arXiv:2402.00382v1 [math.ST] 1 Feb 2024

On the design-dependent suboptimality of the Lasso

Reese Pathak[†] Cong Ma[◇]

[†] Department of Electrical Engineering and Computer Sciences, UC Berkeley
[◇] Department of Statistics, University of Chicago
pathakr@berkeley.edu, congm@uchicago.edu

Abstract

This paper investigates the effect of the design matrix on the ability (or inability) to estimate a sparse parameter in linear regression. More specifically, we characterize the optimal rate of estimation when the smallest singular value of the design matrix is bounded away from zero. In addition to this information-theoretic result, we provide and analyze a procedure which is simultaneously statistically optimal and computationally efficient, based on soft thresholding the ordinary least squares estimator. Most surprisingly, we show that the Lasso estimator—despite its widespread adoption for sparse linear regression—is provably minimax rate-suboptimal when the minimum singular value is small. We present a family of design matrices and sparse parameters for which we can guarantee that the Lasso with *any* choice of regularization parameter—including those which are data-dependent and randomized—would fail in the sense that its estimation rate is suboptimal by polynomial factors in the sample size. Our lower bound is strong enough to preclude the statistical optimality of all forms of the Lasso, including its highly popular penalized, norm-constrained, and cross-validated variants.

1 Introduction

In this paper, we consider the standard linear regression model

$$y = X\theta^\star + w, \tag{1}$$

where $\theta^\star \in \mathbf{R}^d$ is the unknown parameter, $X \in \mathbf{R}^{n\times d}$ is the design matrix, and $w \sim \mathsf{N}\left(0, \sigma^2 I_n\right)$ denotes the stochastic noise. Such linear regression models are pervasive in statistical analysis [19]. To improve model selection and estimation, it is often desirable to impose a *sparsity* assumption on $\theta^\star$—for instance, we might assume that $\theta^\star$ has few nonzero entries or that it has few large entries. This amounts to assuming that for some $p \in [0, 1]$

$$\|\theta^\star\|_p \leqslant R, \tag{2}$$

where $\|\cdot\|_p$ denotes the ℓ_p vector (quasi)norm, and $R > 0$ is the radius of the ℓ_p ball. There has been a flurry of research on this sparse linear regression model (1)-(2) over the last three decades; see the recent books [2, 14, 28, 10, 16] for an overview.

Comparatively less studied, is the effect of the design matrix X on the ability (or inability) to estimate $\theta^\star$ under the sparsity assumption. Intuitively, when X is "close to singular", we would expect that certain directions of $\theta^\star$ would be difficult to estimate. Therefore, in this paper we seek to determine the optimal rate of estimation when the *smallest singular value of X is bounded*. More precisely, we consider the following set of design matrices

$$\mathcal{X}_{n,d}(B) \coloneqq \left\{ X \in \mathbf{R}^{n\times d} : \frac{1}{n} X^\mathsf{T} X \succcurlyeq \frac{1}{B} I_d \right\}, \tag{3}$$

and aim to characterize the corresponding minimax rate of estimation

$$\mathfrak{M}_{n,d}(p,\sigma,R,B) \coloneqq \inf_{\widehat{\theta}} \sup_{\substack{X\in\mathcal{X}_{n,d}(B)\\ \|\theta^\star\|_p\leqslant R}} \mathbf{E}_{y\sim\mathsf{N}(X\theta^\star,\sigma^2 I_n)}\left[\left\|\widehat{\theta}-\theta^\star\right\|_2^2\right]. \tag{4}$$

1.1 A motivation from learning under covariate shift

Although it may seem a bit technical to focus on the dependence of the estimation error on the smallest singular value of the design matrix X, we would like to point out an additional motivation which is more practical and also motivates our problem formulation. This is the problem of linear regression in a well-specified model with covariate shift.

To begin with, recall that under random design, in the standard linear observational model (*i.e.*, without covariate shift) the statistician observes random covariate-label pairs of the form (x, y). Here, the covariate x is drawn from a distribution Q and the label y satisfies $\mathbf{E}[y \mid x] = x^\mathsf{T}\theta^\star$. The goal is to find an estimator $\widehat{\theta}$ that minimizes the out-of-sample excess risk, which takes the quadratic form $\mathbf{E}_{x\sim Q}[((\widehat{\theta}-\theta^\star)^\mathsf{T}x)^2]$. When the covariate distribution Q is isotropic, meaning that $\mathbf{E}_{x\sim Q}[xx^\mathsf{T}] = I$, the out-of-sample excess risk equals the squared ℓ_2 error $\|\widehat{\theta}-\theta^\star\|_2^2$.

Under covariate shift, there is a slight twist to the standard linear regression model previously described, where now the covariates x are drawn from a (source) distribution P that differs from the (target) distribution Q under which we would like to deploy our estimator. Assuming Q is isotropic, the goal is therefore still to minimize the out-of-sample excess risk under Q, which is $\|\theta^\star-\widehat{\theta}\|_2^2$. In general, if $P \neq Q$ and no additional assumptions are made, then learning with covariate shift is impossible in the sense that no estimator can be consistent for the optimal parameter $\theta^\star$. It is therefore common (and necessary) to impose some additional assumptions on the pair (P, Q) to facilitate learning. One popular assumption relates to the likelihood ratio between the source-target pair. It is common to assume that absolute continuity holds so that $Q \ll P$ and that the the likelihood ratio $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is uniformly bounded [21]. Interestingly, it is possible to show that if $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leqslant B$ for P-almost every x, then the semidefinite inequality

$$\mathbf{E}_{x\sim P}[xx^\mathsf{T}] \succcurlyeq \frac{1}{B}I \tag{5}$$

holds [21, 29]. Comparing the inequality (5) to our class $\mathcal{X}_{n,d}(B)$ as defined in display (3), we note that our setup can be regarded as a fixed-design variant of linear regression with covariate shift [20, 9, 30].

1.2 Determining the minimax rate of estimation

We begin with one of our main results, which precisely characterizes the (order-wise) minimax risk $\mathfrak{M}_{n,d}(p,\sigma,R,B)$ of estimating $\theta^\star$ under the sparsity constraint $\|\theta^\star\|_p \leqslant R$ and over the restricted design class $\mathcal{X}_{n,d}(B)$.

Theorem 1. *Let $n \geqslant d \geqslant 1$ and $\sigma, R, B > 0$ be given, and put $\tau_n^2 \coloneqq \frac{\sigma^2 B}{R^2 n}$. There exist two universal constants c_ℓ, c_u satisfying $0 < c_\ell < c_u < \infty$ such that*

(a) if $p \in (0,1]$ and $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$, then

$$c_\ell\, R^2\left(\tau_n^2 \log\left(\mathrm{e}d\tau_n^p\right)\right)^{1-p/2} \leqslant \mathfrak{M}_{n,d}(p,\sigma,R,B) \leqslant c_u\, R^2\left(\tau_n^2 \log\left(\mathrm{e}d\tau_n^p\right)\right)^{1-p/2}, \quad \text{and}$$

(b) if $p = 0$, we denote $s = R \in [d]$, and have

$$c_\ell \frac{\sigma^2 B}{n} s \log\left(\mathrm{e}\frac{d}{s}\right) \leqslant \mathfrak{M}_{n,d}(p, \sigma, s, B) \leqslant c_u \frac{\sigma^2 B}{n} s \log\left(\mathrm{e}\frac{d}{s}\right).$$

The proof of Theorem 1 relies on a reduction to the Gaussian sequence model [16], and is deferred to Section 3.1.

Several remarks on Theorem 1 are in order. The first observation is that Theorem 1 is sharp, apart from universal constants that do not depend on the tuple of problem parameters $(p, n, d, \sigma, s, R, B)$.

