Robust Estimation of a Sparse Linear Model: Provable Guarantees with Non-convexity

Deepak Maurya
Purdue University

Adarsh Barik
Indian Institute of Technology Delhi

Jean Honorio
The University of Melbourne, and ARC OPTIMA

Abstract

In this paper, we address the problem of sparse regression vector estimation in the presence of corrupted samples, with a particular focus on accurately identifying the support. Traditional methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO), often fail in such scenarios, exhibiting inconsistency. To tackle this challenge, we propose a combinatorial, non-convex, and robust variant of the LASSO framework, designed to enhance estimation accuracy under corruption. Our approach is supported by theoretical guarantees, which establish its reliability and robustness. Our method also handles corruption from heavy-tailed distributions, with only a few bounded moments. We validate our theoretical results through extensive experiments, comparing the performance of our method against the LASSO and its other robust variants. These comparisons highlight the efficacy of our framework, demonstrating its practical applicability in sparse regression tasks involving corrupted data.

1 INTRODUCTION

Many modern-day machine learning problems employ linear regression as a fundamental technique. In high-dimensional settings, sparse linear regression is used to model the response variable using high-dimensional predictors to leverage sparsity of the parameter vector. At its core, the task is to estimate a sparse regression parameter vector using a given dataset of predictors and response variables. To promote sparsity, the most intuitive approach is to use LASSO, i.e., add an ℓ_1-regularized term to the least square loss minimization. However, in real-world scenarios, oftentimes we receive samples that do not conform to the linear regression model due to natural or adversarial corruption of samples (Hampel et al., 1986). Multiple studies have explored the robustness of LASSO-type methods in the presence of corrupted samples, and the findings suggest that they are not robust (Hadi and Chatterjee, 2009). Fan and Li (2001); Zou (2006) have shown that variable selection for LASSO can be inconsistent. While Fan and Li (2001) introduced a non-convex penalized likelihood-based method, Zou (2006) proposed the adaptive LASSO. Other researchers, such as Lambert-Lacroix and Zwald (2011); Sun et al. (2020); Dalalyan and Thompson (2019); d'Orsi et al. (2021) have used Huber criterion (Huber, 2011) to deal with corrupted samples.

Robust regression has been a well-studied area in the literature. Readers are encouraged to consult the following works, along with their references, for further insights, though the list is not exhaustive: Bhatia et al. (2015, 2017); Suggala et al. (2019); Prasad et al. (2020); Gao (2020); Awasthi et al. (2022). Similarly, the concept of robustness has also garnered significant attention in the context of sparse linear regression: Nguyen and Tran (2012); Chen et al. (2013); Lozano et al. (2016); Dalalyan and Thompson (2019); Liu et al. (2020); d'Orsi et al. (2021); Norman et al. (2023). Most of these works, besides Chen et al. (2013), focus on providing estimation guarantees aimed at minimizing the estimation error in the ℓ_2-norm, with sample complexity depending on factors such as sparsity, problem dimension, and the number of corrupted samples. Additionally, some studies, such as Nguyen and Tran (2012); Chen et al. (2013), impose further restrictions on the proportion of corrupted samples for their recovery guarantees.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Motivation In this work, our primary goal is to estimate a sparse regression vector and accurately recover its support within the framework of robust LASSO. These objectives have been extensively explored in the context of traditional LASSO, with well-established sample complexity results (Wainwright, 2009; Meinshausen and Bühlmann, 2006). For exact support recovery in LASSO, the sample complexity exhibits polynomial dependence on sparsity while scaling logarithmically with the problem's dimension, enabling existing algorithms to perform effectively in high-dimensional settings.

The seminal work of Chen et al. (2013) introduced non-convex trimming methods to address support recovery under the presence of a small number of corrupted samples. Crucially, their analysis highlighted that convex optimization frameworks are inherently unsuitable for achieving robust support recovery, thereby motivating the need for non-convex formulations. Inspired by this insight, we propose a non-convex approach tailored for robust support recovery. However, non-convex optimization poses inherent challenges, particularly due to the lack of general methods capable of finding global optima efficiently.

To navigate these challenges, our approach focuses on analyzing local solutions to the proposed non-convex problem. We demonstrate that certain local solutions, which satisfy *integrality* properties, are sufficiently close to the ground truth. This allows us to achieve robust support recovery without requiring global optimality. Furthermore, we provide rigorous theoretical guarantees for our approach, addressing both the ℓ_2 estimation error of the sparse regression vector and the exact recovery of its support. Importantly, our method retains desirable sample complexity properties, with polynomial dependence on sparsity and logarithmic dependence on dimensionality, aligning closely with the characteristics of standard LASSO.

1.1 Main contributions and discussion

In this subsection, we present the central results of this work and discuss their significance in the broader context of existing literature. The objective in the subsequent discussion is to estimate a p-dimensional, k-sparse regression vector using n samples, where up to an ε-proportion of the samples may be corrupted.

1. **Exact support recovery:** Our first result establishes the necessary technical condition for achieving exact support recovery in the presence of corrupted samples. We demonstrate that exact support recovery using some specific locally optimal solutions is possible with high probability when the number of samples satisfies $n = \Omega\left(\frac{k^3 \log p}{\varepsilon^2 \log^2 \varepsilon^{-1}}\right)$.

   This finding is particularly noteworthy because it shows that the analysis can be performed on any of the *integral* local optimal points of a non-convex optimization problem, rather than requiring the problem to be solved to global optimality. This broadens the applicability of the method and provides a practical pathway for analysis. Note the polynomial dependence on k and logarithmic dependence on p, which we target in this work.

   The readers might also notice the seemingly counterintuitive dependence of sample complexity on ε. This relationship, however, arises naturally from our analysis and is a well-documented phenomenon in robust statistics. For example, in the context of robust linear regression, Theorem 1.3 in Diakonikolas et al. (2019b) demonstrates a sample complexity lower bound of $\Omega\left(\frac{p}{\varepsilon^2}\text{ polylog}\left(\frac{p}{\varepsilon\delta}\right)\right)$ to recover the regression vector with a confidence level of at least $1-\delta$, when ε-fraction of the samples are corrupted. This inverse relationship between n and ε is also evident in other works, such as Corollary 3.1 in Balakrishnan et al. (2017) and Theorem 4.1 in Liu et al. (2020). Broadly, this dependence stems from the inherent challenge of managing the interplay between corrupted and uncorrupted data. Our analysis carefully balances the distributional properties of two subsets: one containing $\sim (1-\varepsilon)n$ uncorrupted measurements, and the other comprising εn corrupted measurements. By rigorously accounting for the impact of corruption, we derive results that align with established findings in the robust statistics literature.

   Our results hinge on two critical technical criteria:

   (a) To achieve exact support recovery, each non-zero entry in the ground truth $\theta^* \in \mathbb{R}^p$ must satisfy a "minimum weight criterion", which is influenced by the corruption proportion. Specifically, the non-zero entries must be $\Omega\left(\varepsilon \log(1/\varepsilon)\right)$ to counteract the effects of corruption in the observed data. While this criterion is a standard condition in the sparse linear regression literature, as highlighted in prior works such as Wainwright (2009, 2019); Ravikumar et al. (2007, 2010); Daneshmand et al. (2014); Maurya and Honorio (2024), the dependence on ε is unique to our setup. These studies emphasize the necessity of ensuring that true signal components are distinguishable from noise for accurate support identification.
   The central insight in our work is that when the non-zero entries of θ^* exceed this threshold, effectively nullifying the influence of cor-

ruption. This shows that the locally optimal solutions of the proposed non-convex optimization problem are sufficient for exact support recovery. This eliminates the need to achieve global optimality, which is computationally challenging in non-convex settings. By leveraging local solutions, we provide a more practical and computationally feasible pathway for recovering the support, thereby broadening the applicability of our framework to real-world scenarios where corruption is inevitable.

(b) Our analysis further necessitates a mutual-incoherence condition that explicitly accounts for the proportion of corrupted samples in the data. While the mutual-incoherence requirement originates from the classical LASSO framework Wainwright (2009, 2019); Ravikumar et al. (2007, 2010); Daneshmand et al. (2014); Maurya and Honorio (2024), our formulation tailors it to address the challenges introduced by the presence of corruption.

On a high level, this condition ensures that the predictors corresponding to the true support exert a significantly stronger influence on the predictions than the predictors outside the support. Furthermore, we also need the mutual incoherence coefficient to grow as $\omega\left(\varepsilon \log(1/\varepsilon)\right)$, i.e., as the number of corrupted samples increases, the disparity in this influence must grow proportionally. This scaling ensures that the true signal remains distinguishable despite the presence of substantial corruption.

Our insights highlight the interplay between corruption and predictor influence, offering a nuanced understanding of how sparsity and robustness can coexist in high-dimensional regression problems. Our results are expressed directly in terms of the corruption proportion ε, making the dependency on corruption both explicit and transparent.

2. **Estimation error:** Our work establishes an ε-dependent guarantee on the estimation error. Specifically, we demonstrate that when the sample size satisfies $n = \Omega\left(\frac{k^3 \log p}{\varepsilon^2 \log^2 \varepsilon^{-1}}\right)$, any integral local solution is guaranteed to lie within a distance of $\mathcal{O}\left(\sqrt{k}\varepsilon \log(1/\varepsilon)\right)$ from θ^* in the ℓ_2-norm.

3. **Algorithmic contribution:** Our results are presented for algorithms that yield an integral local optimal solution, emphasizing their theoretical and practical significance. However, it is worth noting that typically first-order methods, while capable of converging to a local optimal point, do not inherently guarantee convergence to an integral local optimal point. This distinction is crucial, as integral local optimal points form the foundation for our theoretical guarantees.

   To address this gap and enhance the practical applicability of our approach, we propose a straightforward yet effective black-box construction of an algorithm that outputs an integral local optimal point while using any first-order method as a subroutine. This enhancement guarantees that the solution is an integral local optimal point. By integrating this construction, our framework bridges the theoretical guarantees with real-world algorithmic efficiency, making it accessible for practical implementation while maintaining the robustness and accuracy promised by our theoretical analysis.

4. **Corruption in labels and features:** We provide theoretical guarantees when the corruption can be in both labels and features. Our method handles corruption from any heavy-tailed distribution with only a few bounded moments using a pre-processing step similar to Awasthi et al. (2022). In our pre-processing step, we choose thresholds carefully based on the moments of features, which can be of independent interest, as it is not restricted to sparse linear models.

Our theoretical claims are also validated empirically on synthetic datasets.

2 DATA GENERATION PROCESS

In this section, we introduce the data-generation process for the sparse linear regression problem under the presence of corruption. The uncorrupted samples are generated according to the following procedure:

$$(\forall i \in [n]) \quad y_i = \langle X_i \, , \theta^* \rangle + e_i \ .$$

We assume that $X_i \in \mathbb{R}^p$ is a zero-mean sub-Gaussian random vector (Hsu et al., 2012) with covariance matrix Σ. The additive noise e_i is independently sampled from a zero-mean sub-Gaussian distribution with parameter σ_e. As a result, the samples $(X_i, y_i)_{i\in[n]}$ are generated independently and identically distributed (i.i.d.). The regression vector $\theta^* \in \mathbb{R}^p$ is assumed to be k-sparse, meaning it contains at most k non-zero entries.

Corruption Model: We adopt the strong corruption model from Awasthi et al. (2022) to introduce noise into the dataset. Specifically, the adversary has access to all clean samples and εn samples from $(X_i, y_i)_{i\in[n]}$ can be arbitrarily corrupted.

The corruption proportion ε is then defined as $\varepsilon = \frac{n-m}{n}$, where m denotes the number of clean samples. We

assume $\varepsilon < 0.5$. This model operates under a strong corruption assumption, meaning that any $n - m$ out of the n samples can be arbitrarily corrupted, allowing for maximal flexibility. With the above setup, we are interested in the following question:

Given the ε-corrupted dataset $(X_i, y_i)_{i\in[n]}$, is it possible to have a good sparse estimate of θ^*?

3 A NON-CONVEX FORMULATION

In this section, we propose a non-convex formulation for outlier-robust estimation of the sparse regression parameter vector. Our task is to estimate θ^* using $(X_i, y_i)_{i\in[n]}$. At a high level, we design a combinatorial problem where we choose m out of n available samples.

$$\min_{\mathcal{Q}\subseteq[n],|\mathcal{Q}|\geq m,\theta\in\mathbb{R}^p} \frac{1}{2m}\sum_{i\in\mathcal{Q}}(y_i - \langle X_i\,,\theta\rangle)^2 + \lambda\|\theta\|_1$$

Mathematically, we can formulate the following continuous relaxation of this combinatorial optimization problem:

$$\min_{b\in\mathbb{R}^n,\theta\in\mathbb{R}^p} \frac{1}{2m}\sum_{i=1}^{n} b_i(y_i - \langle X_i\,,\theta\rangle)^2 + \lambda\|\theta\|_1 \quad (1)$$
$$\text{such that, } \sum_{i=1}^{n} b_i = m$$
$$0 \leq b_i \leq 1, \quad \forall i \in [n],$$

where $\lambda \geq 0$ is a regularizer which along with ℓ_1-norm on θ ensures the sparsity of the solution. At a high level, we want $b_i = 1$ to denote an uncorrupted sample. Ideally, we expect exactly m b_i's to take the value 1 and the remaining b_i's to take the value 0.

One can easily show that the optimization problem (1), even with continuous variables, is jointly non-convex with respect to b and θ.

