

We compare the signs of $\widehat{u}_1$ and $\widehat{u}_2$. If $\text{sign}(\widehat{u}_1) = \text{sign}(\widehat{u}_2)$, then by the definition of the penalty function, we have $\text{sign}(g_1) = \text{sign}(g_2)$. Let $\beta := \frac{g_1}{g_1+g_2}$. We multiply equation (F.8) by $1-\beta$ and multiply equation (F.9) by β, then subtract the first equation by the second. Doing so yields that

$$(\widehat{u}_1 - \widehat{u}_2) + 2\sin^2(\alpha)((1-\beta)\widehat{u}_2 - \beta\widehat{u}_1) = \cos(\alpha)\varepsilon_1 + \sin(\alpha)(1-2\beta)\varepsilon_2.$$

The absolute value of the left-hand side is upper bounded by $|\widehat{u}_1 - \widehat{u}_2| + 2\sin^2(\alpha)|(1-\beta)\widehat{u}_2 - \beta\widehat{u}_1|$. Since we have $\sin(\alpha) \le 1/2$, $\text{sign}(\widehat{u}_1) = \text{sign}(\widehat{u}_2)$ and $\beta \in [0,1]$, it is further upper bounded by $|\widehat{u}_1 - \widehat{u}_2| + \sin(\alpha)|\widehat{u}_1 + \widehat{u}_2|$. On the other hand, since $\varepsilon_1 \ge \lambda\gamma_1 + B$ and $|\varepsilon_2| \le B$ hold, the absolute value of the right-hand side is lower bounded by $(\cos(\alpha) - \sin(\alpha))B$. Putting the pieces together, we have

$$|\widehat{u}_1 - \widehat{u}_2| + \sin(\alpha)|\widehat{u}_1 + \widehat{u}_2| \ge (\cos(\alpha) - \sin(\alpha))B.$$

Combining with the relation $a^2 + b^2 \ge \frac{1}{2}(a+b)^2$, we obtain:

$$\begin{aligned}
&\cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 + \sin^2(\alpha)(\widehat{u}_1 + \widehat{u}_2)^2 \\
&\quad \ge \frac{1}{2}(\cos(\alpha)|\widehat{u}_1 - \widehat{u}_2| + \sin(\alpha)|\widehat{u}_1 + \widehat{u}_2|)^2 \\
&\quad \ge \frac{1}{2}\cos^2(\alpha)(\cos(\alpha) - \sin(\alpha))^2 B^2.
\end{aligned}$$

Then combining with the fact $\sin(\alpha) \le 1/2$, we obtain:

$$\cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 + \sin^2(\alpha)(\widehat{u}_1 + \widehat{u}_2)^2 > 0.05\, B^2. \tag{F.10}$$

Next, we consider the case when $\text{sign}(\widehat{u}_1) \ne \text{sign}(\widehat{u}_2)$. For this case, if the absolute value of $\widehat{u}_1$ is greater than B, then we have

$$\cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 \ge \cos^2(\alpha)(\widehat{u}_1)^2 > \frac{3}{4}B^2. \tag{F.11}$$

Otherwise, we assume $|\widehat{u}_1| \le B$, and consequently

$$|g_1| \le \sup_{u\in(0,B]} \rho'_{2i-1}(u) = \gamma_i \le \gamma_1.$$

With this inequality, equation (F.8) implies that

$$|\cos(\alpha)\varepsilon_1 + \sin(\alpha)\varepsilon_2 - \widehat{u}_1 + (1-2\sin^2(\alpha))\widehat{u}_2| \le \frac{\lambda\gamma_1}{2},$$

and consequently:

$$|(\widehat{u}_2 - \widehat{u}_1) - 2\sin^2(\alpha)\widehat{u}_2| \ge \cos(\alpha)\varepsilon_1 - \sin(\alpha)|\varepsilon_2| - \frac{\lambda\gamma_1}{2}.$$