Secondly, it is worth commenting on the sample size restrictions in Theorem 1. For all $p \in [0, 1]$, we have assumed the "low-dimensional" setup that the number of observations n dominates the dimension d.[1] Note that this is necessary for the class of designs $\mathcal{X}_{n,d}(B)$ to be nonempty. On the other hand, for $p > 0$ we additionally require that the sample size is "moderate", i.e., $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$. We make this assumption so that we can focus on what we believe is the "interesting" regime: where neither ordinary least squares nor constant estimators are optimal. Indeed, when $n \geqslant d$ but $\tau_n^2 \geqslant \log^{-1}(\mathrm{e}d)$, it is easily verified that the optimal rate of estimation is on the order R^2; intuitively the effective noise level is too high and no estimator can dominate $\widehat{\theta} \equiv 0$ uniformly. On the other hand, when $n \geqslant d$ but $\tau_n^2 \leqslant d^{-2/p}$, then the ordinary least squares estimator is minimax optimal; intuitively, the noise level is sufficiently small such that there is, in the worst case, no need to shrink on the basis of the ℓ_p constraint to achieve the optimal rate.

Last but not least, as shown in Theorem 1, the optimal rate of estimation depends on the signal-to-noise ratio $\tau_n^{-2} = nR^2/(\sigma^2 B)$. As B increases, the design X becomes closer to singular, estimation of $\theta^\star$, as expected, becomes more challenging. The dependence of our result on B is exactly analagous to the impact of the likelihood ratio bound B appearing in the context of prior work on nonparametric regression under covariate shift [21].

1.3 A computationally efficient estimator

The optimal estimator underlying the proof of Theorem 1 requires computing a d-dimensional Gaussian integral, and therefore is not computationally efficient in general. In this section we propose an estimator that is both computationally efficient and statistically optimal, up to constant factors.

Our procedure is based on the soft thresholding operator: for $v \in \mathbf{R}^d$ and $\eta > 0$, we define

$$\mathsf{S}_\eta(v) \coloneqq \operatorname*{arg\,min}_{u \in \mathbf{R}^d} \left\{ \|u - v\|_2^2 + 2\eta \|u\|_1 \right\}.$$

Note that soft thresholding involves a coordinate-separable optimization problem and has an explicit representation, thus allowing efficient computation. Then we define the *soft thresholded ordinary least squares estimator*

$$\widehat{\theta}_\eta^{\mathsf{STOLS}}(X, y) \coloneqq \mathsf{S}_\eta\Big(\widehat{\theta}^{\mathsf{OLS}}(X, y)\Big), \tag{6}$$

where $\widehat{\theta}^{\mathsf{OLS}}(X, y)$ is the usual ordinary least squares estimate—equal to $(X^\mathsf{T} X)^{-1} X^\mathsf{T} y$ in our case. We have the following guarantees for its performance.

Theorem 2. *The soft thresholded ordinary least squares estimator (6) satisfies*

[1] Notably, this still allows n to be proportional to d, *e.g.*, we can tolerate $n = d$.

(a) in the case $p \in (0,1]$, for any $R > 0$, if $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$, then

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p \leqslant R} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 6\, R^2 \left(\tau_n^2 \log(\mathrm{e}d\tau_n^p)\right)^{1-p/2},$$

with the choice $\eta = \sqrt{2R^2\tau_n^2 \log(\mathrm{e}d\tau_n^p)}$, and

(b) in the case $p = 0$, for any $s \in [d]$,

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 6\, \frac{\sigma^2 B}{n} s \log\left(\mathrm{e}\frac{d}{s}\right),$$

with the choice $\eta = \sqrt{2\frac{\sigma^2 B}{n}\log(\frac{\mathrm{e}d}{s})}$.

The proof is presented in Section 3.2.

Comparing the guarantee in Theorem 2 to the minimax rate in Theorem 1, it is immediate to see that the soft thresholded ordinary least squares estimator is minimax optimal apart from constant factors.

Secondly, we would like to point out a (simple) modification to the soft thresholding ordinary least squares procedure that allows it to be adaptive to the hardness of the particular design matrix encountered. To achieve this, note that $X \in \mathcal{X}_{n,d}(\widehat{B})$ for $\widehat{B} \coloneqq \|\|(X^\mathsf{T}X)^{-1}\|\|_{\mathrm{op}}$. Therefore the results in Theorem 2 continue to hold with B replaced by (a possibly smaller) $\widehat{B}$, provided that the thresholding parameter η is properly adjusted. For instance, in the case with $p = 0$, we have

$$\sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\left[\|\widehat{\theta}_{\widehat{\eta}}^{\mathsf{STOLS}}(X,y) - \theta^\star\|_2^2\right] \leqslant 6\, \frac{\sigma^2 \widehat{B}}{n} s \log\left(\mathrm{e}\frac{d}{s}\right), \tag{7}$$

provided we take $\widehat{\eta} = \sqrt{2\frac{\sigma^2\widehat{B}}{n}\log(\frac{\mathrm{e}d}{s})}$.

Finally, we note that inspecting our proof, the upper bound for $\widehat{\theta}_{\widehat{\eta}}^{\mathsf{STOLS}}(X,y)$ also holds for a larger set of design matrices

$$\mathcal{X}_{n,d}^{\mathsf{diag}}(B) \coloneqq \left\{ X \in \mathbf{R}^{n\times d} : \left(\frac{1}{n}X^\mathsf{T}X\right)^{-1}_{ii} \leqslant B, \quad \text{for } 1 \leqslant i \leqslant d \right\}.$$

Since $\mathcal{X}_{n,d}(B) \subset \mathcal{X}_{n,d}^{\mathsf{diag}}(B)$, this means after combining the lower bounds in Theorem 1 with the guarantees in Theorem 2, we additionally have established the minimax rate over this larger family $\mathcal{X}_{n,d}^{\mathsf{diag}}(B)$.

1.4 Is Lasso optimal?

Arguably, the Lasso estimator [26] is the most widely used estimator for sparse linear regression. Given a regularization parameter $\lambda > 0$, the Lasso is defined to be

$$\widehat{\theta}_\lambda(X,y) \coloneqq \operatorname*{arg\,min}_{\vartheta \in \mathbf{R}^d} \left\{ \frac{1}{n}\|X\vartheta - y\|_2^2 + 2\lambda\|\vartheta\|_1 \right\}. \tag{8}$$

Surprisingly, we show that the Lasso estimator—despite its popularity—is provably *suboptimal* for estimating $\theta^\star$ when $B \gg 1$.

Corollary 1. *The Lasso is minimax suboptimal by polynomial factors in the sample size when $d = n$ and $B = \sqrt{n}$. More precisely,*

(a) if $p \in (0, 1]$, and $\sigma = R = 1$, then we have

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p \leqslant R} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim 1, \quad \text{and}$$

(b) if $p = 0$, and $\sigma = s = 1$, then we have

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim 1.$$

Corollary 1 is in fact a special case of a more general theorem (Theorem 3) to be provided later.

Applying Theorem 1 to the regime considered in Corollary 1, we obtain the optimal rate of estimation

$$\Big(\frac{\sqrt{1 + \log n}}{n^{1/4}}\Big)^{2-p} \ll 1, \qquad \text{for every } p \in [0, 1].$$

As shown, in the worst-case, the multiplicative gap between the performance of the Lasso and a minimax optimal estimator in this scaling regime is at least polynomial in the sample size. As a result, the Lasso is quite strikingly minimax *suboptimal* in this scaling regime.

In fact, the lower bound against Lasso in Corollary 1 is extremely strong. Note that in the lower bound, the Lasso is even allowed to leverage the oracle information $\theta^\star$ to calculate the optimal instance-dependent choice of the regularization parameter (*c.f.,* $\inf_{\lambda>0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2$). As a result, the lower bound applies to any estimator which can be written as the penalized Lasso estimator with data-dependent choice of penalty. Many typical Lasso-based estimators, such as the norm-constrained and cross-validated Lasso, can be written as the penalized Lasso with a data-dependent choice of the penalty parameter λ. For instance, in the case of the norm-constrained Lasso, this holds by convex duality. Thus, we can rule out the minimax optimality of any procedure of this type, in light of Corollary 1.

The separation between the oracle Lasso and the minimax optimal estimator can also be demonstrated in experiments, as shown below in Figure 1.

1.5 Connections to prior work

In this section, we draw connections and comparisons between our work and existing literature.

