The Impact of Contamination and Correlated Design on the Lasso: an Average Case Analysis

Stanislav Minsker[a], Yiqiu Shen[b]

[a] *Department of Mathematics, University of Southern California, 90089, Los Angeles, United States*
[b] *Department of Data Sciences and Operations, University of Southern California, 90089, Los Angeles, United States*

Abstract

We study the prediction problem in the context of the high-dimensional linear regression model. We focus on the practically relevant framework where a fraction of the linear measurements is corrupted while the columns of the design matrix can be moderately correlated. Our findings suggest that for most sparse signals, the Lasso estimator admits strong performance guarantees under more easily verifiable and less stringent assumptions on the design matrix compared to much of the existing literature.

Keywords: Lasso, incoherence, robustness, norms of random submatrices

1. Introduction

Assume that we are given n linear measurements $\boldsymbol{Y} := (Y_1, Y_2, \ldots, Y_n)$ of the p-dimensional vector $\boldsymbol{\beta_*}$, namely, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta_*} + \sqrt{n}\boldsymbol{\theta_*} + \boldsymbol{\xi}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ a fixed "design matrix" such that its rows represent the measurement vectors, $\boldsymbol{\xi}$ is the noise vector with independent sub-Gaussian [1] coordinates with variance σ^2, and $\boldsymbol{\theta_*} \in \mathbb{R}^n$ is the vector of "outliers" that represent additive corruption. We are interested in the situation when (a) $\boldsymbol{\beta_*}$ is sparse, meaning that the cardinality s of its support $\mathcal{S} := \{j \in \{1, \ldots, p\} : \ \boldsymbol{\beta_{*,j}} \neq 0\}$ is much smaller than p, (b) the vector $\boldsymbol{\theta_*}$ of outliers is also sparse, meaning that the cardinality o of its support $\mathcal{O} \subseteq \{1, \ldots, n\}$ is much smaller than n, and (c) the matrix $\boldsymbol{X}$ is allowed to have moderately correlated columns.

Without loss of generality, we will assume that the columns of the design matrix are centered and have length $\sqrt{n}$. Let us note that the form of the corruption term $\sqrt{n}\boldsymbol{\theta}$ is chosen to be consistent with this requirement: indeed, letting $\boldsymbol{I_n}$ be the $n \times n$ identity matrix, we can equivalently express our model as $\boldsymbol{Y} = [\boldsymbol{X} \mid \sqrt{n}\boldsymbol{I_n}] \left(\boldsymbol{\beta_*}^T \, \boldsymbol{\theta_*}^T\right)^T + \boldsymbol{\xi}$. Here, $[\boldsymbol{X} \mid \sqrt{n}\boldsymbol{I_n}]$ is the augmented design matrix with unit columns and $\left(\boldsymbol{\beta_*}^T \, \boldsymbol{\theta_*}^T\right)^T$ is the sparse vector of dimension $p + n$. The celebrated Lasso estimator [11] is the

[1] We say that a random variable Z has sub-Gaussian distribution if $\mathbb{E}e^{tZ} \leq e^{C\mathrm{var}(Z)t^2}$ for all $t \in \mathbb{R}$ and some absolute constant $C > 0$.

solution of the convex optimization problem [2]

$$\hat{\boldsymbol{\beta}}_L \in \arg\min_{\boldsymbol{\beta},\boldsymbol{\theta}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\theta}\|_2^2 + \lambda\left(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\theta}\|_1\right). \tag{1}$$

Note that, unlike the classical sparse linear regression problem, we are only interested in estimating the vector $\boldsymbol{\beta}_*$ and the associated prediction risk $\mathcal{R}(\hat{\boldsymbol{\beta}_L}) := \frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}_*} - \boldsymbol{\beta}_L)\|_2^2$, while treating $\boldsymbol{\theta}_*$ as a nuisance parameter. The idea of relying on Lasso to construct robust regression estimators is not new and goes back at least to the works by Wright et al. [13], Nguyen and Tran [10]. A useful interpretation of Lasso in this context was given by Gannaz [5]: specifically, problem (1) is equivalent to the following one:

$$\hat{\boldsymbol{\beta}}_L \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{\lambda^2 \sum_{i=1}^{n} \Phi\left(\frac{y_i - \boldsymbol{X}^i\boldsymbol{\beta}}{\lambda\sqrt{n}}\right) + \lambda\|\boldsymbol{\beta}\|_1\right\}, \tag{2}$$

where $\Phi(u) = 0.5u^2 \wedge (|u| - 1/2)$ is Huber's loss function. The latter definition gives an intuitively plausible explanation of robustness inherent to $\hat{\boldsymbol{\beta}}_L$. Classical bounds for the Lasso [1] imply that the dependence of $\mathcal{R}(\hat{\boldsymbol{\beta}_L})$ on the number of outliers o is not worse than $O\left(\frac{o\log(n)}{n}\right)$. This bound was improved by Dalalyan and Thompson [4] to $O\left(\left(\frac{o\log(n)}{n}\right)^2\right)$. This aligns with the minimax rate derived by Gao [6], up to the $\log^2(n)$ factor. Recently, it was shown by Minsker et al. [9] that the dependence on the contamination proportion can be further improved to $\left(\frac{o\log(n/o)}{n}\right)^2$ if the Lasso is replaced by the square-root Slope estimator. The approach taken in the previous papers yields guarantees that are uniform with respect to the underlying signal and outlier vectors $\boldsymbol{\beta}_*$ and $\boldsymbol{\theta}_*$, but the price one has to pay are the strict conditions on the design matrix $\boldsymbol{X}$. For instance, to satisfy theses requirements, the prior works have assumed that $\boldsymbol{X}$ has sub-Gaussian rows with covariance structure that satisfies a version of the restricted eigenvalue condition [1]. It is well known that conditions of this type are notoriously hard to verify. Moreover, such assumptions often do not represent the design matrices encountered in applications. An alternative framework for the classical Lasso estimator was proposed and analyzed in [3]: assuming that the matrix $\boldsymbol{X}$ satisfies only a mild and easily verifiable "coherence property," the authors showed that $\mathcal{R}(\hat{\boldsymbol{\beta}}_L)$ admits optimal bounds for *most* (but not all) sparse vectors $\boldsymbol{\beta}_*$, where "most" is understood with respect to the uniform distribution over all choices of supports of given cardinality and all sign patterns of $\boldsymbol{\beta}_*$. In other words, Lasso works well for most "typical," or "average" vectors $\boldsymbol{\beta}_*$, a notion alluded to in the title of the paper. We find this framework particularly appealing for statistical and machine learning applications where the design matrix $\boldsymbol{X}$ is predetermined and often has correlated columns. Further theoretical and empirical evidence for good predictive performance of Lasso in the presence of strong correlations among the features was given by [7, 2]. This motivates our **main goal:**

[2] For simplicity, we assume that the coefficients of β and θ are penalized at the same level λ, although it is possible to introduce two regularization parameters λ_s and λ_o; all our arguments are valid in this case as well, and result in better log-factors.

to understand the instances when robust Lasso (1) *is expected to perform well under weak, verifiable assumptions on the design matrix.* To achieve this goal, we extend the framework proposed by [3] to the contamination framework discussed above.

1.1. Notation

Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\|\boldsymbol{X}\|$ will stand for its spectral norm, $\|\boldsymbol{X}\|_{\infty,2}$ denotes the maximum ℓ_2 norm of a row and $\|\boldsymbol{X}\|_{\max}$ – the largest in absolute value entry of $\boldsymbol{X}$. We denote by $\boldsymbol{X}_j$ and $\boldsymbol{X}^i$ the j-th column and the i-th row of $\boldsymbol{X}$ respectively, while $\boldsymbol{X}_{\mathcal{S}}$, $\boldsymbol{X}^{\mathcal{O}}$ refer to the submatrices obtained by selecting the columns/rows of $\boldsymbol{X}$ indexed by $\mathcal{S} \subseteq [p] := \{1, 2, \dots, p\}$, $\mathcal{O} \subseteq [n]$ respectively. Given a vector $\boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{v}\|$ stands for its ℓ_2 norm and $\|\boldsymbol{v}\|_\infty$ - for its sup-norm. Finally, given $a, b \in \mathbb{R}$, $a \wedge b$ stands for $\min(a, b)$ and $a \vee b$ for $\max(a, b)$.

2. Main results

In this section, we state and discuss our main results. We start with the key definitions and assumptions.

Definition 2.1. Given a matrix $\boldsymbol{X}$ with centered columns, define the **coherence** of $\boldsymbol{X}$ as is maximum correlation between the columns of $\boldsymbol{X}$:

$$\mu(\boldsymbol{X}) = \max_{1\le i<j\le p} \frac{|\langle \boldsymbol{X}_i, \boldsymbol{X}_j\rangle|}{\|\boldsymbol{X}_i\|\|\boldsymbol{X}_j\|}.$$

A matrix $\boldsymbol{X}$ is said to obey the **coherence property** $\mathcal{C}(A_0)$ if $\mu(\boldsymbol{X}) \le A_0 \cdot (\log(p))^{-1}$ for some positive constant A_0.

Note that the property $\mathcal{C}(A_0)$ can be easily verified numerically for a given matrix $\boldsymbol{X}$.

Definition 2.2. A coefficient vector $\boldsymbol{\beta}$ follows a **generic s-sparse model** if

- it support $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta})$ is selected uniformly at random from $[p]$;
- conditionally on $\mathcal{S}$, the signs of individual entries of $\boldsymbol{\beta}$ are independent and uniformly distributed over $\{+1, -1\}$.

Throughout this paper, we make the following assumptions:

Assumption 2.1. $\boldsymbol{X}$ satisfies coherence property $\mathcal{C}(A_0)$. In addition, $\frac{\|\boldsymbol{X}\|_{\max}}{\sqrt{n}} \le A_0 \cdot (\log(p))^{-1}$. These two conditions are equivalent to the requirement that the augmented design matrix $\boldsymbol{M} := [\boldsymbol{X} \mid \boldsymbol{I}_n]$ satisfies $\mathcal{C}(A_0)$.

Assumption 2.2. The vectors $\boldsymbol{\beta}_*$ and $\boldsymbol{\theta}_*$ follow a generic s-sparse model and a generic o-sparse model respectively.

We refer the reader to [3] for an in-depth discussion of the implications of this assumption. We will only remark here that we do not suppose that the randomness is inherent to the signs or the support of $\boldsymbol{\beta}_*$ and $\boldsymbol{\theta}_*$, rather it is just a tool that allows us to describe the “typical” scenario.

Assumption 2.3. The inequality $\sqrt{\frac{s}{p}}\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\| + \sqrt{\frac{o}{n}} < \frac{c}{\sqrt{\log(p)}}$ is satisfied for a sufficiently small absolute constant $c > 0$.

While assumption 2.3 can not be verified directly since s and o are unknown, it gives a good indication of the sparsity and contamination levels that can be tackled for a given design matrix $\boldsymbol{X}$. To get a general idea about the restrictiveness of the last assumption, assume that $\boldsymbol{X}$ has i.i.d. standard normal entries, whence $\|\boldsymbol{X}\|$ is of order $\sqrt{p}+\sqrt{n}$ and $\sqrt{\frac{s}{p}}\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\| \approx \sqrt{\frac{s}{n}}$. Therefore, if s/n is small, $\boldsymbol{X}$ can have much larger norm that a typical matrix with i.i.d. entries. More generally, if the rows of $\boldsymbol{X}$ are sub-Gaussian with an arbitrary covariance matrix Σ such that $\Sigma_{j,j}=1$ for all j, then $\|\boldsymbol{X}\| \le C\|\Sigma\|^{1/2}\sqrt{n}+\sqrt{p}$ with high probability [this can be easily deduced from results by 8, 14]. These examples show that s and o can be as large as $O(n/\log p)$.

