# Confidence Intervals for Linear Models with Arbitrary Noise Contamination

Dong Xie[1], Chao Gao*[1], and John Lafferty[2]

[1] *University of Chicago*
[2] *Yale University*

**Abstract**

We study confidence interval construction for linear regression under Huber's contamination model, where an unknown fraction of noise variables is arbitrarily corrupted. While robust point estimation in this setting is well understood, statistical inference remains challenging, especially because the contamination proportion is not identifiable from the data. We develop a new algorithm that constructs confidence intervals for individual regression coefficients without any prior knowledge of the contamination level. Our method is based on a Z-estimation framework using a smooth estimating function. The method directly quantifies the uncertainty of the estimating equation after a preprocessing step that decorrelates covariates associated with the nuisance parameters. We show that the resulting confidence interval has valid coverage uniformly over all contamination distributions and attains an optimal length of order $O(1/\sqrt{n(1-\epsilon)^2})$, matching the rate achievable when the contamination proportion $\epsilon$ is known. This result stands in sharp contrast to the adaptation cost of robust interval estimation observed in the simpler Gaussian location model.

## 1 Introduction

This paper considers a linear model

$$y_i = X_i^T b + z_i, \qquad \text{for } i = 1, \cdots, n, \tag{1}$$

where the noise variables $\{z_i\}_{i=1}^n$ have an $\epsilon$ fraction of contamination. That is,

$$z_i \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon Q, \tag{2}$$

and they are independent from the $p$-dimensional feature vectors $\{X_i\}_{i=1}^n$. Our goal is to construct a confidence interval for a given coordinate of the regression vector $b \in \mathbb{R}^p$ without imposing any assumption on the contamination distribution $Q$.

The noise distribution (2) is a particular instance of Huber's contamination model (Huber, 1964). Since there is no assumption imposed on $Q$, the contaminated noise variables can just be regarded as arbitrary numbers, which implies the necessity of robust statistical inference. In the regression setting, the linear model (1) with noise following (2) and independent of the design matrix is known

*The research of CG is supported in part by NSF Grants ECCS-2216912 and DMS-2310769, and an Alfred Sloan fellowship.

as robust regression with oblivious contamination, and has been previously studied by Tsakonas et al. (2014); Bhatia et al. (2015, 2017); Suggala et al. (2019); Gao and Lafferty (2020); d'Orsi et al. (2021); d'Orsi et al. (2021). In particular, it was shown by d'Orsi et al. (2021); d'Orsi et al. (2021) that an M-estimator with Huber loss achieves the error bound

$$\|\widehat{b} - b\| = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n(1-\epsilon)^2}}\right). \tag{3}$$

Moreover, the rate (3) is optimal in the sense that a matching minimax lower bound can be established for some $Q$ (d'Orsi et al., 2021). A notable feature of the rate (3) is that consistent estimation is still possible even when the contamination proportion $\epsilon$ is close to 1. This is because in the setting of oblivious contamination, only the response variables $\{y_i\}_{i=1}^n$ are contaminated. In comparison, if the contamination affects both $\{y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$, the optimal estimation rate (3) becomes $\sqrt{\frac{p}{n}} + \epsilon$ (Gao, 2020), which implies inconsistency unless $\epsilon$ is a vanishing function of the sample size $n$.

Although global estimation of $b$ is well understood, the problem of optimal statistical inference for a given coordinate has remained open. When $\widehat{b}$ is computed by M-estimation, its asymptotic normality has been studied by Huber (1973); Yohai and Maronna (1979); Portnoy (1984, 1985); Mammen (1989) under appropriate conditions on sample size $n$ and dimension $p$. Recent developments using random matrix theory even obtain asymptotic distributions when $n$ and $p$ are proportional (Bean et al., 2013; El Karoui et al., 2013; Karoui, 2013; Donoho and Montanari, 2016; Lei et al., 2018). However, given the lack of assumptions on $Q$, deriving an asymptotic distribution under (2) would be impossible, even when $p = 1$.

Beyond M-estimation, another approach to constructing confidence intervals is to use high-probability error bounds for estimation. Given (3), the optimal error rate for estimating a given coordinate of $b$ is expected to be $\frac{1}{\sqrt{n(1-\epsilon)^2}}$. Suppose that one can construct a robust estimator such that

$$|\widehat{b}_1 - b_1| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\epsilon)^2}}\right). \tag{4}$$

Then, the interval $\left[\widehat{b}_1 - \frac{C}{\sqrt{n(1-\epsilon)^2}}, \widehat{b}_1 + \frac{C}{\sqrt{n(1-\epsilon)^2}}\right]$ is guaranteed to cover $b_1$ with some appropriate constant $C > 0$. However, this requires knowledge of the contamination proportion $\epsilon$, which is usually not available in practice. Additionally, the parameter $\epsilon$ is not identifiable in the setting of (2), and thus learning $\epsilon$ from data is impossible. This challenge of adaptive robust confidence interval construction has recently been studied by Luo and Gao (2024) in the simpler setting of a Gaussian location model. Given i.i.d. observations generated from $(1-\epsilon)N(\beta, 1) + \epsilon Q$ with some arbitrary $Q$, it is shown by Luo and Gao (2024) that the optimal length of a confidence interval covering the location parameter $\beta$ is of order $\frac{1}{\sqrt{\log n}} + \frac{1}{\sqrt{\log(1/\epsilon)}}$ when $\epsilon$ is unknown. In contrast, with the knowledge of $\epsilon$, the optimal length is of order $\frac{1}{\sqrt{n}} + \epsilon$. Therefore, ignorance of $\epsilon$ implies a significant adaptation cost in construction of robust confidence intervals. It is natural to suspect that such an adaptation cost is also present in the regression setting (1).

In this paper, we propose an algorithm that computes a confidence interval for a given coordinate of $b$. Our proposed algorithm does not depend on the knowledge of $\epsilon$. We show that the confidence interval not only has the coverage property, but its length is also bounded by $O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\epsilon)^2}}\right)$. According to the lower bound for estimation in d'Orsi et al. (2021), this is the optimal rate even if the value of $\epsilon$ is known. In other words, unlike the negative result of Luo and Gao (2024), there is no adaptation cost in robust confidence interval construction in the regression setting. The length

of the confidence interval automatically shrinks as the quality of the data increases (i.e. a smaller value of $\epsilon$). Moreover, the same theoretical guarantee holds if the Gaussian noise $N(0,1)$ in (2) is replaced by other non-degenerate distributions such as $t$ or Cauchy. Thus, our confidence interval is robust against both heavy tailed noise and arbitrary contamination.

Our algorithm is constructed from a one-dimensional estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} g(\widetilde{x}_i\widehat{b}_1 - \widetilde{y}_i)\widetilde{x}_i = 0, \tag{5}$$

where

$$g(t) = \tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}, \tag{6}$$

and $\widetilde{x}_i, \widetilde{y}_i \in \mathbb{R}$ are transformations of $(X_i, y_i)$ such that $\widetilde{y}_i \approx \widetilde{x}_i b_1 + z_i$ approximately holds. The details of constructing $(\widetilde{x}_i, \widetilde{y}_i)$ will be given in Algorithm 1. The interval estimator is then defined by

$$\left\{\beta : -\frac{C}{\sqrt{n}} \leqslant \frac{1}{n}\sum_{i=1}^{n} g(\widetilde{x}_i\beta - \widetilde{y}_i)\widetilde{x}_i \leqslant \frac{C}{\sqrt{n}}\right\}. \tag{7}$$

In other words, instead of quantifying the uncertainty of a point estimator, the interval (7) directly quantifies the uncertainty of a Z-estimation objective. This powerful idea was recently proposed by Chang and Kuchibhotla (2024). In the setting of robust regression, we are able to characterize the magnitude of $\frac{1}{n}\sum_{i=1}^{n} g(\widetilde{x}_i b_1 - \widetilde{y}_i)\widetilde{x}_i$ uniformly over all contamination distributions $Q$ in (2). Moreover, the choice of $C$ in (7) does not depend on $\epsilon$, so that (7) is adaptive to the unknown contamination distribution and the unknown contamination proportion.

The Z-estimation objective (5) uses the hyperbolic tangent function, which is analytically smooth. In fact, for the purpose of point estimation, replacing $g(\cdot)$ by the derivative of Huber loss would also work. However, by using the smoother hyperbolic tangent, we are able to control the model misspecification error in the approximation $\widetilde{y}_i \approx \widetilde{x}_i b_1 + z_i$ through a Taylor expansion.

It is interesting to note that (7) can also be derived from the general strategy of constructing adaptive robust confidence interval proposed by Luo and Gao (2024). Specifically, as long as one can construct a testing function that is able to distinguish between different values of the parameter together with different contamination proportions, the testing fuction can be inverted into a robust confidence interval that is adaptive to the unknown contamination level $\epsilon$. Although they only stated this framework for location families, it turns out that the same principle can be applied to the regression setting. In particular, one can construct an appropriate test whose inversion recovers the formula (7). We refer the reader to Section 3.5 for more details.

### 1.1 Related Work

Due to the lack of assumption on contamination, confidence interval construction under the noise (2) is required to be *nonparametric*, since there is no asymptotic distribution available. One class of nonparametric methods is based on the boot bootstrap. In particular, bootstrap in linear regression was developed in Freedman (1981), and its variants and theoretical properties were further investigated by Bickel and Freedman (1983); Mammen (1989, 1993). However, bootstrap consistency is usually based on the existence of an asymptotic distribution, which is still violated by (2).

There are several recently proposed methods that impose very few assumptions on the noise variables. The cyclic permutation test (Lei and Bickel, 2021) and the residual permutation test (Wen et al., 2025) only assume the noise variables to be exchangeable, and their tests can be inverted into confidence intervals. However, the theoretical properties under (2) are unknown, and

we find the two methods to be quite conservative in numerical experiments. The HulC procedure (Kuchibhotla et al., 2024) is a general black-box tool to quantify uncertainty of a nearly median-unbiased point estimator. In practice, a large sample size is needed for the confidence interval not to be too wide due to the use of sample splitting in its implementation. Finally, robust universal inference (Park et al., 2023) is a general framework of constructing robust confidence intervals using sample splitting. Applying this framework to a specific model requires the design of a specific testing procedure. Whether this could be done for a linear model with noise of the form (2) is an interesting question to be explored.

### 1.2 Paper Organization

We introduce the new algorithm and state its theoretical guarantee in Section 2. To understand why the proposed method works, we will first discuss the problem and illustrate the main idea for $p = 1$ in Section 3. Then in Section 4, we will discuss how to reduce the original problem to the one-dimensional setting through decorrelation. Extensive numerical experiments will be reported in Section 5. Finally, we collect all technical proofs in Section 6.

### 1.3 Notation

Define $[n] = \{1, \ldots, n\}$ for any positive integer $n$. For any two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there exist constants $c, C > 0$ such that $ca_n \leqslant b_n \leqslant Ca_n$ for all $n$; $a_n \lesssim b_n$ means that $a_n \leqslant Cb_n$ holds for some constant $C > 0$ independent of $n$. For a vector $v$, its $\ell_2$ norm is defined by $\|v\| = \sqrt{\sum_i v_i^2}$. The $p$-dimensional unit sphere is defined by $S_{p-1} = \{v \in \mathbb{R}^p : \|v\| = 1\}$. For a matrix $A$, its operator norm $\|A\|_{\text{op}}$ is defined by the largest singular value. For a set $B$, we use $|B|$ for its cardinality. The CDF of $N(0, 1)$ is denoted by $\Phi(\cdot)$. We use $\mathbb{E}$ and $\mathbb{P}$ for generic expectation and probability operators whenever the distribution is clear from the context.

## 2 Main Results

With $X_i = (x_i, w_i^T)^T$ and $b = (\beta, \theta^T)^T$, we rewrite the linear model (1) as

$$y_i = \beta x_i + \theta^T w_i + z_i, \tag{8}$$

where $\beta \in \mathbb{R}$ is the parameter of interest and $\theta \in \mathbb{R}^{p-1}$ collects all other regression coefficients. Define the Huber loss $\rho(\cdot)$ as

$$\rho(t) = \begin{cases} t^2 & |t| \leqslant 1 \\ 2|t| - 1 & |t| > 1. \end{cases}$$

An algorithm of computing a robust confidence interval of $\beta$ is given below. Algorithm 1 first computes $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^n$ by regressing $x_i$ and $y_i$ on $w_i$ respectively. Then, $\widetilde{x}_i$ and $\widetilde{y}_i$ are defined as the residuals. Since the covariates $\{(x_i, w_i)\}_{i=1}^n$ do not have contamination, regressing $x_i$ on $w_i$ can be done by ordinary least squares. On the other hand, regressing $y_i$ on $w_i$ involves the Huber loss given the presence of outliers in $\{y_i\}_{i=1}^n$. Once we obtain $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^n$, we apply the formula (7) with $C = 1.5\sqrt{\frac{\log(2/\alpha)}{n} \sum_{i=1}^n \widetilde{x}_i^2}$ to obtain the interval $[\widehat{\beta}_L, \widehat{\beta}_R]$.

