Provable Guarantees For Robust Feature Selection in Sparse Linear Models in High-Dimensions

Deepak Maurya[1] **Adarsh Barik**[2] **Jean Honorio**[1,3]

[1]Department of Computer Science, Purdue University, USA
[2]Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India
[3]School of Computing and Information Systems, The University of Melbourne, ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Australia

Abstract

We study feature selection (support recovery) in sparse linear models, including linear and logistic regression, under a strong adversarial contamination model where an adversary can arbitrarily corrupt a constant fraction of samples in the high-dimensional regime. Most of the existing methods target one or two aspects of this interesting problem: (1) achieving robustness against *strong* adversaries, (2) handling adversarial corruption in *both* features and labels, and (3) performing feature selection in *high dimensions*. Our approach tackles all three issues simultaneously. We propose a trimmed maximum likelihood estimator with ℓ_1-regularization, leading to a non-convex relaxation of an NP-hard combinatorial optimization problem. We prove that any locally optimal solution to this non-convex problem achieves *near-minimax* optimal statistical rates, up to logarithmic factors. Critically, our method recovers the true support in a computationally efficient manner despite the intractability of the exact formulation. The resulting sample complexity scales only logarithmically with the ambient dimension, making it well-suited for high-dimensional settings. Furthermore, our framework accommodates heavy-tailed noise distributions under mild moment assumptions, requiring only that the fourth moment be bounded. We also validate our theoretical findings empirically.

1 INTRODUCTION

High-dimensional linear regression is a fundamental learning problem, with applications in genomics, neuroscience, and econometrics. In essence, the goal in this problem is to estimate a sparse regression coefficient vector from observed predictor-response data. In high-dimensional learning problems, the ambient dimension p is much larger than the sample size n, making the sparsity of the model important for statistical efficiency and interpretability.

At the same time, real-world datasets are frequently contaminated by outliers, heavy-tailed noise, or adversarial corruptions, arising from data poisoning or system failures. These challenges motivate the learning of *robust* sparse regression in high dimensions ($p \gg n$ and $n = \Omega(\log(p))$).

In this work, we propose a practical polynomial-time algorithm for robust feature selection of linear models in high-dimensional setting under strong adversarial contamination model. Our main contributions are twofold:

1. **Robustness (Near-Optimal Error Rate)**: Our estimator achieves a near-minimax optimal error rate, differing from the information-theoretic minimax rate by only a logarithmic factor. Achieving the exact minimax rate of $\mathcal{O}(\varepsilon)$ error (where ε is the fraction of contaminated samples) is NP-hard in general [Gao, 2020]. Our polynomial-time algorithm attains an error of $\mathcal{O}(\varepsilon \log(1/\varepsilon))$, which is close to the minimax optimum up to a log-factor. This is a practically desirable trade-off between statistical optimality and computational tractability.
2. **High-Dimensional Efficiency**: Our sample complexity is also logarithmic in the number of variables, meaning $n = \Omega(\log(p))$. This is critical for high-dimensional problem as $p \gg n$ and is known to be minimax optimal even for the classical case of no adversary [Wainwright, 2009]. Thus, our method matches the optimal sample size dependence on p despite the presence of strong adversarial corruption.

Next, we briefly discuss the challenging aspects of problem, our algorithmic and theoretical contributions, and clarify the generalized assumptions of our setting in comparison to prior work.

Challenges: There are two primary challenges in our problem: (i) Feature selection in high dimensions, and (ii) Robustness to adversarial outliers. Even in the absence of an

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4381–4426.

Table 1: Our work focuses on the most challenging combination of assumptions: (1) an adaptive (strong) adversary, which can design corruptions after observing the clean data (more challenging than the Huber or oblivious models); (2) corruption in both the features $\mathbf{x}$ and responses y (harder than corruption in labels alone); and (3) the high-dimensional regime ($p \gg n$), which is more challenging than the low-dimensional setting. As shown, only a few prior works address the same difficult setting as ours, and we handle strictly more general conditions (adaptive adversary, label *and* feature corruption, high p) than most of the existing literature. In particular for linear regression, while Chen et al. [2013] and Liu et al. [2020] also consider the strong adversary with feature and label corruption in high dimensions, we provide a more comprehensive comparison in Table 2.

Paper (Linear Regression)	**Adversarial Model**	**Outliers in** (y, X)	**HD**	**Paper (Logit Models)**
Dalalyan and Thompson [2019]	Huber	(✓, ✗)	✓	Cantoni and Ronchetti [2001]
Diakonikolas et al. [2023a]	Huber	(✓, ✓)	✗	-
Balakrishnan et al. [2017], Gao [2020]	Huber	(✓, ✓)	✓	-
McWilliams et al. [2014], Bhatia et al. [2017]	Oblivious/Weak	(✓, ✗)	✗	Diakonikolas et al. [2023b]
Wright and Ma [2010], Nasrabadi et al. [2011], Dalalyan and Chen [2012], Nguyen and Tran [2013], Suggala et al. [2019], d'Orsi et al. [2021]	Oblivious/Weak	(✓, ✗)	✓	Yang et al. [2013]
Bhatia et al. [2015]	Adaptive/Strong	(✓, ✗)	✗	Zarifis et al. [2025]
Karmalkar and Price [2018], Minsker et al. [2024]	Adaptive/Strong	(✓, ✗)	✓	-
Klivans et al. [2018], Jambulapati et al. [2021], Awasthi et al. [2022], Pensia et al. [2025]	Adaptive/Strong	(✓, ✓)	✗	Feng et al. [2014], Awasthi et al. [2022], Mukhoty et al. [2023]
Chen et al. [2013], Liu et al. [2020] and **Ours***	Adaptive/Strong	(✓, ✓)	✓	**Ours***

adversary, identifying the correct support of a sparse linear model is an NP-hard problem (often referred to as best-subset selection). Likewise, detecting and removing outliers is NP-hard even in low-dimensional regression. Combining these, it is clear that robust feature selection in high dimensions (with adversarially corrupted data) is extremely challenging. This double difficulty motivates our focus on approximation algorithms and relaxations that can provably succeed in this setting.

Our Algorithmic Contribution: To tackle the aforementioned challenges, we introduce an estimator based on trimming outliers and L1-regularization. In particular, our algorithm can be viewed as solving a *trimmed* (outlier-pruned) likelihood maximization with an ℓ_1 penalty. While finding the global optimum of this formulation is combinatorially hard, we show that even any local optimum of our nonconvex relaxation enjoys the near-optimal error guarantees mentioned above. In other words, we obtain a *nearly optimal* solution in a *computationally efficient* manner, despite the underlying problem being NP-hard in its exact form.

Near-Optimality of Our Results: Notably, our theoretical guarantees nearly reach the information-theoretic limits of the problem. Gao [2020] established a minimax-optimal error rate of $\mathcal{O}(\varepsilon)$ by using a multivariate depth-based estimator for the easier case of the Huber adversarial model, but that approach is computationally intractable in high dimensions. In contrast, our estimator achieves an error bound of $\mathcal{O}(\varepsilon \log(1/\varepsilon))$, which is near-optimal up to the logarithmic factor. This result is derived by analyzing a local optimum of a nonconvex relaxation of the NP-hard trimmed sparse regression problem. Thus, we attain essentially near-optimal qualitative robustness guarantee running in polynomial time, as compared to the optimal method which is NP-hard.

1.1 RELATED WORKS

A plethora of existing research on robust linear regression makes various assumptions to tackle this problem. We briefly survey the most relevant literature, using the three aspects: adversarial model, contamination in both labels and features, and high-dimensionality to differentiate them, as summarized in Table 1.

Adversarial contamination models: We consider the most challenging adversarial model among the three major adversarial models studied in ML literature. In the *ε-Huber* contamination model, each sample is drawn *independently* from a mixture $(1-\varepsilon)P + \varepsilon Q$, where P is the true (clean) distribution and Q is an arbitrary outlier distribution. A stronger model is the *oblivious/weak* adversary, which can corrupt an ε-fraction of the samples, but does so *without* knowledge of the dataset (i.e. the corruptions are fixed or random, independent of the actual data). The most challenging setting is an *adaptive/strong* adversary, who can inspect the entire clean dataset (or even the underlying model) and then arbitrarily replace ε fraction of the samples with worst-

Table 2: We present a detailed comparison with Chen et al. [2013], Liu et al. [2020], who also address robust regression under strong (adaptive) adversarial corruption in both features and labels in high-dimensional settings. Our work is novel and an improvement in terms of generality of assumptions, sample complexity, and tighter estimation guarantees. **Generalized Assumptions:** Chen et al. [2013] assume sub-Gaussian features with a variance proxy of $\frac{1}{n}\sigma^2$, which is more restrictive than the standard σ^2 sub-Gaussian assumption used in Liu et al. [2020] and our work. Liu et al. [2020] additionally assume Gaussian features and an r-sparse covariance matrix. In contrast, our method only assumes sub-Gaussian features and does not require covariance sparsity; our mutual incoherence (MI) condition applies more generally. **Sample Complexity:** In high-dimensional regimes ($p \gg n$), achieving $n = \Omega(\log p)$ is crucial. While Liu et al. [2020] achieve this under sparsity assumptions on the covariance, their sample complexity becomes $\Omega(p^2 \log p)$ when these assumptions are relaxed. Our method maintains the logarithmic sample complexity without requiring covariance sparsity. **Asymptotic Error Bounds:** We derive sharp ℓ_∞ bounds for feature selection, while Chen et al. [2013] and Liu et al. [2020] report ℓ_2 error bounds. Our ℓ_∞ bounds are particularly well suited to the feature selection, enabling precise control over coordinate-wise errors. Here, n denotes the number of samples, p the number of features, k the number of nonzero entries in the true parameter vector, and ε denotes the fraction of samples corrupted by the adversary.

Paper (Linear Regression)	Assumptions: Features($\mathbf{x}$)	Assumptions: Covariance Matrix($\mathbf{x}$)	Sample Complexity	Asymptotic Error Bound $\boldsymbol{\Delta} = \mathbf{w} - \mathbf{w}^\star$
Chen et al. [2013]	sub-Gaussian($\frac{1}{n}\sigma^2$)	$\mathrm{Cov}(\mathbf{x}) = \frac{1}{n}\mathbf{I}$	$\Omega(\log p)$ ✓	$\|\boldsymbol{\Delta}\|_2 \le \mathcal{O}(\|\mathbf{w}^\star\|_2 \sqrt{k}\log(p)\varepsilon)$
	sub-Gaussian($\frac{1}{n}\sigma^2$)	$\mathrm{Cov}(\mathbf{x}) = \frac{1}{n}\boldsymbol{\Sigma}$	$\Omega(\log p)$ ✓	$\|\boldsymbol{\Delta}\|_2 \le \mathcal{O}(\|\mathbf{w}^\star\|_2 k\log(p)\varepsilon)$
Liu et al. [2020]	Gaussian(σ^2)	$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}$ is r-sparse	$\tilde{\Omega}(\log(p))$ ✓	$\|\boldsymbol{\Delta}\|_2 \le \tilde{\mathcal{O}}(\sqrt{\varepsilon})$
	Gaussian(σ^2)	$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$	$\tilde{\Omega}(p^2\log(p))$ ✗	$\|\boldsymbol{\Delta}\|_2 \le \tilde{\mathcal{O}}(\sqrt{\varepsilon})$
Ours*	sub-Gaussian(σ^2)	$\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$, MI	$\Omega(\log(p))$ ✓	$\|\boldsymbol{\Delta}\|_\infty \le \mathcal{O}(\varepsilon\log(1/\varepsilon))$

case corruptions. Our work operates under this strongest adversarial model, which strictly generalizes the other two models.

Feature & label corruption: Another challenging aspect of our work is that we consider adversarial contamination in both features and labels instead of just labels. This is very critical as it increases the complexity of the problem significantly. For example, consistent robust regression may be possible in case of the contamination in labels only Bhatia et al. [2017] but may not be possible if both labels and features are corrupted Gao [2020]. Our setting allows corruption in both labels and features, which is more realistic and challenging than the label-only case considered in some of the literature.

High-dimensional setting: Finally, methods differ in whether they address the high-dimensional regime (p comparable to or much larger than n). Techniques that work in low dimensions (where $n \gg p$) often break down when $p \gg n$ because key matrices become singular and standard linear algebra tools fail. In this work, we consider the high-dimensional setting.

As seen from Table 1, by addressing the most general and challenging setting, our work subsumes many of the special cases studied previously. To the best of our knowledge, existing results for high-dimensional logistic regression and Logit family models in general do not address adaptive (strong) adversarial corruptions simultaneously affecting both features and labels. Our work fills this gap.

For linear regression, the closest prior works to our setting are those of Chen et al. [2013] and Liu et al. [2020], which also consider a strong adversary corrupting both features and labels in a high-dimensions. Hence, we present a careful deeper comparison with Chen et al. [2013] and Liu et al. [2020] in Table 2 and briefly discuss it below.

1.1.1 Comparison of our results with Chen et al. [2013], Liu et al. [2020]

Our method improves upon these works along several key dimensions:

Generalized Assumptions: Chen et al. [2013] assume sub-Gaussian features with variance proxy parameter as $\frac{1}{n}\sigma^2$ and covariance matrix of form $\frac{1}{n}\Sigma$. This is significantly smaller due to $1/n$ factor as compared to standard σ^2 sub-Gaussian and covariance matrix Σ assumption used in Liu et al. [2020] and our work. Also, Liu et al. [2020] assumes Gaussian features and an r-sparse covariance matrix, whereas our method allows general sub-Gaussian covariates without requiring sparsity in the covariance. Our mutual incoherence (MI) condition is broadly applicable and does not require sparsity in covariance matrix.

Sample Complexity: In high-dimensional settings ($p \gg n$), it is critical to achieve sample complexity that scales logarithmically with p. While Liu et al. [2020] achieve this rate

under sparsity assumptions on the covariance, their required sample size becomes quadratic in p for non-sparse covariance matrix. In contrast, our method maintains $\Omega(\log p)$ sample complexity under mutual incoherence condition which allows for a non-sparse covariance matrix. This is critical in robustness as the covariance matrix is not known a priori to be sparse due to adversarial contamination.

Sharper Estimation Guarantees: Finally, we derive ℓ_∞ error bounds for support recovery, in contrast to the ℓ_2 bounds reported in Chen et al. [2013], Liu et al. [2020]. The ℓ_∞ guarantee offers finer control over coordinate-wise estimation error, making it especially well-suited for the feature selection task. Using norm inequalities such as $\|\mathbf{\Delta}\|_\infty \leq \|\mathbf{\Delta}\|_2$ may lead to weaker bounds as $\|\mathbf{\Delta}\|_2$ bounds the entire vector instead of the maximum absolute entry like $\|\mathbf{\Delta}\|_\infty$ directly. For example, as presented in Table 2, the presence of terms like $\|\mathbf{w}^\star\|_2$ makes the error bound by Chen et al. [2013] weaker.

Notation: We represent sets using calligraphic symbols, such as $\mathcal{P}$. Furthermore, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For a vector, $\mathbf{a}_i$ represents the i^{th} element of vector $\mathbf{a}$. A superscript of (i) like $\mathbf{a}^{(i)}$ denotes i^{th} sample of vector $\mathbf{a}$. In the case of a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, we denote the sub-matrix with rows in $\mathcal{P} \subseteq [p]$ and columns in $\mathcal{Q} \subseteq [q]$ as $\mathbf{A}_{\mathcal{P}\mathcal{Q}} \in \mathbb{R}^{|\mathcal{P}|\times|\mathcal{Q}|}$. We denote the induced ℓ_∞ norm using $|||\mathbf{A}|||_\infty = \max_{i\in[m]} \sum_{j\in[k]} |\mathbf{A}_{ij}|$ for a matrix $\mathbf{A} \in \mathbb{R}^{m\times k}$. Similarly, $|||\mathbf{A}|||_2$ denotes the spectral norm and $\|\mathbf{A}\|_\infty$ denotes the entrywise ℓ_∞ norm of a matrix. The minimum and maximum eigenvalues are denoted using $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$. For a vector $\mathbf{w} \in \mathbb{R}^p$, $\mathcal{S}(\mathbf{w}) = \{i \in [p] \mid \mathbf{w}_i \neq 0\}$ represents the support of $\mathbf{w}$. Likewise, $\mathcal{S}^c(\mathbf{w}) = [p] \setminus \mathcal{S}(\mathbf{w})$ denotes the non-support.

2 DATA GENERATION PROCESS

We start by formally defining the data generating process for a generalized linear model.

Definition 2.1. (Generalized Linear Model) The probability density function of the labels y for given feature $\mathbf{x} \in \mathbb{R}^p$ and parameter $\mathbf{w} \in \mathbb{R}^p$ is

$$f(y|\mathbf{w}^\intercal\mathbf{x}) = c(y)\exp\left(y\mathbf{w}^\intercal\mathbf{x} - b\left(\mathbf{w}^\intercal\mathbf{x}\right)\right), \qquad (1)$$

where $b : \mathbb{R} \mapsto \mathbb{R}$ is the cumulant function and $c : \mathbb{R} \mapsto \mathbb{R}$ is used for normalization. The first and second order derivatives of the function $b(\cdot)$ are called mean and variance function respectively, as $\mathbb{E}[y|\mathbf{w}^\intercal\mathbf{x}] = b'(\mathbf{w}^\intercal\mathbf{x})$ and $\text{var}[y|\mathbf{w}^\intercal\mathbf{x}] = b''(\mathbf{w}^\intercal\mathbf{x})$. Table 3 lists distribution like Gaussian which leads to linear regression and logit family, which covers logistic, binomial, categorical, and multinomial regression.

Algorithm 1 Robust MoM Filtering

Input: $\{\mathbf{x}^{(i)}\}_{i=1}^n \subset \mathbb{R}^p$, δ, ε
Output: Remaining index set $\mathcal{R}$
Set $B = \mathcal{O}(\log(p/\delta))$, $T = \mathcal{O}\left(\sqrt{\log(p/\varepsilon)}\right)$, $\mathcal{R} \leftarrow \emptyset$
Partition $[n]$ into blocks $G^{(1)}, \ldots, G^{(B)}$
for $j = 1$ **to** p **do**
  $\widehat{\mu}_{1j} \leftarrow \underset{b\in[B]}{\text{median}}\left(\frac{1}{|G^{(b)}|}\sum_{i\in G^{(b)}} \mathbf{x}_j^{(i)}\right)$
  $\widehat{\mu}_{2j} \leftarrow \underset{b\in[B]}{\text{median}}\left(\frac{1}{|G^{(b)}|}\sum_{i\in G^{(b)}} \left(\mathbf{x}_j^{(i)}\right)^2\right)$
  $\widehat{\sigma}_j \leftarrow \sqrt{\widehat{\mu}_{2j} - \widehat{\mu}_{1j}^2}$
end for
$\mathcal{R} \leftarrow \left\{ i \in [n] \;\middle|\; \left\|(\mathbf{x}^{(i)} - \widehat{\mu}_1) \oslash \widehat{\sigma}\right\|_\infty \leq T \right\}$

We assume covariates $\mathbf{x}^{(i)} \in \mathbb{R}^p$ for all clean samples to be a zero mean sub-Gaussian random vector [Hsu et al., 2012] with covariance $\mathbf{\Sigma}$. Further, labels y are generated using the GLM model specified in Definition 2.1. We assume n independent and identically distributed (i.i.d.) samples are generated.

Adaptive/Strong Contamination Model: Further, an adversary inspects these n clean samples and replaces ε proportion of the samples arbitrarily.

Problem of Interest: We are given a data set of n i.i.d samples which constitutes of at least $m < n$ clean samples following GLM specified in Definition 2.1 and εn arbitrary samples introduced by adversary. We are interested in estimating a sparse GLM with $\|\mathbf{w}\|_0 \leq k$ specified in Definition 2.1 from these n samples.

Formally speaking, we derive the sample complexity for feature selection based on the true parameter vector $\mathbf{w}$ in Definition 2.1.

3 OUR NOVEL FILTERING

In this section, we start by discussing our novel median-of-means based pre-filtering step. As we know that the clean samples are generated from a sub-Gaussian distribution, we use this information to remove extreme outliers or heavy-tailed noise introduced by adversary.

Algorithm 1 is discussed as follows. Also note that the symbol $\oslash$ in definition of set $\mathcal{R}$ denotes coordinate-wise division.

We use a median-of-means (MoM)–based filtering procedure to robustly pre-process the feature vectors in the presence of heavy-tailed noise and adversarial outliers. The

Table 3: Cumulant function, its derivatives, and loss functions in GLMs

Model	$c(y)$	$b(\theta)$	$b'(\theta)$	$b''(\theta)$	$f(y\|\theta)$	$-\log(f(y\|\theta=\mathbf{w}^\intercal\mathbf{x}^{(i)}))$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$	$\frac{\theta^2}{2}$	θ	1	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y-\theta)^2}{2}\right)$	$C\left(y-\mathbf{w}^\intercal\mathbf{x}^{(i)}\right)^2$
Logit	1	$\log(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	$\frac{e^\theta}{(1+e^\theta)^2}$	$\frac{e^{y\theta}}{1+e^\theta}$	$\log\left(1+e^{\mathbf{w}^\intercal\mathbf{x}^{(i)}}\right)-y\mathbf{w}^\intercal\mathbf{x}^{(i)}$

samples are first partitioned into multiple disjoint blocks, and coordinate-wise robust estimates of the mean and second moment are computed via medians over block averages. These estimates are used to construct a robust scale estimate for each feature. A sample is retained if its coordinate-wise normalized deviation, measured relative to the robust mean and scale, is uniformly bounded by a threshold. This filtering step removes samples exhibiting anomalously large deviations in any coordinate while preserving the bulk of the uncontaminated data.

Algorithm 1 provides a robust filtering mechanism under adversarial contamination, assuming that at least $(1-\varepsilon)$ fraction of the samples follow sub-Gaussian distribution. The median of means based estimator is a computationally efficient way of robustly estimating mean and variance. We just require $B=\mathcal{O}(\log(p/\delta))$ disjoint blocks to obtain ℓ_∞-bounds. The standardized residuals behave like normalized sub-Gaussian random variables. The threshold $T=\mathcal{O}\left(\sqrt{\log\frac{p}{\varepsilon}}\right)$ is chosen to ensure that approximately no clean sample is removed. The $\log(p)$ factor in both number of blocks, B and threshold, T appears due to the union bound over p variables. At the end, we can guarantee with high probability that the algorithm discards most corrupted samples while preserving the majority of the clean samples.

We formalize the preceding intuition through rigorous theoretical guarantees for the proposed filtering procedure. In particular, the following theoretical claims are presented in Appendix 9:

1. We derive high-probability bounds on the accuracy of the robust estimators for the first-order moment $\widehat{\mu}_1$ and the second-order moment $\widehat{\mu}_2$ in Theorem 9.1 and Corollary 9.2, respectively.

2. We show that the resulting coordinate-wise variance estimator is accurate with high probability in Lemma 9.3.

3. We prove that the filtering procedure preserves all uncontaminated samples with high probability in Lemma 9.4.

Note that we do not use the pre-processed samples further. We use Algorithm 1 to obtain the set of well behaved samples which are used without any pre-processing further.

4 OUR PROPOSED ALGORITHM

In this section, we present the ℓ_1-regularized trimmed maximum likelihood estimator.

4.1 OUR REGULARIZED TRIMMED MLE

The trimmed MLE problem with ℓ_1 regularization is defined below.

Definition 4.1. (Regularized Trimmed MLE) The optimization problem is presented below:

$$\begin{aligned}&\left(\tilde{\mathbf{w}},\tilde{\mathbf{t}}\right)=\underset{\mathbf{w},\mathbf{t}}{\arg\min}\;\; l(\mathbf{w},\mathbf{t})+\lambda\left\|\mathbf{w}\right\|_1,\\&\text{such that } \mathbf{1}^\intercal\mathbf{t}=m,\;0\le\mathbf{t}_i\le 1,\;\forall\, i\in[n],\\&l(\mathbf{w},\mathbf{t}):=-\frac{1}{m}\sum_{i=1}^{n}\mathbf{t}_i\log\left(f\left(y^{(i)}|\mathbf{w}^\intercal\mathbf{x}^{(i)}\right)\right)\end{aligned}\tag{2}$$

where $f(\cdot)$ is defined in Eq. (1) and λ is the regularization parameter.

Further, we claim Eq. (2) is non-convex in $(\mathbf{w},\mathbf{t})$. Hence, achieving global optima is not guaranteed.

Lemma 4.2. *If* $\exists i,\;\mathbf{x}^{(i)}\neq 0$ *for* $i\in[n]$, *then* $l(\mathbf{w},\mathbf{t})$ *in Eq.* (2) *is non-convex in* $(\mathbf{w},\mathbf{t})$.

Hence 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.

5 OUR MAIN THEOREM

Before stating our main theorem, we state the standard mutual incoherence assumption prevalent in support recovery literature [Wainwright, 2009, Ravikumar et al., 2010, 2011, Daneshmand et al., 2014].