2.1. Noiseless model

Our first result applies to the simpler framework where the dense noise $\boldsymbol{\xi}$ is absent:

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}_*+\sqrt{n}\boldsymbol{\theta}_*. \tag{3}$$

In this case, we replace Lasso with the following constrained ℓ_1-minimization problem:

$$\begin{aligned} \min_{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p,\bar{\boldsymbol{\theta}}\in\mathbb{R}^n} \quad & \|\bar{\boldsymbol{\beta}}\|_1+\|\bar{\boldsymbol{\theta}}\|_1 \\ \text{subject to} \quad & \boldsymbol{Y}=\boldsymbol{X}\bar{\boldsymbol{\beta}}+\sqrt{n}\bar{\boldsymbol{\theta}}. \end{aligned} \tag{4}$$

Theorem 2.1. *There exist absolute constants $C_1, c_2>0$ with the following property: with probability at least $1-C_1p^{-c_2}$, $(\boldsymbol{\beta}_*,\boldsymbol{\theta}_*)$ is the unique solution of the problem* (4).

The proof of this result, formally given in section E.1 of the supplement, is based on the analysis of random matrices obtained by sampling the columns of $\boldsymbol{X}$. Our main technical contribution is the extension of the tools introduced in [12] to the case of non-uniform sampling of columns.

2.2. The general model

Next, we consider the general model $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}_*+\sqrt{n}\boldsymbol{\theta}_*+\boldsymbol{\xi}$ and the estimator (1). Note that, without loss of generality, we can and will assume that $\boldsymbol{\theta}_{*,j}\neq 0 \iff \boldsymbol{\xi}_{j}=0$ (to see this, note that we can redefine the vector of outliers via $\boldsymbol{\theta}'_{*,j}:=\boldsymbol{\theta}_{*,j}+\boldsymbol{\xi}_{j}/\sqrt{n}$ if both are non-zero). The following is our most general result.

Theorem 2.2. *Suppose that assumptions* (2.1), (2.2) *and* (2.3) *hold with $A_0\le 1/16$. Then there exist absolute constants $C_j, c_j>0,\ j=1,\dots,3$ such that*

$$\begin{aligned} \frac{1}{n}\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_L-\boldsymbol{\beta}_*\right)\right\|_2^2 \le C_1\sigma^2\frac{s\log(p)}{n} \\ +C_2\sigma^2\left(\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max}\frac{(s+o)^2\log(p)}{n}\wedge\frac{o\log(p)}{n}\right) \end{aligned} \tag{5}$$

with probability at least $1-C_3p^{1-c_3/A_0}$ whenever $\lambda\ge 4\sqrt{2(\log(p)+\log(n))/n}$.

Remark 2.3. The constant c_3 in the statement above can be chosen to satisfy $c_3\ge 1/4$, whence $1-c_3/A_0\le -3$.

It is easy to see that the quantity $\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max}\frac{(s+o)^2\log(p)}{n}$ is the leading term controlling the dependence of the predictio risk on o whenever $\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max} \ll \frac{1}{s\vee o}$. In the best case scenario, $\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max}$ is of order $O(n^{-1/2})$, whence the necessary conditions for our bound to yield improvements over the standard Lasso analysis take the form $\max(s,o) \ll \sqrt{n}$. The situation can be significantly better if the additive dense noise $\boldsymbol{\xi}$ is "small" compared to the magnitude of the outliers $\boldsymbol{\theta}_*$ and the regression coefficients $\boldsymbol{\beta}_*$. We describe this favorable scenario next.

Theorem 2.4. *Suppose that assumptions* (2.1)*,* (2.2) *and* (2.3) *hold. In addition, suppose that the maximal standard deviation σ of the additive noise vector satisfies*

$$\sigma \le \frac{2}{35}\sqrt{\frac{n}{\log(pn)}}\min\left(\min_{i\in\mathcal{S}}|\boldsymbol{\beta}_{*i}|, \min_{i\in\mathcal{O}}|\boldsymbol{\theta}_{*i}|\right). \tag{6}$$

Then, whenever $\lambda \ge 4\sqrt{2(\log(p)+\log(n))/n}$, the estimator (1) *satisfies the inequality*

$$\frac{1}{n}\left\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_*\right)\right\|_2 \le C_1\sigma^2\frac{s\log(p)}{n} \tag{7}$$

with probability at least $1 - C_2 p^{1-c/A_0}$ for absolute constants $C_1, C_2 > 0$ and $c \ge 1/4$.

In other words, the prediction error *does not depend on the number of outliers as long as the regression coefficients and these outliers are large relative to the typical magnitude of the dense additive noise*. Let us remark that condition 6 does not preclude the coefficients of $\boldsymbol{\theta}_*$ from being small: indeed, in the generic o-sparse mode, $\boldsymbol{\theta}_{*j}$ is a sub-Gaussian random variable for each j with standard deviation $\left|\boldsymbol{\theta}_{*j}\right|$. Therefore, if $\sqrt{n}\left|\boldsymbol{\theta}_{*j}\right|$ is small, we can simply treat it as an element of the vector $\boldsymbol{\xi}$. In other words, condition 6 can be viewed as a requirement on the existence of a partition of the outliers into sets of elements with "large" and "small" magnitude so that the inequality 6 is valid. Theorem 2.4 is a corollary of the well-known fact that Lasso is able to recover the sign pattern and the locations of non-zero elements of the unknown coefficients if they are sufficiently large. In the framework considered above, this fact is formally stated below.

Theorem 2.5. *Under the assumptions of Theorem 2.4,*

$$\begin{aligned} \operatorname{supp}(\hat{\boldsymbol{\beta}}_L) &= \operatorname{supp}(\boldsymbol{\beta}_*), \quad \operatorname{sign}(\hat{\boldsymbol{\beta}}_{L,i}) = \operatorname{sign}(\boldsymbol{\beta}_{*},_i)\ \forall i \in S \\ \operatorname{supp}(\hat{\boldsymbol{\theta}}) &= \operatorname{supp}(\boldsymbol{\theta}_*), \quad \operatorname{sign}(\hat{\boldsymbol{\theta}}, j) = \operatorname{sign}(\boldsymbol{\theta}_{*},_j)\ \forall j \in \mathcal{O}, \end{aligned} \tag{8}$$

and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_\infty \le 3.5\lambda, \quad \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|_\infty \le 3.5\lambda$, with probability at least $1 - Cp^{1-c/A_0}$ for some absolute constants $C > 0$ and $c \ge 1/4$.

Remark 2.6. We showed that the Lasso estimator admits strong performance guarantees in the presence of gross measurement errors and dense additive noise for most "typical" vectors of outliers and regression coefficients when the design is moderately correlated. While our results are not uniform, they hold under mild assumptions and can shed light on the success of Lasso observed by the practitioners. It would be interesting to understand whether it is possible to deduce optimal performance guarantees without additional assumptions on the magnitude of the quantities involved.

References

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

[2] Bühlmann, P., Rütimann, P., Van De Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. Journal of Statistical Planning and Inference, 143(11):1835–1858.

[3] Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. The Annals of Statistics, 37(5A).

[4] Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. Advances in Neural Information Processing Systems, 32.

[5] Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. Statistics and Computing, 17:293–310.

[6] Gao, C. (2020). Robust regression via multivariate regression depth. Bernoulli, 26(2):1139–1170.

[7] Hebiri, M. and Lederer, J. (2012). How correlations influence Lasso prediction. IEEE Transactions on Information Theory, 59(3):1846–1854.

[8] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, 23(1):110–133.

[9] Minsker, S., Ndaoud, M., and Wang, L. (2024). Robust and tuning-free sparse linear regression via square-root slope. SIAM Journal on Mathematics of Data Science, 6(2):428–453.

[10] Nguyen, N. H. and Tran, T. D. (2013). Robust lasso with missing and grossly corrupted observations. IEEE Transactions on Information Theory, 59(4):2036–2058.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

[12] Tropp, J. A. (2008). Norms of random submatrices and sparse approximation. Comptes Rendus Mathematique, 346(23-24):1271–1274.

[13] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2008). Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2):210–227.

[14] Zhivotovskiy, N. (2024). Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. Electronic Journal of Probability, 29:1–28.

Supplementary Material
The Impact of Contamination and Correlated Design on the Lasso: an Average Case Analysis

A. Double rejective sampling and Poissonization

First, let us recall the definition of the Poisson sampling.

Definition A.1. (**Poisson Sampling**) Let $\delta_1, \dots, \delta_K$ be a sequence of K independent Bernoulli random variables with success probabilities $p_1, \dots, p_K$ such that $\sum_{j=1}^{K} p_j = S$, where S is a positive integer. We say the the random set I follows the Poisson sampling model if

$$I \overset{\mathrm{d}}{=} \{i \mid \delta_i = 1\}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution. It is easy to see that for any $\mathcal{I} \subseteq [K]$,

$$\mathbb{P}(I = \mathcal{I}) = \prod_{i \in \mathcal{I}} p_i \prod_{j \notin \mathcal{I}} (1 - p_j).$$

Furthermore, assume that $K = K_1 + K_2$ and define

$$I_1 := \{i \mid \delta_i = 1, i = 1, \dots, K_1\}, \quad I_2 := \{i \mid \delta_i = 1, i = K_1 + 1, \dots, K\} \tag{9}$$

Then I_1 and I_2 are independent random sets. Assume that $S < K_1 \wedge K_2$. Similar to the rejective sampling model used by [6], we define a model to accommodate to our needs.

Definition A.2. (**Double Rejective Sampling**) Let $\delta_1, \dots, \delta_K$ denote a sequence of $K = K_1 + K_2$ independent Bernoulli random variables with success probabilities $p_j,\ j = 1, \dots, K$ such that $\sum_{j=1}^{K_1} p_j = S_1 \in \mathbb{N}, \sum_{j=K_1+1}^{K} p_j = S_2 \in \mathbb{N}$, and denote by $\mathbb{P}$ the probability measure of the corresponding double Poisson sampling model. We say a random set follows a double rejective sampling model with parameters $(K_1, K_2, (p_i)_{i=1}^K)$ if for all $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2 \subset [K]$ is chosen with the following probability

$$\begin{aligned}\mathbb{P}_{S_1,S_2}(\mathcal{I}) &:= \mathbb{P}(\mathcal{I} \mid |\mathcal{I}_1| = S_1, |\mathcal{I}_2| = S_2) \\ &= \begin{cases} c' \prod_{i \in \mathcal{I}} p_i \prod_{j \notin \mathcal{I}} (1 - p_j) & \text{if } |\mathcal{I}_1| = S_1, |\mathcal{I}_2| = S_2 \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Recall the following lemma that is proven in [6, Lemma 7]. Recall that $[K] := \{1, \dots, K\}$ and let $\mathcal{P}(\mathcal{I})$ denotes the power set of $\mathcal{I}$.