Linear regression with elliptical or no constraints. Without any parameter restrictions, the exact minimax rate for linear regression when error is measured in the ℓ_2 norm along with the dependence on the design matrix is known: it is given by $\sigma^2 \mathsf{Tr}((X^\mathsf{T} X)^{-1})$ [19]. These results match our intuition that as the smallest singular value of X decreases, the hardness of estimating $\theta^\star$ increases. It is also worth mentioning that the design matrix does not play a role, apart from being invertible, in determining the optimal rate for the in-sample prediction error. The rate is given uniformly by $\frac{\sigma^2 d}{n}$ when $n \geqslant d$ [15].

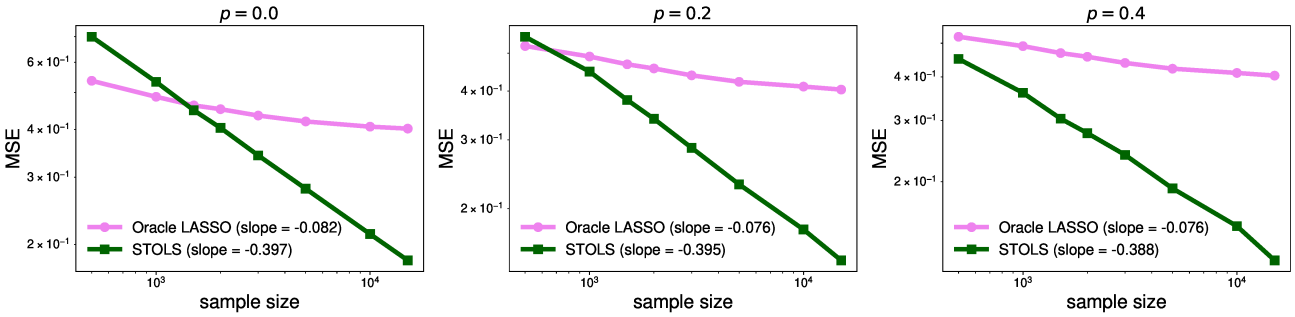

Figure 1. Numerical simulation demonstrating suboptimality of the Lasso, with the oracle choice of regularization $\widehat{\lambda} \in \arg\min_{\lambda>0} \|\widehat{\theta}_\lambda(X,y) - \theta^\star\|_2^2$ versus the soft thresholded ordinary least squares (STOLS) procedure as defined in display (6). We have simulated the lower bound instance from Corollary 1; for each sample size n, we simulate oracle Lasso and STOLS on a pair $(X_n, \theta_n^\star)$, with the dimension $d = n$ and lower singular value bound $B = \sqrt{n}$. Further details on the simulation are provided in Appendix B.

On the other hand, with ℓ_2- or elliptical parameter constraints the minimax rate in both fixed and random design was established in the recent paper [23]. Although that work shows the dependence on the design, the rate is not explicit and the achievable result requires potentially solving a semidefinite program. More explicit results were previously derived under proportional asymptotics in the paper [6], in the restricted setting of Gaussian isotropic random design. The author is able to establish the asymptotic minimaxity of a particular ridge regression estimator. These type of results are not immediately useful for our work, since they are based on linear shrinkage procedures which are known to be minimax-suboptimal even in orthogonal design, in the ℓ_p setting for $p < 2$ [7].

Gaussian sequence model and sparse linear regression. In the case of orthogonal design, *i.e.*, when $\frac{1}{n}X^\mathsf{T}X = I_d$, the minimax risk of estimating sparse $\theta^\star$ is known; see [7]. It can also be shown that Lasso, with optimally-tuned parameter, can achieve the (order-wise) optimal estimation rate [16]. This roughly corresponds to the case with $B = 1$ in our consideration. Our work generalizes this line of research in the sense that we characterize the optimal rate of estimation over the larger family of design matrices $\frac{1}{n}X^\mathsf{T}X \succcurlyeq \frac{1}{B}I_d$. In stark contrast to the Gaussian sequence model, Lasso is no longer optimal, even with the oracle knowledge of the true parameter.

Without assuming an orthogonal design, [3] provides a design-dependent lower bound in the exact sparsity case (*i.e.*, $p = 0$). The lower bound depends on the design through its Frobenius norm $\|X\|_\mathsf{F}^2$. Similarly, in the weak sparsity case, [24] provides lower bounds depending on the maximum column norm of the design matrix. However, matching upper bounds are not provided in this general design case. In contrast, using the minimum singular value of X (*c.f.,* the parameter B) allows us to obtain matching upper and lower bounds in sparse regression.

Suboptimality of Lasso. The suboptimality of Lasso for minimizing the *prediction* error has been noted in the case of exact sparsity (*i.e.*, $p = 0$). To our knowledge, previous studies required a carefully chosen design matrix which was highly-correlated. For instance, it was shown that for certain highly-correlated designs the Lasso can achieve only a slow rate $(1/\sqrt{n})$, while information-theoretically, the optimal rate is faster $(1/n)$; see for instance the papers [27, 17]. Additionally in the paper [4], the authors exhibit the failure of Lasso for a *fixed* regularization parameter, which does not necessarily rule out the optimality of other Lasso variants. Similarly, in the paper [11], it

is shown via a correlated design matrix and a 2-sparse vector, that the norm-constrained version of Lasso can only achieve a slow rate in terms of the prediction error. Again, this result does not rule out the optimality of other variants of the Lasso. In addition, in the paper [5], there is an example for which Lasso with any fixed (*i.e.*, independent from the observed data) choice of regularization would fail to achieve the optimal rate. Again, this fails to rule out data-dependent choices of regularization or other variants of the Lasso. In our work, we are able to rule out the optimality of the Lasso by considering a simple diagonal design matrix which exhibits no correlations among the columns. Nonetheless, for any $p \in [0, 1]$, we show that the Lasso will fall short of optimality by polynomial factors in the sample size. Our result also simultaneously rules out the optimality of constrained, penalized, and even data-dependent variants of the Lasso, in contrast to the literature described above.

Covariate shift. As mentioned previously, our work is also related to linear regression under covariate shift [20, 30, 9]. The statistical analysis of covariate shift, albeit with an asymptotic nature, dates back to the seminal work by Shimodaira [25]. Recently, nonasymptotic minimax analysis of covariate shift has gained much attention in unconstrained parametric models [12], nonparametric classification [18], and also nonparametric regression [22, 21, 29].

2 A closer look at the failure mode of Lasso

In this section, we take a closer look at the failure instance for Lasso. We will investigate the performance of the Lasso on diagonal design matrices $X_\alpha \in \mathbf{R}^{n \times d}$ which satisfy, when $d = 2k$,

$$\frac{1}{n} X_\alpha^\mathsf{T} X_\alpha = \begin{pmatrix} \frac{\alpha}{B} I_k & 0 \\ 0 & \frac{1}{B} I_k \end{pmatrix}.$$

Thus, this matrix has condition number α and satisfies $X_\alpha \in \mathcal{X}_{n,d}(B)$ for all $\alpha \geqslant 1$. As our proof of Theorem 1 reveals, from an information-theoretic perspective, the hardest design matrix X_α is with the choice $\alpha = 1$: when all directions have the worst possible signal-to-noise ratio. Strikingly, this is not the case for the Lasso: there are in fact choices of $\alpha \gg 1$ which are even harder for the Lasso.

Theorem 3. *Fix $n \geqslant d \geqslant 2$ and let $\sigma, B > 0$ be given. For $\alpha \geqslant 1$, on the diagonal design X_α,*

(a) if $p \in (0, 1]$ and $R > 0$, then there is a vector $\theta^\star \in \mathbf{R}^d$ such that $\|\theta^\star\|_p \leqslant R$ but

$$\mathbf{E}_{y \sim \mathsf{N}(X_\alpha \theta^\star, \sigma^2 I_n)} \Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 \Big] \geqslant \frac{9}{20000} \Big(\frac{\sigma^2 B d}{n\alpha} \wedge R^2 \Big(\frac{\sigma^2 B}{R^2 n} \alpha \Big)^{1 - p/2} \wedge R^2 \Big), \quad \text{and}$$

(b) if $p = 0$ and $s \in [d]$, then there is a vector $\theta^\star \in \mathbf{R}^d$ which is s-sparse but

$$\mathbf{E}_{y \sim \mathsf{N}(X_\alpha \theta^\star, \sigma^2 I_n)} \Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 \Big] \geqslant \frac{9}{20000} \Big(\frac{\sigma^2 B d}{n\alpha} \wedge \frac{\sigma^2 B s}{n} \alpha \Big).$$

The proof of Theorem 3 is presented in Section 3.3. We now make several comments on the implications of this result.