Proposition 1. *The optimization problem (1) is non-convex with respect to $b_1, \ldots, b_n, \theta$.*

Proof. It suffices to show the non-convexity of the function $g(b,\theta) = \sum_{i=1}^{n} b_i(y_i - \langle X_i, \theta\rangle)^2$, which can be viewed as a special case when $\lambda = 0$. First, we compute the partial derivatives of g:

$$\frac{\partial g}{\partial b_i} = (y_i - \langle X_i, \theta\rangle)^2, \quad \forall i \in [n],$$
$$\frac{\partial g}{\partial \theta} = -2\sum_{i=1}^{n} b_i X_i(y_i - \langle X_i, \theta\rangle)\ .$$

Table 1: KKT conditions

KKT Condition	Equation	
Stationarity:	$\frac{1}{2m}(y_i - \langle X_i, \theta\rangle)^2 + \mu - \alpha_i + \beta_i = 0, \forall i \in [n]$	(2)
	$-\frac{1}{m}\sum_{i=1}^{n} b_i(y_i - \langle X_i, \theta\rangle)X_i + \lambda\zeta = 0$	(3)
Complementary slackness:	$\mu\left(\sum_{i=1}^{n} b_i - m\right) = 0$	(4)
	$\alpha_i b_i = 0, \ \forall i \in [n]$	(5)
	$\beta_i(b_i - 1) = 0, \ \forall i \in [n]$	(6)
Dual feasibility:	$\alpha_i \geq 0, \ \beta_i \geq 0, \ \forall i \in [n]$	(7)
Primal feasibility:	$\sum_{i=1}^{n} b_i = m$	(8)
	$0 \leq b_i \leq 1, \ \forall i \in [n]$	(9)

Next, we define the function $G(b, \bar{b}, \theta, \bar{\theta})$ as follows:

$$G(b,\bar{b},\theta,\bar{\theta}) = g(b,\theta) - g(\bar{b},\bar{\theta}) - \sum_{i=1}^{n}\frac{\partial g(\bar{b},\bar{\theta})}{\partial b_i}(b_i - \bar{b}_i)$$
$$- \langle\frac{\partial g(\bar{b},\bar{\theta})}{\partial\theta}, \theta - \bar{\theta}\rangle.$$

To prove the non-convexity of $g(b,\theta)$, we aim to construct specific instances of (b,θ) and $(\bar{b},\bar{\theta})$ such that $G(b,\bar{b},\theta,\bar{\theta})$ takes both positive and negative values. Consider the following assignments:

$$b = (0,\ldots,0), \quad \bar{b} = \left(\frac{1}{2},\ldots,\frac{1}{2}\right),$$
$$\theta_i = \begin{cases} u, & \text{if } i = t, \\ 0, & \text{otherwise,} \end{cases} \qquad \bar{\theta}_i = \begin{cases} w, & \text{if } i = t, \\ 0, & \text{otherwise.} \end{cases}$$

Under these settings, the function simplifies to: $G(b,\bar{b},\theta,\bar{\theta}) = -\sum_{i=1}^{n}(y_i - X_{it}w)X_{it}(u - w)$.

If $X_{it} = 0$, we can modify the index t to ensure that $X_{it} \neq 0$. By fixing w, we can select an appropriate value of u such that $G(b,\bar{b},\theta,\bar{\theta})$ alternates between positive and negative values. This establishes the non-convexity of $g(b,\theta)$. □

Due to the inherent non-convexity of problem (1), attaining a global optimal solution is generally not possible. Consequently, our focus shifts to identifying local optimal solutions, often referred to as Karush-Kuhn-Tucker (KKT) points. These solutions are characterized by satisfying the KKT conditions, which serve as necessary optimality criteria for constrained optimization problems. Below, we outline the complete set of KKT conditions for problem (1).

In Table 1, $\mu, \alpha_i, \beta_i, i \in [n]$ are dual variables and ζ denotes an element of subdifferential set of $\|\theta\|_1$. The analysis presented in the following sections relies

extensively on these KKT conditions. Specifically, we introduce a distinct subset of KKT points, referred to as "integral" KKT points, which play a critical role in our theoretical and practical framework.

Definition 1 (Integral KKT points)**.** A KKT point (b, θ) that satisfies all the conditions listed in Table 1 is referred to as an integral KKT point if $b_i \in \{0, 1\}$ for all $i \in [n]$.

If a KKT point is not an integral KKT point, we call it non-integral KKT point. We now present and prove several properties related to the integral properties of KKT points of problem (1), which are instrumental to our analysis.

Lemma 1. *Let (b, θ) be a non-integral KKT point of problem (1), then $\exists$ a primal feasible point $(\bar{b}, \theta)$ satisfying (8) and (9) such that $\bar{b}_i \in \{0, 1\}, \forall i \in [n]$ and*

$$\sum_{i=1}^{n} b_i (y_i - \langle X_i \, , \theta \rangle)^2) = \sum_{i=1}^{n} \bar{b}_i (y_i - \langle X_i \, , \theta \rangle)^2) \; . (10)$$

Proof. We define a set P in the following way: $P = \{i \in [n] \mid 0 < b_i < 1\}$. It follows from equations (5) and (6) that $\alpha_i = 0$ and $\beta_i = 0$ for all $i \in P$. Therefore, following equation(2):

$$\frac{1}{2m}(y_i - \langle X_i \, , \theta \rangle)^2 = -\mu, \quad \forall i \in P \tag{11}$$

Moreover, since $\sum_{i=1}^{n} b_i = m$, it immediately follows that $\sum_{i \in P} b_i$ is an integer. We pick $\sum_{i \in P} b_i$ entries from P and for each such entry j, assign $\bar{b}_j = 1$. For the remaining entries in P, we assign $\bar{b}_j = 0$. For all other indices j that are not in P, $\bar{b}_j = b_j$. It is easy to see that $(\bar{b}, \theta)$ is primal feasible. Additionally, equation (11) ensures that equation (10) holds. □

The result of Lemma 1 is useful in developing an algorithm that always produces an integral KKT point of problem (1). Next, let $b^* \in \{0, 1\}^n$ is defined such that $b_i^* = 1$ if i is an uncorrupted sample and $b_i^* = 0$, otherwise. At a high level, the next lemma establishes local optimality of the integral KKT points.

Lemma 2. *Let (b, θ) be an integral KKT point, then*

$$\sum_{i=1}^{n} b_i (y_i - \langle X_i \, , \theta \rangle)^2 \leq \sum_{i=1}^{n} b_i^* (y_i - \langle X_i \, , \theta \rangle)^2 \tag{12}$$

Proof. Let $P = \{i \in [n] \mid b_i = 1\}$. Since (b, θ) is an integral KKT point, we have that $P_c = \{i \in [n] \mid b_i = 0\} = [n] \setminus P$. Using equation (5), (8), and multiplying equation (2) with b_i and summing it over $i \in [n]$:

$$\frac{1}{2m} \sum_{i=1}^{n} b_i (y_i - \langle X_i \, , \theta \rangle)^2 = -\mu m - \sum_{i \in P} \beta_i \; .$$

Similarly, multiplying equation (2) with b_i^* and summing it over $i \in [n]$, we get:

$$\frac{1}{2m} \sum_{i=1}^{n} b_i^* (y_i - \langle X_i \, , \theta \rangle)^2 = -\mu m - \sum_{i \in P} \beta_i b_i^* + \sum_{i \in P_c} \alpha_i b_i^*$$

Since $b_i^* \in \{0, 1\}$, $\sum_{i \in P} \beta_i b_i^* \leq \sum_{i \in P} \beta_i$ and $\sum_{i \in P_c} \alpha_i b_i^* \geq 0$, the final result follows. □

4 ESTIMATION AND SUPPORT RECOVERY

In this section, we present our main result and provide a comprehensive theoretical analysis. Our approach is based on the well-established primal-dual witness framework Wainwright (2009); Ravikumar et al. (2007, 2010); Daneshmand et al. (2014). However, our analysis incorporates additional arguments to address the challenges posed by suboptimal local solutions and the presence of corrupted samples in the dataset. While we build upon the standard assumptions of the primal-dual witness framework, these assumptions are adapted to ensure robustness against data corruption. Before detailing the assumptions, we clarify some key notations. Specifically, we define the support of the true regression vector θ^* as the set S, where $S = \{i \in [p] \mid \theta_i^* \neq 0\}$. The set S^c represents the complement of S, defined as $S^c = [p] \setminus S$. For a matrix $A \in \mathbb{R}^{p \times q}$, its restriction to rows $B \subseteq [p]$ and columns $C \subseteq [q]$, where $|B| = r$ and $|C| = s$, is denoted as $A_{BC} \in \mathbb{R}^{r \times s}$. Let Σ denote the population covariance matrix of the samples. The analysis in this work adopts the following assumptions, borrowed from the primal-dual witness framework, to facilitate robust estimation and support recovery.

Assumption 1 (Positive Definiteness)**.** There exists an $\alpha_1 = \Omega\left(1 + \frac{\varepsilon \log(1/\varepsilon)}{1-\varepsilon}\right)$, and $\alpha_2 > 0$ such that $\alpha_1 I \preceq \Sigma_{SS} \preceq \alpha_2 I$, where I is an identity matrix.

Assumption 2 (Mutual Incoherence)**.** For some $\kappa \in (0, 1]$ with $\kappa = \omega(\varepsilon \log(1/\varepsilon))$, $\|\Sigma_{S^c S} \Sigma_{SS}^{-1}\|_\infty \leq 1 - \kappa$, where $\|\cdot\|_\infty$ denotes matrix-induced ℓ_∞-norm.

Assumption 3 (Minimum Weight Assumption)**.** For each entry in the support, i.e., $\forall i \in S$, $|\theta_i^*| = \Omega\left(\sqrt{\frac{k \log(p)}{n}} + \varepsilon \log(1/\varepsilon)\right)$.

Note the presence of ε in Assumptions 1, 2, and 3. This dependence is to overcome the effects of the corrupted labels in the dataset. With these assumption in place, we are ready to state our main result for the simpler case of corruption only in the labels y. We discuss similar results for the case of corruption in both labels and features later in Section 6.

Theorem 1 (Informal). *Let $(\hat{b}, \theta)$ be an integral KKT point of the problem (1). Under assumptions 1, 2 and 3, and choosing $\lambda = \Omega\left(\frac{(2-\kappa)\sigma\sigma_e\sqrt{k}\varepsilon\log(1/\varepsilon)}{\kappa}\right)$ and $n = \Omega\left(\frac{k^3\log p}{\varepsilon^2\log^2(1/\varepsilon)}\right)$, we claim following with probability at least $1 - \mathcal{O}(1/p)$:*

1. *Support of θ matches exactly with support of θ^*.*
2. *Furthermore, the estimation error between θ and θ^* is bounded by*

$$\|\theta - \theta^*\|_\infty \leq \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon)\right) . \tag{13}$$

Proof Overview. The proof primarily adheres to the standard primal-dual witness framework. We begin by assuming that the true support of θ^*, denoted as S, is known and that the given solution θ shares the same support S. The goal is to construct a primal-dual witness solution $\left((\hat{b}, \theta), \zeta\right)$ and demonstrate that $\|\zeta_S\|_\infty \leq 1$ and $\|\zeta_{S^c}\|_\infty < 1$. The first inequality is trivial, as recall that ζ is an element of the subdifferential set of $\|\theta\|_1$. Therefore, it follows that $\|\zeta\|_\infty \leq 1$. For a comprehensive introduction to the primal-dual witness method, readers may refer to Wainwright (2009); Ravikumar et al. (2007, 2010).

The primary challenge in this proof arises from the presence of corrupted samples. To address this, we partition the n samples into four disjoint sets. These sets form the foundation for managing the complexities introduced by label corruption and for ensuring the validity of the constructed solution. The sets are defined as follows:

$$\begin{aligned}
J^{cc} &= \left\{(X_i, y_i) \mid b_i^* = 1, \hat{b}_i = 1, i \in [n]\right\}, \\
J^{co} &= \left\{(X_i, y_i) \mid b_i^* = 1, \hat{b}_i = 0, i \in [n]\right\}, \\
J^{oc} &= \left\{(X_i, y_i) \mid b_i^* = 0, \hat{b}_i = 1, i \in [n]\right\}, \\
J^{oo} &= \left\{(X_i, y_i) \mid b_i^* = 0, \hat{b}_i = 0, i \in [n]\right\}.
\end{aligned}$$

For convenience, we also define $\hat{J}^c = \left\{(X_i, y_i) \mid \hat{b}_i = 1, i \in [n]\right\} = J^{cc} \cup J^{oc}$. Using the stationarity condition (3), we can write:

$$\begin{aligned}
-\frac{1}{m}\sum_{(X_i,y_i)\in\hat{J}^c} X_i(y_i - \langle X_i\,, \theta\rangle) + \lambda\zeta = 0\,, \\
-\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} X_i(y_i - \langle X_i\,, \theta\rangle) + \lambda\zeta = \\
\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} X_i(y_i - \langle X_i\,, \theta\rangle)
\end{aligned} \tag{14}$$

For all $(X_i, y_i) \in J^{cc}$, we can substitute $y_i = \langle X_i\,, \theta^*\rangle + e_i$. Observing that θ has support S, we can write equation (14) in two parts. First part only considers the entries in S:

$$\begin{aligned}
\widehat{\Sigma}_{SS}^{J^{cc}}\,(\underline{\theta} - \underline{\theta}^*) = -\lambda\zeta_S + \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i \\
+\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,, \underline{\theta}\rangle)
\end{aligned} \tag{15}$$

where $\widehat{\Sigma}^{J^{cc}} = \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} X_i X_i^T$ and $\underline{\theta} \in \mathbb{R}^k$, $\underline{\theta}^* \in \mathbb{R}^k$, and $\widetilde{X}_i \in \mathbb{R}^k$ denote θ, θ^* and X_i restricted to the entries in S respectively. Equation (15) only considers the entries in S. Similarly, equation (16) below only considers the entries in S^c:

$$\begin{aligned}
\zeta_{S^c} = \frac{1}{\lambda}\Bigg(-\widehat{\Sigma}_{S^cS}^{J^{cc}}(\underline{\theta} - \underline{\theta}^*) + \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \overline{X}_i e_i \\
+\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \overline{X}_i(y_i - \langle\widetilde{X}_i\,, \underline{\theta}\rangle\Bigg)
\end{aligned} \tag{16}$$

where $\overline{X}_i$ denotes X_i restricted to the entries in S^c. Equation (15), after using norm inequalities, provides ℓ_∞ error bound for entries in the support:

$$\begin{aligned}
\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \underbrace{\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_\infty}_{G_1}\Big(\|\lambda\zeta_S\|_\infty \\
\underbrace{+\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty}_{G_2\text{:due to noise}} \\
\underbrace{+\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,, \underline{\theta}\rangle)\|_\infty}_{G_3\text{:due to corruption}}\Big) .
\end{aligned}$$

Observe that $\|\lambda\zeta_S\|\infty \leq \lambda$, which follows directly from the definition of ζ_S. With our proposed choice of n (note that $m = (1-\varepsilon)n$), controlling G_1 and G_2 is relatively straightforward. This is because neither of these terms involves any corrupted samples. Additionally, X_i and e_i are drawn from subGaussian distributions, ensuring their tails are well-behaved. Furthermore, the spectral properties of Σ_{SS}, as guaranteed by Assumption 1, facilitate the analysis by providing stability and boundedness for $\widehat{\Sigma}$ (See Lemmas 5 and 6).