Lemma A.1. *Let $f : \mathcal{P}([K]) \to \{0, 1\}$ be such that for all $\mathcal{I}, \mathcal{J} \in \mathcal{P}([K])$,*

$$f(\mathcal{I}) \leq f(\mathcal{J}) \text{ if } \mathcal{I} \subseteq \mathcal{J}.$$

Then for $0 \leq T \leq K - 1$ we have

$$\mathbb{P}(f(\mathcal{I}) = 1 \mid |\mathcal{I}| = T) \leq \mathbb{P}(f(\mathcal{I}) = 1 \mid |\mathcal{I}| = T + 1)$$

The next lemma states that probabilities with respect to the double rejective sampling model can be controlled in terms of the probabilities associated with the Poisson sampling process. Therefore, we can use the Poisson sampling model whenever it is convenient, and the bound of the lemma allows to automatically transfer the results to the case of double rejective sampling.

Lemma A.2. *The following inequality holds:* $\mathbb{P}_{S_1,S_2}(f(\mathcal{I}) = 1) \leq 4\mathbb{P}(f(\mathcal{I}) = 1)$.

Proof. Note that

$$\begin{aligned}
&\mathbb{P}(f(\mathcal{I}) = 1) \\
&= \sum_{k_1=0}^{K_1} \sum_{k_2=0}^{K_2} \mathbb{P}(f(\mathcal{I}) = 1 \mid |\mathcal{I}_1| = k_1, |\mathcal{I}_2| = k_2) \times \mathbb{P}(|\mathcal{I}_1| = k_1, |\mathcal{I}_2| = k_2) \\
&\geq \sum_{k_1=S_1}^{K_1} \sum_{k_2=S_2}^{K_2} \mathbb{P}(f(\mathcal{I}) = 1 \mid |\mathcal{I}_1| = k_1, |\mathcal{I}_2| = k_2) \times \mathbb{P}(|\mathcal{I}_1| = k_1)\mathbb{P}(|\mathcal{I}_2| = k_2) \\
&\geq \mathbb{P}(f(\mathcal{I}) = 1 \mid |\mathcal{I}_1| = S_1, |\mathcal{I}_2| = S_2) \times \sum_{k_1=S_1}^{K_1} \mathbb{P}(|\mathcal{I}_1| = k_1) \sum_{k_2=S_2}^{K_2} \mathbb{P}(|\mathcal{I}_2| = k_2) \\
&= \mathbb{P}_{S_1,S_2}(f(\mathcal{I}) = 1) \cdot \mathbb{P}(|\mathcal{I}_1| \geq S_1) \cdot \mathbb{P}(|\mathcal{I}_2| \geq S_2) \geq \frac{1}{4}\mathbb{P}_{S_1,S_2}(f(\mathcal{I}) = 1),
\end{aligned}$$

where the first inequality comes from the independence of $\mathcal{I}_1$ and $\mathcal{I}_2$, the second one comes from Lemma A.1 and the last one is implied by the fact that if the mean number of successes of K independent trials is an integer S, the median is also S [proved by 5]. □

B. Norms of random submatrices

Let Assumption 2.1 and Assumption 2.2 be satisfied. Moreover, let the set $I \subseteq [p+n]$ be selected according to the double rejective sampling model with parameters $K_1 = p$ and $K_2 = n$, $p_j = s/p$ for $j = 1, \dots, p$ and $\delta_j = o/n$ for $j = p+1, \dots, p+n$. We let $\boldsymbol{M}_I$ stand for the submatrix of $\boldsymbol{M}$ with columns indexed by $I = \mathcal{S} \cup \mathcal{O}$, and $\boldsymbol{X}_\mathcal{S}$ be the corresponding submatrix of $\boldsymbol{X}$ indexed by $\mathcal{S}$. Then we have the following bounds.

Lemma B.1. *With probability at least* $1 - p^{-2\log(p)}$,

$$\|(\boldsymbol{X}_\mathcal{S}^\top \boldsymbol{X}_\mathcal{S}/n)^{-1}\| \leq 2. \tag{10}$$

Moreover, with probability at least $1 - Cp^{1-1/(4A_0)}$,

$$\|(\boldsymbol{M}_\mathcal{I}^\top \boldsymbol{M}_\mathcal{I}/n)^{-1}\| \leq 2. \tag{11}$$

To prove these bounds, let us define, for any $\mathcal{I} \in [p+n]$,

$$f(\mathcal{I}) = \mathbb{1}\{\|\boldsymbol{M}_\mathcal{I}^\top \boldsymbol{M}_\mathcal{I}/n - \mathbf{I}_{s+o}\| \geq r\}$$

where $r \in (0,1)$. In view of Lemma A.2 we have that

$$\mathbb{P}_{S_1,S_2}(\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{s+o}\| \geq r) \leq 4\mathbb{P}(\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{s+o}\| \geq r). \tag{12}$$

The following two lemmas are among the key technical results of the paper.

Lemma B.2. *Assume $\mathcal{I} \subset [p+n]$ is chosen according to the double rejective sampling model with $K_1 = p, K_2 = n, S_1 = s$ and $S_2 = o$. Then, if*

$$\sqrt{s/p}\|\boldsymbol{X}/\sqrt{n}\| + \sqrt{o/n} \leq c/\sqrt{\log(p)}$$

for $c \in (0, 1/8e^2)$ and $2A_0 < 1$, then $\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{s+o}\| < r$ with probability at least $1 - Cp^{1-1/(4A_0)}$ for some positive constant $r \in (0,1), C$.

Note that an immediate implication is that whenever the inequality of Lemma B.2 holds, $\|\boldsymbol{M}_{\mathcal{I}}/\sqrt{n}\| > \sqrt{1-r}$.

Lemma B.3. *The following inequality holds with probability at least $1 - p^{-2\log 2}$:*

$$\max_{i \in \mathcal{S}^c} \|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_i/n\|_2 \leq C_1 \cdot (\sqrt{\log(p)})^{-1}. \tag{13}$$

Moreover, $\max_{i \in \mathcal{I}^c} \|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i/n\|_2 \leq C_2 \cdot (\sqrt{\log(p)})^{-1}$ *with probability at least* $1 - p^{1-1/A_0^2}$.

The proofs of Lemmas B.2 and B.3 are given in the supplementary material.

C. Essential technical lemmas

Recall that $\lambda \geq 4\sigma\sqrt{\frac{2(\log(p)+\log(n))}{n}}$.

Lemma C.1. *With probability at least $1 - 2p^{-1}$,*

$$\left\|\frac{\boldsymbol{X}^\top \boldsymbol{\xi}}{n}\right\|_\infty \leq 2\sigma\sqrt{\frac{2\log(p)}{n}}. \tag{14}$$

Moreover, with probability at least $1 - 4p^{-1}$

$$\left\|\frac{\boldsymbol{M}^\top \boldsymbol{\xi}}{n}\right\|_\infty \leq 2\sigma\sqrt{\frac{2(\log(p)+\log(n))}{n}} = \frac{\sqrt{2}}{2}\lambda \tag{15}$$

as long as $n \geq 5$.

Proof. Note that

$$\|\boldsymbol{M}^\top \boldsymbol{\xi}\|_\infty = \|\boldsymbol{X}^\top \boldsymbol{\xi}\|_\infty + \sqrt{n}\|\boldsymbol{\xi}\|_\infty$$

Since the noise $\boldsymbol{\xi}$ consists of i.i.d sub-Gaussian entries with variance at most σ^2,

$$\mathbb{P}\left(\max_{i \in [n]} |\boldsymbol{\xi}_i| > \sigma\sqrt{2(\log n + t)}\right) \leq 2e^{-t}$$

Take $t = \log(p)$, we have with probability at least $1 - 2p^{-1}$,

$$\|\boldsymbol{\xi}\|_\infty = \max_{i\in[n]} |\boldsymbol{\xi}_i| \le 2\sigma\sqrt{\log(p) + \log(n)}$$

Since $\|\boldsymbol{X}^\top\boldsymbol{\xi}\|_\infty = \max_{i\in[p]} |\boldsymbol{X}_i^\top\boldsymbol{\xi}|$ and for each i, $\boldsymbol{X}_i^\top\boldsymbol{\xi}$ is a sub-Gaussian random variable with variance at most $\|\boldsymbol{X}_i\|^2\sigma^2 = n\sigma^2$, with probability at least $1 - 2p^{-1}$, $\|\boldsymbol{X}^\top\boldsymbol{\xi}\|_\infty \le 2\sigma\sqrt{n\log(p)}$. □

Next, we will establish the so-called "complementary size" [4] condition for the augmented design matrix.

Lemma C.2. *With probability at least* $1 - Cp^{-c(A_0)}$,

$$\frac{1}{n}\|\boldsymbol{M}_{\mathcal{I}^c}^\top\boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{\xi}\|_\infty + 2\lambda\|\boldsymbol{M}_{\mathcal{I}^c}^\top\boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}})^{-1}\operatorname{sign}(\beta_{\mathcal{I}})\|_\infty \le \frac{1}{2}\lambda \quad (16)$$

for some absolute constant C and $c(A_0)$ such that $c(A_0) \to \infty$ as $A_0 \to 0$.

Proof. For each $i \in \mathcal{I}^c$, let

$$W_i = (\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i, W_i' = \boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i.$$

Define the event E via

$$E = \{\max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\|_2 \le c_1(\sqrt{\log(p)})^{-1}\} \cup \{\|(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \le 2\} \quad (17)$$

for some absolute constant c_1. Previously, we have shown that

$$\mathbb{P}(E) \ge 1 - C_0 p^{-c_0}$$

for some constants C_0, c_0 depending on A_0 appearing in the coherence condition. Then on the event E, $\|\boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \le \sqrt{2n}$ and for each $i \in \mathcal{I}^c$,

$$\max_{i\in\mathcal{I}^c}\|W_i\|_2 \le \|(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\|_2 \le (c_2\sqrt{\log(p)})^{-1} \quad (18)$$

and

$$\max_{i\in\mathcal{I}^c}\|W_i'\|_2 \le \|\boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\|_2 \le \sqrt{n}(c_2'\sqrt{\log(p)})^{-1}. \quad (19)$$

Therefore, by Hoeffding's inequality, for each i and $t > 0, u > 0$,

$$\begin{aligned}\mathbb{P}(|\langle W_i, \operatorname{sign}(\theta_{\mathcal{I}})\rangle| > t \mid E) &\le 2e^{-\frac{t^2}{2\sum_{j\in\mathcal{I}}(W_i)_j^2}} \\ &\le 2e^{-t^2/2\max_i\|W_i\|_2^2} \quad (20) \\ &\le 2e^{-c_2^2t^2\log(p)/2} = p^{-c_2^2t^2/2}\end{aligned}$$

and

$$
\begin{aligned}
\mathbb{P}(|\langle W_i', \boldsymbol{\xi}\rangle| > \sqrt{n}u \mid E) &\leq 2e^{-\frac{nu^2}{2\sum_{j\in\mathcal{I}}(W_i')_j^2}} \\
&\leq 2e^{-nu^2/2\max_i \|W_i'\|_2^2} \\
&\leq 2e^{-c_2' u^2 \log(p)/2} = p^{-c_2'^2 u^2/2}.
\end{aligned} \tag{21}
$$

Union bound implies that

$$
\mathbb{P}(\|\langle W_i, \operatorname{sign}(\theta_{\mathcal{I}})\rangle\|_\infty > t \mid E) \leq 2(n+p-s-o)p^{-c_2 t^2/2} \leq 4p^{1-c_2^2 t^2/2} \tag{22}
$$

and

$$
\mathbb{P}(\|\langle W_i', \boldsymbol{\xi})\rangle\|_\infty > \sqrt{n}u \mid E) \leq 2(n+p-s-o)p^{-c_2' u^2/2} \leq 4p^{1-c_2'^2 u^2/2}. \tag{23}
$$