To state the theoretical guarantee of Algorithm 1, we impose the following assumption on the random design.

**Assumption A.** *There exist constants $K_1, K_2, K_3 > 0$, such that for any $v \in S_{p-1}$, the random vector $X_i = (x_i, w_i^T)^T \in \mathbb{R}^p$ satisfies*

**Algorithm 1:** Constructing Confidence Interval for $\beta$

**Input** : $\{(x_i, w_i, y_i)\}_{i=1}^n$

**Output:** $\widehat{\beta}_L$ and $\widehat{\beta}_R$

$\widehat{\alpha} \leftarrow \arg\min_\alpha \frac{1}{n}\sum_{i=1}^n (x_i - \alpha^T w_i)^2$;

For $i \in [n]$,

$\widetilde{x}_i \leftarrow x_i - \widehat{\alpha}^T w_i$;

$\widehat{\gamma} \leftarrow \arg\min_\gamma \min_\beta \frac{1}{n}\sum_{i=1}^n \rho(y_i - \beta\widetilde{x}_i - \gamma^T w_i)$;

For $i \in [n]$,

$\widetilde{y}_i \leftarrow y_i - \widehat{\gamma}^T w_i$;

$\widehat{\beta}_L \leftarrow \inf\left\{\beta : \frac{\sum_{i=1}^n g(\widetilde{x}_i\beta - \widetilde{y}_i)\widetilde{x}_i}{\sqrt{\sum_{i=1}^n \widetilde{x}_i^2}} \geqslant -1.5\sqrt{\log(2/\alpha)}\right\}$;

$\widehat{\beta}_R \leftarrow \sup\left\{\beta : \frac{\sum_{i=1}^n g(\widetilde{x}_i\beta - \widetilde{y}_i)\widetilde{x}_i}{\sqrt{\sum_{i=1}^n \widetilde{x}_i^2}} \leqslant 1.5\sqrt{\log(2/\alpha)}\right\}$.

1. $\mathbb{P}\left(v^T X_i > t\right) = \mathbb{P}\left(v^T X_i < -t\right) \leqslant \exp\left(-\frac{t^2}{K_1^2}\right)$ *for all* $t > 0$.

2. $\mathbb{E}|v^T X_i|^2 \mathbb{1}\{|v^T X_i| \leqslant K_2\} \geqslant K_3^{-1}$.

In words, for each direction $v \in S_{p-1}$, the projection $v^T X_i$ is a symmetric sub-Gaussian random variable whose distribution is not degenerate in a neighborhood of zero.

**Theorem 1.** *Consider i.i.d. samples* $\{(x_i, w_i, y_i)\}_{i=1}^n$ *generated from the linear model (8), with* $\{(x_i, w_i)\}_{i=1}^n$ *and* $\{(z_i)\}_{i=1}^n$ *independent from each other and satisfying Assumption A and (2). For any* $\alpha \in (0,1)$*, there exist constants* $c > 0$ *and* $C > 0$ *depending on* $\alpha$*, such that the interval* $[\widehat{\beta}_L, \widehat{\beta}_R]$ *computed by Algorithm 1 satisfies*

$$\mathbb{P}\left(\beta \in [\widehat{\beta}_L, \widehat{\beta}_R] \text{ and } \widehat{\beta}_R - \widehat{\beta}_L \leqslant \frac{C}{\sqrt{n(1-\epsilon)^2}}\right) \geqslant 1 - \alpha,$$

*as long as* $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.

**Remark 1.** *The noise condition (2) assumes that* $z_i$*'s are generated from a standard* $N(0,1)$ *if they are not drawn from contamination. Our proofs would remain the same if* $N(0,1)$ *is replaced by any distribution satisfying* $\mathbb{P}(|z_i| \leqslant 1) \gtrsim 1$*, which includes most heavy tailed distributions such as t and Cauchy.*

Theorem 1 also implies that the point estimator $\widehat{\beta} = \frac{\widehat{\beta}_L + \widehat{\beta}_R}{2}$ achieves the error bound $O_{\mathbb{P}}\left(\frac{1}{\sqrt{n(1-\epsilon)^2}}\right)$, which is known to be optimal given the lower bound of estimation in d'Orsi et al. (2021). This immediately implies that the length of $[\widehat{\beta}_L, \widehat{\beta}_R]$ is also rate optimal, since a smaller interval length would imply an even better point estimator. Moreover, Algorithm 1 does not require the knowledge of $\epsilon$, and thus it is adaptive to the unknown contamination proportion. Since the lower bound of d'Orsi et al. (2021) holds even if the value of $\epsilon$ is given, Theorem 1 implies that there is no adaptation cost in terms of the rate. Compared with the necessity of adaptation cost under the Gaussian location model (Luo and Gao, 2024), the result of Theorem 1 is unexpected.

Assumption A on the covariate assumes symmetry and sub-Gaussianity. The symmetry condition can be avoided by incorporating a preprocessing step of computing the pairwise difference

$y_i - y_j = (X_i - X_j)^T\beta + (z_i - z_j)$. The new covariate $X_i - X_j$ certainly satisfies symmetry regardless of the distribution of $X_i$. The sub-Gaussianity condition is imposed for the convenience of certain technical arguments in the proof, and we do not think it is necessary for the result to hold. It can possibly be relaxed with some lengthier and more tedious arguments.

We finally give a brief discussion on the condition $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$ for some sufficiently small constant $c > 0$. When $\epsilon$ is bounded away from 1, the condition allows $p$ to be a growing function of $n$ as long as $n \gg p^2$. The requirement that the sample size needs to exceed squared degrees of freedom has been previously noted in the high-dimensional literature when signal is very sparse (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014), and can be shown to be necessary for interval estimation with parametric rate (Ren et al., 2015; Cai and Guo, 2017; Verzelen and Gassiat, 2018). When the signal is dense, however, $n \gg p^2$ is actually not needed for confidence interval construction when the data has no contamination (Wen et al., 2025). Whether or not $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$ can be further relaxed under Huber's contamination model (2) is an interesting theoretical question to be explored in the future.

In the next two sections, we will introduce the idea behind each step of Algorithm 1. We will first consider a simpler problem of $p = 1$ without any nuisance parameter in Section 3, and explain why the formula (7) has coverage property under (2) with arbitrary contamination. Section 4 will motivate the formulas of $\widetilde{x}_i$ and $\widetilde{y}_i$ and explain why Algorithm 1 works in the original model (8).

# 3 Adaptive Confidence Interval for Univariate Regression

In this section, we will introduce general methodology of confidence interval construction in a univariate version of (8) without nuisance parameter. The extension to the multivariate setting will be considered in Section 4. To be specific, we consider i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ that are generated by a linear model

$$y_i = \beta x_i + z_i, \tag{9}$$

where $x_i \sim N(0,1)$ and $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q$ are independent from each other. Our goal is to construct a confidence interval when the contamination proportion $\epsilon$ is unknown.

## 3.1 A Symmetric Location Model

To motivate the proposed method for (9), we first study a very simple location model with symmetric error,

$$y_i = \beta + z_i, \tag{10}$$

where $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q_0$ for some distribution $Q_0$ that is symmetric around zero. In other words, $z_i \sim Q_0$ implies $-z_i \sim Q_0$. We emphasize that such symmetry assumption is not assumed for the error distribution in (9).

Since $\beta$ is the center of a symmetric distribution regardless of what $Q_0$ is, a natural estimator for $\beta$ is the sample median $\widehat{\beta} = \text{Median}(\{y_i\}_{i=1}^n)$, which achieves the error bound[1]

$$\mathbb{P}\left(|\widehat{\beta} - \beta| \leqslant 3.5\frac{1}{\sqrt{n(1-\epsilon)^2}}\right) \geqslant 0.95.$$

[1]The non-asymptotic bound follows the same analysis of sample median in Chen et al. (2018). The explicit constant 3.5 can be obtained by assuming that $n(1-\epsilon)^2$ is sufficiently large. Moreover, the rate $\frac{1}{\sqrt{n(1-\epsilon)^2}}$ is minimax optimal given the existence of a symmetric distribution $Q_0$ such that $(1-\epsilon)N(0,1) + \epsilon Q_0 = N(0,(1-\epsilon)^{-2})$ (d'Orsi et al., 2021).

This high-probability error bound immediately implies that

$$\left[\widehat{\beta} - 3.5\frac{1}{\sqrt{n(1-\epsilon)^2}}, \widehat{\beta} + 3.5\frac{1}{\sqrt{n(1-\epsilon)^2}}\right] \tag{11}$$

is a valid confidence interval that has coverage probability at least 0.95. However, since (11) depends on the knowledge of $\epsilon$, it cannot be used when $\epsilon$ is unknown.

For any $s \in [0,1]$, define the left and right empirical quantile functions by

$$\begin{aligned} q_n^-(s) &= \inf\left\{t : \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{y_i \leqslant t\} \geqslant s\right\}, \\ q_n^+(s) &= \sup\left\{t : \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{y_i < t\} \leqslant s\right\}. \end{aligned}$$

Then, the sample median can be written as $\widehat{\beta} = \frac{q_n^-\left(\frac{1}{2}\right)+q_n^+\left(\frac{1}{2}\right)}{2}$. Instead of quantifying the uncertainty of the median as in (11), we could directly quantify the uncertainty of the quantile level by considering the interval

$$\left[q_n^-\left(\frac{1}{2} - \sqrt{\frac{\log(2/\alpha)}{2n}}\right), q_n^+\left(\frac{1}{2} + \sqrt{\frac{\log(2/\alpha)}{2n}}\right)\right]. \tag{12}$$

Interestingly, the interval (12) covers $\beta$ with probability at least $1-\alpha$ regardless of the value of $\epsilon$. Indeed, the event that the interval (12) covers the true $\beta$ in (10) is equivalent to

$$\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{z_i \leqslant 0\} \geqslant \frac{1}{2} - \sqrt{\frac{\log(2/\alpha)}{2n}} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{z_i < 0\} \leqslant \frac{1}{2} + \sqrt{\frac{\log(2/\alpha)}{2n}}, \tag{13}$$

which hold by Hoeffding's inequality using the symmetry of $z_i$. In fact, the random variable $\sum_{i=1}^n \mathbb{1}\{z_i \leqslant 0\}$ is almost distribution free, since it follows Binomial$(n, 1/2)$ unless $Q_0$ has a point mass at 0. One can even replace the $\sqrt{\frac{\log(2/\alpha)}{2n}}$ in (12) with an appropriate quantity computed from the CDF of Binomial$(n, 1/2)$ to achieve an almost exact coverage level $1-\alpha$.

The interval (12) was originally considered in Scheffe and Tukey (1945) as a nonparametric confidence interval for location. Though the definition of (12) does not depend on the knowledge of $\epsilon$, its length still adapts to the unknown $\epsilon$.

**Proposition 1.** *Consider i.i.d. samples from the location model with error distribution $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q_0$ for some distribution $Q_0$ that is symmetric around zero. For any $\alpha \in (0,1)$, there exist constants $c > 0$ and $C > 0$ depending on $\alpha$, such that the interval (12) covers $\beta$ with probability at least $1-\alpha$. Moreover, the length of the interval (12) is bounded by $\frac{C}{\sqrt{n(1-\epsilon)^2}}$ with probability at least $1-\alpha$, as long as $\frac{1}{\sqrt{n(1-\epsilon)^2}} \leqslant c$.*

### 3.2 Confidence Interval for Z-Estimation

Scheffe and Tukey's interval (12) is specifically designed for the inference of population quantiles. To extend the idea to the regression setting (9), it is helpful to understand (12) from a more general perspective.