Bounding G_3 becomes a little trickier as it involves corrupted samples (details in Lemma 7). By changing norms, observe that :

$$G_3 \leq \frac{\varepsilon}{(1-\varepsilon)}\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,, \underline{\theta}\rangle)\|_2.$$

Further, we use the Cauchy-Schwarz inequality to get:

$$G_3 \leq \frac{\varepsilon}{(1-\varepsilon)} \left(\underbrace{\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} \left\| \widetilde{X}_i \right\|_2^2}_{G_{31}} \right)^{1/2} \left(\underbrace{\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle)^2}_{G_{32}} \right)^{1/2} \tag{17}$$

The first term G_{31}, can be bounded using the resilience condition from Awasthi et al. (2022). The second term G_{32} still contains the corrupted labels. We utilize the result from Lemma 2 to deal with G_{32}. Specifically, after removing entries in J^{cc} from both sides of inequality (12), we get

$$\begin{aligned} G_{32} &= \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle)^2 \\ &\leq \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{co}} (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle)^2 \, . \end{aligned} \tag{18}$$

Observe that the RHS in equation (18) does not depend on corrupted samples and thus can be handled with appropriate subGaussian bounds. Combining all the results together, we arrive at $\|\underline{\theta} - \underline{\theta}^*\|_\infty = \mathcal{O}\left(\sigma\varepsilon \log(1/\varepsilon)\right)$ for sufficiently small ε satisfying $C\sqrt{k}\sigma\sigma_e\varepsilon\log(1/\varepsilon) < 1/2$. For more details, kindly refer to proof of Lemma 4 in the appendix.

With equation (13) in place, we turn our attention to equation (16). Using (15) we can write:

$$\begin{aligned} \|\zeta_{S^c}\|_\infty = \| \underbrace{\widehat{\Sigma}^{J^{cc}}_{S^cS} \widehat{\Sigma}^{J^{cc}-1}_{SS}}_{G_4\text{: mutual incoherence}} \zeta_S - \frac{\widehat{\Sigma}^{J^{cc}}_{S^cS} \widehat{\Sigma}^{J^{cc}-1}_{SS}}{\lambda} \\ \left(\frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle) + \frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i \right) \\ + \frac{1}{\lambda} \left(\frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \overline{X}_i e_i + \frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \overline{X}_i (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle \right) \|_\infty . \end{aligned} \tag{19}$$

Leveraging the subGaussian properties of X_i and e_i and Lemma 7, we address the impact of corrupted labels. All terms on the right-hand side of equation (19), except G_4, are effectively bounded. Establishing a bound on G_4, however, necessitates the use of a finite-sample variant of the mutual incoherence condition outlined in Assumption 2. The presence of corrupted samples introduces a dependency on the corruption proportion ε. Specifically, the following result can be derived from Lemma 15:

$$\left\| \widehat{\Sigma}^{J^{cc}}_{S^cS} \widehat{\Sigma}^{J^{cc}-1}_{SS} \right\|_\infty \leq 1 - \frac{\kappa}{2} + \mathcal{O}\left(\varepsilon \log(1/\varepsilon)\right) . \tag{20}$$

This imposes a more stringent requirement on κ in Assumption 2. Specifically, κ must scale as $\omega(\varepsilon \log(1/\varepsilon))$ to sufficiently mitigate the effects of corruption. Once, equation (20) is established, we can bound ζ_{S^c} as:

$$\|\zeta_{S^c}\|_\infty \leq 1 - \frac{\kappa}{4} + \mathcal{O}\left(\varepsilon \log(1/\varepsilon)\right) < 1 \, . \tag{21}$$

The proof is complete after combining results from equations (13) and (21). □

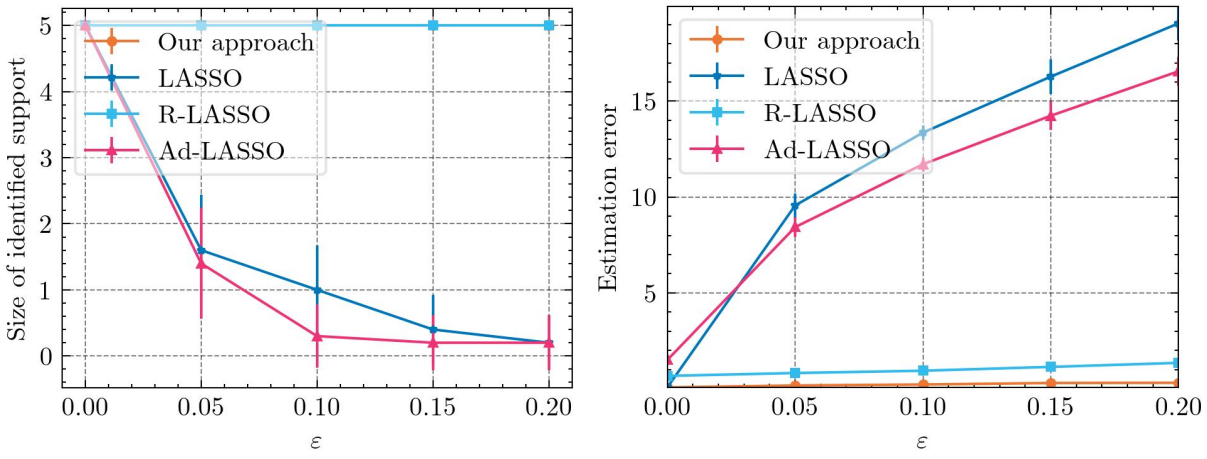
5 CONSTRUCTING AN ALGORITHM

Our theoretical analysis in Section 4 hinges on identifying an integral KKT point. However, most existing methods are only equipped to guarantee convergence to a potentially non-integral KKT point. Consider an algorithm, referred to as GET-A-KKT-POINT, which outputs such a KKT point. Leveraging the insights from Lemma 1, we propose a black-box framework that modifies the output of GET-A-KKT-POINT to produce an integral KKT point.

Algorithm 1 GET-AN-INTEGRAL-KKT-POINT

1: **Input:** $\lambda, \mathcal{D} = (X_i, y_i)_{i\in[n]}$
2: **Initialize:**
$(b, \theta) \leftarrow$ GET-A-KKT-POINT$((b_{\text{init}}, \theta_{\text{init}}), \lambda, \mathcal{D})$
3: **while** (b, θ) is not an integral KKT point **do**
4: Construct an integral primal feasible point $(\bar{b}, \theta)$ using Lemma 1
5: $(b, \theta) \leftarrow$ GET-A-KKT-POINT$\big((\bar{b}, \theta), \lambda, \mathcal{D}\big)$
6: **end while**

Since Lemma 1 constructs an integral primal feasible point quite efficiently, the computational complexity of GET-AN-INTEGRAL-KKT-POINT is influenced by two primary factors: the complexity of GET-A-KKT-POINT and the number of iterations executed within the while loop. If it is ensured that each iteration of the while loop decreases the objective function in (1) by at least a fixed amount, δ, then, using an argument similar to that in Awasthi et al. (2022), one can establish that the while loop terminates after a finite number of iterations. Consequently, the efficiency of GET-AN-INTEGRAL-KKT-POINT is intrinsically tied to the efficiency of GET-A-KKT-POINT.



(a) Support recovery, $p = 2000, n = 1000, k = 5$

(b) Estimation error, $p = 2000, n = 1000, k = 5$

Figure 1: (Left) Support recovery against varying amounts of corruption proportion in a high dimension low sample regime, (Right) Estimation error against varying amounts of corruption proportion in a high dimension low sample regime.

6 CORRUPTION IN LABELS AND FEATURES

In this section, we discuss more complicated setting of corruption in labels and features. We use a pre-processing step similar to the literature (Awasthi et al., 2022) to deal with corruption in X.

Basically, we first ensure all the features have zero mean by subtracting the mean. Note that this operation does not change the coefficients of the sparse linear model. The next step is to remove any samples with large absolute values. Formally speaking, we remove sample $i \in [n]$ if $|X_{ij}| > \tau$, for all $j \in [p]$. We choose the value of threshold τ very carefully based on the generalized assumptions on moments of X as shown in Lemma 3. This result is not restricted to linear models and can be used in other works also.

Lemma 3 (Informal). *Let* $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ *for any* $j \in [p]$. *Then with high probability of* $1-\delta$, *we claim* $A \leq \tau$, *where*

- $\tau = C\sigma\sqrt{\frac{\log(1/\delta)}{n}}$, *if* $X_{ij} \sim subGaussian(\sigma), \forall i \in [n]$, *implying all moments are bounded.*
- $\tau = \frac{var(X_j)}{\delta}$, *if only second order moment of* X_{ij} *is bounded for all* $i \in [n]$.
- $\tau = \left(\frac{\mu_{2q}}{\delta}\right)^{1/q}$, *if any even order moment, say* $2q$ *of* $|X_{ij}|$, *denoted by* μ_{2q} *is bounded. Here,* $q \geq 1$.

The above lemma bounds the quantity A, even when the corruption or the features X are sampled from some heavy-tailed distribution having only a few bounded moments. Now, we are equipped to present our main result for exact support recovery.

Theorem 2 (Informal). *Under assumptions 1, 2, and 3, if we choose* $\lambda = \Omega\left(\frac{(2-\kappa)\sigma\sigma_\epsilon\varepsilon\sqrt{k\tau\log(1/\varepsilon)}}{\kappa}\right)$, $n \geq \Omega(\frac{k^3\log p}{\varepsilon^2\log^2(1/\varepsilon)})$, *then with probability at least* $1-\mathcal{O}(\frac{1}{p})-\delta$, *we claim* $\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon)\right)$. *Here,* τ *is a function of* δ *chosen based on the moments of features* X *as shown in Lemma 3.*

Proof Overview. The proof is very similar to proof of Theorem 1 discussed in Section 4. The only difference is bounding the quantity G_{31} defined in Eq. (17). Instead of using the resilience condition, we use the pre-processing step and Lemma 3. For more details, please refer to Section A.2 in the appendix. □

Comparing Theorem 2 from Theorem 1, we can infer that the corruption in features influences the regularization parameter as it is a function of threshold τ. From Lemma 3, we can also note that the value of τ (and hence λ) is smaller when higher order moments are bounded. This makes sense intuitively as all the moments being bounded make the distribution well behaved like subGaussian.

In the next section, we proceed to experiments to validate the theoretical results discussed in this section.

7 EXPERIMENTAL VALIDATION

In this section, we empirically validate our theoretical findings. We compare our method with LASSO, R-

LASSO (Chen et al., 2013), and Ad-LASSO (Lambert-Lacroix and Zwald, 2011; Zou, 2006). To obtain a KKT point, we employ an alternating minimization approach, akin to the method proposed in Awasthi et al. (2022), with LASSO serving as a key subroutine.

Data generating process: We conduct experiments with parameters $p = 2000, k = 5$, and $n = 1000$. Notably, this regime is characterized by a limited number of samples, specifically $n < p$. The predictors X_i were sampled from $\mathcal{N}(0, \Sigma)$, where $\Sigma_{ii} = 1 \, \forall i \in [p]$ and $\Sigma_{ij} = 0.4 \, \forall i, j \in [p], i \neq j$. The non-zero entries of θ^* were drawn uniformly at random from $\{-1, 1\}$, and corrupted labels were selected uniformly at random from $[-2, 2]$.

Estimation: Following our theoretical framework, the regularization parameter λ was set to $\sqrt{\frac{k \log p}{n}} + \varepsilon \log(1/\varepsilon)$, where ε represents the corruption proportion. Further, we briefly discuss the hyperparameters chosen for baselines. We use Algorithm 4 ("Robust Lasso") from Chen et al. (2013) and assume the true hyperparameters n_1 and $R = \|\beta^*\|_2$ are known, which favors the baseline. All baselines Chen et al. (2013); Lambert-Lacroix and Zwald (2011) were run for 500 gradient-descent iterations or until convergence. The weights for Ad-LASSO were chosen as $1./\theta_{\text{ols}}$, where θ_{ols} denotes the ordinary least squares solution.

The experiments were carried out for varying values of ε, with each point in the results representing the average of 30 independent runs. For this setup, we vary the corruption proportion ε within the range $[0.05, 0.2]$. We compute the size of the identified support over 30 independent runs. Estimation error was assessed as $\|\widehat{\theta} - \theta^*\|_2$, quantifying the deviation of the estimated vector $\widehat{\theta}$ from the ground truth θ^*.

Observation: Figure 1 highlights the robustness of our approach in terms of both support recovery and estimation error. The degradation in performance becomes noticeable only at higher corruption proportions. For comparison, we also present the performance of standard LASSO as a baseline, demonstrating how corrupted samples can significantly compromise the recovery guarantees of LASSO.

Figure 1a demonstrates that both our approach and R-LASSO exhibit robustness in support recovery within the high-dimensional, low-sample regime. In contrast, other baselines, such as LASSO and Ad-LASSO, show a marked decline in performance. Additionally, Figure 1b underscores that while both R-LASSO and our method accurately recover the support, our approach achieves a lower estimation error.

We conduct additional experiments for various settings of (n, p, k) and outlier proportions. We make a similar observation that our method has better support recovery or lower estimation error from these experiments. The results are presented in Section C of the appendix.

8 CONCLUSION AND FUTURE DIRECTION

In this paper, we addressed the challenge of robust sparse linear regression by establishing theoretical guarantees for integral local solutions to a non-convex optimization problem. Using the primal-dual witness framework, we derived sufficient conditions for exact support recovery and accurate parameter estimation. We also introduced a simple, yet effective pre-processing step to handle corruptions from heavy-tailed distributions with only a few bounded moments. An exciting direction for future work is extending our analysis to more complex settings, such as generalized linear models or Gaussian graphical models.

References

Awasthi, P., Das, A., Kong, W., and Sen, R. (2022). Trimmed maximum likelihood estimation for robust generalized linear model. In *Advances in Neural Information Processing Systems*.

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. *Advances in Neural Information Processing Systems*, 30.

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. *Advances in neural information processing systems*, 28.

Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International conference on machine learning*, pages 774–782. PMLR.

Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. *Advances in neural information processing systems*, 32.

Daneshmand, H., Gomez-Rodriguez, M., Song, L., and Schoelkopf, B. (2014). Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-Thresholding Algorithm. In *International Conference on Machine Learning*, pages 793–801.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019a). Robust estimators

in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864.

Diakonikolas, I., Kong, W., and Stewart, A. (2019b). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM.

d'Orsi, T., Liu, C.-H., Nasser, R., Novikov, G., Steurer, D., and Tiegel, S. (2021). Consistent estimation for pca and sparse regression with oblivious outliers. *Advances in Neural Information Processing Systems*, 34:25427–25438.

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

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139 – 1170.