Take $t = 1/8$ and $u = 1/4$, and note that with probabilities at least $1-4p^{1-c_2^2t^2/2}-\mathbb{P}(E)$ and $1-4p^{1-c_2'^2u^2/2}-\mathbb{P}(E)$ respectively,

$$
\|\boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_\infty = \max_{i\in\mathcal{I}^c}\langle W_i, \boldsymbol{\gamma}_{\mathcal{I}}\rangle \leq \frac{1}{8} \tag{24}
$$

and

$$
\frac{1}{n}\|\boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{M}_{\mathcal{I}}(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{\xi}\|_\infty = \max_{i\in\mathcal{I}^c}\langle W_i', \boldsymbol{\xi}\rangle \leq \frac{1}{4}\lambda. \tag{25}
$$

Finally, we have that

$$
\begin{aligned}
&\mathbb{P}(\{\|\langle W_i, \operatorname{sign}(\theta_{\mathcal{I})}\rangle\|_\infty > 1/8\} \cup \{\|\langle W_i', \boldsymbol{\xi}\rangle\|_\infty > 1/4\lambda\}) \\
&\leq \mathbb{P}(\{\|\langle W_i, \operatorname{sign}(\theta_{\mathcal{I})}\rangle\|_\infty > 1/8\} \cup \{\|\langle W_i', \boldsymbol{\xi}\rangle\|_\infty > 1/4\lambda\} \mid E) + \mathbb{P}(E^c) \\
&\leq \mathbb{P}(\{\|\langle W_i, \operatorname{sign}(\theta_{\mathcal{I})}\rangle\|_\infty > 1/8\} \mid E) + \mathbb{P}(\{\|\langle W_i', \boldsymbol{\xi}\rangle\|_\infty > 1/4\lambda\} \mid E) + \mathbb{P}(E^c) \\
&\leq 1 - C_3 p^{-c_3}
\end{aligned} \tag{26}
$$

for some positive constant C_3 and c_3. □

Lemma C.3. *Let $\Pi_{\mathcal{J}}$ be the projection matrix onto the space spanned by the columns of $M_{\mathcal{I}}$.*

$$
\|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{\xi}\|_\infty \leq \lambda/2,
$$

$$
\|\boldsymbol{M}_{\mathcal{I}^c}^\top(\mathbf{I}-\Pi_{\mathcal{J}})\boldsymbol{\xi}/\sqrt{n}\|_\infty \leq \lambda
$$

with probability at least $1 - C_1 p^{-C_2} - p^{-1}$ for some positive constants C_1 and C_2.

Proof. In the proof of Lemma C.2 we showed that $\|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\| \leq \sqrt{2}$ with probability at least $1-C_1p^{-C_2}$ for some positive constant C_1, C_2. For each $i \in \mathcal{I}$ define U_i to be the i-th row of $(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top$, $\max_{i\in\mathcal{I}}\|U_i\|_2 = \max_{i\in\mathcal{I}}\|e_i^\top(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\|_2 \leq \|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^\top\| \leq \sqrt{2}$.

Then Hoeffding's inequality gives that

$$
\begin{aligned}
\mathbb{P}(|\langle U_i, \boldsymbol{\xi}\rangle| > \lambda/2) &\leq 2e^{-\lambda^2/8\|U_i\|_2^2} \\
&\leq 2e^{-\lambda^2/8\max_{i\in\mathcal{I}}\|U_i\|_2^2} \leq 2p^{-2}
\end{aligned}
$$

and the union bound yields that with probability at least $1 - (s + o)p^{-2}$,

$$\|(\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}})^{-1}\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{\xi}\|_{\infty} = \max_{i\in\mathcal{I}} |\langle U_i, \boldsymbol{\xi}\rangle| \leq \lambda/2.$$

For the other term, $\|\boldsymbol{M}_{\mathcal{I}^c}^{\top}(\mathbf{I} - \Pi_{\mathcal{I}})\boldsymbol{\xi}/\sqrt{n}\|_{\infty} = \max_{i\in\mathcal{I}^c}((\mathbf{I} - \Pi_{\mathcal{I}})\boldsymbol{M}_i)^{\top}\boldsymbol{\xi}/\sqrt{n} =: \max_{i\in\mathcal{I}^c} U_i'$, each U_i' is a sub-Gaussian random variable with variance at most $\|(\mathbf{I} - \Pi_{\mathcal{I}})\boldsymbol{M}_i/\sqrt{n}\|_2^2 \leq \|\boldsymbol{M}_i/\sqrt{n}\|_2^2 = 1$. Therefore, with probability at least $1 - 2p^{-2}$, $U_i' \leq 4\sqrt{2\log(p)/n}$ and by the union bound,

$$\|\boldsymbol{M}_{\mathcal{I}^c}^{\top}(\mathbf{I} - \Pi_{\mathcal{I}})\boldsymbol{\xi}/\sqrt{n}\|_{\infty} \leq \lambda,$$

with probability at least $1 - 4p^{-1} > 1 - 2(n + p - s - o)p^{-1}$. □

Lemma C.4. *With probability at least* $1 - 2(s + o)p^{-1} - p^{1-c/A_0}$,

$$\|(\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_{\infty} \leq 2.5,$$

where $c > 0$ *is an absolute constant.*

Proof. Notice that

$$\begin{aligned}\|(\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1}\operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_{\infty} &\leq \|\operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_{\infty} + \|((\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1} - \mathbf{I}_{s+o})\operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_{\infty} \\ &= 1 + \max_{i\in\mathcal{I}}\langle W_i, \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\rangle,\end{aligned}$$

where W_i is the i-th row of the matrix $(\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1} - \mathbf{I}_{s+o}$. Define $\boldsymbol{A} := \mathbf{I}_{s+o} - \boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}$, then

$$(\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1} = \sum_{k\geq 0}\boldsymbol{A}^k, \quad \boldsymbol{A}^0 = \mathbf{I}_{s+o}$$

$$W_i = ((\boldsymbol{M}_{\mathcal{I}}^{\top}\boldsymbol{M}_{\mathcal{I}}/n)^{-1} - \mathbf{I}_{s+o})e_i = \sum_{k>0}\boldsymbol{A}^k e_k$$

$$\|W_i\| \leq \sum_{i>0}\|\boldsymbol{A}^k e_i\| \leq \|\boldsymbol{A}e_i\|\sum_{k\geq 0}\|\boldsymbol{A}\|^k = \frac{\|\boldsymbol{A}e_i\|}{1 - \|\boldsymbol{A}\|} \leq 2\|\boldsymbol{A}e_i\|$$

last inequality coming from the proof of Lemma B.1 where we showed that with probability at least $1 - Cp^{1-1/(4A_0)}$, $\|\boldsymbol{A}\| < 1/2$. Recall in the proof of Lemma B.1 we defined $H = \boldsymbol{M}^{\top}\boldsymbol{M}/n - \mathbf{I}_{n+p}$. Then $\boldsymbol{A} = H_{\mathcal{I}}$. Combining inequality (12) with Lemma 21 in [6], we deduce that

$$\mathbb{P}_{S_1,S_2}(\|\boldsymbol{A}\|_{\infty,2} > (\log(p))^{-1/2}) \leq 4(n + p)(e\log(p)\|HW\|_{\infty,2}^2)^{\frac{1}{\mu^2\log(p)}}. \quad (27)$$

Recall that in Section B we showed that $\|HW\|_{\infty,2} \leq \sqrt{\frac{s}{p}}\|X\| + \sqrt{\frac{o}{n}} < c_1(\log(p))^{-1/2}$, moreover, we have also assumed that $\mu = \mu(X) \leq A_0 \cdot (\log(p))^{-1}$. Therefore

$$\max_{i\in\mathcal{I}}\|\boldsymbol{A}e_i\| = \|\boldsymbol{A}\|_{\infty,2} \leq (\log(p))^{-1/2}$$

with probability at least $1 - Cp^{1-1/(A_0^2\log(p))}$. Repeating to the proof of bound (24), we can show that whenever $\max_i\|W_i\| \leq 2(\log(p))^{-1/2}$,

$$\max_{i\in\mathcal{I}}\langle W_i, \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\rangle \leq \sqrt{2} < 1.5 \quad (28)$$

with probability at least $1 - 2(s + o)p^{-1}$. □

D. Proof of Lemma B.2

Proof. The result for the simple rejective sampling model that yields the inequality (10) was proved by [4]. To extend the result to the double rejective sampling case needed to prove (10), define the hollow Gram matrix $\boldsymbol{H} := \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{n+p}$ and the weight matrix $\boldsymbol{W} := \operatorname{diag}\left((\sqrt{p_i})_{i=1}^{p+n}\right)$. Notice that $\boldsymbol{M}/\sqrt{n}$ has columns of norm 1, we can combine inequality (12) with proof of Corollary 4 in [6] to deduce that for all $r \geq 2e^2\|\boldsymbol{W}\boldsymbol{H}\boldsymbol{W}\|$,

$$\mathbb{P}_{S_1,S_2}(\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{n+p}\| \leq r) \leq C \cdot p \cdot \exp\left(-\min\left\{\frac{r^2}{4e^2\|HW\|_{\infty,2}^2}, \frac{r}{2\mu}\right\}\right) \tag{29}$$

where $C < 10^3$ is a positive constant and $\|\cdot\|_{\infty,2}$ is the maximum ℓ_2 norm of a row. Recall $\mu = \mu(\boldsymbol{X}) \leq A_0 \cdot (\log(p))^{-1}$. In our case,

$$\boldsymbol{H}\boldsymbol{W} = \begin{bmatrix} \sqrt{\frac{s}{p}}(\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p) & \sqrt{\frac{o}{n}}\boldsymbol{X}^\top/\sqrt{n} \\ \sqrt{\frac{s}{p}}\boldsymbol{X}/\sqrt{n} & 0 \end{bmatrix},$$

$$\boldsymbol{W}\boldsymbol{H}\boldsymbol{W} = \begin{bmatrix} \frac{s}{p}(\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p) & \sqrt{\frac{so}{pn}}\boldsymbol{X}^\top/\sqrt{n} \\ \sqrt{\frac{so}{pn}}\boldsymbol{X}/\sqrt{n} & 0 \end{bmatrix}.$$

Then, since $\|\boldsymbol{X}_i/\sqrt{n}\| = 1$ for each column of $\boldsymbol{X}$,

$$\begin{aligned} \|\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p\|_{\infty,2} = \max \|\boldsymbol{X}_i^\top \boldsymbol{X}/n - e_i\|_2 &\leq \max \|\boldsymbol{X}_i/\sqrt{n}\|\|\boldsymbol{X}/\sqrt{n}\| + 1 \\ &= \|\boldsymbol{X}\|/\sqrt{n} + 1. \end{aligned} \tag{30}$$

and $\|\boldsymbol{X}\|_{\infty,2}/\sqrt{n} \leq \|\boldsymbol{X}\|/\sqrt{n}$. Therefore,

$$\begin{aligned} \|\boldsymbol{H}\boldsymbol{W}\|_{\infty,2} &\leq \sqrt{\frac{s}{p}}\|\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p\|_{\infty,2} + \sqrt{\frac{o}{n}}\|\boldsymbol{X}^\top\|_{\infty,2}/\sqrt{n} \\ &\leq 2\left(\sqrt{\frac{s}{p}}\|\boldsymbol{X}\|/\sqrt{n} + \sqrt{\frac{o}{n}}\right). \end{aligned} \tag{31}$$