We emphasize that the dependence of the Lasso on the parameter α, which governs the condition number of the matrix X_α, is suboptimal, as revealed by Theorem 3. At a high-level, large α should

only make the ability to estimate $\theta^\star$ *easier*—it effectively increases the signal-to-noise ratio in certain directions. This can also be seen from Theorem 1: the conditioning of the design matrix *does not* enter into the worst-case rate of estimation when the bottom signular value of X is bounded. Nonetheless, Theorem 3 shows that the Lasso actually can suffer when the condition number α is large.

Proof of Corollary 1. We now complete the proof of Corollary 1 given Theorem 3.

Maximizing over the parameter $\alpha \geqslant 1$ appearing in our result, we can determine a particularly nasty configuration of the conditioning of the design matrix for the Lasso. Doing so, we find that for $p \in (0, 1]$ and $R > 0$ that

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_p \leqslant R} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim R^2 \bigg(\Big(\frac{\sigma^2 B\sqrt{d}}{R^2 n}\Big)^{\frac{4-2p}{4-p}} \wedge 1\bigg).$$

This is exhibited by considering the lower bound in Theorem 3 with the choice $\alpha^\star(p) = (\tau_n^2 d^{2/p})^{p/(4-p)}$. On the other hand, if $p = 0$, we have for $s \in [d]$ that

$$\sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\|\theta^\star\|_0 \leqslant s} \mathbf{E}\Big[\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2\Big] \gtrsim \frac{\sigma^2 B}{n}\sqrt{sd}$$

The righthand side above is exhibited by considering the lower bound with the choice $\alpha^\star(0) = \sqrt{d/s}$.

The proof is completed by setting $d = n$, $B = \sqrt{n}$, and $\sigma = R = 1$.

3 Proofs

In this section, we present the proofs for the main results of this paper. We start with introducing a few useful notations. For a positive integer k, we define $[k] \coloneqq \{1, \ldots, k\}$. For a real number x, we define $\lfloor x \rfloor$ to be the largest integer less than or equal to x and $\{x\}$ to be the fractional part of x.

3.1 Proof of Theorem 1

Our proof is based on a decision-theoretic reduction to the Gaussian sequence model. It holds in far greater generality, and so we actually prove a more general claim which could be of interest to other linear regression problems on other parameter spaces or with other loss functions.

To develop the claim, we first need to introduce notation. Let $\Theta \subset \mathbf{R}^d$ denote a parameter space, and let $\ell \colon \Theta \times \mathbf{R}^d \to \mathbf{R}$ be a given loss function. We define two minimax rates,

$$\mathfrak{M}_{\mathrm{seq}}(\Theta, \ell, \nu) \coloneqq \inf_{\widehat{\mu}} \sup_{\mu^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(\mu^\star, \nu^2 I_d)} \Big[\ell\big(\mu^\star, \widehat{\mu}(y)\big)\Big], \quad \text{and} \tag{9a}$$

$$\mathfrak{M}_{\mathrm{reg}}(\Theta, \ell, \sigma, B, n) \coloneqq \inf_{\widehat{\theta}} \sup_{X \in \mathcal{X}_{n,d}(B)} \sup_{\theta^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(X\theta^\star, \sigma^2 I_n)} \mathbf{E}\Big[\ell\big(\theta^\star, \widehat{\theta}(X, y)\big)\Big]. \tag{9b}$$

The definitions above correspond to the minimax rates of estimation over the ℓ_p ball of radius $R > 0$ in $\mathbf{R}^d$ for the Gaussian sequence model, in the case of definition (9a), and for n-sample linear

regression with B-bounded design, in the case of definition (9b). The infima range over measurable functions of the observation vector y, in both cases.

The main result we need is the following statistical reduction from linear regression to mean estimation in the Gaussian sequence model.

Proposition 1 (Reduction to sequence model)**.** *Fix $n, d \geqslant 1$ and $\sigma, B > 0$. Let $\Theta \subset \mathbf{R}^d$, and $\ell \colon \Theta \times \mathbf{R}^d \to \mathbf{R}$ be given. If $\ell(\theta, \cdot) \colon \mathbf{R}^d \to \mathbf{R}$ is a convex function for each $\theta \in \Theta$, then*

$$\mathfrak{M}_{\mathrm{reg}}(\Theta, \ell, \sigma, B, n) = \mathfrak{M}_{\mathrm{seq}}\left(\Theta, \ell, \sqrt{\frac{\sigma^2 B}{n}}\right).$$

Deferring the proof of Proposition 1 to Section 3.1.1 for the moment, we note that it immediately implies Theorem 1. Indeed, we set $\Theta = \Theta_{d,p}(R)$ and $\ell = \ell_{\mathrm{sq}}$ where

$$\Theta_{d,p}(R) \coloneqq \{\theta \in \mathbf{R}^d : \|\theta\|_p \leqslant R\}, \quad \text{and} \quad \ell_{\mathrm{sq}}(\theta, \widehat{\theta}) = \|\widehat{\theta} - \theta\|_2^2.$$

With these choices, we obtain

$$\mathfrak{M}_{n,d}(p, \sigma, R, B) = \mathfrak{M}_{\mathrm{reg}}(\Theta_{d,p}(R), \ell_{\mathrm{sq}}, \sigma, B, n) = \mathfrak{M}_{\mathrm{seq}}\left(\Theta_{d,p}, \ell_{\mathrm{sq}}, \sqrt{\frac{\sigma^2 B}{n}}\right),$$

where the final equality follows from Proposition 1. The righthand side then corresponds to estimation in the ℓ_2 norm over the Guassian sequence model with parameter space corresponding to an ℓ_p ball in $\mathbf{R}^d$, which is a well-studied problem [7, 1, 16]; we thus immediately obtain the result via classical results (for a precise statement with explicit constants, see Propositions 2 and 3 presented in Appendix A).

3.1.1 Proof of Proposition 1

We begin by lower bounding the regression minimax risk by the sequence model minimax risk. Indeed, let $X_\star$ be such that $\frac{1}{n} X_\star^{\mathsf{T}} X_\star = \frac{1}{B} I_d$. Then we have

$$\begin{aligned}
\mathfrak{M}_{\mathrm{reg}}(\Theta, \ell, \sigma, B, n) &\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta} \mathbf{E}_{y \sim \mathsf{N}(X_\star \theta^\star, \sigma^2 I_n)} \left[\ell\big(\theta^\star, \widehat{\theta}\big)\right] \\
&\geqslant \inf_{\widehat{\theta}} \sup_{\theta^\star \in \Theta} \mathbf{E}_{z \sim \mathsf{N}\left(\theta^\star, \frac{\sigma^2 B}{n} I_d\right)} \left[\ell\big(\theta^\star, \widehat{\theta}\big)\right] \\
&= \mathfrak{M}_{\mathrm{seq}}\left(\Theta, \ell_{\mathrm{sq}}, \sqrt{\frac{\sigma^2 B}{n}}\right).
\end{aligned}$$

The penultimate equality follows by noting that in the regression model, $\mathcal{P}_{X_\star} \coloneqq \{\mathsf{N}\left(X_\star \theta, \sigma^2 I_d\right) : \theta \in \Theta\}$, the ordinary least squares (OLS) estimate is a sufficient statistic. Therefore, by the Rao-Blackwell Theorem, there exists a minimax optimal estimator which is only a function of the OLS estimate. For any $\theta^\star$, the ordinary least squares estimator has the distribution, $\mathsf{N}\left(\theta^\star, \frac{\sigma^2 B}{n} I_d\right)$, which provides this equality.