Hadi, A. S. and Chatterjee, S. (2009). *Sensitivity analysis in linear regression*. John Wiley & Sons.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., and Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. Wiley-Interscience; New York.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hsu, D., Kakade, S., Zhang, T., et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17.

Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer.

Jambulapati, A., Li, J., and Tian, K. (2020). Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33:15689–15701.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2020). High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR.

Lozano, A. C., Meinshausen, N., and Yang, E. (2016). Minimum distance lasso for robust high-dimensional regression. *Electronic Journal of Statistics*, 10(1):1296–1340.

Maurya, D. and Honorio, J. (2024). A theoretical study of the effects of adversarial attacks on sparse regression. *Transactions on Machine Learning Research*.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462.

Nguyen, N. H. and Tran, T. D. (2012). Robust lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59(4):2036–2058.

Norman, T., Weinberger, N., and Levy, K. Y. (2023). Robust linear regression for general feature distribution. In *International Conference on Artificial Intelligence and Statistics*, pages 2405–2435. PMLR.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007). Spam: Sparse Additive Models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 1201–1208. Curran Associates Inc.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional Ising Model Selection Using L1-Regularized Logistic Regression. *The Annals of Statistics*, 38(3):1287–1319.

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. (2019). Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265.

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

Wainwright, M. J. (2009). Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No, but we provide all the experimental details to reproduce the main results]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable, no special computing infrastructure was used.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material: Robust Estimation of a Sparse Linear Model: Provable Guarantees with Non-convexity

A Formal Statements and Proofs of Theorems and Lemmas

A.1 Proof of Lemma 4

Lemma 4. *Under assumptions 1 and 2 and if we choose* $\lambda = \Omega\left(\frac{2-\kappa}{\kappa}\left(\sigma\sigma_e\sqrt{\frac{k\log p}{n}} + \frac{\sigma\sigma_e\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)$ *and* $n \geq \Omega(\frac{k^3\log p}{\varepsilon^2\log^2(1/\varepsilon)\tau_1(\alpha_1,\kappa,\sigma,\Sigma)})$*, then with probability at least* $1-\mathcal{O}(\frac{1}{p})$*,* $\|\underline{\theta}-\underline{\theta}^*\|_\infty \leq \delta_{n,\varepsilon}$*, where* $\delta_{n,\varepsilon} = \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon)\right)$*. Here* $\tau_1(\alpha_1,\kappa,\sigma,\Sigma)$ *is a constant which is independent of* p, k, m *and* n*.*

Proof. We define the sets:

$$J^{cc} = \Big\{(X_i,y_i) \mid b_i^* = 1, \hat{b}_i = 1, i \in [n]\Big\}, \qquad J^{co} = \Big\{(X_i,y_i) \mid b_i^* = 1, \hat{b}_i = 0, i \in [n]\Big\},$$
$$J^{oc} = \Big\{(X_i,y_i) \mid b_i^* = 0, \hat{b}_i = 1, i \in [n]\Big\}, \qquad J^{oo} = \Big\{(X_i,y_i) \mid b_i^* = 0, \hat{b}_i = 0, i \in [n]\Big\},$$
$$\hat{J}^c = \Big\{(X_i,y_i) \mid \hat{b}_i = 1, i \in [n]\Big\} = J^{cc} \cup J^{oc}.$$

Let J^{*c} denote the set of n clean samples before any corruption. So, $J^{cc} = J^{*c} \setminus J^{co}$.

The optimization problem can be rewritten in the following way:

$$\underline{\theta} = \arg\min_{\widetilde{\theta}\in\mathbb{R}^k} \frac{1}{2m}\sum_{(X_i,y_i)\in\hat{J}^c} (y_i - \langle\widetilde{X}_i\,,\widetilde{\theta}\rangle)^2 + \lambda\|\widetilde{\theta}\|_1\,.$$

Then $\underline{\theta}$ must satisfy the stationarity KKT condition, i.e.,

$$-\frac{1}{m}\sum_{(X_i,y_i)\in\hat{J}^c} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle) + \lambda\zeta = 0\,,$$
$$-\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle) + \lambda\zeta = \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\,,$$

where ζ is an element of the subdifferential set of $\|\widetilde{\theta}\|_1$ at $\underline{\theta}$. Since $(X_i,y_i) \in J^{cc}$, we can substitute $y_i = \langle\widetilde{X}_i\,,\underline{\theta}^*\rangle + e_i$.

$$\widehat{\Sigma}_{SS}^{J^{cc}}(\underline{\theta} - \underline{\theta}^*) = -\lambda\zeta + \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i + \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle),$$

where $\widehat{\Sigma}_{SS}^{J^{cc}} = \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i\widetilde{X}_i^T$.

With little manipulation and by use of norm-inequalities, we can rewrite this as

$$\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_\infty \Bigg(\|\lambda\zeta\|_\infty + \|\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty + \|\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\|_\infty\Bigg)$$

where $m = n(1-\varepsilon)$ and ε denotes the outlier proportion. Note that the set J^{cc} will contain at least $n(1-\varepsilon)$ clean samples.

We observe that by using $n = \Omega(\frac{k^2 \log(p)}{\alpha_1^4 f(\varepsilon)})$, where $f(\varepsilon) = \varepsilon \log(1/\varepsilon)$ and Weyl's inequality in Lemma 5, we have $\| \left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1} \|_\infty \leq \frac{3}{2}\|\Sigma_{SS}^{-1}\|_\infty$. Thus,

$$\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \frac{3}{2}\|\Sigma_{SS}^{-1}\|_\infty \left(\lambda + \|\frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty + \|\frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle)\|_\infty \right) \tag{22}$$

We bound $\|\frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty$ using Lemma 6. We can bound the last term in the RHS using Lemma 7. We arrive at:

$$\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \lambda + \left(\lambda + 128 \frac{\sigma\sigma_e \varepsilon \log(1/\varepsilon)}{1-\varepsilon} \right) + C\sqrt{k}\sigma\sigma_e \varepsilon \log(1/\varepsilon) \|\underline{\theta} - \underline{\theta}^*\|_\infty.$$

This can be further simplified to following, if $C\sqrt{k}\sigma\sigma_e\varepsilon\log(1/\varepsilon) < 1/2$

$$\begin{aligned}
\|\underline{\theta} - \underline{\theta}^*\|_\infty &\leq \mathcal{O}\left(\frac{\sigma\sigma_e}{1 - \sqrt{k}\sigma\sigma_e\varepsilon\log(1/\varepsilon)} \sqrt{\frac{k\log(p)}{n}} + \frac{\sigma\sigma_e}{1 - \sqrt{k}\sigma\sigma_e\varepsilon\log(1/\varepsilon)} \varepsilon \log(1/\varepsilon) \right) \\
&\leq \mathcal{O}\left(\sigma\sigma_e \sqrt{\frac{k\log(p)}{n}} + \sigma\varepsilon\log(1/\varepsilon) \right) \\
&\leq \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon) \right).
\end{aligned}$$

□

A.2 Proof of Theorem 2

Theorem 2: Under assumptions 1 and 2 and if we choose $\lambda = \Omega\left(\frac{(2-\kappa)\sigma\sigma_e\varepsilon\sqrt{k\tau\log(1/\varepsilon)}}{\kappa} \right)$, $n \geq \Omega(\frac{k^3 \log p}{\varepsilon^2 \log^2(1/\varepsilon)})$, then with probability at least $1 - \mathcal{O}(\frac{1}{p}) - \delta$, we claim $\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon)\right)$. Here, τ is a function of δ chosen based on the moments of features X as shown in Lemma 3:

- $\tau = C\sigma\sqrt{\frac{\log(1/\delta)}{n}}$, if $X_{ij} \sim \text{subGaussian}(\sigma), \forall i \in [n]$, implying all moments are bounded.
- $\tau = \frac{\text{var}(X_j)}{\delta}$, if only second order moment of X_{ij} is bounded for all $i \in [n]$.
- $\tau = \left(\frac{\mu_4}{\delta}\right)^{1/2}$, if only the fourth order moment of $|X_{ij}|$, denoted by μ_4 is bounded.
- $\tau = \left(\frac{\mu_6}{\delta}\right)^{1/3}$, if only the sixth order moment of $|X_{ij}|$, denoted by μ_6 is bounded.
- $\tau = \left(\frac{\mu_{2q}}{\delta}\right)^{1/q}$, if any even order moment, say $2q$ of $|X_{ij}|$, denoted by μ_{2q} is bounded.

Proof. We start from Eq. (22) of Lemma 4:

$$\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \frac{3}{2}\|\Sigma_{SS}^{-1}\|_\infty \left(\lambda + \|\frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty + \|\frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i (y_i - \langle \widetilde{X}_i \, , \underline{\theta} \rangle)\|_\infty \right)$$

We bound $\|\frac{1}{m} \sum_{(X_i,y_i)\in J^{cc}} \widetilde{X}_i e_i\|_\infty$ using Lemma 5 as done in Lemma 4. We can bound the last term in the RHS using Lemma 8.

$$\|\underline{\theta} - \underline{\theta}^*\|_\infty \leq \lambda + \left(\lambda + 128 \frac{\sigma\sigma_e \varepsilon \log(1/\varepsilon)}{1-\varepsilon} \right) + C\sigma_e \varepsilon \sqrt{k\tau\log(1/\varepsilon)} \, \|\underline{\theta} - \underline{\theta}^*\|_\infty .$$

This can be further simplified to following, if $C\sigma_e\varepsilon\sqrt{k\tau\log(1/\varepsilon)} < 1/2$

$$\begin{aligned}\|\underline{\theta} - \underline{\theta}^*\|_\infty &\leq \mathcal{O}\left(\frac{\sigma\sigma_e}{1-\sigma_e\varepsilon\sqrt{k\tau\log(1/\varepsilon)}}\sqrt{\frac{\log(p)}{n}} + \frac{\sigma\sigma_e}{1-\sigma_e\varepsilon\sqrt{k\tau\log(1/\varepsilon)}}\varepsilon\log(1/\varepsilon)\right)\\ &\leq \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\varepsilon\log(1/\varepsilon)\right)\\ &\leq \mathcal{O}\left(\sigma\varepsilon\log(1/\varepsilon)\right).\end{aligned}$$

The strict dual feasibility can be assured by appropriately tweaking the proof of Lemma 14. We use Lemma 8 in the proof of Lemma 14 instead of Lemma 5. One can still ensure strict dual feasibility by choosing $\lambda = \Omega\left(\frac{(2-\kappa)\sigma\sigma_e\varepsilon\sqrt{k\tau\log(1/\varepsilon)}}{\kappa}\right)$.

Note that the value of τ can be selected as suggested in Lemma 3.

□

A.3 Proof of Lemma 5

Lemma 5. *If $n = \Omega(\frac{k^2\log(p)}{\alpha_1^4\varepsilon\log(1/\varepsilon)})$, then with high probability of $1-\mathcal{O}(1/p)$, we claim $\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_\infty \leq \frac{3}{2}\|\Sigma_{SS}^{-1}\|_\infty$.*

Proof. We start by applying the triangle inequality:

$$\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_\infty \leq \|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1} - \Sigma_{SS}^{-1}\|_\infty + \|\Sigma_{SS}^{-1}\|_\infty.$$

Further, we use simple norm inequalities to bound

$$\begin{aligned}\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1} - \Sigma_{SS}^{-1}\|_\infty &\leq \sqrt{k}\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1} - \Sigma_{SS}^{-1}\|_2\\ &\leq \sqrt{k}\left\|\Sigma_{SS}^{-1}\right\|_2\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right) - \Sigma_{SS}\|_2\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_2.\end{aligned}$$

Further using Lemma 16, we can claim $\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_2 \leq \left(1-\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)^{-1}\alpha_1^{-1}$ with high probability. We know that $\varepsilon\log(1/\varepsilon) \leq 0.16$ for $0 < \varepsilon \leq 0.5$. For sufficiently small ε, we can assume $\left(1-\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)^{-1}$ is bounded. Hence, $\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_2 \leq \frac{1}{c_1\alpha_1}$ for some constant $0 < c_1 < 1$.

Using arguments similar to Lemma 16, we can also bound $\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right) - \Sigma_{SS}\|_2 \leq \frac{c_1\alpha_1^2}{2\sqrt{k}}\|\Sigma_{SS}^{-1}\|_\infty$ with high probability of $1-\mathcal{O}(1/p)$ if $n = \Omega(\frac{k^2\log(p)}{\alpha_1^4 f(\varepsilon)})$ for sufficiently small ε, where $f(\varepsilon) = \varepsilon\log(1/\varepsilon)$. So, we can finally claim

$$\begin{aligned}\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_\infty &\leq \sqrt{k}\left\|\Sigma_{SS}^{-1}\right\|_2\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right) - \Sigma_{SS}\|_2\|\left(\widehat{\Sigma}_{SS}^{J^{cc}}\right)^{-1}\|_2 + \|\Sigma_{SS}^{-1}\|_\infty\\ &\leq \sqrt{k}\frac{1}{c_1\alpha_1^2}\frac{c_1\alpha_1^2}{2\sqrt{k}}\|\Sigma_{SS}^{-1}\|_\infty + \|\Sigma_{SS}^{-1}\|_\infty = \frac{3}{2}\|\Sigma_{SS}^{-1}\|_\infty.\end{aligned}$$

□

A.4 Proof of Lemma 6

Lemma 6. *If $\lambda \geq 8\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}$ and $n = \Omega(k^2\log(p)/(\varepsilon\log(1/\varepsilon)))$, then $\|\frac{1}{n(1-\varepsilon)}\sum_{(X_i,y_i)\in J^{cc}}\widetilde{X}_i e_i\|_\infty \leq \left(\frac{\lambda}{(1-\varepsilon)} + 128\frac{\sigma\sigma_e\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)$ with probability at least $1-\mathcal{O}(\frac{1}{p})$.*

Proof. We take the i-th entry of $\frac{1}{n(1-\varepsilon)}\sum_{(X_j,y_j)\in J^{cc}}\widetilde{X}_j e_j$ for some $i\in S$, i.e., $|\frac{1}{n(1-\varepsilon)}\sum_{(X_j,y_j)\in J^{cc}}X_{ji}e_j|$. We use the standard approach used in Lemma 4.3 of Diakonikolas et al. (2019a) or Lemma 8 of Jambulapati et al. (2020):

$$\frac{1}{n(1-\varepsilon)}\sum_{(X_j,y_j)\in J^{cc}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}=\frac{n}{n(1-\varepsilon)}\frac{1}{n}\sum_{(X_j,y_j)\in J^{*c}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}-\frac{n\varepsilon}{n(1-\varepsilon)}\frac{1}{n\varepsilon}\sum_{(X_j,y_j)\in J^{co}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}.$$

Recall that X_{ji} is a sub-Gaussian random variable with parameter σ and e_j is a sub-Gaussian random variable with parameter σ_e. Then, $\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}$ is a sub-exponential random variable with parameters $(4\sqrt{2},2)$. Using the concentration bounds for the sum of independent sub-exponential random variables (Wainwright, 2019), we can write:

$$\mathbf{P}(|\frac{1}{n}\sum_{(X_j,y_j)\in J^{*c}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_1)\leq 2\exp(-\frac{nt_1^2}{64}),\quad 0\leq t_1\leq 8.$$

Taking a union bound across $i\in S$:

$$\mathbf{P}(\exists i\in S\mid|\frac{1}{n}\sum_{(X_j,y_j)\in J^{*c}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_1)\leq 2k\exp(-\frac{nt_1^2}{64}),\quad 0\leq t_1\leq 8.$$

We select $t_1=\frac{\lambda}{\sigma\sigma_e}$, where $\lambda>8\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}$ to claim $\|\frac{1}{n}\sum_{(X_i,y_i)\in J^{*c}}\widetilde{X}_i e_i\|_\infty\leq\lambda$ with probability $1-\mathcal{O}(\frac{1}{p})$.