To this end, we define

$$Z_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}(\beta - y_i).$$

The sample median can also be regarded as a Z-estimator, since it is an approximate solution to the equation $Z_n(\beta) = 0$. Using the objective function of Z-estimation, we can write (12) as

$$\left\{\beta : -\sqrt{\frac{2\log(2/\alpha)}{n}} \leqslant Z_n(\beta) \leqslant \sqrt{\frac{2\log(2/\alpha)}{n}}\right\}. \tag{14}$$

An advantage of the formula (14) is that it is applicable in settings beyond the one-dimensional location model (10), in particular when quantile function is not available. For example, for the regression model (9), one can simply use the same formula (14) with $Z_n(\beta)$ chosen to be an appropriate score function for regression. Construction of general confidence sets based on Z-estimation has recently been considered by Chang and Kuchibhotla (2024) even when the parameter of interest is multivariate. It turns out that this is a correct framework for us to derive robust confidence intervals for (9) when $\epsilon$ is unknown.

### 3.3 Application in Robust Regression

Return to the one-dimensional linear model (9). We will construct a confidence interval for $\beta$ using ideas developed in Section 3.1 and Section 3.2.

One straightforward way of analyzing (9) is to consider the ratio statistic

$$\frac{y_i}{x_i} = \beta + \frac{z_i}{x_i}. \tag{15}$$

Since the covariate $x_i$ follows $N(0,1)$ and the noise variable $z_i$ is independent from $x_i$, the random variable $\frac{z_i}{x_i}$ is symmetric around zero regardless of the noise distribution. Therefore, (15) is an instance of the symmetric location model (9). Even if the distribution of $x_i$ is not necessarily $N(0,1)$ or not even symmetric, one could still deal with a linear model with symmetric design by preprocessing the data with pairwise difference $y_i - y_j = \beta(x_i - x_j) + z_i - z_j$.

With the reduction to a symmetric location model, the confidence interval (12) can be applied with quantile functions computed from the ratio statistics $\{y_i/x_i\}_{i=1}^n$. Similar to Proposition 1, the coverage and length properties of this construction continue to hold for the regression model (9). Unfortunately, when it comes to the multivariate setting (8) with the presence of nuisance, the idea of taking ratio does not apply.

On the other hand, the framework of Z-estimation has more flexibility. A natural robust estimator for (9) is median regression which minimizes

$$\frac{1}{n}\sum_{i=1}^{n} |y - x_i\beta|. \tag{16}$$

The gradient of (16) is given by

$$Z_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}(x_i\beta - y_i)x_i. \tag{17}$$

Median regression can also be defined as a Z-estimator that solves the equation $Z_n(\beta) = 0$. Similar to (14), one can use the interval

$$\left\{\beta : -\frac{C}{\sqrt{n}} \leqslant Z_n(\beta) \leqslant \frac{C}{\sqrt{n}}\right\}, \tag{18}$$

with the function (17) for some constant $C$ that depends on the coverage probability. To choose $C$, we note that the event that the interval (18) contains the true $\beta$ is equivalent to

$$\left|\frac{1}{n}\sum_{i=1}^{n}\operatorname{sign}(z_i)x_i\right| \leqslant \frac{C}{\sqrt{n}}.$$

Since $\{x_i\}$ and $\{z_i\}$ are independent, the random variable $\frac{1}{n}\sum_{i=1}^{n}\operatorname{sign}(z_i)x_i$ is stochastically dominated by $N(0, n^{-1})$, and thus the Gaussian quantile $C = \Phi^{-1}(1-\alpha/2)$ guarantees coverage. More generally, one could also take $C = \sqrt{\frac{2}{n}\sum_{i=1}^{n} x_i^2 \log(2/\alpha)}$ so that the coverage guarantee holds beyond the Gaussian design by applying Hoeffding's inequality to the self-normalizing sum.

### 3.4 A Smooth Objective

For the general multivariate linear model (8), in order to estimate $\beta$ or construct its confidence interval, our strategy is to reduce (8) to the one-dimensional setting (9). This is implemented through a decorrelation technique to be introduced in Section 4. Our technique will construct pairs $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{n}$ from (8) such that

$$\widetilde{y}_i \approx \beta \widetilde{x}_i + z_i. \tag{19}$$

In other words, the pairs $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{n}$ follow the linear model (9) approximately. When applying (18) with $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{n}$, one also needs to account for the model misspecification error in (19) for the coverage property. With the Z-estimation objective given by (17), the effect of model misspecification error is very hard to control due to the nonsmoothness of $\operatorname{sign}(\cdot)$.

To overcome this difficulty, we consider a smooth version of (17), given by

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i\beta - y_i)x_i, \tag{20}$$

where $g(\cdot)$ is the hyperbolic tangent function defined by (6). We note that $g(\cdot)$ can be regarded as a smoothed sign function. The corresponding M-estimator of (20) minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \log\cosh(y_i - x_i\beta),$$

where $\cosh(t) = \frac{e^t + e^{-t}}{2}$. The function $\log\cosh(\cdot)$ is very similar to Huber loss since $\log\cosh(t) \sim \frac{t^2}{2}$ as $t \to 0$ and $\log\cosh(t) \sim |t|$ as $t \to \infty$. Unlike Huber loss, $\log\cosh(\cdot)$ has bounded derivatives of all orders. Thus, when applying (20) to (19), the effect of model misspecification can be controlled by Taylor expansion.

**Proposition 2.** *Consider i.i.d. samples $\{(x_i, y_i)\}_{i=1}^{n}$ generated from the linear model (9), with $x_i \sim N(0,1)$ and $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q$ independent from each other. For any $\alpha \in (0,1)$, there exist constants $c > 0$ and $C > 0$ depending on $\alpha$, such that the interval*

$$\left\{\beta : -\sqrt{\frac{2\log(2/\alpha)}{n}} \leqslant \frac{\frac{1}{n}\sum_{i=1}^{n} g(x_i\beta - y_i)x_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}} \leqslant \sqrt{\frac{2\log(2/\alpha)}{n}}\right\}, \tag{21}$$

*covers $\beta$ with probability at least $1-\alpha$. Moreover, the length of the interval (21) is bounded by $\frac{C}{\sqrt{n(1-\epsilon)^2}}$ with probability at least $1-\alpha$, as long as $\frac{1}{\sqrt{n(1-\epsilon)^2}} \leqslant c$.*

### 3.5 Why is Adaptation Possible?

The paper Luo and Gao (2024) considers confidence interval construction for

$$y_1, \cdots, y_n \overset{iid}{\sim} (1-\epsilon)N(\beta, 1) + \epsilon Q, \tag{22}$$

where $Q$ is some arbitrary contamination distribution. This model can be regarded as (10) without the symmetry condition on the error distribution. While the optimal length of confidence interval is of order $\frac{1}{\sqrt{n}} + \epsilon$ when $\epsilon$ is known, Luo and Gao (2024) proves that the optimal length slows down to $\frac{1}{\sqrt{\log(n)}} + \frac{1}{\sqrt{\log(1/\epsilon)}}$ when $\epsilon$ is unknown.

The adaptation cost caused by unknown $\epsilon$ in the setting of (22) can be explained by the difficulty of solving the following hypothesis testing problem,

$$\begin{aligned} H_0 &: y_1, \cdots, y_n \overset{iid}{\sim} P \in \{(1-\epsilon_{\max})N(\beta, 1) + \epsilon_{\max} Q : Q\}, \\ H_1 &: y_1, \cdots, y_n \overset{iid}{\sim} P \in \{(1-\epsilon)N(\beta + r, 1) + \epsilon Q : Q\}, \end{aligned} \tag{23}$$

where $\epsilon_{\max}$ is some constant that is strictly smaller than $1/2$. It was shown by Luo and Gao (2024) that an adaptive confidence interval under (22) directly leads to a solution to (23). To be specific, one can reject the null whenever $\beta$ is not contained in the interval. Then, as long as the parameter $r$ under the alternative is greater than the interval length, both Type-1 and Type-2 errors can be controlled. Consequently, the smallest $r$ such that (23) can be solved serves as a lower bound for the length of adaptive confidence interval under (22). When $r \lesssim \frac{1}{\sqrt{\log(n)}} + \frac{1}{\sqrt{\log(1/\epsilon)}}$, Luo and Gao (2024) showed that there exist $Q_0$ and $Q_1$ such that

$$(1-\epsilon_{\max})N(\beta, 1) + \epsilon_{\max} Q_0 \approx (1-\epsilon)N(\beta + r, 1) + \epsilon Q_1$$

in the sense that the total variation distance between the two product distributions is sufficiently small; it is impossible to distinguish null from alternative.

Analogously, there is also a hypothesis testing problem associated with the construction of adaptive confidence intervals under (9). Let us use $P_{\epsilon,\beta,Q}$ for the joint distribution of $(x_i, y_i)$ under the linear model (9). Recall that $x_i \sim N(0,1)$ and $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q$, and they are independent from each other. The testing problem is

$$\begin{aligned} H_0 &: (x_1, y_1), \cdots, (x_n, y_n) \overset{iid}{\sim} P \in \{P_{1,\beta,Q} : Q\}, \\ H_1 &: (x_1, y_1), \cdots, (x_n, y_n) \overset{iid}{\sim} P \in \{P_{\epsilon,\beta+r,Q} : Q\}. \end{aligned} \tag{24}$$

Here, we can take $\epsilon_{\max} = 1$ since $\beta$ is identifiable in (9). Unlike (23), there exists a testing function for (24) with small Type-1 and Type-2 error as long as $r \gtrsim \frac{1}{\sqrt{n(1-\epsilon)^2}}$, which exactly matches the minimax rate of estimating $\beta$ under (9).

It turns out that an informative testing statistic is the function $Z_n(\beta) = \frac{1}{n}\sum_{i=1}^n \text{sign}(x_i\beta - y_i)x_i$ in (17). Under null, $Z_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \text{sign}(z_i)x_i$ so that it has zero expectation. Under alternative, $Z_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \text{sign}(x_i r + z_i)x_i$, where $z_i \sim (1-\epsilon)N(0,1) + \epsilon Q$. Thus, the expectation of $Z_n(\beta)$ under alternative is at most

$$-(1-\epsilon)\mathbb{E}_{x,z \overset{iid}{\sim} N(0,1)}\left[\text{sign}(xr + z)x\right] \lesssim -(1-\epsilon)r.$$

Then, by rejecting the null whenever $Z_n(\beta) < -\frac{C}{\sqrt{n}}$, one can solve the testing problem (24) as long as the difference between means under null and alternative exceeds the noise level, which is

$(1-\epsilon)r \gtrsim \frac{1}{\sqrt{n}}$, equivalent to the condition $r \gtrsim \frac{1}{\sqrt{n(1-\epsilon)^2}}$. Similarly, the testing statistic (20) also works under the same condition.

Luo and Gao (2024) also shows that any testing procedure solving (24) can be inverted into an adaptive confidence interval for (9) by collecting all $\beta$ such that the null hypothesis in (24) is not rejected. The rejection region $Z_n(\beta) < -\frac{C}{\sqrt{n}}$ then immediately leads to the one-sided interval $\left\{\beta : Z_n(\beta) \geqslant -\frac{C}{\sqrt{n}}\right\}$. By replacing the $\beta + r$ in the alternative of (24) with $\beta - r$, one can also obtain $\left\{\beta : Z_n(\beta) \leqslant \frac{C}{\sqrt{n}}\right\}$ in a similar way. The intersection of the two intervals recovers (18). Using (20) instead of $Z_n(\beta)$ also recovers (21).

# 4 Extension to Multivariate Regression

For the general multivariate setting (8), our main idea to construct an adaptive confidence interval for $\beta$ is a reduction from (8) to (9) so that the methodology introduced in Section 3 can be applied.

## 4.1 Decorrelating Nuisance Parameters

One naive way of regarding (8) as a univariate model is to treat $\theta^T w_i + z_i$ as a noise variable and directly apply (21) to the pairs $\{(x_i, y_i)\}_{i=1}^n$. However, there are two problems with this approach. First, $\theta^T w_i + z_i$ is not independent from $x_i$. Second, the variance of $\theta^T w_i + z_i$ when $z_i$ is not contaminated is $O(\|\theta\|^2 + \sigma^2)$, which can be arbitrarily large.