We now turn to the upper bound. Let $\widehat{\theta}^{\mathsf{OLS}}(X, y) = (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} y$ denote the ordinary least squares estimate. For any estimator $\widehat{\mu}$, we define

$$\widetilde{\theta}(X, y) \coloneqq \mathbf{E}_{\xi \sim \mathsf{N}(0, W)} \left[\widehat{\mu}\big(\widehat{\theta}^{\mathsf{OLS}}(X, y) + \xi\big)\right], \qquad \text{where} \quad W = \frac{\sigma^2 B}{n} I_d - \sigma^2 (X^{\mathsf{T}} X)^{-1}. \tag{10}$$

Note that for any $X \in \mathcal{X}_{n,d}(B)$, we have $W \geqslant 0$. Additionally, by Jensen's inequality, for any $X \in \mathcal{X}_{n,d}(B)$, and any $\theta^\star \in \Theta$,

$$\begin{aligned}\mathbf{E}_{y\sim\mathsf{N}(X\theta^\star,\sigma^2 I_n)}\left[\ell\big(\theta^\star,\widetilde{\theta}(X,y)\big)\right] &\leqslant \mathbf{E}_{y\sim\mathsf{N}(X\theta^\star,\sigma^2 I_n)}\,\mathbf{E}_{\xi\sim\mathsf{N}(0,W)}\left[\ell\big(\theta^\star,\widehat{\mu}\big(\widehat{\theta}^{\mathsf{OLS}}(X,y)+\xi\big)\big)\right]\\ &= \mathbf{E}_{z\sim\mathsf{N}\left(\theta^\star,\frac{\sigma^2 B}{n}I_d\right)}\left[\ell\big(\theta^\star,\widehat{\mu}(z)\big)\right]\end{aligned}$$

Passing to the supremum over $\theta^\star \in \Theta$ on each side and then taking the infimum over measurable estimators, we immediately see that the above display implies

$$\mathfrak{M}_{\mathrm{reg}}(\Theta,\ell,\sigma,B,n) \leqslant \mathfrak{M}_{\mathrm{seq}}\left(\Theta,\ell,\sqrt{\frac{\sigma^2 B}{n}}\right),$$

as needed.

3.2 Proof of Theorem 2

We begin by bounding the risk for soft thresholding procedures, based on a rescaling and monotonicity argument and applying results from [16]. To state it, we need to define the quantitites

$$(\nu_n^\star)^2 \coloneqq \frac{\sigma^2 B}{n} \quad \text{and} \quad \rho_n(\theta,\eta) \coloneqq d\,(\nu_n^\star)^2 \mathrm{e}^{-(\eta/\nu_n^\star)^2/2} + \sum_{i=1}^d \theta_i^2 \wedge (\nu_n^\star)^2 + \sum_{i=1}^d \theta_i^2 \wedge \eta^2.$$

Then we have the following risk bound.

Lemma 1. *For any $\theta^\star \in \mathbf{R}^d$ and any $\eta > 0$ we have*

$$\sup_{X\in\mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y)-\theta^\star\|_2^2\right] \leqslant \rho_n(\theta^\star,\eta).$$

We now define for $\zeta > 0$ and a subset $\Theta \subset \mathbf{R}^d$,

$$T(\zeta,\Theta) \coloneqq \sup_{\theta^\star\in\Theta} \sum_{i=1}^d (\theta_i^\star)^2 \wedge \zeta^2.$$

Lemma 1 then yields with the choice $\eta = \gamma\nu_n^\star$ for some $\gamma \geqslant 1$ that

$$\sup_{\theta^\star\in\Theta}\sup_{X\in\mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y)-\theta^\star\|_2^2\right] \leqslant 3\left[d(\nu_n^\star)^2\mathrm{e}^{-\gamma^2/2} \vee T(\gamma\nu_n^\star,\Theta)\right]. \tag{11}$$

We bound the map T for the ℓ_p balls of interest. To state the bound, we use the shorthand Θ_p for the radius-R ℓ_p ball in $\mathbf{R}^d$ centered at the origin for $p \neq 0$, and for $p = 0$, the set of s-sparse vectors in $\mathbf{R}^d$, for $s \in [d]$.

Lemma 2. *Let $d \geqslant 1$ be fixed. We have the following relations:*

(a) in the case $p \in (0,1]$, we have for any $\zeta > 0$,

$$T(\zeta,\Theta_p) \leqslant R^2\left[\left(\frac{\zeta}{R}\right)^2 d \wedge \left(\frac{\zeta}{R}\right)^{2-p} \wedge 1\right]$$

for any $R > 0$, and

(b) in the case $p = 0$, we have for any $\zeta > 0$,

$$T(\zeta, \Theta_p) = \zeta^2 s,$$

for any $s \in [d]$.

To complete the argument, we now split into the two cases of hard and weak sparsity.

When $p = 0$: Combining inequality (11) together with Lemma 2, we find for $\eta = \gamma\nu_n^\star$, $\gamma \geqslant 1$, that

$$\sup_{\theta^\star \in \Theta} \sup_{X \in \mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X, y) - \theta^\star\|_2^2\right] \leqslant 3(\nu_n^\star)^2\left[d\mathrm{e}^{-\gamma^2/2} \vee \gamma^2 s\right] = 6(\nu_n^\star)^2 s \log\left(\mathrm{e}\frac{d}{s}\right),$$

where the last equality holds with $\gamma^2 = 2\log(\mathrm{e}d/s)$.

When $p \in (0, 1]$: Combining inequality (11) together with Lemma 2, we find for $\eta = \gamma\nu_n^\star$, $\gamma \geqslant 1$

$$\sup_{\theta^\star \in \Theta} \sup_{X \in \mathcal{X}_{n,d}(B)} \mathbf{E}\left[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X, y) - \theta^\star\|_2^2\right] \leqslant 3R^2\left[d\tau_n^2\mathrm{e}^{-\gamma^2/2} \vee \left(\gamma^{2-p}\tau_n^{2-p} \wedge 1\right)\right] \tag{12}$$

Above, we used $\tau_n^2 R^2 = (\nu_n^\star)^2$ and $\gamma^2\tau_n^2 d \geqslant \gamma^{2-p}\tau_n^{2-p}$, which holds since $\gamma \geqslant 1$ and $\tau_n^2 \geqslant d^{-2/p}$. If we take $\gamma^2 = 2\log(\mathrm{e}d\tau_n^p)$, then note $\gamma^2 \geqslant 1$ by $\tau_n^2 \geqslant d^{-2/p}$ and the term in brackets in inequality (12) satisfies

$$\begin{aligned} d\tau_n^2\mathrm{e}^{-\gamma^2/2} \vee \left(\gamma^{2-p}\tau_n^{2-p} \wedge 1\right) &= \frac{\tau_n^{2-p}}{\mathrm{e}} \vee \left((2\tau_n^2\log(\mathrm{e}d\tau_n^p))^{1-p/2} \wedge 1\right) \\ &\leqslant 2\left[\tau_n^{2-p} \vee \left((\tau_n^2\log(\mathrm{e}d\tau_n^p))^{1-p/2} \wedge 1\right)\right] \\ &= 2(\tau_n^2\log(\mathrm{e}d\tau_n^p))^{1-p/2}, \end{aligned}$$

which follows by $\tau_n^2 \in [d^{-2/p}, \log^{-1}(\mathrm{e}d)]$.

Thus, to complete the proof of Theorem 2 we only need to provide the proofs of the lemmas used above.