Similarly, focusing on the second term, we can claim using sub-exponential tail bounds:

$$\mathbf{P}(\exists i\in S\mid|\frac{1}{n\varepsilon}\sum_{(X_j,y_j)\in J^{co}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_2)\leq 2k\exp(-\frac{n\varepsilon t_2}{64}).$$

Let $\mathcal{M}$ represent the set of all samples. Let $\mathcal{J}^{cc}=\{J^{cc}\mid J^{cc}\subseteq\mathcal{M},|J^{cc}|=n(1-\varepsilon)\}$. Further, taking a union bound over all the sets of size $n(1-\varepsilon)$

$$\mathbf{P}(\exists J^{cc}\subset\mathcal{J}^{cc},\exists i\in S\mid|\frac{1}{n\varepsilon}\sum_{(X_j,y_j)\in J^{co}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_2)\leq 2k\binom{n}{n\varepsilon}\exp(-\frac{n\varepsilon t_2}{64}).$$

Further, we use $\log\binom{n}{n(1-\varepsilon)}=\mathcal{O}(n\varepsilon\log(1/\varepsilon))$

$$\mathbf{P}(\exists J^{cc}\subset\mathcal{J}^{cc},\exists i\in S\mid|\frac{1}{n\varepsilon}\sum_{(X_j,y_j)\in J^{co}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_2)\leq 2k\exp(-\frac{n\varepsilon t_2}{64}+n\varepsilon\log(1/\varepsilon)).\tag{23}$$

Further we take $t_2=128\log(1/\varepsilon)>8$ and $n=\Omega(k^2\log(p)/(\varepsilon\log(1/\varepsilon)))$.

$$\mathbf{P}(\exists J^{cc}\subset\mathcal{J}^{cc},\exists i\in S\mid|\frac{1}{n\varepsilon}\sum_{(X_j,y_j)\in J^{co}}\frac{\widetilde{X}_{ji}}{\sigma}\frac{e_j}{\sigma_e}|\geq t_2)\leq 2k\exp(-n\varepsilon\log(1/\varepsilon))\leq\mathcal{O}(\frac{1}{p}).\tag{24}$$

Further, we use the triangle inequality to arrive at:

$$\|\frac{1}{n(1-\varepsilon)}\sum_{(X_i,y_i)\in J^{cc}}\widetilde{X}_i e_i\|_\infty\leq\left(\frac{\lambda}{1-\varepsilon}+128\frac{\sigma\sigma_e\varepsilon}{1-\varepsilon}\log(1/\varepsilon)\right).$$

which gives the desired result.

□

Next we bound $\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i-\langle\widetilde{X}_i,\underline{\theta}\rangle)\|_\infty$.

A.5 Proof of Lemma 7

Lemma 7. *If* $n = \Omega\left(\frac{k\log(p)}{\varepsilon^2\log^2(1/\varepsilon)}\right)$, *then with probability at least* $1-\mathcal{O}\left(\frac{1}{p}\right)$, *we claim* $\|\frac{1}{n(1-\varepsilon)}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\|_\infty \le C\sqrt{k}\sigma\sigma_e\varepsilon\log(1/\varepsilon)\,\|\underline{\theta}-\underline{\theta}^*\|_\infty$.

Proof. We start with algebraic manipulation

$$\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\|_\infty \le \frac{n\varepsilon}{n(1-\varepsilon)}\times\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\|_\infty.$$

We focus on the j^{th} entry of $\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)$, which is denoted by $\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_{ij}(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)$. Further we apply Cauchy-Schwarz inequality for this j^{th} entry:

$$\begin{aligned}\left|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_{ij}(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\right| &\le \left(\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_{ij}^2\right)^{1/2}\left(\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)^2\right)^{1/2}\\ &\le \left(\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_{ij}^2\right)^{1/2}\left(\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)^2\right)^{1/2},\end{aligned}$$

where we have used the optimality of the solution in the last step. Note that J^{co} denotes the set of clean samples. Hence we substitute $(y_i - \langle\widetilde{X}_i\,,\underline{\theta}^*\rangle)^2 = e_i^2$ to simplify $(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)^2$:

$$(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)^2 = (\langle\widetilde{X}_i\,,\underline{\theta}-\underline{\theta}^*\rangle)^2 + e_i^2 + 2\langle e_i\widetilde{X}_i\,,\underline{\theta}-\underline{\theta}^*\rangle. \tag{25}$$

Further, we use norm inequalities to claim:

$$\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}(\langle\widetilde{X}_i\,,\underline{\theta}-\underline{\theta}^*\rangle)^2 \le \left\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}\widetilde{X}_i\widetilde{X}_i^T\right\|_2\|\underline{\theta}-\underline{\theta}^*\|_2^2 \le k\left\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}\widetilde{X}_i\widetilde{X}_i^T\right\|_2\|\underline{\theta}-\underline{\theta}^*\|_\infty^2.$$

The first term can be easily bounded to $\mathcal{O}(\log(1/\varepsilon))$ with probability $1-O(1/p)$ using Lemma 8 of Jambulapati et al. (2020) if $n \ge \frac{k\log(p)}{(\varepsilon\log(1/\varepsilon))^2}$.

Similarly, for $\langle e_i\widetilde{X}_i\,,\underline{\theta}-\underline{\theta}^*\rangle$

$$\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}\langle e_i\widetilde{X}_i\,,\underline{\theta}-\underline{\theta}^*\rangle \le \left\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}e_i\widetilde{X}_i\right\|_2\|\underline{\theta}-\underline{\theta}^*\|_2 \le k\left\|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}e_i\widetilde{X}_i\right\|_\infty\|\underline{\theta}-\underline{\theta}^*\|_\infty.$$

The first term of the above equation can be bound using Eq. (23) as $\mathcal{O}(\log(1/\varepsilon))$.

As e_i is assumed to be sub-Gaussian, e_i^2 is sub-exponential, and we use sub-exponential concentration inequalities for $t > \sigma_e^2$:

$$\mathbf{P}\left[\left|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}e_i^2 - \mathbb{E}[e_i^2]\right| \ge t_1\right] \le 2\exp\left(-cn\varepsilon\min\left(\frac{t_1^2}{\sigma_e^4},\frac{t_1}{\sigma_e^2}\right)\right) = 2\exp\left(-cn\varepsilon\frac{t_1}{\sigma_e^2}\right).$$

Let $\mathcal{M}$ represent the set of all samples. Let $\mathcal{J}^{cc} = \{J^{cc} \mid J^{cc}\subseteq\mathcal{M}, |J^{cc}| = n(1-\varepsilon)\}$. Further, taking a union bound over all the sets of size $n(1-\varepsilon)$ and using $\log\binom{n}{n(1-\varepsilon)} = \mathcal{O}(n\varepsilon\log(1/\varepsilon))$

$$\mathbf{P}\left[\exists J^{cc}\subset\mathcal{J}^{cc}, \left|\frac{1}{n\varepsilon}\sum_{(X_i,y_i)\in J^{co}}e_i^2 - \mathbb{E}[e_i^2]\right| \ge t_1\right] \le 2\exp\left(n\varepsilon\left(\frac{-ct_1}{\sigma_e^2}+\log(1/\varepsilon)\right)\right).$$

Further, we take $t_1 = \frac{2\sigma_e^2 \log(1/\varepsilon)}{c}$ and $n \geq \frac{\log(p)}{\varepsilon \log(1/\varepsilon)}$.

Substituting the above bounds in Eq. (25), we obtain the following bound with high probability:

$$\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{co}} (y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle)^2 = k\sigma_e^2 \mathcal{O}(\log(1/\varepsilon)) \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty^2 .$$

Similarly, for $\widetilde{X}_{ij}^2$, we claim the following using sub-exponential tail bounds:

$$\mathbf{P}\left[\exists J^{cc} \subset \mathcal{J}^{cc}, \left| \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_{ij}^2 - \mathbb{E}[\widetilde{X}_{ij}^2] \right| \geq t_2 \right] \leq 2\exp\left(n\varepsilon \left(\frac{-ct_2}{\sigma^2} + \log(1/\varepsilon) \right) \right).$$

So, we can claim the following with probability $1 - \mathcal{O}\left(\frac{1}{p}\right) - 2\exp\left(n\varepsilon\left(\frac{-ct_2}{\sigma^2} + \log(1/\varepsilon)\right)\right)$:

$$\left| \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_{ij}(y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle) \right| \leq \sqrt{k}\sqrt{t_2}\sigma_e \mathcal{O}(\sqrt{\log(1/\varepsilon)}) \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty .$$

Taking the union bound across k entries, we can claim the following

$$\| \frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle) \|_\infty \leq \frac{\varepsilon}{1-\varepsilon} \sqrt{k}\sqrt{t_2}\sigma_e \mathcal{O}(\sqrt{\log(1/\varepsilon)}) \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty ,$$

with probability $1 - \mathcal{O}\left(\frac{1}{p}\right) - 2k\exp\left(n\varepsilon\left(\frac{-ct_2}{\sigma^2} + \log(1/\varepsilon)\right)\right)$. Further, we select $t_2 = \frac{2\sigma^2 \log(1/\varepsilon)}{c}$ and take $n \geq \frac{2\log(p)}{\varepsilon \log(1/\varepsilon)}$ to bound the above probability as $1 - \mathcal{O}\left(\frac{1}{p}\right)$. We finally arrive at:

$$\begin{aligned}
\| \frac{1}{m} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle) \|_\infty &\leq \frac{\varepsilon}{1-\varepsilon} \sqrt{k \left(\frac{2\sigma_e^2 \log(1/\varepsilon)}{c} \right) \left(\frac{2\sigma^2 \log(1/\varepsilon)}{c} \right)} \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty \\
&\leq \frac{\varepsilon}{1-\varepsilon} C\sqrt{k}\sigma\sigma_e \log(1/\varepsilon) \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty \\
&\leq C\sqrt{k}\sigma\sigma_e \varepsilon \log(1/\varepsilon) \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty .
\end{aligned}$$

□

A.6 Proof of Lemma 8

Lemma 8. *If* $n = \Omega\left(\frac{k\log(p)}{\varepsilon^2 \log^2(1/\varepsilon)}\right)$ *and* $\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2 \leq \tau$, *then we claim* $\| \frac{1}{n(1-\varepsilon)} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_i(y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle) \|_\infty \leq C\sigma_e \varepsilon \sqrt{k\tau \log(1/\varepsilon)} \left\| \underline{\theta} - \underline{\theta}^* \right\|_\infty$.

Proof. The Cauchy-Schwarz inequality inequality in the proof of Lemma 7 can be generalized using Hölder's inequality:

$$\begin{aligned}
\left| \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} \widetilde{X}_{ij}(y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle) \right| &\leq \left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^r \right)^{1/r} \left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle|^q \right)^{1/q} \\
&\leq \left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^r \right)^{1/r} \left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |y_i - \langle \widetilde{X}_i \,, \underline{\theta} \rangle|^2 \right)^{1/2} ,
\end{aligned}$$

where $\frac{1}{r} + \frac{1}{q} = 1$. To use Hölder's inequality, we need $r \geq 1$ and $q \geq 1$. In the last step, we assume $q \in [1, 2]$ and use the fact that power means function is monotonically increasing in order. Hence, depending on the value of q, the value of r can lie in the interval $[2, \infty)$.

The second term $\sum_{(X_i,y_i)\in J^{oc}} |y_i - \langle \widetilde{X}_i\,, \underline{\theta}\rangle|^2$ can be bounded using the optimality of the solution as done in the proof of Lemma 7. The first term of $\left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^r\right)^{1/r}$ can be bounded depending on the assumptions on the features $\widetilde{X}_{ij}$.

Before doing that, we utilize a simple transformation using vector norm inequalities for any $r \geq 2$:

$$\left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^r\right)^{1/r} \leq \left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2\right)^{1/r} \left(\max_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|\right)^{1-2/r}$$

We obtain a high probability upper bound for the first term under various generalized settings. The bound on the second term is obtained due to the preprocessing step. In the preprocessing step, we remove any rows if $\max_{i\in[n]} X_{ij}^2 > \tau$ for all $j \in [p]$. The value of the threshold, denoted by τ is critical and is actually the bound on the first term, which is derived later. Hence, the overall bound turns out to be:

$$\left(\frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^r\right)^{1/r} \leq (\tau)^{1/r} \left(\sqrt{\tau}\right)^{1-2/r} = \sqrt{\tau}$$

Using the above equation, we arrive at the desired result. In the next lemma, we show the derivation for the threshold τ. □

Lemma 3 Let $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ for any $j \in [p]$. Then with high probability of $1-\delta$, we claim $A \leq \tau$, where

- $\tau = C\sigma\sqrt{\frac{\log(1/\delta)}{n}}$, if $X_{ij} \sim \text{subGaussian}(\sigma), \forall i \in [n]$, implying all moments are bounded.
- $\tau = \frac{\text{var}(X_j)}{\delta}$, if only second order moment of X_{ij} is bounded for all $i \in [n]$.
- $\tau = \left(\frac{\mu_4}{\delta}\right)^{1/2}$, if only the fourth order moment of $|X_{ij}|$, denoted by μ_4 is bounded.
- $\tau = \left(\frac{\mu_6}{\delta}\right)^{1/3}$, if only the sixth order moment of $|X_{ij}|$, denoted by μ_6 is bounded.
- $\tau = \left(\frac{\mu_{2q}}{\delta}\right)^{1/q}$, if any even order moment, say $2q$ of $|X_{ij}|$, denoted by μ_{2q} is bounded.