To address these two issues, Algorithm 1 considers the transformation

$$\widetilde{x}_i = x_i - \widehat{\alpha}^T w_i, \tag{25}$$
$$\widetilde{y}_i = y_i - \widehat{\gamma}^T w_i, \tag{26}$$

so that (8) can be equivalently written as

$$\widetilde{y}_i = \beta \widetilde{x}_i + (\gamma - \widehat{\gamma})^T w_i + z_i, \tag{27}$$

where $\gamma$ stands for the vector $\beta\widehat{\alpha} + \theta$. Since $\widehat{\alpha}$ in (25) is the least-squares estimator computed from regressing $\{x_i\}_{i=1}^n$ on $\{w_i\}_{i=1}^n$, it leads to the following identity

$$\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i w_i = 0. \tag{28}$$

Even though $\widetilde{x}_i$ and $(\gamma - \widehat{\gamma})^T w_i + z_i$ are still not independent in (27), the independence between $\widetilde{x}_i$ and $z_i$ and the uncorrelatedness between $\widetilde{x}_i$ and $w_i$ are sufficient for us to apply (21). We emphasize that the uncorrelatedness condition (28) is only in the sample level, which turns out to be particularly convenient when bounding the model missspecification error. Moreover, the $\widehat{\gamma}$ in (26) a robust estimator for $\gamma = \beta\widehat{\alpha} + \theta$, which will ensure that the scale of the noise variable $(\gamma - \widehat{\gamma})^T w_i + z_i$ is well controlled.

Applying the formula (20) to $\{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^n$, we define

$$G_n(\beta) = \frac{1}{n}\sum_{i=1}^n g\left(\widetilde{x}_i\beta - \widetilde{y}_i\right)\widetilde{x}_i. \tag{29}$$

Similar to (21), we use the confidence interval

$$\left\{\beta : -1.5\sqrt{\frac{\log(2/\alpha)}{n}} \leqslant \frac{G_n(\beta)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}} \leqslant 1.5\sqrt{\frac{\log(2/\alpha)}{n}}\right\}, \tag{30}$$

which is exactly the interval $[\widehat{\beta}_L, \widehat{\beta}_R]$ computed by Algorithm 1. The constant 1.5 in (30) is chosen to be slightly greater than the $\sqrt{2}$ used in (21) to account for the model misspecification caused by $(\gamma - \widehat{\gamma})^T w_i$.

## 4.2 Coverage

In order that the interval (30) still has the coverage property, we need to understand the behavior of $G_n(\beta)$ when $\beta$ is the true parameter that generates the data. Plugging (27) into (29), we obtain

$$G_n(\beta) = -\frac{1}{n}\sum_{i=1}^n g\left((\gamma - \widehat{\gamma})^T w_i + z_i\right)\widetilde{x}_i.$$

To analyze the effect of model missspecification caused by $(\gamma - \widehat{\gamma})^T w_i$, we use the smoothness of $g(\cdot)$ and have

$$-G_n(\beta) - \frac{1}{n}\sum_{i=1}^n g(z_i)\widetilde{x}_i = \frac{1}{n}\sum_{i=1}^n g'(z_i)(\gamma - \widehat{\gamma})^T w_i \widetilde{x}_i + \frac{1}{2n}\sum_{i=1}^n g''(\xi_i)|(\gamma - \widehat{\gamma})^T w_i|^2 \widetilde{x}_i, \tag{31}$$

where $\xi_i$ is some number between $(\gamma - \widehat{\gamma})^T w_i + z_i$ and $z_i$. We will separately bound the two terms on the right hand side of (31).

For the first term, we have

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^n g'(z_i)(\gamma - \widehat{\gamma})^T w_i \widetilde{x}_i\right| \\
= & \left|\frac{1}{n}\sum_{i=1}^n \left(g'(z_i) - \mathbb{E}g'(z_i)\right)(\gamma - \widehat{\gamma})^T w_i \widetilde{x}_i\right| \qquad (32) \\
\leqslant & \|\widehat{\gamma} - \gamma\| \left\|\frac{1}{n}\sum_{i=1}^n \left(g'(z_i) - \mathbb{E}g'(z_i)\right)\widetilde{x}_i w_i\right\| \\
\lesssim & \|\widehat{\gamma} - \gamma\|\sqrt{\frac{p}{n}}, \qquad (33)
\end{aligned}$$

where the last bound holds with high probability. The equality (32) uses the identity $\frac{1}{n}\sum_{i=1}^n (\mathbb{E}g'(z_i))\widetilde{x}_i w_i = 0$, which is by (28) and the fact that $\{z_i\}_{i=1}^n$ are identically distributed. The last step (33) is by Lemma 5, which bounds the $\ell_2$ norm of a $p$-dimensional vector by $O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right)$ since each of its entry is of order $O_{\mathbb{P}}(n^{-1/2})$.

For the second term, we have

$$\begin{aligned}
& \left|\frac{1}{2n}\sum_{i=1}^n g''(\xi_i)|(\gamma - \widehat{\gamma})^T w_i|^2 \widetilde{x}_i\right| \\
\leqslant & \frac{1}{2n}\sum_{i=1}^n |(\gamma - \widehat{\gamma})^T w_i|^2 |\widetilde{x}_i| \qquad (34)
\end{aligned}$$

$$\leqslant \frac{\|\widehat{\gamma}-\gamma\|^2}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}|\widetilde{x}_i|w_i w_i^T\right\|_{\mathrm{op}}$$
$$\lesssim \|\widehat{\gamma}-\gamma\|^2, \tag{35}$$

with high probability. The inequality (34) is by the fact $\|g''\|_\infty \leqslant 1$ and the operator norm bound used in (35) is by Lemma 5.

Combining the bounds for the two terms on the right hand side of (31), we can write down the following result.

**Lemma 1.** *Under the setting of Theorem 1, for any constant $\delta \in (0,1)$, there exist some constants $c>0$ and $C>0$ depending on $\delta$ such that*

$$|G_n(\beta)| \leqslant \left|\frac{1}{n}\sum_{i=1}^{n} g(z_i)\,\widetilde{x}_i\right| + C\left(\|\widehat{\gamma}-\gamma\|\sqrt{\frac{p}{n}} + \|\widehat{\gamma}-\gamma\|^2\right),$$

*with probability at least $1-\delta$ as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.*

Therefore, to guarantee coverage, it suffices to show that

$$\left|\frac{1}{n}\sum_{i=1}^{n} g(z_i)\,\widetilde{x}_i\right| \leqslant 1.45\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2}\sqrt{\frac{\log(2/\alpha)}{n}}, \tag{36}$$

and

$$C\left(\|\widehat{\gamma}-\gamma\|\sqrt{\frac{p}{n}} + \|\widehat{\gamma}-\gamma\|^2\right) \leqslant \frac{1}{20}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2}\sqrt{\frac{\log(2/\alpha)}{n}}. \tag{37}$$

The first bound (36) can be established in a similar way to Proposition 2 using concentration of self-normalizing sums. The second bound (37) is implied by the following estimation error bound.

**Lemma 2.** *Under the setting of Theorem 1, for any constant $\delta \in (0,1)$, there exist some constants $c>0$ and $C>0$ depending on $\delta$ such that*

$$\|\widehat{\gamma}-\gamma\|^2 \leqslant C\frac{p}{n(1-\epsilon)^2},$$

*with probability at least $1-\delta$ as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.*

With Lemma 2, the bound (37) holds as long as $\frac{p^2}{n(1-\epsilon)^4}$ is sufficiently small. By Lemma 1, we immediately obtain the coverage guarantee of the confidence interval.

**Theorem 2.** *Under the setting of Theorem 1, for any constant $\alpha \in (0,1)$, the interval $[\widehat{\beta}_L, \widehat{\beta}_R]$ computed by Algorithm 1 satisfies*

$$\mathbb{P}\left(\beta \in [\widehat{\beta}_L, \widehat{\beta}_R]\right) \geqslant 1-\alpha,$$

*as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$ for some $c>0$ depending on $\alpha$.*

### 4.3 Length

To bound the length of the confidence interval (30), it suffices to show for any $\widetilde{\beta} \in \mathbb{R}$ such that $|\widetilde{\beta} - \beta| > r$, $\widetilde{\beta}$ does not belong to the set (30) in the sense that

$$\left| \frac{G_n(\widetilde{\beta})}{\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}} \right| > 1.5\sqrt{\frac{\log(2/\alpha)}{n}}. \tag{38}$$

This implies that the interval length is at most $2r$. By triangle inequality,

$$|G_n(\widetilde{\beta})| \geqslant |-G_n(\widetilde{\beta}) + G_n(\beta)| - |G_n(\beta)|. \tag{39}$$

It suffices to lower bound $|-G_n(\widetilde{\beta}) + G_n(\beta)|$ and upper bound $|G_n(\beta)|$. The latter is directly implied by the coverage guarantee that $\beta$ belongs to (30). That is,

$$|G_n(\beta)| \lesssim \frac{1}{\sqrt{n}}. \tag{40}$$

To lower bound $|-G_n(\widetilde{\beta}) + G_n(\beta)|$, we note that the function $g(\cdot)$ satisfies the inequality

$$(g(t+\Delta) - g(t))\Delta \geqslant c\left(|\Delta|^2 \wedge |\Delta|\right) \mathbb{1}\{|t| \leqslant 1\}. \tag{41}$$

To see why (41) is true, let us consider the two cases $|\Delta| \leqslant 1$ and $|\Delta| > 1$ separately. By symmetry, we could assume $\Delta \geqslant 0$ without loss of generality. When $|\Delta| \leqslant 1$, the function $g(\cdot)$ is locally linear so that $g(t+\Delta) - g(t) \gtrsim \Delta$. When $|\Delta| > 1$, the monotonicity of $g(\cdot)$ implies $g(t+\Delta) - g(t) \gtrsim 1$.

Applying (41) to the difference $|(-G_n(\widetilde{\beta}) + G_n(\beta))(\beta - \widetilde{\beta})|$, we have

$$\begin{aligned}
& |(-G_n(\widetilde{\beta}) + G_n(\beta))(\beta - \widetilde{\beta})| \\
= & \left| \frac{1}{n}\sum_{i=1}^n \left( g\left((\gamma - \widehat{\gamma})^T w_i + z_i + \widetilde{x}_i(\beta - \widetilde{\beta})\right) - g\left((\gamma - \widehat{\gamma})^T w_i + z_i\right)\right) \widetilde{x}_i(\beta - \widetilde{\beta}) \right| \\
\geqslant & c\frac{1}{n}\sum_{i=1}^n \left( |\widetilde{x}_i(\widetilde{\beta} - \beta)|^2 \wedge |\widetilde{x}_i(\widetilde{\beta} - \beta)| \right) \mathbb{1}\left\{ |(\gamma - \widehat{\gamma})^T w_i + z_i| \leqslant 1 \right\} \\
\geqslant & c\left( |\widetilde{\beta} - \beta| \wedge |\widetilde{\beta} - \beta|^2 \right) \frac{1}{n}\sum_{i=1}^n (|\widetilde{x}_i| \wedge \widetilde{x}_i^2) \mathbb{1}\left\{ |(\gamma - \widehat{\gamma})^T w_i + z_i| \leqslant 1 \right\}.
\end{aligned} \tag{42}$$

By Lemma 5, we have

$$\frac{1}{n}\sum_{i=1}^n (|\widetilde{x}_i| \wedge \widetilde{x}_i^2) \mathbb{1}\left\{ |(\gamma - \widehat{\gamma})^T w_i + z_i| \leqslant 1 \right\} \gtrsim 1 - \epsilon, \tag{43}$$

with high probability. Intuitively, with the norm of $\gamma - \widehat{\gamma}$ controlled by Lemma 2, the random variable $\mathbb{1}\left\{ |(\gamma - \widehat{\gamma})^T w_i + z_i| \leqslant 1 \right\}$ has expectation at least of order $1 - \epsilon$ by (2), and thus (43) holds. Combining (42) and (43), we have

$$|-G_n(\widetilde{\beta}) + G_n(\beta)| \gtrsim (1-\epsilon)\left(1 \wedge |\widetilde{\beta} - \beta|\right). \tag{44}$$

Together with (39) and (40), the inequality (38) holds as long as $|\widetilde{\beta} - \beta| \gtrsim \frac{1}{\sqrt{n(1-\epsilon)^2}}$.

**Theorem 3.** *Under the setting of Theorem 1, for any constant $\alpha \in (0,1)$, there exist some constants $c > 0$ and $C > 0$ depending on $\alpha$ such that the interval $[\widehat{\beta}_L, \widehat{\beta}_R]$ computed by Algorithm 1 satisfies*

$$\mathbb{P}\left(\widehat{\beta}_R - \widehat{\beta}_L \leqslant \frac{C}{\sqrt{n(1-\epsilon)^2}}\right) \geqslant 1-\alpha,$$

*as long as* $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.