3.2.1 Proof of Lemma 1

Note that if $z = \widehat{\theta}^{\mathsf{OLS}}(X, y)$ then $\widehat{\theta}_\eta^{\mathsf{STOLS}}(X, y) = \mathsf{S}_\eta(z) = \mathsf{S}_\eta(\theta^\star + \xi)$ where $\xi \sim \mathsf{N}\left(0, \frac{\sigma^2}{n}\Sigma_n^{-1}\right)$, where we recall $\Sigma_n \coloneqq (1/n)X^\mathsf{T}X$. We now recall some classical results regarding the soft thresholding estimator. Let us write for $\lambda > 0$ and $\mu \in \mathbf{R}$,

$$\begin{aligned} r_\mathrm{S}(\lambda, \mu) &\coloneqq \mathbf{E}_{y \sim \mathsf{N}(\mu,1)}\left[\left(\mathsf{S}_\lambda(y) - \mu\right)^2\right], \quad \text{and,} \\ \widetilde{r}_\mathrm{S}(\lambda, \mu) &\coloneqq \mathrm{e}^{-\lambda^2/2} + \left(1 \wedge \mu^2\right) + \left(\lambda^2 \wedge \mu^2\right). \end{aligned}$$

Using $(a+b) \wedge c \leqslant a \wedge c + b \wedge c$ for nonnegative $a, b, c \geqslant 0$, Lemma 8.3 and the inequalities $r_\mathrm{S}(\lambda, 0) \leqslant 1 + \lambda^2$ and $r_\mathrm{S}(\lambda, 0) \leqslant \mathrm{e}^{-\lambda^2/2}$ on page 219 of the monograph [16], we find that $r_\mathrm{S}(\lambda, \mu) \leqslant \widetilde{r}_\mathrm{S}(\lambda, \mu)$. Define $\nu_i^2 \coloneqq \frac{\sigma^2}{n}(\Sigma_n^{-1})_{ii}$. Using the fact that $(\mathsf{S}_\eta(z))_i = \mathsf{S}_{\frac{\eta}{\nu_i}}\left(\frac{z_i}{\nu_i}\right)$ for $i \in [d]$, we obtain

$$\mathbf{E}[((\mathsf{S}_\eta(z))_i - \theta_i^\star)^2] = \nu_i^2 r_\mathrm{S}\left(\frac{\eta}{\nu_i}, \frac{\theta_i^\star}{\nu_i}\right) \leqslant \nu_i^2 \widetilde{r}_\mathrm{S}\left(\frac{\eta}{\nu_i}, \frac{\theta_i^\star}{\nu_i}\right).$$

Summing over the coordinates yields

$$\begin{aligned}\mathbf{E}\Big[\|\widehat{\theta}_\eta^{\mathsf{STOLS}}(X,y)-\theta^\star\|_2^2\Big] &\leqslant \sum_{i=1}^d \nu_i^2 \widetilde{r}_{\mathrm{S}}\Big(\frac{\eta}{\nu_i},\frac{\theta_i^\star}{\nu_i}\Big)\\ &= \sum_{i=1}^d \nu_i^2 \mathrm{e}^{-(\eta/\nu_i)^2/2} + \big((\theta_i^\star)^2 \wedge \nu_i^2\big) + \big((\theta_i^\star)^2 \wedge \eta^2\big)\\ &\leqslant \rho_n(\theta^\star,\eta),\end{aligned}$$

where the last inequality follows by noting that both $\nu \mapsto \nu^2 \mathrm{e}^{-(\eta/\nu)^2/2}$ and $\nu \mapsto \theta^2 \wedge \nu^2$ are non-decreasing functions of $\nu > 0$. Noting that this inequality holds uniformly on $X \in \mathcal{X}_{n,d}(B)$ and passing to the supremum yields the claim.

3.2.2 Proof of Lemma 2

The proof of claim (b) is immediate, so we focus on the case $p \in (0,1]$, $R > 0$. We consider three cases for the tuple (R,ζ,p,d). Combination of all three cases will yield the claim.

When $R \geqslant \zeta d^{1/p}$: Evidently, for each θ such that $\|\theta\|_p \leqslant R$, we have

$$\sum_{i=1}^d \theta_i^2 \wedge \zeta^2 \leqslant \zeta^2 d.$$

When $R \leqslant \zeta$: This case is immediate, since $\theta \in \Theta_p$ implies $\|\theta\|_2 \leqslant \|\theta\|_p \leqslant R \leqslant \zeta$.

When $\zeta \leqslant R \leqslant \zeta d^{1/p}$: In this case, by rescaling and putting $\varepsilon \coloneqq \frac{\zeta^2}{R^2}$, we have

$$\sup_{\|\theta\|_p \leqslant R} \sum_{i=1}^d \theta_i^2 \wedge \zeta^2 = R^2 \Bigg[\sup_{\lambda \in \Delta_d} \sum_{i:\lambda_i \geqslant \varepsilon^{p/2}} \varepsilon + \sum_{i:\lambda_i < \varepsilon^{p/2}} \lambda_i^{2/p}\Bigg] = \varepsilon R^2 \Big(\Big\lfloor \varepsilon^{-p/2}\Big\rfloor + \{\varepsilon^{-p/2}\}^{2/p}\Big)$$

where above Δ_d denotes the probability simplex in $\mathbf{R}^d$. Noting that $\varepsilon \leqslant 1$ and $p \leqslant 1$ we have

$$\Big(\Big\lfloor \varepsilon^{-p/2}\Big\rfloor + \{\varepsilon^{-p/2}\}^{2/p}\Big) \leqslant \varepsilon^{-p/2},$$

which in combination with the previous display shows that

$$\sup_{\|\theta\|_p \leqslant R} \sum_{i=1}^d \theta_i^2 \wedge \zeta^2 \leqslant R^2 \varepsilon^{1-p/2}.$$

To conclude, now note that $R^2\varepsilon^{1-p/2} = R^p \zeta^{2-p}$.

3.3 Proof of Theorem 3

Since X_α has nonzero entries only on the diagonal, we can derive an explicit representation of the Lasso estimate, as defined in display (8). To develop this, we first recall the notion of the *soft thresholding operator*, which is defined by a parameter $\eta > 0$ and then satisfies

$$\mathsf{S}_\eta(v) \coloneqq \arg\min_{u \in \mathbf{R}} \Big\{ (u-v)^2 + 2\eta|u| \Big\}.$$

We then start by stating the following lemma which is crucial for our analysis. It is a straightforward consequence of the observation that

$$\widehat{\theta}_\lambda(X_\alpha, y)_i = \begin{cases} \mathsf{S}_{\lambda B/\alpha}(z_i) & 1 \leqslant i \leqslant k \\ \mathsf{S}_{\lambda B}(z_i) & k+1 \leqslant i \leqslant d \end{cases}, \tag{13}$$

where we have defined the independent random variables $z_i \sim \mathsf{N}\left(\theta_i^\star, \frac{\sigma^2 B}{n\alpha}\right)$ if $i \leqslant k$ and $z_i \sim \mathsf{N}\left(\theta_i^\star, \frac{\sigma^2 B}{n}\right)$ otherwise.

Lemma 3. *Let $\theta^\star \in \mathbf{R}^d$. Then for the design matrix X_α, we have*

$$\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star\|_2^2 = \sum_{i=1}^k \Big(\mathsf{S}_{\lambda B/\alpha}(z_i) - \theta_i^\star\Big)^2 + \sum_{i=k+1}^d \Big(\mathsf{S}_{\lambda B}(z_i) - \theta_i^\star\Big)^2.$$

We will now focus on vectors $\theta^\star(\eta) = (0_k, \eta, 0_{d-2k})$, which are parameterized by $\eta \in \mathbf{R}^k$. For these vectors, we can further lower bound the best risk as

$$\inf_{\lambda > 0} \|\widehat{\theta}_\lambda(X_\alpha, y) - \theta^\star(\eta)\|_2^2 \geqslant T_1 \wedge T_2(\eta) \tag{14}$$

where we have defined

$$\overline{\lambda} = \sqrt{\frac{\sigma^2}{n}\frac{\alpha}{B}}, \quad T_1 \coloneqq \inf_{\lambda \leqslant \overline{\lambda}} \sum_{i=1}^k \Big(\mathsf{S}_{\lambda B/\alpha}(z_i)\Big)^2 \quad \text{and} \quad T_2(\eta) \coloneqq \inf_{\lambda \geqslant \overline{\lambda}} \sum_{i=k+1}^{2k} \Big(\mathsf{S}_{\lambda B}(z_i) - \eta_i\Big)^2.$$

We now move to lower bound T_1 and $T_2(\eta)$ by auxiliary, independent random variables.