Proof. We show the proof for a very generalized setting. Let $\phi : [0,\infty) \to [0,\infty)$ be a monotonically increasing and non-negative function. Let $A = \frac{\sum_i X_{ij}^2}{n} \geq 0$ for any $j \in [p]$. Without loss of generality, we assume all the columns of X are zero-mean. Note that all the columns can be made zero-mean by subtracting the mean. This operation will not change the coefficients of the sparse linear model.

By using the Markov inequality, we claim:

$$\Pr[A > \tau] = \Pr[\phi(A) > \phi(\tau)] \leq \frac{\mathbb{E}[\phi(A)]}{\phi(\tau)} = \delta, \tag{26}$$

where $\tau > 0$ and $\delta > 0$. Further, we can claim $\tau = \phi^{-1}\left(\frac{\mathbb{E}[\phi(A)]}{\delta}\right)$.

Further, for all five claims in the statement, we use a different function ϕ to derive τ. The proof for each lemma can be seen from the other lemmas summarized below:

1. For the subGaussian case, please refer to Lemma 9.
2. For the second order moment bounded case, please refer to Lemma 10.

3. For the fourth order moment bounded case, please refer to Lemma 11.
4. For the sixth order moment bounded case, please refer to Lemma 12.
5. For any even order moment bounded case, please refer to Lemma 13.

□

Lemma 9. *Let* $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ *and* $X_{ij} \sim subGaussian(\sigma), \forall i \in [n]$*, then with probability* $1-\delta$*, we claim* $A \leq \tau$*, where* $\tau = C\sigma^2 \frac{\log(1/\delta)}{n}$*.*

Proof. We use $\phi(z) = e^{tz}$ for some $t > 0$ in Eq. (26). Also note that if X is sub-Gaussian, then X^2 is sub-exponential. Hence, we use sub-exponential concentration bounds:

$$\Pr\Big(A \ > \ \tau\Big) \leq \exp\Big(- c\, n \min\Big\{\frac{\tau^2}{\sigma^4}, \frac{\tau}{\sigma^2}\Big\}\Big),$$

For the heavier tail side, we will have

$$\tau \geq C\sigma^2 \frac{\log(1/\delta)}{n}.$$

□

Lemma 10. *Let* $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ *and let* X_{ij} *be sampled from some heavy-tailed distribution with only second order moment bounded, then with probability* $1-\delta$*, we claim* $A \leq \tau$*, where* $\tau = \frac{var(X_j)}{\delta}$*.*

Proof. In this case, we consider $\phi(z) = z$ in Eq. (26), hence we have

$$\tau = \frac{\mathbb{E}[\frac{1}{n}\sum_i X_{ij}^2]}{\delta} = \frac{\mathrm{var}(X_j)}{\delta}.$$

□

Lemma 11. *Let* $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ *and let* X_{ij} *be sampled from some heavy-tailed distribution with only fourth moment bounded and denoted by* μ_4*, then with probability* $1-\delta$*, we claim* $A \leq \tau$*, where* $\tau = \left(\frac{\mu_4}{\delta}\right)^{1/2}$*.*

Proof. Assume 4^{th} moment is bounded. Consider $\phi(z) = z^2$ in Eq. (26) and therefore $\phi^{-1}(z) = \sqrt{z}$. So we have

$$\tau = \sqrt{\frac{\mathbb{E}\left[\left(\frac{\sum_i X_{ij}^2}{n}\right)^2\right]}{\delta}}.$$

For the ease of notation, we drop the subscript j. For simplicity, we assume the samples are independent. Further we assume the fourth-order moments for each variable are bounded as shown below:

$$\mathbb{E}\left[X_i^4\right] \leq \mu_4, \qquad \forall i \in [n].$$

Further, we can simplify

$$\mathbb{E}[A^2] = \mathbb{E}\left[\left(\frac{\sum_i X_i^2}{n}\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_i X_i^4 + \sum_{i\neq j} X_i^2 X_j^2\right] = \frac{1}{n^2}\left(n\mathbb{E}[X_i^4] + n(n-1)\mathbb{E}[X_i^2 X_j^2]\right).$$

Further, we use Hölder's inequality to arrive at:

$$\mathbb{E}[A] \leq \frac{1}{n^2}\left(n\mathbb{E}[X_i^4] + n(n-1)\sqrt{\mathbb{E}[X_i^4]\mathbb{E}[X_j^4]}\right) \leq \mathbb{E}[X_i^4] \leq \mu_4.$$

Therefore the final expression of τ simplifies to:

$$\tau = \left(\frac{\mu_4}{\delta}\right)^{1/2}.$$

□

Lemma 12. *Let $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ and let X_{ij} be sampled from some heavy-tailed distribution with only sixth moment bounded and denoted by μ_6, then with probability $1-\delta$, we claim $A \le \tau$, where $\tau = \left(\frac{\mu_6}{\delta}\right)^{1/3}$.*

Proof. Consider $\phi(z) = z^3$ in Eq. (26) and hence $\phi^{-1}(z) = z^{\frac{1}{3}}$. Hence, we have

$$\tau = \left(\frac{\mathbb{E}\left[\left(\frac{\sum_i X_{ij}^2}{n}\right)^3\right]}{\delta}\right)^{1/3} = \frac{1}{\delta^{1/3}}\left(\mathbb{E}\left[\left(\frac{\sum_i X_{ij}^2}{n}\right)^3\right]\right)^{1/3}.$$

For the ease of notation, we drop the subscript j. For simplicity, we assume the samples are independent. Further, we assume the variance, fourth-order, and sixth-order moments for each variable are bounded as shown below:

$$\mathbb{E}\left[X_i^6\right] \le \mu_6, \qquad \forall i \in [n].$$

We can compute $\mathbb{E}[A]$ and $\mathbb{E}[A^2]$ as done previously. So we proceed to $\mathbb{E}[A^3]$

$$\begin{aligned}\mathbb{E}[A^3] = \mathbb{E}\left[\left(\frac{\sum_i X_i^2}{n}\right)^3\right] &= \frac{1}{n^3}\left(\mathbb{E}\left[\sum_{m_1+m_2+\cdots+m_n=3}\binom{3}{m_1 m_2\cdots m_n}\prod_{i=1}^{n}\left(X_i^2\right)^{m_i}\right]\right)\\ &= \frac{1}{n^3}\left(\sum_{i=1}^{n}\mathbb{E}[X_i^6] + 3\sum_{i\neq j}\mathbb{E}[X_i^4 X_j^2] + 6\sum_{i<j<k}\mathbb{E}[X_i^2 X_j^2 X_k^2]\right).\end{aligned}$$

Further using Hölder's inequality, we arrive at:

$$\begin{aligned}\mathbb{E}[A^3] &\le \frac{1}{n^3}\left(n\mu_6 + 3\sum_{i\neq j}\left(\mathbb{E}[X_i^{4\times\frac{6}{4}}]\right)^{\frac{4}{6}}\left(\mathbb{E}[X_j^{2\times\frac{6}{2}}]\right)^{\frac{2}{6}} + 6\sum_{i<j<k}\left(\mathbb{E}[X_i^{2\times\frac{6}{2}}]\right)^{\frac{2}{6}}\left(\mathbb{E}[X_j^{2\times\frac{6}{2}}]\right)^{\frac{2}{6}}\left(\mathbb{E}[X_k^{2\times\frac{6}{2}}]\right)^{\frac{2}{6}}\right)\\ &\le \frac{1}{n^3}\left(n\mu_6 + 3n(n-1)\mu_6 + 6n(n-1)(n-2)\mu_6\right) \le \mu_6.\end{aligned}$$

Here, we have assumed that the samples are independent in the above proof. Hence, the final expression for τ simplifies to

$$\tau = \left(\frac{\mu_6}{\delta}\right)^{1/3}.$$

□

Lemma 13. *Let $A = \frac{1}{n\varepsilon} \sum_{(X_i,y_i)\in J^{oc}} |\widetilde{X}_{ij}|^2$ and let X_{ij} be sampled from some heavy-tailed distribution with only some even order moment, say $2q$ bounded and denoted by μ_{2q}, then with probability $1-\delta$, we claim $A \le \tau$, where $\tau = \left(\frac{\mu_{2q}}{\delta}\right)^{1/q}$.*

Proof. Consider $\phi(z) = z^q$ in Eq. (26), and $r := 2q$. We assume the r^{th} moment is bounded and denoted by μ_r.

$$\tau = \left(\frac{\mathbb{E}\left[\left(\frac{\sum_i X_i^2}{n}\right)^q\right]}{\delta}\right)^{1/q} = \frac{1}{\delta^{1/q}}\left(\mathbb{E}\left[\left(\frac{\sum_i X_i^2}{n}\right)^q\right]\right)^{1/q}.$$

Hence, we start with the computation of $\mathbb{E}[A^q]$

$$\mathbb{E}[A^q] = \mathbb{E}\left[\left(\frac{\sum_i X_i^2}{n}\right)^q\right] = \frac{1}{n^q}\left(\mathbb{E}\left[\sum_{m_1+m_2+\cdots+m_n=q}\binom{q}{m_1 m_2\cdots m_n}\prod_{i=1}^{n}\left(X_i^2\right)^{m_i}\right]\right)$$
$$= \frac{1}{n^q}\left(\sum_{m_1+m_2+\cdots+m_n=q}\binom{q}{m_1 m_2\cdots m_n}\mathbb{E}\left[\prod_{i=1}^{n}X_i^{2m_i}\right]\right).$$

Further, we use Hölder's inequality,

$$\mathbb{E}[A^q] \le \frac{1}{n^q}\left(\sum_{m_1+m_2+\cdots+m_n=q}\binom{q}{m_1 m_2\cdots m_n}\prod_{i=1}^{n}\left(\mathbb{E}\left[X_i^{2m_i\times\frac{q}{m_i}}\right]\right)^{\frac{m_i}{q}}\right)$$
$$= \frac{1}{n^q}\left(\sum_{m_1+m_2+\cdots+m_n=q}\binom{q}{m_1 m_2\cdots m_n}\prod_{i=1}^{n}\left(\mathbb{E}\left[X_i^{2q}\right]\right)^{\frac{m_i}{q}}\right)$$
$$\le \frac{1}{n^q}\left(\sum_{m_1+m_2+\cdots+m_n=q}\binom{q}{m_1 m_2\cdots m_n}\prod_{i=1}^{n}\left(\mu_r\right)^{\frac{m_i}{q}}\right)$$
$$\le \frac{1}{n^q}\times n^q\times \mu_r^{\frac{\sum_i^n m_i}{q}} = \mu_r.$$

where we have used $r = 2q$, and the r^{th} moment is bounded and denoted by μ_r. We have also used multinomial theorem in the last step.

Hence, the final expression of τ is

$$\tau = \left(\frac{\mu_{2q}}{\delta}\right)^{1/q}.$$

□

A.7 Proof of Lemma 14

Lemma 14. *Under assumptions 1, 2 and after choosing* $\lambda \ge \Omega\left(\frac{2-\kappa}{\kappa}\left(\sigma\sigma_e\sqrt{\frac{k\log p}{n}} + \frac{\sigma\sigma_e\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)$ *and* $n \ge \Omega(\frac{k^3\log^2 p}{\varepsilon\log(1/\varepsilon)})$*, the following bound on* $\bar{\zeta}$ *from (16) holds with high probability* $\|\bar{\zeta}\|_\infty < 1$*, which ensures strict dual feasibility.*

Proof. We start by rewriting equation (16). Let $\bar{\zeta} = \zeta_{S^c}$ for the ease of notation.

$$\bar{\zeta} = \frac{1}{\lambda}\left(-\widehat{\Sigma}_{S^cS}^{J^{cc}}(\underline{\theta}-\underline{\theta}^*) + \frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}}\overline{X}_i e_i + \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\overline{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle\right)$$
$$= \widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\zeta - \frac{\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}}{\lambda}\left(\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}}\widetilde{X}_i e_i + \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\right)$$
$$+ \frac{1}{\lambda}\left(\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}}\overline{X}_i e_i + \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\overline{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle\right).$$

To bound the first term, we use Lemma 15 and the mutual incoherence condition.

The second term can be bounded using arguments similar to Lemma 6 and Lemma 7:

$$\left\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}}\widetilde{X}_i e_i\right\|_\infty + \left\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\widetilde{X}_i(y_i - \langle\widetilde{X}_i\,,\underline{\theta}\rangle)\right\|_\infty = \mathcal{O}\left(\frac{\lambda\kappa}{8\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty} + \frac{\sigma\sigma_e\sqrt{k}\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right).$$

Specifically, we can choose $t_1 = \frac{\lambda\kappa}{8\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty}$ in Lemma 6.

Similarly, we can use Lemma 6 with $t_1 = \frac{\lambda\kappa}{8}$ and union bound over $(p-k)$ entries to claim

$$\left\|\frac{1}{m}\sum_{(X_i,y_i)\in J^{cc}}\overline{X}_i e_i + \frac{1}{m}\sum_{(X_i,y_i)\in J^{oc}}\overline{X}_i(y_i - \langle\widetilde{X}_i, \underline{\theta}\rangle\right\|_\infty \leq \mathcal{O}\left(\frac{\lambda\kappa}{8} + \frac{\sigma\sigma_e\sqrt{k}\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right),$$

with high probability of $1-\mathcal{O}(1/p)$ if $n = \Omega\left(\frac{k^2\log(p)}{\varepsilon\log(1/\varepsilon)}\right)$.

Using the above results, we can claim:

$$\left\|\bar{\zeta}\right\|_\infty \leq 1 - \frac{\kappa}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) + \frac{\kappa}{4} + \mathcal{O}\left(\frac{\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty \sigma\sigma_e\sqrt{k}\varepsilon\log(1/\varepsilon)}{\lambda(1-\varepsilon)}\right).$$

We take $\lambda \geq \frac{8\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty \sigma\sigma_e\varepsilon\log(1/\varepsilon)}{\kappa(1-\varepsilon)} = \Omega\left(\frac{(2-\kappa)\sigma\sigma_e\sqrt{k}\varepsilon\log(1/\varepsilon)}{\kappa}\right)$ to claim:

$$\left\|\bar{\zeta}\right\|_\infty \leq 1 - \frac{\kappa}{4} + \mathcal{O}\left(\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right) < 1,$$

where we have used $\kappa = \Omega(\sigma\sigma_e\varepsilon\log(1/\varepsilon))$ in the last step. This completes the proof for strict dual feasibility. □

Lemma 15. *If Assumption 2 holds and* $n = \frac{k^3\log(p)}{\varepsilon\log(1/\varepsilon)}$, *then* $\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty \leq 1 - \frac{\kappa}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)$ *with probability at least* $1-\mathcal{O}(1/p)$.