Assumption 5.1 (Mutual Incoherence)**.** For some $\gamma\in(0,1]$ with $\gamma=\omega\left(\varepsilon\log(1/\varepsilon)\right)$, $\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\|_\infty\le 1-\gamma$, where $\|\cdot\|_\infty$ is the norm matrix induced ∞-norm.

Assumption 5.1 requires that the non-support features are not too strongly correlated with the support features, relative

to the correlations among the support features themselves. Intuitively, if some $j \in S^c$ is nearly a linear combination of the support features, then coordinate j is statistically indistinguishable from a true-support coordinate even with infinite clean data.

Next, we present the informal version of our main theorem for Gaussian model (linear regression) and Logit model (example: logistic regression).

Theorem 5.2 (Informal)**.** *Let* $(\tilde{\mathbf{t}}, \tilde{\mathbf{w}})$ *be a KKT point of the problem* (2)*. Under assumption 5.1, if* $\lambda = \Omega\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon)\right)$ *and* $n = \Omega\left(\frac{k^2\log(p)}{\varepsilon^2\log^2(1/\varepsilon)}\right)$*, then with probability of at least* $1 - \mathcal{O}\left(\frac{1}{p}\right)$

1. Support of $\tilde{\mathbf{w}}$ matches exactly with support of $\mathbf{w}^\star$.
2. The error bound for Gaussian model (linear regression):

$$\|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon)\right).$$

3. The error bound for Logit model (logistic regression):

$$\|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon) + \sqrt{g(\sigma^2, \varepsilon, p)}\right),$$

where $g(\sigma^2, \varepsilon, p)$ is bounded as

$$g(\sigma^2, \varepsilon, p) = \begin{cases} \mathcal{O}\left(\sigma_a^2\log(1/\varepsilon)\right) & \mathbf{x}_a^{(i)} \sim \text{sub-Gaussian}(\sigma_a) \\ \mathcal{O}\left(\sigma^2\log(p/\varepsilon)\right) & \mathbf{x}_a^{(i)} \sim \text{heavy-tailed with bounded } 4^{\text{th}}\text{-moment} \end{cases}$$

Here, $\mathbf{x}_a^{(i)}$ denotes the sample contaminated by the adversary.

Proof Sketch: For linear regression, please refer to Lemma 12.1 and Lemma 12.2 in Appendix 12. For logit models, please refer to Lemma 13.1 and Lemma 13.2 in Appendix 13. Lemma 12.1 and Lemma 13.1 prove the first claim: support of $\tilde{\mathbf{w}}$ matches exactly with support of $\mathbf{w}^\star$. The upper bound on the ℓ_∞ norm of the parameter vector is proved in Lemma 12.2 and Lemma 13.2. The proof also relies on going from any non-integral KKT point to integral point, which can be done easily as shown in Lemma 11.1.

Before diving into the mathematical details of the proof, we briefly discuss our main theorem.

Interpretation of Theorem 5.2. The error bound decomposes into two regimes. The term $\sqrt{\log(p)/n}$ is the standard statistical price of high-dimensional estimation, while $\epsilon\log(1/\epsilon)$ is injected by the strong adversary, which is unavoidable up to the logarithmic factor. The additional term $g(\sigma^2, \epsilon, p)$ arises only for logit models and is defined casewise according to the tail behaviour of the covariates: under sub-Gaussian covariates $g = O(\sigma_a^2\log(1/\epsilon))$, whereas under heavy-tailed covariates with only a bounded fourth moment $g = O(\sigma^2\log(p/\epsilon))$. The extra $\log(p)$ factor in the heavy-tailed case is a consequence of the median-of-means filter (Algorithm 1), which must threshold uniformly over all p coordinates. For linear regression, $\nabla_w^2 l(\nu, \tilde{t}^{J^{cc}})$ is independent of ν, which is precisely why no analogue of g appears in this case.

6 PROOF SKETCH

The Lagrangian can be written as:

$$L(\mathbf{w}, \mathbf{t}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu) = l(\mathbf{w}, \mathbf{t}) + \boldsymbol{\alpha}^\intercal(-\mathbf{t}) + \boldsymbol{\beta}^\intercal(\mathbf{t} - \mathbf{1}_n) + \mu(-\mathbf{1}^\intercal\mathbf{t} + m),$$

where $(\mathbf{w}, \mathbf{t})$ are primal variables and $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu)$ are dual variables.

6.1 STATIONARITY CONDITION

First, we define the sets:

$$\begin{aligned} J^{cc} &= \{i \mid \mathbf{t}_i^* = 1, \tilde{\mathbf{t}}_i = 1, i \in [n]\}, \\ J^{co} &= \{i \mid \mathbf{t}_i^* = 1, \tilde{\mathbf{t}}_i = 0, i \in [n]\}, \\ J^{oc} &= \{i \mid \mathbf{t}_i^* = 0, \tilde{\mathbf{t}}_i = 1, i \in [n]\}, \\ J^{oo} &= \{i \mid \mathbf{t}_i^* = 0, \tilde{\mathbf{t}}_i = 0, i \in [n]\}, \\ \hat{J}^c &= \{i \mid \tilde{\mathbf{t}}_i = 1, i \in [n]\} = J^{cc} \cup J^{oc}, \\ J^{*c} &= \{i \mid \mathbf{t}_i^* = 1, i \in [n]\} = J^{cc} \cup J^{co}. \end{aligned}$$

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

The first order stationary condition with respect to $\mathbf{w}$, stated in Eq. (6) can be re-stated for $(\tilde{\mathbf{w}}, \tilde{\mathbf{t}})$ and sub-differential $\tilde{\mathbf{z}} \in \partial\|\tilde{\mathbf{w}}\|$ as

$$\nabla l_{\mathbf{w}}\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\hat{J}^c}\right) + \lambda\tilde{\mathbf{z}} = \mathbf{0}_{p\times 1},$$

where

$$\nabla l_{\mathbf{w}}(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\mathcal{J}}) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}\left(b'(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right)$$

and the decision variable $\tilde{\mathbf{t}}_i^{\mathcal{J}} = 1$ if $i \in \mathcal{J}$ and 0 otherwise.

Using mean value theorem and after some algebraic manipulations which are detailed in Section 11.1, we arrive at the following

$$\begin{bmatrix} \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{J^{cc}} & \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}^c}^{J^{cc}} \\ \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}} & \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}^c}^{J^{cc}} \end{bmatrix}\begin{bmatrix} \tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}_{\mathcal{S}}^\star \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{g}}_{\mathcal{S}}^{J^{cc}} \\ \hat{\mathbf{g}}_{\mathcal{S}^c}^{J^{cc}} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{r}}_{\mathcal{S}}^{J^{cc}} \\ \hat{\mathbf{r}}_{\mathcal{S}^c}^{J^{cc}} \end{bmatrix} + \varepsilon\begin{bmatrix} \hat{\mathbf{h}}_{\mathcal{S}}^{J^{oc}} \\ \hat{\mathbf{h}}_{\mathcal{S}^c}^{J^{oc}} \end{bmatrix} + \lambda\begin{bmatrix} \tilde{\mathbf{z}}_{\mathcal{S}} \\ \tilde{\mathbf{z}}_{\mathcal{S}^c} \end{bmatrix} = \mathbf{0}_{p\times 1},$$

where we use $\hat{\mathbf{r}}$ to denote $\hat{\mathbf{r}}_{\mathcal{S}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}})$ for the ease of notation, and $\hat{\mathbf{g}}$ and $\hat{\mathbf{Q}}$ are defined as follows:

$$\hat{\mathbf{g}}^{\mathcal{J}} = \nabla_{\mathbf{w}} l\left(\mathbf{w}^{\star}, \tilde{\mathbf{t}}^{\mathcal{J}}\right), \ \hat{\mathbf{Q}}^{\mathcal{J}} = \nabla^2_{\mathbf{w}} l\left(\mathbf{w}^{\star}, \tilde{\mathbf{t}}^{\mathcal{J}}\right), \quad (3)$$
$$\hat{\mathbf{h}}^{\mathcal{J}} = \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\mathcal{J}}\right),$$

and where the remainder $\mathbf{r}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}})$ is defined as:

$$\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}) := \left(\nabla^2_{\mathbf{w}} l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{cc}}\right) - \nabla^2_{\mathbf{w}} l\left(\mathbf{w}^{\star}, \tilde{\mathbf{t}}^{J^{cc}}\right)\right)(\tilde{\mathbf{w}} - \mathbf{w}^{\star}).$$

Here $\boldsymbol{\nu} = a\mathbf{w} + b\mathbf{w}^{\star}$, where $a \geq 0$, $b \geq 0$ and $a + b = 1$. Note that $\hat{\mathbf{r}} = \mathbf{0}$ for linear regression as $\nabla^2_{\mathbf{w}} l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{cc}}\right) = 1$ and $\nabla^2_{\mathbf{w}} l\left(\mathbf{w}^{\star}, \tilde{\mathbf{t}}^{J^{cc}}\right) = 1$ for linear regression. After some algebraic manipulations, we obtain:

$$\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^{\star}_{\mathcal{S}} = -\hat{\mathbf{Q}}^{J^{cc}-1}_{\mathcal{S}\mathcal{S}}\left(\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} + \lambda \mathbf{z}_{\mathcal{S}}\right) \quad (4)$$

$$\mathbf{z}_{\mathcal{S}^c} = \frac{1}{\lambda}\left(-\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}^c} - \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}^c} - \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}^c}\right)$$
$$+ \frac{1}{\lambda}\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}^{-1}_{\mathcal{S}\mathcal{S}}\left(\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} + \lambda \mathbf{z}_{\mathcal{S}}\right).$$

Further, we use norm inequalities and the triangle inequality on the RHS and bound the individual terms as shown in Eq. (18) and Eq. (19). Ensuring $\|\mathbf{z}_{\mathcal{S}^c}\|_{\infty} < 1$ helps to show that support of $\tilde{\mathbf{w}}$ matches exactly with support of $\mathbf{w}^{\star}$. Please refer to Appendix 12 and Appendix 13 for more details.

7 EMPIRICAL VALIDATION

We present empirical analysis to validate the theoretical claims of the paper on sparse linear and logistic regression under adversarial contamination on synthetic datasets and also test our proposed method on real-world cancer datasets. A detailed description of the experimental setup, data generation process, and hyperparameter settings for all methods is provided in Appendix 14 and Appendix 15.

Baselines and Metrics. For linear regression, the directly comparable baselines as per Table 2 are Robust LASSO [Chen et al., 2013] and RSGE [Liu et al., 2020]. We additionally include LASSO [Tibshirani, 1996] to study the effect of trimming, and methods suited for non-sparse models in low dimensions: Trimmed MLE [Awasthi et al., 2022] and SVAM [Mukhoty et al., 2023].

For logistic regression, there are no directly comparable existing methods. We therefore compare against LASSO GLM [Friedman et al., 2010] and methods suited for non-sparse, low-dimensional settings: Robust Logistic Regression (RoLR) [Feng et al., 2014], Trimmed MLE [Awasthi et al., 2022], and SVAM [Mukhoty et al., 2023].

We evaluate all methods using two complementary metrics:

1. Support recovery rate: It is defined as the fraction of trials achieving *exact* support recovery.
2. F1 score: It captures partial support recovery enabling finer-grained comparison when *exact* recovery is rare.

7.1 LINEAR REGRESSION

Low-Dimensional Regime. We begin with a low-dimensional setting ($n = 200$, $p = 30$, $k = 5$) under random adversarial corruption. As shown in Tables 6 and 7 in Appendix 14.3.1, our proposed method achieves consistently strong support recovery rates and F1 scores across all corruption levels, comparing favorably with existing baselines. We note that RSGE [Liu et al., 2020] delivers competitive recovery quality but relies on solving a semidefinite program (SDP) at each iteration, resulting in runtimes on the order of 10^5-10^6 milliseconds per run (Table 8, Appendix 14.3.2), which limits its scalability. Hence, for high-dimensional setting, we compare against other methods.

High-Dimensional Regime. We set $n = 200$, $p = 400$, $k = 5$ and evaluate across $\varepsilon \in \{0.05, 0.1, 0.15, 0.2\}$ under three corruption models.

1. **Random adversarial corruption.** The adversary corrupts data randomly, without knowledge of the clean dataset. Our method achieves high support recovery and F1 scores across all ε (Tables 11 and 12 in Appendix 14.3.3).
2. **Adaptive/strong adversarial corruption [Chen et al., 2013].** The adversary has full knowledge of the clean data and true support, deliberately targeting support recovery failure. We note that our experiments use constant per-sample variance rather than the variance decaying as $O(1/n)$ assumed in Chen et al. [2013], which makes the corruption model more challenging. Table 4 shows that our method achieves high F1 scores across all ε. A detailed discussion is presented in Appendix 14.4.
3. **Heavy-tailed corruption.** We validate our theoretical robustness to heavy-tailed noise (bounded fourth moment) using log-normal and Student's t distribution. Our method achieves high support recovery and F1 scores under both distributions (Tables 13–16 in Appendix 14.5).

Remark. In Table 4, LASSO and Robust LASSO achieve higher F1 scores than Trimmed MLE and SVAM, which is consistent due to ℓ_1-regularization. The stronger F1 scores of our method relative to Robust LASSO are consistent with our more general assumption of $\mathbf{x}^{(i)} \sim$ sub-Gaussian(σ^2), compared to the more restrictive sub-Gaussian$(\frac{1}{n}\sigma^2)$ in Chen et al. [2013] (see Table 2).

7.2 LOGISTIC REGRESSION

Low-Dimensional Regime. For $n = 200$, $p = 30$, $k = 5$ under random adversarial corruption, our method achieves high support recovery rates and F1 scores (Appendix 15.3.1). We note that methods without ℓ_1-regularization (RoLR, Trimmed MLE, SVAM) may attain non-zero F1 scores even when the support recovery rate is zero, since F1 accounts for partial overlap with the true support.

Table 4: Linear Regression: F1 scores (mean $\pm$ standard deviation over 100 runs) across varying corruption proportions (ε) for the high-dimensional regime ($n = 200$, $p = 400$, $k = 5$) under the adaptive/strong adversarial contamination model of Chen et al. [2013] (Appendix 14.4). Methods with ℓ_1-regularization (LASSO, Robust LASSO, Ours) tend to perform well on sparse models. Our method and LASSO outperform Robust LASSO as the later restrictively assumes $\mathbf{x}^{(i)} \sim$ sub-Gaussian$(\frac{1}{n}\sigma^2)$ instead of just sub-Gaussian(σ^2) as discussed in Table 2. Numbers smaller than 10^{-4} are reported as 0.

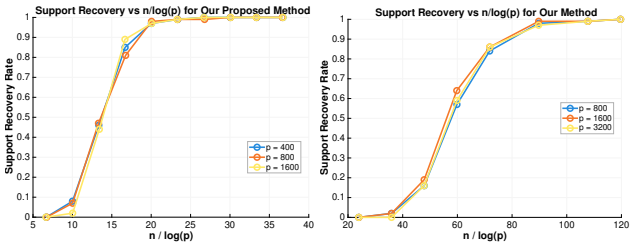
ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.543 ± 0.132	0.182 ± 0.024	0.02 ± 0.011	0.025 ± 0	1 ± 0
0.1	0.535 ± 0.132	0.103 ± 0.044	0.02 ± 0.01	0.025 ± 0	0.998 ± 0.013
0.15	0.549 ± 0.116	0.043 ± 0.036	0.018 ± 0.011	0.025 ± 0	0.975 ± 0.069
0.2	0.542 ± 0.128	0.063 ± 0.039	0.015 ± 0.012	0.025 ± 0	0.831 ± 0.16

Table 5: Logistic Regression: F1 scores (mean $\pm$ standard deviation over 100 runs) across varying corruption proportions (ε) for the high-dimensional regime ($n = 600$, $p = 800$, $k = 5$) under the adaptive/strong adversarial corruption model (Appendix 15.4). The strong performance of LASSO GLM highlights the importance of ℓ_1-regularization for sparse models in high dimensions. Our method combines ℓ_1-regularization with robust trimming, achieving high F1 scores across all corruption levels. Methods without regularization (RoLR, Trimmed MLE, SVAM) are better suited to the low-dimensional setting (Table 18). Numbers smaller than 10^{-4} are reported as 0.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.996 ± 0.018	0 ± 0	0.023 ± 0	0.018 ± 0	0.998 ± 0.013
0.1	0.962 ± 0.052	0 ± 0	0.024 ± 0	0.018 ± 0	0.992 ± 0.026
0.15	0.932 ± 0.066	0 ± 0	0.026 ± 0	0.018 ± 0	0.994 ± 0.024
0.2	0.816 ± 0.101	0 ± 0	0.027 ± 0.001	0.018 ± 0	0.929 ± 0.081

High-Dimensional Regime: We set $n = 600$, $p = 800$, $k = 5$ and follow the same progression of corruption models as above. Under random adversarial corruption (Appendix 15.3), strong adversarial corruption (Table 5 and Appendix 15.4), and heavy-tailed corruption from both log-normal and Student's t-distributions (Appendix 15.5), our proposed method consistently achieves high support recovery rates and F1 scores across all corruption proportions. This demonstrates that the strong empirical performance observed for linear regression carries over to the logistic regression model.

Remark. The F1 scores of RoLR, Trimmed MLE, and SVAM are higher in the low-dimensional case (Table 18) than in the high-dimensional case (Table 20), which is consistent with the fact that these methods do not incorporate ℓ_1-regularization and are therefore better suited to low-dimensional settings. The strong performance of LASSO GLM in Table 5 further highlights the importance of regularization for sparse models in high dimensions. Our method combines ℓ_1-regularization with robust trimming, which together contribute to its strong support recovery performance.



(a) Linear regression (b) Logistic regression

Figure 1: Support recovery probability versus $n/\log(p)$ under strong contamination model for linear regression (Appendix 14.6) and logistic regression (Appendix 15.6). The overlap of support recovery curves for different p provides concrete empirical evidence that the sample complexity scales proportionally with $\log(p)$, in agreement with our theoretical claims.

7.3 LOGARITHMIC SAMPLE COMPLEXITY

Finally, we empirically validate our theoretical claim that the required sample size scales as $n = \Omega(\log p)$. For both linear and logistic regression, we fix $\varepsilon = 0.1$ for adaptive/strong adversary. We vary n and p jointly such that $n/\log(p)$ takes

identical values across $p \in \{400, 800, 1600\}$ (linear regression) and $p \in \{800, 1600, 3200\}$ (logistic regression). As shown in Figure 1, the support recovery curves for different values of p largely overlap when plotted against $n/\log(p)$. This overlap of support recovery curves across different ambient dimensions provides concrete empirical evidence that the required sample size for reliable support recovery scales proportionally with $\log(p)$, in agreement with our theoretical claims. The same behavior is observed under random and heavy-tailed corruption, as shown in Figures 5 and 6 (Appendix 14.6 and 15.6).

7.4 REAL-WORLD EXPERIMENTS

We evaluated our proposed method on two public cancer datasets: TCGA and NCI-60.

TCGA: The Cancer Genome Atlas (TCGA) is a public cancer genomics resource from NCI/NHGRI. We used a gene-expression matrix with dimension 2360 samples $\times$ 187 genes, and the response was the clinical stage variable. Our proposed method identified the support as `FAM107A` and `FOXP1`. These genes provide a compact cancer-stage association signature as verified by independent biology literature:

1. `FAM107A`: Zhang et al. [2023] reported that FAM107A acts as a tumor suppressor in esophageal squamous cell carcinoma, with reduced expression associated with poorer prognosis and more aggressive clinicopathological features.
2. `FOXP1`: Wlodarska et al. [2005] reported that FOXP1 is highly expressed in a subset of diffuse large B-cell lymphoma and is recurrently targeted by genomic aberrations. Xiao et al. [2016] further evaluated decreased FOXP1 protein expression across multiple tumor types in a systematic review and meta-analysis.

NCI-60: It is a public cancer cell-line pharmacogenomics panel from the U.S. National Cancer Institute Developmental Therapeutics Program. We used a gene-expression matrix with dimension 60 cell lines $\times$ 2000 genes, and the response was the Methotrexate drug activity z-score [Luna et al., 2016].

Our proposed method identified the support as `TRAP1`, `HNRNPA1L2`, `CCT7`, `PLK1`, `E2F4`, `RPSA`, `EIF4A1`, and `NOP2`. These genes form a compact pharmacogenomic signature for drug-response modeling and biomarker hypothesis generation, as discussed below:

- `TRAP1`: Matassa et al. [2018] reviewed the role of TRAP1 in cancer metabolism, including its context-dependent role as an oncogene or tumor suppressor.
- `HNRNPA1L2`: Lee et al. [2024] reported that H3.3-G34W in giant cell tumor of bone functionally aligns with the exon choice repressor hnRNPA1L2, supporting a cancer-relevant RNA-processing role.
- `CCT7`: Huang et al. [2022] reported that overexpression of CCT7 has diagnostic and prognostic value in hepatocellular carcinoma.
- `PLK1`: Liu et al. [2017] reviewed PLK1 as a potential therapeutic target in cancer.
- `E2F4`: Zheng et al. [2021] reported that E2F4 is an indicator of poor prognosis and is related to immune infiltration in hepatocellular carcinoma.
- `RPSA`: Limone et al. [2023] reviewed the role of the PrPC-RPSA interaction in cancer biology; RPSA is also known as the 37/67 kDa laminin receptor.
- `EIF4A1`: Zhao et al. [2019] reported that aberrant PDCD4/eIF4A1 signaling predicts postoperative recurrence in early-stage oral squamous cell carcinoma.
- `NOP2`: Yang et al. [2023] reported a functional role of the m5C RNA methyltransferase NOP2 in high-grade serous ovarian cancer.

8 CONCLUSION & FUTURE WORK

We studied robust feature selection in sparse generalized linear models under the strongest adversarial contamination setting, where an adaptive adversary corrupts both features and labels in the high-dimensional regime. Our method combines ℓ_1-regularization with trimmed maximum likelihood estimation, yielding a non-convex relaxation of an NP-hard problem. We show that any KKT point achieves exact support recovery under mutual incoherence and satisfies sharp ℓ_∞ error bounds tailored to feature selection. Empirically, we validate robustness under random, strong adaptive, and heavy-tailed contamination, and observe support recovery behavior consistent with the predicted logarithmic scaling and also verify our proposed method on real-world datasets.

Our estimator attains near-minimax robustness while preserving logarithmic sample complexity in the ambient dimension, matching the optimal dependence on dimensionality even in the absence of adversarial corruption. The framework applies to both linear and logit models and extends to heavy-tailed covariates under bounded fourth-moment assumptions.

The residual $\log(1/\epsilon)$ factor in the rate $O(\epsilon \log(1/\epsilon))$ appears intrinsic to computationally efficient estimators under the strong adversary. Statistical-query lower bounds for robust Gaussian mean estimation exhibit an analogous $\Omega(\epsilon\sqrt{\log(1/\epsilon)})$ barrier [Diakonikolas et al., 2017], giving evidence that the gap to the optimal $O(\epsilon)$ rate of the (NP-hard) estimator of Gao [2020] may be unavoidable without further assumptions, which remains an open problem. Future work includes extending these guarantees to broader nonlinear models and further tightening the robustness–computation trade-offs.

References

Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In *Advances in Neural Information Processing Systems*, 2022.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212. PMLR, 2017.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Eva Cantoni and Elvezio Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International conference on machine learning*, pages 774–782. PMLR, 2013.

Arnak Dalalyan and Yin Chen. Fused sparsity and robust estimation for linear models with unknown variance. *Advances in Neural Information Processing Systems*, 25, 2012.

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

Hadi Daneshmand, Manuel Gomez-Rodriguez, Le Song, and Bernhard Schoelkopf. Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-Thresholding Algorithm. In *International Conference on Machine Learning*, pages 793–801, 2014.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Ilias Diakonikolas, Daniel Kane, Ankit Pensia, and Thanasis Pittas. Near-optimal algorithms for gaussians with huber contamination: Mean estimation and linear regression. *Advances in Neural Information Processing Systems*, 36: 43384–43422, 2023a.

Ilias Diakonikolas, Sushrut Karmalkar, Jong Ho Park, and Christos Tzamos. Distribution-independent regression for generalized linear models with oblivious corruptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5453–5475. PMLR, 2023b.

Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Tommaso d'Orsi, Chih-Hung Liu, Rajai Nasser, Gleb Novikov, David Steurer, and Stefan Tiegel. Consistent estimation for pca and sparse regression with oblivious outliers. *Advances in Neural Information Processing Systems*, 34:25427–25438, 2021.

Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. *Advances in neural information processing systems*, 27, 2014.

Jerome H Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33: 1–22, 2010.

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

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Xinghua Huang, Huaxiang Wang, Fengfeng Xu, Lizhi Lv, Ruling Wang, Bin Jiang, Tingting Liu, Huanzhang Hu, and Yi Jiang. Overexpression of chaperonin containing TCP1 subunit 7 has diagnostic and prognostic value for hepatocellular carcinoma. *Aging*, 14(2):747–769, 2022. doi: 10.18632/aging.203809.

Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33:15689–15701, 2020.

Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34:4475–4488, 2021.

Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. *arXiv preprint arXiv:1809.08055*, 2018.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

Eunbi Lee, Yoon Jung Park, and Anders M. Lindroth. H3.3-G34W in giant cell tumor of bone functionally aligns with the exon choice repressor hnRNPA1L2. *Cancer Gene Therapy*, 31:1177–1185, 2024. doi: 10.1038/s41417-024-00776-6.

Adriana Limone, Valentina Maggisano, Daniela Sarnataro, and Stefania Bulotta. Emerging roles of the cellular prion protein (PrPC) and 37/67 kDa laminin receptor (RPSA) interaction in cancer biology. *Cellular and Molecular Life Sciences*, 80:207, 2023. doi: 10.1007/s00018-023-04844-2.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pages 411–421. PMLR, 2020.

Zhixian Liu, Qingrong Sun, and Xiaosheng Wang. PLK1, a potential target for cancer therapy. *Translational Oncology*, 10(1):22–32, 2017. doi: 10.1016/j.tranon.2016.10.003.

Augustin Luna, Vinodh N. Rajapakse, Felipe G. Sousa, Jianjiong Gao, Nikolaus Schultz, Sandeep Varma, William C. Reinhold, Chris Sander, and Yves Pommier. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. *Bioinformatics*, 32(8):1272–1274, 2016. doi: 10.1093/bioinformatics/btv701.

Domenico S. Matassa, Ilaria Agliarulo, Rosanna Avolio, Michele Landriscina, and Franca Esposito. TRAP1 regulation of cancer metabolism: dual role as oncogene or tumor suppressor. *Genes*, 9(4):195, 2018. doi: 10.3390/genes9040195.

Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. *Advances in Neural Information Processing Systems*, 27, 2014.

Stanislav Minsker, Mohamed Ndaoud, and Lang Wang. Robust and tuning-free sparse linear regression via square-root slope. *SIAM Journal on Mathematics of Data Science*, 6(2):428–453, 2024.

Bhaskar Mukhoty, Debojyoti Dey, and Purushottam Kar. Corruption-tolerant algorithms for generalized linear models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9243–9250, 2023.

Nasser Nasrabadi, Trac Tran, and Nam Nguyen. Robust lasso with missing and grossly corrupted observations. *Advances in Neural Information Processing Systems*, 24, 2011.

Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association*, 120(550):1002–1013, 2025.

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

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(none):935 – 980, 2011. doi: 10.1214/11-EJS631.

Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

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

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

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

Iwona Wlodarska, E Veyt, P De Paepe, Peter Vandenberghe, P Nooijen, Ivan Theate, Lucienne Michaux, Xavier Sagaert, Peter Marynen, Anne Hagemeijer, et al. Foxp1, a gene highly expressed in a subset of diffuse large b-cell lymphoma, is recurrently targeted by genomic aberrations. *Leukemia*, 19(8):1299–1305, 2005.

John Wright and Yi Ma. Dense error correction via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

Jian Xiao, Bixiu He, Yong Zou, Xi Chen, Xiaoxiao Lu, Mingxuan Xie, Wei Li, Shuya He, Shaojin You, and Qiong Chen. Prognostic value of decreased foxp1 protein expression in various tumors: a systematic review and meta-analysis. *Scientific reports*, 6(1):30437, 2016.

Eunho Yang, Ambuj Tewari, and Pradeep Ravikumar. On robust estimation of high dimensional generalized linear models. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

Shuang Yang, Di Zhou, Chao Zhang, Jing Xiang, and Xiaojun Xi. Function of m5C RNA methyltransferase NOP2 in high-grade serous ovarian cancer. *Cancer Biology & Therapy*, 24(1):2263921, 2023. doi: 10.1080/15384047.2023.2263921.

Nikos Zarifis, Puqian Wang, Ilias Diakonikolas, and Jelena Diakonikolas. Robustly learning monotone generalized linear models via data augmentation. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pages 5921–5990. PMLR, 2025.

Fuzhen Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

Jiale Zhang, Shouyin Di, Mingyang Li, Yanxin Dong, Shun Xie, Taiqian Gong, Peizhen Hu, Qingge Jia, and Boshi Fan. Fam107a as a tumor suppressor in esophageal squamous carcinoma inhibits growth and metastasis. *Pathology-Research and Practice*, 252:154945, 2023.

Mengxiang Zhao, Liang Ding, Yan Yang, Sheng Chen, Nisha Zhu, Yong Fu, Yanhong Ni, and Zhiyong Wang. Aberrant expression of PDCD4/eIF4A1 signal predicts postoperative recurrence for early-stage oral squamous cell carcinoma. *Cancer Management and Research*, 11: 9553–9562, 2019. doi: 10.2147/CMAR.S223273.

Qiuxian Zheng, Qiang Fu, Jia Xu, Xinyu Gu, Haibo Zhou, and Chen Zhi. Transcription factor E2F4 is an indicator of poor prognosis and is related to immune infiltration in hepatocellular carcinoma. *Journal of Cancer*, 12(6): 1792–1803, 2021. doi: 10.7150/jca.51616.

Supplementary Material: Provable Guarantees For Robust Feature Selection in Sparse Linear Models in High-Dimensions

Deepak Maurya[1] **Adarsh Barik**[2] **Jean Honorio**[1,3]

[1]Department of Computer Science, Purdue University, USA
[2]Department of Computer Science and Engineering, Indian Institute of Technology Delhi, India
[3]School of Computing and Information Systems, The University of Melbourne, ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Australia

9 PROOFS RELATED TO MEDIAN-OF-MEANS BASED FILTERING ALGORITHM

Theorem 9.1 (Robust ℓ_∞ Mean Estimation via Median-of-Means)**.** *Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be i.i.d. with coordinate population means μ_j and finite coordinate population variances $\sigma_j^2 < \infty$ for $j = 1, \ldots, p$. Fix $\delta \in (0, 1/3)$ and choose the number of blocks $B \geq 12 \log\left(\frac{2p}{\delta}\right), q := \lfloor n/B \rfloor$. Partition $\{1, \ldots, n\}$ into B disjoint blocks $G^{(1)}, \ldots, G^{(B)}$ of size q, form block means coordinate-wise $\bar{X}_j^{(b)} := \frac{1}{|G^{(b)}|} \sum_{i \in G^{(b)}} \mathbf{x}_j^{(i)}$, and define the* coordinate-wise Median-of-Means (MoM) *estimator*

$$\widehat{\mu}_{1j} \;:=\; \text{median}\big\{\bar{X}_j^{(1)}, \ldots, \bar{X}_j^{(B)}\big\}, \qquad j = 1, \ldots, p.$$

Then there exists an absolute constant $C > 0$ such that, with probability at least $1 - \delta/2$

$$\|\widehat{\mu}_1 - \mu_1\|_\infty \;\leq\; C\, \sigma_{\max} \sqrt{\frac{\log(2p/\delta)}{n}}, \qquad \sigma_{\max} := \max_j \sigma_j.$$

Proof. We start by applying the Chebyshev inequality for coordinate j for general block b:

$$\Pr\left[|\bar{X}_j^{(b)} - \mu_{1j}| > t_j\right] \leq \frac{\text{Var}(\bar{X}_j^{(b)})}{t_j^2} \;=\; \frac{\sigma_j^2}{q\, t_j^2} \;=:\; r$$

Further, we define a "bad" block as

$$I_j^{(b)} \;:=\; \mathbf{1}\{|\bar{X}_j^{(b)} - \mu_{1j}| > t_j\}, \quad S_j \;:=\; \sum_{b=1}^{B} I_j^{(b)}$$

The median of means estimate $\widehat{\mu}_j$ should be a good estimate if at least $B/2$ blocks estimate the population mean approximately. This can be formalized as

$$\Pr\left[|\widehat{\mu}_{1j} - \mu_{1j}| > t_j\right] \leq \Pr\left[S_j \geq \frac{B}{2}\right].$$

The indicator variable $I^{(1)}, I^{(2)}, \ldots, I^{(B)}$ are independent Bernoulli random variables with mean at most $r = \sigma_j^2/(qt_j^2)$. Hence, $\mathbb{E}[S] = Br$. We use the multiplicative version of Chernoff bound

$$\Pr\left[S \geq (1+\delta)\mathbb{E}[S]\right] \leq \exp\left(-\frac{\delta^2}{2+\delta}\mathbb{E}[S]\right)$$

We substitute $\delta = \frac{1}{2r} - 1$ in the above equation to arrive at:

$$\Pr\left[S \geq \frac{B}{2}\right] = \Pr\left[S \geq (1+\delta)\mathbb{E}[S]\right] \leq \exp\left(-\frac{\delta^2}{2+\delta} \cdot \mathbb{E}[S]\right) \leq \exp\left(-\frac{1}{3} \cdot \frac{B}{4}\right) = \exp\left(-\frac{B}{12}\right)$$

In this last step, we have assumed $\delta \geq 1$ and hence $\frac{\delta^2}{2+\delta} \geq \frac{1}{3}$. We ensure $\delta \geq 1$ by controlling $r \leq \frac{1}{4}$ through block size q as shown later. We can claim:

$$\Pr\left[|\widehat{\mu}_{1j} - \mu_{1j}| > t_j\right] \leq \Pr\left[S_j \geq \frac{B}{2}\right] \leq \exp\left(-\frac{B}{12}\right)$$

Further taking union bound along p variables, we can claim:

$$\Pr\left[\|\widehat{\mu}_1 - \mu_1\|_\infty > t\right] \leq \exp\left(-\frac{B}{12} + \log(p)\right)$$

Further, we choose $B = 12\log(\frac{2p}{\delta})$

$$\Pr\left[\|\widehat{\mu}_1 - \mu_1\|_\infty > t\right] \leq \exp\left(-\log\left(\frac{2p}{\delta}\right) + \log(p)\right) = \frac{\delta}{2}$$

Finally the condition $r \leq \frac{1}{4}$ leads to

$$\frac{\sigma^2}{qt^2} \leq \frac{1}{4} \implies t \geq \frac{2\sigma}{\sqrt{q}} = 2\sigma\sqrt{\frac{B}{n}} = 2\sigma\sqrt{\frac{12\log\left(\frac{2p}{\delta}\right)}{n}}$$

Hence, with high probability of at least $1 - \delta/2$, we claim

$$\|\widehat{\mu}_1 - \mu_1\|_\infty \leq 2\sigma\sqrt{\frac{12\log\left(\frac{2p}{\delta}\right)}{n}}$$

which completes the proof.

□

Corollary 9.2 (Robust ℓ_∞ Second Order Moment Estimation via Median-of-Means)**.** *Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be i.i.d. with coordinate-wise second order moment as μ_{2j} and finite coordinate-wise fourth-order moment $\mu_{4j} < \infty$ for $j = 1, \ldots, p$. Fix $\delta \in (0, 1/3)$ and choose the number of blocks $B \geq 12\log\left(\frac{2p}{\delta}\right)$, $q := \lfloor n/B \rfloor$. Let $\mathbf{z}_j^{(i)} = \left(\mathbf{x}_j^{(i)}\right)^2$. Partition $\{1, \ldots, n\}$ into B disjoint blocks $G^{(1)}, \ldots, G^{(B)}$ of size q, form block means coordinate-wise $\bar{Z}_j^{(b)} := \frac{1}{|G^{(b)}|}\sum_{i \in G^{(b)}} \mathbf{z}_j^{(i)}$, and define the* coordinate-wise Median-of-Means (MoM) *estimator*

$$\widehat{\mu}_{2j} := \operatorname{median}\left\{\bar{Z}_j^{(1)}, \ldots, \bar{Z}_j^{(B)}\right\}, \qquad j = 1, \ldots, p.$$

Then there exists an absolute constant $C > 0$ such that, with probability at least $1 - \delta/2$

$$\|\widehat{\mu}_2 - \mu_2\|_\infty \;\leq\; C\sqrt{\mu_{4\max}}\sqrt{\frac{\log(2p/\delta)}{n}}, \qquad \mu_{4\max} := \max_j \mu_{4j}.$$

Proof. The proof is very similar to the proof of Theorem 9.1. Substitute $\mathbf{x}^{(i)}$ with $\mathbf{z}^{(i)}$ in the proof of Theorem 9.1. □

Lemma 9.3 (Robust ℓ_∞ Variance Estimation via Median-of-Means)**.** *Let the variance estimate, $\widehat{\sigma^2} := \widehat{\mu}_2 - \hat{\mu}_1^2$, where $\widehat{\mu}_1$ and $\hat{\mu}_2$ are the estimates of first order and second order moment as defined in Theorem 9.1 and Corollary 9.2. Then we claim the following with high probability of $1 - \delta$*

$$\left\|\widehat{\sigma^2} - \sigma^2\right\|_\infty \leq \mathcal{O}\left(C\sqrt{\frac{\log(2p/\delta)}{n}}\right)$$

where C is a constant that depends on the maximum value of μ_{4j}, σ_j and $|\mu_j|$ for $j \in \{1, 2, \ldots, p\}$.

Proof. As per the definition of variance estimate, we have $\widehat{\sigma^2} = \widehat{\mu}_2 - \hat{\mu}_1^2$, where $\widehat{\mu}_2$ and $\hat{\mu}_1$ are defined in Corollary 9.2 and Theorem 9.1 respectively.

By using the definition of variance we know,

$$\begin{aligned}\widehat{\sigma^2}_j - \sigma_j^2 &= \widehat{\mu}_{2j} - \mathbb{E}\left[\left(\mathbf{x}_j^{(i)}\right)^2\right] - \left(\hat{\mu}_{1j}^2 - \left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)^2\right) \\ &= \widehat{\mu}_{2j} - \mathbb{E}\left[\left(\mathbf{x}_j^{(i)}\right)^2\right] - \left(\hat{\mu}_{1j} - \left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)\right)\left(\hat{\mu}_{1j} + \left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)\right)\end{aligned}$$

Using norm inequalities, we can claim:

$$\begin{aligned}|\widehat{\sigma^2}_j - \sigma_j^2| &\le \left|\widehat{\mu}_{2j} - \mathbb{E}\left[\left(\mathbf{x}_j^{(i)}\right)^2\right]\right| + \left|\left(\hat{\mu}_{1j} - \left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)\right)\right|\left|\left(\hat{\mu}_{1j} + \left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)\right)\right| \\ &\overset{(i)}{\le} \mathcal{O}\left(\sqrt{\mu_{4j}}\sqrt{\frac{\log(2p/\delta)}{n}}\right) + \mathcal{O}\left(\sigma_j\sqrt{\frac{\log(2p/\delta)}{n}}\right)\left(|\hat{\mu}_{1j}| + \left|\left(\mathbb{E}\left[\mathbf{x}_j^{(i)}\right]\right)\right|\right) \\ &\overset{(ii)}{\le} \mathcal{O}\left(\sqrt{\mu_{4j}}\sqrt{\frac{\log(2p/\delta)}{n}}\right) + \mathcal{O}\left(\sigma_j\sqrt{\frac{\log(2p/\delta)}{n}}\right)\left(2|\mu_j| + \mathcal{O}\left(\sigma_j\sqrt{\frac{\log(2p/\delta)}{n}}\right)\right) \\ &\le \mathcal{O}\left(C_j\sqrt{\frac{\log(2p/\delta)}{n}}\right)\end{aligned}$$

where C_j is a constant that depends on μ_{4j}, σ_j and $|\mu_j|$. In step (i), we have used Corollary 9.2 for the first term and Theorem 9.1 for second term. In step (ii), we use Theorem 9.1 again.

Further, we can claim the following with probability at least $1-\delta$:

$$\left\|\widehat{\sigma^2} - \sigma^2\right\|_\infty \le \mathcal{O}\left(C\sqrt{\frac{\log(2p/\delta)}{n}}\right)$$

where C is a constant that depends on the maximum value of μ_{4j}, σ_j and $|\mu_j|$ across all $j \in \{1, 2, \ldots, p\}$.

□

Lemma 9.4 (Clean-sample preservation under MoM-standardized filtering)**.** *Let an unknown set $G \subset [n]$ with $|G| = (1-\varepsilon)n$ exist such that $\{\mathbf{x}^{(i)}\}_{i\in G}$ are i.i.d. with mean $\mu \in \mathbb{R}^p$ follow sub-Gaussian distribution with variances $\sigma_j^2 > 0$ and bounded fourth moments $\mathbb{E}[(\mathbf{x}_j^{(i)} - \mu_j)^4] \le C_4\sigma_j^4$. The remaining εn samples are arbitrary.*

Let $\widehat{\mu}_{1j}$ and $\widehat{\sigma}_j$ be the coordinate-wise median-of-means estimators constructed using $B = \mathcal{O}(\log(p/\delta))$ blocks, and define

$$U_j^{(i)} := \frac{\mathbf{x}_j^{(i)} - \widehat{\mu}_{1j}}{\widehat{\sigma}_j}, \qquad T := \mathcal{O}\left(\sqrt{\log\frac{p}{\varepsilon}}\right).$$

Assume that, with probability at least $1-\delta$, the estimators satisfy

$$\max_{j\le p}|\widehat{\mu}_{1j} - \mu_j| \le a\,\sigma_j, \qquad (1-\eta)\sigma_j \le \widehat{\sigma}_j \le (1+\eta)\sigma_j, \qquad \forall j \in [p], \tag{5}$$

for constants $a, \eta \in (0, 1/2)$. Then there exist universal constants $c_0, c_1 > 0$ such that, with probability at least $1 - \delta - \exp(-c_1\varepsilon n)$, the filtering criteria of pruning sample i if $\|U^{(i)}\|_\infty > T$ removes at most $c_0\,\varepsilon n$ clean samples.

Proof. Let $\mathcal{E}$ denote the event on which (5) holds, implying mean and variance are close to true value. This is already proved in Theorem 9.1 and Lemma 9.3. Hence, $\Pr[\mathcal{E}] \ge 1 - \delta$.

Fix a clean index $i \in G$ and a coordinate $j \in [p]$. Under the event $\mathcal{E}$,

$$|U_j^{(i)}| = \left|\frac{\mathbf{x}_j^{(i)} - \mu_j}{\widehat{\sigma}_j} + \frac{\mu_j - \widehat{\mu}_{1j}}{\widehat{\sigma}_j}\right| \le \frac{|\mathbf{x}_j^{(i)} - \mu_j|}{(1-\eta)\sigma_j} + \frac{a}{1-\eta}.$$

We know the first term $Z_j^{(i)} := \frac{\mathbf{x}_j^{(i)} - \mu_j}{\sigma_j}$ is standard sub-Gaussian. Therefore, for any $t > a/(1-\eta)$,

$$\mathbb{P}\Big(|U_j^{(i)}| > t \,\Big|\, \mathcal{E}\Big) \le \mathbb{P}\Big(|Z_j^{(i)}| > (1-\eta)t - a\Big) \le 2\exp\left(-\frac{((1-\eta)t-a)^2}{2}\right).$$

Applying a union bound over $j \in [p]$ yields

$$\mathbb{P}\Big(\|U^{(i)}\|_\infty > t \,\Big|\, \mathcal{E}\Big) \le 2p\exp\left(-\frac{((1-\eta)t-a)^2}{2}\right).$$

Setting $t = T = \mathcal{O}\left(\sqrt{\log(p/\varepsilon)}\right)$ and choosing a, η sufficiently small, we obtain

$$\mathbb{P}\Big(\|U^{(i)}\|_\infty > T \,\Big|\, \mathcal{E}\Big) \le c\varepsilon$$

for a universal constant $c \in (0,1)$. Let the pruned samples be defined as:

$$R_i := \mathbf{1}\{\|U^{(i)}\|_\infty > T\}, \qquad R_G := \sum_{i \in G} R_i.$$

Conditioned on event $\mathcal{E}$, the variables $\{R_i\}_{i\in G}$ are independent with $\mathbb{E}[R_G \mid \mathcal{E}] \le c\varepsilon|G| = c\varepsilon(1-\varepsilon)n \le c\varepsilon n$.

We use multiplicative Chernoff bound:

$$\Pr\left[R_G \ge (1+\delta)\mathbb{E}[R_G] \,|\, \mathcal{E}\right] \le \exp\left(-\frac{\delta^2}{2+\delta}\mathbb{E}[R_G \,|\, \mathcal{E}]\right).$$

We select $\delta = 1$ for simplicity to arrive at:

$$\Pr\left[R_G \ge 2c\varepsilon|G| \;|\mathcal{E}\right] \le \exp(-c_1\varepsilon|G|) \le \exp(-c_1\varepsilon n).$$

Unconditioning completes the proof:

$$\Pr\left[R_G \ge 2c\varepsilon|G|\right] \le \Pr[\mathcal{E}^c] + \Pr\left[R_G \ge 2c\varepsilon|G| \;|\mathcal{E}\right] \le \delta + \exp(-c_1\varepsilon n).$$

□

10 FORMAL PROOFS AND THEOREM STATEMENTS

10.1 PROOF OF LEMMA 4.2 TO CLAIM NON-CONVEXITY OF TRIMMED MLE OBJECTIVE FUNCTION

Lemma 4.2: If $\exists i,\ \mathbf{x}^{(i)} \neq 0$ for $i \in [n]$, then $l(\mathbf{w}, \mathbf{t})$ in Eq. (2) is non-convex in $(\mathbf{w}, \mathbf{t})$.

Proof. We derive the Hessian matrix for the loss function in Eq. (2) with respect to $(\mathbf{w}, \mathbf{t})$:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 l(\tilde{\mathbf{w}},\tilde{\mathbf{t}})}{\partial \mathbf{w}^2} & \frac{\partial l(\tilde{\mathbf{w}},\tilde{\mathbf{t}})}{\partial \mathbf{w}\partial \mathbf{t}} \\ \frac{\partial l(\tilde{\mathbf{w}},\tilde{\mathbf{t}})}{\partial \mathbf{t}\partial \mathbf{w}} & \frac{\partial^2 l(\tilde{\mathbf{w}},\tilde{\mathbf{t}})}{\partial \mathbf{t}^2} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{M} \\ \mathbf{M}^\intercal & \mathbf{B} \end{bmatrix},$$

where the matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{M} \in \mathbb{R}^{p\times n}$, and $\mathbf{B} \in \mathbb{R}^{n\times n}$ is given by:

$$\mathbf{A} = \frac{1}{m}\sum_{i=1}^{n} \tilde{\mathbf{t}}_i b''\left(\mathbf{w}_{\mathcal{S}}^{\star\intercal}\mathbf{x}_{\mathcal{S}}^{(i)}\right)\mathbf{x}_{\mathcal{S}}^{(i)}\mathbf{x}_{\mathcal{S}}^{(i)\intercal}, \quad \mathbf{B} = \mathbf{0}_{n\times n}, \quad \mathbf{M}_{\cdot,i} = \left(b'\left(\mathbf{w}^\intercal\mathbf{x}^{(i)}\right) - y^{(i)}\right)\mathbf{x}^{(i)},$$

where $\mathbf{M}_{\cdot,i}$ denotes the i^{th} column. Also note that $\mathbf{M} \neq \mathbf{0}$, as $\exists i,\ \mathbf{x}^{(i)} \neq 0$ for $i \in [n]$ as stated in the lemma. This is very mild and reasonable assumption because if $\mathbf{x}^{(i)} = \mathbf{0}, \forall i \in [n]$, then the model is unidentifiable even if there are no outliers.