# 5 Numerical Experiments

## 5.1 Numerical Setups

In this section, we evaluate the performance of Algorithm 1, together with several competitors, in the following synthetic data sets. The observations $\{(y_i, X_i)\}_{i=1}^n$ are generated according to the linear model $y_i = X_i^T b + z_i$. The covariates are sampled from $X_i \overset{iid}{\sim} N(0, \Sigma)$ with $\Sigma_{jk} = 0.6^{|j-k|}$. We consider two types of noise distributions:

1. **Gaussian**: $z_i \overset{iid}{\sim} (1-\epsilon)N(0,1) + \epsilon\left(\frac{1}{2}N(10^2,1) + \frac{1}{2}N(10^4,1)\right)$.
2. **Cauchy**: $z_i \overset{iid}{\sim} (1-\epsilon)C(0,1) + \epsilon\left(\frac{1}{2}N(10^2,1) + \frac{1}{2}N(10^4,1)\right)$.

The Cauchy distribution $C(0,1)$ has density $\frac{1}{\pi(1+t^2)}$, and thus the second setting above involves both heavy tail and outliers. We set $p = 20$ and vary $n \in \{200, 1000\}$ and $\epsilon \in [0, 0.8]$. Due to translation invariance of the linear model, we set $b = 0$ without loss of generality.

The threshold level $1.5\sqrt{\log(2/\alpha)}$ used in Algorithm 1 is to guarantee non-asymptotic coverage in Theorem 1. One may wonder whether replacing $1.5\sqrt{\log(2/\alpha)}$ with the more aggressive Gaussian quantile $\Phi^{-1}(1-\alpha/2)$ works better in practice. While the Gaussian approximation works in the asymptotics of $\frac{p^2}{n(1-\epsilon)^4} \to 0$, our numerical results indicates that it should be used with caution when the sample size $n$ is small or $\epsilon$ is large. In addition to this variation of Algorithm 1, we also implement several methods in the literature. This includes the residual bootstrap (RB) of Freedman (1981), the restructured regression (RR) of Zhu and Bradic (2018), the cyclic permutation test (CPT) of Lei and Bickel (2021), the convex hull method (HulC) of Kuchibhotla et al. (2024) and the residual permutation test (RPT) of Wen et al. (2025). We also report the performance of the classical confidence interval computed from t-statistic centering at the ordinary least-squares estimator (OLS). Among the above methods, RR (Zhu and Bradic, 2018), CPT (Lei and Bickel, 2021) and RPT (Wen et al., 2025) are originally proposed to test the null hypothesis that a coefficient is zero, which also lead to confidence intervals by inverting the tests.

## 5.2 Implementation Details

The residual bootstrap (Freedman, 1981) is implemented with Huber regression

$$\min_b \frac{1}{n}\sum_{i=1}^n \rho(y_i - X_i^T b), \tag{45}$$

where $\rho(\cdot)$ is the same Huber loss used in Algorithm 1. The boostrap sample size is taken as 200. The restructured regression method has two versions in Zhu and Bradic (2018), and we implement the one with the knowledge of the design covariance. The cyclic permutation test (Lei and Bickel, 2021)

is implemented with the recommended preordering computed using a genetic algorithm. The genetic algorithm is computed with 1000 random samples. The HulC method (Kuchibhotla et al., 2024) is also implemented with the Huber regression (45), together with a recommended randomized choice of sample splitting to prevent the interval from being overly conservative. The residual permutation test (Wen et al., 2025) is implemented with the more powerful version using the empirical p-values, and the tuning parameters are set as recommended in the original work.

### 5.3 Results and Discussion

We report the coverage and length properties of confidence interval construction for the first regression coefficient with level $1 - \alpha = 0.95$. Figure 1 reports the coverage probabilities, and Figure 2 reports the average lengths of the intervals with coverage. Each point in the plots is computed from 500 independent experiments. Besides RB, RR, CPT, HulC, RPT and OLS, our methods are denoted as Alg1 for the original version of Algorithm 1 and Alg1-GA for the variant of Algorithm 1 with $1.5\sqrt{\log(2/\alpha)}$ replaced by $\Phi^{-1}(1-\alpha/2)$ using Gaussian approximation. Since CPT requires $n/p \geqslant \alpha^{-1} - 1$ (Lei and Bickel, 2021), it is not included in the setting with $n = 200$ and $p = 20$.

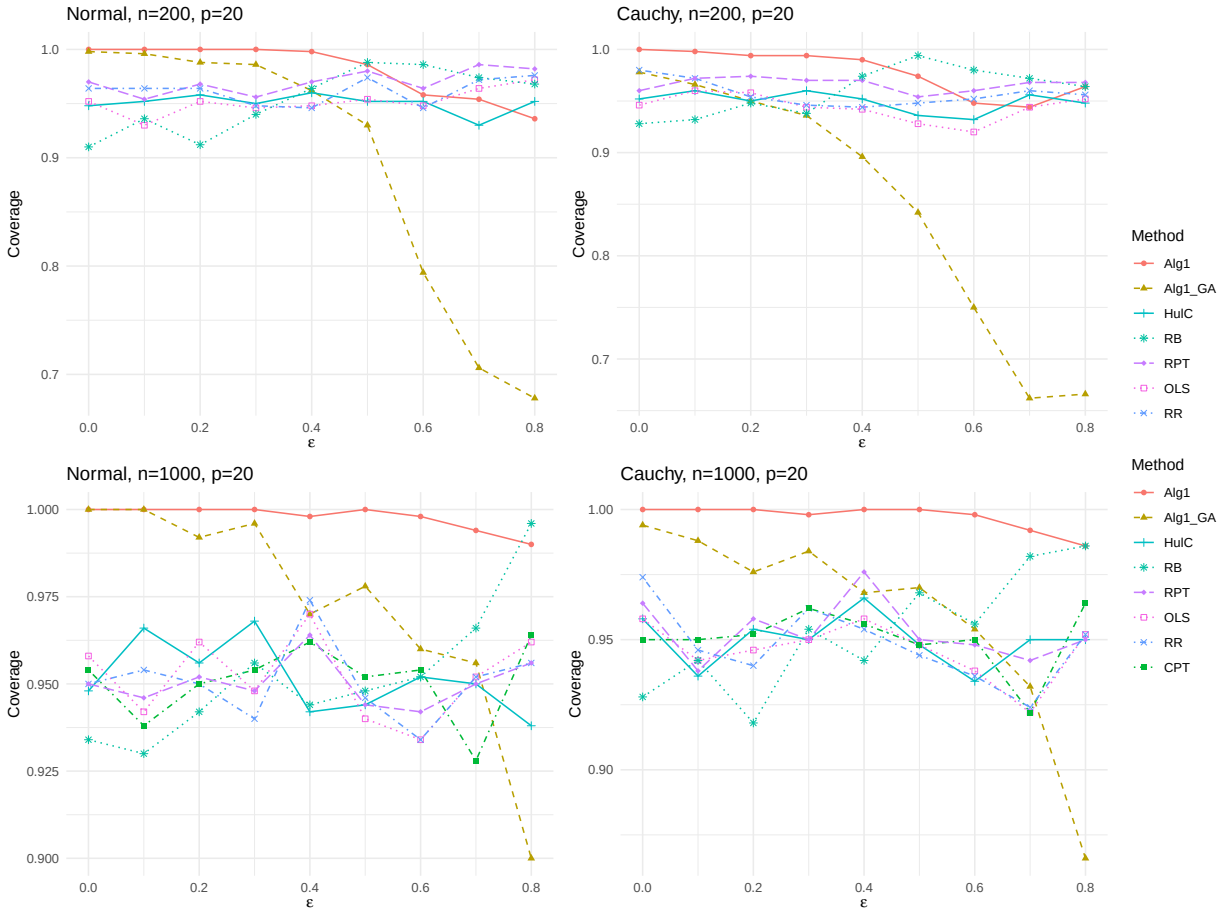


**Figure 1:** Coverage against $\epsilon$ across settings.

Figure 1 shows that most methods achieve coverage at levels around 0.95. In particular, since the proposed Algorithm 1 (Alg1) is designed to have non-asymptotic coverage guarantee, its coverage probability as a function of $\epsilon$ consistently stays above the nominal level, while other methods all

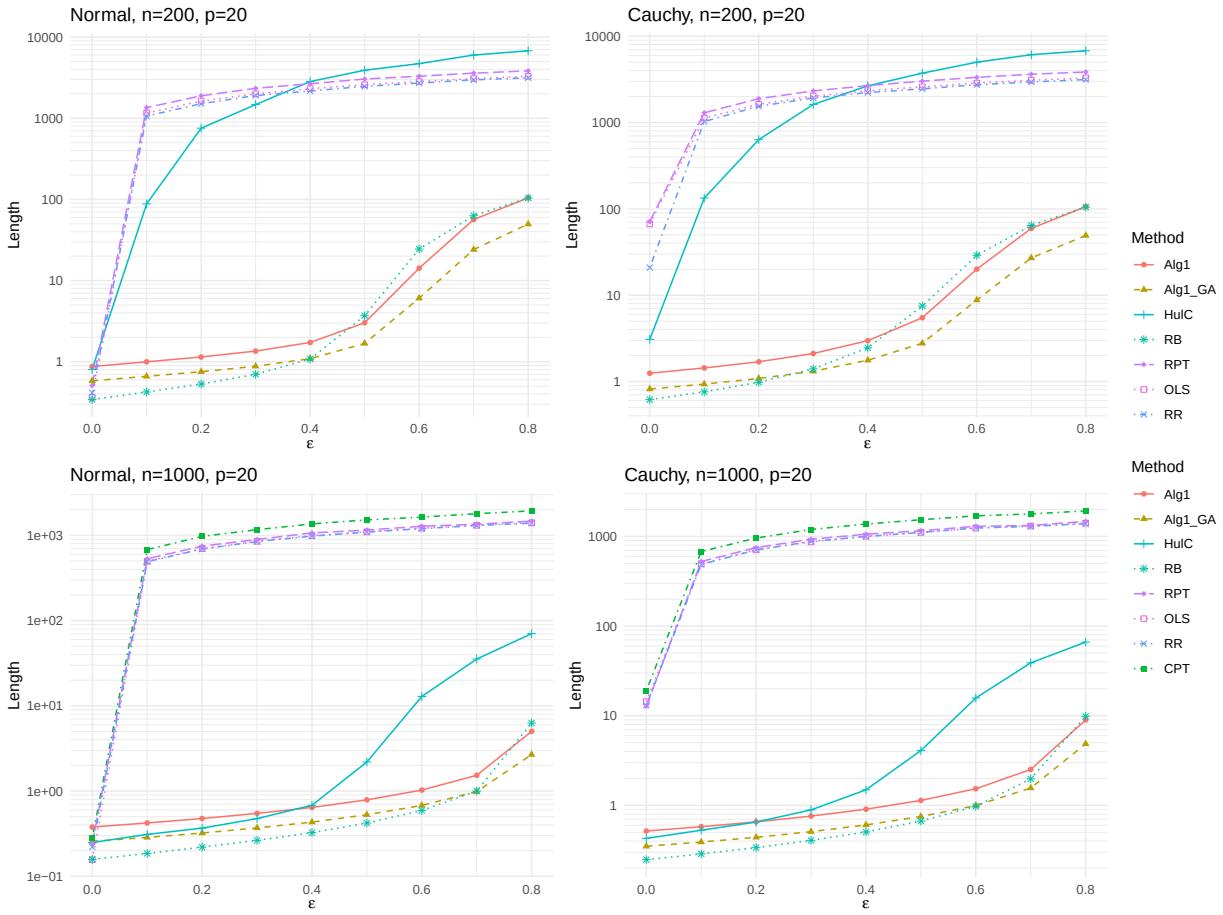


**Figure 2:** Average interval length against $\epsilon$ across settings.

have fluctuations. In comparison, Alg1-GA, which is the variant of Algorithm 1 with $1.5\sqrt{\log(2/\alpha)}$ replaced by $\Phi^{-1}(1-\alpha/2)$, has coverage below the nominal level for large values of $\epsilon$. This indicates that Gaussian approximation does not hold when the data has a large proportion of outliers, and the threshold level $1.5\sqrt{\log(2/\alpha)}$ based on non-asymptotic concentration is necessary.