Lemma 4 (Lower bound on T_1)**.** *Then, for any $\eta \in \mathbf{R}^k$ if $\theta^\star = \theta^\star(\eta)$, we have*

$$T_1 \geqslant \frac{1}{4}\frac{\sigma^2 B}{n\alpha} Z \qquad \textit{where} \quad Z \coloneqq \sum_{i=1}^k \mathbf{1}\Big\{\tfrac{|z_i|}{\sqrt{\sigma^2 B/(n\alpha)}} \geqslant 3/2\Big\}.$$

Lemma 5 (Lower bound on $T_2(\eta)$)**.** *Fix $\eta \in \mathbf{R}^k$ if $\theta^\star = \theta^\star(\eta)$, and suppose that*

$$0 \leqslant \eta_i \leqslant 2\sqrt{\frac{\sigma^2 B\alpha}{n}} \quad \textit{for all } i \in [k].$$

Then, we have

$$T_2(\eta) \geqslant \frac{1}{4}\sum_{i=1}^k \eta_i^2 W_i \quad \textit{where} \quad W_i \coloneqq \mathbf{1}\{z_{k+i} \leqslant \eta_i\}$$

Note that Z is distributed as a Binomial random variable: $Z \sim \mathsf{Bin}\,(k, p)$ where $p \coloneqq \mathbf{P}\{|\mathsf{N}\,(0,1)| \geqslant 3/2\}$. Similarly, W_i are Bernoulli: we have $W_i \sim \mathsf{Ber}\,(1/2)$.

Lower bound for $p > 0$: We consider two choices of η. First suppose that $4\tau_n^2\alpha \leqslant 1$. Then, we will consider $\eta = R\delta \mathbf{1}_\ell$ where

$$\delta \coloneqq 2\tau_n\sqrt{\alpha} \quad \text{and} \quad \ell \coloneqq k \wedge \lfloor \delta^{-p} \rfloor$$

For this choice of η we have by assumption that $\tau_n^2 \geqslant d^{-2/p}$ that $\ell \geqslant (1/2)\delta^{-p}$ and so

$$T_2(\eta) \geqslant \frac{1}{4}R^2\delta^2\ell\overline{W_\ell} \geqslant \frac{R^2}{8}\delta^{2-p}\overline{W_\ell} \geqslant \frac{R^2}{8}\Big(\tau_n^2\alpha\Big)^{1-p/2}\overline{W_\ell}$$

Above, $\overline{W_\ell} \coloneqq (1/\ell)\sum_{i=1}^{\ell} W_i$. On the other hand, if $4\tau_n^2\alpha \geqslant 1$, we take $\eta' = Re_1$, and we consequently obtain

$$T_2(\eta') \geqslant \frac{R^2}{4}W_1$$

Taking $\delta = 1/2$ in Lemma 4, let us define

$$c_1 \coloneqq \min_{1\leqslant\ell\leqslant k} \mathbf{P}\left\{\overline{W_\ell} \geqslant \frac{1}{2}\right\} \quad \text{and} \quad c_2 \coloneqq \mathbf{P}\Big\{Z \geqslant kp\Big\}.$$

Let us take

$$\theta_\alpha^\star \coloneqq \begin{cases} \theta^\star(\eta) & 4\tau_n^2\alpha \leqslant 1 \\ \theta^\star(\eta') & 4\tau_n^2\alpha > 1 \end{cases}.$$

Then combining Lemmas 4 and 5 and the lower bounds on $T_2(\eta), T_2(\eta')$ above, we see that

$$\mathbf{E}_{y\sim\mathsf{N}(X_\alpha\theta_\alpha^\star,\sigma^2 I_n)}\Big[\inf_{\lambda>0}\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta_\alpha^\star\|_2^2\Big] \geqslant \frac{c_2c_1p}{16}\Big(\frac{\sigma^2Bd}{n\alpha} \wedge R^2\Big(\frac{\sigma^2B}{R^2n}\alpha\Big)^{1-p/2} \wedge R^2\Big) \tag{15}$$

where above we have used $k \geqslant d/4$.

Lower bound when $p = 0$: In this case, we let $s' = s \wedge k$. Note that $s' \geqslant s/4$. We then set $\eta = 2\sqrt{\frac{\sigma^2B\alpha}{n}}\mathbf{1}_{s'}$, and this yields the lower bound

$$T_2(\eta) \geqslant \frac{1}{8}\frac{\sigma^2Bs}{n}\alpha\overline{W_{s'}}$$

In this case, we have, after combining this bound with the bound on T_1 that for $\theta_\alpha^\star \coloneqq \theta^\star(\eta)$ as defined above,

$$\mathbf{E}_{y\sim\mathsf{N}(X_\alpha\theta_\alpha^\star,\sigma^2 I_n)}\Big[\inf_{\lambda>0}\|\widehat{\theta}_\lambda(X_\alpha, y) - \theta_\alpha^\star\|_2^2\Big] \geqslant \frac{c_2c_1p}{16}\Big(\frac{\sigma^2Bd}{n\alpha} \wedge \frac{\sigma^2Bs}{n}\alpha\Big) \tag{16}$$

The proof of Theorem 3 is complete after combining inequalities (15) (16), and the following lemma.

Lemma 6. *The constant factor $c \coloneqq \frac{c_1c_2p}{16}$ is lower bounded as $c \geqslant \frac{9}{20000}$.*

We conclude this section by proving the lemmas above.

3.3.1 Proof of Lemma 4

For the first term, T_1, we note that for any $\lambda \leqslant \overline{\lambda}$ we evidently have for each $i \in [k]$ and for any $\zeta > 0$, that

$$\Big(\mathsf{S}_{\lambda B/\alpha}(z_i)\Big)^2 = (|z_i| - \lambda B/\alpha)_+^2 \geqslant (|z_i| - \overline{\lambda} B/\alpha)_+^2 \geqslant \zeta^2 \frac{\sigma^2 B}{n\alpha} \mathbf{1}\Big\{\frac{|z_i|}{\sqrt{\sigma^2 B/(n\alpha)}} \geqslant 1 + \zeta\Big\}$$

Summing over $i \in [k]$, and taking $\zeta = 1/2$, we thus obtain the claimed almost sure lower bound.

3.3.2 Proof of Lemma 5

Fix any i such that $1 \leqslant i \leqslant k$. For any fixed $\lambda \geqslant \overline{\lambda}$, note

$$S_{\lambda B}(z_{k+i}) \notin [\eta_i/2, 3\eta_i/2] \quad \text{implies} \quad \Big|S_{\lambda B}(z_{k+i}) - \eta_i\Big| \geqslant \frac{\eta_i}{2}.$$

Note that the condition $S_{\lambda B}(z_{k+i}) \notin [\eta_i/2, 3\eta_i/2]$ is equivalent to $z_{k+i} \notin [\eta_i/2 + \lambda B, 3\eta_i/2 + \lambda B]$, Therefore, if $z_{k+i} \leqslant \eta_i/2 + \overline{\lambda} B$, then then for all $\lambda \geqslant \overline{\lambda}$ we have $|S_{\lambda B}(z_{k+i}) - \eta_i| \geqslant \frac{\eta_i}{2}$. Equivalently, we have that

$$T_2 \geqslant \frac{1}{4} \sum_{i=1}^{k} \eta_i^2 \mathbf{1}\{z_{k+i} \leqslant \eta_i/2 + \overline{\lambda} B\} \geqslant \frac{1}{4} \sum_{i=1}^{k} \eta_i^2 \mathbf{1}\{z_{k+i} \leqslant \eta_i\} \tag{17}$$

The final relation uses the distribution of z_{k+i} and

$$\frac{-\eta_i/2 + \overline{\lambda} B}{\sqrt{\sigma^2 B/n}} = \sqrt{\alpha} - \frac{1}{2}\sqrt{\alpha}\sqrt{\frac{n\eta_i^2}{\alpha\sigma^2 B}} \geqslant 0$$

which holds by assumption that $\eta_i^2 \leqslant 4\frac{\sigma^2 B}{n}\alpha$.

3.3.3 Proof of Lemma 6

Evidently $c_1 \geqslant 1/2$ by symmetry. On the other hand, since $p \leqslant 1/2$, we have by anticoncentration results for Binomial random variables [13, Theorem 6.4] that $c_2 \geqslant p$. Therefore all together, $c \geqslant p^2/32$. Note that by standard lower bounds for the Gaussian tail [8, Theorem 1.2.6], we have

$$p \geqslant \frac{10}{27} \mathrm{e}^{-9/8} \geqslant \frac{3}{25},$$

which provides our claimed bound.