Using Lemma 12.12, we can claim that $\mathbf{A} \succ 0$. Further, we use Schur Complement (page 34 of Zhang [2006]) to claim that $\mathbf{H}$ can be positive definite if and only if:

$$\mathbf{B} - \mathbf{M}^\intercal\mathbf{A}^{-1}\mathbf{M} = \mathbf{0}_{n\times n} - \mathbf{M}^\intercal\mathbf{A}^{-1}\mathbf{M} \succeq 0,$$

which is not possible as $\mathbf{A} \succ 0$. Hence, $\mathbf{H}$ is not positive definite. □

11 KKT CONDITIONS

The Lagrangian can be stated as:

$$L\left(\mathbf{w}, \mathbf{t}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu\right) = l\left(\mathbf{w}, \mathbf{t}\right) + \boldsymbol{\alpha}^{\intercal}(-\mathbf{t}) + \boldsymbol{\beta}^{\intercal}\left(\mathbf{t} - \mathbf{1}_n\right) + \mu\left(-\mathbf{1}^{\intercal}\mathbf{t} + m\right),$$

where $(\mathbf{w}, \mathbf{t})$ are primal variables and $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mu)$ are dual variables. The KKT conditions are given below:

1. The stationarity conditions with respect to $\mathbf{w}$ and $\mathbf{t}$ lead to:

$$\frac{1}{m}\sum_{i=1}^{n} \mathbf{t}_i \left(b'\left(\mathbf{w}^{\intercal}\mathbf{x}^{(i)}\right)\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) + \lambda \mathbf{z} = \mathbf{0}_{p\times 1}, \tag{6}$$

$$\nabla_{\mathbf{t}} l(\mathbf{w}, \mathbf{t}) - \boldsymbol{\alpha} + \boldsymbol{\beta} - \mu \mathbf{1}_n = \mathbf{0}_{n\times 1}, \tag{7}$$

   where $b'(\cdot)$ denotes the derative of function $b(\cdot)$ of GLM defined in Eq. (2), $\nabla_{\mathbf{t}} l(\mathbf{w}, \mathbf{t})$ denotes the partial derivative of the likelihood function with respect to $\mathbf{t}$, and $\mathbf{z}$ denotes the sub-differential of $\|\mathbf{w}\|_1$, whose entries are given by:

$$\mathbf{z}_i = \begin{cases} \text{sign}(\mathbf{w}_i) & \text{if } \mathbf{w}_i \neq 0, \\ \in [-1, 1] & \text{if } \mathbf{w}_i = 0. \end{cases}$$

   It should be noted that $\nabla_{\mathbf{t}} l(\mathbf{w}, \mathbf{t})$ is $n-$dimensional column vector, whose i^{th} term is given by likelihood at sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$:

$$\nabla_{\mathbf{t}_i} l(\mathbf{w}, \mathbf{t}) = -\log\left(f(y^{(i)}|\mathbf{w}^{\intercal}\mathbf{x}^{(i)})\right) = b\left(\mathbf{w}^{\intercal}\mathbf{x}^{(i)}\right) - y^{(i)}\mathbf{w}^{\intercal}\mathbf{x}^{(i)} - \log\left(c\left(y^{(i)}\right)\right).$$

2. Complementary slackness conditions are as follows:

$$\boldsymbol{\alpha} \odot (-\mathbf{t}) = \mathbf{0}_n, \quad \boldsymbol{\beta} \odot \left(\mathbf{t} - \mathbf{1}_n\right) = \mathbf{0}_n, \quad \mu\left(-\mathbf{t}^{\intercal}\mathbf{1}_n + m\right) = 0. \tag{8}$$

3. Dual feasibility conditions can be stated as:

$$\boldsymbol{\alpha} \geq \mathbf{0}_n, \quad \boldsymbol{\beta} \geq \mathbf{0}_n, \quad \|\mathbf{z}\|_{\infty} \leq 1. \tag{9}$$

4. Primal feasibility condition can be expressed as:

$$0 \leq \mathbf{t}_i \leq 1 \;\; \forall i \in [n], \quad \mathbf{1}^{\intercal}\mathbf{t} = m. \tag{10}$$

Let a non-integer KKT point where any $\mathbf{t}_i$ is not an integer. We claim that for any non-integer KKT point, there exist a KKT integer point.

Lemma 11.1. *Let $(\tilde{\mathbf{t}}, \tilde{\mathbf{w}})$ be a non-integral KKT point of problem (4.1). Then there exists a primal feasible point $(\bar{\mathbf{t}}, \tilde{\mathbf{w}})$ satisfying the primal constraints such that $\bar{\mathbf{t}}_i \in \{0, 1\}$ for all $i \in [n]$ and*

$$\sum_{i=1}^{n} \tilde{\mathbf{t}}_i \, \ell_i(\tilde{\mathbf{w}}) = \sum_{i=1}^{n} \bar{\mathbf{t}}_i \, \ell_i(\tilde{\mathbf{w}}). \tag{11}$$

Proof. Define $P := \{i \in [n] \mid 0 < \tilde{\mathbf{t}}_i < 1\}$. From complementary slackness for $0 < \mathbf{t}_i < 1$, we have

$$\boldsymbol{\alpha}_i = 0, \quad \boldsymbol{\beta}_i = 0, \quad \forall i \in P.$$

The stationarity condition with respect to t_i yields

$$\ell_i(\tilde{\mathbf{w}}) = -\mu, \quad \forall i \in P, \tag{12}$$

where μ is the multiplier associated with $\mathbf{1}^{\top}\mathbf{t} = m$. Hence all indices in P have identical loss values.

Since $\sum_{i=1}^{n} \tilde{\mathbf{t}}_i = m$, and $\tilde{\mathbf{t}}_i \in \{0, 1\}$ for $i \notin P$, it follows that $\sum_{i\in P} \tilde{\mathbf{t}}_i \in \mathbb{Z}$.

Construct $\bar{\mathbf{t}}$ by selecting exactly $\sum_{i\in P} \tilde{\mathbf{t}}_i$ indices in P and setting them to 1, assigning 0 to the remaining indices in P, and keeping $\bar{\mathbf{t}}_i = \tilde{\mathbf{t}}_i$ for $i \notin P$. Then $\sum_{i=1}^{n} \bar{\mathbf{t}}_i = m$, so $(\bar{\mathbf{t}}, \tilde{\mathbf{w}})$ is primal feasible. Using (12), the total weighted loss over P is preserved, which implies (11). □

11.1 STATIONARITY CONDITION

First, we re-define

$$\tilde{\mathbf{w}} := \underset{\mathbf{w}_{\mathcal{S}^c}=\mathbf{0}}{\arg\min}\, l\left(\mathbf{w}, \tilde{\mathbf{t}}\right) + \lambda \left\|\mathbf{w}\right\|_1, \text{ where } \quad \tilde{\mathbf{t}}_i = \begin{cases} 1 & i \in J^{cc}, \\ 0 & i \in \bar{\mathcal{I}}^{c_m}. \end{cases} \tag{13}$$

We define the sets:

$$\begin{aligned} J^{cc} &= \left\{i \mid b_i^* = 1, \tilde{\mathbf{t}}_i = 1, i \in [n]\right\}, \qquad J^{co} = \left\{i \mid b_i^* = 1, \tilde{\mathbf{t}}_i = 0, i \in [n]\right\}, \\ J^{oc} &= \left\{i \mid b_i^* = 0, \tilde{\mathbf{t}}_i = 1, i \in [n]\right\}, \qquad J^{oo} = \left\{i \mid b_i^* = 0, \tilde{\mathbf{t}}_i = 0, i \in [n]\right\}, \\ \hat{J}^c &= \left\{i \mid \tilde{\mathbf{t}}_i = 1, i \in [n]\right\} = J^{cc} \cup J^{oc}, \quad J^{*c} = \{i \mid b_i^* = 1, i \in [n]\} = J^{cc} \cup J^{co}. \end{aligned}$$

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

The first order stationary condition with respect to $\mathbf{w}$, stated in Eq. (6) can be re-stated for $\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}\right)$ and sub-differential $\tilde{\mathbf{z}} \in \partial \left\|\tilde{\mathbf{w}}\right\|$ as

$$\nabla l_{\mathbf{w}}\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\hat{J}^c}\right) + \lambda \tilde{\mathbf{z}} = \mathbf{0}_{p\times 1},$$

where $\nabla l_{\mathbf{w}}\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\mathcal{Q}}\right) = \frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\left(b'(\tilde{\mathbf{w}}^\intercal \mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right)$ and the decision variable $\tilde{\mathbf{t}}_i^Q = 1$ if $i \in Q$ and 0 otherwise.

Adding and subtracting $\nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}\right)$ from the above equation, we arrive at:

$$\nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}\right) - \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) + \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) + \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{oc}}\right) + \lambda \tilde{\mathbf{z}} = \mathbf{0}_{p\times 1}. \tag{14}$$

Further, we use mean value theorem to simplify $\nabla l_{\mathbf{w}}\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}\right) - \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right)$ as shown below:

$$\nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}\right) = \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) + \nabla^2_{\mathbf{w}} l(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{cc}})\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right),$$

where $\boldsymbol{\nu}$ is some point between $\tilde{\mathbf{w}}$ and $\mathbf{w}^\star$. More formally, $\boldsymbol{\nu} = \mathbf{w}^\star + s\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right)$ for some $s \in [0, 1]$. Also note that $\text{support}(\boldsymbol{\nu}) = \mathcal{S}$. Adding and subtracting $\nabla^2_{\mathbf{w}} l(\mathbf{w}^\star, \tilde{\mathbf{t}})$ from the above equation leads to:

$$\nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}\right) - \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) = \nabla^2_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right)\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right) + \mathbf{r}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}), \tag{15}$$

where the remainder $\mathbf{r}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}})$ is defined as:

$$\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}) := \left(\nabla^2_{\mathbf{w}} l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{cc}}\right) - \nabla^2_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right)\right)\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right).$$

Note that $\hat{\mathbf{r}} = \mathbf{0}$ for linear regression as $\nabla^2_{\mathbf{w}} l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{cc}}\right) = 1$ and $\nabla^2_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) = 1$ for linear regression.

Substituting Eq. (15) in Eq. (14), we arrive at:

$$\nabla^2_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right)\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right) + \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{cc}}\right) + \mathbf{r}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}) + \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{oc}}\right) + \lambda \tilde{\mathbf{z}} = \mathbf{0}_{p\times 1}.$$

Further, we separate the above p equations in $\mathcal{S}$ and $\mathcal{S}^c$, and set $\mathbf{w}_{\mathcal{S}^c} = \mathbf{w}^\star_{\mathcal{S}} = \mathbf{0}$ as per definition in Eq. (13) to obtain

$$\begin{bmatrix} \hat{\mathbf{Q}}^{J^{cc}}_{\mathcal{S}\mathcal{S}} & \hat{\mathbf{Q}}^{J^{cc}}_{\mathcal{S}\mathcal{S}^c} \\ \hat{\mathbf{Q}}^{J^{cc}}_{\mathcal{S}^c\mathcal{S}} & \hat{\mathbf{Q}}^{J^{cc}}_{\mathcal{S}^c\mathcal{S}^c} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^\star_{\mathcal{S}} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} \\ \hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}^c} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} \\ \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}^c} \end{bmatrix} + \varepsilon \begin{bmatrix} \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} \\ \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}^c} \end{bmatrix} + \lambda \begin{bmatrix} \tilde{\mathbf{z}}_{\mathcal{S}} \\ \tilde{\mathbf{z}}_{\mathcal{S}^c} \end{bmatrix} = \mathbf{0}_{p\times 1},$$

where we use $\hat{\mathbf{r}}$ to denote $\hat{\mathbf{r}}_{\mathcal{S}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}})$ for the ease of notation, and $\hat{\mathbf{g}}$ and $\hat{\mathbf{Q}}$ are defined as follows:

$$\begin{aligned} \hat{\mathbf{g}}^{\mathcal{J}} &= \nabla_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{\mathcal{J}}\right) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \tilde{\mathbf{t}}_i^{\mathcal{J}}\left(b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}\left(b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) - y^{(i)}\right)\mathbf{x}^{(i)}, \\ \hat{\mathbf{Q}}^{\mathcal{J}} &= \nabla^2_{\mathbf{w}} l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{\mathcal{J}}\right) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \tilde{\mathbf{t}}_i^{\mathcal{J}} b''(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})\mathbf{x}^{(i)}\mathbf{x}^{(i)\intercal} = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} b''(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})\mathbf{x}^{(i)}\mathbf{x}^{(i)\intercal}, \\ \hat{\mathbf{h}}^{\mathcal{J}} &= \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{\mathcal{J}}\right) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \tilde{\mathbf{t}}_i^{\mathcal{J}}\left(b'(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) = \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}}\left(b'(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)}) - y^{(i)}\right)\mathbf{x}^{(i)}. \end{aligned} \tag{16}$$

After some algebraic manipulations, we obtain:

$$\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^{\star}_{\mathcal{S}} = -\hat{\mathbf{Q}}^{J^{cc}-1}_{\mathcal{S}\mathcal{S}} \left(\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} + \lambda \mathbf{z}_{\mathcal{S}} \right), \tag{17}$$

$$\mathbf{z}_{\mathcal{S}^c} = \frac{1}{\lambda} \left(-\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}^c} - \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}^c} - \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}^c} + \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}} \hat{\mathbf{Q}}^{-1}_{\mathcal{S}\mathcal{S}} \left(\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} + \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} + \lambda \mathbf{z}_{\mathcal{S}} \right) \right).$$

Using $\|\mathbf{z}_{\mathcal{S}}\|_\infty \leq 1$ and applying triangle inequality with other basic norm inequality, we arrive at:

$$\|\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^{\star}_{\mathcal{S}}\|_\infty \leq \left\| \hat{\mathbf{Q}}^{J^{cc}-1}_{\mathcal{S}\mathcal{S}} \right\|_\infty \left(\left\| \hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}} \right\|_\infty + \left\| \hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}} \right\|_\infty + \left\| \hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}} \right\|_\infty + \lambda \right) \tag{18}$$

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq \left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}} \hat{\mathbf{Q}}^{-1}_{\mathcal{S}\mathcal{S}} \right\|_\infty + \frac{1}{\lambda} \left(1 + \left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}} \hat{\mathbf{Q}}^{-1}_{\mathcal{S}\mathcal{S}} \right\|_\infty \right) \left(\left\| \hat{\mathbf{g}}^{J^{cc}} \right\|_\infty + \left\| \hat{\mathbf{r}}^{J^{cc}} \right\|_\infty + \left\| \hat{\mathbf{h}}^{J^{oc}} \right\|_\infty \right). \tag{19}$$

Definition 11.2 (Integral KKT point)**.** A KKT point $(\tilde{\mathbf{t}}, \tilde{\mathbf{w}})$ of problem in Eq. (4.1), i.e., a point satisfying the stationarity conditions in Eq. (6), Eq. (7), the complementary slackness conditions Eq. (8), the dual feasibility conditions Eq. (9), and the primal feasibility conditions Eq. (10), is called an *integral* KKT point if $\tilde{\mathbf{t}}_i \in \{0, 1\}$ for all $i \in [n]$. If $\tilde{\mathbf{t}}_i \in (0, 1)$ for some $i \in [n]$, it is called a *non-integral* KKT point.

For brevity, we denote the per-sample loss by $\ell_i(\mathbf{w}) := \nabla_{\mathbf{t}_i} l(\mathbf{w}, \mathbf{t}) = -\frac{1}{m} \log \left(f\left(y^{(i)} \mid \mathbf{w}^\intercal \mathbf{x}^{(i)} \right) \right)$, so that the trimmed objective can be written as $l(\mathbf{w}, \mathbf{t}) = \sum_{i=1}^n \mathbf{t}_i \, \ell_i(\mathbf{w})$. We further let $\mathbf{t}^* \in \{0, 1\}^n$ denote the ground-truth clean-sample indicator, with $\mathbf{t}^*_i = 1$ if sample i is uncorrupted and $\mathbf{t}^*_i = 0$ otherwise; equivalently $J^{*c} = \{i \in [n] \mid \mathbf{t}^*_i = 1\}$ is the set of clean samples, so that $\mathbf{1}^\intercal \mathbf{t}^* = (1 - \varepsilon)n \geq m$.

Lemma 11.3 (Local optimality of integral KKT points)**.** *Let $(\tilde{\mathbf{t}}, \tilde{\mathbf{w}})$ be an integral KKT point of the problem in Eq. (4.1). Then*

$$\sum_{i=1}^n \tilde{\mathbf{t}}_i \, \ell_i(\tilde{\mathbf{w}}) \;\leq\; \sum_{i=1}^n \mathbf{t}^*_i \, \ell_i(\tilde{\mathbf{w}}).$$

Equivalently, since $l(\tilde{\mathbf{w}}, \mathbf{t}) = \sum_{i=1}^n \mathbf{t}_i \, \ell_i(\tilde{\mathbf{w}})$, we have $l(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}) \leq l(\tilde{\mathbf{w}}, \mathbf{t}^*)$.

Proof. Since $(\tilde{\mathbf{t}}, \tilde{\mathbf{w}})$ is integral, let $P := \left\{ i \in [n] \mid \tilde{\mathbf{t}}_i = 1 \right\}$ and $P_c := \left\{ i \in [n] \mid \tilde{\mathbf{t}}_i = 0 \right\} = [n] \setminus P$. The stationarity condition with respect to $\mathbf{t}$ in Eq. (7) reads, component-wise, $\ell_i(\tilde{\mathbf{w}}) = \mu + \boldsymbol{\alpha}_i - \boldsymbol{\beta}_i$. Multiplying by $\tilde{\mathbf{t}}_i$, summing over $i \in [n]$, and using the complementary slackness $\boldsymbol{\alpha}_i \tilde{\mathbf{t}}_i = 0$ together with the primal feasibility $\sum_{i=1}^n \tilde{\mathbf{t}}_i = m$,

$$\sum_{i=1}^n \tilde{\mathbf{t}}_i \, \ell_i(\tilde{\mathbf{w}}) = \mu \, m - \sum_{i \in P} \boldsymbol{\beta}_i. \tag{20}$$

Similarly, multiplying by $\mathbf{t}^*_i$, summing, and using $\sum_{i=1}^n \mathbf{t}^*_i = m$,

$$\sum_{i=1}^n \mathbf{t}^*_i \, \ell_i(\tilde{\mathbf{w}}) = \mu \, m - \sum_{i \in P} \boldsymbol{\beta}_i \mathbf{t}^*_i + \sum_{i \in P_c} \boldsymbol{\alpha}_i \mathbf{t}^*_i, \tag{21}$$

where we used $\boldsymbol{\alpha}_i = 0$ for $i \in P$ (from $\boldsymbol{\alpha}_i \tilde{\mathbf{t}}_i = 0$) and $\boldsymbol{\beta}_i = 0$ for $i \in P_c$ (from $\boldsymbol{\beta}_i (\tilde{\mathbf{t}}_i - 1) = 0$). Since $\mathbf{t}^*_i \in \{0, 1\}$, $\boldsymbol{\beta}_i \geq 0$, and $\boldsymbol{\alpha}_i \geq 0$, we have $\sum_{i \in P} \boldsymbol{\beta}_i \mathbf{t}^*_i \leq \sum_{i \in P} \boldsymbol{\beta}_i$ and $\sum_{i \in P_c} \boldsymbol{\alpha}_i \mathbf{t}^*_i \geq 0$. Combining with Eq. (20)–(21) gives $\sum_{i=1}^n \tilde{\mathbf{t}}_i \, \ell_i(\tilde{\mathbf{w}}) \leq \sum_{i=1}^n \mathbf{t}^*_i \, \ell_i(\tilde{\mathbf{w}})$. □

12 CONSTRUCTION OF PRIMAL VARIABLES AND VERIFYING STRICT DUAL FEASIBILITY

Lemma 12.1 (Strict Dual Feasibility for Linear Regression to show $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$)**.** *If $\lambda = \Omega \left(\zeta \sigma \sigma_e \left(\sqrt{\frac{\log(p)}{n}} + \varepsilon \log(1/\varepsilon) \right) \right)$ and $n = \frac{k^2 \log(p)}{\varepsilon \log(1/\varepsilon)}$, then with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$ we claim $\|\mathbf{z}_{\mathcal{S}^c}\|_\infty < 1$ and hence $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$.*

Proof. We start from Eq. (19) and use $\hat{\mathbf{r}} = 0$ for linear regression

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty + \frac{1}{\lambda}\left(1 + \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty\right)\left(\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty + 0 + \left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty\right)$$

Further, applying Lemma 12.8 to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty$, we arrive at:

$$\begin{aligned}\|\mathbf{z}_{\mathcal{S}^c}\|_\infty &\leq \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty + \frac{1}{\lambda}\left(1 + \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{-1}\right\|_\infty\right)\left(\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty + 0 + \left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty\right) \\ &\leq 1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) + \frac{1}{\lambda}\left(1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)\right)\left(\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty + \left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty\right)\end{aligned}$$

Further, we use Lemma 12.3 to bound $\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty$:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq 1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) + \frac{1}{\lambda}\left(1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)\right)\left(\mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right) + \left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty\right)$$

Next, we use Lemma 12.4 to bound $\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty$:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq 1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) + \frac{1}{\lambda}\left(1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)\right)\mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$$

Further using $\lambda = \frac{\left(1-\frac{\gamma}{2}+\mathcal{O}(\varepsilon\log(1/\varepsilon))\right)}{\frac{\gamma}{4}}\Omega\left(\zeta\sigma\sigma_e\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon)\right)\right) = \Omega\left(\frac{\zeta\sigma\sigma_e}{\gamma}\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon)\right)\right)$. Here, we have used $1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) \leq 1 - \frac{\gamma}{4} \leq 1$ for $\gamma = \Omega(\varepsilon\log(1/\varepsilon))$. Finally, we arrive at:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq 1 - \frac{\gamma}{4} < 1.$$

By using dual norm properties, we can ensure $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$. □

Lemma 12.2 (Parameter Vector Bound for Linear Regression)**.** *If* $\lambda = \Omega\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$ *and* $n = \Omega(\frac{k^2\log(p)}{\varepsilon^2\log^2(1/\varepsilon)})$*, then with probability of at least* $1 - \mathcal{O}\left(\frac{1}{p}\right)$*:*

$$\|\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$$

Proof. We start by applying triangle inequality in Eq. (17):

$$\|\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \left\|\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{J^{cc}-1}\right\|_\infty\left(\left\|\hat{\mathbf{g}}_{\mathcal{S}}^{J^{cc}}\right\|_\infty + 0 + \left\|\hat{\mathbf{h}}_{\mathcal{S}}^{J^{oc}}\right\|_\infty + \lambda\right)$$

where we have used $\hat{\mathbf{r}} = \mathbf{0}$ for linear regression and $\|\mathbf{z}_{\mathcal{S}}\|_\infty \leq 1$. Next, we use Lemma 12.7 to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}^{J^{cc}-1}\right\|_\infty$ as $\mathcal{O}\left(\left\|\mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right\|_\infty\right) = \mathcal{O}(\zeta)$. The terms $\left\|\hat{\mathbf{g}}_{\mathcal{S}}^{J^{cc}}\right\|_\infty$ and $\left\|\hat{\mathbf{h}}_{\mathcal{S}}^{J^{oc}}\right\|_\infty$ can be bounded using Lemma 12.3 and Lemma 12.4 respectively to arrive at:

$$\|\tilde{\mathbf{w}}_{\mathcal{S}} - \mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon) + \zeta\lambda\right).$$

Further using the choice of the regularization parameter $\lambda = \Theta\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$ leads to the desired result.

□

Lemma 12.3 (Gradient bound over selected clean samples for Linear regression)**.** *If* $\lambda > 8\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}$ *and* $n = \Omega\left(\frac{k^2\log(p)}{\varepsilon\log(1/\varepsilon)}\right)$*, then for the case of linear regression:*

$$\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty \leq \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$$

with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$.

Proof. We use the standard approach in Lemma 4.3 of Diakonikolas et al. [2019] or Lemma 8 of Jambulapati et al. [2020] to express $\hat{\mathbf{g}}_j^{J^{cc}}$ as:

$$\hat{\mathbf{g}}_j^{J^{cc}} = \frac{1}{1-\varepsilon}\hat{\mathbf{g}}_j^{J^{*c}} - \frac{\varepsilon}{1-\varepsilon}\hat{\mathbf{g}}_j^{J^{co}}$$

Further we use triangle inequality to arrive at:

$$\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty \leq \left\|\hat{\mathbf{g}}^{J^{*c}}\right\|_\infty + \varepsilon\left\|\hat{\mathbf{g}}^{J^{co}}\right\|_\infty$$

For linear regression, we have $b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) = \mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}$

$$\hat{\mathbf{g}}^{J^{*c}} = \frac{1}{|J^{*c}|}\sum_{i\in J^{*c}}\left(b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) = \frac{1}{|J^{*c}|}\sum_{i\in J^{*c}}\left(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)} - y^{(i)}\right)\mathbf{x}^{(i)} = \frac{1}{n}\sum_{i\in J^{*c}} e^{(i)}\mathbf{x}^{(i)}$$

As $\mathbf{x}^{(i)}$ and $e^{(i)}$ are sub-Gaussian variables, their product is sub-exponential. We use sub-exponential tail bounds to claim

$$\Pr\left[\left\|\frac{1}{n}\sum_{i\in J^{*c}}\frac{e^{(i)}}{\sigma_e}\frac{\mathbf{x}^{(i)}}{\sigma}\right\|_\infty \geq t_1\right] \leq 2p\exp\left(-\frac{nt_1^2}{64}\right), \qquad 0 \leq t_1 \leq 8$$

We select $t_1 = 8\sqrt{\frac{\log(p)}{n}}$, $\|\frac{1}{n}\sum_{i\in J^{*c}} e^{(i)}\mathbf{x}^{(i)}\|_\infty \leq 8\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}$ with probability $1 - \mathcal{O}(\frac{1}{p})$.