Figure 2 shows that the length properties are drastically different across different methods. The performances of restructured regression (RR), cyclic permutation test (CPT) and residual permutation test (RPT) are similar to that of OLS. These methods do not seem to be robust, given the interval lengths reported in settings with large $\epsilon$ or Cauchy noise. Though CPT and RPT are proved to achieve coverage properties even when $\epsilon = 1$ (Lei and Bickel, 2021; Wen et al., 2025), the length guarantees require a moment condition (Wen et al., 2025) and our experiments show that they do not work with arbitrary outliers. The convex hull method (HulC) performs reasonably well in the setting of $n = 1000$ and $p = 20$ for $\epsilon$'s that are not too large. However, when $n = 200$, it is not competitive due to the necessary sample splitting step in the interval construction. The classical residual bootstrap (RB) works surprisingly well across all settings. It is both efficient when $\epsilon$ is small and robust when $\epsilon$ is large, even though there is not theoretical guarantee for residual bootstrap and we do not think it is a provable method for arbitrary contamination distribution. Compared with Algorithm 1, the residual bootstrap tends to undercover even when $\epsilon$ is very small. The most severe drawback of residual bootstrap is its computational cost, which is quite typical for a method based on resampling. Its computational time is around 25–180 times that of Algorithm 1

depending on the outlier configuration.

In summary, we conclude that Algorithm 1 behaves consistently well in all settings in terms of both coverage and length properties, which agrees with the theoretical guarantee established in Theorem 1. Compared with residual bootstrap and other resampling algorithms, Algorithm 1 has far better computational cost, which makes it more scalable for large data sets.

# 6 Proofs

This section collects all technical proofs. We will first derive error bounds of $\widehat{\alpha}$ and $\widehat{\gamma}$ in Section 6.1. The main results (Theorem 2, Theorem 3 and Theorem 1) will be proved in Section 6.2. We will then prove Proposition 1 and Proposition 2 in Section 6.3. Finally, additional technical lemmas will be proved in Section 6.4.

## 6.1 Bounding Errors of $\widehat{\alpha}$ and $\widehat{\gamma}$

Writing the joint covariance of $X_i = (x_i, w_i^T)^T$ as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xw} \\ \Sigma_{wx} & \Sigma_{ww} \end{pmatrix}.$$

The least-squares estimator $\widehat{\alpha}$ can be regarded as the sample version of $\alpha = \Sigma_{ww}^{-1}\Sigma_{wx}$.

**Lemma 3.** *Under the setting of Theorem 1, for any constant $\delta \in (0,1)$, there exist some constants $c > 0$ and $C > 0$ depending on $\delta$ such that*

$$\|\widehat{\alpha} - \alpha\| \leqslant C\sqrt{\frac{p}{n}},$$

*with probability at least $1-\delta$ as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.*

*Proof.* Assumption A implies $\|\Sigma\|_{\text{op}} \leqslant 2K_1^2$ and $\left\|\Sigma^{-1}\right\|_{\text{op}} \leqslant K_3$. Define

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T = \begin{pmatrix} \widehat{\Sigma}_{xx} & \widehat{\Sigma}_{xw} \\ \widehat{\Sigma}_{wx} & \widehat{\Sigma}_{ww} \end{pmatrix}.$$

Given the sub-Gaussianity in Assumption A, we have $\left\|\widehat{\Sigma} - \Sigma\right\|_{\text{op}} \leqslant C_1\sqrt{\frac{p}{n}}$ with high probability (see Lemma 9 stated in Section 6.4). Since $\widehat{\alpha} = \widehat{\Sigma}_{ww}^{-1}\widehat{\Sigma}_{wx}$, we have

$$\|\widehat{\alpha} - \alpha\| \leqslant \left\|\widehat{\Sigma}_{ww}^{-1}\right\|_{\text{op}} \|\widehat{\Sigma}_{wx} - \Sigma_{wx}\| + \left\|\widehat{\Sigma}_{ww}^{-1} - \Sigma_{ww}^{-1}\right\|_{\text{op}} \|\Sigma_{wx}\|,$$

where $\|\widehat{\Sigma}_{wx} - \Sigma_{wx}\| \leqslant \left\|\widehat{\Sigma} - \Sigma\right\|_{\text{op}} \leqslant C_1\sqrt{\frac{p}{n}}$ and $\|\Sigma_{wx}\| \leqslant \|\Sigma\|_{\text{op}} \leqslant 2K_1^2$. Since $\left\|\widehat{\Sigma}_{ww} - \Sigma_{ww}\right\|_{\text{op}} \leqslant \left\|\widehat{\Sigma} - \Sigma\right\|_{\text{op}} \leqslant C_1\sqrt{\frac{p}{n}}$, we know that both $\left\|\widehat{\Sigma}_{ww}\right\|_{\text{op}}$ and $\left\|\widehat{\Sigma}_{ww}^{-1}\right\|_{\text{op}}$ are bounded by constants. Therefore, $\left\|\widehat{\Sigma}_{ww}^{-1} - \Sigma_{ww}^{-1}\right\|_{\text{op}} \leqslant \left\|\widehat{\Sigma}_{ww}^{-1}\right\|_{\text{op}} \left\|\widehat{\Sigma}_{ww} - \Sigma_{ww}\right\|_{\text{op}} \left\|\Sigma_{ww}^{-1}\right\|_{\text{op}} \leqslant C_2\sqrt{\frac{p}{n}}$. Combining these bounds, we obtain the desired rate for $\|\widehat{\alpha} - \alpha\|$. □

Next, we will prove Lemma 2 that bounds the error of $\widehat{\gamma}$. This requires the following lemma on the random vector $\widetilde{X}_i = (\widetilde{x}_i, w_i^T)^T \in \mathbb{R}^p$.

**Lemma 4.** *Under the setting of Theorem 1, for any constant $\delta \in (0,1)$, there exist some constants $C, c > 0$ depending on $\delta$ such that*

1. $\left\| \frac{1}{n}\sum_{i=1}^n \rho'(z_i)\widetilde{X}_i \right\| \leqslant C\sqrt{\frac{p}{n}}$,
2. $\inf_{\|v\|=1} \frac{1}{n}\sum_{i=1}^n (|v^T\widetilde{X}_i|^2 \wedge |v^T\widetilde{X}_i|)\mathbb{1}\{|z_i| \leqslant 1\} \geqslant c(1-\epsilon)$,

*with probability at least $1-\delta$.*

The proof of Lemma 4 will be given in Section 6.4. Now we are ready to prove Lemma 2.

*Proof of Lemma 2.* Recall that $\gamma = \beta\widehat{\alpha} + \theta$, and we can write $y_i = \beta\widetilde{x}_i + \gamma^T w_i + z_i$. Define $\eta = (\beta, \gamma^T)^T \in \mathbb{R}^p$, and the model becomes

$$y_i = \widetilde{X}_i^T \eta + z_i.$$

Similarly, define $\widehat{\eta} = (\widehat{\beta}, \widehat{\gamma}^T)^T \in \mathbb{R}^p$, where $\widehat{\beta}$ and $\widehat{\gamma}$ solve the optimization $\min_{\beta,\gamma} \frac{1}{n}\sum_{i=1}^n \rho(y_i - \beta\widetilde{x}_i - \gamma^T w_i)$. Since $\|\widehat{\gamma} - \gamma\| \leqslant \|\widehat{\eta} - \eta\|$, it suffices to bound $\|\widehat{\eta} - \eta\|$. The definition of $\widehat{\eta}$ implies

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^n \rho'(y_i - \widetilde{X}_i^T\widehat{\eta})\widetilde{X}_i \\
&= \frac{1}{n}\sum_{i=1}^n \rho'(\widetilde{X}_i^T(\eta - \widehat{\eta}) + z_i)\widetilde{X}_i \\
&= \frac{1}{n}\sum_{i=1}^n \rho'(z_i)\widetilde{X}_i + \frac{1}{n}\sum_{i=1}^n (\rho'(\widetilde{X}_i^T(\eta - \widehat{\eta}) + z_i) - \rho'(z_i))\widetilde{X}_i. \qquad (46)
\end{aligned}$$

It is straightforward to check that $\rho'(\cdot)$ also satisfies the inequality (41). That is,

$$(\rho'(t+\Delta) - \rho'(t))\Delta \geqslant c\left(|\Delta|^2 \wedge |\Delta|\right)\mathbb{1}\{|t| \leqslant 1\},$$

for some constant $c > 0$. Together with Lemma 4, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (\rho'(\widetilde{X}_i^T(\eta - \widehat{\eta}) + z_i) - \rho'(z_i))\widetilde{X}_i^T(\eta - \widehat{\eta}) \\
\geqslant\ & c\frac{1}{n}\sum_{i=1}^n \left(|\widetilde{X}_i^T(\eta - \widehat{\eta})|^2 \wedge |\widetilde{X}_i^T(\eta - \widehat{\eta})|\right)\mathbb{1}\{|z_i| \leqslant 1\} \\
\geqslant\ & c\|\eta - \widehat{\eta}\|^2 \wedge \|\eta - \widehat{\eta}\| \inf_{\|v\|=1} \frac{1}{n}\sum_{i=1}^n (|v^T\widetilde{X}_i|^2 \wedge |v^T\widetilde{X}_i|)\mathbb{1}\{|z_i| \leqslant 1\} \\
\geqslant\ & c_2(1-\epsilon)\left(\|\eta - \widehat{\eta}\|^2 \wedge \|\eta - \widehat{\eta}\|\right). \qquad (47)
\end{aligned}$$

On the other hand,

$$\left|\frac{1}{n}\sum_{i=1}^n \rho'(z_i)\widetilde{X}_i(\eta - \widehat{\eta})\right| \leqslant \|\eta - \widehat{\eta}\| \left\|\frac{1}{n}\sum_{i=1}^n \rho'(z_i)\widetilde{X}_i\right\| \leqslant C\sqrt{\frac{p}{n}}\|\eta - \widehat{\eta}\|, \qquad (48)$$

where the last inequality is again from Lemma 4. Combining (46), (47) and (48), we have $c_2(1-\epsilon)\left(\|\eta - \widehat{\eta}\| \wedge 1\right) \leqslant C\sqrt{\frac{p}{n}}$, which leads to the desired conclusion when $\frac{p}{n(1-\epsilon)^2}$ is small. □

## 6.2 Proof of Main Results

We first state the following two lemmas, whose proofs will be given in Section 6.4.

**Lemma 5.** *Under the setting of Theorem 1, for any constant $\delta \in (0,1)$, there exist some constants $C, c, c' > 0$ depending on $\delta$ such that*

1. $\left\| \frac{1}{n}\sum_{i=1}^n \left(g'(z_i) - \mathbb{E}g'(z_i)\right)\widetilde{x}_i w_i \right\| \leqslant C\sqrt{\frac{p}{n}}$,
2. $\left\| \frac{1}{n}\sum_{i=1}^n |\widetilde{x}_i| w_i w_i^T \right\|_{\mathrm{op}} \leqslant C$,
3. $\frac{1}{n}\sum_{i=1}^n (|\widetilde{x}_i| \wedge \widetilde{x}_i^2)\mathbb{1}\left\{|(\gamma - \widehat{\gamma})^T w_i + z_i| \leqslant 1\right\} \geqslant c'(1-\epsilon)$,
4. $c' \leqslant \frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2 \leqslant C$,

*with probability at least $1-\delta$ as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$.*

**Lemma 6.** *Under the setting of Theorem 1, for any constant $\alpha \in (0,1)$,*

$$\mathbb{P}\left( \left| \frac{\frac{1}{n}\sum_{i=1}^n g(z_i)\widetilde{x}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}} \right| > 1.45\sqrt{\frac{\log(2/\alpha)}{n}} \right) \leqslant 0.97\alpha,$$

*as long as $\frac{p^2}{n(1-\epsilon)^4} \leqslant c$ for some constant $c > 0$ depending on $\alpha$.*

*Proof of Theorem 2.* By Lemma 1, it suffices to bound (36) and (37), which are implied by Lemma 6 and Lemma 2, respectively. Combining the bounds and choosing the $\delta$ in Lemma 1 to be sufficiently small, we have $\frac{|G_n(\beta)|}{\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}} \leqslant 1.5\sqrt{\frac{\log(2/\alpha)}{n}}$ with probability at least $1-\alpha$, which implies coverage. □