Notice that the signs of $\widehat{u}_2 - \widehat{u}_1$ and $2\sin^2(\alpha)\widehat{u}_2$ are equal, and the former term has a greater absolute value than the later, whence we have the upper bound $|(\widehat{u}_2 - \widehat{u}_1) - 2\sin^2(\alpha)\widehat{u}_2| \leq |\widehat{u}_2 - \widehat{u}_1|$. As a result, we have:

$$|\widehat{u}_2 - \widehat{u}_1| \geq \cos(\alpha)\varepsilon_1 - \sin(\alpha)|\varepsilon_2| - \frac{\lambda\gamma_1}{2}.$$

In order to lower bound the right-hand side, we use the conditions $\varepsilon_1 \geq \lambda\gamma_1 + B$ and $|\varepsilon_2| \leq B$, and combine with the fact that $\sin(\alpha) \leq 1/2$. Doing so yields

$$\cos^2(\alpha)(\widehat{u}_1 - \widehat{u}_2)^2 \geq \cos^2(\alpha)(\cos(\alpha) - \sin(\alpha))^2 B^2 > 0.1\, B^2. \tag{F.12}$$

Combining inequalities (F.10), (F.11), and (F.12) completes the proof.

Acknowledgements

This work was partially supported by grants NSF grant DMS-1107000, NSF grant CIF-31712-23800, Air Force Office of Scientific Research Grant AFOSR-FA9550-14-1-0016, and by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-15-1-2670.

References

[1] Regularized least-squares regression using lasso or elastic net algorithms. http://www.mathworks.com/help/stats/lasso.html.

[2] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. MR2860324

[3] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1995.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. MR2061575

[6] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, 2011.

[7] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

[8] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[9] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, 2009. MR2543688

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

[11] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Wiley-Interscience, New York, 1983. MR0709590

[12] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23:552–581, 2017.

[13] S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003. MR1943859

[14] V. F. Dem'yanov and V. N. Malozemov. *Introduction to Minimax*. Courier Dover Publications, 1990.

[15] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

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

[17] R. Foygel and N. Srebro. Fast rate and optimistic rate for ℓ_1-regularized regression. Technical report, Toyoto Technological Institute, 2011. arXiv:1108.037v1.

[18] L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[19] D. Ge, Z. Wang, Y. Ye, and H. Yin. Strong NP-hardness result for regularized l_q-minimization problems with concave penalty functions. *arXiv preprint arXiv:1501.00622*, 2015.

[20] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. *GIST: General iterative shrinkage and thresholding for non-convex sparse learning*. Tsinghua University, 2013. URL http://www.public.asu.edu/~jye02/Software/GIST.

[21] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[22] M. I. Jordan. On statistics, computation and scalability. *Bernoulli*, 19:1378–1390, 2013.

[23] K. C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27(3):320–341, 1983.

[24] J. B. Lasserre. An explicit exact SDP relaxation for nonlinear 0-1 programs. In *Integer Programming and Combinatorial Optimization*, pages 293–303. Springer, 2001.

[25] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[26] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[27] R. Mifflin. *A Modification and an Extension of Lemaréchal's Algorithm for Nonsmooth Minimization*. Springer, 1982. MR0654692

[28] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[29] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[30] A. Nemirovski. Topics in non-parametric statistics. In P. Bernard, editor, *Ecole d'été de Probabilities de Saint-Flour XXVIII*, Lecture notes in Mathematics. Springer, 2000.

[31] M. Pilanci, M. J. Wainwright, and L. El Ghaoui. Sparse learning via Boolean relaxations. *Mathematical Programming*, 151:63–87, 2015.

[32] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 99:2241–2259, 2010.

[33] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[34] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv:1003.2990*, 2010.

[35] H. D. Sherali and W. P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3:411–430, 1990.

[36] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

[37] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[38] S. A. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[39] M. J. Wainwright. Constrained forms of statistical minimax: Computation, communication and privacy. In *Proceedings of the International Congress of Mathematicians*, Seoul, Korea, 2014.

[40] P. Wolfe. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable Optimization*, pages 145–173. Springer, 1975. MR0448896

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

[42] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *Proceedings of the Conference on Computational Learning Theory (COLT)*, 2014.

[43] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.