Acknowledgements

RP gratefully acknowledges partial support from the ARCS Foundation via the Berkeley Fellowship. Additionally, RP thanks Martin J. Wainwright for helpful comments, discussion, and references in the preparation of this manuscript. CM was partially supported by the National Science Foundation via grant DMS-2311127.

References

[1] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268, 2001.

[2] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[3] Emmanuel J Candes and Mark A Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[4] Emmanuel J. Candès and Yaniv Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.

[5] Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

[6] Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

[7] David L. Donoho and Iain M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields*, 99(2):277–303, 1994.

[8] Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019.

[9] Benjamin Eyre, Elliot Creager, David Madras, Vardan Papyan, and Richard Zemel. Out of the ordinary: Spectrally adapting regression for covariate shift. *arXiv preprint arXiv:2312.17463*, 2023.

[10] Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. CRC press, 2020.

[11] Rina Foygel and Nathan Srebro. Fast-rate and optimistic-rate error bounds for ℓ_1-regularized regression. *arXiv preprint arXiv:1108.0373*, 2011.

[12] Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

[13] Spencer Greenberg and Mehryar Mohri. Tight lower bound on the probability of a binomial exceeding its expectation. *Statist. Probab. Lett.*, 86:91–98, 2014.

[14] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the Lasso and generalizations*. CRC press, 2015.

[15] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

[16] Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Book manuscript, September 2019.

[17] Jonathan A. Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*, pages 550–561. 2022.

[18] Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

[19] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[20] Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pages 6164–6174. PMLR, 2021.

[21] Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

[22] Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR, 2022.

[23] Reese Pathak, Martin J. Wainwright, and Lin Xiao. Noisy recovery from random linear observations: Sharp minimax rates under elliptical constraints, 2023.

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

[25] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

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

[27] Sara van de Geer. On tight bounds for the Lasso. *J. Mach. Learn. Res.*, 19:Paper No. 46, 48, 2018.

[28] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[29] Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

[30] Xuhui Zhang, Jose Blanchet, Soumyadip Ghosh, and Mark S Squillante. A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pages 3794–3820. PMLR, 2022.

A Results in the Gaussian sequence model

In this section, we collect classical results regarding the nonasymptotic minimax rate of estimation for Gaussian sequence model over the origin-centered ℓ_p balls, $p \in [0, 1]$. All of the results in this section are based on the monograph [16]. We use the following notation to specify the minimax rate of interest,

$$\mathfrak{M}\Big(p, d, R, \varepsilon\Big) \coloneqq \inf_{\widehat{\mu}} \sup_{\substack{\mu^\star \in \mathbf{R}^d \\ \|\mu^\star\|_p \leqslant R}} \mathbf{E}_{y \sim \mathsf{N}(\mu^\star, \varepsilon^2)} \Big[\|\widehat{\mu}(y) - \mu^\star\|_2^2 \Big].$$

As usual, the infimum ranges over measurable estimators from the observation vector $y \in \mathbf{R}^d$ to an estimate $\widehat{\mu}(y) \in \mathbf{R}^d$. Throughout, we use the notation $\tau \coloneqq \frac{\varepsilon}{R}$ for the inverse signal-to-noise ratio.

Proposition 2 (Minimax rate of estimation when $0 < p \leqslant 1$)**.** *Fix an integer $d \geqslant 1$. Let $p \in (0, 1]$. If $R, \varepsilon > 0$ satisfy*

$$\frac{1}{d^{2/p}} \leqslant \tau^2 \leqslant \frac{1}{1 + \log d}, \qquad \text{where} \quad \tau = \frac{\varepsilon}{R},$$

then

$$\frac{7}{2000}\, R^2 \left(\tau^2 \log(\mathrm{e} d \tau^p)\right)^{1-p/2} \leqslant \mathfrak{M}(p, d, R, \varepsilon) \leqslant 1203\, R^2 \left(\tau^2 \log(\mathrm{e} d \tau^p)\right)^{1-p/2}.$$

The upper and lower bounds are taken from Theorem 11.7 in the monograph [16]. Although the constants are not made explicit in their theorem statement, the upper bound constant is obtained via their Theorem 11.4, setting their parameters as $\zeta = \frac{23}{4}, \gamma = 2\mathrm{e}, \beta = 0$. Similarly, the lower bound constant is implicit in their proof of Theorem 11.7.

We now turn to the minimax rate in the special case that $p = 0$.

Proposition 3 (Minimax rate of estimation when $p = 0$)**.** *Suppose that $d \geqslant 1$ and $s \in [d]$. Then for any $\varepsilon > 0$ we have*

$$\frac{3}{500}\, \varepsilon^2\, s \log\left(\mathrm{e} \frac{d}{s}\right) \leqslant \mathfrak{M}(p, d, s, \varepsilon) \leqslant 2\, \varepsilon^2\, s \log\left(\mathrm{e} \frac{d}{s}\right),$$

provided that $p = 0$.

The proof of the above claim is omitted as it is a straightforward combination of the standard minimax rate $\varepsilon^2 k$ for the unconstrained Normal location model in a k-dimensional problem (this provides a useful lower bound when $s \geqslant d/2$ or when $d = 1$) and the result in Proposition 8.20 in the monograph [16].

B Details for experiments in Figure 1

For each choice of p, we simulate the oracle Lasso and STOLS procedures on instances $(X_n, \theta_n^\star)$ indexed by the sample size $n \in \{1000, 2000, 3000, 5000, 10000, 15000\}$. The matrix $X_n \in \mathbf{R}^{n \times n}$ is block diagonal and given by

$$\frac{1}{\sqrt{n}} X_n = \begin{pmatrix} I_{n/2 \times n/2} & 0 \\ 0 & n^{-1/4} I_{n/2 \times n/2} \end{pmatrix},$$

When $p = 0$, $\theta_n^\star = 2e_{n/2+1}$ and when $p \neq 0$, $\theta_n^\star = e_{n/2+1}$. In the figures, we are plotting the average performance of the oracle Lasso and STOLS procedures, as measured by ℓ_2 error, when applied to the data (X_n, y), where $y \sim \mathsf{N}(X_n \theta_n^\star, I_n)$. The average is taken over 1000 trials for $n < 10,000$. In the case $n \geqslant 10,000$ due to memory constraints we only run 300 trials.

The STOLS procedure is implemented as described in Section 1.3. On the other hand, the oracle Lasso procedure is implemented by a slightly more involved procedure. Our goal is to compute

$$\widehat{\theta}_{\widehat{\lambda}}(X, y) \quad \text{where} \quad \widehat{\lambda} \in \operatorname*{arg\,min}_{\lambda > 0} \|\widehat{\theta}_\lambda(X, y) - \theta^\star\|_2^2,$$

where the Lasso is defined as in display (8). To do this, we can use the fact that the Lasso regularization path is piecewise linear. That is, there exist knot points $0 = \lambda_0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$ such that the knot points $\widehat{\theta}_i \coloneqq \widehat{\theta}_{\lambda_i}(X, y)$ satisfy $\|\theta_i\|_0 > \|\theta_{i+1}\|_0$. Moreover, we have

$$\{\widehat{\theta}_\lambda(X, y) : \lambda \in (\lambda_i, \lambda_{i+1})\} = \{\widehat{\theta}_i + \alpha(\widehat{\theta}_{i+1} - \widehat{\theta}_i) : \alpha \in (0, 1)\}.$$

That is, we can compute the set of Lasso solutions between the knot points by taking all convex combinations of knot points. Therefore the distance between the oracle Lasso solution and the true parameter $\theta^\star$ satisfies,

$$\|\widehat{\theta}_{\widehat{\lambda}}(X, y) - \theta^\star\|_2^2 = \min_i \min_{\alpha \in [0,1]} \|\widehat{\theta}_i + \alpha(\widehat{\theta}_{i+1} - \widehat{\theta}_i) - \theta^\star\|_2^2.$$

We are able to compute the righthand side of the display above by noting that for each i the inner minimization problem is a quadratic function of the univariate parameter α and therefore can be minimized explicitly.

Code: The code has been released at the following public repository,

`https://github.com/reesepathak/lowerlassosim`.

In particular, the repository contains a `Python` program which runs simulations of STOLS and oracle Lasso on the lower bound instance described above for any desired choice of $p \in [0, 1]$.