For $\boldsymbol{W}\boldsymbol{H}\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{M}^\top \boldsymbol{M}/n)\boldsymbol{W} - \boldsymbol{W}^2$, we see that

$$\boldsymbol{W}(\boldsymbol{M}^\top \boldsymbol{M}/n)\boldsymbol{W} = \begin{bmatrix} \frac{s}{p}\boldsymbol{X}^\top \boldsymbol{X}/n & \sqrt{\frac{so}{pn}}\boldsymbol{X}^\top/\sqrt{n} \\ \sqrt{\frac{so}{pn}}\boldsymbol{X}/\sqrt{n} & \frac{o}{n}I_p \end{bmatrix} \succeq 0.$$

In view of Corollary 3.5 in [2],

$$\boldsymbol{W}(\boldsymbol{M}^\top \boldsymbol{M}/n)\boldsymbol{W}/n \leq \frac{s}{p}\|\boldsymbol{X}^\top \boldsymbol{X}\|/n + \frac{o}{n} = \frac{s}{p}\|\boldsymbol{X}\|^2/n + \frac{o}{n}.$$

On the other hand, $\boldsymbol{W}^2$ is a diagonal matrix with first p diagonal entries equal to s/p and last n diagonal entries equal to o/n. By Weyl's inequality,

$$\|\boldsymbol{W}\boldsymbol{H}\boldsymbol{W}\| \leq \|\boldsymbol{W}(\boldsymbol{M}^\top \boldsymbol{M}/n)\boldsymbol{W}\| + \lambda_{\min}(-\boldsymbol{W}^2) \leq \|\boldsymbol{W}(\boldsymbol{M}^\top \boldsymbol{M}/n)\boldsymbol{W}\|.$$

Equation (29) is equivalent to stating that with probability at least $1 - C \cdot pe^{-t}$,

$$\begin{aligned}\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n - \mathbf{I}_{s+o}\| &\leq \max(2e\sqrt{t}\|\boldsymbol{H}\boldsymbol{W}\|_{\infty,2}, 2t\mu) \\ &\leq \max\left(2e\sqrt{t}\left(\sqrt{\frac{s}{p}}\|\boldsymbol{X}\|/\sqrt{n} + \sqrt{\frac{o}{n}}\right), \frac{2tA_0}{\log(p)}\right). \end{aligned} \tag{32}$$

Set $r = \max\left(2e\sqrt{t}\left(\sqrt{\frac{s}{p}}\|\boldsymbol{X}\| + \sqrt{\frac{o}{n}}\right), \frac{2tA_0}{\log(p)}\right)$ and $t = c \cdot \log(p)$. If

$$\sqrt{\frac{s}{p}}\|\boldsymbol{X}/\sqrt{n}\| + \sqrt{\frac{o}{n}} < \frac{\sqrt{c}A_0}{e\sqrt{\log(p)}} \tag{33}$$

and $c < 1/(2A_0)$, then $r = \frac{2tA_0}{\log(p)} < 1$ and with probability at least $1 - C \cdot p^{1-c}$, $\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}} - \mathbf{I}_{s+o}\| \leq r$. We need $c > 1$ for p^{1-c} to be small, thus require $A_0 < 1/2$.

We still need to ensure that $r < 1$ for $r \geq 2e^2\|\boldsymbol{W}\boldsymbol{H}\boldsymbol{W}\|$. A sufficient condition is $2e^2\|\boldsymbol{W}\boldsymbol{H}\boldsymbol{W}\| < 1$, which is satisfied if we require that $\frac{s}{p}\|\boldsymbol{X}\|^2/n + \frac{o}{n} < c'$ for some $c' \leq 0.125e^{-2}$.

In particular, taking $r = 1/2$ and $t = \frac{1}{4A_0}\log(p)$, we have that with probability at least $1 - Cp^{1-1/(4A_0)}$, $\|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \leq 2$ given that $A_0 \leq 1/4$. □

E. Proofs

In this section, we state the main technical results related to the norms of matrices obtained by choosing the random columns of the augmented design matrix

$$\boldsymbol{M} := \left[\boldsymbol{X}\ \sqrt{n}\boldsymbol{I}\right].$$

Then we explain the implications of these bounds for the Lasso.

E.1. Proof of Theorem 2.1

Recall that $\boldsymbol{M} = [\boldsymbol{X} \mid \sqrt{n}\boldsymbol{I}]$, and denote $\boldsymbol{\gamma} = (\boldsymbol{\beta_*}^\top, \boldsymbol{\theta_*}^\top)^\top$. Moreover, let $\mathcal{I}$ stand for the support of $\boldsymbol{\gamma}$. Then $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{\gamma}$, and we need to show that $\boldsymbol{\gamma}$ is the unique solution to the problem

$$\min_{\bar{\boldsymbol{\gamma}} \in \mathbb{R}^{p+n}} \quad \|\bar{\boldsymbol{\gamma}}\|_1 \qquad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{M}\bar{\boldsymbol{\gamma}}. \tag{34}$$

We will apply the following lemma.

Lemma E.1 (Lemma 3.2 in [3]). *Assume that*

$$\left\|\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n\right)^{-1}\right\| \leq 2, \quad \max_{i \in \mathcal{I}^c}\left\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i/n\right\|_2 \leq 1. \tag{35}$$

Suppose there exists $\boldsymbol{v} \in \mathbb{R}^{n+p}$ *in the row space of* $\boldsymbol{M}$ *obeying the inequalities*

$$\|\boldsymbol{v}_{\mathcal{I}} - \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\|_2 \le 1/4, \quad \|\boldsymbol{v}_{\mathcal{I}^c}\|_\infty \le 1/4. \tag{36}$$

Then $\boldsymbol{\gamma}$ *is the unique solution to problem* (34).

In section (B), we showed in particular that conditions (35) hold with probability at least $1 - C_1 p^{-c_1}$ for some positive constants c_1, C_1. Next, let

$$\boldsymbol{v} = \left(\boldsymbol{M}^\top \boldsymbol{M}_{\mathcal{I}}/n\right)\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n\right)^{-1} \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) \in \mathbb{R}^{n+p}. \tag{37}$$

We want to show that $\boldsymbol{v}$ satisfies relations (36) with high probability. Let E be the event on which inequalities (35) hold. On this event, $\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}$ is invertible and

$$\begin{aligned} \boldsymbol{v}_{\mathcal{I}} &= \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1} \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) = \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}}), \\ \boldsymbol{v}_{\mathcal{I}^c} &= \boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{M}_{\mathcal{I}}\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1} \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) \end{aligned} \tag{38}$$

For each $i \in \mathcal{I}^c$, define $\boldsymbol{W}_i = \left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i$. Moreover, on E

$$\begin{aligned} \|\boldsymbol{W}_i\|_2 = \left\|\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i\right\|_2 \le \left\|\left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1}\right\| \left\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i\right\|_2 \\ \le C_1/\sqrt{\log(p)}, \end{aligned} \tag{39}$$

where we used the relations $\|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}\| \in (1/2, 3/2)$ and $\|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i\|_2 \le C/\log(p)$.

Recall that in the generic sparse model, the signs of the entries of $\boldsymbol{\gamma}$ are iid Rademacher random variables. In view of this fact, Hoeffding's inequality gives that

$$\begin{aligned} \mathbb{P}\left(|\langle \boldsymbol{W}_i, \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\rangle| > t|E\right) &\le 2e^{-\frac{t^2}{2\sum_{j\in\mathcal{I}}(\boldsymbol{W}_i)_j^2}} \\ &\le 2e^{-t^2/2\max_i \|\boldsymbol{W}_i\|_2^2} \le 2e^{-c_2^2 t^2 \log(p)/2} = p^{-c_2^2 t^2/2}. \end{aligned} \tag{40}$$

Taking $t = 1/4$ and applying the union bound, we deduce that

$$\|\boldsymbol{v}_{\mathcal{I}^c}\|_\infty = \max_{i\in\mathcal{I}^c}\left\langle \left(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}\right)^{-1} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i, \operatorname{sign}(\boldsymbol{\gamma})\right\rangle \le 1/4 \tag{41}$$

with probability at least $1 - 2(n+p-s-o)p^{-c_2 t^2/2} \ge 1 - 4p^{1-c_2^2 t^2/2}$.

E.2. Proof of Theorem 2.5

Recall the definitions of $\boldsymbol{M}$ and $\boldsymbol{\gamma}$, and let $\boldsymbol{\Gamma} = \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$, where $\hat{\boldsymbol{\gamma}}$ is the solution to the augmented Lasso problem $\min_{\bar{\boldsymbol{\gamma}}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{M}\bar{\boldsymbol{\gamma}}\|_2^2 + \lambda\|\bar{\boldsymbol{\gamma}}\|_1$. Recall that $\mathcal{S} = \operatorname{supp}(\boldsymbol{\beta}_*)$, $\mathcal{O} = \operatorname{supp}(\boldsymbol{\theta}_*)$ and $\mathcal{I} = \mathcal{S} \cup \mathcal{O}$. Inspired by the idea of Wainwright [7], we want to show that under our assumptions, the solution of the Lasso problem restricted to the supports of $\boldsymbol{\beta}_*$ and $\boldsymbol{\theta}_*$ defined via

$$\breve{\boldsymbol{\gamma}}_{\mathcal{I}} \in \arg\min_{\boldsymbol{\gamma}_{\mathcal{I}} \in \mathbb{R}^{s+o}} \frac{1}{2n} \|y - \boldsymbol{M}_{\mathcal{I}} \boldsymbol{\gamma}_{\mathcal{I}}\|_2^2 + \lambda \|\boldsymbol{\beta}_{\mathcal{S}}\|_1 + \lambda \|\boldsymbol{\theta}_{\mathcal{O}}\|_1 \tag{42}$$

is indeed the non-zero part of the solution to the original, unrestricted Lasso problem. Moreover, we will show that for each non-zero element $\boldsymbol{\gamma}_i$, if $|\boldsymbol{\gamma}_i| > \tau$ for some threshold τ to be determined later, then $\mathrm{sign}(\breve{\boldsymbol{\gamma}}_i) = \mathrm{sign}(\boldsymbol{\gamma}_i)$.

Karush-Kuhn-Tucker (KKT) optimality conditions imply that for any $\boldsymbol{z} \in \partial \|\hat{\gamma}\|_1$, $-\boldsymbol{M}^\top (y - \boldsymbol{M}\hat{\gamma})/n + \boldsymbol{\lambda} \otimes \boldsymbol{z} = 0$ which yields the relation

$$\frac{1}{n} \boldsymbol{M}^\top \boldsymbol{M} (\hat{\gamma} - \boldsymbol{\gamma}) - \frac{1}{n} \boldsymbol{M}^\top \boldsymbol{\xi} + \boldsymbol{\lambda} \otimes \boldsymbol{z} = 0 \tag{43}$$

where $\otimes$ represents the element-wise multiplication.