*Proof of Theorem 3.* Following the arguments in Section 4.3, it suffices to show (38) holds uniformly over all $\widehat{\beta}$ such that $|\widetilde{\beta} - \beta| > r$. Since the function $G_n(\cdot)$ is monotone, we only need to show (38) for $\widetilde{\beta} \in \{\beta - r, \beta + r\}$. Since $|G_n(\widetilde{\beta})| \geqslant |-G_n(\widetilde{\beta}) + G_n(\beta)| - |G_n(\beta)|$, we need to lower bound $|-G_n(\widetilde{\beta}) + G_n(\beta)|$ and upper bound $|G_n(\beta)|$. Applying Theorem 2 with its $\alpha$ replaced by some small $\delta$ to be determined later, we have $|G_n(\beta)| \leqslant \frac{C_1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}$ with probability at least $1-\delta$. Next, we use (42) followed by Lemma 5, and obtain $|-G_n(\widetilde{\beta}) + G_n(\beta)| \geqslant c_1(1-\epsilon)\left(1 \wedge |\widetilde{\beta} - \beta|\right)$. Therefore, for $\widetilde{\beta} \in \{\beta - r, \beta + r\}$, we have $|G_n(\widetilde{\beta})| \geqslant c_1(1-\epsilon)(1 \wedge r) - \frac{C_1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}$. In order that (38) holds, we need $c_1(1-\epsilon)(1 \wedge r) \geqslant \frac{C_1 + 1.5\sqrt{\log(2/\alpha)}}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2}$. Since $\frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^2$ is bounded by a constant according to Lemma 5, it is sufficient to set $r = \frac{C_2}{\sqrt{n(1-\epsilon)^2}}$ for some sufficiently large constant $C_2 > 0$. In summary, (38) holds for $\widetilde{\beta} \in \{\beta - r, \beta + r\}$ with probability at least $1 - C'\delta$. The proof is complete by setting $\delta = \alpha/C'$. □

*Proof of Theorem 1.* With slight modifications of the proofs of Theorem 2 and Theorem 3, it can be shown that $\beta \in [\widehat{\beta}_L, \widehat{\beta}_R]$ holds with probability at least $1 - 0.99\alpha$ and $\widehat{\beta}_R - \widehat{\beta}_L \leqslant \frac{C}{\sqrt{n(1-\epsilon)^2}}$ holds with probability at least $1 - 0.01\alpha$. Therefore, by union bound, the two events hold simultaneiously with probability at least $1 - \alpha$. □

### 6.3 Proofs of Proposition 1 and Proposition 2

*Proof of Proposition 1.* The coverage property is equivalent to (13), which is implied by Hoeffding's inequality (Lemma 7 stated in Section 6.4). To derive the length guarantee, it suffices to show that for any $|\widetilde{\beta} - \beta| > \frac{C}{2\sqrt{n(1-\epsilon)^2}}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \leqslant \widetilde{\beta}\} < \frac{1}{2} - \sqrt{\frac{\log(2/\alpha)}{2n}} \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i < \widetilde{\beta}\} > \frac{1}{2} + \sqrt{\frac{\log(2/\alpha)}{2n}},$$

and thus $\widehat{\beta}$ does not belong to (12). When $\widetilde{\beta} - \beta > \frac{C}{2\sqrt{n(1-\epsilon)^2}}$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i < \widetilde{\beta}\} &\geqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{y_i < \beta + \frac{C}{2\sqrt{n(1-\epsilon)^2}}\right\} \\
&\geqslant \mathbb{P}_{z\sim(1-\epsilon)N(0,1)+\epsilon Q_0}\left(z < \frac{C}{2\sqrt{n(1-\epsilon)^2}}\right) - \frac{C_1}{\sqrt{n}} \qquad (49) \\
&\geqslant \frac{1}{2} + (1-\epsilon)\mathbb{P}\left(0 < N(0,1) < \frac{C}{2\sqrt{n(1-\epsilon)^2}}\right) - \frac{C_1}{\sqrt{n}} \\
&\geqslant \frac{1}{2} + \sqrt{\frac{\log(2/\alpha)}{2n}},
\end{aligned}$$

where (49) holds with high probability by Hoeffding's inequality and the last inequality holds with a sufficiently large $C > 0$. The other direction $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \leqslant \widetilde{\beta}\} < \frac{1}{2} - \sqrt{\frac{\log(2/\alpha)}{2n}}$ implied by $\widetilde{\beta} - \beta < -\frac{C}{2\sqrt{n(1-\epsilon)^2}}$ is by a symmetric argument. □

*Proof of Proposition 2.* The coverage property follows the same argument in the proof of Lemma 6. For the length guarantee, it suffices to show

$$\frac{\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i\widetilde{\beta} - y_i)x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}} > \sqrt{\frac{2\log(2/\alpha)}{n}}$$

holds for all $|\widetilde{\beta} - \beta| > \frac{C}{2\sqrt{n(1-\epsilon)^2}}$. This can be established by the same argument in the proof of Theorem 3. □

### 6.4 Proofs of Technical Lemmas

We will prove Lemma 1, Lemma 4, Lemma 5 and Lemma 6 in this section. We first state some concentration bounds that will be needed.

**Lemma 7** (Hoeffding (1963)). *Suppose $x_1, \cdots, x_n$ are independent zero-mean random variables such that $a_i \leqslant x_i \leqslant b_i$ for all $i \in [n]$. Then,*

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

*for any $t > 0$.*

**Lemma 8** (Proposition 5.16 of Vershynin (2010)). *Suppose $x_1, \cdots, x_n$ are independent zero-mean sub-exponential random variables such that $\max_{i\in[n]} \mathbb{E}\exp(x_i/K) \leqslant e$ for some constant $K > 0$. Then, there exists some constant $C > 0$ such that*

$$\mathbb{P}\left(\sum_{i=1}^{n} x_i \geqslant t\right) \leqslant \exp\left(-C\left(\frac{t^2}{n} \wedge t\right)\right),$$

*for any $t > 0$.*

**Lemma 9** (Theorem 2.1 of Al-Ghattas et al. (2025)). *Suppose $X_1, \cdots, X_n$ are i.i.d. vectors satisfying Assumption A. For any constant $\delta \in (0,1)$ and $d \in \{2,4\}$, there exists some constant $C > 0$ depending on $\delta$ such that*

$$\sup_{v\in S_{p-1}} \left|\frac{1}{n}\sum_{i=1}^{n}\left((v^T X_i)^d - \mathbb{E}(v^T X_i)^d\right)\right| \leqslant C\left(\sqrt{\frac{p}{n}} + \frac{p^{d/2}}{n}\right),$$

*with probability at least $1-\delta$.*

*Proof of Lemma 1.* The expansion (31) and the arguments in Section 4.2 imply

$$|G_n(\beta)| \leqslant \left|\frac{1}{n}\sum_{i=1}^{n} g(z_i)\,\widetilde{x}_i\right| + \|\widehat{\gamma}-\gamma\|\left\|\frac{1}{n}\sum_{i=1}^{n}\left(g'(z_i) - \mathbb{E}g'(z_i)\right)\widetilde{x}_i w_i\right\| + \frac{\|\widehat{\gamma}-\gamma\|^2}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}|\widetilde{x}_i| w_i w_i^T\right\|_{\mathrm{op}}.$$

This immediately implies the desired bound using Lemma 5. □

*Proof of Lemma 4.* The data generating process (2) can be equivalently described as follows. We first sample $I_i \overset{iid}{\sim} \text{Bernoulli}(1-\epsilon)$. Then, sample $z_i|I_i \sim I_i N(0,1) + (1-I_i)Q$. Define the set $I = \{i \in [n] : I_i = 1, |z_i| \leqslant 0.5\}$ and $n_0 = |I|$. Besides, we define $\breve{x}_i = x_i - \alpha^T w_i$ and $\breve{X}_i^T = (\breve{x}_i, w_i^T)^T$. Since the $(1,1)$ entry of $\Sigma^{-1}$ equals $(\Sigma_{xx} - \Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx})^{-1}$, $\Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx} \leqslant \Sigma_{xx} \leqslant \|\Sigma\|_{\mathrm{op}} \leqslant 2K_1^2$ and

$$\|\alpha\|^2 = \Sigma_{xw}\Sigma_{ww}^{-2}\Sigma_{wx} \leqslant \left\|\Sigma_{ww}^{-\frac{1}{2}}\Sigma_{wx}\right\|^2 \left\|\Sigma_{ww}^{-1}\right\|_{\mathrm{op}} = \Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx}\left\|\Sigma_{ww}^{-1}\right\|_{\mathrm{op}} \leqslant 2K_1^2 K_3.$$

Thus, we can write $\breve{x}_i$ as $u^T X_i$ with $\|u\|$ bounded by some constant. This implies the vector $\breve{X}_i$ also satisfies Assumption A with constants $\breve{K}_1$, $\breve{K}_2$ and $\breve{K}_3$.

By triangle inequality, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)\widetilde{X}_i\right\| \leqslant \left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)\breve{X}_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)(\widetilde{X}_i - \breve{X}_i)\right\|.$$

Since $\breve{X}_i$ is symmetric and is independent of $z_i$, and thus

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)\breve{X}_i\right\|^2 = \sum_{j=1}^{p}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)\breve{X}_{ij}\right)^2 = \frac{1}{n^2}\sum_{j=1}^{p}\sum_{i=1}^{n}\mathbb{E}\left(\rho'(z_i)\breve{X}_{ij}\right)^2 \leqslant C_1\frac{p}{n}, \tag{50}$$

which implies $\left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)\breve{X}_i\right\| \leqslant C_2\sqrt{\frac{p}{n}}$ with high probability by Markov's inequality. By Lemma 3, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)(\widetilde{X}_i - \breve{X}_i)\right\| = \left|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)(\widehat{\alpha}-\alpha)^T w_i\right| \leqslant \|\widehat{\alpha}-\alpha\|\left\|\frac{1}{n}\sum_{i=1}^{n}\rho'(z_i)w_i\right\| \leqslant C_3\sqrt{\frac{p}{n}},$$

with high probability. Combining the above displays, we obtain the first inequality in Lemma 4.

For the second inequality, we note that $n(1-\epsilon)^2$ is large enough when $\frac{p^2}{n(1-\epsilon)^4}$ is small. By Chebyshev's inequality, we have

$$\left|\frac{n_0}{n} - (1-\epsilon)(\Phi(0.5)-\Phi(-0.5))\right| \leqslant C_4\sqrt{\frac{\mathsf{Var}(\mathbb{1}\{|z_i| \leqslant 0.5, I_i = 1\})}{n}} \leqslant \frac{(1-\epsilon)(\Phi(0.5)-\Phi(-0.5))}{2},$$

with high probability, which implies

$$\frac{n_0}{n} \geqslant \frac{1}{2}(1-\epsilon)(\Phi(0.5)-\Phi(-0.5)). \tag{51}$$

For any $v \in S_{p-1}$ and some constant $\kappa$ to be specified later, by Lipschitz continuity we have

$$\begin{aligned}
&\frac{1}{n_0}\sum_{i\in I}|\widetilde{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2 \\
\geqslant\ & \frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2 - \frac{1}{n_0}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right| \\
\geqslant\ & \frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2 - \frac{1}{n_0}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right| - \frac{1}{n_0}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right| 2\sqrt{\left|v^T\check{X}_i\right|} \\
\geqslant\ & \frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2 - \frac{1}{n_0}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right| - \frac{1}{n_0\kappa}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right|^2 - \frac{\kappa}{n_0}\sum_{i\in I}\left|v^T\check{X}_i\right| \\
\geqslant\ & \frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2 - \sqrt{\frac{1}{n_0}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right|^2} - \frac{1}{n_0\kappa}\sum_{i\in I}\left|v^T(\check{X}_i - \widetilde{X}_i)\right|^2 - \kappa\sqrt{\frac{1}{n_0}\sum_{i\in I}\left|v^T\check{X}_i\right|^2}.
\end{aligned} \tag{52}$$

The inequality (52) uses the fact that $f(x) = x^2 \wedge a$ is $\sqrt{2a}$-Lipschitz. Let $\mathcal{V}$ be a covering of $S_{p-1}$ such that for any $v \in S_{p-1}$, there exists some $v' \in \mathcal{V}$ s.t. $\|v - v'\| \leqslant \Delta < 1$. It is known that $\mathcal{V}$ can be constructed such that $|\mathcal{V}| \leqslant \left(\frac{3}{\Delta}\right)^p$. By the same argument used in the above display, we have

$$\begin{aligned}
&\frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2 \\
\geqslant\ & \frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v'| \wedge |\check{X}_i^T v'|^2 - \sqrt{\frac{1}{n_0}\sum_{i\in I}\left|(v-v')^T\check{X}_i\right|^2} - \frac{1}{n_0\kappa}\sum_{i\in I}\left|(v-v')^T\check{X}_i\right|^2 - \kappa\sqrt{\frac{1}{n_0}\sum_{i\in I}\left|v'^T\check{X}_i\right|^2}
\end{aligned}$$