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

$$\Pr\left[\left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\frac{e^{(i)}}{\sigma_e}\frac{\mathbf{x}^{(i)}}{\sigma}\right\|_\infty \geq t_2\right] \leq 2p\exp\left(-\frac{n\varepsilon t_2}{64}\right).$$

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 $n(1-\varepsilon)$

$$\Pr\left[\exists J^{cc} \subset \mathcal{J}^{cc}, \left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\frac{e^{(i)}}{\sigma_e}\frac{\mathbf{x}^{(i)}}{\sigma}\right\|_\infty \geq t_2\right] \leq 2p\binom{n}{n\varepsilon}\exp\left(-\frac{n\varepsilon t_2}{64}\right).$$

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

$$\Pr\left[\exists J^{cc} \subset \mathcal{J}^{cc}, \left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\frac{e^{(i)}}{\sigma_e}\frac{\mathbf{x}^{(i)}}{\sigma}\right\|_\infty \geq t_2\right] \leq 2p\exp\left(-\frac{n\varepsilon t_2}{64} + n\varepsilon\log(1/\varepsilon)\right). \tag{22}$$

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

$$\Pr\left[\exists J^{cc} \subset \mathcal{J}^{cc}, \left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\frac{e^{(i)}}{\sigma_e}\frac{\mathbf{x}^{(i)}}{\sigma}\right\|_\infty \geq t_2\right] \leq 2p\exp(-n\varepsilon\log(1/\varepsilon)) \leq \mathcal{O}\left(\frac{1}{p}\right). \tag{23}$$

Further, we use the triangle inequality to arrive at:

$$\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty \leq (\lambda + 128\sigma\sigma_e\varepsilon\log(1/\varepsilon)) = \mathcal{O}\left(\lambda + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right).$$

which gives the desired result.

□

Lemma 12.4 (Gradient bound over selected corrupted samples for linear regression)**.** *If* $n \geq \frac{k\log(p)}{\varepsilon\log(1/\varepsilon)}$*, then with probability at least* $1-\mathcal{O}\left(\frac{1}{p}\right)$ *for the particular case of linear regression, we claim*

$$\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty \leq \mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right),$$

where

$$g(\sigma^2,\varepsilon,p) = \begin{cases} \mathcal{O}\left(\sigma_m^2\log(1/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{sub-Gaussian}(\sigma_a),\ \sigma_m = \max(\sigma,\sigma_a) \\ \mathcal{O}\left(\sigma^2\log(p/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{heavy-tailed with bounded } 4^{th}\textit{-moment} \end{cases} \tag{24}$$

Proof. We want to bound

$$\hat{\mathbf{h}}^{J^{oc}} = \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{oc}}\right) = \frac{1}{|J^{oc}|}\sum_{i\in J^{oc}}\left(b'(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) = \frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)\mathbf{x}^{(i)}$$

For any j^{th} entry, we can apply Cauchy-Schwarz inequality

$$\left|\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)\mathbf{x}_j^{(i)}\right| \leq \left(\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2\right)^{1/2}\left(\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\mathbf{x}_j^{(i)2}\right)^{1/2}$$

To bound the first term in the RHS, we use the optimality of the solution in Lemma 12.6. The second term can be bounded using Lemma 12.5. The overall bound can be simplified as:

$$\begin{aligned}\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty &\leq \sqrt{k(\sigma+\sigma_e)}\mathcal{O}(\sqrt{\log(1/\varepsilon)}) \times \sqrt{g(\sigma^2,\varepsilon,p)} \times \|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_\infty \\ &\leq \mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right).\end{aligned}$$

where $g(\sigma^2,\varepsilon,p)$ is defined in Eq. (29) and we have used $\sqrt{k(\sigma+\sigma_e)}\mathcal{O}(\sqrt{\log(1/\varepsilon)}) \times \sqrt{g(\sigma^2,\varepsilon,p)} \leq \frac{1}{2}$ for sufficiently small value of ε. The term $\|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_\infty$ is bounded using Lemma 12.2.

Further, we can take union bound across p variables to derive the bound for $\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty$.

□

Lemma 12.5. *If* $n \geq \frac{k\log(p)}{\varepsilon\log(1/\varepsilon)}$*, then with probability at least* $1-\mathcal{O}\left(\frac{1}{p}\right)$*, we claim*

$$\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\mathbf{x}_j^{(i)2} \leq g(\sigma^2,\varepsilon,p)$$

where

$$g(\sigma^2,\varepsilon,p) = \begin{cases} \mathcal{O}\left(\sigma_a^2\log(1/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{sub-Gaussian}(\sigma_a) \\ \mathcal{O}\left(\sigma^2\log(p/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{heavy-tailed with bounded } 4^{th}\textit{-moment} \end{cases}$$

Proof. First, we consider if the features are $\mathbf{x}^{(i)}$ are drawn from sub-Gaussian distribution, in which the median of means based filtering is not required. This may happen if the adversary does not corrupt features or if the adversary is inserting sub-Gaussian corruption with variance proxy parameter σ_a.

As $\mathbf{x}^{(i)}$ is sub-Gaussian, its square will be sub-exponential. Formally speaking, $\frac{1}{n\varepsilon}\sum_{iJ^{oc}}\mathbf{x}_j^{(i)2} \sim SE(\frac{4\sqrt{2}\sigma_m}{\sqrt{n\varepsilon}}, \frac{4\sigma_m}{n\varepsilon})$, where SE denotes sub-exponential distribution and $\sigma_m = \max(\sigma,\sigma_a)$. Further, using sub-exponential tail bounds, we claim:

$$\mathbb{P}\left[\left|\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\mathbf{x}_j^{(i)2} - \mathbb{E}\left(\mathbf{x}_j^{(i)2}\right)\right| \geq \delta\right] \leq 2\exp\left(\frac{-n\varepsilon\delta}{64\sigma_m^2}\right).$$

Further, we take union bound across all subsets of size εn and use $\log\binom{n}{n(1-\varepsilon)} = \mathcal{O}(n\varepsilon\log(1/\varepsilon))$ like done in Eq. (22). Hence, we claim $\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\mathbf{x}_j^{(i)2} = \mathcal{O}\left(\sigma_m^2\log(1/\varepsilon)\right)$ with high probability by choosing $\delta = \sigma_m^2 + \log(1/\varepsilon)$ if $n \geq \frac{C\log(p)}{\varepsilon\log(1/\varepsilon)}$.

If the features (in corrupted samples), meaning $\mathbf{x}^{(i)}$ do not follow sub-Gaussian distribution, then we can use Algorithm 1 to filtering samples with ℓ_∞ based threshold. The direct bound will be $\mathcal{O}\left(\sigma_m^2\log(p/\varepsilon)\right)$. Note that this results holds for any heavy tailed distribution with only 4^{th}-bounded moment.

□

Lemma 12.6 (Objective function bound over selected corrupted samples for linear regression)**.** *If* $n = \Omega\left(\frac{k^2\log(p)}{(\varepsilon\log(1/\varepsilon))^2}\right)$, *then with probability at least* $1 - \mathcal{O}\left(\frac{1}{p}\right)$, *we claim the following for the particular case of linear regression:*

$$\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2 \leq k(\sigma+\sigma_e)\mathcal{O}(\log(1/\varepsilon))\left\|\tilde{\mathbf{w}} - \mathbf{w}^\star\right\|_\infty^2.$$

Proof. By using the optimality of the solution, we can claim:

$$\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2 \leq \frac{1}{n\varepsilon}\sum_{i\in J^{co}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2$$

Note that J^{co} denotes the set of clean samples. Hence we substitute $\left(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)} - y^{(i)}\right)^2 = e_i^2$:

$$\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2 = \left(\mathbf{x}^{(i)\intercal}\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right)\right)^2 + e_i^2 + 2e_i\mathbf{x}^{(i)\intercal}\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right)$$

We analyze the sum for each term:

$$\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\left(\mathbf{x}^{(i)\intercal}\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right)\right)^2 \leq \left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\mathbf{x}^{(i)}\mathbf{x}^{(i)\intercal}\right\|_2\left\|\tilde{\mathbf{w}} - \mathbf{w}^\star\right\|_2^2 \leq k\left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\mathbf{x}^{(i)}\mathbf{x}^{(i)\intercal}\right\|_2\left\|\tilde{\mathbf{w}} - \mathbf{w}^\star\right\|_\infty^2$$

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

We proceed in a similar manner for the last term:

$$\frac{1}{n\varepsilon}\sum_{i\in J^{co}}e_i\mathbf{x}^{(i)\intercal}\left(\tilde{\mathbf{w}} - \mathbf{w}^\star\right) \leq k\left\|\frac{1}{n\varepsilon}\sum_{i\in J^{co}}e_i\mathbf{x}^{(i)}\right\|_\infty\left\|\tilde{\mathbf{w}} - \mathbf{w}^\star\right\|_\infty$$

The first term of the above equation can be bound using Eq. (22) as $\mathcal{O}(\sigma\sigma_e\log(1/\varepsilon))$. As e_i is assumed to be sub-Gaussian, e_i^2 is sub-exponential, and we use sub-exponential concentration inequalities for $t > \sigma_e^2$:

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

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

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

Further, we take $t_1 = \frac{2\sigma_e^2\log(1/\varepsilon)}{c}$ and $n \geq \frac{\log(p)}{\varepsilon\log(1/\varepsilon)}$. Using the above bounds, we can claim:

$$\frac{1}{n\varepsilon}\sum_{i\in J^{co}}\left(\tilde{\mathbf{w}}^\intercal\mathbf{x}^{(i)} - y^{(i)}\right)^2 \leq k(\sigma+\sigma_e)\mathcal{O}(\log(1/\varepsilon))\left\|\tilde{\mathbf{w}} - \mathbf{w}^\star\right\|_\infty^2.$$

□

Lemma 12.7. *If $m > Ck\log(p)$ for a sufficiently large constant $C > 0$, $|b''(\cdot)| < \gamma_2 < \infty$, and $0 < C_{\min} \le \Lambda_{\min}\left(\mathbf{Q}^{\star}_{\mathcal{SS}}\right)$, then with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$, the induced ℓ_∞ norm of the inverse restricted Hessian over the clean-selected samples J^{cc} satisfies*

$$\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty \;\le\; \mathcal{O}\left(\left\|\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty\right) \;=\; \mathcal{O}(\zeta),$$

where $\zeta \coloneqq \left\|\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty$ denotes the induced ℓ_∞ norm of the inverse population Hessian on the support.

Proof. We decompose $\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}}$ into two terms with summation over set J^{*c} and J^{co}:

$$\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1} = \left(\frac{1}{1-\varepsilon}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}} - \frac{\varepsilon}{1-\varepsilon}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right)^{-1} = (1-\varepsilon)\sum_{i=0}^{\infty}\left(\varepsilon\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right)^i\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1},$$

which is valid whenever $\varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty < 1$. Taking the induced ℓ_∞ norm and using its sub-multiplicativity, the Neumann series is dominated by a geometric series:

$$\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty \;\le\; (1-\varepsilon)\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\sum_{i=0}^{\infty}\left(\varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty\right)^i = \frac{(1-\varepsilon)\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty}{1-\varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty}.$$

Further, we use Lemma 12.11 to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty$ as $\frac{3}{2}\left\|\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty$ with high probability of $1 - \mathcal{O}\left(\frac{1}{p}\right)$. The other term of $\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty$ can be bounded using ideas similar to Lemma 12.10. Hence, for sufficiently small ε the denominator equals $1 - \mathcal{O}(\varepsilon)$, and we arrive at the desired result. □

12.1 PROOFS RELATED TO MUTUAL INCOHERENCE WHICH HELP TO SHOW STRICT DUAL FEASIBILITY

Lemma 12.8. *If Assumption 5.1 holds and $n = \frac{k^2\log(p)}{\varepsilon\log(1/\varepsilon)}$, then $\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty \le 1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)$ with probability at least $1 - \mathcal{O}(1/p)$.*

Proof. We start by applying the Woodbury matrix identity for $\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}$ to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty$

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

Hence, the overall term simplifies to:

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

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

$$\begin{aligned}\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty \le{}& \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty + \varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{co}}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\\ &+ \left(\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}\right\|_\infty + \varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{co}}\right\|_\infty\right)\sum_{i=1}^{\infty}\left(\varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty\right)^i.\end{aligned}$$

We can bound the first term in the RHS using Lemma 12.9 as $1-\frac{\gamma}{2}$. We can use ideas similar to Lemma 12.3 to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c-1}}\right\|_\infty \leq \frac{3}{2}\left\|\Sigma_{\mathcal{SS}}^{-1}\right\|_\infty$ with high probability of $1-\mathcal{O}(\frac{1}{p})$, if $n=\Omega(k^2\log(p))$.

Similarly, $\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{co}}\right\|_\infty$ can be bounded as $\mathcal{O}(\log(1/\varepsilon))$ using arguments similar to Lemma 12.3. We apply sub-exponential tail bounds and algebraic details are similar to Eq. (23).

The last term involves a geometric series and can be bounded if the common ratio term,

$$\varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c-1}}\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}\right\|_\infty=\mathcal{O}\left(\varepsilon\log(1/\varepsilon)\left\|\Sigma_{\mathcal{SS}}^{-1}\right\|_\infty\right)<1.$$

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

$$\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{cc}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}-1}\right\|_\infty\leq 1-\frac{\gamma}{2}+\mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right),$$

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

Lemma 12.9. *(Mutual Incoherence) If $m>\frac{C\gamma_2^2\sigma^2k^2\log(p)}{\gamma^2}$, then the sample version of mutual incoherence can be upper bounded as shown below with probability at least $1-\mathcal{O}\left(\frac{1}{p}\right)$:*

$$\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\leq 1-\frac{\gamma}{2}.$$

Proof. Let $\mathbf{Q}^\star:=\mathbf{\Sigma}\left(\mathbf{w}^\star\right)$ be the population covariance matrix. Further, by some algebraic manipulation and use of triangle inequality, we can easily state:

$$\begin{aligned}
&\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\leq q_1+q_2+q_3+q_4,\\
&\text{where}\quad q_1:=\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}-\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}-\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty,\qquad q_2:=\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}-\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\right\|_\infty\left\|\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty,\\
&\qquad\quad q_3:=\left\|\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\right\|_\infty\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}-\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty,\qquad q_4:=\left\|\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\mathbf{Q}_{\mathcal{SS}}^{\star-1}\right\|_\infty.
\end{aligned}$$

By using population mutual incoherence specified in Assumption 5.1, we can directly claim $q_4\leq 1-\gamma$.

To upper bound q_1, we substitute $\delta=\sqrt{\frac{\gamma}{6}}$ in Lemma 12.10 and Lemma 12.11 and claim that $q_1\leq\frac{\gamma}{6}$ if $m>\frac{C\gamma_2^2\sigma^2k^2\log(p)}{\gamma}$ for a sufficiently large constant $C>0$. Similarly, we substitute appropriate value of δ in Lemma 12.10 and Lemma 12.11 to obtain $q_2\leq\frac{\gamma}{6}$ and $q_3\leq\frac{\gamma}{6}$. Finally, this helps to verify strict dual fesability:

$$\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}-1}\right\|_\infty\leq q_1+q_2+q_3+q_4\leq\frac{\gamma}{6}+\frac{\gamma}{6}+\frac{\gamma}{6}+1-\gamma=1-\frac{\gamma}{2}.$$

□

Lemma 12.10. *For any $\delta\in\left(0,8k(1+4\gamma_2\sigma^2)\max_{i\in[p]}\mathbf{\Sigma}_{ii}^{\star2}\right)$, we claim:*

$$\mathbb{P}\left[\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}-\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\right\|_\infty\leq\delta\right]\geq 1-4k(p-k)\exp\left(\frac{-m\delta^2}{128k^2(1+4\gamma_2\sigma^2)^2\max_{i\in[p]}\mathbf{\Sigma}_{ii}^{\star2}}\right).$$

Proof. We start by basic norm properties and union bound:

$$\mathbb{P}\left[\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}^{J^{*c}}-\mathbf{Q}_{\mathcal{S}^c\mathcal{S}}^\star\right\|_\infty\geq\delta\right]\leq\mathbb{P}\left[\max_{i\in\mathcal{S}^c,j\in\mathcal{S}}|\hat{\mathbf{Q}}_{ij}-\mathbf{Q}_{ij}^\star|\geq\frac{\delta}{k}\right]\leq(p-k)k\mathbb{P}\left[|\hat{\mathbf{Q}}_{ij}-\mathbf{Q}_{ij}^\star|\geq\frac{\delta}{k}\right].$$

Further, we note $\hat{\mathbf{Q}}$ defined in Eq. (16) consists of summation across clean samples only. Also note that $\hat{\mathbf{Q}}$ can be interpreted as the mean of m covariance matrices computed from m clean samples. Further note that the i^{th} sample used to generate

such covariance matrix, denoted by $\mathbf{z}^{(i)} := \sqrt{b''(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})}\mathbf{x}^{(i)}$. As $b''(\cdot)$ is assumed to be bounded with constant $\gamma_2 < \infty$, we can claim $\frac{\mathbf{z}_j}{\sqrt{\mathbf{\Sigma}^\star_j}}$ is zero-mean sub-Gaussian random variable with variance proxy parameter $k_2\sigma^2$. This helps us to invoke Lemma 1 from Ravikumar et al. [2010] to claim:

$$\mathbb{P}\left[\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}^c\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}^c\mathcal{S}}\right|\right|\right|_\infty \geq \delta\right] \leq 4(p-k)k\exp\left(\frac{-m\delta^2}{128k^2(1+4\gamma_2\sigma^2)^2\max\limits_{i\in[p]}\mathbf{\Sigma}^{\star 2}_{ii}}\right),$$

for all $\delta \in \left(0, 8k(1+4\gamma_2\sigma^2)\max\limits_{i\in[p]}\mathbf{\Sigma}^{\star 2}_{ii}\right)$. □

Lemma 12.11. *If* $m > \frac{Ck\log(p)}{\delta^2}$, $|b''(\cdot)| < \gamma_2 < \infty$, *and* $0 < C_{\min} \leq \Lambda_{\min}\left(\mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right)$, *then*

$$\mathbb{P}\left[\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}-1}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_\infty \leq \delta\right] \geq 1 - \mathcal{O}\left(\frac{1}{p}\right).$$

Proof. Using some basic algebraic manipulations and sub-multiplicative property of norms, we obtain:

$$\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}-1}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_\infty \leq \sqrt{k}\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \left|\left|\left|\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \left|\left|\left|\mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2.$$

Further, we bound each of the terms in the RHS of the above equation. Starting from the third term, we can directly claim $\left|\left|\left|\mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \leq \frac{1}{C_{\min}}$ as $C_{\min} \leq \Lambda_{\min}\left(\mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right)$. To upper bound the first term, we use the Courant-Fischer formula along with basic infimum properties to claim:

$$\Lambda_{\min}\left(\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}}\right) \geq \Lambda_{\min}\left(\mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right) - \left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \geq C_{\min} - \left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2. \tag{27}$$

Note that the sample covariance matrix $\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}}$ is constructed from sub-Gaussian vectors as discussed in the proof of Lemma 12.10. Hence, to upper bound $\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2$, we use Proposition 2.1 from Vershynin [2012]:

$$\mathbb{P}\left[\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \geq \epsilon\right] \leq 2\exp\left(-c\epsilon^2 m + k\right), \tag{28}$$

for any $\epsilon > 0$ and some constant $c > 0$. We can claim $\Lambda_{\min}\left(\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}}\right) > \frac{C_{\min}}{2}$ by substituting $\epsilon = \frac{C_{\min}}{2}$ in the above equation, if $m > Ck\log(p)$ for a sufficiently large constant $C > 0$. Hence, we can claim:

$$\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}-1}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right|\right|\right|_\infty \leq \sqrt{k}\frac{2}{C_{\min}}\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2 \frac{1}{C_{\min}}.$$

Further, we substitute $\epsilon = \frac{\delta C^2_{\min}}{2\sqrt{k}}$ in Eq. (28) to bound $\left|\left|\left|\hat{\mathbf{Q}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} - \mathbf{Q}^\star_{\mathcal{S}\mathcal{S}}\right|\right|\right|_2$ and claim that it holds with high probability if $m > \frac{Ck\log(p)}{\delta^2}$ for a sufficiently large constant $C > 0$. This completes the proof for the claimed result. □

12.2 PROOFS RELATED TO MINIMUM AND MAXIMUM EIGENVALUE

Lemma 12.12. *For* $n > Ck\log(p)/(\varepsilon\log(1/\varepsilon))$, *the minimum eigenvalue of sample covariance can be lower bounded as shown below with probability at least* $1 - \mathcal{O}\left(\frac{1}{p}\right)$*:*

$$\Lambda_{\min}\left(\frac{1}{m}\sum_{i\in J^{cc}}\tilde{\mathbf{t}}_i b''\left(\mathbf{w}^{\star\intercal}_{\mathcal{S}}\mathbf{x}^{(i)}_{\mathcal{S}}\right)\mathbf{x}^{(i)}_{\mathcal{S}}\mathbf{x}^{(i)\intercal}_{\mathcal{S}}\right) \geq \frac{C_{\min}}{2}.$$

Proof. First, for the ease of the notation, we define:

$$\hat{\mathbf{Q}}^{\mathcal{J}}_{\mathcal{S}\mathcal{S}} := \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} b''\left(\mathbf{w}^{\star\intercal}_{\mathcal{S}}\mathbf{x}^{(i)}_{\mathcal{S}}\right)\mathbf{x}^{(i)}_{\mathcal{S}}\mathbf{x}^{(i)\intercal}_{\mathcal{S}}$$

The proof is very similar to the proof of Lemma 12.11. Basically, we use the Courant-Fischer formula to arrive at Eq. (27).

$$\Lambda_{\min}\left(\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}}\right) \geq \Lambda_{\min}\left(\mathbf{Q}_{\mathcal{SS}}^{\star}\right) - \left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}} - \mathbf{Q}_{\mathcal{SS}}^{\star}\right\|_2 \geq C_{\min} - \left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}} - \mathbf{Q}_{\mathcal{SS}}^{\star}\right\|_2$$

To bound the second term, we decompose the sum over samples in J^{cc} to all clean samples minus the sum over clean samples treated as outliers:

$$\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}} = \hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}} - \hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}}$$

Further, subtracting both sides from $\mathbf{Q}_{\mathcal{SS}}^{\star}$ and using triangle inequality, we claim:

$$\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}} - \mathbf{Q}_{\mathcal{SS}}^{\star}\right\|_2 \leq \left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{*c}} - \mathbf{Q}_{\mathcal{SS}}^{\star}\right\|_2 + \varepsilon\left\|\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{co}} - \mathbf{Q}_{\mathcal{SS}}^{\star}\right\|_2$$

The first term in the RHS can be upper bounded by using Proposition 2.1 of Vershynin [2012]. For the second term involving J^{co}, we use the lighter tail of sub-exponential distribution as there are fewer terms in the set J^{co}. We also take a union bound across all the sets of size εn as done in Lemma 12.3 to arrive at the bound of $\mathcal{O}(\log(1/\varepsilon))$ if $n = \frac{Ck\log(p)}{\varepsilon\log(1/\varepsilon)}$.

Hence, we claim:

$$\Lambda_{\min}\left(\hat{\mathbf{Q}}_{\mathcal{SS}}^{J^{cc}}\right) \geq C_{\min} - \left(\frac{C_{\min}}{2} + \varepsilon\log(1/\varepsilon)C_{\min}\right) = \mathcal{O}\left(\frac{1}{2} - \varepsilon\log(1/\varepsilon)\right)C_{\min}$$

□

Lemma 12.13. *If $m > Ck\log(p)$, then $\hat{\mathbf{Q}}_{\mathcal{SS}} \succ 0$, with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$.*

Proof. To prove that the minimum eigenvalue of Hessian is strictly positive. Note that after Eq. (28), we show $\Lambda_{\min}\left(\hat{\mathbf{Q}}_{\mathcal{SS}}\right) > \frac{C_{\min}}{2}$. Note that $C_{\min} > 0$. Hence we arrive at $\Lambda_{\min}\left(\hat{\mathbf{Q}}_{\mathcal{SS}}\right) > 0$, which leads to $\hat{\mathbf{Q}}_{\mathcal{SS}} \succ 0$. □

13 LOGIT MODELS PROOFS

Lemma 13.1 (Strict Dual Feasibility for Logit Family Models to show $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$)**.** *If $\lambda = \Omega\left(\zeta\sigma\sigma_e\left(\sqrt{\frac{\log(p)}{n}} + \varepsilon\log(1/\varepsilon)\right)\right)$ and $n = \frac{k^2\log(p)}{\varepsilon^2\log^2(1/\varepsilon)}$, then with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$ we claim $\|\mathbf{z}_{\mathcal{S}^c}\|_\infty < 1$ and hence $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$.*

Proof. We start from Eq. (19)

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty + \frac{1}{\lambda}\left(1 + \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty\right)\left(\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty + \left\|\hat{\mathbf{r}}^{J^{cc}}\right\|_\infty + \left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty\right)$$

Further, we use the bound for individual terms $\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty$ and $\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty$ from Lemma 13.3 and Lemma 13.4 respectively:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq$$
$$\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty + \frac{1}{\lambda}\left(1 + \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty\right)\left(\mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon)\right) + \left\|\hat{\mathbf{r}}^{J^{cc}}\right\|_\infty + \sqrt{Cg(\sigma^2,\varepsilon,p)}\right)$$

We can bound $\left\|\hat{\mathbf{r}}^{J^{cc}}\right\|_\infty$ using Lemma 13.5 to arrive at:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty + \frac{1}{\lambda}\left(1 + \left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty\right)\mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon) + \sqrt{Cg(\sigma^2,\varepsilon,p)}\right)$$

Further, we apply Lemma 12.8 to bound $\left\|\hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}\hat{\mathbf{Q}}_{\mathcal{SS}}^{-1}\right\|_\infty$ as $1 - \frac{\gamma}{2} + \mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right)$.