We want to show that the vector $(\hat{\boldsymbol{\gamma}}_{\mathcal{I}}, \boldsymbol{0})$ satisfies the KKT conditions. We can write equation (43) in a block form:

$$\frac{1}{n} \begin{bmatrix} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}} & \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}^c} \\ \boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{M}_{\mathcal{I}} & \boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{M}_{\mathcal{I}^c} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\gamma}}_{\mathcal{I}} - \boldsymbol{\gamma}_{\mathcal{I}} \\ 0 \end{bmatrix} - \frac{1}{n} \begin{bmatrix} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{\xi} \\ \boldsymbol{M}_{\mathcal{I}^c}^\top \boldsymbol{\xi} \end{bmatrix} + \boldsymbol{\lambda} \otimes \begin{bmatrix} \boldsymbol{z}_{\mathcal{I}} \\ \boldsymbol{z}_{\mathcal{I}^c} \end{bmatrix} = 0 \tag{44}$$

Lemma E.2. *Define $\breve{\boldsymbol{z}} \in \mathbb{R}^{s+o}$ via*

$$\breve{\boldsymbol{z}}_i = \begin{cases} \mathrm{sign}(\boldsymbol{\gamma}_i), & i \in \mathcal{I} \\ \frac{1}{\lambda} \langle \boldsymbol{M}_{\mathcal{I}} \breve{\boldsymbol{\Gamma}}_{\mathcal{I}}, \boldsymbol{M}_i \rangle - \frac{1}{\lambda n} \boldsymbol{M}_i^\top \boldsymbol{\xi}, & i \in \mathcal{I}^c \end{cases} . \tag{45}$$

. Then $\|\breve{\boldsymbol{z}}_{\mathcal{I}^c}\|_\infty < 1$ with probability at least $1 - Cp^{1-cA_0}$ for some constants C, c.

Therefore, with high probability $\|\breve{\boldsymbol{z}}_{\mathcal{I}^c}\|_\infty < 1$, which implies that with the same high probability, $\breve{\boldsymbol{z}} \in \partial \|\hat{\gamma}\|_1$ and the solution of the Lasso problem is $\hat{\boldsymbol{\gamma}} = (\breve{\boldsymbol{\gamma}}_{\mathcal{I}}, \boldsymbol{0})$, and $\breve{\boldsymbol{z}} \in \partial \|\hat{\gamma}\|_1$. Moreover, $\mathrm{supp}(\hat{\gamma}) \subseteq \mathrm{supp}(\boldsymbol{\gamma})$.

Define $\breve{\boldsymbol{\Gamma}}$ via

$$\breve{\boldsymbol{\Gamma}}_{\mathcal{I}} := (\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n)^{-1} \left(\frac{1}{n} \boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{\xi} - \boldsymbol{\lambda}_{\mathcal{I}} \otimes \breve{\boldsymbol{z}}_{\mathcal{I}} \right) \tag{46}$$

Note that $\breve{\boldsymbol{\Gamma}}$ is the candidate for the error vector $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$. Finally, let us determine the threshold τ. The main idea is that whenever the non-zero coefficient satisfies $|\boldsymbol{\gamma}_i| > \tau$, then $|\hat{\gamma}_i| = |\boldsymbol{\gamma}_i + \breve{\boldsymbol{\Gamma}}_i| > 0$ and $\mathrm{sign}(\boldsymbol{\gamma}_i) = \mathrm{sign}(\hat{\gamma}_i)$. We have that

$$\begin{aligned} \|\breve{\boldsymbol{\Gamma}}_{\mathcal{I}}\|_\infty &\leq \|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}})^{-1} \boldsymbol{M}_{\mathcal{I}} \boldsymbol{\xi}\|_\infty + \lambda \left\| (\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n)^{-1} \begin{bmatrix} \mathrm{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) \\ 0 \end{bmatrix} \right\|_\infty \\ &\leq 5\sqrt{2}/8\lambda + 2.5\lambda < 3.5\lambda. \end{aligned} \tag{47}$$

Therefore, if $\tau \geq 3.5\lambda$, then the sign consistency holds.

E.3. Proof of Theorem 2.4

Let $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Then Theorem 2.5 implies that $\mathrm{supp}(\boldsymbol{\Delta}) = \mathcal{S}$. Therefore, on the event $E := \{\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}/n\| \in (1/2, 3/2)\} \cup \{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty \leq 3.5\lambda\}$,

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \leq \|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}/n\| \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{\mathcal{S}}\|_2^2 &\leq \frac{3}{2} s\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty^2 \\ &\leq \frac{3}{2} s(3.5\lambda)^2 \leq \frac{147}{8} s\lambda_s^2. \quad (48)\end{aligned}$$

In sections (B) and (E.2), we proved that event E occurs with probability at least $1 - C_2 p^{1-c/A_0}$. This implies the claim of the theorem.

E.4. Proof of Lemma B.3

Proof. The first inequality follows from the results in [4]. To prove the second bound, we first notice that

$$\max_{i \in \mathcal{I}^c} \|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i/n\| \leq \|(\boldsymbol{M}^\top \boldsymbol{M}/n - \mathbf{I}_{n+p})_{\mathcal{I}}\|_{\infty,2} \quad (49)$$

where $\|\cdot\|_{\infty,2}$ is the maximum ℓ_2 norm of a row of a matrix. Next, we will need the following corollary of Lemma 6 in [6].

Lemma E.3. *Let $H \in \mathbb{R}^{d \times K}$. Assume that $\mathcal{I}$ is chosen according to the double rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$, $W = \mathrm{diag}\left((\sqrt{p_i})_{i=1}^K\right)$. Then for all $v > 0$,*

$$\mathbb{P}_{S_1,S_2}\left(\|\boldsymbol{H}_{\mathcal{I}}\|_{\infty,2} \geq v\right) \leq 4K\left(e\frac{\|\boldsymbol{H}\boldsymbol{W}\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}. \quad (50)$$

In our case, $d = n+p, K = n+p, \boldsymbol{H} = \boldsymbol{M}^\top \boldsymbol{M}/n - \mathbf{I}_{n+p}$ and

$$\boldsymbol{H}\boldsymbol{W} = \begin{bmatrix} \sqrt{\frac{s}{p}}(\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p) & \sqrt{\frac{o}{n}}\boldsymbol{X}^\top/\sqrt{n} \\ \sqrt{\frac{s}{p}}\boldsymbol{X}/\sqrt{n} & 0 \end{bmatrix}.$$

Therefore, since each column of $\boldsymbol{X}$ has norm $\sqrt{n}$,

$$\begin{aligned}\|\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p\|_{\infty,2} = \max_{i \in [p]} \|\boldsymbol{X}_i^\top \boldsymbol{X}/n - e_i\|_2 &= \max_{i \in [p]} \|\boldsymbol{X}_i^\top \boldsymbol{X}/n\|_2 \\ &\leq \max_{i \in [p]} \|\boldsymbol{X}_i/\sqrt{n}\| \|\boldsymbol{X}/\sqrt{n}\| = \|\boldsymbol{X}/\sqrt{n}\|. \quad (51)\end{aligned}$$

Combined with the fact that $\|\boldsymbol{X}\|_{\infty,2} \leq \|\boldsymbol{X}\|$, we have

$$\begin{aligned}&\|\boldsymbol{H}\boldsymbol{W}\|_{\infty,2} \\ &\leq \max\left(\left\|\sqrt{\frac{s}{p}}(\boldsymbol{X}^\top \boldsymbol{X}/n - \mathbf{I}_p)\right\|_{\infty,2} + \left\|\sqrt{\frac{o}{n}}\boldsymbol{X}^\top/\sqrt{n}\right\|_{\infty,2}, \left\|\sqrt{\frac{s}{p}}\boldsymbol{X}/\sqrt{n}\right\|_{\infty,2}\right) \\ &\leq \max\left(\sqrt{\frac{s}{p}}\|\boldsymbol{X}/\sqrt{n}\| + \sqrt{\frac{o}{n}}, \sqrt{\frac{s}{p}}\|\boldsymbol{X}/\sqrt{n}\|\right) \leq \sqrt{\frac{s}{p}}\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\| + \sqrt{\frac{o}{n}}.\end{aligned}$$

Now take $v := e(\sqrt{s/p}\|\boldsymbol{X}/\sqrt{n}\| + \sqrt{o/n})$ and recall that in view of assumption (2.2), $\mu \leq A_0/\log(p)$ and $\sqrt{s/p}\|\boldsymbol{X}/\sqrt{n}\| + \sqrt{o/n}+ \leq c/\sqrt{\log(p)}$, so that

$$\begin{aligned}&\mathbb{P}_{S_1,S_2}\left(\max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\| \geq e\left(\sqrt{\frac{s}{p}}\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\| + \sqrt{\frac{o}{n}}\right)\right)\\ &\leq 4(n+p)\exp\left(-e^2\mu^{-2}\left(\sqrt{\frac{s}{p}}\left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\| + \sqrt{\frac{o}{n}}\right)^2\right)\\ &\leq 8\exp(\log(p) - e^2(\log(p))^2A_0^{-2}\cdot c^2(\log(p))^{-1}) \leq 8p^{1-e^2c^2/A_0^2}.\end{aligned} \tag{52}$$

In other words, with probability at least $1 - 8p^{1-e^2c^2A_0^{-2}}$, $\max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\| \leq ec/\sqrt{\log(p)}$. Take $c = 1/16e$ to deduce that

$$\mathbb{P}_{S_1,S_2}\left(\max_{i\in\mathcal{I}^c}\|\boldsymbol{M}_{\mathcal{I}}^\top\boldsymbol{M}_i/n\| \geq \frac{1}{16}(\sqrt{\log(p)})^{-1}\right) \leq p^{1-(A_0)^{-2}/128}. \tag{53}$$

This probability is small because previously we require that $A_0 < 1/16$. □

E.5. Proof of Theorem 2.2

Proof. Without loss of generality, we will assume that $\sigma^2 = 1$.

Define $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*$ and $\boldsymbol{\Theta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*$. Since $\hat{\boldsymbol{\beta}}$ solves problem (1), we have the inequality

$$\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2 + \lambda\|\boldsymbol{\beta}_*\|_1 \tag{54}$$

Notice that

$$\begin{aligned}&\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2\\ &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\hat{\boldsymbol{\theta}} - \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)\|_2^2\\ &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 - 2\langle\boldsymbol{X}\boldsymbol{\Delta}, Y - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\boldsymbol{\theta}_* - \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)\rangle\\ &= \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 - 2\langle\boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\xi}\rangle - 2\langle\boldsymbol{X}\boldsymbol{\Delta}, \sqrt{n}\boldsymbol{\Theta}\rangle\end{aligned}$$

Plugging this relation in (54) and rearranging terms, we notice that the term $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_* - \sqrt{n}\hat{\boldsymbol{\theta}}\|_2^2$ cancels out, thus

$$\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 \leq \frac{1}{n}\langle\boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\xi}\rangle + \lambda(\|\boldsymbol{\beta}_*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) - \frac{1}{\sqrt{n}}\langle\boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\Theta}\rangle. \tag{55}$$

We will estimate the first two terms and the last display separately. First, we recall the following result due to Candès and Plan [4].

Theorem E.4. *Suppose that*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \qquad \hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\frac{1}{2n}\|y - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Suppose that the matrix $\boldsymbol{X}$ obeys the coherence property, and assume that $\boldsymbol{\beta}$ is chosen from the generic s-sparse model, $\boldsymbol{\xi}$ has independent sub-Gaussian entries and $s \leq c_0 p/\left[\|\boldsymbol{X}\|^2 \log(p)\right]$ for some absolute constant $c_0 > 0$. Then when $\lambda = 2\sqrt{2\log(p)/n}$,

$$\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2 \leq C_1\sigma^2 \cdot \lambda^2 \cdot s \tag{56}$$

with probability at least $1 - 6p^{-2\log 2} - p^{-1}(2\pi\log(p))^{-1/2}$.