Combining the above two displays, we have

$$\begin{aligned}
&\inf_{\|v\|=1}\frac{1}{n_0}\sum_{i\in I}|\widetilde{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2 \\
\geqslant\ & \inf_{\|v\|\in\mathcal{V}}\frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v'| \wedge |\check{X}_i^T v'|^2 - \sqrt{\frac{1}{n_0}\sum_{i\in I}|(\hat{\alpha}-\alpha)^T w_i|^2} - \frac{1}{n_0\kappa}\sum_{i\in I}|(\hat{\alpha}-\alpha)^T w_i|^2 - \kappa\sqrt{\left\|\check{\Sigma}\right\|_{\mathrm{op}}} \\
& -\Delta\sqrt{\left\|\check{\Sigma}\right\|_{\mathrm{op}}} - \frac{\Delta^2}{\kappa}\left\|\check{\Sigma}\right\|_{\mathrm{op}} - \kappa\sqrt{\left\|\check{\Sigma}\right\|_{\mathrm{op}}},
\end{aligned}$$

where $\check{\Sigma} = \frac{1}{n_0}\sum_{i\in I}\check{X}_i\check{X}_i^T$. By (51), we know that $\frac{p}{n_0} \leqslant \frac{p}{n}\frac{2}{(1-\epsilon)(\Phi(0.5)-\Phi(-0.5))}$ is small when $\frac{p^2}{n(1-\epsilon)^4}$ is small. Since $n_0$ and $I$ are independent of $\{\check{X}_i\}_{i=1}^n$, we have $\left\|\check{\Sigma}\right\|_{\mathrm{op}} \leqslant \left\|\check{\Sigma} - \mathsf{Cov}(\check{X}_i)\right\|_{\mathrm{op}} + \left\|\mathsf{Cov}(\check{X}_i)\right\|_{\mathrm{op}} \leqslant$

$C_5$ by Lemma 9 and the sub-Gaussianity of $\check{X}_i$. By Lemma 3 and Markov's inequality, we have $\frac{1}{n_0}\sum_{i\in I}\left|(\hat{\alpha}-\alpha)^T w_i\right|^2 \leqslant \|\hat{\alpha}-\alpha\|^2 \frac{1}{n_0}\sum_{i\in I}\|w_i\|^2 \leqslant C_6\frac{p^2}{n}$. In addition,

$$\mathbb{E}(|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2) \geqslant \left(\frac{1}{\check{K}_2}\wedge 1\right)\mathbb{E}(|\check{X}_i^T v|^2 \mathbb{1}\{|v^T\check{X}_i| \leqslant \check{K}_2\}) \geqslant \frac{\check{K}_3}{\check{K}_2}\wedge \check{K}_3.$$

Since $|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2$ is sub-exponential, by Lemma 8 and union bound there exists some constant $C_7$ such that

$$\sup_{\|v\|\in\mathcal{V}}\left|\frac{1}{n_0}\sum_{i\in I}|\check{X}_i^T v'| \wedge |\check{X}_i^T v'|^2 - \mathbb{E}(|\check{X}_i^T v| \wedge |\check{X}_i^T v|^2)\right| \leqslant C_7\left(\sqrt{\frac{p\log(1/\Delta)}{n_0}} + \frac{p\log(1/\Delta)}{n_0}\right)$$

with high probability. In the end, choosing $\kappa = \Delta$ to be a sufficiently small constant, then when $\frac{p^2}{n(1-\epsilon)^4}$ is small, there exists some constant $c>0$ such that

$$\inf_{\|v\|=1}\frac{1}{n_0}\sum_{i\in I}|\widetilde{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2 \geqslant c$$

with high probability by combining the above bounds with (51). Using (51) again, we obtain

$$\inf_{\|v\|=1}\frac{1}{n}\sum_{i=1}^n\left(|\widetilde{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2\right)\mathbb{1}\{|z_i|\leqslant 1\} \geqslant \frac{n_0}{n}\inf_{\|v\|=1}\frac{1}{n_0}\sum_{i\in I}|\widetilde{X}_i^T v| \wedge |\widetilde{X}_i^T v|^2 \geqslant c_1(1-\epsilon).$$

The proof is complete. □

*Proof of Lemma 5.* Recall the definitions of $\check{x}_i$ and $\check{X}_i$ in the proof of Lemma 4. By triangle inequality,

$$\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right)\widetilde{x}_i w_i\right\| \leqslant \left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right)\check{x}_i w_i\right\| + \left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right)(\widetilde{x}_i-\check{x}_i) w_i\right\|.$$

Using a similar argument to (50) together with Markov's inequality, we have

$$\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right)\check{x}_i w_i\right\| \leqslant C_1\sqrt{\frac{p}{n}}.$$

Moreover, Cauchy-Schwarz and Lemma 3 imply

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right)(\widetilde{x}_i-\check{x}_i) w_i\right\| \\
\leqslant\ & \|\hat{\alpha}-\alpha\|\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right) w_i w_i^T\right\|_{\mathrm{op}} \\
\leqslant\ & C_2\sqrt{\frac{p}{n}}\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right) w_i w_i^T\right\|_{\mathrm{op}},
\end{aligned}$$

where

$$\left\|\frac{1}{n}\sum_{i=1}^n\left(g'(z_i)-\mathbb{E}g'(z_i)\right) w_i w_i^T\right\|_{\mathrm{op}}$$

$$
\begin{aligned}
&= \sup_{\|v\|=1} \frac{1}{n}\sum_{i=1}^{n}\left(g'(z_i)-\mathbb{E}g'(z_i)\right)(v^T w_i)^2 \\
&\leqslant \frac{1}{2n}\sum_{i=1}^{n}\left(g'(z_i)-\mathbb{E}g'(z_i)\right)^2 + \sup_{\|v\|=1}\frac{1}{2n}\sum_{i=1}^{n}(v^T w_i)^4 \\
&\leqslant C_3,
\end{aligned}
$$

by Lemma 9. Combining the above high-probability bounds, we obtain the first inequality of Lemma 5.

The second inequality follows directly from Lemma 9 and the fourth inequality in Lemma 5 which will be proved later,

$$
\left\|\frac{1}{n}\sum_{i=1}^{n}|\widetilde{x}_i| w_i w_i^T\right\|_{\text{op}} = \sup_{\|v\|=1}\frac{1}{n}\sum_{i=1}^{n}(v^T w_i)^2|\widetilde{x}_i| \leqslant \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2 + \sup_{\|v\|=1}\frac{1}{n}\sum_{i=1}^{n}(v^T w_i)^4\right) \leqslant C_4.
$$

To prove the third inequality of Lemma 5, we first claim that

$$
\max_{1\leqslant i\leqslant n}\|w_i\| \leqslant C_5(\sqrt{p}+\sqrt{\log n}) \tag{53}
$$

holds with high probability. To show (53), recall the $\Delta$-net $\mathcal{V}$ used in the proof of Lemma 4. With $\Delta$ being a sufficiently small constant, we have $\max_{1\leqslant i\leqslant n}\|w_i\| \leqslant 2\max_{1\leqslant i\leqslant n}\max_{v\in\mathcal{V}}|v^T w_i|$, and (53) follows by a union bound argument with Assumption A. Combining (53) with Lemma 2, we have

$$
\max_{1\leqslant i\leqslant n}|(\gamma-\widehat{\gamma})^T w_i| \leqslant \|\gamma-\widehat{\gamma}\|\max_{1\leqslant i\leqslant n}\|w_i\| \leqslant C_6\sqrt{\frac{p}{n(1-\epsilon)^2}}(\sqrt{p}+\sqrt{\log n}) \leqslant \frac{1}{2}. \tag{54}
$$

The event (54) implies

$$
\frac{1}{n}\sum_{i=1}^{n}(|\widetilde{x}_i|\wedge\widetilde{x}_i^2)\mathbb{1}\left\{|(\gamma-\widehat{\gamma})^T w_i + z_i| \leqslant 1\right\} \geqslant \frac{1}{n}\sum_{i=1}^{n}(|\widetilde{x}_i|\wedge\widetilde{x}_i^2)\mathbb{1}\left\{|z_i|\leqslant 0.5\right\},
$$

which is lower bounded by $c(1-\epsilon)$ following the same analysis leading to the second inequality of Lemma 4.

For the last inequality of Lemma 5, we first use triangle inequality and Cauchy-Schwarz, and

$$
\left|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\breve{x}_i^2}\right| \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i-\breve{x}_i)^2} \leqslant \|\alpha-\hat{\alpha}\|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|w_i\|^2},
$$

which is at most $C_7\sqrt{\frac{p^2}{n}}$ by Lemma 3. Under Assumption A, it is straightforward to check that $\frac{1}{n}\sum_{i=1}^{n}\breve{x}_i^2 \asymp 1$. The proof is complete. □

*Proof of Lemma 6.* Let $\xi_i \overset{i.i.d}{\sim}$ Rademacher and define $X_i^s = X_i\xi_i$. Since $X_i$ is symmetric, for any measurable set $A$,

$$
\mathbb{P}(X_i^s\in A|\xi_i=1) = \mathbb{P}(X_i\in A) = \mathbb{P}(-X_i\in A) = \mathbb{P}(X_i^s\in A|\xi_i=-1),
$$

which implies $X_i^s$'s are independent of $\xi_i$'s. Using the notation $\widetilde{x}_i^s = \xi_i\widetilde{x}_i$, we have

$$
\frac{\frac{1}{n}\sum_{i=1}^{n}g(z_i)\widetilde{x}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widetilde{x}_i^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\xi_i g(z_i)\widetilde{x}_i^s}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{x}_i^s)^2}}.
$$

Define $u_i = \frac{g(z_i)\widetilde{x}_i^s}{\sqrt{\frac{1}{n}\sum_{i=1}^n (\widetilde{x}_i^s)^2}}$, and then by Lemma 7, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i u_i\right| \geqslant t \middle| u\right) \leqslant 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n u_i^2}\right) \leqslant 2\exp\left(-\frac{t^2}{2n}\right).$$

for any $t > 0$. The same bound holds for $\mathbb{P}\left(|\sum_{i=1}^n \xi_i u_i| \geqslant t\right)$ as well. By verifying that $2\exp\left(-\frac{t^2}{2n}\right) \leqslant 0.97\alpha$ with $t = 1.45\sqrt{n\log\frac{2}{\alpha}}$, the proof is complete. □

## References


Al-Ghattas, O., Chen, J., and Sanz-Alonso, D. (2025). Sharp concentration of simple random tensors. *arXiv preprint arXiv:2502.16916*.

Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568.

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

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

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, pages 28–48.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646.

Chang, W. and Kuchibhotla, A. K. (2024). Confidence sets for $z$-estimation problems using self-normalization. *arXiv preprint arXiv:2407.12278*.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969.

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

d'Orsi, T., Novikov, G., and Steurer, D. (2021). Consistent regression when oblivious outliers overwhelm. In *International Conference on Machine Learning*, pages 2297–2306. PMLR.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

Freedman, D. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228.

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

Gao, C. and Lafferty, J. (2020). Model repair: Robust recovery of over-parameterized statistical models. *arXiv preprint arXiv:2005.09912*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Karoui, N. E. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*.

Kuchibhotla, A. K., Balakrishnan, S., and Wasserman, L. (2024). The hulc: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):586–622.

Lei, L. and Bickel, P. J. (2021). An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412.

Lei, L., Bickel, P. J., and El Karoui, N. (2018). Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields*, 172:983–1079.

Luo, Y. and Gao, C. (2024). Adaptive robust confidence intervals. *arXiv preprint arXiv:2410.22647*.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, 17(1):382–400.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285.

Park, B., Balakrishnan, S., and Wasserman, L. (2023). Robust universal inference. *arXiv preprint arXiv:2307.04034*.

Portnoy, S. (1984). Asymptotic behavior of $m$-estimators of $p$ regression parameters when $p^2/n$ is large. i. consistency. *The Annals of Statistics*, 12(4):1298–1309.

Portnoy, S. (1985). Asymptotic behavior of $m$ estimators of $p$ regression parameters when $p^2/n$ is large; ii. normal approximation. *The Annals of Statistics*, 13(4):1403–1417.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026.

Scheffe, H. and Tukey, J. W. (1945). Non-parametric estimation. i. validation of order statistics. *The Annals of Mathematical Statistics*, 16(2):187–192.

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

Tsakonas, E., Jaldén, J., Sidiropoulos, N. D., and Ottersten, B. (2014). Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Verzelen, N. and Gassiat, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710.

Wen, K., Wang, T., and Wang, Y. (2025). Residual permutation test for regression coefficient testing. *The Annals of Statistics*, 53(2):724–748.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of m-estimators for the linear model. *The Annals of Statistics*, 7(2):258–268.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242.

Zhu, Y. and Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600.