Further, using regularization parameter $\lambda = \frac{\left(1-\frac{\gamma}{2}+\mathcal{O}(\varepsilon\log(1/\varepsilon))\right)}{\frac{\gamma}{4}}\Omega\left(\sigma\sigma_e\left(\sqrt{\frac{\log(p)}{n}}+\varepsilon\log(1/\varepsilon)\right)+\sqrt{Cg(\sigma^2,\varepsilon,p)}\right) = \Omega\left(\frac{\sigma\sigma_e}{\gamma}\left(\sqrt{\frac{\log(p)}{n}}+\varepsilon\log(1/\varepsilon)\right)+\sqrt{Cg(\sigma^2,\varepsilon,p)}\right)$. Here, we have used $1-\frac{\gamma}{2}+\mathcal{O}\left(\varepsilon\log(1/\varepsilon)\right) \leq 1-\frac{\gamma}{4} \leq 1$ for $\gamma = \Omega(\varepsilon\log(1/\varepsilon))$. Finally, we arrive at:

$$\|\mathbf{z}_{\mathcal{S}^c}\|_\infty \leq 1-\frac{\gamma}{4} < 1.$$

By using dual norm properties, we can ensure $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$. □

Lemma 13.2 (Parameter Vector Bound for Logit Family Models)**.** *If regularization parameter* $\lambda = \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\sigma\sigma_e\varepsilon\log(1/\varepsilon)+\sqrt{C}\sqrt{g(\sigma^2,\varepsilon,p)}\right)$, $n = \Omega\left(\frac{k\log(p)}{\varepsilon^2\log^2(1/\varepsilon)}\right)$, *then with probability at least* $1-\mathcal{O}\left(\frac{1}{p}\right)$*:*

$$\|\tilde{\mathbf{w}}_{\mathcal{S}}-\mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)+\zeta\sqrt{C}\sqrt{g(\sigma^2,\varepsilon,p)}\right).$$

Proof. We start by applying Eq. (18)

$$\|\tilde{\mathbf{w}}_{\mathcal{S}}-\mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \left\|\hat{\mathbf{Q}}^{J^{cc}-1}_{\mathcal{S}\mathcal{S}}\right\|_\infty\left(\left\|\hat{\mathbf{g}}^{J^{cc}}_{\mathcal{S}}\right\|_\infty+\left\|\hat{\mathbf{r}}^{J^{cc}}_{\mathcal{S}}\right\|_\infty+\left\|\hat{\mathbf{h}}^{J^{oc}}_{\mathcal{S}}\right\|_\infty+\lambda\right)$$

Further, we use the bound for individual terms $\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty$ and $\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty$ from Lemma 13.3 and Lemma 13.4 respectively. We also use Lemma 12.7 to bound $\left\|\hat{\mathbf{Q}}^{J^{cc}-1}_{\mathcal{S}\mathcal{S}}\right\|_\infty$ as $\mathcal{O}\left(\left\|\mathbf{Q}^{\star-1}_{\mathcal{S}\mathcal{S}}\right\|_\infty\right) = \mathcal{O}(\zeta)$. We finally arrive at:

$$\|\tilde{\mathbf{w}}_{\mathcal{S}}-\mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \mathcal{O}(\zeta)\left(\mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)+\sqrt{C}\times\sqrt{g(\sigma^2,\varepsilon,p)}+\left\|\hat{\mathbf{r}}^{J^{cc}}\right\|_\infty+\lambda\right)$$

We further bound $\left\|\hat{\mathbf{r}}^{J^{cc}}\right\|_\infty$ using Lemma 13.5 as $\mathcal{O}\left(\zeta\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\zeta\sigma\sigma_e\varepsilon\log(1/\varepsilon)+\zeta\sqrt{C}\sqrt{g(\sigma^2,\varepsilon,p)}\right)$. We arrive at:

$$\|\tilde{\mathbf{w}}_{\mathcal{S}}-\mathbf{w}^\star_{\mathcal{S}}\|_\infty \leq \mathcal{O}(\zeta)\left(\mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)+\sqrt{C}\times\sqrt{g(\sigma^2,\varepsilon,p)}+\lambda\right)$$

Further, using our regularization parameter as $\lambda = \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\sigma\sigma_e\varepsilon\log(1/\varepsilon)+\sqrt{C}\sqrt{g(\sigma^2,\varepsilon,p)}\right)$, we arrive at the desired result.

□

Lemma 13.3 (Gradient bound over selected clean samples for logit family models)**.** *If* $\lambda > 8\sigma R\sqrt{\frac{\log(p)}{n}}$ *and* $n = \Omega\left(\frac{k^2\log(p)}{\varepsilon\log(1/\varepsilon)}\right)$*, then for the case of logit models:*

$$\left\|\hat{\mathbf{g}}^{J^{cc}}\right\|_\infty \leq \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}}+\sigma\sigma_e\varepsilon\log(1/\varepsilon)\right)$$

with probability at least $1-\mathcal{O}\left(\frac{1}{p}\right)$.

Proof. For logit family models, we have $b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) = \frac{\exp\left(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}\right)}{1+\exp\left(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}\right)}$ and $y^{(i)} \leq R$, where R is a constant. For example, in logistic regression we have $y^{(i)} \in \{0,1\} \leq 1$. We revisit

$$\hat{\mathbf{g}}^{J^{*c}} = \frac{1}{|J^{*c}|}\sum_{i\in J^{*c}}\left(b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})-y^{(i)}\right)\mathbf{x}^{(i)}$$

As both random variables $0 < b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) < 1$ and $y^{(i)} < R$ are bounded. Hence, we can consider them sub-Gaussian. Hence, $-R < b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) - y^{(i)} < 1 - R$ is also bounded and hence sub-Gaussian.

For the ease of notation let $e^{(i)} := b'(\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}) - y^{(i)}$. As $\mathbf{x}^{(i)}$ and $e^{(i)}$ are sub-Gaussian variables, their product is sub-exponential. The rest of the proof is similar to the proof of Lemma 12.3. □

Lemma 13.4 (Gradient bound over selected corrupted samples for logit family models)**.** *If $n \geq \frac{k\log(p)}{\varepsilon\log(1/\varepsilon)}$, then with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$ for the particular case of logit family models with $y_i \leq C$, we claim*

$$\left\|\hat{\mathbf{h}}^{J^{oc}}\right\|_\infty \leq \sqrt{C} \times \sqrt{g(\sigma^2, \varepsilon, p)}$$

where C is a constant and

$$g(\sigma^2, \varepsilon, p) = \begin{cases} \mathcal{O}\left(\sigma_a^2 \log(1/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{sub-Gaussian}(\sigma_a) \\ \mathcal{O}\left(\sigma^2 \log(p/\varepsilon)\right) & \text{if } \mathbf{x}^{(i)} \sim \textit{heavy-tailed with bounded } 4^{th}\textit{-moment} \end{cases} \tag{29}$$

Proof. For the particular case of logit family models, we want to bound

$$\hat{\mathbf{h}}^{J^{oc}} = \nabla_{\mathbf{w}} l\left(\tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{oc}}\right) = \frac{1}{|J^{oc}|}\sum_{i\in J^{oc}}\left(b'(\tilde{\mathbf{w}}^{\intercal}\mathbf{x}^{(i)})\mathbf{x}^{(i)} - y^{(i)}\mathbf{x}^{(i)}\right) = \frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\frac{1}{1+\exp\left(-\tilde{\mathbf{w}}^{\intercal}\mathbf{x}^{(i)}\right)} - y^{(i)}\right)\mathbf{x}^{(i)}$$

where denotes the sigmoid function For any j^{th} entry, we can apply Cauchy-Schwarz inequality

$$\left|\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\frac{1}{1+\exp\left(-\tilde{\mathbf{w}}^{\intercal}\mathbf{x}^{(i)}\right)} - y^{(i)}\right)\mathbf{x}_j^{(i)}\right| \leq \left(\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\left(\frac{1}{1+\exp\left(-\tilde{\mathbf{w}}^{\intercal}\mathbf{x}^{(i)}\right)} - y^{(i)}\right)^2\right)^{1/2}\left(\frac{1}{n\varepsilon}\sum_{i\in J^{oc}}\mathbf{x}_j^{(i)2}\right)^{1/2}$$

We know that $\frac{1}{1+\exp\left(-\tilde{\mathbf{w}}^{\intercal}\mathbf{x}^{(i)}\right)} \in [0,1]$ for all samples irrespective of them being corrupted. We also know that $y^{(i)}$ is bounded for logit models by some constant C. Hence the first term can be bounded as C. For the second term, we use Lemma 12.5.

□

Lemma 13.5. *If $n \geq \frac{k\log(p)}{\varepsilon\log(1/\varepsilon)}$, then with probability at least $1 - \mathcal{O}\left(\frac{1}{p}\right)$ for the particular case of logit family models*

$$\left\|\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}})\right\|_\infty \leq \mathcal{O}\left(\sigma\sigma_e\sqrt{\frac{\log(p)}{n}} + \sigma\sigma_e\varepsilon\log(1/\varepsilon) + \sqrt{C}\sqrt{g(\sigma^2, \varepsilon, p)}\right)$$

Proof. Recall that

$$\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{*c}}) = \left(\nabla_{\mathbf{w}}^2 l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{*c}}\right) - \nabla_{\mathbf{w}}^2 l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{*c}}\right)\right)(\tilde{\mathbf{w}} - \mathbf{w}^\star).$$

By the triangle inequality,

$$\left\|\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{*c}})\right\|_\infty \leq \left\|\nabla_{\mathbf{w}}^2 l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{*c}}\right)(\tilde{\mathbf{w}} - \mathbf{w}^\star)\right\|_\infty + \left\|\nabla_{\mathbf{w}}^2 l\left(\mathbf{w}^\star, \tilde{\mathbf{t}}^{J^{*c}}\right)(\tilde{\mathbf{w}} - \mathbf{w}^\star)\right\|_\infty. \tag{30}$$

We bound the first term; the second follows identically.

For logit models,

$$\nabla_{\mathbf{w}}^2 l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}^{J^{*c}}\right) = \frac{1}{n}\sum_{i\in J^{*c}} b''(\boldsymbol{\nu}^{\intercal}\mathbf{x}^{(i)})\,\mathbf{x}^{(i)}\mathbf{x}^{(i)\intercal}, \qquad 0 \leq b''(\cdot) \leq \frac{1}{4}.$$

Let $\boldsymbol{\Delta} := \tilde{\mathbf{w}} - \mathbf{w}^\star$ and hence $\boldsymbol{\Delta}_{\mathcal{S}^c} = 0$. Then

$$\nabla^2_{\mathbf{w}} l\left(\boldsymbol{\nu}, \tilde{\mathbf{t}}\right) \boldsymbol{\Delta} = \begin{bmatrix} \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \\ \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \end{bmatrix} \boldsymbol{\Delta}_{\mathcal{S}},$$

where

$$\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) := \frac{1}{n} \sum_{i \in J^{*c}} b''(\boldsymbol{\nu}^\intercal \mathbf{x}^{(i)})\, \mathbf{x}^{(i)}_{\mathcal{S}} \mathbf{x}^{(i)\intercal}_{\mathcal{S}} \in \mathbb{R}^{k \times k}, \qquad \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) := \frac{1}{n} \sum_{i \in J^{*c}} b''(\boldsymbol{\nu}^\intercal \mathbf{x}^{(i)})\, \mathbf{x}^{(i)}_{\mathcal{S}^c} \mathbf{x}^{(i)\intercal}_{\mathcal{S}} \in \mathbb{R}^{(p-k) \times k}.$$

Coordinates in $\mathcal{S}$: For $j \in \mathcal{S}$,

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \boldsymbol{\Delta}_{\mathcal{S}} \right\|_\infty \leq \left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \right\|_\infty \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty \leq \sqrt{k} \left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty, \tag{31}$$

where we have used basic matrix norm inequalities. Since $\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \succeq 0$ and $b''(\cdot) \leq \frac{1}{4}$,

$$\hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \preceq \frac{1}{4} \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}}, \qquad \text{where} \quad \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} := \frac{1}{n} \sum_{i \in J^{*c}} \mathbf{x}^{(i)}_{\mathcal{S}} \mathbf{x}^{(i)\intercal}_{\mathcal{S}}.$$

Therefore $\left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \right\|_2 \leq \frac{1}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} \right\|_2$. Substituting into (31),

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \boldsymbol{\Delta}_{\mathcal{S}} \right\|_\infty \leq \frac{\sqrt{k}}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty \leq \frac{\sqrt{k}}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}} \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty. \tag{32}$$

Coordinates in $\mathcal{S}^c$: For $j \in \mathcal{S}^c$,

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \boldsymbol{\Delta}_{\mathcal{S}} \right\|_\infty \leq \left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \right\|_\infty \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty \leq \sqrt{k} \left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty. \tag{33}$$

where we have used matrix norm inequalities. Let

$$\hat{\mathbf{Q}}(\boldsymbol{\nu}) = \begin{bmatrix} \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) & \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}^c}(\boldsymbol{\nu}) \\ \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) & \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}^c}(\boldsymbol{\nu}) \end{bmatrix} = \frac{1}{n} \sum_{i \in J^{*c}} b''(\boldsymbol{\nu}^\intercal \mathbf{x}^{(i)})\, \mathbf{x}^{(i)} \mathbf{x}^{(i)\intercal} \succeq 0.$$

For a PSD block matrix,

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \right\|^2_2 \leq \left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}^c}(\boldsymbol{\nu}) \right\|_2 \left\| \hat{\mathbf{Q}}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\nu}) \right\|_2 \leq \frac{1}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}^c\mathcal{S}^c} \right\|_2 \frac{1}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} \right\|_2,$$

where we have again used $b''(\cdot) \leq \frac{1}{4}$ and as per our notation defined earlier $\widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}^c\mathcal{S}^c} := \frac{1}{n} \sum_{i \in J^{*c}} \mathbf{x}^{(i)}_{\mathcal{S}^c} \mathbf{x}^{(i)\intercal}_{\mathcal{S}^c}$. Hence

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \right\|_2 \leq \frac{1}{4} \sqrt{\left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}\mathcal{S}} \right\|_2 \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}}_{\mathcal{S}^c\mathcal{S}^c} \right\|_2} \leq \frac{1}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}} \right\|_2.$$

Combining with Eq. (33),

$$\left\| \hat{\mathbf{Q}}_{\mathcal{S}^c\mathcal{S}}(\boldsymbol{\nu}) \boldsymbol{\Delta}_{\mathcal{S}} \right\|_\infty \leq \frac{\sqrt{k}}{4} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}} \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty. \tag{34}$$

Applying Eq. (32) and Eq. (34) to both terms in Eq. (30) yields the following bound for $\left\| \hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}) \right\|_\infty$.

$$\left\| \hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{*c}}) \right\|_\infty \leq \frac{\sqrt{k}}{2} \left\| \widehat{\boldsymbol{\Sigma}}^{J^{*c}} \right\|_2 \|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty.$$

We actually want to bound the sum of terms in the set J^{cc}. Hence, we express the sum over J^{cc} as the difference between sums over the set J^{*c} and J^{co}. Further, we apply triangle inequality to arrive at:

$$\left\| \hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}}) \right\|_\infty \leq \left\| \hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{*c}}) \right\|_\infty + \varepsilon \left\| \hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{co}}) \right\|_\infty.$$

Using the similar arguments for the set J^{co}, we can arrive at:

$$\left\|\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}})\right\|_\infty \leq \frac{\sqrt{k}}{2}\left(\left\|\widehat{\boldsymbol{\Sigma}}^{J^{*c}}\right\|_2 + \varepsilon\left\|\widehat{\boldsymbol{\Sigma}}^{J^{co}}\right\|_2\right)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty .$$

Further, we can bound the spectral norm to arrive at:

$$\left\|\hat{\mathbf{r}}(\boldsymbol{\nu}, \tilde{\mathbf{w}}, \tilde{\mathbf{t}}^{J^{cc}})\right\|_\infty \leq \frac{\sqrt{k}}{2} C_{\max}\left(1 + \varepsilon\log(1/\varepsilon)\right)\|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty$$

Further, we use $\frac{\sqrt{k}}{2} C_{\max}\left(1 + \varepsilon\log(1/\varepsilon)\right) \leq \frac{1}{2}$ for sufficiently small ε and Lemma 13.2 to bound $\|\boldsymbol{\Delta}_{\mathcal{S}}\|_\infty$ and arrive at the desired result.

□

14 EXPERIMENTS: LINEAR REGRESSION

In this section, we provide experimental settings in a detailed manner.

14.1 EXPERIMENTAL SETUP: DATA GENERATION PROCESS

We describe the data generation process first for linear regression in a detailed manner as follows:

1. **Support Generation:** A support set $\mathcal{S}$ of size $k = 5$ is generated uniformly at random from $\{1, 2, \ldots, p\}$, where p denotes the number of features. The nonzero entries of the parameter vector $\mathbf{w}^\star_{\mathcal{S}}$ are generated by sampling uniformly from $[-1.2, -0.1] \cup [0.1, 1.2]$ to avoid values close to zero. For the non-support coordinates, we set $\mathbf{w}^\star_{\mathcal{S}^c} = 0$.
2. **Clean Data:** We next describe the generation of clean samples $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$. Each feature vector $\mathbf{x}^{(i)} \in \mathbb{R}^p$ is generated independently from $\mathcal{N}(0, I_p)$. The response variable is generated according to the linear model

   $$y^{(i)} = \mathbf{w}^{\star\intercal}\mathbf{x}^{(i)} + e^{(i)},$$

   where the noise term $e^{(i)} \sim \mathcal{N}(0, \sigma_e^2)$ with $\sigma_e = 0.1$.
3. **Adversarial Corruption:** Next, we generate samples corrupted by an adversary. The number of corrupted samples depends on the corruption proportion, while their construction depends on the adversarial corruption model. We consider the following corruption models:
   (a) Random adversarial corruption,
   (b) Strong adversarial corruption model of Chen et al. [2013],
   (c) Heavy-tailed adversarial corruption.

   The specific details for each corruption model are described in the later sections.
4. **Estimation:** The resulting dataset, consisting of both clean and adversarially corrupted samples, is provided to all algorithms for support estimation. We compare classical LASSO [Tibshirani, 1996], Robust LASSO [Chen et al., 2013], RSGE [Liu et al., 2020], Trimmed MLE [Awasthi et al., 2022], SVAM [Mukhoty et al., 2023], and our proposed method.
5. **Evaluation Metrics:** We consider two metrics to evaluate the performance of all estimation methods:
   (a) Support recovery probability: Steps 1-4 are repeated multiple times to estimate the probability of exact support recovery, defined as the event $\mathcal{S}(\hat{\mathbf{w}}) = \mathcal{S}(\mathbf{w}^\star)$. The support recovery probability is computed as the fraction of successful recoveries over the total number of repetitions.
   (b) F1 score: We additionally evaluate partial support recovery using the F1 score between the estimated support $\mathcal{S}(\hat{\mathbf{w}})$ and the true support $\mathcal{S}(\mathbf{w}^\star)$. Let

   $$\text{Precision} = \frac{|\mathcal{S}(\hat{\mathbf{w}}) \cap \mathcal{S}(\mathbf{w}^\star)|}{|\mathcal{S}(\hat{\mathbf{w}})|}, \qquad \text{Recall} = \frac{|\mathcal{S}(\hat{\mathbf{w}}) \cap \mathcal{S}(\mathbf{w}^\star)|}{|\mathcal{S}(\mathbf{w}^\star)|}.$$

   The F1 score is defined as the harmonic mean of precision and recall:

   $$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

   The reported F1 score corresponds to the average across all repetitions.

14.2 ESTIMATION ALGORITHMS AND HYPERPARAMETERS

In this section, we describe the hyperparameter settings of the proposed and existing methods used in the evaluation to ensure reproducibility.

1. **LASSO** [Tibshirani, 1996]: This corresponds to the classical LASSO estimator, which utilizes all available samples during estimation. The method requires a regularization parameter, which we set to 0.1 in all experiments.

2. **Robust LASSO** [Chen et al., 2013]: We implement Algorithm 4 (Robust LASSO) from Chen et al. [2013]. We assume that the algorithm has access to the true hyperparameters n_1 and $R = \|\beta^*\|_1$, which favors this baseline. Here, n_1 denotes the number of corrupted samples and R denotes the ℓ_1 norm of the true parameter vector.

3. **RSGE** [Liu et al., 2020]: This method is based on the Robust Sparse Gradient Estimator (RSGE). We use Algorithm 1 from Liu et al. [2020], which requires the hard-thresholding parameter $\tilde{k}$. In our experiments, we provide $\tilde{k} = k$, where k is the true sparsity level, which again favors this method. The maximum number of iterations is set to $T = 40$.

4. **Trimmed MLE** [Awasthi et al., 2022]: This approach performs estimation by trimming potentially corrupted samples. Accordingly, we provide the true corruption proportion ε to the algorithm. Following Theorem 5.1 of Awasthi et al. [2022], we set the convergence parameter $\eta = \varepsilon^2$ and $R = 10000$. Algorithm 2 in Awasthi et al. [2022] additionally requires an estimate of the covariance matrix. We use the sample covariance matrix so that the setting remains comparable with other estimation methods. The algorithm also has a preprocessing step (Algorithm 4 of Dong et al. [2019]), which involves the hyperparameters γ_1, γ_2, and C. In our experiments, we set $\gamma_1 = 0.1$, $\gamma_2 = 0.1$, and $C = 20$.

5. **SVAM** [Mukhoty et al., 2023]: We implement Algorithm 1 from Mukhoty et al. [2023], termed as Sequential Variance-Altered MLE (SVAM). The hyperparameters are selected following the publicly available implementation provided by the authors. Specifically, we use $\alpha = 0.15$, step size $\eta = 1$, and initialize the regression vector $\mathbf{w}_{\text{init}}$ with a random vector. The maximum number of iterations is set to 40. The parameter β is selected from the set $[2, 1, 0.5, 0.1, 0.01]$ as done in Mukhoty et al. [2023].

6. **Our Proposed Algorithm:** Our algorithm requires the corruption proportion like Chen et al. [2013], Awasthi et al. [2022] and hence we supply the true corruption proportion. For the medians of means based pre-processing step, we use the number of batches as $B = (\log(p/\delta))$, $T = 6\sqrt{\log(p/\varepsilon)}$, where $\delta = 10^{-3}$ is a constant and ε denotes the corruption proportion.. We chose the regularization parameter $\lambda = 0.1$. We used maximum number of iterations as 40. We also had a convergence criteria to stop earlier if $\frac{\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\|_2}{\|\mathbf{w}^{(i)}\|_2} < 10^{-6}$.

14.3 RANDOM ADVERSARIAL CONTAMINATION IN BOTH FEATURES AND LABELS

In this section, we briefly describe the random adversarial contamination model and explain how corruption is introduced in both features and labels. Under this model, the corrupted feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^p$ are generated independently from $\mathcal{N}(\mathbf{0}, 5\mathbf{I}_p)$, while the corresponding corrupted responses $y^{(i)}$ are generated independently from $\mathcal{N}(0, 5)$, independent of $\mathbf{x}^{(i)}$.

14.3.1 Low-dimensional Regime

We start with the simple case of $n = 200$, $p = 30$, $k = 5$, $\sigma = 1$, and $\sigma_e = 0.1$ for data generation. For We consider five values of corruption proportions, $\varepsilon \in \{0.05, 0.1, 0.15, 0.2\}$.

Further, the data with contamination is passed to all the methods for support estimation. We repeat 30 trials of this procedure to compute the empirical support recovery. The results for support recovery are shown in Figure 2. We also show the mean and the standard deviation from 30 runs. As multiple lines in Figure 2 are overlapping for support recovery metric, we also present the F1 score in Figure 3.

As the lines in Figure 3 largely overlap, we report the mean and standard deviation of the support recovery rate and F1 score in Tables 6 and 7, respectively.