Proof. We start by applying the Woodbury matrix identity for $\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}$ to bound $\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty$

$$\widehat{\Sigma}_{SS}^{J^{cc}-1} = \left(\frac{1}{1-\varepsilon}\widehat{\Sigma}_{SS}^{J^{*c}} - \frac{\varepsilon}{1-\varepsilon}\widehat{\Sigma}_{SS}^{J^{co}}\right)^{-1} = (1-\varepsilon)\left(\widehat{\Sigma}_{SS}^{J^{*c-1}} - \sum_{i=1}^{\infty}\left(\varepsilon\widehat{\Sigma}_{SS}^{J^{*c-1}}\widehat{\Sigma}_{SS}^{J^{co}}\right)^i\right).$$

Hence, the overall term simplifies to:

$$\begin{aligned}\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1} &= \left(\widehat{\Sigma}_{S^cS}^{J^{*c}} - \varepsilon\widehat{\Sigma}_{S^cS}^{J^{co}}\right) \times \left(\widehat{\Sigma}_{SS}^{J^{*c-1}} - \sum_{i=1}^{\infty}\left(\varepsilon\widehat{\Sigma}_{SS}^{J^{*c-1}}\widehat{\Sigma}_{SS}^{J^{co}}\right)^i\right) \\ &= \widehat{\Sigma}_{S^cS}^{J^{*c}}\widehat{\Sigma}_{SS}^{J^{*c-1}} - \varepsilon\widehat{\Sigma}_{S^cS}^{J^{co}}\widehat{\Sigma}_{SS}^{J^{*c-1}} - \left(\widehat{\Sigma}_{S^cS}^{J^{*c}} - \varepsilon\widehat{\Sigma}_{S^cS}^{J^{co}}\right)\sum_{i=1}^{\infty}\left(\varepsilon\widehat{\Sigma}_{SS}^{J^{*c-1}}\widehat{\Sigma}_{SS}^{J^{co}}\right)^i.\end{aligned}$$

We apply the triangle inequality to bound the ℓ_∞ norm.

$$\begin{aligned}\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty &\leq \left\|\widehat{\Sigma}_{S^cS}^{J^{*c}}\widehat{\Sigma}_{SS}^{J^{*c-1}}\right\|_\infty + \varepsilon\left\|\widehat{\Sigma}_{S^cS}^{J^{co}}\right\|_\infty\left\|\widehat{\Sigma}_{SS}^{J^{*c-1}}\right\|_\infty \\ &\quad + \left(\left\|\widehat{\Sigma}_{S^cS}^{J^{*c}}\right\|_\infty + \varepsilon\left\|\widehat{\Sigma}_{S^cS}^{J^{co}}\right\|_\infty\right)\sum_{i=1}^{\infty}\left(\varepsilon\left\|\widehat{\Sigma}_{SS}^{J^{*c-1}}\right\|_\infty\left\|\widehat{\Sigma}_{SS}^{J^{co}}\right\|_\infty\right)^i.\end{aligned}$$

We can bound the first term in the RHS using Lemma 17 as $1-\frac{\kappa}{2}$. We can use ideas similar to Lemma 5 to bound $\left\|\widehat{\Sigma}_{SS}^{J^{*c-1}}\right\|_\infty \leq \frac{3}{2}\left\|\Sigma_{SS}^{-1}\right\|_\infty$ with high probability of $1-\mathcal{O}(\frac{1}{p})$, if $n = \Omega(k^2\log(p))$. Similarly, $\left\|\widehat{\Sigma}_{S^cS}^{J^{co}}\right\|_\infty$ can be bounded as $\mathcal{O}(\log(1/\varepsilon))$ using arguments similar to Lemma 6. We apply sub-exponential tail bounds and algebraic details are similar to Eq. (24). The last term involves a geometric series and can be bounded if the common ratio term, $\varepsilon\left\|\widehat{\Sigma}_{SS}^{J^{*c-1}}\right\|_\infty\left\|\widehat{\Sigma}_{SS}^{J^{co}}\right\|_\infty = \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\left\|\Sigma_{SS}^{-1}\right\|_\infty\right) < 1$.

The overall term can be simplified as follows for sufficiently small ε

$$\left\|\widehat{\Sigma}_{S^cS}^{J^{cc}}\widehat{\Sigma}_{SS}^{J^{cc}-1}\right\|_\infty \leq 1 - \frac{\kappa}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right),$$

with high probability of $1-\mathcal{O}(1/p)$ if $n = \Omega\left(\frac{k^3\log(p)}{\varepsilon\log(1/\varepsilon)}\right)$. □

B Auxiliary Lemmas and Proofs

Lemma 16. *If Assumption 1 holds and* $n = \Omega(\frac{k+\log p}{\alpha_1^2 f(\varepsilon)})$*, where* $f(\varepsilon) = \varepsilon\log(1/\varepsilon)$*, then* $\left(1-\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)\alpha_1 I \preceq \widehat{\Sigma}_{SS}^{J^{cc}} \preceq \left(1+\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)\alpha_2 I$ *with probability at least* $1-\mathcal{O}(\frac{1}{p})$.

Proof. By the Courant-Fischer variational representation Horn and Johnson (2012):

$$\begin{aligned}\min_{\text{eig}}(\Sigma_{SS}) = \min_{\|y\|_2=1} y^T\Sigma_{SS}y &= \min_{\|y\|_2=1} y^T(\Sigma_{SS}-\widehat{\Sigma}_{SS}^{J^{cc}}+\widehat{\Sigma}_{SS}^{J^{cc}})y\\ &\le y^T(\Sigma_{SS})-\widehat{\Sigma}_{SS}^{J^{cc}}+\widehat{\Sigma}_{SS}^{J^{cc}})y\\ &= y^\intercal(\Sigma_{SS}-\widehat{\Sigma}_{SS}^{J^{cc}})y + y^\intercal\widehat{\Sigma}_{SS}^{J^{cc}}y.\end{aligned}$$

It follows that

$$\min_{\text{eig}}(\widehat{\Sigma}_{SS}^{J^{cc}}) \ge \alpha_1 - \|\Sigma_{SS}-\widehat{\Sigma}_{SS}^{J^{cc}}\|_2.$$

Further, we can proceed in a similar manner as compared to Lemma 6

$$\|\widehat{\Sigma}_{SS}^{J^{cc}}-\Sigma_{SS}\|_2 \le \frac{1}{1-\varepsilon}\|\widehat{\Sigma}_{SS}^{J^{*c}}-\Sigma_{SS}\|_2 + \frac{\varepsilon}{1-\varepsilon}\|\widehat{\Sigma}_{SS}^{J^{co}}-\Sigma_{SS}\|_2.$$

The first term $\|\widehat{\Sigma}_{SS}^{J^{*c}}-\Sigma_{SS}\|_2$ can be bounded using Proposition 2.1 in Vershynin (2012) for sub-Gaussian random variables.

For the second term, we use the heavier tail arising from sub-exponential concentration inequality to claim the following with high probability of $1-\mathcal{O}(\frac{1}{p})$:

$$\|\widehat{\Sigma}_{SS}^{J^{cc}}-\Sigma_{SS}\|_2 \le \frac{1}{1-\varepsilon}\frac{\alpha_1}{2} + \frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}.$$

Hence, the upper bound on the minimum eigenvalue will be

$$\min_{\text{eig}}(\widehat{\Sigma}_{SS}^{J^{cc}}) \ge \alpha_1 - \frac{\alpha_1/2}{1-\varepsilon} - \frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon} \ge \left(1-\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)\alpha_1.$$

with high probability of $1-\mathcal{O}(\frac{1}{p})$ if $n = \Omega(\frac{k+\log(p)}{\alpha_1^2\varepsilon\log(1/\varepsilon)})$. In the last step, we have used Assumption 1 in the last step, meaning $\alpha_1 = \Omega\left(1+\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)$.

Similarly, it can be shown that $\text{eig}_{\max}(\widehat{\Sigma}_{SS}^{J^{cc}}) \le \left(1+\mathcal{O}\left(\frac{\varepsilon\log(1/\varepsilon)}{1-\varepsilon}\right)\right)\alpha_2$ with probability at least $1-\mathcal{O}(\frac{1}{p})$. □

Lemma 17. *If Assumption 2 holds and* $n = \Omega(\frac{k^3(\log k+\log p)}{\tau(\alpha_1,\kappa,\sigma,\Sigma)})$*, then* $\|\widehat{\Sigma}_{S^cS}^{J^{*c}}\widehat{\Sigma}_{SS}^{J^{*c}-1}\| \le 1-\frac{\kappa}{2}$ *with probability at least* $1-\mathcal{O}(\frac{1}{p})$ *where* $\tau(\alpha_1,\kappa,\sigma,\Sigma)$ *is a constant independent of* n, p *and* k.

Proof. Before we prove the result of Lemma 17, we will prove a helper lemma.

Lemma 18. *If Assumption 2 holds then for some* $\delta > 0$*, the following inequalities hold:*

$$\begin{aligned}&\mathbf{P}(\|\widehat{\Sigma}_{S^cS}^{J^{*c}}-\Sigma_{S^cS}\|_\infty \ge \delta) \le 4(p-k)k\exp(-\frac{n\delta^2}{128k^2(1+4\sigma^2)\max_l\Sigma_{ll}^2}),\\ &\mathbf{P}(\|\widehat{\Sigma}_{SS}^{J^{*c}}-\Sigma_{SS}\|_\infty \ge \delta) \le 4k^2\exp(-\frac{n\delta^2}{128k^2(1+4\sigma^2)\max_l\Sigma_{ll}^2}),\\ &\mathbf{P}(\|(\widehat{\Sigma}_{SS}^{J^{*c}})^{-1}-(\Sigma_{SS})^{-1}\|_\infty \ge \delta) \le 2\exp(-\frac{c\delta^2\alpha_1^4 n}{4k}+k) + 2\exp(-\frac{c\alpha_1^2 n}{4}+k).\end{aligned} \tag{27}$$

Proof. Let A_{ij} be (i,j)-th entry of $\widehat{\Sigma}^{J^{*c}}_{S^cS} - \Sigma_{S^cS}$. Clearly, $\mathbb{E}(A_{ij}) = 0$. By using the definition of the $\|\cdot\|_\infty$ norm, we can write:

$$\mathbf{P}(\|\widehat{\Sigma}^{J^{*c}}_{S^cS} - \Sigma_{S^cS}\|_\infty \geq \delta) = \mathbf{P}(\max_{i\in S^c} \sum_{j\in S} |A_{ij}| \geq \delta) \leq (p-k)\mathbf{P}(\sum_{j\in S} |A_{ij}| \geq \delta)$$
$$\leq (p-k)k\mathbf{P}(|A_{ij}| \geq \frac{\delta}{k}),$$

where the second last inequality comes as a result of the union bound across entries in S^c and the last inequality is due to the union bound across entries in S. Recall that $X_i, i \in \{1,\ldots,p\}$ are zero mean random variables with covariance Σ and each $\frac{X_i}{\sqrt{\Sigma_{ii}}}$ is a sub-Gaussian random variable with parameter σ. Using the results from Lemma 1 of Ravikumar et al. (2010), for some $\delta \in (0, k\max_l \Sigma_{ll} 8(1+4\sigma^2))$, we can write:

$$\mathbf{P}(|A_{ij}| \geq \frac{\delta}{k}) \leq 4\exp(-\frac{n\delta^2}{128k^2(1+4\sigma^2)\max_l \Sigma^2_{ll}}).$$

Therefore,

$$\mathbf{P}(\|\widehat{\Sigma}^{J^{*c}}_{S^cS} - \Sigma_{S^cS}\|_\infty \geq \delta) \leq 4(p-k)k\exp(-\frac{n\delta^2}{128k^2(1+4\sigma^2)\max_l \Sigma^2_{ll}}).$$

Similarly, we can show that

$$\mathbf{P}(\|\widehat{\Sigma}^{J^{*c}}_{SS} - \Sigma_{SS}\|_\infty \geq \delta) \leq 4k^2\exp(-\frac{n\delta^2}{128k^2(1+4\sigma^2)\max_l \Sigma^2_{ll}}).$$

Next, we will show that the third inequality in (27) holds. Note that

$$\|(\widehat{\Sigma}^{J^{*c}}_{S^cS})^{-1} - (\Sigma_{S^cS})^{-1}\|_\infty = \|(\Sigma_{SS})^{-1}(\Sigma_{SS} - \widehat{\Sigma}^{J^{*c}}_{SS})(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1}\|_\infty$$
$$\leq \sqrt{k}\|(\Sigma_{SS})^{-1}(\Sigma_{SS} - \widehat{\Sigma}^{J^{*c}}_{SS})(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1}\|_2$$
$$\leq \sqrt{k}\|(\Sigma_{SS})^{-1}\|_2\|(\Sigma_{SS} - \widehat{\Sigma}^{J^{*c}}_{SS})\|_2\|(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1}\|_2.$$

Note that $\|\Sigma_{SS}\|_2 \geq \alpha_1$, thus $\|(\Sigma_{SS})^{-1}\|_2 \leq \frac{1}{\alpha_1}$. Similarly, $\|\Sigma_{SS}\|_2 \geq \frac{\alpha_1}{2}$ with probability at least $1-2\exp(-\frac{c\alpha_1^2 n}{4}+k)$. We also have $\|(\Sigma_{SS} - \widehat{\Sigma}^{J^{*c}}_{SS})\|_2 \leq \epsilon$ with probability at least $1-2\exp(-c\epsilon^2 n + k)$. Taking $\epsilon = \delta\frac{\alpha_1^2}{2\sqrt{k}}$, we get

$$\mathbf{P}(\|(\Sigma_{SS} - \widehat{\Sigma}^{J^{*c}}_{SS})\|_2 \geq \delta\frac{\alpha_1^2}{2\sqrt{k}}) \leq 2\exp(-\frac{c\delta^2\alpha_1^4 n}{4k} + k).$$

It follows that $\|(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1} - (\Sigma_{SS})^{-1}\|_\infty \leq \delta$ with probability at least $1-2\exp(-\frac{c\delta^2\alpha_1^4 n}{4k}+k) - 2\exp(-\frac{cC^2_{\min} n}{4}+k)$. □

Now we are ready to show that the statement of Lemma 17 holds using the results from Lemma 18. We will rewrite $\widehat{\Sigma}^{J^{*c}}_{S^cS}(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1}$ as the sum of four different terms:

$$\widehat{\Sigma}^{J^{*c}}_{S^cS}(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1} = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 \triangleq \widehat{\Sigma}^{J^{*c}}_{S^cS}((\widehat{\Sigma}^{J^{*c}}_{SS})^{-1} - (\Sigma_{SS})^{-1}),$$
$$T_2 \triangleq (\widehat{\Sigma}^{J^{*c}}_{S^cS} - \Sigma_{S^cS})(\Sigma_{SS})^{-1},$$
$$T_3 \triangleq (\widehat{\Sigma}^{J^{*c}}_{S^cS} - \Sigma_{S^cS})((\widehat{\Sigma}^{J^{*c}}_{SS})^{-1} - (\Sigma_{SS})^{-1}),$$
$$T_4 \triangleq \Sigma_{S^cS}(\Sigma_{SS})^{-1}.$$

Then it follows that $\|\widehat{\Sigma}^{J^{*c}}_{S^cS}(\widehat{\Sigma}^{J^{*c}}_{SS})^{-1}\|_\infty \leq \|T_1\|_\infty + \|T_2\|_\infty + \|T_3\|_\infty + \|T_4\|_\infty$. Now, we will bound each term separately. First, recall that Assumption 2 ensures that $\|T_4\|_\infty \leq 1-\kappa$.