The proof of this theorem mainly concerns the following conditions on $\boldsymbol{X}$ and $\boldsymbol{\beta}$:

- *Invertibility condition:* the matrix $\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}$ is invertible and

$$\left\| \left(\frac{\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}}{n} \right)^{-1} \right\| \leq 2.$$

- *Orthogonality condition:* the noise correlation is bounded,

$$\left\| \frac{\boldsymbol{X}^\top \boldsymbol{\xi}}{n} \right\|_\infty \leq (1/2)\lambda.$$

- *Complementary size condition:* the following inequality holds

$$\frac{1}{n}\|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{\xi}\|_\infty + 2\lambda\|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1}\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty \leq (1/2)\lambda. \tag{57}$$

The conditions stated above are proved to hold with high probability in [4]. Notice that under Assumption 2.1, Assumption 2.2 and Assumption 2.3, and that $\boldsymbol{\beta}_*$ in assumed to come from the generic s-sparse model, the original design matrix and coefficient vector $(\boldsymbol{X}, \boldsymbol{\beta})$ still satisfies these conditions. Therefore for the first two terms on the right-hand side of the inequality (55), we can repeat the steps of the proof in [4]. Define $v = \frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{\xi} - \lambda\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}})$, then

$$\begin{aligned}
&\frac{1}{n}\langle \boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\xi}\rangle + \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \\
&\leq \frac{1}{n}\langle \boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\xi}\rangle - \lambda(\langle \boldsymbol{\Delta}_{\mathcal{S}}, \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}}) + \|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1) \\
&\leq \langle \boldsymbol{\Delta}_{\mathcal{S}}, \frac{1}{n}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{\xi} - \lambda\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}})\rangle - (1 - 1/2)\lambda\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \\
&\leq \langle \boldsymbol{\Delta}_{\mathcal{S}}, v\rangle - (1 - 1/2)\lambda\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \\
&\leq C\left\| \frac{\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}}{n} \right\|_\infty \cdot s\lambda,
\end{aligned} \tag{58}$$

where the last inequality comes from the following sequence of inequalities due to Candès and Plan [4] and given here for completeness:

$$\begin{aligned}\langle \boldsymbol{\Delta}_{\mathcal{S}}, v\rangle &= \langle (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})\boldsymbol{\Delta}_{\mathcal{S}}, v\rangle \\ &= \langle (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})\boldsymbol{\Delta}_{\mathcal{S}}, (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\rangle \\ &= \langle \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}, (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\rangle - \langle (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}^c})\boldsymbol{\Delta}_{\mathcal{S}^c}, (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\rangle \\ &:= A_1 - A_2.\end{aligned} \tag{59}$$

We have that

$$\begin{aligned}A_1 &= \langle \boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}, (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\rangle \\ &\le \|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty \|(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}/n)^{-1} v\|_1 \\ &\le \sqrt{s}\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty \|(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}/n)^{-1} v\|_2 \\ &\le s\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty \|(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}}/n)^{-1}\| \|v\|_\infty \\ &\le 3\lambda_s s\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty\end{aligned} \tag{60}$$

where the last inequality coming from Lemma C.1 and the bound $\|v\|_\infty \le (1/2+1)\lambda_s$. For A_2, note that

$$\begin{aligned}|A_2| &= |\langle (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}^c})\boldsymbol{\Delta}_{\mathcal{S}^c}, (\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\rangle| \\ &= |\langle \boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v, \boldsymbol{\Delta}_{\mathcal{S}^c}\rangle| \\ &\le \|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\|_\infty.\end{aligned} \tag{61}$$

Due to the complementary size condition (C.2),

$$\begin{aligned}&\|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1} v\|_\infty \\ &\le \frac{1}{n}\|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1}\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{\xi}\|_\infty + 2\lambda\|\boldsymbol{X}_{\mathcal{S}^c}^\top \boldsymbol{X}_{\mathcal{S}}(\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}_{\mathcal{S}})^{-1}\operatorname{sign}(\boldsymbol{\beta}_{\mathcal{S}})\|_\infty \\ &\le (1/2)\lambda_s.\end{aligned} \tag{62}$$

Therefore,

$$|A_2| \le 1/2\lambda_s \|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1. \tag{63}$$

Putting together the bounds obtained above, we deduce that

$$\begin{aligned}&\langle \boldsymbol{\Delta}_{\mathcal{S}}, v\rangle - (1-1/2)\lambda\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \\ &\le 3\lambda_s s\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty + 1/2\lambda_s\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 - 1/2\lambda_s\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 \\ &\le 3\lambda_s s\|\boldsymbol{X}_{\mathcal{S}}^\top \boldsymbol{X}\boldsymbol{\Delta}/n\|_\infty.\end{aligned} \tag{64}$$

We also need the following lemma:

Lemma E.5. *The following inequality holds:*

$$\frac{1}{n}\|\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\Delta}\| \le \frac{3}{2}\lambda + \frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1. \tag{65}$$

Proof. The optimality conditions for the problem (1) give that

$$-\frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \sqrt{n}\hat{\boldsymbol{\theta}}) + \lambda\partial\|\hat{\boldsymbol{\beta}}\|_1 = 0 \tag{66}$$

Substituting $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_* + \sqrt{n}\boldsymbol{\theta}_* + \boldsymbol{\xi}$ into equation (66), rearranging terms and taking infinity norm of both sides, we have in view of the triangle inequality that

$$\frac{1}{n}\|\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\Delta}\|_\infty \le \frac{1}{n}\|\boldsymbol{X}^\top\boldsymbol{\xi}\|_\infty + \frac{1}{\sqrt{n}}\|\boldsymbol{X}^\top\boldsymbol{\Theta}\|_\infty + \lambda\|\partial\|\hat{\boldsymbol{\beta}}\|_1\|_\infty. \tag{67}$$

By orthogonality condition, with high probability $\|\boldsymbol{X}^\top\boldsymbol{\xi}\|/n \le (1/2)\lambda$. The definition of subgradient gives that the $\|\cdot\|_\infty$ norm of any vector in $\partial\|\hat{\boldsymbol{\beta}}\|_1$ is at most 1. Moreover,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}^\top\boldsymbol{\Theta}\|_\infty \le \frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1.$$

Therefore, inequality (58) yields that

$$\frac{1}{n}\langle\boldsymbol{X}\boldsymbol{\Delta},\boldsymbol{\xi}\rangle + \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1) \le C\cdot s\lambda^2 + C\cdot s\lambda\frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1. \tag{68}$$

□

It remains to check that

$$\begin{aligned}
& -\frac{1}{\sqrt{n}}\langle\boldsymbol{X}\boldsymbol{\Delta},\boldsymbol{\Theta}\rangle + C\cdot s\lambda\frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1 \\
\le{}& \frac{1}{\sqrt{n}}|\boldsymbol{\Theta}^\top\boldsymbol{X}\boldsymbol{\Delta}| + C\cdot s\lambda\frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1 \\
\le{}& \frac{1}{\sqrt{n}}\|\boldsymbol{X}\boldsymbol{\Delta}\|_\infty\|\boldsymbol{\Theta}\|_1 + C\cdot s\lambda\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1 \\
\le{}& \frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1(\|\boldsymbol{\Delta}\|_1 + C\cdot s\lambda) \\
\le{}& \frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}(\|\boldsymbol{\Delta}\|_1 + \|\boldsymbol{\Theta}\|_1)^2 + \frac{1}{\sqrt{n}}C^2\cdot s^2\lambda^2.
\end{aligned} \tag{69}$$

where the last inequality is due to the relation that for $a, b, c > 0$, $2a(b+c) \le a^2 + (b+c)^2 \le a^2 + 2b^2 + 2c^2 \le 2a^2 + 2b^2 + 2c^2 \le 2(a+b)^2 + 2c^2$.

Define $\boldsymbol{M}_{\mathcal{I}} = [\boldsymbol{X}_{\mathcal{S}}|\sqrt{n}\mathbf{I}_{\mathcal{O}}]$. Bickel et al. [1] show that whenever $\lambda \ge c\sigma\sqrt{\log(p)/n}$, $c > 2\sqrt{2}$, $(\boldsymbol{\Delta},\boldsymbol{\Theta})$ is in the cone of vectors satisfying

$$\|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_1 + \|\boldsymbol{\Theta}_{\mathcal{O}^c}\|_1 \le 3(\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + \|\boldsymbol{\Theta}_{\mathcal{O}}\|_1)$$

with probability at least $1 - p^{1-c^2}$. Therefore,

$$\begin{aligned}
(\|\boldsymbol{\Delta}\|_1 + \|\boldsymbol{\Theta}\|_1)^2 &\le 16(\|\boldsymbol{\Delta}_{\mathcal{S}}\|_1 + \|\boldsymbol{\Theta}_{\mathcal{O}}\|_1)^2 \le 16(s+o)\left\|\begin{bmatrix}\boldsymbol{\Delta}_{\mathcal{S}}\\ \boldsymbol{\Theta}_{\mathcal{O}}\end{bmatrix}\right\|_2^2 \\
&\le 16(s+o)\left\|\left(\frac{\boldsymbol{M}^\top\boldsymbol{M}}{n}\right)^{-1}\right\|\cdot\frac{1}{n}\left\|\boldsymbol{M}_{\mathcal{I}}\begin{bmatrix}\boldsymbol{\Delta}_{\mathcal{S}}\\ \boldsymbol{\Theta}_{\mathcal{O}}\end{bmatrix}\right\|_2^2 \\
&\le 32(s+o)\cdot\frac{1}{n}\|\boldsymbol{M}(\hat{\gamma} - \gamma)\|_2^2.
\end{aligned} \tag{70}$$

The last inequality was derived as follows. Denote $\boldsymbol{\gamma}_{\mathcal{I}} = \begin{bmatrix} \boldsymbol{\Delta}_{\mathcal{S}} \\ \boldsymbol{\Theta}_{\mathcal{O}} \end{bmatrix}$, $\boldsymbol{\gamma}_{\mathcal{I}^c} = \begin{bmatrix} \boldsymbol{\Delta}_{\mathcal{S}^c} \\ \boldsymbol{\Theta}_{\mathcal{O}^c} \end{bmatrix}$ and let $\hat{\boldsymbol{\gamma}}_{\mathcal{I}}$, $\hat{\boldsymbol{\gamma}}_{\mathcal{I}^c}$ be defined similarly. Then

$$\|\boldsymbol{M}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2 = \|\boldsymbol{M}_{\mathcal{I}}\boldsymbol{\gamma}_{\mathcal{I}}\|_2^2 + \|\boldsymbol{M}_{\mathcal{I}^c}\boldsymbol{\gamma}_{\mathcal{I}^c}\|_2^2 + 2\langle \boldsymbol{M}_{\mathcal{I}}\boldsymbol{\gamma}_{\mathcal{I}}, \boldsymbol{M}_{\mathcal{I}^c}\boldsymbol{\gamma}_{\mathcal{I}^c}\rangle \tag{71}$$

so that

$$2|\langle \boldsymbol{M}_{\mathcal{I}}\boldsymbol{\gamma}_{\mathcal{I}}, \boldsymbol{M}_{\mathcal{I}^c}\boldsymbol{\gamma}_{\mathcal{I}^c}\rangle| \leq \|\boldsymbol{M}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2.$$