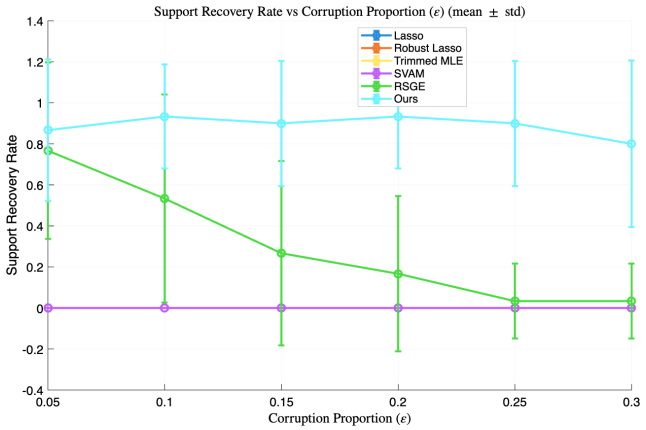


Figure 2: Support Recovery Rate vs Corruption Proportion

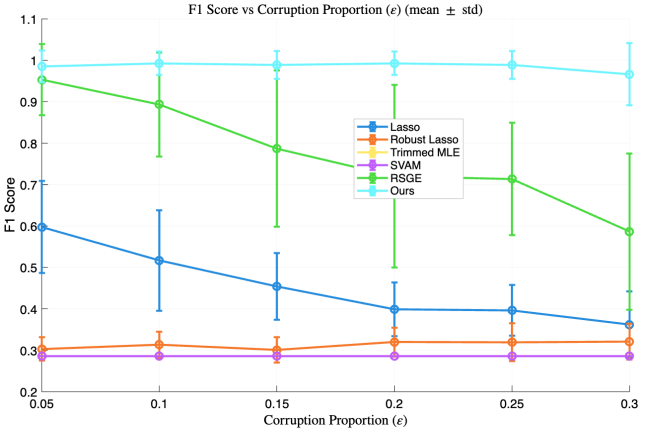


Figure 3: F1 score vs Corruption Proportion

Figure 4: Support Recovery Rate and F1 score vs corruption proportion shows that our method performs better than existing methods. It also shows that the support recovery rate or F1 score decreases as the corruption proportion is increased.

Table 6: Support recovery rates (mean $\pm$ standard deviation over 30 runs) of competing methods across varying corruption proportions (ε), with $n = 200$, $p = 30$, $k = 5$, $\sigma = 1$, and $\sigma_e = 0.1$. Our method consistently achieves high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	RSGE	**Ours**
0.05	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.867 ± 0.346	0.933 ± 0.254
0.1	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.5 ± 0.509	0.865 ± 0.321
0.15	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.3 ± 0.466	0.933 ± 0.254
0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.167 ± 0.379	0.863 ± 0.346

Table 7: F1 scores (mean $\pm$ standard deviation over 30 runs) of competing methods across varying corruption proportions (ε), with $n = 200$, $p = 30$, $k = 5$, $\sigma = 1$, and $\sigma_e = 0.1$. Our method consistently maintains high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	RSGE	**Ours**
0.05	0.652 ± 0.124	0.313 ± 0.055	0.286 ± 0	0.286 ± 0	0.973 ± 0.069	0.993 ± 0.028
0.1	0.472 ± 0.081	0.309 ± 0.039	0.286 ± 0	0.286 ± 0	0.873 ± 0.153	0.985 ± 0.038
0.15	0.468 ± 0.096	0.305 ± 0.022	0.286 ± 0	0.286 ± 0	0.807 ± 0.153	0.993 ± 0.028
0.2	0.421 ± 0.067	0.317 ± 0.047	0.286 ± 0	0.286 ± 0	0.767 ± 0.149	0.985 ± 0.038

We make the following observations from Tables 6 and 7:

1. As shown in Table 6, support recovery is a binary metric. It is 1 if the true support is exactly recovered and 0 otherwise. This makes it a stringent metric. We therefore additionally report the F1 score to enable a more comprehensive comparison across all methods.
2. Table 7 provides a more discriminative comparison among estimation methods. Notably, our proposed method achieves the higher F1 scores across all corruption proportion. This demonstrates its strong ability to recover the true support under varying levels of corruption.
3. As expected, both the support recovery rate and F1 score decrease as the corruption proportion increases, reflecting the increasing difficulty of the estimation problem under heavier contamination. Notably, our proposed method maintains relatively better performance even at higher corruption levels.

In this section, we demonstrated the strong empirical performance of the proposed algorithm in the low-dimensional regime,

whereas all theoretical guarantees in this work are established for the high-dimensional setting. In the next section, we compare the computational runtime of existing methods and demonstrate why RSGE Liu et al. [2020] does not scale well to high-dimensional regimes. We then continue the empirical evaluation in the high-dimensional setting by comparing our method with all existing approaches except RSGE.

14.3.2 Comparison of Time Taken By Each Method

We compare the time taken by each method in the estimation process. The mean and standard deviation from the time consumed from 30 runs is presented in Table 8.

Table 8: Comparison of time consumed by each algorithm in milliseconds for increasing dimension. We present the mean and standard deviation from 30 runs. The results show that our approach takes significantly less time as compared to our closest baseline of RGSE.

p	LASSO	Robust LASSO	Trimmed MLE	SVAM	RSGE	**Ours**
10	0.2877 ± 1.0289	0.4459 ± 0.2152	0.1502 ± 0.1521	0.3127 ± 0.1473	17770 ± 31650	0.6223 ± 1.0552
20	0.0778 ± 0.0112	1.2749 ± 0.1691	0.1527 ± 0.0860	0.4747 ± 0.0635	77790 ± 47490	0.5805 ± 0.2096
30	0.0969 ± 0.0344	2.4420 ± 0.0747	0.2312 ± 0.0309	0.6805 ± 0.0152	177260 ± 92140	0.6702 ± 0.2097
40	0.1131 ± 0.0124	4.1777 ± 0.1609	0.4304 ± 0.0481	0.9270 ± 0.0210	526280 ± 179540	0.8243 ± 0.2283

14.3.3 High-dimensional Regime

In this section, we increase the ambient dimension $p = 400$ with other parameters same as $n = 200$, $k = 5$, $\sigma = 1$, and $\sigma_e = 0.1$.

In Section 14.3.2, we discussed that RSGE [Liu et al., 2020] takes significantly more time due to the computationally expensive semi-definite dual program (SDP) solved in each iteration. Hence, in this section, we only consider other existing methods of LASSO, Robust LASSO [Chen et al., 2013], Trimmed MLE [Awasthi et al., 2022], SVAM [Mukhoty et al., 2023]. We also increased the number of runs to 100 as these methods much better runtime.

Further, we compute support recovery rate and F1 scores by these methods and report them in Table 9 and Table 10 respectively. From Table 9, we observe that the support recovery rate of existing methods is 0. Hence, we also compare the F1 scores in Table 10. For both metrics, we observe that our method consistently performs impressively across increasing corruption proportions.

Table 9: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 200$, $p = 400$, and $k = 5$ under random adversarial corruption. Our method consistently achieves high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.01 ± 0.1	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.1	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.94 ± 0.239
0.15	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.58 ± 0.496
0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.08 ± 0.273

Table 10: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) for high-dimensional regime with $n = 200$, $p = 400$, and $k = 5$ under random adversarial corruption. Our method consistently maintains high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.402 ± 0.147	0.2 ± 0.038	0.025 ± 0.011	0.025 ± 0	1 ± 0
0.1	0.302 ± 0.058	0.175 ± 0.037	0.024 ± 0.011	0.025 ± 0	0.988 ± 0.05
0.15	0.251 ± 0.047	0.145 ± 0.035	0.026 ± 0.011	0.025 ± 0	0.853 ± 0.218
0.2	0.242 ± 0.048	0.12 ± 0.042	0.022 ± 0.014	0.025 ± 0	0.539 ± 0.216

In this section, we verify that our theoretical guarantees hold in the high-dimensional regime under random adversarial corruption, which does not correspond to adaptive or strong adversarial contamination. One of the key contributions of our work, in addition to handling the high-dimensional setting, is robustness against adaptive/strong adversarial contamination. Therefore, in the next section, we evaluate the performance of our proposed method and existing approaches under the adaptive/strong adversarial contamination model introduced by Chen et al. [2013].

14.4 ADAPTIVE/STRONG ADVERSARIAL CONTAMINATION MODEL FROM Chen et al. [2013]

In this section, we consider the adaptive/strong adversarial contamination model proposed by Chen et al. [2013] to avoid providing any unintended advantage to our proposed method due to the random adversarial corruption considered in the previous experiments. The adversarial contamination in Chen et al. [2013] is constructed using knowledge of the true support and the clean data, thereby explicitly targeting the support recovery task and making it significantly more challenging. We briefly describe their adversarial contamination model below.

They compute optimal $\boldsymbol{\theta}^*$ with respect from clean data in the true non-support using:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p-k} : \|\boldsymbol{\theta}\|_1 \leq \|\boldsymbol{\beta}^*\|_1}{\arg\min} \left\| \mathbf{y}^A - \mathbf{X}^A_{\Lambda^c} \boldsymbol{\theta} \right\|_2 .$$

Here, Λ^c denotes the complement of the support set and superscript A in $\mathbf{y}^A, \mathbf{X}^A$ denotes the actual (clean) samples in the notation from Chen et al. [2013]. Further, the outliers in features for support are chosen as $\mathbf{X}^O_\Lambda = \frac{3}{\sqrt{n}} \mathbf{A}$, where $\mathbf{A}$ is a random ± 1 matrix of dimension $n_1 \times k$. Here n_1 denotes the number of outliers.

Note that Chen et al. [2013] considers $\mathbf{X}^O_\Lambda = \frac{3}{\sqrt{n}} \mathbf{A}$ instead of $\mathbf{X}^O_\Lambda = \mathbf{A}$ like us because Section 6.1.1 of Chen et al. [2013] considers the per sample variance in $\mathcal{O}\left(\frac{1}{n}\right)$. We consider this assumption of per sample variance decaying with respect to number of samples instead of constant quite restrictive. Hence, we intentionally select a more challenging adversarial model with $\mathbf{X}^O_\Lambda = \mathbf{A}$.

Next Chen et al. [2013] select the label outliers using $\mathbf{y}^O = \mathbf{X}^O_\Lambda(-\boldsymbol{\beta}^*)$. Further, the outlier features for non-support are chosen as follows. For $i = 1, \ldots, n_1$, set $\mathbf{X}^O_{i,\Lambda^c} = \dfrac{y^O_i}{\mathbf{B}_i^\top \boldsymbol{\theta}^*} \mathbf{B}_i^\top$ where $\mathbf{B}_i$ is a $(p-k)$-vector with i.i.d. standard Gaussian entries.

We provide this adversarially contaminated data to the estimation algorithms. The mean and standard deviation are reported in Table 11. As multiple methods lead to zero support recovery, we also compute the F1 scores, which is presented in Table 12.

Table 11: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from strong contamination model by Chen et al. [2013] discussed in Section 14.4. Our method consistently achieves high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.1	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.98 ± 0.141
0.15	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.86 ± 0.349
0.2	0.01 ± 0.1	0 ± 0	0 ± 0	0 ± 0	0.37 ± 0.485

Table 12: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from strong contamination model by Chen et al. [2013] discussed in Section 14.4. Our method consistently maintains high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.543 ± 0.132	0.182 ± 0.024	0.02 ± 0.011	0.025 ± 0	1 ± 0
0.1	0.535 ± 0.132	0.103 ± 0.044	0.02 ± 0.01	0.025 ± 0	0.998 ± 0.013
0.15	0.549 ± 0.116	0.043 ± 0.036	0.018 ± 0.011	0.025 ± 0	0.975 ± 0.069
0.2	0.542 ± 0.128	0.063 ± 0.039	0.015 ± 0.012	0.025 ± 0	0.831 ± 0.16

We make the following inferences from Table 11 and Table 12:

1. As compared to the random adversarial attack in Table 9 and Table 10, the reported support recovery rate and F1 scores in Table 11 and Table 12 are smaller in general. This makes sense because we consider the strong adversarial model by Chen et al. [2013] in Table 11 and Table 12, and hence the performance is supposed to drop.
2. From Table 11 and Table 12, we can easily infer from that our proposed method performs well across all the corruption proportion even for the strong adversarial attack considered by Chen et al. [2013].
3. From Table 11 and Table 12, We also observe that our method performs better than Robust LASSO proposed by Chen et al. [2013] on the adversarial corruption model considered by Chen et al. [2013]. This happens because we considered the more practical and challenging adversary with constant per sample variance instead of decay per sample variance like $\mathcal{O}\left(\frac{1}{n}\right)$ considered by Chen et al. [2013].

In this section, we verified our theoretical guarantees under the adaptive/strong adversarial contamination model in the high-dimensional regime. Another key contribution of our work is robustness to heavy-tailed distributions, which was not considered either in the previous section or in the experiments based on the adversarial model of Chen et al. [2013]. Therefore, in the next section, we evaluate our method to further validate the theoretical claims under heavy-tailed data distributions.

14.5 OUR PROPOSED METHOD CAN HANDLE HEAVY-TAILED DISTRIBUTIONS

One of the key contributions of our work is robustness to heavy-tailed distributions requiring only bounded fourth-order moments. Therefore, in this section, we evaluate the performance of our proposed method and existing approaches when both the features and labels are adversarially contaminated using heavy-tailed distributions, such as the log-normal distribution and the Student's t-distribution.

14.5.1 Corruption based on Log-Normal distribution

We first briefly describe the data generating process. In the adversarial corruption setting, the corrupted samples are generated from a log-normal distribution, whose logarithm follows a normal distribution. Specifically, for any corrupted sample i, we generate

$$\log\left(\mathbf{x}^{(i)}\right) \sim \mathcal{N}(\mathbf{0}, 5\mathbf{I}), \quad \log\left(y^{(i)}\right) \sim \mathcal{N}(0, 5).$$

Consequently, the corrupted samples follow a heavy-tailed distribution with all moments finite.

Next, we generate both clean and corrupted datasets under varying corruption proportions. The support recovery rate and F1 score achieved by existing methods and our proposed method are reported in Table 13 and Table 14, respectively.

Table 13: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from Log-Normal distribution. Our method consistently achieves high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.02 ± 0.141	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.1	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.99 ± 0.1
0.15	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.96 ± 0.197
0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.64 ± 0.482

Table 14: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from Log-Normal distribution. Our method consistently maintains high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.486 ± 0.216	0.219 ± 0.046	0.024 ± 0.011	0.025 ± 0	1 ± 0
0.1	0.371 ± 0.124	0.178 ± 0.039	0.023 ± 0.012	0.025 ± 0	0.999 ± 0.009
0.15	0.327 ± 0.119	0.148 ± 0.039	0.024 ± 0.013	0.025 ± 0	0.996 ± 0.019
0.2	0.311 ± 0.102	0.112 ± 0.046	0.025 ± 0.012	0.025 ± 0	0.953 ± 0.071

From Table 13 and Table 14, we observe that our method consistently achieves high support recovery rates and F1 scores across varying corruption levels. This demonstrates the robustness of the proposed approach to adversarial corruption arising from heavy-tailed distributions in the high-dimensional regime.

In the next subsection, we consider another corruption setting based on a different heavy-tailed distribution.

14.5.2 Corruption based on student's t-distribution

First, we briefly describe the adversarial contamination model. Any corrupted sample $\mathbf{x}^{(i)}$ is generated using the random variable $\mathbf{z}^{(i)}$ described as follows:

$$\mathbf{z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad v^{(i)} \sim \chi^2_{\mathrm{df}}, \quad \mathbf{x}^{(i)} = \frac{\mathbf{z}^{(i)}}{\sqrt{v_i/\mathrm{df}}},$$

where χ^2_{df} denotes the chi-square distribution with degrees of freedom $= \mathrm{df}$. We chose $\mathrm{df} = 5$ in our experiments. Hence, the fourth moment is bounded. The corrupted samples for the labels $y^{(i)}$ are also generated in a similar manner.

We provide this adversarially corrupted dataset to all estimation methods and report the corresponding support recovery rates and F1 scores in Table 15 and Table 16, respectively.

Table 15: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from student's t-distribution with degree of freedom = 3. Our method consistently achieves high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.4 ± 0.492	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.1	0.06 ± 0.239	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.15	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.97 ± 0.171
0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.8 ± 0.402

Table 16: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε) from student's t-distribution with degree of freedom = 3. Our method consistently maintains high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO	Robust LASSO	Trimmed MLE	SVAM	**Ours**
0.05	0.843 ± 0.181	0.208 ± 0.04	0.025 ± 0.012	0.025 ± 0	1 ± 0
0.1	0.578 ± 0.19	0.184 ± 0.033	0.023 ± 0.011	0.025 ± 0	1 ± 0
0.15	0.425 ± 0.139	0.15 ± 0.036	0.022 ± 0.011	0.025 ± 0	0.997 ± 0.016
0.2	0.35 ± 0.108	0.121 ± 0.05	0.023 ± 0.013	0.025 ± 0	0.971 ± 0.068

From Table 15 and Table 16, we observe that our method achieves high support recovery rates and F1 scores across all corruption proportions in the high-dimensional setting under heavy-tailed corruption generated using the Student's t-distribution. This further demonstrates the robustness of the proposed approach to adversarial contamination arising from heavy-tailed distributions such as the Student's t-distribution.

In the next section, we verify one of the main claims of our work about logarithmic sample complexity with respect to ambient dimension under adaptive/strong adversarial corruption, which is particularly important in high-dimensional settings.

14.6 EMPIRICAL VALIDATION OF OUR SAMPLE COMPLEXITY BEING LOGARITHMIC IN AMBIENT DIMENSION, MEANING $n = \Omega(\log(p))$

We first describe the experimental setup used to empirically verify the logarithmic sample complexity with respect to the ambient dimension under the three adversarial contamination models considered in the previous sections. In this experiment, we fix the corruption proportion to $\varepsilon = 0.1$ and consider three values of $p \in \{400, 800, 1600\}$. For each value of p, the experiment is repeated across multiple sample sizes n.

The sample sizes are chosen such that the quantity $n/\log(p)$ remains approximately constant across different values of p. Specifically, we consider:

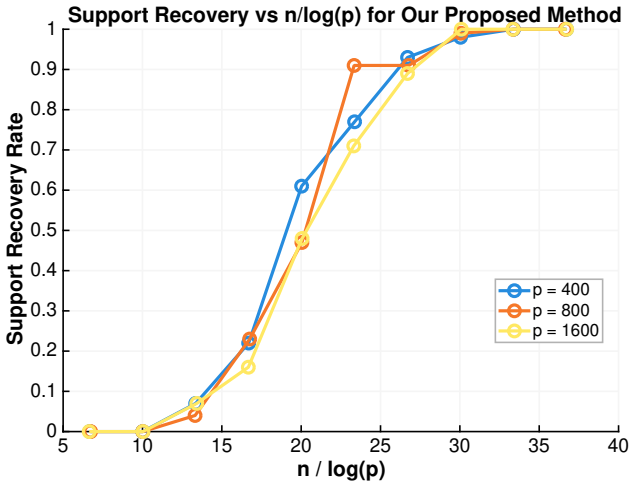
1. For $p = 400$, $n = \{40, 60, 80, 100, 120, 140, 160, 180, 200, 220\}$.
2. For $p = 800$, $n = \{45, 67, 89, 112, 134, 156, 179, 201, 223, 245\}$.
3. For $p = 1600$, $n = \{49, 74, 99, 123, 148, 172, 197, 222, 246, 271\}$.

Note that all settings correspond to the challenging high-dimensional regime since $n < p$ in all cases. For the above configurations, the corresponding values of $n/\log(p)$ are approximately $\{6.6, 10, 13.3, 16.6, 20, 23.3, 26.7, 30, 33.3, 36.7\}$.

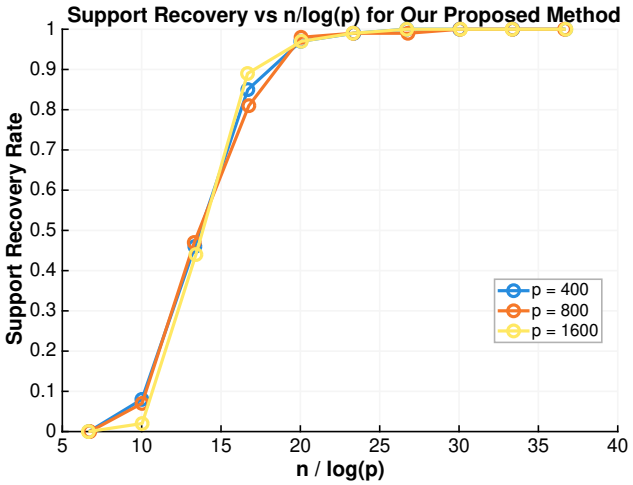
Each experiment is repeated 100 times to estimate the support recovery probability. The objective of this experiment is to examine whether the support recovery probability remains similar for identical values of $n/\log(p)$ across different values of dimension p. Such behavior would empirically validate that successful support recovery depends logarithmically on the dimension p, rather than exhibiting any other dependence, like polynomial or exponential. This is very critical for high-dimension.

We perform this evaluation under the different adversarial contamination models introduced earlier. The resulting support recovery plots are shown in Figure 5. In particular:

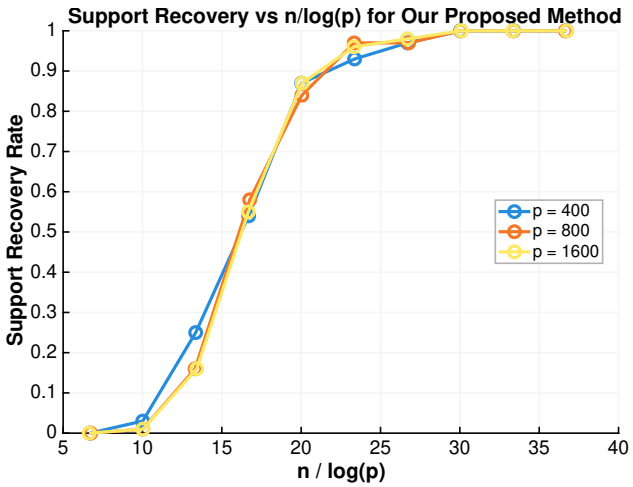
1. Random adversarial corruption (Figure 5a),
2. Strong adversarial contamination model of Chen et al. [2013] (Figure 5b),
3. Heavy-tailed corruption using the log-normal distribution (Figure 5c) and the Student's t-distribution (Figure 5d).



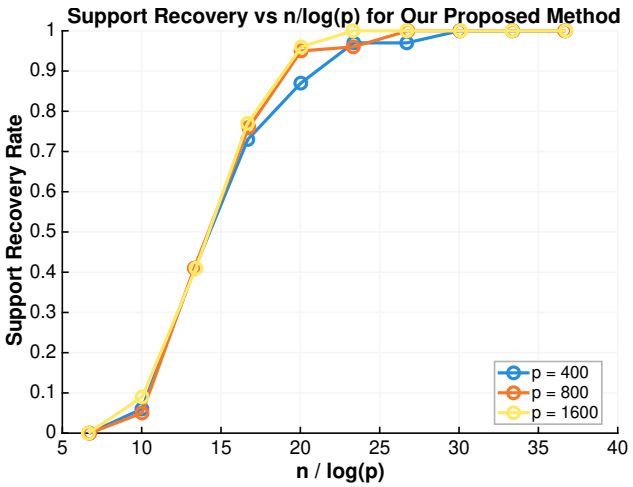
(a) Random adversarial corruption



(b) Strong adversarial corruption model by Chen et al. [2013]



(c) Heavy-tailed corruption using Log-Normal distribution



(d) Heavy-tailed corruption using Student's t-distribution

Figure 5: Support recovery probability versus $n/\log(p)$ under different adversarial corruption models for linear regression.

Across all settings, the curves corresponding to different values of $p \in \{400, 800, 1600\}$ largely overlap when plotted against $n/\log(p)$. This overlap of support recovery curves across different ambient dimensions provides strong empirical evidence that the required sample size for reliable support recovery scales proportionally with $\log(p)$, in agreement with our theoretical claims.

Having validated the theoretical claims for linear regression, we next examine whether similar behavior holds for logistic regression.

15 EXPERIMENTS: LOGISTIC REGRESSION

15.1 EXPERIMENTAL SETUP: DATA GENERATION PROCESS

We describe the data generation process first for logistic regression in a detailed manner as follows:

1. **Support Generation:** A support set $\mathcal{S}$ of size $k = 5$ is generated uniformly at random from $\{1, 2, \ldots, p\}$, where p denotes the number of features. The nonzero entries of the parameter vector $\mathbf{w}^\star_{\mathcal{S}}$ are generated by sampling uniformly from $[-1.5, -1] \cup [1, 1.5]$ to avoid values close to zero. For the non-support coordinates, we set $\mathbf{w}^\star_{\mathcal{S}^c} = 0$.
2. **Clean Data:** We next describe the generation of clean samples $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$. Each feature vector $\mathbf{x}^{(i)} \in \mathbb{R}^p$ is generated

independently from $\mathcal{N}(0, I_p)$. The response variable is generated according to the logistic regression model

$$\Pr(y^{(i)} = 1 \mid \mathbf{x}^{(i)}) = \frac{1}{1 + \exp(-\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)})},$$

and $y^{(i)} \in \{0, 1\}$ is sampled from the corresponding Bernoulli distribution.

3. **Adversarial Corruption:** Next, we generate samples corrupted by an adversary. The number of corrupted samples depends on the corruption proportion, while their construction depends on the adversarial corruption model. We consider the following corruption models:
   (a) Random adversarial corruption,
   (b) Strong adversarial corruption model,
   (c) Heavy-tailed adversarial corruption.