Controlling T_1. We can rewrite T_1 as,

$$T_1 = -\Sigma_{S^cS}(\Sigma_{SS})^{-1}(\widehat{\Sigma}_{SS}^{J^{*c}} - \Sigma_{SS})(\widehat{\Sigma}_{SS}^{J^{*c}})^{-1}$$

then,

$$\begin{aligned}
\|T_1\|_\infty &= \|\Sigma_{S^cS}(\Sigma_{SS})^{-1}(\widehat{\Sigma}_{SS}^{J^{*c}} - \Sigma_{SS})(\widehat{\Sigma}_{SS}^{J^{*c}})^{-1}\|_\infty \\
&\leq \|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty \|(\widehat{\Sigma}_{SS}^{J^{*c}} - \Sigma_{SS})\|_\infty \|(\widehat{\Sigma}_{SS}^{J^{*c}})^{-1}\|_\infty \\
&\leq (1-\kappa)\|(\widehat{\Sigma}_{SS}^{J^{*c}} - \Sigma_{SS})\|_\infty \sqrt{k}\|(\widehat{\Sigma}_{SS}^{J^{*c}})^{-1}\|_2 \\
&\leq (1-\kappa)\|(\widehat{\Sigma}_{SS}^{J^{*c}} - \Sigma_{SS})\|_\infty \frac{2\sqrt{k}}{\alpha_1} \\
&\leq \frac{\kappa}{6}.
\end{aligned}$$

The last inequality holds with probability at least $1 - 2\exp(-\frac{c\alpha_1^2 n}{4} + k) - 4k^2\exp(-\frac{n\alpha_1^2\kappa^2}{18432(1-\kappa)^2k^3(1+4\sigma^2)\max_l \Sigma_{ll}^2})$ by taking $\delta = \frac{\alpha_1\kappa}{12(1-\kappa)\sqrt{k}}$.

Controlling T_2. Recall that $T_2 = (\widehat{\Sigma}_{S^cS}^{J^{*c}} - \Sigma_{S^cS})(\Sigma_{SS})^{-1}$. Thus,

$$\|T_2\|_\infty \leq \sqrt{k}\|(\Sigma_{SS})^{-1}\|_2 \|(\widehat{\Sigma}_{S^cS}^{J^{*c}} - \Sigma_{S^cS})\|_\infty \leq \frac{\sqrt{k}}{\alpha_1}\|(\widehat{\Sigma}_{S^cS}^{J^{*c}} - \Sigma_{S^cS})\|_\infty \leq \frac{\kappa}{6}.$$

The last inequality holds with probability at least $1 - 4(p-k)k\exp(-\frac{n\alpha_1^2\kappa^2}{4608k^3(1+4\sigma^2)\max_l \Sigma_{ll}^2})$ by choosing $\delta = \frac{\alpha_1\kappa}{6\sqrt{k}}$.

Controlling T_3. Note that,

$$\|T_3\|_\infty \leq \|(\widehat{\Sigma}_{S^cS}^{J^{*c}} - \Sigma_{S^cS})\|_\infty \|((\widehat{\Sigma}_{SS}^{J^{*c}})^{-1} - (\Sigma_{SS})^{-1})\|_\infty \leq \frac{\kappa}{6}.$$

The last inequality holds with probability at least $1 - 4(p-k)k\exp(-\frac{n\kappa}{768k^2(1+4\sigma^2)\max_l \Sigma_l l^2}) - 2\exp(-\frac{c\kappa\alpha_1^4 n}{24k} + k) - 2\exp(-\frac{c\alpha_1^2 n}{4} + k)$ by choosing $\delta = \sqrt{\frac{\kappa}{6}}$ in the first and third inequality of equation (27). By combining all the above results, we prove Lemma 17. □

C Additional Experiments

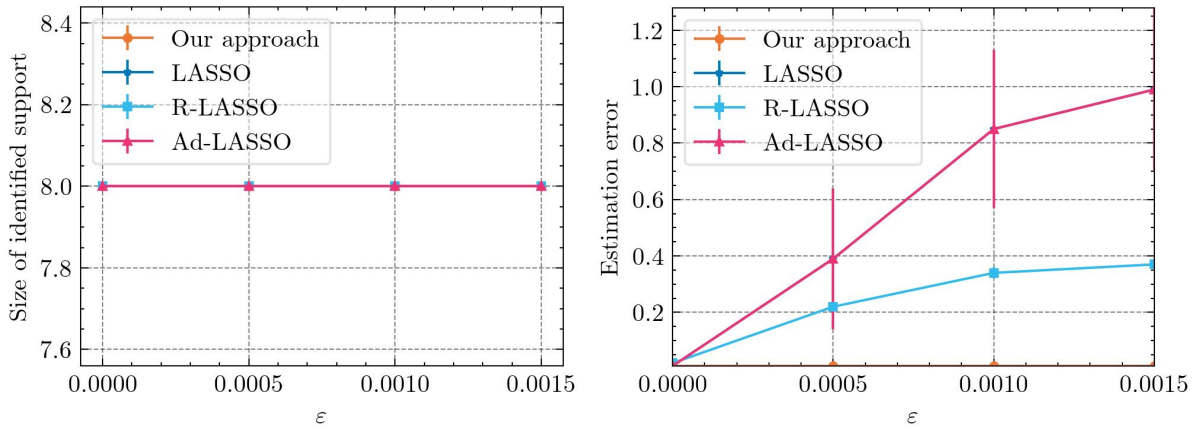
In this section, we present additional experiments to validate our theoretical findings. We compare our method with LASSO, R-LASSO (Chen et al., 2013), and Ad-LASSO (Lambert-Lacroix and Zwald, 2011; Zou, 2006). To obtain a KKT point, we employ an alternating minimization approach, akin to the method proposed in Awasthi et al. (2022), with LASSO serving as a key subroutine.

The predictors X_i were sampled from $\mathcal{N}(0, \frac{1}{n}\Sigma)$, where $\Sigma = I$. The additive noise e_i was independently drawn from $\mathcal{N}(0, \frac{\sigma_e^2}{n})$, with $\sigma_e = 2$. The non-zero entries of θ^* were uniformly sampled from $\{-1, 1\}$, and corrupted labels were randomly chosen from $[-2, 2]$. Following our theoretical framework, the regularization parameter λ for our method was set as $\lambda = \frac{\sigma_e}{n}\left(\sqrt{\frac{\log p}{n}} + \sqrt{k}\varepsilon\log(1/\varepsilon)\right)$, where ε represents the corruption proportion. The scaling factor $\frac{\sigma_e}{n}$ accounts for the variance of X_i and e_i. Consistent with Chen et al. (2013), we estimated the support of θ^* by identifying the indices corresponding to the largest k entries in the recovered regression vector. Estimation error was computed using $\|\widehat{\theta} - \theta^*\|$, measuring the deviation of the estimated vector $\widehat{\theta}$ from the ground truth θ^*.

We conducted experiments across three distinct regimes, and the results are summarized below:

C.1 Small ε regime

In this regime we vary the ε in $[0, 0.0015]$. The experiments were conducted in a setting with $p = 500, k = 8$, and $n = k^3 \log p = 3182$. Each point in the plot represents the average of 30 independent runs.



(a) Support recovery in small ε regime

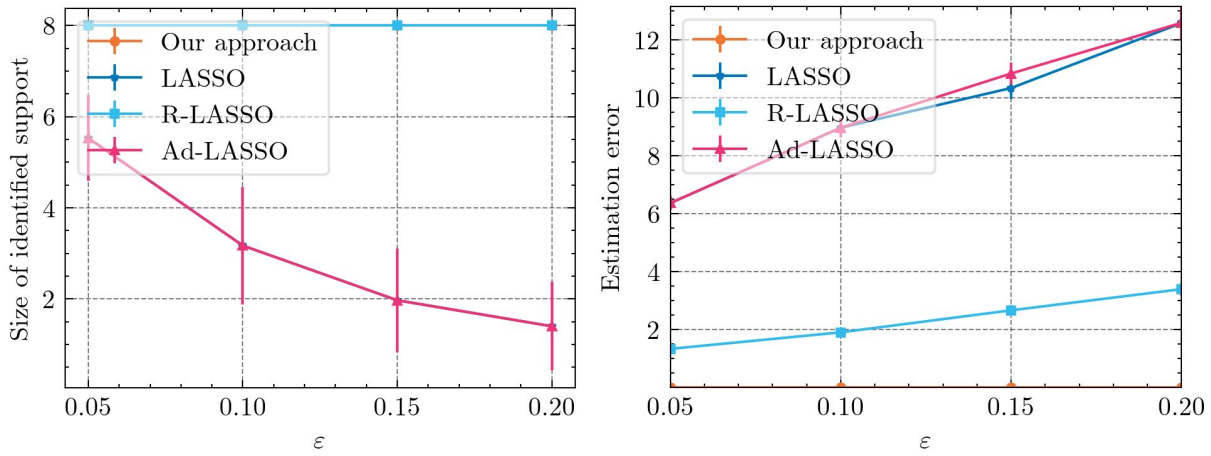
(b) Estimation error in small ε regime

Figure 2: (left) Support recovery against varying amounts of small corruption proportion, (Right) Estimation error against varying amounts of small corruption proportion.

Figure 2a demonstrates that in the small ε regime, all methods exhibit robustness to corruption with respect to support recovery. However, Figure 2b reveals that while all methods perform well, our approach achieves the lowest estimation error.

C.2 Large ε regime

In this regime we vary the ε in $[0.05, 0.2]$. The experiments were conducted in a setting with $p = 500, k = 8$, and $n = k^3 \log p = 3182$. Each point in the plot represents the average of 30 independent runs.



(a) Support recovery in large ε regime

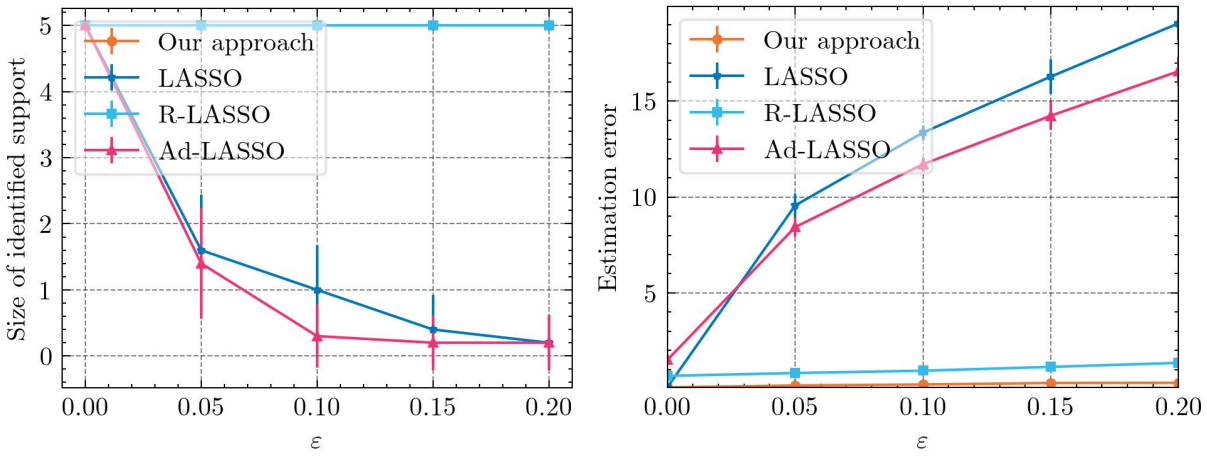
(b) Estimation error in large ε regime

Figure 3: (left) Support recovery against varying amounts of large corruption proportion, (Right) Estimation error against varying amounts of large corruption proportion.

Figure 3a illustrates that both our method and R-LASSO maintain robustness in support recovery even in the presence of large corruption proportions. In contrast, the performance of other baselines, such as LASSO and Ad-LASSO, deteriorates significantly as ε increases. Furthermore, Figure 3b highlights that while both R-LASSO and our approach successfully recover the support, our method achieves a notably lower estimation error.

C.3 High-dimensional low sample regime

In this regime, we conduct experiments with parameters $p = 2000, k = 5$, and $n = 1000$. Notably, this regime is characterized by a limited number of samples, specifically $n < p$. Each point in the plot represents the average outcome from 30 independent runs. For this setup, we vary the corruption proportion ε within the range $[0.05, 0.2]$.



(a) Support recovery, $p = 2000, n = 1000, k = 5$

(b) Estimation error, $p = 2000, n = 1000, k = 5$

Figure 4: (Left) Support recovery against varying amounts of corruption proportion in a high dimension low sample regime, (Right) Estimation error against varying amounts of corruption proportion in a high dimension low sample regime.

Figure 4a demonstrates that both our approach and R-LASSO exhibit robustness in support recovery within the high-dimensional, low-sample regime. In contrast, other baselines, such as LASSO and Ad-LASSO, show a marked decline in performance. Additionally, Figure 4b underscores that while both R-LASSO and our method accurately recover the support, our approach achieves a lower estimation error.