Therefore,

$$\begin{aligned} \|\boldsymbol{M}_{\mathcal{I}}\boldsymbol{\gamma}_{\mathcal{I}}\|_2^2 = \|\boldsymbol{M}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2 - \|\boldsymbol{M}_{\mathcal{I}^c}\boldsymbol{\gamma}_{\mathcal{I}^c}\|_2^2 - 2\langle \boldsymbol{M}_{\mathcal{I}}\boldsymbol{\gamma}_{\mathcal{I}}, \boldsymbol{M}_{\mathcal{I}^c}\boldsymbol{\gamma}_{\mathcal{I}^c}\rangle \\ \leq 2\|\boldsymbol{M}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2. \end{aligned} \tag{72}$$

If instead of $(\boldsymbol{X}, \boldsymbol{\beta})$, in the previous section we showed that the same conditions hold for $(\boldsymbol{M}, \boldsymbol{\gamma})$, then we can use Theorem E.4 to derive the fact that the inequality $\frac{1}{n}\|\boldsymbol{M}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2 \leq C_1'\sigma^2 \cdot \lambda^2 \cdot (s+o)$ holds with high probability. Therefore, inequality (69) implies that

$$\begin{aligned} -\frac{1}{\sqrt{n}}\langle \boldsymbol{X}\boldsymbol{\Delta}, \boldsymbol{\Theta}\rangle + C \cdot s\lambda\frac{1}{\sqrt{n}}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{\Theta}\|_1 &\leq \left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max}(64(s+o)^2\lambda^2 + C^2s^2\lambda^2) \\ &\leq \left\|\frac{\boldsymbol{X}}{\sqrt{n}}\right\|_{\max} C'(s+o)^2\lambda^2 \end{aligned} \tag{73}$$

for some C', hence the result follows.

□

E.6. Proof of Lemma E.2

Proof. Let $i \in \mathcal{I}^c$ and recall the definition of $W_i = (\boldsymbol{M}_{\mathcal{I}_1}^\top \boldsymbol{M}_{\mathcal{I}_1})^{-1}\boldsymbol{M}_{\mathcal{I}_1}^\top \boldsymbol{M}_i$ and the event E

$$E := \{\max_{i\in\mathcal{I}^c} \|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i/n\|_2 \leq c_1(\sqrt{\log(p)})^{-1}\} \cup \{\|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \leq 2\}. \tag{74}$$

In section (B), we proved that event E occurs with high probability and that on E,

$$\max_{i\in\mathcal{I}^c} \|W_i\|_2 \leq \|(\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_{\mathcal{I}}/n)^{-1}\| \max_{i\in\mathcal{I}^c} \|\boldsymbol{M}_{\mathcal{I}}^\top \boldsymbol{M}_i/n\|_2 \leq (c_2\sqrt{\log(p)})^{-1}. \tag{75}$$

Therefore, for each $i \in \mathcal{I}^c$,

$$\check{\boldsymbol{z}}_i = \left\langle W_i, \begin{bmatrix} \text{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) \\ 0 \end{bmatrix} \right\rangle + \boldsymbol{M}_i^\top \Pi_{\mathcal{I}^\perp}\left(\frac{\boldsymbol{\xi}}{\lambda_i n}\right), \tag{76}$$

where $\Pi_{\mathcal{I}^\perp} = \boldsymbol{I} - \Pi_{\mathcal{I}}$ is the projection matrix onto the subspace perpendicular to column space of $\boldsymbol{M}_{\mathcal{I}}$.

For the first term, we see that

$$\left|\left\langle W_i, \begin{bmatrix} \text{sign}(\boldsymbol{\gamma}_{\mathcal{I}}) \\ 0 \end{bmatrix} \right\rangle\right| = \langle (W_i)_{\mathcal{I}}, \text{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\rangle.$$

Employing Hoeffding's inequality, we see that on event E, for all i and any $t > 0$,

$$\begin{aligned}\mathbb{P}\left(|\langle (W_i)_{\mathcal{I}}, \operatorname{sign}(\boldsymbol{\gamma}_{\mathcal{I}})\rangle| > t|\right) \leq 2e^{-t^2/2\|W_i\|_2^2} &\leq 2e^{-t^2/2\max_i \|W_i\|_2^2} \\ &\leq 2e^{-c_2^2 t^2 \log(p)/2} = p^{-c_2 t^2/2}. \quad (77)\end{aligned}$$

Next, take $t = 1/4$ and apply the union bound to see that with probability at least $1 - 4p^{1-c_2^2 t^2/2} - \mathbb{P}(E)$,

$$\max_{i \in \mathcal{I}_1^c} |\langle W_i, \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{I}})\rangle| \leq \frac{1}{4}. \quad (78)$$

For the second term in (76), since the eigenvalues of projection matrices can be only 0 or 1,

$$\max_{i \in \mathcal{I}_1^c} \left| \boldsymbol{M}_i^\top \Pi_{\mathcal{I}^\perp} \left(\frac{\boldsymbol{\xi}}{\lambda_i n} \right) \right| \leq \left\| \boldsymbol{M}^\top \left(\frac{\boldsymbol{\xi}}{n} \right) \right\|_\infty \leq 5/8 \quad (79)$$

with probability at least $1 - 4p^{-1}$ in view of Lemma C.1 in the appendix. We conclude that for each $i \in \mathcal{I}^c$ and on the event E

$$|\breve{\boldsymbol{z}}_i| \leq 1/4 + 5/8 < 1. \quad (80)$$

□

F. Numerical Studies

We performed a numerical simulation to illustrate the obtained theoretical results and demonstrate the robustness of augmented Lasso estimator. We set $n = 1000$ and $p = 1100$, and performed two experiments. For the first experiment, sparsity level $s \in \{5, 15, 25\}$ and standard deviation of noise ξ is $\sigma = 1$. For the second experiment, sparsity level $s = 25$ and the noise standard deviation $\sigma \in \{0.1, 1, 2, 5\}$. The parameters of the problem are defined as follows: sample size $n = 1000$, dimension of the vector of regression coefficients $p = 1100$, and the fraction of outliers $o/n \in (0, 0.15)$ with an increment 0.03. The design matrix X has independent rows sampled from the multivariate normal distribution $\sim \mathcal{N}(0, \Sigma)$ where Σ has block-diagonal structure consisting of 5×5 identical blocks (220 blocks overall) with off-diagonal elements equal to $1/\log(p) \approx 0.14$ and diagonal elements being 1. The ℓ_2 norm of the columns of X has been normalized to be $\sqrt{n} = \sqrt{1000} \approx 31.62$. The optimization problem (1) was solved using the **cvxr** package, where the tuning parameter λ is set as $\lambda = 4\sqrt{2(\log(p)/n + \log(n)/n)} \approx 0.66$. Moreover, **(i)** the vector of coefficients β and the vector of outliers θ is drawn from the generic s and o models with magnitudes of the coefficients having uniform distributions $\operatorname{Unif}(8, 16)$ and $\operatorname{Unif}(16, 32)$ respectively, and **(ii)** the fraction o/n of outliers ranging between 0 and 0.15. The mean squared error (MSE) was approximated using 10 independent repetitions of the experiment. Results of the simulation are shown in figure Figure 1 and Figure 2.

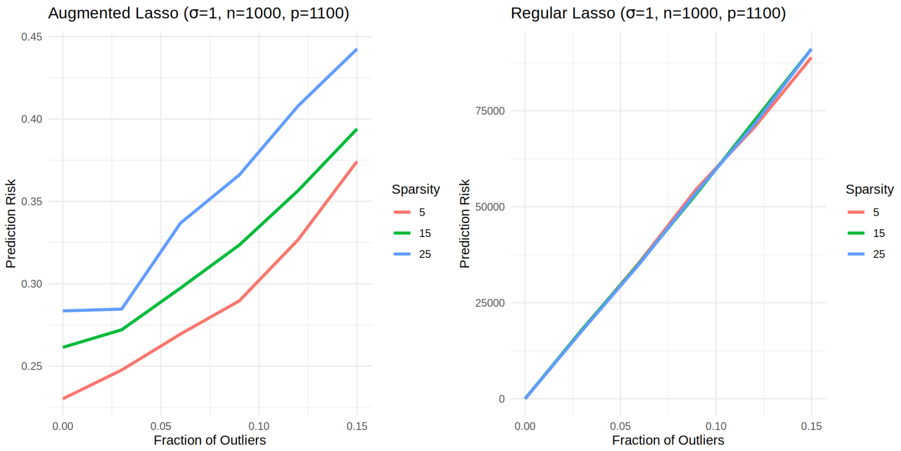

Figure 1: Inspection of the plot reveals that the prediction risk of the standard Lasso (right) increases as the number of outliers grows, but the risk of a robust version of Lasso grows much slower, thus confirming the theoretical results of the paper.

References

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

[2] Bourin, J.-C. and Lee, E.-Y. (2012). Unitary orbits of Hermitian operators with convex or concave functions. Bulletin of the London Mathematical Society, 44(6):1085–1102.

[3] Candes, E. J. and Plan, Y. (2011). A probabilistic and ripless theory of compressed sensing. IEEE Transactions on Information Theory, 57(11):7235–7254.

[4] Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by ℓ_1 minimization. The Annals of Statistics, 37(5A).

[5] Jogdeo, K. and Samuels, S. M. (1968). Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. The Annals of Mathematical Statistics, 39(4):1191–1195.

[6] Ruetz, S. and Schnass, K. (2021). Submatrices with nonuniformly selected random supports and insights into sparse approximation. SIAM Journal on Matrix Analysis and Applications, 42(3):1268–1289.

[7] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Transactions on Information Theory, 55(5):2183–2202.

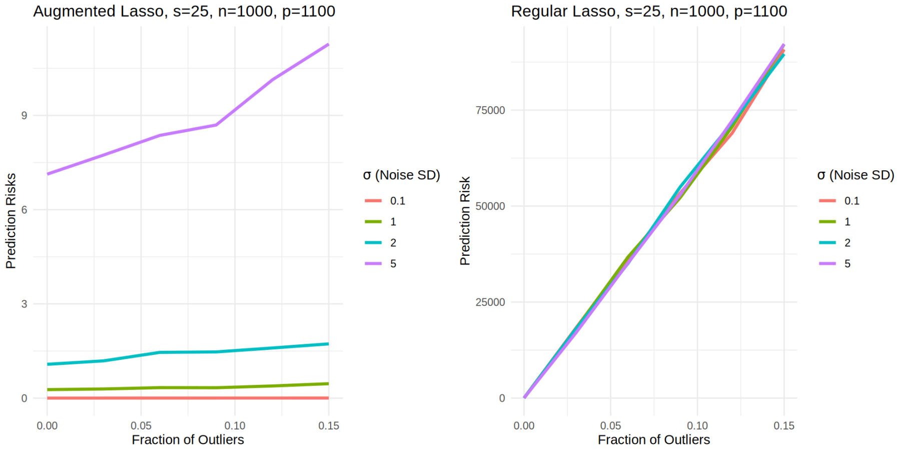

Figure 2: Inspection of the plot reveals that the risk of robust version of Lasso grows linearly with the noise variance, with nearly exact recovery when noise is close to 0.