   The specific details for each corruption model are described in the later sections.
4. **Estimation:** The resulting dataset, consisting of both clean and adversarially corrupted samples, is provided to all algorithms for support estimation. We compare classical LASSO GLM [Friedman et al., 2010], Robust Logistic Regression (RoLR) [Feng et al., 2014], Trimmed MLE [Awasthi et al., 2022], SVAM [Mukhoty et al., 2023], and our proposed method.
5. **Evaluation Metrics:** We consider two metrics to evaluate the performance of all estimation methods:
   (a) Support recovery probability: Steps 1-4 are repeated multiple times to estimate the probability of exact support recovery, defined as the event $\mathcal{S}(\hat{\mathbf{w}}) = \mathcal{S}(\mathbf{w}^\star)$. The support recovery probability is computed as the fraction of successful recoveries over the total number of repetitions.
   (b) F1 score: We additionally evaluate partial support recovery using the F1 score between the estimated support $\mathcal{S}(\hat{\mathbf{w}})$ and the true support $\mathcal{S}(\mathbf{w}^\star)$. Let

   $$\text{Precision} = \frac{|\mathcal{S}(\hat{\mathbf{w}}) \cap \mathcal{S}(\mathbf{w}^\star)|}{|\mathcal{S}(\hat{\mathbf{w}})|}, \qquad \text{Recall} = \frac{|\mathcal{S}(\hat{\mathbf{w}}) \cap \mathcal{S}(\mathbf{w}^\star)|}{|\mathcal{S}(\mathbf{w}^\star)|}.$$

   The F1 score is defined as the harmonic mean of precision and recall:

   $$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

   The reported F1 score corresponds to the average across all repetitions.

15.2 ESTIMATION ALGORITHMS AND HYPERPARAMETERS

In this section, we describe the hyperparameter settings of the proposed and existing methods used in the evaluation to ensure reproducibility.

1. **LASSO GLM** [Friedman et al., 2010]: This corresponds to the classical LASSO estimator for generalized linear models. The algorithm utilizes all available samples during estimation. The method requires a regularization parameter, which we set to 0.05 in all experiments.
2. **Robust Logistic Regression (RoLR)** [Feng et al., 2014]: We implement Algorithm 1 from Feng et al. [2014]. We assume that the algorithm has access to the true hyperparameters n_1. Here, n_1 denotes the number of corrupted samples. Another threshold parameter T is chosen as $4\sqrt{\frac{\log(p)}{n} + \frac{\log(n)}{n}}$ as suggested in Algorithm 1 of Feng et al. [2014].
3. **Trimmed MLE** [Awasthi et al., 2022]: This approach performs estimation by trimming potentially corrupted samples. Accordingly, we provide the true corruption proportion ε to the algorithm. Following Theorem 5.3 of Awasthi et al. [2022], we set the convergence parameter $\eta = \varepsilon^2$ and $R = 2\,\|\beta^*\|$. Algorithm 2 in Awasthi et al. [2022] additionally requires an estimate of the covariance matrix. We use the sample covariance matrix so that the setting remains comparable with other estimation methods. The algorithm also has a preprocessing step (Algorithm 4 of Dong et al. [2019]), which involves the hyperparameters γ_1, γ_2, and C. In our experiments, we set $\gamma_1 = 0.1$, $\gamma_2 = 0.1$, and $C = 20$.

4. **SVAM** [Mukhoty et al., 2023]: We implement Algorithm 1 from Mukhoty et al. [2023], termed as Sequential Variance-Altered MLE (SVAM). The hyperparameters are selected following the publicly available implementation provided by the authors. Specifically, we use $\alpha = 0.3$, step size $\eta = 1.1$, and initialize the regression vector $\mathbf{w}_{\text{init}}$ with a random vector. The maximum number of iterations is set to 2500 as suggested by Mukhoty et al. [2023]. The parameter β is selected from the set $[0.01, 1, 10]$ as done in Mukhoty et al. [2023].

5. **Our Proposed Algorithm:** Our algorithm requires the corruption proportion like Feng et al. [2014], Awasthi et al. [2022] and hence we supply the true corruption proportion. For the medians of means based pre-processing step, we use the number of batches as $B = (\log(p/\delta))$, $T = 6\sqrt{\log(p/\varepsilon)}$, where $\delta = 10^{-3}$ is a constant and ε denotes the corruption proportion. We chose the regularization parameter $\lambda = 0.05$. We used maximum number of iterations as 40. We also had a convergence criteria to stop earlier if $\frac{\left\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\right\|_2}{\left\|\mathbf{w}^{(i)}\right\|_2} < 10^{-6}$.

15.3 RANDOM ADVERSARIAL CONTAMINATION IN BOTH FEATURES AND LABELS

In this section, we briefly describe the random adversarial contamination model and explain how corruption is introduced in both features and labels for logistic regression. Under this model, the corrupted feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^p$ are generated independently from $\mathcal{N}(\mathbf{0}, 5\mathbf{I}_p)$, while the corresponding corrupted responses $y^{(i)}$ are generated independently from a Bernoulli distribution with parameter $1/2$, i.e., $y^{(i)} \sim \text{Bernoulli}(1/2)$, independently of $\mathbf{x}^{(i)}$.

15.3.1 Low-Dimensional Regime

We focus on the low-dimensional regime with $n = 200$, $p = 30$, and sparsity level $k = 5$. We generate both clean and adversarially contaminated samples and use them as input to the estimation algorithms. The experiment is repeated 100 times to compute the empirical support recovery rate and F1 score. The results are reported in Table 17 and Table 18, respectively.

Table 17: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for low dimensional regime with $n = 200$, $p = 30$, and $k = 5$ under random adversarial corruption in Section 15.3.1. Our method delivers high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	Lasso GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.52 ± 0.505	0 ± 0	0 ± 0	0 ± 0	0.72 ± 0.454
0.1	0.36 ± 0.485	0 ± 0	0 ± 0	0 ± 0	0.54 ± 0.503
0.15	0.22 ± 0.418	0 ± 0	0 ± 0	0 ± 0	0.32 ± 0.471
0.2	0.06 ± 0.24	0 ± 0	0 ± 0	0 ± 0	0.16 ± 0.37

Table 18: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for low dimensional regime with $n = 200$, $p = 30$, and $k = 5$ under random adversarial corruption in Section 15.3.1. Our method delivers high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	Lasso GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.947 ± 0.062	0.313 ± 0.015	0.313 ± 0.021	0.338 ± 0.028	0.971 ± 0.05
0.1	0.901 ± 0.093	0.343 ± 0.023	0.312 ± 0.017	0.36 ± 0.031	0.95 ± 0.057
0.15	0.853 ± 0.127	0.37 ± 0.027	0.307 ± 0.015	0.37 ± 0.044	0.909 ± 0.077
0.2	0.772 ± 0.136	0.381 ± 0.035	0.311 ± 0.017	0.37 ± 0.046	0.834 ± 0.108

Form Table 17 and Table 18, we note that the F1 scores may be non-zero even when the support recovery rate is zero for methods like RoLR, Trimmed MLE, and SVAM.

Next, we proceed to the high-dimensional regime.

15.3.2 High-Dimensional Regime

We focus on the high-dimensional regime with $n = 600$, $p = 800$, and sparsity level $k = 5$. We generate both clean and adversarially contaminated samples and use them as input to the estimation algorithms. The experiment is repeated 100 times to compute the empirical support recovery rate and F1 score. The results are reported in Table 19 and Table 20, respectively. For both metrics, we observe that our method performs impressively across increasing corruption proportions, demonstrating its robustness.

Table 19: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under random adversarial corruption in Section 15.3.2. Our method delivers high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.9 ± 0.303	0 ± 0	0 ± 0	0 ± 0	0.96 ± 0.198
0.1	0.76 ± 0.431	0 ± 0	0 ± 0	0 ± 0	0.86 ± 0.351
0.15	0.56 ± 0.501	0 ± 0	0 ± 0	0 ± 0	0.92 ± 0.274
0.2	0.48 ± 0.505	0 ± 0	0 ± 0	0 ± 0	0.74 ± 0.443

Table 20: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under random adversarial corruption in Section 15.3.2. Our method delivers high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.991 ± 0.028	0 ± 0	0.023 ± 0	0.018 ± 0	0.996 ± 0.018
0.1	0.976 ± 0.044	0 ± 0	0.024 ± 0	0.018 ± 0	0.987 ± 0.032
0.15	0.951 ± 0.061	0 ± 0	0.026 ± 0.001	0.018 ± 0	0.99 ± 0.035
0.2	0.926 ± 0.088	0 ± 0	0.028 ± 0.001	0.017 ± 0	0.973 ± 0.046

In this section, we verify our theoretical guarantees for logistic regression in the high-dimensional regime. We considered the random adversarial contamination model, which does not correspond to adaptive or strong adversarial corruption.

Beyond handling high-dimensional settings, a key contribution of our work is robustness to adaptive/strong adversarial contamination. Accordingly, in the next section, we evaluate the performance of our proposed method and competing approaches under the adaptive/strong adversarial contamination model.

15.4 ADAPTIVE/STRONG ADVERSARIAL CONTAMINATION

In this section, we consider the adaptive/strong adversarial contamination model motivated from Chen et al. [2013] to avoid providing any unintended advantage to our proposed method due to the random adversarial corruption considered in the previous experiments. The adversarial contamination in Chen et al. [2013] proposed for linear regression is constructed using knowledge of the true support and the clean data, thereby explicitly targeting the support recovery task and making it significantly more challenging. We extend this adversarial contamination model for logistic regression.

Since the adversary is adaptive (strong), we assume it has knowledge of the true support and non-support. Let the support size be k. The adversary selects a set $\mathcal{S}_{adv}$ by choosing k coordinates uniformly at random from the true non-support.

Let the adversarial regression vector be denoted by $\mathbf{w}^{adv}$. The non-zero entries $\mathbf{w}^{adv}_{\mathcal{S}_{adv}}$ are sampled independently and uniformly from $[-1.5, -1] \cup [1, 1.5]$, ensuring that the magnitudes are bounded away from zero. For the remaining coordinates, we set $\mathbf{w}^{adv}_{\mathcal{S}^c_{adv}} = 0$.

Under this model, the corrupted feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^p$ are generated independently from $\mathcal{N}(\mathbf{0}, 5\mathbf{I}_p)$. The response variable follows the logistic regression model using $y^{(i)} = \text{sign}\left(\mathbf{x}^{(i)\intercal}\mathbf{w}^{adv}\right)$. Hence, the adversary has designed the

contamination by using a similar logistic model on different support. This may confuse the learning algorithm in selecting the original support.

We generate clean and adversarially contaminated samples under varying corruption proportions and use them as input to the estimation methods. The experiments are repeated 100 times to compute the empirical support recovery rate and F1 score. The results are reported in Table 21 and Table 22, respectively.

Across both metrics, our proposed method achieves consistently higher scores, highlighting its robustness to adversarial contamination.

Table 21: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the adaptive/strong adversarial corruption model in Section 15.4. Our method delivers high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.96 ± 0.198	0 ± 0	0 ± 0	0 ± 0	0.98 ± 0.141
0.1	0.62 ± 0.49	0 ± 0	0 ± 0	0 ± 0	0.92 ± 0.274
0.15	0.42 ± 0.499	0 ± 0	0 ± 0	0 ± 0	0.94 ± 0.24
0.2	0.02 ± 0.141	0 ± 0	0 ± 0	0 ± 0	0.46 ± 0.503

Table 22: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the adaptive/strong adversarial corruption model in Section 15.4. Our method delivers high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.996 ± 0.018	0 ± 0	0.023 ± 0	0.018 ± 0	0.998 ± 0.013
0.1	0.962 ± 0.052	0 ± 0	0.024 ± 0	0.018 ± 0	0.992 ± 0.026
0.15	0.932 ± 0.066	0 ± 0	0.026 ± 0	0.018 ± 0	0.994 ± 0.024
0.2	0.816 ± 0.101	0 ± 0	0.027 ± 0.001	0.018 ± 0	0.929 ± 0.081

In this section, we verified our theoretical guarantees under the adaptive/strong adversarial contamination model in the high-dimensional regime for logistic regression. Another key contribution of our work is robustness to heavy-tailed distributions, which was not considered either in the previous section or in the experiments based on the adaptive/strong adversarial model considered in this section. Therefore, in the next section, we evaluate our method to further validate the theoretical claims under heavy-tailed data distributions.

15.5 OUR METHOD CAN HANDLE HEAVY-TAILED NOISE

One of the key contributions of our work is robustness to heavy-tailed distributions requiring only bounded fourth-order moments. Therefore, in this section, we evaluate the performance of our proposed method and existing approaches when both the features and labels are adversarially contaminated using heavy-tailed distributions, such as the log-normal distribution and the Student's t-distribution.

15.5.1 Corruption based on Log-Normal distribution

We first briefly describe the data generating process. In the adversarial corruption setting, the corrupted samples are generated from a log-normal distribution, whose logarithm follows a normal distribution. Specifically, for any corrupted sample i, we generate using $\log\left(\mathbf{x}^{(i)}\right) \sim \mathcal{N}(\mathbf{0}, 5\mathbf{I})$. Consequently, the corrupted samples follow a heavy-tailed distribution with all moments finite. For labels, we use $y^{(i)} \sim \text{sign}\left(-\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}\right)$. Hence, the adversary is using logistic regression model with parameter vector $-\mathbf{w}^{\star}$. This may confuse the learning algorithm in selecting the original support. This strong contamination model motivated from Feng et al. [2014].

Next, we generate both clean and corrupted datasets under varying corruption proportions and provide that as an input to the estimation methods. The support recovery rate and F1 score achieved by existing methods and our proposed method are reported in Table 23 and Table 24, respectively.

Table 23: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the heavy-tailed corruption model in Section 15.5.1. Our method delivers high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.46 ± 0.503	0 ± 0	0 ± 0	0 ± 0	1 ± 0
0.1	0.32 ± 0.471	0 ± 0	0 ± 0	0 ± 0	0.98 ± 0.141
0.15	0.14 ± 0.351	0 ± 0	0 ± 0	0 ± 0	0.92 ± 0.274
0.2	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0.46 ± 0.503

Table 24: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the heavy-tailed corruption model in Section 15.5.1. Our method delivers high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.933 ± 0.08	0 ± 0	0.023 ± 0	0.018 ± 0	1 ± 0
0.1	0.895 ± 0.102	0 ± 0	0.024 ± 0.001	0.018 ± 0	0.998 ± 0.016
0.15	0.81 ± 0.129	0 ± 0	0.026 ± 0.002	0.017 ± 0	0.991 ± 0.03
0.2	0.608 ± 0.2	0 ± 0	0.027 ± 0.003	0.017 ± 0	0.923 ± 0.083

From Table 23 and Table 24, we observe that our method consistently achieves high support recovery rates and F1 scores across varying corruption levels for logistic regression. This demonstrates the robustness of the proposed approach to adversarial corruption arising from heavy-tailed distributions in the high-dimensional regime for logistic regression.

In the next subsection, we consider another corruption setting based on a different heavy-tailed distribution.

15.5.2 Corruption based on student t-student distribution

First, we briefly describe the adversarial contamination model. Any corrupted sample $\mathbf{x}^{(i)}$ is generated using the random variable $\mathbf{z}^{(i)}$ described as follows:

$$\mathbf{z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad v^{(i)} \sim \chi^2_{\mathrm{df}}, \quad \mathbf{x}^{(i)} = \frac{\mathbf{z}^{(i)}}{\sqrt{v_i/\mathrm{df}}},$$

where χ^2_{df} denotes the chi-square distribution with degrees of freedom $= \mathrm{df}$. We chose $\mathrm{df} = 5$ in our experiments. Hence, the fourth moment is bounded.

For labels, we use $y^{(i)} \sim \mathrm{sign}\left(-\mathbf{w}^{\star\intercal}\mathbf{x}^{(i)}\right)$. Hence, the adversary is using logistic regression model with parameter vector $-\mathbf{w}^\star$. This may confuse the learning algorithm in selecting the original support. This strong contamination model motivated from Feng et al. [2014].

Next, we generate both clean and corrupted datasets under varying corruption proportions and provide that as an input to the estimation methods. The support recovery rate and F1 score achieved by existing methods and our proposed method are reported in Table 25 and Table 26, respectively.

Table 25: Support recovery rates (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the heavy-tailed corruption model in Section 15.5.2. Our method delivers high support recovery across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.98 ± 0.141	0 ± 0	0 ± 0	0 ± 0	0.98 ± 0.141
0.1	0.88 ± 0.328	0 ± 0	0 ± 0	0 ± 0	0.96 ± 0.198
0.15	0.52 ± 0.505	0 ± 0	0 ± 0	0 ± 0	0.82 ± 0.388
0.2	0.04 ± 0.198	0 ± 0	0 ± 0	0 ± 0	0.36 ± 0.485

Table 26: F1 scores (mean $\pm$ standard deviation over 100 runs) of competing methods across varying corruption proportions (ε), for high dimensional regime with $n = 600$, $p = 800$, and $k = 5$ under the heavy-tailed corruption model in Section 15.5.2. Our method delivers high F1 scores across all corruption levels, demonstrating its robustness to increasing data corruption.

ε	LASSO GLM	RoLR	Trimmed MLE	SVAM	**Ours**
0.05	0.998 ± 0.013	0 ± 0	0.022 ± 0	0.017 ± 0	0.998 ± 0.013
0.1	0.989 ± 0.03	0 ± 0	0.022 ± 0	0.017 ± 0	0.996 ± 0.018
0.15	0.943 ± 0.066	0 ± 0	0.023 ± 0	0.017 ± 0	0.973 ± 0.068
0.2	0.794 ± 0.122	0 ± 0	0.023 ± 0	0.016 ± 0	0.879 ± 0.127

From Table 25 and Table 26, we observe that our method achieves high support recovery rates and F1 scores across all corruption proportions in the high-dimensional setting under heavy-tailed corruption generated using the Student's t-distribution. This further demonstrates the robustness of the proposed approach to adversarial contamination arising from heavy-tailed distributions such as the Student's t-distribution.

In the next section, we verify one of the main claims of our work about logarithmic sample complexity with respect to ambient dimension under adaptive/strong adversarial corruption, which is particularly important in high-dimensional settings.

15.6 EMPIRICAL VALIDATION OF OUR SAMPLE COMPLEXITY BEING LOGARITHMIC IN AMBIENT DIMENSION, MEANING $n = \Omega(\log(p))$

The experimental setup is very similar to the case of linear regression discussed in Section 15.6 but we describe it here in detail for completeness and reproducibility.

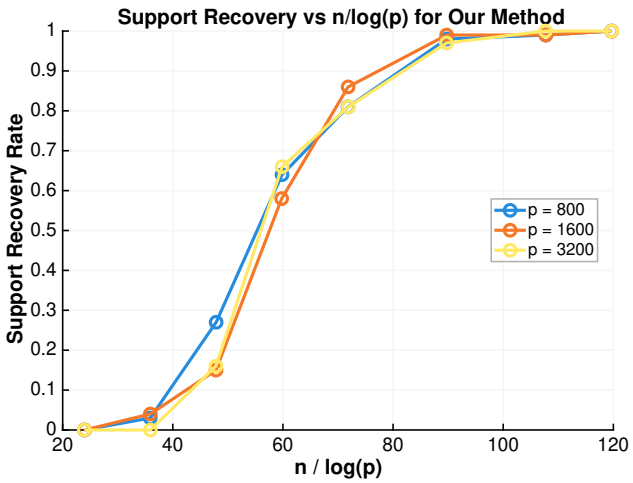
We empirically verify the logarithmic sample complexity with respect to the ambient dimension under the three adversarial contamination models considered in the previous sections. In this experiment, we fix the corruption proportion to $\varepsilon = 0.1$ and consider three values of $p \in \{800, 1600, 3200\}$. For each value of p, the experiment is repeated across multiple sample sizes n.

The sample sizes are chosen such that the quantity $n/\log(p)$ remains approximately constant across different values of p. Specifically, we consider:

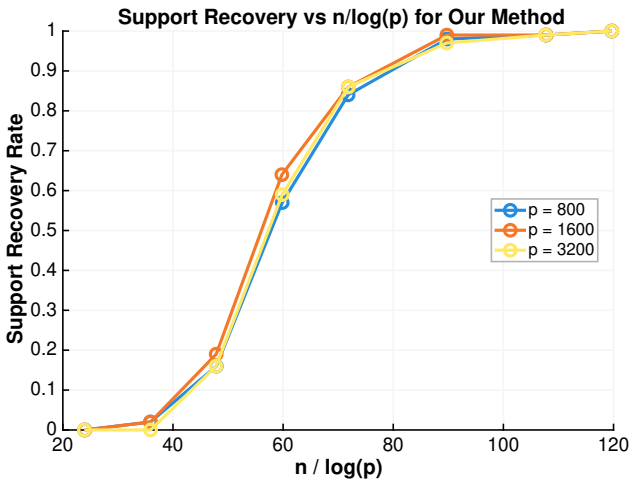
1. For $p = 800$, we $n = \{160, 240, 320, 400, 480, 600, 720, 790\}$.
2. For $p = 1600$, $n = \{177, 265, 353, 441, 530, 662, 795, 883\}$.
3. For $p = 3200$, $n = \{193, 290, 386, 483, 580, 724, 869, 966\}$.

Note that all settings correspond to the challenging high-dimensional regime since $n < p$ in all cases. For the above configurations, the corresponding values of $n/\log(p)$ are approximately $\{26.70, 40.05, 53.41, 66.76, 80.11, 100.14, 120.17, 133.52\}$.

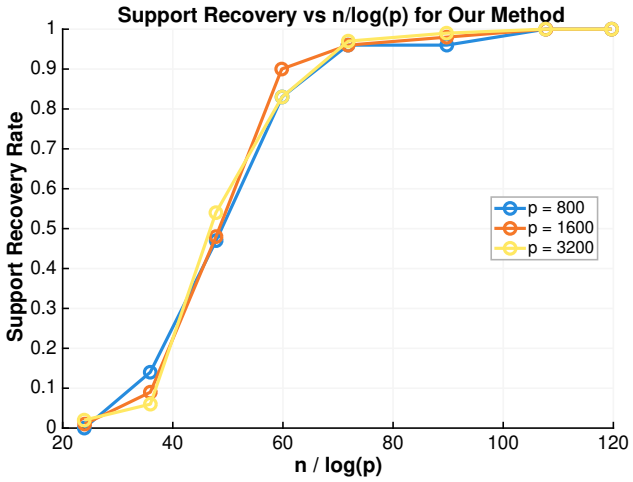
Each experiment is repeated 100 times to estimate the support recovery probability. The objective of this experiment is to examine whether the support recovery probability remains similar for identical values of $n/\log(p)$ across different values



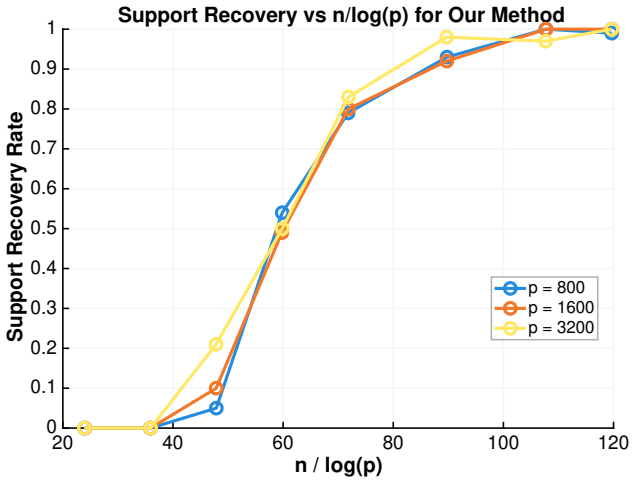
(a) Random adversarial corruption



(b) Strong adversarial corruption



(c) Heavy-tailed corruption using Log-Normal distribution



(d) Heavy-tailed corruption using Student-t distribution

Figure 6: Support recovery probability versus $n/\log(p)$ under different adversarial corruption models for logistic regression.

of dimension p. Such behavior would empirically validate that successful support recovery depends logarithmically on the dimension p, rather than exhibiting any other dependence, like polynomial or exponential. This is very critical for high-dimension.

We perform this evaluation under the different adversarial contamination models introduced earlier. The resulting support recovery plots are shown in Figure 5. In particular:

1. Support recovery plots under random adversarial corruption are presented in Figure 6a.
2. Support recovery plots for strong adversarial contamination model discussed in Section 15.4 is presented in Figure 6b.
3. Support recovery plots for heavy-tailed corruption using the log-normal distribution and the Student's t-distribution are presented in Figure 6c and Figure 6d respectively.

Across all settings, the curves corresponding to different values of $p \in \{800, 1600, 3200\}$ largely overlap when plotted against $n/\log(p)$. This overlap of support recovery curves across different ambient dimensions provides concrete empirical evidence that the required sample size for reliable support recovery scales proportionally with $\log(p)$, in agreement with our theoretical claims.

Hence, we have verified all the theoretical claims empirically for linear regression and logistic regression.

All the experiments were run on a machine with a 2.2 GHz Quad-Core Intel Core i7 processor with RAM of 16 GB.